School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

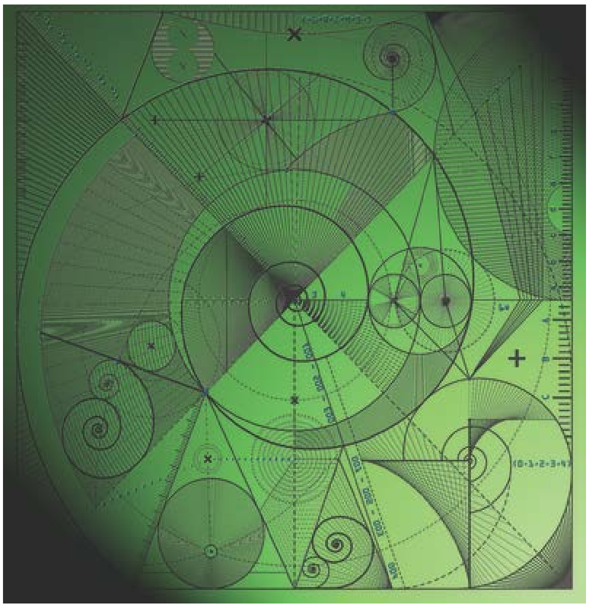
Preprint MPS-2015-06

1 April 2015

Bayesian model comparison with intractable likelihoods

by

Richard G. Everitt, Adam M. Johansen, Ellen Rowing and Melina Evdemon-Hogan



Bayesian model comparison with intractable likelihoods

Richard G. Everitt, Adam M. Johansen, Ellen Rowing and Melina Evdemon-Hogan

1st April 2015

Abstract

Markov random field models are used widely in computer science, statistical physics and spatial statistics and network analysis. However, Bayesian analysis of these models using standard Monte Carlo methods is not possible due to their intractable likelihood functions. Several methods have been developed that permit exact, or close to exact, simulation from the posterior distribution. However, estimating the evidence and Bayes' factors (BFs) for these models remains challenging in general. This paper describes new random weight importance sampling and sequential Monte Carlo methods for estimating BFs that use simulation to circumvent the evaluation of the intractable likelihood, and compares them to existing methods. In some cases we observe an advantage in the use of *biased* weight estimates; an initial investigation into the theoretical and empirical properties of this class of methods is presented.

1 Introduction

There has been much recent interest in performing Bayesian inference in models where the posterior is intractable. Specifically, we have the situation where the posterior distribution $\pi(\theta|y) \propto p(\theta)f(y|\theta)$, cannot be evaluated pointwise. In most applications this intractability occurs due to the intractability of the likelihood, i.e. $f(y|\theta)$ cannot be evaluated pointwise. Examples of cases where this situation arises include:

1. the use of big data sets, where $f(y|\theta)$ consists of a product of a large number of terms;

2. the existence of a large number of latent variables x, so that $f(y|\theta)$ is known only as a high dimensional integral $f(y|\theta) = \int_x f(y,x|\theta)dx$;

3. when $f(y|\theta) = \frac{1}{Z(\theta)}\gamma(y|\theta)$, with $Z(\theta)$ being an intractable normalising constant (INC) for the tractable term $\gamma(y|\theta)$ (e.g. when f factorises as a Markov random field);

4. where it is possible to sample from $f(\cdot|\theta)$, but not to evaluate it, such as when the distribution of the data given θ is modelled by a complex stochastic computer model.

Each of these (overlapping) situations has been considered in some detail in previous work and each has inspired different methodologies. The feature common to every approach is that some approximation is introduced into either the likelihood under consideration, or the Monte Carlo simulation algorithm used to simulate from the posterior.

In this paper we focus on the third case, in which the likelihood has an INC. This is an important problem in its own right (Girolami *et al.* (2013) refer to it as "one of the main challenges to methodology for computational statistics currently"). There exist several competing methodologies for inference in this setting (see Everitt (2012)). In particular, the *"exact" approaches* of Møller *et al.* (2006) and Murray *et al.* (2006) exploit the decomposition $f(y|\theta) = \frac{1}{Z(\theta)}\gamma(y|\theta)$, whereas *"simulation based" methods* such as approximate Bayesian computation (ABC) (Grelaud *et al.*, 2009) do not depend upon such a decomposition and can be applied more generally: to situation 1 in Picchini and Forman (2013); situations 2 and 3 (e.g. Everitt (2012)) and situation 4 (e.g. Wilkinson (2013)).

This paper considers the problem of Bayesian model comparison in the presence of an INC. We explore both exact and simulation-based methods, and find that elements of both approaches may also be more generally applicable. Specifically:

- For exact methods we find that approximations are required to allow practical implementation, and this leads us to investigate the use of approximate weights in importance sampling (IS) and sequential Monte Carlo (SMC). We examine the use of both *exact-approximate* approaches (as in Fearnhead *et al.* (2010)) and also "*inexact-approximate*" methods, in which complete flexibility is allowed in the approximation of weights, at the cost of losing the exactness of the method. This work is a natural counterpart to Alquier *et al.* (2014), which examines the analogue of this question (concerning the acceptance probability) for Markov chain Monte Carlo (MCMC) algorithms. These generally applicable methods, "noisy MCMC" (Alquier *et al.*, 2014) and "noisy SMC" (this paper) have some potential to address situations 1-3.

- One of the simulation based methods that we consider is "synthetic likelihood" (SL) (Wood, 2010). In the applications considered here we find this to be a viable alternative to ABC. Our results are suggestive that this, and related methods, may find success in scenarios in which ABC is more usually applied.

In the remainder of this section we briefly outline the problem, and outline methods for, parameter inference in the presence of an INC. We then outline the problem of Bayesian model comparison, before discussing methods for tackling this in the rest of the paper.

1.1 Parameter inference

In this section we consider the problem of simulating from $\pi(\theta|y) \propto p(\theta)\gamma(y|\theta)/Z(\theta)$ using MCMC. This problem has been well studied, and such models are termed *doubly intractable* because the acceptance probability in the Metropolis-Hastings (MH) algorithm

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

cannot be evaluated due to the presence of the INC. We first review exact methods for simulating from such a target in sections 1.1.1, 1.1.2 and 1.1.3, before looking at simulation-based methods in sections 1.1.4 and 1.1.5. The methods described here in the context of MCMC form the basis for the methods for evidence estimation we develop in the rest of the paper.

1.1.1 Single and multiple auxiliary variable methods

Møller *et al.* (2006) avoid the evaluation of the INC by augmenting the target distribution with an extra variable u that lies on the same space as y, and use an MH algorithm with target distribution

$$\pi(\theta, u|y) \propto q_u(u|\theta, y) f(y|\theta) p(\theta), \tag{2}$$

where q_u is some (normalised) arbitrary distribution. As the MH proposal in (θ, u)-space they use

$$(\theta^*, u^*) \sim f(u^*|\theta^*) q(\theta^*|\theta), \tag{3}$$

giving an acceptance probability of

$$\min \left\{ 1, \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \frac{p(\theta^*)}{p(\theta)} \frac{\gamma(y|\theta^*)}{\gamma(y|\theta)} \frac{q_u(u^*|\theta^*, y)}{\gamma(u^*|\theta^*)} \frac{\gamma(u|\theta)}{q_u(u|\theta, y)} \right\}. \tag{4}$$

Note that, by viewing $q_u(u^*|\theta^*, y)/\gamma(u^*|\theta^*)$ as an unbiased IS estimator of $1/Z(\theta^*)$, this algorithm can be seen as an instance of the *exact-approximations* described in Beaumont (2003) and Andrieu and Roberts (2009), where it is established that if an unbiased estimator of a target density is used appropriately in an MH algorithm, the θ-marginal of the invariant distribution of this chain is the target distribution of interest. This automatically suggests extensions to the *single auxiliary variable (SAV)* method described above, where M importance points are used to instead give the estimate

$$\widehat{\frac{1}{Z(\theta)}} = \frac{1}{M} \sum_{m=1}^{M} \frac{q_u(u^{(m)}|\theta, y)}{\gamma(u^{(m)}|\theta)}. \tag{5}$$

Andrieu and Vihola (2012) show that the reduced variance of this estimator leads to a reduced asymptotic variance of estimators from the resultant Markov chain. The variance of the IS estimator is strongly dependent on an appropriate choice of IS target $q_u(\cdot|\theta, y)$, which should ideally have lighter tails than $f(\cdot|\theta)$. Møller *et al.* (2006) suggest that a reasonable choice may be $q_u(\cdot|\theta, y) = f(\cdot|\widehat{\theta})$, where $\widehat{\theta}$ is the maximum likelihood estimator of θ. However, in practice $q_u(\cdot|\theta, y)$ can be difficult to choose well, particularly when y lies on a high dimensional space. Motivated by this, annealed importance sampling (AIS) (Neal, 2001) can be used as an alternative to IS, as suggested in Murray *et al.* (2006) (who names this approach the *multiple auxiliary variable (MAV)* method). AIS makes use of a sequence of K targets, which in Murray *et al.* (2006) are chosen to be

$$f_k(\cdot|\theta, \widehat{\theta}, y) \propto \gamma_k(\cdot|\theta, \widehat{\theta}, y) = \gamma(\cdot|\theta)^{(K+1-k)/(K+1)} + q_u(\cdot|\theta, y)^{k/(K+1)}, \tag{6}$$

between $f(\cdot|\theta)$ and $q_u(\cdot|\theta, y)$. After the initial draw $u_{K+1} \sim f(\cdot|\theta)$, the auxiliary point is taken through a sequence of K MCMC moves which successively have target $f_k(\cdot|\theta, \widehat{\theta}, y)$ for $k = K : 1$. The resultant IS estimator is given by

$$\widehat{\frac{1}{Z(\theta)}} = \frac{1}{M} \sum_{m=1}^{M} \prod_{k=0}^{K} \frac{\gamma_{k+1}(u_k^{(m)}|\theta, \widehat{\theta}, y)}{\gamma_k(u_k^{(m)}|\theta, \widehat{\theta}, y)}. \tag{7}$$

This estimator has a lower variance (although at a higher computational cost) than the corresponding IS estimator.

1.1.2 Exchange algorithms

An alternative approach to avoiding the ratio of INCs in equation 1 is given by Murray *et al.* (2006), in which it is suggested to use the acceptance probability

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

where $u \sim f(\cdot|\theta^*)$, motivated by the intuitive idea that $\gamma(u|\theta)/\gamma(u|\theta^*)$ is a single point IS estimator of $Z(\theta)/Z(\theta^*)$. This method is shown to have the correct invariant distribution, as is the extension in which AIS is used in place of IS. A potential extension might seem to be using multiple importance points $\{u^{(m)}\}_{m=1}^{M} \sim f(\cdot|\theta^*)$ to obtain an estimator of $Z(\theta)/Z(\theta^*)$ that has a smaller variance, with the aim of improving the statistical efficiency of estimators based on the resultant Markov chain. This scheme is shown to work well empirically in Alquier *et al.* (2014). However, this chain does not have the desired target as its invariant distribution. Instead it can be seen as part of a wider class of algorithms that use a noisy estimate of the acceptance probability: *noisy Monte Carlo* algorithms (also referred to as *"inexact-approximations"* in Girolami *et al.* (2013)). Alquier *et al.* (2014) shows that under uniform ergodicity of the ideal chain, a bound on the expected difference between the noisy and true acceptance probabilities can lead to bounds on the distance between the desired target distribution and the iterated noisy kernel. It also describes additional noisy MCMC algorithms for approximately simulating from the posterior, based on Langevin dynamics.

1.1.3 Russian Roulette and other approaches

Girolami *et al.* (2013) use series-based approximations to intractable target distributions within the exact-approximation framework, where "Russian Roulette" methods from the physics literature are used to ensure the unbiasedness of truncations of infinite sums. These methods do not require exact simulation from $f(\cdot|\theta^*)$, as is do the SAV and exchange approaches described in the previous two sections. However, SAV and exchange are often implemented in practice by generating the auxiliary variables by taking the final point of a long "internal" MCMC run in place of exact simulation (e.g Caimo and Friel (2011)). For finite runs of the internal MCMC, this approach will not have exactly the desired invariant distribution, but Everitt (2012) shows that under regularity conditions the bias introduced by this approximation tends to zero as the run length of the internal MCMC increases: the same proof holds for the use of an MCMC chain for the simulation within an ABC-MCMC or SL-MCMC algorithm, as described in sections 1.1.4 and 1.1.5. Although the approach of Girolami *et al.* (2013) is exact, they comment that it is significantly more computationally expensive than this approximate approach. For this reason, we do not pursue Russian Roulette approaches further in this paper.

When a rejection sampler is available for simulating from $f(\cdot|\theta^*)$, Rao *et al.* (2013) introduce an alternative exact algorithm that has some favourable properties compared to the exchange algorithm. Since a rejection sampler is not available in many cases, we do not pursue this approach further.

1.1.4 Approximate Bayesian computation

ABC (Tavaré *et al.*, 1997) refers to methods that aim to approximate an intractable likelihood $f(y|\theta)$ through the integral

$$\widetilde{f}(S(y)|\theta) = \int \pi_\epsilon \left(S(u)|S(y)\right) f\left(u|\theta\right) \left|\frac{\partial S_i}{\partial u_j}(u)\right| du, \tag{8}$$

where $S(\cdot)$ gives a vector of summary statistics, $|\partial S_i/\partial u_j(u)|$ denotes the Jacobian determinant arising from the change of variable, and $\pi_\epsilon\left(\cdot|S(y)\right)$ is a density centred at $S(y)$ with bandwidth ϵ. As $\epsilon \rightarrow 0$, this distribution becomes more concentrated around $S(y)$, so that in the case where $S(\cdot)$ gives sufficient statistics for estimating θ, as $\epsilon \rightarrow 0$ the approximate posterior becomes closer to the true posterior. This approximation is used within standard Monte Carlo methods for simulating from the posterior. For example, it may be used within an MCMC algorithm (known as ABC-MCMC (Marjoram *et al.*, 2003)), where using an exact-approximation argument it can be seen that it is sufficient in the calculation of the acceptance probability to use the Monte Carlo approximation

$$\widehat{f}_\epsilon(S(y)|\theta^*) = \frac{1}{M}\sum_{m=1}^{M} \pi_\epsilon\left(S\left(u^{(m)}\right)|S(y)\right) \tag{9}$$

for the likelihood at θ^* at each iteration, where $\{u^{(m)}\}_{m=1}^{M} \sim f(\cdot|\theta^*)$. Whilst the exact-approximation argument means that there is no additional bias due to this Monte Carlo approximation, the approximation introduced through using a tolerance $\epsilon > 0$ or insufficient summary statistics may be large. For this reason it might be considered a last resort to use ABC on likelihoods with an INC, but previous success on these models (e.g Grelaud *et al.* (2009) and Everitt (2012)) lead us to consider them further in this paper.

1.1.5 Synthetic likelihood

ABC is essentially using, based on simulations from f, a nonparameteric estimator of $f_S\left(S|\theta\right)$, the distribution of the summary statistics of the data given θ. In some situations, a parametric model might be more appropriate. For example, it might be that the statistic consists of the sum of independent random variables, in which case a Central Limit Theorem (CLT) might imply that it would be appropriate to assume that $f_S\left(S|\theta\right)$ is multivariate Gaussian.

The SL approach (Wood, 2010) proceeds by making exactly this Gaussian assumption and uses this approximate likelihood within an MCMC algorithm (SL-MCMC), The parameters (the mean and variance) of this approximating distribution for a given θ^* are estimated based on the summary statistics of simulations $\{u^{(m)}\}_{m=1}^{M} \sim f(\cdot|\theta^*)$. Concretely, the estimate of the likelihood is

$$\widehat{f}_{\text{SL}}\left(S(y)|\theta\right) = \mathcal{N}\left(S(y); \widehat{\mu}_\theta, \widehat{\Sigma}_\theta\right), \tag{10}$$

where

$$\widehat{\mu}_\theta = \frac{1}{M}\sum_{m=1}^{M} S\left(u^{(m)}\right), \tag{11}$$

$$\widehat{\Sigma}_\theta = \frac{ss^T}{M-1}, \tag{12}$$

with $s = \left(S\left(u_1\right) - \widehat{\mu}_\theta, ..., S\left(u_M\right) - \widehat{\mu}_\theta\right)$. Wood (2010) applies this method in a setting where the summary statistics are regression coefficients, motivated by their distribution being approximately normal. One of the approximations inherent in this method, as in ABC, is the use of summary statistics rather than the whole dataset. However, unlike ABC, there is no need to choose a bandwidth ϵ: this approximation is replaced with that arising from the discrepancy between the normal approximation and the exact distribution of the chosen summary statistic. An additional approximation in SL arises since, even if the summary statistic distribution were truly Gaussian, $\widehat{f}_{\text{SL}}$ is not an unbiased estimate of it, so the exact-approximation results do not apply. Rather, this is a special case of noisy MCMC, and we do not expect the additional bias introduced by estimating the parameters of $\widehat{f}_{\text{SL}}$ to have large effects on the results, even if the parameters are estimated via an internal MCMC chain targeting $f(\cdot|\theta^*)$ as described in section 1.1.3.

SL is related to a number of other simulation based algorithms under the umbrella of Bayesian indirect inference (Drovandi and Pettitt, 2013). This suggests a number of extensions to some of the methods presented in this paper that we do not explore here.

1.2 Bayesian model comparison

The main focus of this paper is estimating the *marginal likelihood* (aka the *evidence*)

$$p(y) = \int p(\theta)f(y|\theta)d\theta$$

and *Bayes' factors*: ratios of evidences for different models (M_1 and M_2, say)

$$\text{BF}_{12} = \frac{p(y|M_1)}{p(y|M_2)}.$$

These quantities cannot usually be estimated reliably from MCMC output, and commonly used methods for estimating them require $f(y|\theta)$ to be tractable in θ. This leads Friel (2013) to label their estimation as *"triply intractable"* when f has an INC. To our knowledge the only published approach to estimating the evidence for such models is in Friel (2013), with this paper also giving one of the only approaches to estimating BFs in this setting. For estimating BFs, ABC provides a viable alternative (Grelaud *et al.*, 2009), as long as the approximations in this approach described in section 1.1.4 are not too large.

The methods in Friel (2013) are based on Chib's approximation,

$$\widehat{p}(y) = \frac{f(y|\widetilde{\theta})p(\widetilde{\theta})}{\widehat{\pi}(\widetilde{\theta}|y)}, \tag{13}$$

where $\widetilde{\theta}$ can be an arbitrary value of θ and $\widehat{\pi}$ is an approximation to the posterior distribution. Such an approximation is intractable when f has an INC. Friel (2013) devises a "population" version of the exchange algorithm that simulates points $\theta^{(p)}$ from the posterior distribution, and which also gives an estimate $\widehat{Z}(\theta^{(p)})$ of the INC at each of these points. The points $\theta^{(p)}$ can be used to find a kernel density approximation $\widehat{\pi}$, and estimates $\widehat{Z}(\theta^{(p)})$ of the INC. These are then used in a number of evaluations of equation 13 at points (generated by the population exchange algorithm) in a region of high posterior density, which are then averaged to find an estimate of the evidence. This method has a number of useful properties (including that it may be a more efficient approach for parameter inference than the standard exchange algorithm), but for evidence estimation it suffers the limitation of using a kernel density estimate which means that, as noted in the paper, its use is limited to low-dimensional parameter spaces.

In this paper we explore the alternative approach of methods based on IS, making use of the likelihood approximations described earlier in this section. These IS methods are outlined in section 2. However, IS is also not readily applicable to high dimensional parameter spaces, so in section 3 we look at extensions to the IS methods based on SMC. In section 2 we note the good empirical performance of an inexact-approximate method; such approaches are examined in more detail in section 4. We note that all of the algorithms presented later in the paper are viable alternatives to the MCMC approaches to parameter estimation described in this section, and may outperform the corresponding MCMC approach in some cases. In particular they all automatically make use of a population of points, an idea previously explored in the MCMC context by Caimo and Friel (2011) and Friel (2013). In section 5 we draw conclusions.

2 Importance sampling approaches

In this section we investigate the use of IS for estimating the evidence and BFs for models with INCs. We consider an "ideal" importance sampler that simulates P points $\left\{\theta^{(p)}\right\}_{p=1}^{P}$ from a proposal $q(\cdot)$ and calculates their weight, in the presence of an INC, using

$$\widetilde{w}^{(p)} = \frac{p(\theta^{(p)})\gamma(y|\theta^{(p)})}{q(\theta^{(p)})Z(\theta^{(p)})}, \tag{14}$$

with an estimate of the evidence given by

$$\widehat{p}(y) = \frac{1}{P} \sum_{p=1}^{P} \widetilde{w}^{(p)}. \tag{15}$$

To estimate a BF we simply take the ratio of estimates of the evidence for the two models under consideration. However, the presence of the INC in the weight expression in equation 14 means that importance samplers cannot be directly implemented for these models. To circumvent this problem we will investigate the use of the techniques described in section 1.1 in importance sampling. We begin by looking at exact-approximation based methods in section 2.1. We then examine the use to approximate likelihoods based on simulation, including ABC and SL in section 2.2, before looking at the performance of all of these methods on a toy example in section 2.3. Finally, in sections 2.4 and 2.5 we examine applications to exponential random graph models (ERGMs) and Ising models, the latter of which leads us to consider the use of inexact-approximations in IS.

2.1 Auxiliary variable IS

To avoid the evaluation of the INC in equation 14, we propose the use of the auxiliary variable method used in the MCMC context in section 1.1.1. Specifically, consider IS using the SAV target

$$p(\theta, u|y) \propto q_u(u|\theta, y) f(y|\theta) p(\theta),$$

noting that it has the same evidence as $p(\theta|y)$, with proposal

$$q(\theta, u) = f(u|\theta) q(\theta).$$

This results in weights

$$\begin{aligned} \widetilde{w}^{(p)} &= \frac{q_u(u|\theta^{(p)}, y)\gamma(y|\theta^{(p)})p(\theta^{(p)})}{\gamma(u|\theta^{(p)})q(\theta^{(p)})} \frac{Z(\theta^{(p)})}{Z(\theta^{(p)})} \\ &= \frac{\gamma(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})} \frac{q_u(u|\theta^{(p)}, y)}{\gamma(u|\theta^{(p)})}, \end{aligned}$$

which can be used in equation 15 to obtain an estimate of the evidence.

In this method, which we will refer to as single auxiliary variable IS (SAVIS), we may view $q_u(u|\theta^{(p)}, y)/\gamma(u|\theta^{(p)})$ as an unbiased importance sampling (IS) estimator of $1/Z(\theta^{(p)})$. Although we are using an unbiased estimator of the weights in place of the ideal weights, the result is still an exact importance sampler. SAVIS is an exact-approximate IS method, as seen previously in Fearnhead *et al.* (2010), Chopin *et al.* (2013) and Tran *et al.* (2013). As in the MCMC setting, to ensure the variance of estimators produced by this scheme is not large we must ensure the variance of estimator of $1/Z(\theta^{(p)})$ is small. Thus in practice we found extensions to this basic algorithm were useful: using multiple u importance points for each proposed $\theta^{(p)}$ as in equation 5; and using AIS, rather than simple IS, for estimating $1/Z(\theta^{(p)})$ as in equation 7 (giving an algorithm that we refer to as multiple auxiliary variable IS (MAVIS), in common with the terminology in Murray *et al.* (2006)). Using $q_u(\cdot|\theta, y) = f(\cdot|\widehat{\theta})$, as described in section 1.1.1, and γ_k as in equation 6, we obtain

$$\widehat{\frac{1}{Z(\theta)}} = \frac{1}{Z(\widehat{\theta})} \frac{1}{M} \sum_{m=1}^{M} \prod_{k=0}^{K} \frac{\gamma_{k+1}(u_k^{(m)}|\theta^*, \theta, y)}{\gamma_k(u_k^{(m)}|\theta^*, \theta, y)}. \tag{16}$$

In this case the (A)IS methods are being used as unbiased estimators of the ratio $Z(\widehat{\theta})/Z(\theta)$.

2.2 Simulation based methods

Didelot *et al.* (2011) investigate the use of the ABC approximation when using IS for estimating marginal likelihoods. In this case the weight equation becomes

$$\widetilde{w}^{(p)} = \frac{p(\theta^{(p)}) \frac{1}{R} \sum_{r=1}^{R} \pi_\epsilon(S(x_r^{(p)})|S(y))}{q(\theta^{(p)})},$$

where $\left\{x_r^{(p)}\right\}_{r=1}^{R} \sim f(\cdot|\theta^{(p)})$, and using the notation from section 1.1.4. However, using these weights within equation 15 gives an estimate for $p(S(y))$ rather than, as desired, an estimate of the evidence $p(y)$. The only way to obtain an estimate of the evidence from ABC is to use the full data, rather than taking summary statistics.

Fortunately, there are cases in which ABC may be used to estimate BFs. Didelot *et al.* (2011) establishes that, for the BF for two exponential family models: if $S_1(y)$ is sufficient for the parameters in model 1 and $S_2(y)$ is sufficient for the parameters in model 2, then using $S(y) = (S_1(y), S_2(y))$ gives

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

Outside the exponential family, making an appropriate choice of summary statistics can be more involved (Robert *et al.*, 2011; Prangle *et al.*, 2013; Marin *et al.*, 2013).

Just as in the parameter estimation case, the use of a tolerance $\epsilon > 0$ results in estimating an approximation to the true BF. An alternative approximation, not previously used in model comparison, is to use SL (as described in section 1.1.5). In this case the weight equation becomes

$$\widetilde{w}^{(p)} = \frac{p(\theta^{(p)}) \mathcal{N}\left(S(y); \widehat{\mu}_{\theta^{(p)}}, \widehat{\Sigma}_{\theta^{(p)}}\right)}{q(\theta^{(p)})},$$

where $\widehat{\mu}_\theta, \widehat{\Sigma}_\theta$ are given by equations 11 and 12. As in parameter estimation, this approximation is only appropriate if the normality assumption is reasonable. The issue of the choice of summary statistics is the same as in the ABC case.

2.3 Toy example

In this section we have introduced three alternative methods for estimating BFs: MAVIS, ABC and SL. To further understand their properties we now investigate the performance of each method on a toy example.

Consider i.i.d. observations $y = \{y_i\}_{i=1}^{n=100}$ of a discrete random variable that takes values in $\mathbb{N}$. For such a dataset, we will find the BF for the models

1. $y|\theta \sim \text{Poisson}(\theta)$, $\theta \sim \text{Exp}(1)$

$$\begin{aligned} f_1\left(\{y_i\}_{i=1}^{n}|\theta\right) &= \prod_{i=1}^{n} \frac{\lambda^{y_i}\exp(-\lambda)}{y_i!} \\ &= \frac{1}{\exp(n\lambda)} \prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!} \end{aligned}$$

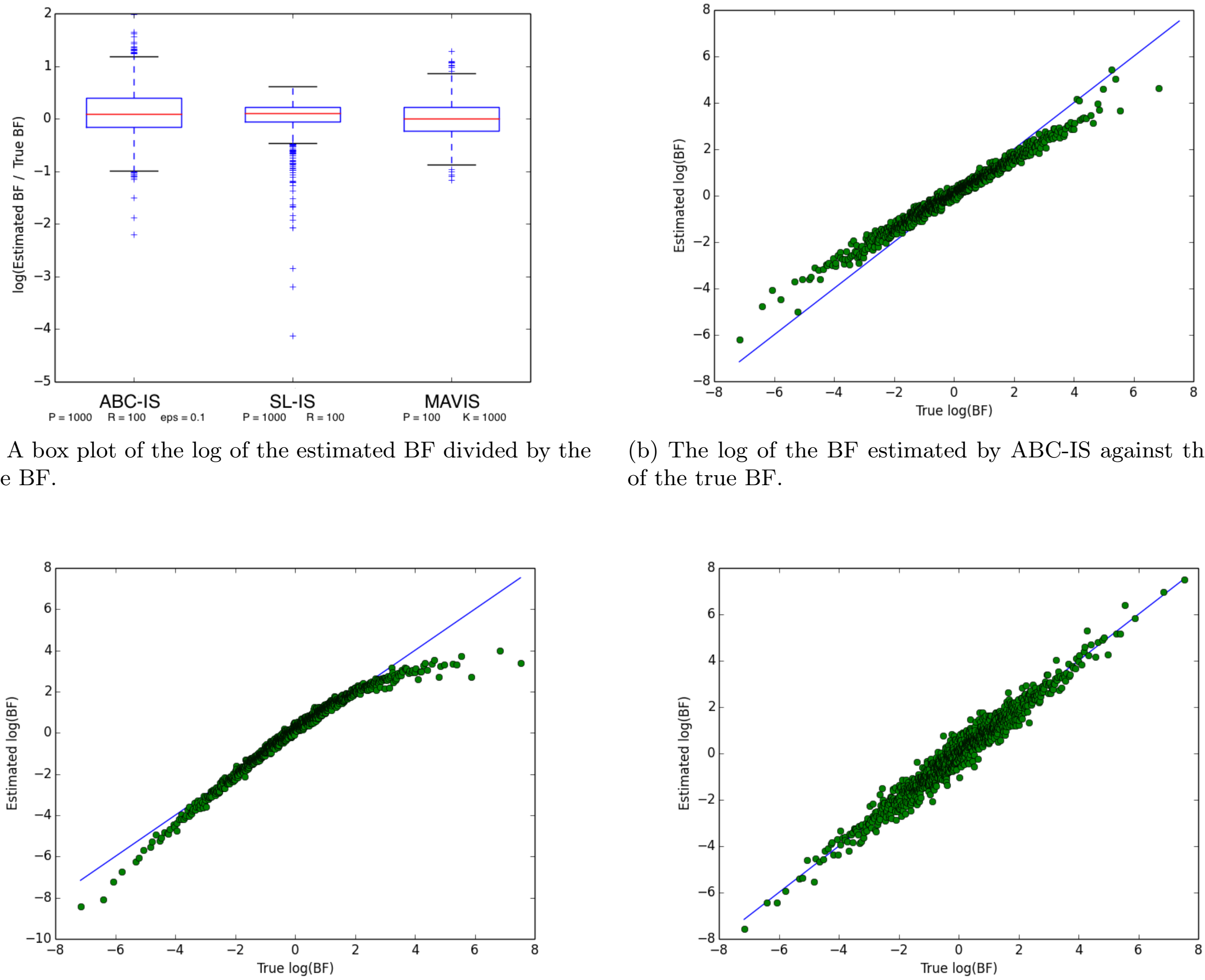
2. $y|\theta \sim \text{Geometric}(\theta)$, $\theta \sim \text{Unif}(0,1)$

$$\begin{aligned} f_2\left(\{y_i\}_{i=1}^{n}|\theta\right) &= \prod_{i=1}^{n} p(1-p)^{y_i} \\ &= \frac{1}{p^{-n}} \prod_{i=1}^{n} (1-p)^{y_i}. \end{aligned}$$

In both cases we have rewritten the likelihoods f_1 and f_2 in the form $\gamma(y|\theta)/Z(\theta)$ in order to use MAVIS. Due to the use of conjugate priors the BF for these two models can be found analytically. As in Didelot *et al.* (2011) we simulated (using an approximate rejection sampling scheme) 1000 datasets for which $\frac{p(y|M_1)}{p(y|M_1)+p(y|M_2)}$ roughly uniformly cover the interval [0.01,0.99], to ensure that testing is performed in a wide range of scenarios. For each algorithm we used the same computational effort, in terms of the number of simulations (10^5) from the likelihood.

Our results are shown in figure 1, with the algorithm-specific parameters being given in figure 1a. We note that we achieved better results for MAVIS when: devoting more computational effort to the estimation of $1/Z(\theta)$ (thus we used only 100 importance points in θ-space, compared to 1000 for the other algorithms); and using more intermediate bridging distributions in the AIS, rather than multiple importance points (thus, in equation 16 we used $K = 1000$ and $M = 1$). In the ABC case we found that reducing ϵ much further than 0.1 resulted in many importance points with zero weight. From the box plots in figure 1a, we might infer that overall SL has outperformed the other methods, but be concerned about the number of outliers. Figures 1b to 1d shed more light on the situations in which each algorithm performs well.

In figure 1b we observe that the non-zero ϵ results in a bias in the BF estimates (represented by the shallower slope in the estimated BFs compared to the true values). In this example we conclude that ABC has worked quite well, since the bias is only pronounced in situations where the true BF favours one model strongly over the other, and this conclusion would not be affected



(a) A box plot of the log of the estimated BF divided by the true BF.

(b) The log of the BF estimated by ABC-IS against the log of the true BF.

(c) The log of the BF estimated by SL-IS against the log of the true BF.

(d) The log of the BF estimated by MAVIS against the log of the true BF.

Figure 1: Bayes' factors for the Poisson and geometric models.

by the bias. For this reason it might be more relevant in this example to consider the deviations from the shallow slope, which are likely due to the Monte Carlo variance in the estimator (which becomes more pronounced as ϵ is reduced). We see that the choice of ϵ essentially governs a bias-variance trade-off, and that the difficulty in using the approach more generally is that it is not easy to evaluate whether a choice of ϵ that ensures a low variance also ensures that the bias is not significant in terms of affecting the conclusions that might be drawn from the estimated BF (see section 2.4). Figure 1c suggests that SL has worked extremely well (in terms of having a low variance) for the most important situations, where the BF is close to 1. However, we note that the large biases introduced due to the limitation of the Gaussian assumption when the BF is far from 1. Figure 1d indicates that there is little or no bias when using MAVIS, but that the variance (due to using IS on the relatively high-dimensional u-space) negatively impacts on the results.

These results highlight that the three methods will be most effective in slightly different situations. The approximations in ABC and SL introduce a bias, the effect of which might be difficult to assess. In ABC (assuming sufficient statistics) this bias can be reduced by an increased computational effort allowing a smaller ϵ, however it is essentially impossible to assess when this bias is "small enough". SL is the simplest method to implement, and seems to work well in a wide variety of situations, but the advice in Wood (2010) should be followed in checking that the assumption of normality is appropriate. MAVIS is limited by the need to perform importance sampling on the high-dimensional (θ, u) space (a problem avoided by using summary statistics in the other methods and mitigated by the possibility of using AIS), but the bias is less of an issue, and the method is also able to estimate the evidence for a single model.

2.4 Application to social networks

In this section we use our methods to compare the evidence for two alternative ERGMs for the Gamaneg data (figure 2), previously analysed in Friel (2013). An ERGM has the general form

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

where $S(y)$ is a vector of statistics of a network y and θ is a parameter vector of the same length. We take $S(y) = \#$ of edges in model 1 and $S(y) = (\#$ of edges, $\#$ of two stars) in model 2 . As in Friel (2013) we use the prior $p(\theta) = \mathcal{N}(\theta; 0, 25I)$.

Using a computational budget of 10^5 simulations from the likelihood (each simulation consisting of an internal MCMC run of length 1000 as a proxy for an exact sampler, as described in section 1.1.3), Friel (2013) finds that the evidence for model 1 is $\sim 37\times$ that for model 2. Using the same computational budget for our methods, consisting of 1000 importance points (with 100 simulations from the likelihood for each point), we obtained the results shown in Table 1.

This example highlights the issue with the bias-variance trade-off in ABC, with $\epsilon = 0.1$ having too large a bias and $\epsilon = 0.05$ having too large a variance. SL performs well - in this particular case the Gaussian assumption appears to be appropriate. One might expect this, since the statistics are

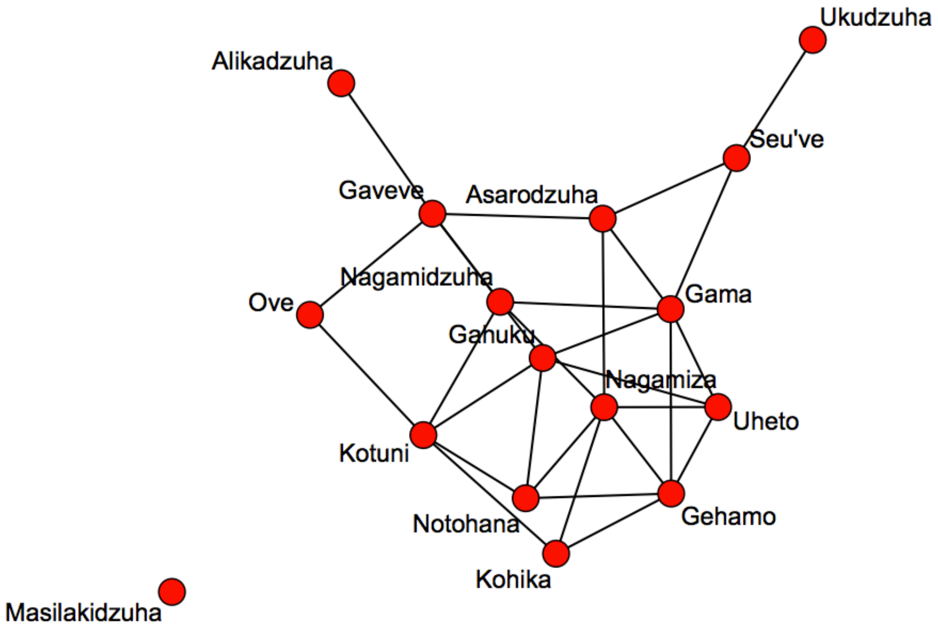


Figure 2: The Gamaneg data.

	ABC ($\epsilon = 0.1$)	ABC ($\epsilon = 0.05$)	SL	MAVIS
$\frac{\hat{p}(y\|M_1)}{\hat{p}(y\|M_2)}$	4	20	40	41

Table 1: Model comparison results for Gamaneg data. Note that the ABC ($\epsilon = 0.05$) estimate was based upon just 5 sample points of non-zero weight. MAVIS also provides estimates of the individual evidence ($\log [\widehat{p}(y|M_1)] = -69.6$, $\log [\widehat{p}(y|M_2)] = -73.3$).

sums of random variables. However, we note that this is not usually the case for ERGMs, particularly when modelling large networks, and that SL is a much more appropriate method for inference in the ERGMs with local dependence (Schweinberger and Handcock, 2015). We might expect a more sophisticated ABC approach to exhibit improved performance, possibly outperforming SL. However, the appeal of SL is in its simplicity, and we find it to be a useful method for obtaining good results with minimal tuning.

2.5 Application to Ising models

The implementation of MAVIS in the previous section is not an exact-approximate method for two reasons:

1. An internal MCMC chain was used in place of an exact sampler;
2. The $1/Z(\widehat{\theta})$ term in equation 16 was estimated before running this algorithm (by using a standard SMC method to estimate $Z(\widehat{\theta})$, and taking the reciprocal) with this fixed estimate being used throughout.

However, in practice, we tend to find that such "inexact-approximations" do not introduce large errors into Bayes' factor estimates, particularly when compared to standard implementations of ABC (as seen in the previous section). We investigate some of the theoretical aspects of such approximations in section 4. In the current section we investigate this type of approach further empirically, using data simulated from Ising models. In particular we 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 for which accurate estimates (via Friel and Rue (2007)) of the evidence serve as a ground truth against which to compare our methods.

The approach we take is to look at several different methods for estimating the $Z(\theta^*)/Z(\theta)$ term in the weight equation 14. We use IS and AIS (as already introduced), but note a potential weakness in these approaches, in that they use $f(\cdot|\theta^{(p)})$ as an IS proposal without guarantees that this distribution has heavier tails than the IS "target" $f(\cdot|\widehat{\theta})$. As an alternative we investigate the use of bridge sampling (Meng and Wong, 1996). In particular we use the estimate

$$\frac{\widehat{Z(\widehat{\theta})}}{Z(\theta)} = \left(\sum_{m=1}^{M/2}\left[\frac{1}{Z(\widehat{\theta})}\frac{\gamma(u_1^{(m)}|\widehat{\theta})}{\gamma(u_1^{(m)}|\theta)}\right]^{1/2}\right) \Bigg/ \left(\sum_{m=1}^{M/2}\left[Z(\widehat{\theta})\frac{\gamma(u_2^{(m)}|\theta)}{\gamma(u_2^{(m)}|\widehat{\theta})}\right]^{1/2}\right), \tag{17}$$

where the $u_1^{(m)} \sim f(\cdot|\theta)$ and $u_2^{(m)} \sim f(\cdot|\widehat{\theta})$ are simulated by taking the final points of long internal MCMC runs. This bridge sampling estimate results from the use of the geometric bridging distribution $\gamma^*\left(\cdot|\theta,\widehat{\theta}\right) = \left[\gamma^*\left(\cdot|\widehat{\theta}\right)\gamma^*\left(\cdot|\theta\right)\right]^{1/2}$, and we will refer to the resultant algorithm as "bridged SAVIS". This method essentially "averages" IS estimators of $Z(\theta^*)/Z(\theta)$ and $Z(\theta)/Z(\theta^*)$, only one of which may have a high variance due to the use of a light tailed proposal. However, although we might expect the bridge sampling estimates to be lower variance than those from IS, the effect of

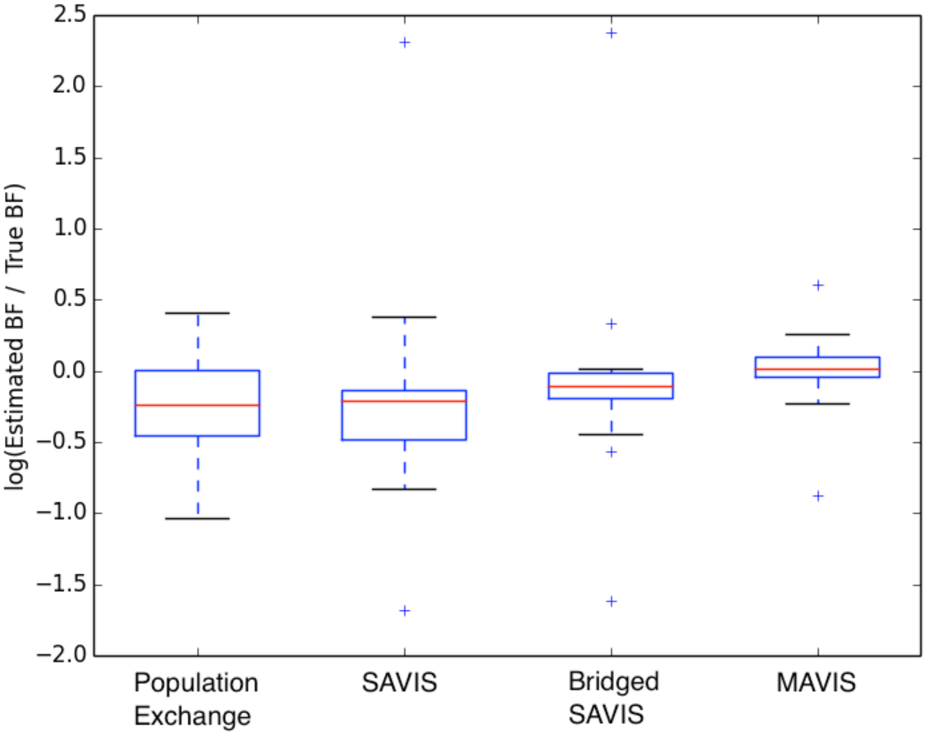


Figure 3: Box plots of the results of population exchange, SAVIS, bridged SAVIS and MAVIS on the Ising data.

its use within the "external" IS method is more unpredictable since the estimates from equation 17 are not unbiased.

We configure our methods such that they have a similar computational cost to those in Friel (2013), using a total 10^5 simulations from the likelihood (each of which is the final point of an internal MCMC run of length 200) to estimate one Bayes' factor, in order that we may compare our results with those in that paper. Here, estimating a marginal likelihood is done in three stages: firstly $\widehat{\theta}$ is estimated; followed by $Z(\widehat{\theta})$, then finally the marginal likelihood. We took $\widehat{\theta}$ to be the posterior expectation, estimated from a run of the exchange algorithm of 10^4 iterations. $Z(\widehat{\theta})$ was then estimated using SMC with an MCMC move, with 200 particles and 100 targets, with the ith target being $\gamma_i(\cdot|\theta) = \gamma_i\left(\cdot|\frac{i\theta}{100}\right)$, employing stratified resampling when the effective sample size (ESS; Kong *et al.* (1994)) falls below 100. The importance proposal used in all cases was a multivariate normal distribution, with mean and variance taken to be the sample mean and variance from the initial run of the exchange algorithm. This proposal would clearly not be appropriate in high dimensions, but is reasonable for the low dimensional parameter spaces considered here. We then examine three marginal likelihood estimation methods, each with 100 importance points: (a) SAVIS with $M = 200$; (b) bridged SAVIS with $M = 200$; and (c) MAVIS with $M = 1$ and $K = 200$. Figure 3 shows the results produced by these methods in comparison with those from Friel (2013).

We observe that AIS offers an improvement over IS, but also that the bridged IS method performs better than IS despite the additional bias in estimating $Z(\theta^*)/Z(\theta)$ due to its improved variance. As in Alquier *et al.* (2014), we observe that it may be useful to move away from the exact-approximate approaches, and in this case, to simply use the best available estimator of $Z(\theta^*)/Z(\theta)$ (taking into account its statistical and computational efficiency) regardless of whether it is unbiased. We examine this strategy further in section 4. In this example we might also consider alternative approaches, for example linked IS Neal (2005) (which is unbiased), or a nested sampling (Skilling, 2006) style approach (which is not).

2.6 Discussion

In this section we have compared the use of ABC-IS, SL-IS, MAVIS (and alternatives) for estimating marginal likelihoods and Bayes' factors. The use of ABC for model comparison has received much attention, with much of the discussion centring around appropriate choices of summary statistics. We have avoided this in our examples by using exponential family models, but in general this remains an issue affecting both ABC and SL. It is the use of summary statistics that makes ABC and SL unable to provide evidence estimates. However, it is the use of summary statistics, usually essential in these settings, that provides ABC and SL with an advantage over MAVIS, in which importance sampling must be performed over the high dimensional data-space. Despite this disadvantage, MAVIS avoids the approximations made in the simulation based methods (illustrated in figures 1b to 1d), with the accuracy depending primarily on the quality of the estimate of $1/Z$ used. In section 2.5 we saw that there can be advantages of using biased, but lower variance estimates in place of standard IS.

The main weakness of all of the methods described in this section is that they are all based on standard IS and are thus not practical for use when θ is high dimensional. In the next section we examine the use of SMC samplers as an extension to IS for use on triply intractable problems and in this framework discuss further the effect of inexact-approximations.

3 Sequential Monte Carlo approaches

SMC samplers (Del Moral *et al.*, 2006) are a generalisation of IS, in which the problem of choosing an appropriate proposal distribution in IS is avoided by performing IS sequentially on a sequence of target distributions, starting at a target that is easy to simulate from, and ending at the target of interest. In standard IS the number of Monte Carlo points required in order to obtain a particular accuracy increases exponentially with the dimension of the space, but Beskos *et al.* (2011) show (under appropriate regularity conditions) that the use of SMC circumvents this problem and can thus be practically useful in high dimensions.

In this section we introduce SMC algorithms for simulating from doubly intractable posteriors which have the by-product that, like IS, they also produce estimates of marginal likelihoods. We note that, although here we focus on estimating the evidence, the SMC sampler approaches based here are a natural alternative to the MCMC methods described in section 1.1. and inherently use a "population" of Monte Carlo points (shown to be beneficial on these models by Caimo and Friel (2011)). In section 3.1 we describe these algorithms, before examining an application to estimating the precision matrix of a Gaussian distribution in high dimensions in section 3.2.

3.1 SMC samplers in the presence of an INC

This section introduces two alternative SMC samplers for use on doubly intractable target distributions. The first, marginal SMC, directly follows from the IS methods in the previous section. The second, SMC-MCMC, requires a slightly different approach, but is more computationally efficient. Finally we briefly discuss simulation-based SMC samplers in section 3.1.4.

To begin, we introduce notation that is common to all algorithms that we discuss. SMC samplers perform sequential IS using P "particles" $\theta^{(p)}$, each having (normalised) weight $w^{(p)}$, using a sequence of targets π_0 to π_T, with π_T being the distribution of interest, in our case $\pi(\theta|y) \propto p(\theta)f(y|\theta)$. In this section we will take $\pi_t(\theta|y) \propto p(\theta)f_t(y|\theta) = p(\theta)\gamma_t(y|\theta)/Z_t(\theta)$. At target t, a "forward" kernel $K_t(\cdot|\theta_{t-1}^{(p)})$ is used to move particle $\theta_{t-1}^{(p)}$ to $\theta_t^{(p)}$, with each particle then being reweighted to give unnormalised weight

$$\begin{aligned}\widetilde{w}_t^{(p)} &= \frac{p(\theta_t^{(p)})f_t(y|\theta_t^{(p)})}{p(\theta_{t-1}^{(p)})f_{t-1}(y|\theta_{t-1}^{(p)})}\frac{L_{t-1}(\theta_t^{(p)},\theta_{t-1}^{(p)})}{K_t(\theta_{t-1}^{(p)},\theta_t^{(p)})} &(18)\\ &= \frac{p(\theta_t^{(p)})\gamma_t(y|\theta_t^{(p)})}{p(\theta_{t-1}^{(p)})\gamma_{t-1}(y|\theta_{t-1}^{(p)})}\frac{Z_{t-1}(\theta_{t-1}^{(p)})}{Z_t(\theta_t^{(p)})}\frac{L_{t-1}(\theta_t^{(p)},\theta_{t-1}^{(p)})}{K_t(\theta_{t-1}^{(p)},\theta_t^{(p)})}. &(19)\end{aligned}$$

Here, L_{t-1} represents a "backward" kernel that we chose differently in the alternative algorithms below. We note the presence of the INC, which means that this algorithm cannot be implemented in practice in it current form. The weights are then normalised to give $\left\{w_t^{(p)}\right\}$, then a resampling step is carried out. In the following sections the focus is on the reweighting step, since this is the main difference between the different algorithms. For more detail on these methods, see Del Moral *et al.* (2007).

Zhou *et al.* (2013) describe how BFs can be estimated directly by SMC samplers, simply by taking π_1 to be one model and π_T to be the other (with the π_t being intermediate distributions). This idea is also explored for Gibbs random fields in Friel (2013). However, the empirical results in Zhou *et al.* (2013) suggest that in some cases this method does not necessarily perform better than estimating marginal likelihoods for the two models separately and taking the ratio of the estimates. Here we do not investigate these algorithms further, but note that they offer an alternative to estimating the marginal likelihood separately.

3.1.1 Marginal SMC

The first method we describe results from the use of an approximation to the optimal backward kernel (Peters, 2005; Klaas *et al.*, 2005). In this case the weight update is

$$\begin{aligned}
\widetilde{w}_t^{(p)} &= \frac{p(\theta_t^{(p)}) f_t(y|\theta_t^{(p)})}{\sum_{r=1}^{P} w_{t-1}^{(r)} K_t(\theta_t^{(p)}|\theta_{t-1}^{(r)})} && (20) \\
&= \frac{p(\theta_t^{(p)}) \gamma_t(y|\theta_t^{(p)})}{Z_t(\theta_t^{(p)}) \sum_{r=1}^{P} w_{t-1}^{(r)} K_t(\theta_t^{(p)}|\theta_{t-1}^{(r)})} && (21)
\end{aligned}$$

for an arbitrary forward kernel K_t. This method is quite widely used, but is a little non-standard in terms of the framework of Del Moral *et al.* (2006), since it uses a Monte Carlo estimate of an importance weight defined on the marginal θ-space at target t, compared to the usual weight on the entire past history of each particle. This results in a computational complexity of $O(P^2)$ (although as noted by Klaas *et al.* (2005) that this may often be reduced to $O(P \log(P))$ at the cost of negligible bias), compared to $O(P)$ for a standard SMC method, but we include it here since we notice that equation 21 contains the term $1/Z(\cdot)$ just as does the corresponding expression (equation 14) for standard IS. This leads us to consider the same approach for avoiding the calculation of $Z(\cdot)$ as in section 2.1. Namely, we employ the SAV target and proposal within the SMC algorithm.

We still have an intractable normalising constant in the denominator. Now let us try the SAVM posterior, where in target t we use the distribution q_u for the auxiliary variable u_t, and the SAVM proposal, where $u_t^{(p)} \sim f_t(\cdot|\theta_t^{(p)})$. In this case, the weight update is

$$\begin{aligned}
\widetilde{w}_t^{(p)} &= \frac{q_u(u_t^{(p)}|\theta_t^{(p)}, y) p(\theta_t^{(p)}) f_t(y|\theta_t^{(p)})}{\sum_{r=1}^{P} K_t(\theta_t^{(p)}|\theta_{t-1}^{(r)}) f_t(u_t^{(p)}|\theta_t^{(p)}) w_{t-1}^{(r)}} \\
&= \frac{q_u(u_t^{(p)}|\theta_t^{(p)}, y) p(\theta_t^{(p)}) \gamma_t(y|\theta_t^{(p)})}{\gamma_t(u_t^{(p)}|\theta_t^{(p)}) \sum_{r=1}^{P} K_t(\theta_t^{(p)}|\theta_{t-1}^{(r)}) w_{t-1}^{(r)}}.
\end{aligned}$$

We see that no normalising constant appears in this weight update, but the algorithm is $O(P^2)$.

3.1.2 SMC with an MCMC kernel

Suppose we were able to use a reversible MCMC kernel K_t with invariant distribution $\pi_t(\theta|y) \propto p(\theta) f_t(y|\theta)$, and choose the L_{t-1} kernel to be the time reversal of K_t with respect to its invariant distribution, we obtain the following incremental weight:

$$\widetilde{w}_t^{(p)} = \frac{p(\theta_t^{(p)})f_t(y|\theta_t^{(p)})}{p(\theta_{t-1}^{(p)})f_{t-1}(y|\theta_{t-1}^{(p)})}\frac{L_{t-1}(\theta_t^{(p)},\theta_{t-1}^{(p)})}{K_t(\theta_{t-1}^{(p)},\theta_t^{(p)})} \tag{22}$$

$$= \frac{p(\theta_t^{(p)})f_t(y|\theta_t^{(p)})}{p(\theta_{t-1}^{(p)})f_{t-1}(y|\theta_{t-1}^{(p)})}\frac{p(\theta_{t-1}^{(p)})f_t(y|\theta_{t-1}^{(p)})}{p(\theta_t^{(p)})f_t(y|\theta_t^{(p)})} \tag{23}$$

$$= \frac{\gamma_t(y|\theta_{t-1}^{(p)})}{\gamma_{t-1}(y|\theta_{t-1}^{(p)})}\frac{Z_{t-1}(\theta_{t-1}^{(p)})}{Z_t(\theta_{t-1}^{(p)})}. \tag{24}$$

Once again, we cannot evaluate this incremental weight due to the presence of a ratio of normalising constants. Also, such an MCMC kernel cannot generally be directly constructed - the MH update itself involves evaluating the ratio of intractable normalising constants. However, appendix A shows that precisely the same weight update results when using either SAV or exchange MCMC moves in place of a direct MCMC step.

In order that this approach may be implemented we might consider, in the spirit of the approximations suggested in section 2, using an estimate of the ratio term $Z_{t-1}(\theta_{t-1}^{(p)})/Z_t(\theta_{t-1}^{(p)})$. For example, an unbiased IS estimate is given by

$$\widehat{\frac{Z_{t-1}(\theta_{t-1}^{(p)})}{Z_t(\theta_{t-1}^{(p)})}} = \frac{1}{M}\sum_{m=1}^{M}\frac{\gamma_{t-1}(u_t^{(m,p)}|\theta_{t-1}^{(p)})}{\gamma_t(u_t^{(m,p)}|\theta_{t-1}^{(p)})}, \tag{25}$$

where $u_t^{(p,m)} \sim f_t(\cdot|\theta_{t-1}^{(p)})$. Although this estimate is unbiased, we note that the resultant algorithm does not have precisely the same extended space interpretation as the methods in Del Moral *et al.* (2006). Appendix B gives an explicit construction for this case, which incorporates a pseudomarginal-type construction (Andrieu and Roberts, 2009).

3.1.3 Data point tempering

For the SMC approach to be efficient we require the sequence of likelihoods $\{\pi_t\}$ to be chosen such that π_0 is easy to simulate from, π_T is the target of interest and the π_t provide a "route" between them. For the applications in this paper we found the data tempering approach of Chopin (2002) to be particularly useful. Suppose that the data y consists of N points, and that N is exactly divisible by T for ease of exposition. We then propose to take $\pi_0(\theta|y) = p(\theta)$ and for $t = 1, ...T$ $\pi_t(\theta|y) = p(\theta)f_t(y|\theta)$ with

$$f_t(y|\theta) = f\left(y_{1:Nt/T}|\theta\right), \tag{26}$$

i.e. essentially we add in N/T data points for each increment of t. On this sequence of targets we then propose to use the SMC sampler with an MCMC kernel as described in the previous section. The only slightly non-standard point is the estimation of $Z_{t-1}(\theta_{t-1}^{(p)})/Z_t(\theta_{t-1}^{(p)})$, since in this case

$Z_{t-1}(\theta_{t-1}^{(p)})$ and $Z_t(\theta_{t-1}^{(p)})$ are the normalising constants of distributions on different spaces. We use

$$\widehat{\frac{Z_{t-1}(\theta_{t-1}^{(p)})}{Z_t(\theta_{t-1}^{(p)})}} = \frac{1}{M}\sum_{m=1}^{M} \frac{\gamma_{t-1}(v_{t,1}^{(m,p)}|\theta_{t-1}^{(p)})q_w(w_{t,1}^{(m,p)})}{\gamma_t(u_{t,1}^{(m,p)}|\theta_{t-1}^{(p)})} \tag{27}$$

where $u_{t,1}^{(m,p)} \sim f_t(\cdot|\theta_{t-1}^{(p)})$ and $v_{t,1}^{(m,p)}$ and $w_{t,1}^{(m,p)}$ are subvectors of $u_{t,1}^{(m,p)}$. $w_{t,1}^{(m,p)}$ is in the space of the additional variables added when moving from f_{t-1} to f_t (providing the argument in an arbitrary auxiliary distribution $q_w(\cdot)$) and $v_{t,1}^{(m,p)}$ is in the space of the existing variables. For $t = 1$ this becomes

$$\widehat{\frac{1}{Z_1(\theta_0^{(p)})}} = \frac{1}{M}\sum_{m=1}^{M} \frac{q_w(u_{1,1}^{(m,p)})}{\gamma_1(u_{1,1}^{(m,p)}|\theta_0^{(p)})} \tag{28}$$

with $u_{1,1}^{(m,p)} \sim f_t(.|\theta_0^{(p)})$.

Analogous to the SAV method, a sensible choice for $q_w(w)$ might be to use $f\left(w|\widehat{\theta}\right)$, where w is on the same space as N/T data points. The normalising constant for this distribution needs to be known to calculate the importance weight in equation 31 so, as earlier, we advocate estimating this in advance of running the SMC sampler (aside from when the data points are added one at a time - in this case the normalising constant may usually be found analytically). Note that if y does not consist of i.i.d. points, it is useful to choose the order in which data points are added such that the same q_w (each with the same normalising constant) can be used in every weight update. For example, in an Ising model, the requirement would be to add the same shape grid of variables at each target.

3.1.4 Simulation-based SMC samplers

Section 2.2 describes how the ABC and SL approximations may be used within IS. The same approximate likelihoods may be used in SMC. In ABC (Sisson *et al.*, 2007), where the sequence of targets is chosen to be $\pi_t(\theta) \propto p(\theta)\widehat{f}_{\epsilon_t}(y|\theta)$ with a decreasing sequence ϵ_t, this idea has proved a particularly useful alternative to MCMC for exploring ABC posterior distributions, whilst also providing estimates of Bayes' factors (Didelot *et al.*, 2011). The use of SMC with SL does not appear to have been explored previously. One might expect SMC to be useful in this context (using, for example, the sequence of targets $\pi_t(\theta) \propto p(\theta)\widehat{f}_{\text{SL}}^{(t/T)}(S(y)|\theta)$), even if it is not as obviously useful as in the ABC setting.

3.2 Application to precision matrices

3.2.1 Model

In this section we examine the performance of our SMC sampler for estimating the evidence in an example in which θ is high dimensional (relative to the other examples considered in this paper). We consider the case in which $\theta = \Sigma^{-1}$ is an unknown precision matrix, $f(y|\theta)$ is the d-dimensional multivariate Gaussian distribution with zero mean and $p(\theta)$ is a Wishart distribution $\mathcal{W}(\nu, V)$ with parameters $\nu \geq d$ and $V \in \mathbb{R}^{d\times d}$. Suppose we observe n i.i.d. observations $y = \{y_i\}_{i=1}^{n}$, where $y_i \in \mathbb{R}^d$. The true evidence can be calculated analytically, and is given by

$$p(y) = \frac{1}{\pi^{nd/2}} \frac{\Gamma_d(\frac{\nu+n}{2})}{\Gamma_d(\frac{\nu}{2})} \frac{\left|\left(V^{-1} + \sum_{i=1}^{n} y_i y_i^T\right)^{-1}\right|^{\frac{\nu+n}{2}}}{|V|^{\frac{\nu}{2}}}, \tag{29}$$

where Γ_d denotes the d-dimensional gamma function. For ease of implementation, we parametrise the precision using a Cholesky decomposition $\Sigma^{-1} = LL'$ with L a lower triangular matrix whose (i, j)'th element is denoted a_{ij}.

As in section 2.3, we write $f(y|\theta)$ as $\gamma(y|\theta)/Z(\theta)$ as follows

$$\begin{aligned} f\left(\{y_i\}_{i=1}^{n} \mid \Sigma^{-1}\right) &= \prod_{i=1}^{n} (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} y_i' \Sigma^{-1} y_i\right) \\ &= \left[(2\pi)^d |\Sigma|\right]^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i' \Sigma^{-1} y_i\right), \end{aligned} \tag{30}$$

where in some of the experiments that follow, $Z(\theta) = \left[(2\pi)^d |\Sigma|\right]^{n/2}$ is treated as if it is an INC. In the Wishart prior, we take $\nu = 10 + d$ and $V = I_d$.

3.2.2 Data, algorithm description and results

Taking $d = 10$, $n = 30$ points were simulated using $y_i \sim \mathcal{MVN}\left(0_d, 0.1 \times I_d\right)$. The parameter space is thus 55-dimensional, motivating the use of an SMC sampler in place of IS or the population exchange method, neither of which are suited to this problem. In the SMC sampler, where we used $P = 10,000$ particles, the sequence of targets is given by data point tempering. Specifically, the sequence of targets is to use $p(\Sigma^{-1})$ when $t = 0$ and $p(\Sigma^{-1}) f\left(\{y_i\}_{i=1}^{t} \mid \Sigma^{-1}\right)$ for $t = 1, ..., T(= n)$. The parameter space consists of $\{a_{ij} \mid 1 \leq j \leq i \leq d\}$. We use single component MH kernels to update each of the parameters, with one (deterministic) sweep consisting of an update of each in turn. For each a_{ij} we use a Gaussian random walk proposal, where at target t, the variance for the proposal used for a_{ij} is taken to be the sample variance of a_{ij} at target $t - 1$. For updating the weights of each particle we used equation 27, where we chose $q_w(\cdot) = f\left(\cdot \mid \widehat{\Sigma^{-1}}\right)$ with $\widehat{\Sigma^{-1}}$ the maximum likelihood estimate of the precision Σ^{-1}, and chose $M = 200$ "internal" importance sampling points. Systematic resampling was performed when the effective sample size (ESS) fell below $P/2$.

We estimated the evidence 10 times using the SMC sampler and compared the statistical properties of each algorithm using these estimates. For our simulated data, the log of the true evidence was -89.43. Over the 10 runs of the SMC sampler, the median log evidence was -89.34. A summary is given in table 2.

Statistic	Minimum	1st Quartile	Median	3rd Quartile	Maximum
log evidence	-90.01	-89.51	-89.35	-88.92	-88.37

Table 2: Summary statistics of the log evidence over 10 runs of the SMC sampler.

4 IS and SMC with biased weights

The example in section 2.5 indicates that in practice it may sometimes be advantageous to use biased rather than unbiased estimates of importance weights within a random weight IS algorithm: an observation that is somewhat analogous to that made in Alquier *et al.* (2014) in the context of MCMC. This section provides an initial exploration as to whether this might be a useful strategy in general. In section 4.1 we consider the situation of IS, and in section 4.2 that of SMC.

4.1 IS with biased weights

In order to analyse the behaviour of importance sampling with biased weights, we consider estimating the normalising constant using biased importance sampling. In particular, consider the estimation of $Z(\theta)$, via equation 15, by sampling from a proposal distribution $U_1, \ldots, U_P \overset{iid}{\sim} q(\cdot|\theta)$ and computing:

$$\hat{Z}^{-1}_{\text{bis}}(\theta) = \frac{1}{P}\sum_{p=1}^{P} w'(U_p|\theta),$$

where the biased randomised weight admits an additive decomposition,

$$w'(u|\theta) := \frac{p(\theta)\gamma(u|\theta)}{Z(\theta)q(u)} + b(u|\theta) + V_{u|\theta},$$

in which $b(u|\theta)$ is a deterministic function describing the bias of the weights and $V_{u|\theta}$ is a random variable (more precisely, there is an independent copy of such a random variable associated with every particle), which conditional upon u and θ is of mean zero and variance $\sigma^2_\theta(u)$. This decomposition will not generally be available in practice, but is flexible enough to allow the formal description of many settings of interest. For instance, one might consider the algorithms presented in section 2 by setting $b(u|\theta)$ to the (conditional) expected value of the difference between the approximate and exact weights and $V_{u|\theta}$ to the difference between the approximate weights and their expected value.

We have immediately that the bias of such an estimate is $\mathbb{E}_{q(\cdot|\theta)}[b(\cdot|\theta)]$ and, by a simple application of the law of total variance, its variance is

$$\text{Var}_{q(\cdot|\theta)}(w'(\cdot|\theta))/P = \text{Var}_{q(\cdot|\theta)}\left[p(\theta)\gamma(\cdot|\theta)/Z(\theta)q(\cdot|\theta) + b(\cdot|\theta)\right]/P + \mathbb{E}_{q(\cdot|\theta)}[\sigma^2_\theta(\cdot)]/P.$$

Consequently, the mean squared error of this estimate is:

$$\text{MSE}(w') = \text{Var}_{q(\cdot|\theta)}\left[p(\theta)\gamma(\cdot|\theta)/Z(\theta)q(\cdot|\theta) + b(\cdot|\theta)\right]/P + \mathbb{E}_{q(\cdot|\theta)}[\sigma^2_\theta(\cdot)]/P + \mathbb{E}_{q(\cdot|\theta)}[b(\cdot|\theta)]^2.$$

If we compare such a biased estimator with a second estimator in which we use the same proposal distribution but instead use an unbiased weight

$$\tilde{w}(u|\theta) := \frac{p(\theta)\gamma(u|\theta)}{Z(\theta)q(u|\theta)} + \tilde{V}_{u|\theta},$$

where $\tilde{V}_{u|\theta}$ has conditional expectation zero and variance $\tilde{\sigma}_\theta^2(u)$, then it's clear that the biased estimator has smaller mean squared error for small enough samples if it has sufficiently smaller variance, i.e., when (assuming $\mathbb{E}_{q(\cdot|\theta)}[b(\cdot|\theta)]^2 > 0$, otherwise one estimator dominates the other for all sample sizes):

$$\begin{aligned}
&\mathrm{Var}_{q(\cdot|\theta)}\left[\frac{p(\theta)\gamma(\cdot|\theta)}{Z(\theta)q(\cdot|\theta)} + b(\cdot|\theta)\right]/P + \mathbb{E}_{q(\cdot|\theta)}[\sigma_\theta^2]/P + \mathbb{E}_{q(\cdot|\theta)}[b(\cdot|\theta)]^2 \\
&<\mathrm{Var}_{q(\cdot|\theta)}\left[\frac{p(\theta)\gamma(\cdot|\theta)}{Z(\theta)q(\cdot|\theta)}\right]/P + \mathbb{E}_q[\tilde{\sigma}_\theta^2]/P \\
\Leftrightarrow P &< \frac{\mathrm{Var}_{q(\cdot|\theta)}\left[\frac{p(\theta)\gamma(\cdot|\theta)}{Z(\theta)q(\cdot|\theta)}\right] + \mathbb{E}_q[\tilde{\sigma}_\theta^2] - \mathrm{Var}_{q(\cdot|\theta)}\left[\frac{p(\theta)\gamma(\cdot|\theta)}{Z(\theta)q(\cdot|\theta)} + b(\cdot|\theta)\right] - \mathbb{E}_{q(\cdot|\theta)}[\sigma_\theta^2]}{\mathbb{E}_{q(\cdot|\theta)}[b(\cdot|\theta)]^2} \\
&< \frac{\mathbb{E}_{q(\cdot|\theta)}[\tilde{\sigma}_\theta^2 - \sigma_\theta^2] - \mathrm{Var}_{q(\cdot|\theta)}\left[b(\cdot|\theta)\right] - 2\mathrm{Cov}_{q(\cdot|\theta)}\left[\frac{p(\theta)\gamma(\cdot|\theta)}{Z(\theta)q(\cdot|\theta)}, b(\cdot|\theta)\right]}{\mathbb{E}_{q(\cdot|\theta)}[b(\cdot|\theta)]^2}.
\end{aligned}$$

In the artificially simple setting in which $b(u|\theta) = b_0$ is constant, this would mean that the biased estimator would have smaller MSE for samples smaller than the ratio of the difference in variance to the square of that bias suggesting that qualitatively a biased estimator might be better if the square of the average bias is small in comparison to the variance reduction that it provides. Given a family of increasingly expensive biased estimators with progressively smaller bias, one could envisage using this type of argument to manage the trade-off between less biased estimators and larger sample sizes. See section 2.5 for an empirical investigation of importance sampling estimators with biased weights.

4.2 SMC with biased weights

4.2.1 Error bounds

We now examine the effect of such inexact weights on estimates produced by SMC samplers. By way of theoretical motivation of such an approach, we demonstrate that under strong assumptions on the mixing of the sampler, if the approximation error is sufficiently small, then this error can be controlled uniformly over the iterations of the algorithm and will *not* accumulate unboundedly over time (and that it can in principle be made arbitrarily small by making the relative bias small enough for the desired level of accuracy). We do not here consider the particle system itself, but rather the sequence of distributions which are being approximated by Monte Carlo in the approximate version of the algorithm and in the idealised algorithm being approximated. The Monte Carlo approximation of this sequence can then be understood as a simple mean field approximation and its convergence has been well studied, see for example Del Moral (2004).

In order to do this, we make a number of identifications in order to allow the consideration of the approximation in an abstract manner. We allow $\widetilde{G}_t$ to denote the incremental weight function at time t, and G_t to denote the *exact* weight function which it approximates (any auxiliary random variables needed in order to obtain this approximation are simply added to the state space and their sampling distribution to the transition kernel). The transition kernel M_t combines the proposal distribution of the SMC algorithm together with the sampling distribution of any needed auxiliary variables. We allow x to denote the full collection of variables sampled during an iteration of the sampler, which is assumed to exist on the same space during each iteration of the sampler.

We employ the following assumptions (we assume an infinite sequence of algorithm steps and associated target distributions, proposals and importance weights; naturally, in practice only a finite number would be employed but this formalism allows for a straightforward statement of the result):

A1 (Bounded Relative Approximation Error) There exists $\gamma < \infty$ such that:

$$\sup_{t\in\mathbb{N}} \sup_{x} \frac{|G_t(x) - \widetilde{G}_t(x)|}{\widetilde{G}_t(x)} \leq \gamma.$$

A2 (Strong Mixing; slightly stronger than a global Doeblin condition) There exists $\epsilon(M) > 0$ such that:

$$\sup_{t\in\mathbb{N}} \inf_{x,y} \frac{dM_t(x,\cdot)}{dM_t(y,\cdot)} \geq \epsilon(M).$$

A3 (Control of Potential) There exists $\epsilon(G) > 0$ such that:

$$\sup_{t\in\mathbb{N}} \inf_{x,y} \frac{G_t(x)}{G_t(y)} \geq \epsilon(G).$$

The first of these assumptions controls the error introduced by employing an inexact weighting function; the others ensure that the underlying dynamic system is sufficiently ergodic to forget it's initial conditions and hence limit the accumulation of errors. We demonstrate below that the combination of these properties suffices to transfer that stability to the approximating system.

We consider the behaviour of the distributions η_p and $\tilde{\eta}_p$ which correspond to the target distributions at iteration p of the exact and approximating algorithms, prior to reweighting, at iteration p in the following proposition, the proof of which is provided in Appendix C, which demonstrates that if the approximation error, γ, is sufficiently small then the accumulation of error over time is controlled:

Proposition 1 (Uniform Bound on Total-Variation Discrepancy). *If A1, A2 and A3 hold then:*

$$\sup_{n\in\mathbb{N}} \|\eta_n - \widetilde{\eta}_n\|_{TV} \leq \frac{4\gamma(1-\epsilon(M))}{\epsilon^3(M)\epsilon(G)}.$$

This result is not intended to do any more than demonstrate that, qualitatively, such forgetting can prevent the accumulation of error even in systems with “biased” importance weighting potentials. In practice, one would wish to make use of more sophisticated ergodicity results such as those of Whiteley (2013), within this framework to obtain results which are somewhat more broadly applicable: assumptions A2 and A3 are very strong, and are used only because they allow stability to be established simply. Although this result is, in isolation, too weak to justify the use of the approximation schemes introduced here in practice it seems sufficient, together with the empirical results presented below, to suggest that further investigation of such approximations might be warranted.

4.2.2 Empirical results

We use the precision example introduced in section 3.2.1 to investigate the effect of using biased weights in SMC samplers. Specifically we take $d = 1$ and use a simulated dataset y where $n = 500$ points were simulated using $y_i \sim \mathcal{N}(0, 0.1)$. In this case there is only a single parameter to estimate, λ_1, and we examine the bias of estimates of the evidence using three alternative SMC samplers, each of which use a data-tempered sequence of targets (adding one data point at each target). The first SMC sampler, which we refer to as *exact*, is the same as that used in section 3.2.2, which uses an unbiased IS weight estimate, here with $M = 20$ “internal” importance sampling points. The second, which we refer to as *inexact*, uses a biased bridge estimator instead, specifically we use in place of equation 27

$$\frac{\widehat{Z_{t-1}(\theta^{(p)}_{t-1})}}{Z_t(\theta^{(p)}_{t-1})} = \left(\sum_{m=1}^{M/2}\left[\frac{\gamma_{t-1}(v^{(m,p)}_{t,1}|\theta^{(p)}_{t-1})q_w(w^{(m,p)}_{t,1})}{\gamma_t(u^{(m,p)}_{t,1}|\theta^{(p)}_{t-1})}\right]^{1/2}\right) \Bigg/ \left(\sum_{m=1}^{M/2}\left[\frac{\gamma_t(u^{(m,p)}_{t,2}|\theta^{(p)}_{t-1})}{\gamma_{t-1}(v^{(m,p)}_{t,2}|\theta^{(p)}_{t-1})q_w(w^{(m,p)}_{t,2})}\right]^{1/2}\right), \tag{31}$$

where $v^{(m,p)}_{t,2} \sim f_{t-1}(.|\theta^{(p)}_{t-1})$, $w^{(m,p)}_{t,2} \sim q_w(.)$ so that $u^{(m,p)}_{t,2} = \left(v^{(m,p)}_{t,2}, w^{(m,p)}_{t,2}\right)$, and $u^{(m,p)}_{t,1} \sim f_t(.|\theta^{(p)}_{t-1})$ with $v^{(m,p)}_{t,1}$ and $w^{(m,p)}_{t,1}$ being the corresponding subvectors of $u^{(m,p)}_{t,1}$. The third, which we refer to as *perfect,* is the method where the true value of $Z_{t-1}(\theta^{(p)}_{t-1})/Z_t(\theta^{(p)}_{t-1})$ is used in the weight update. Each SMC sampler was run 50 times, using 500 particles.

Here we are concerned with examining the bias introduced by using the inexact SMC sampler compared to the perfect one; to give context we also compare the exact sampler to the perfect one. Figure 4 shows the difference in the mean log evidence estimate at each target between both the exact and inexact SMC algorithms and the perfect algorithm. We observe that, as expected, there is no indication of bias accumulating when using the exact algorithm, but that bias does accumulate in the inexact algorithm. This accumulation of bias means that one should exercise caution in the use of SMC samplers with biased weights. In this case we observe that the bias does not accumulate sufficiently to give poor estimates of the evidence: the standard deviation of the final log evidence estimate over the SMC sampler runs is approximately 0.1, so the bias is not large by comparison.

Further, motivated by the theoretical argument presented previously, we investigate the effect of improving the mixing of the kernel used within the SMC. In this model the exact posterior is

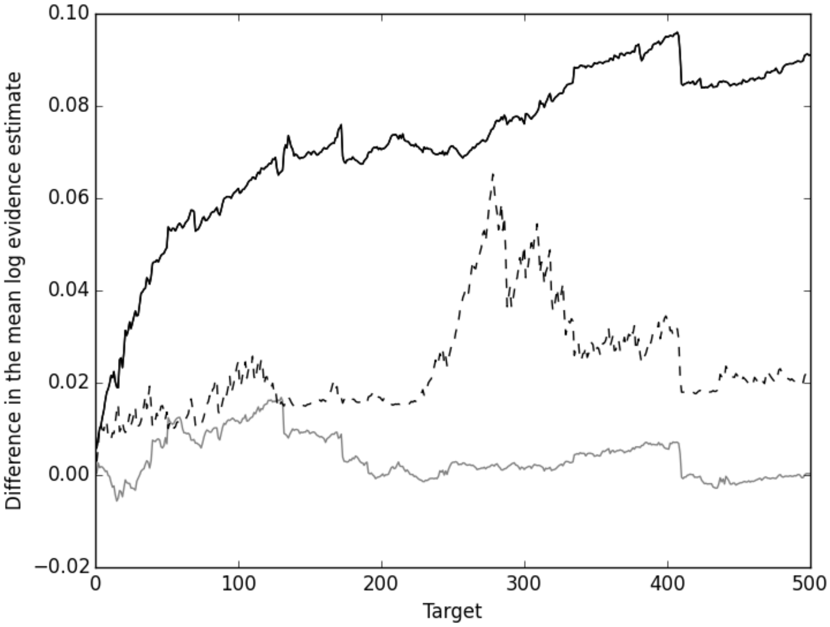


Figure 4: The difference between the mean log evidence estimates of the exact (grey solid) and inexact (black solid) SMC algorithms using MCMC kernels with the perfect SMC algorithm, and also the corresponding difference when using the inexact algorithm with perfect mixing (dashed).

available at each SMC target, so we may replace the use of an MCMC move to update the parameter with a direct simulation from the posterior. In this extreme case, there is no dependence between each particle and its history; we refer to this as "perfect mixing". Figure 4 shows the bias of the inexact algorithm using an MCMC kernel and also the corresponding bias of the inexact algorithm with perfect mixing. We observe that improved mixing substantially decreases the bias in the evidence estimates from the algorithm.

4.3 Discussion

In section 2.5 we observed clearly that the use of biased weights in IS can be useful for estimating the evidence in doubly intractable models, but we have not observed the same for SMC with biased weights. When applied to the precision example in section 3.2, an inexact sampler (using the bridge estimator) did not outperform the exact sampler, despite the mean square error of the biased bridge weight estimates being substantially improved compared to the unbiased IS estimate. Over 10 runs the mean square error in the log evidence was 0.34 for the inexact sampler, compared to 0.28 for the exact sampler. This experience suggests that samplers with biased weights should be used with caution: weight estimates with low variance do not guarantee good performance due to the accumulation of bias in the SMC.

However, the theoretical and empirical investigation in this section suggests that this idea is worth further investigation, possibly for situations involving some of the other intractable likelihoods listed in section 1. Our results suggest that improved mixing can help combat the accumulation of bias, which may implies that there may be situations where it is useful to perform many iterations of a kernel at a particular target, rather than the more standard approach of using many intermediate targets at each of which a single iteration of a kernel is used. Other variations are also possible, such as the calculation of fast cheap biased weights at each target in order only to adaptively decide when to resample, with more accurate weight estimates (to ensure accurate resampling and accurate estimates based on the particles) only calculated when the method chooses to resample.

5 Conclusions

This paper describes several IS and SMC approaches for estimating the evidence in models with INCs that outperform previously described approaches. These methods may also prove to be useful alternatives to MCMC for parameter estimation. Several of the ideas in the paper are also applicable more generally, in particular the use of synthetic likelihood in the IS context and the notion of using biased weight estimates in IS and SMC. We advise caution in the use of biased weights in SMC due to the potential for bias to accumulate, but also note that this accumulated bias may be small compared to errors resulting from commonly accepted approximate techniques such as ABC.

References

Alquier P, Friel N, Everitt RG, Boland A (2015) Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. Statistics and Computing In press.

Andrieu C, Roberts GO (2009) The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics 37(2):697–725

Andrieu C, Vihola M (2012) Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. arXiv (1210.1484)

Beaumont MA (2003) Estimation of population growth or decline in genetically monitored populations. Genetics 164(3):1139–1160

Beskos A, Crisan D, Jasra A, Whiteley N (2011) Error Bounds and Normalizing Constants for Sequential Monte Carlo in High Dimensions. arXiv (1112.1544)

Caimo A, Friel N (2011) Bayesian inference for exponential random graph models. Social Networks 33:41–55

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

Chopin N, Jacob PE, Papaspiliopoulos O (2013) SMC^2: an efficient algorithm for sequential analysis of state space models. Journal of the Royal Statistical Society: Series B 75(3):397–426

Del Moral P (2004) Feynman-Kac formulae: genealogical and interacting particle systems with applications. Probability and Its Applications, Springer, New York

Del Moral P, Doucet A, Jasra A (2006) Sequential Monte Carlo samplers. Journal of the Royal Statistical Society: Series B 68(3):411–436

Del Moral P, Doucet A, Jasra A (2007) Sequential Monte Carlo for Bayesian computation. Bayesian Statistics 8:115–148

Didelot X, Everitt RG, Johansen AM, Lawson DJ (2011) Likelihood-free estimation of model evidence. Bayesian Analysis 6(1):49–76

Drovandi CC, Pettitt AN, Lee A (2014) Bayesian indirect inference using a parametric auxiliary model. Statistical Science

Everitt RG (2012) Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks. Journal of Computational and Graphical Statistics 21(4):940–960

Fearnhead P, Papaspiliopoulos O, Roberts GO, Stuart AM (2010) Random-weight particle filtering of continuous time processes. Journal of the Royal Statistical Society Series B 72(4):497–512

Friel N (2013) Evidence and Bayes factor estimation for Gibbs random fields. Journal of Computational and Graphical Statistics 22(3):518–532

Friel N, Rue H (2007) Recursive computing and simulation-free inference for general factorizable models. Biometrika 94(3):661–672

Girolami MA, Lyne AM, Strathmann H, Simpson D, Atchade Y (2013) Playing Russian Roulette with Intractable Likelihoods. arXiv (1306.4032)

Grelaud A, Robert CP, Marin JM (2009) ABC likelihood-free methods for model choice in Gibbs random fields. Bayesian Analysis 4(2):317–336

Klaas M, de Freitas N, Doucet A (2005) Toward practical N^2 Monte Carlo: The marginal particle filter. In: Proceedings of the 20th International Conference on Uncertainty in Artificial Intelligence

Kong A, Liu JS, Wong WH (1994) Sequential imputations and Bayesian missing data problems. JASA 89(425):278–288

Marin JM, Pillai NS, Robert CP, Rousseau J (2014) Relevant statistics for Bayesian model choice. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(5):833–859

Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America 100(26):15,324–15,328

Meng Xl, Wong WH (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica 6:831–860

Møller J, Pettitt AN, Reeves RW, Berthelsen KK (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika 93(2):451–458

Murray I, Ghahramani Z, MacKay DJC (2006) MCMC for doubly-intractable distributions. In: Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI), pp 359–366

Neal RM (2001) Annealed importance sampling. Statistics and Computing 11(2):125–139

Neal RM (2005) Estimating Ratios of Normalizing Constants Using Linked Importance Sampling. arXiv (0511.1216)

Nicholls GK, Fox C, Watt AM (2012) Coupled MCMC With A Randomized Acceptance Probability. arXiv (1205.6857)

Peters GW (2005) Topics in Sequential Monte Carlo Samplers. M.Sc. thesis, Unviersity of Cambridge

Picchini U, Forman JL (2013) Accelerating inference for diffusions observed with measurement error and large sample sizes using Approximate Bayesian Computation: A case study. arXiv (1310.0973)

Prangle D, Fearnhead P, Cox MP, Biggs PJ, French NP (2014) Semi-automatic selection of summary statistics for ABC model choice. Statistical Applications in Genetics and Molecular Biology 13(1):67–82

Rao V, Lin L, Dunson DB (2013) Bayesian inference on the Stiefel manifold. arXiv pp 1–32

Robert CP, Cornuet JM, Marin JM, Pillai NS (2011) Lack of confidence in approximate Bayesian computation model choice. Proceedings of the National Academy of Sciences of the United States of America 108(37):15,112–7

Schweinberger M, Handcock M (2015) Local dependence in random graph models: characterization, properties and statistical inference. Journal of the Royal Statistical Society: Series B In press.

Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America 104(6):1760–1765

Skilling J (2006) Nested sampling for general Bayesian computation. Bayesian Analysis 1(4):833–859

Tavaré S, Balding DJ, Griffiths RC, Donnelly PJ (1997) Inferring Coalescence Times From DNA Sequence Data. Genetics 145(2):505–518

Tran MN, Scharth M, Pitt MK, Kohn R (2013) IS^2 for bayesian inference in latent variable models. arXiv (1309.3339)

Whiteley N (2013) Stability properties of some particle filters. Annals of Applied Probability 23(6):2500–2537

Wilkinson RD (2013) Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Statistical Applications in Genetics and Molecular Biology 12(2):129–141

Wood SN (2010) Statistical inference for noisy nonlinear ecological dynamic systems. Nature 466(August):1102–1104

Zhou Y, Johansen AM, Aston JAD (2013) Towards automatic model comparison: An adaptive sequential Monte Carlo approach. arXiv (1303.3123)

A Using SAV and exchange MCMC within SMC

A.1 Weight update when using SAV-MCMC

Let us consider the SAVM posterior, with K being the MCMC move used in SAVM. In this case the weight update is

$$\begin{aligned}
\widetilde{w}_k^{(p)} &= \frac{p(\theta_t^{(p)})f_t(y|\theta_t^{(p)})q_u(u_t^{(p)}|\theta_t^{(p)},y)}{p(\theta_{t-1}^{(p)})f_{t-1}(y|\theta_{t-1}^{(p)})q_u(u_{t-1}^{(p)}|\theta_{t-1}^{(p)},y)}\frac{L_{t-1}((\theta_t^{(p)},u_t^{(p)}),(\theta_{t-1}^{(p)},u_{t-1}^{(p)}))}{K_t((\theta_{t-1}^{(p)},u_{t-1}^{(p)}),(\theta_t^{(p)},u_t^{(p)}))} \\
&= \frac{p(\theta_t^{(p)})f_t(y|\theta_t^{(p)})q_u(u_t^{(p)}|\theta_t^{(p)},y)}{p(\theta_{t-1}^{(p)})f_{t-1}(y|\theta_{t-1}^{(p)})q_u(u_{t-1}^{(p)}|\theta_{t-1}^{(p)},y)}\frac{p(\theta_{t-1}^{(p)})f_t(y|\theta_{t-1}^{(p)})q_u(u_{t-1}^{(p)}|\theta_{t-1}^{(p)},y)}{p(\theta_t^{(p)})f_t(y|\theta_t^{(p)})q_u(u_t^{(p)}|\theta_t^{(p)},y)} \\
&= \frac{\gamma_t(y|\theta_{t-1}^{(p)})}{\gamma_{t-1}(y|\theta_{t-1}^{(p)})}\frac{Z_{t-1}(\theta_{t-1}^{(p)})}{Z_t(\theta_{t-1}^{(p)})},
\end{aligned}$$

which is the same update as if we could use MCMC directly.

A.2 Weight update when using the exchange algorithm

Nicholls *et al.* (2012) show the exchange algorithm, when set up to target $\pi_t(\theta|y) \propto p(\theta)f_t(y|\theta)$ in the manner described in section 1.1.2, simulates a transition kernel that is in detailed balance with $\pi_t(\theta|y)$. This follows from showing that it satisfies a "very detailed balance" condition, which takes account of the auxiliary variable u. The result is that the derivation of the weight update follows exactly that of equations 22-24.

B An extended space construction for the random weight SMC method in section 3.1.2

The following extended space construction justifies the use of the "approximate" weights in equation 25 via an explicit sequential importance (re)sampling argument along the lines of Del Moral *et al.*

(2006), albeit with a slightly different sequence of target distributions.

Consider an actual sequence of target distributions $\{\pi_t\}_{t\geq 0}$. Assume we seek to approximate a normalising constant during every iteration by introducing additional variables $u_t = (u_t^1, \ldots, u_t^M)$ during iteration $t > 0$.

Define the sequence of target distributions:

$$\begin{aligned}&\widetilde{\pi}_t\left(\widetilde{x}_t = (\theta_0, \theta_1, u_1, \ldots, \theta_t, u_t)\right)\\ :=&\pi_t(\theta_t) \prod_{s=0}^{t-1} L_s(\theta_{s+1}, \theta_s) \prod_{s=1}^{t} \frac{1}{M} \sum_{m=1}^{M} \left[f_{s-1}(u_s^m|\theta_{s-1}) \prod_{q\neq m} f_s(u_s^m|\theta_{s-1}) \right]\end{aligned}$$

where L_s has the same rôle and interpretation as it does in a standard SMC sampler.

Assume that at iteration t the auxiliary variables u_t^m are sampled independently (conditional upon the associated value of the parameter, θ_{t-1})and identically according to $f_t(\cdot|\theta_{t-1})$ and that K_t denotes the incremental proposal distribution at iteration t, just as in a standard SMC sampler.

In the absence of resampling, each particle has been sampled from the following proposal distribution at time t:

$$\widetilde{\mu}_t(\widetilde{x}_t) = \mu_0(\theta_0) \prod_{s=1}^{t} K_s(\theta_{s-1}, \theta_s) \prod_{s=1}^{t} \prod_{m=1}^{M} f_s(u_s^m|\theta_{s-1})$$

and hence its importance weight should be:

$$\begin{aligned}W_t(\widetilde{x}_t) =& \frac{\pi_t(\theta_t) \prod_{s=0}^{t-1} L_s(\theta_{s+1}, \theta_s) \prod_{s=1}^{t} \frac{1}{M} \sum_{m=1}^{M} \left[f_{s-1}(u_s^m|\theta_{s-1}) \prod_{q\neq m} f_s(u_s^m|\theta_{s-1}) \right]}{\mu_0(\theta_0) \prod_{s=1}^{t} K_s(\theta_{s-1}, \theta_s) \prod_{s=1}^{t} \prod_{m=1}^{M} f_s(u_s^m|\theta_{s-1})}\\ =& \frac{\pi_t(\theta_t) \prod_{s=0}^{t-1} L_s(\theta_{s+1}, \theta_s)}{\mu_0(\theta_0) \prod_{s=1}^{t} K_s(\theta_{s-1}, \theta_s)} \frac{\prod_{s=1}^{t} \frac{1}{M} \sum_{m=1}^{M} \left[f_{s-1}(u_s^m|\theta_{s-1}) \prod_{q\neq m} f_s(u_s^m|\theta_{s-1}) \right]}{\prod_{s=1}^{t} \prod_{m=1}^{M} f_s(u_s^m|\theta_{s-1})}\\ =& \frac{\pi_t(\theta_t) \prod_{s=0}^{t-1} L_s(\theta_{s+1}, \theta_s)}{\mu_0(\theta_0) \prod_{s=1}^{t} K_s(\theta_{s-1}, \theta_s)} \prod_{s=1}^{t} \frac{1}{M} \sum_{m=1}^{M} \frac{f_{s-1}(u_s^m|\theta_{s-1})}{f_s(u_s^m|\theta_{s-1})}\\ =& W_{t-1}(\widetilde{x}_{t-1}) \cdot \frac{\pi_t(\theta_t) L_{t-1}(\theta_t, \theta_{t-1})}{\pi_{t-1}(\theta_{t-1}) K_t(\theta_{t-1}, \theta_t)} \frac{1}{M} \sum_{m=1}^{M} \frac{f_{t-1}(u_t^m, \theta_{t-1})}{f_t(u_t^m|\theta_{t-1})},\end{aligned}$$

which yields the natural sequential importance sampling interpretation. The validity of the incorporation of resampling follows by standard arguments.

If one has that $\pi_t(\theta_t) \propto p(\theta_t) f_t(y|\theta_t) = p(\theta_t)\gamma_t(y|\theta_t)/Z_t(\theta_t)$ and employs the time reversal of K_t for L_{t-1} then one arrives at an incremental importance weight, at time t of:

$$\frac{p(\theta_t) f_t(y|\theta_{t-1})}{p(\theta_{t-1}) f_{t-1}(y|\theta_{t-1})} \frac{1}{M} \sum_{m=1}^{M} \frac{f_{t-1}(u_t^m|\theta_{t-1})}{f_t(u_t^m|\theta_{t-1})} = \frac{p(\theta_t)\gamma_t(y|\theta_{t-1})}{p(\theta_{t-1})\gamma_{t-1}(y|\theta_{t-1})} \frac{1}{M} \sum_{m=1}^{M} \frac{\gamma_{t-1}(u_t^m|\theta_{t-1})}{\gamma_t(u_t^m|\theta_{t-1})}$$

yielding the algorithm described in section 3.1.2 as an exact SMC algorithm on the described extended space.

C Proof of SMC Sampler Error Bound

A little notation is required. We allow $(E, \mathcal{E})$ to denote the common state space of the sampler during each iteration, $\mathcal{C}_b(E)$ the collection of continuous, bounded functions from E to $\mathbb{R}$, and $\mathcal{P}(E)$ the collection of probability measures on this space. We define the Boltzmann-Gibbs operator associated with a potential function $G : E \to (0, \infty)$ as a mapping, $\Psi_G : \mathcal{P}(E) \to \mathcal{P}(E)$, weakly via the integrals of any function $\varphi \in \mathcal{C}_b(E)$

$$\int \varphi(x)\Psi_G(\eta)(dx) = \frac{\int \eta(dx)G(x)\varphi(x)}{\int \eta(dx')G(x')}.$$

The integral of a set A under a probability measure η is written $\eta(A)$ and the expectation of a function φ of $X \sim \eta$ is written $\eta(\varphi)$. The supremum norm on $\mathcal{C}_b(E)$ is defined $||\varphi||_\infty = \sup_{x\in E} \varphi(x)$ and the total variation distance on $\mathcal{P}(E)$ is $||\mu - \nu||_{\text{TV}} = \sup_A(\nu(A) - \mu(A))$. Markov kernels, $M : E \to \mathcal{P}(E)$ induce two operators, one on integrable functions and the other on (probability) measures:

$$\begin{aligned}
\forall \varphi \in& \mathcal{C}_b(E) : & M(\varphi)(\cdot) :=& \int M(\cdot, dy)\varphi(y) \\
\forall \mu \in& \mathcal{P}(E) : & (\mu M)(\cdot) :=& \int \mu(dx)M(x, \cdot).
\end{aligned}$$

Having established this notation, we note that we have the following recursive definition of the distributions we consider:

$$\widetilde{\eta}_0 = \eta_0 =: M_0 \qquad \eta_{t\geq 1} = \Psi_{G_{t-1}}(\eta_{t-1}) \qquad \widetilde{\eta}_{t\geq 1} = \Psi_{\widetilde{G}_{t-1}}(\widetilde{\eta}_{t-1})$$

and for notational convenience define the transition operators as

$$\Phi_t(\eta_{t-1}) = \Psi_{G_{t-1}}(\eta_{t-1})M_t \qquad \widetilde{\Phi}_t(\widetilde{\eta}_{t-1}) = \Psi_{\widetilde{G}_{t-1}}(\widetilde{\eta}_{t-1})M_t.$$

We make use of the (nonlinear) dynamic semigroupoid, which we define recursively, via it's action on a generic probability measure η, for $t \in \mathbb{N}$:

$$\Phi_{t,t}(\eta) = \eta \qquad \Phi_{t-1,t}(\eta) = \Phi_t(\eta) \qquad \Phi_{s,t} = \Phi_t(\Phi_{s,t-1}(\eta)) \text{ for } s < t,$$

with $\widetilde{\Phi}_{s,t}$ defined correspondingly.

We begin with a lemma which allows us to control the discrepancy introduced by Bayesian updating of a measure with two different likelihood functions.

Lemma 1. *Approximation Error*

If A1. holds, then $\forall \eta \in \mathcal{P}(E)$ and any $t \in \mathbb{N}$:

$$||\Psi_{\widetilde{G}_t}(\eta) - \Psi_{G_t}(\eta)||_{TV} \leq 2\gamma.$$

Proof. Let $\Delta_t := \widetilde{G}_t - G_t$ and consider a generic $\varphi \in \mathcal{C}_b(E)$:

$$\begin{aligned}
(\Psi_{\widetilde{G}_t}(\eta) - \Psi_{G_t}(\eta))(\varphi) =& \frac{\eta(\widetilde{G}_t\varphi)}{\eta(\widetilde{G}_t)} - \frac{\eta(G_t\varphi)}{\eta(G_t)} \\
=& \frac{\eta(G_t)\eta(\widetilde{G}_t\varphi) - \eta(\widetilde{G}_t)\eta(G_t\varphi)}{\eta(\widetilde{G}_t)\eta(G_t)} \\
=& \frac{\eta(G_t)\eta((G_t + \Delta_t)\varphi) - \eta((G_t + \Delta_t))\eta(G_t\varphi)}{\eta(\widetilde{G}_t)\eta(G_t)} \\
=& \frac{\eta(G_t)\eta(\Delta_t\varphi) - \eta(\Delta_t)\eta(G_t\varphi)}{\eta(\widetilde{G}_t)\eta(G_t)}
\end{aligned}$$

Considering the absolute value of this discrepancy, making using of the triangle inequality:

$$\begin{aligned}
\left|(\Psi_{\widetilde{G}_t}(\eta) - \Psi_{G_t}(\eta))(\varphi)\right| =& \left|\frac{\eta(G_t)\eta(\Delta_t\varphi) - \eta(\Delta_t)\eta(G_t\varphi)}{\eta(\widetilde{G}_t)\eta(G_t)}\right| \\
\leq& \left|\frac{\eta(\Delta_t\varphi)}{\eta(\widetilde{G}_t)}\right| + \left|\frac{\eta(\Delta_t)}{\eta(\widetilde{G}_t)}\right|\left|\frac{\eta(G_t\varphi)}{\eta(G_t)}\right|
\end{aligned}$$

Noting that G_t is strictly positive, we can bound $|\eta(G_t\varphi)|/\eta(G_t)$ with $\eta(G_t|\varphi|)/\eta(G_t)$ and thus with $\|\varphi\|_\infty$ and apply a similar strategy to the first term:

$$\begin{aligned}
\left|(\Psi_{\widetilde{G}_t}(\eta) - \Psi_{G_t}(\eta))(\varphi)\right| \leq& \left|\frac{\eta(|\Delta_t|)\,\|\varphi\|_\infty}{\eta(\widetilde{G}_t)}\right| + \left|\frac{\eta(\Delta_t)}{\eta(\widetilde{G}_t)}\right|\left|\frac{\eta(G_t|\varphi|)}{\eta(G_t)}\right| \\
\leq& \gamma\,\|\varphi\|_\infty + \gamma\,\|\varphi\|_\infty = 2\gamma\,\|\varphi\|_\infty\,.
\end{aligned}$$

(noting that $\eta(|\Delta_t|)/\eta(\widetilde{G_t}) < \gamma$ by integration of both sides of A1). □

We now demonstrate that, if the local approximation error at each iteration of the algorithm(characterised by γ) is sufficiently small then it does not accumulate unboundedly as the algorithm progresses.

Proof of Proposition 1

Proof. We begin with a telescopic decomposition (mirroring the strategy employed for analysing particle approximations of these systems in Del Moral (2004)):

$$\eta_t - \widetilde{\eta}_t = \sum_{s=1}^{t} \Phi_{s-1,t}(\widetilde{\eta}_{s-1}) - \Phi_{s,t}(\widetilde{\eta}_s).$$

We thus establish (noting that $\widetilde{\eta}_0 = \eta_0$):

$$\eta_t - \widetilde{\eta}_t = \sum_{s=1}^{t} \Phi_{s,t}(\Phi_s(\widetilde{\eta}_{s-1})) - \Phi_{s,t}(\widetilde{\Phi}_s(\widetilde{\eta}_{s-1})). \tag{32}$$

Turning our attention to an individual term in this expansion, noting that:

$$\Phi_s(\eta)(\varphi) = \Psi_{G_{s-1}}(\eta) M_s(\varphi) \qquad \widetilde{\Phi}_s(\eta)(\varphi) = \Psi_{\widetilde{G}_{s-1}}(\eta) M_s(\varphi)$$

we have, by application of a standard Dobrushin contraction argument and Lemma 1

$$\begin{aligned}
(\Phi_s(\widetilde{\eta}_{s-1}) - \widetilde{\Phi}_s(\widetilde{\eta}_{s-1}))(\varphi) =& \Psi_{G_{s-1}}(\widetilde{\eta}_{s-1}) M_s(\varphi) - \Psi_{\widetilde{G}_{s-1}}(\widetilde{\eta}_{s-1}) M_s(\varphi) \\
\left\| \Phi_s(\widetilde{\eta}_{s-1}) - \widetilde{\Phi}_s(\widetilde{\eta}_{s-1}) \right\|_{\mathrm{TV}} \leq& (1-\epsilon(M)) \left\| \Psi_{G_{s-1}}(\widetilde{\eta}_{s-1}) - \Psi_{\widetilde{G}_{s-1}}(\widetilde{\eta}_{s-1}) \right\|_{\mathrm{TV}} \\
\leq& 2\gamma(1-\epsilon(M))
\end{aligned} \tag{33}$$

which controls the error introduced instantaneously during each step.

We now turn our attention to controlling the accumulation of error. We make use of (Del Moral, 2004, Proposition 4.3.6) which, under assumptions A2 and A3, allows us to deduce that for any probability measures μ, ν:

$$\|\Phi_{s,s+k}(\mu) - \Phi_{s,s+k}(\nu)\|_{\mathrm{TV}} \leq \beta(\Phi_{s,s+k}) \|\mu - \nu\|_{\mathrm{TV}}$$

where

$$\beta(\Phi_{s,s+k}) = \frac{2}{\epsilon(M)\epsilon(G)} (1 - \epsilon^2(M))^k.$$

Returning to decomposition (32), applying the triangle inequality and this result, before finally inserting (33) we arrive at:

$$\begin{aligned}
\|\eta_t - \widetilde{\eta}_t\|_{\mathrm{TV}} \leq& \sum_{s=1}^{t} \left\| \Phi_{s,t}(\Phi_s(\widetilde{\eta}_{s-1})) - \Phi_{s,t}(\widetilde{\Phi}_s(\widetilde{\eta}_{s-1})) \right\|_{\mathrm{TV}} \\
\leq& \sum_{s=1}^{t} \frac{2(1-\epsilon^2(M))^{t-s}}{\epsilon(M)\epsilon(G)} \left\| \Phi_s(\widetilde{\eta}_{s-1}) - \widetilde{\Phi}_s(\widetilde{\eta}_{s-1}) \right\|_{\mathrm{TV}} \\
\leq& \sum_{s=1}^{t} \frac{2(1-\epsilon^2(M))^{t-s}}{\epsilon(M)\epsilon(G)} \cdot 2\gamma(1-\epsilon(M)) \\
=& \frac{4\gamma(1-\epsilon(M))}{\epsilon(M)\epsilon(G)} \sum_{s=1}^{t} (1-\epsilon^2(M))^{t-s}
\end{aligned}$$

This is trivially bounded over all t by the geometric series and a little rearrangement yields the result:

$$\begin{aligned}\frac{4\gamma(1-\epsilon(M))}{\epsilon(M)\epsilon(G)}\sum_{s=0}^{\infty}(1-\epsilon^2(M))^s =& \frac{4\gamma(1-\epsilon(M))}{\epsilon(M)\epsilon(G)}\frac{1}{1-(1-\epsilon^2(M))}\\ =&\frac{4\gamma(1-\epsilon(M))}{\epsilon^3(M)\epsilon(G)}.\end{aligned}$$

□