Bayesian model comparison with intractable likelihoods

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

1st April 2015

Abstract

Markov random eld 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. Speci cally, we have the situation where the posterior distribution (jy) / p( )f (yj)

cannot be evaluated due to the presence of the INC. We rst 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

$$(\;; u | y) \propto \; q_u(u| \;; y) f(y| \;) p(\;); \tag{2}$$

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

$$(\;; u) \quad f(u | \;) q(\;| \;); \tag{3}$$

giving an acceptance probability of

$$\min \quad 1; \frac{q(\;| \;)}{q(\;| \;)} \frac{p(\;)}{p(\;)} \frac{(y| \;)}{(y| \;)} \frac{q_u(u| \;; y)}{(u| \;)} \frac{(u| \;)}{q_u(u| \;; y)} \quad : \tag{4}$$

Note that, by viewing $q_u(u| \;; y) = (u| \;)$ as an unbiased IS estimator of $1 = Z(\;)$, 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 -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

$$\widehat{\frac{1}{Z(\;)}} = \frac{1}{M} \sum_{m=1}^{M} \frac{q_u(u^{(m)}| \;; y)}{(u^{(m)}| \;)}: \tag{5}$$

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 $q_u(\;| \;; y)$, which should ideally have lighter tails than $f(\;| \;)$. Møller et al. (2006) suggest that a reasonable choice may be $q_u(\;| \;; y) = f(\;| \;\widehat{\;})$, where $\widehat{\;}$ is the maximum likelihood estimator of . However, in practice $q_u(\;| \;; y)$ can be di cult 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 alterne79091 Tf 5.424 0 Td [d17874.25ma between $f(\,|\,)$ and $q_u(\,|\,;y)$. After the initial draw u_{K+1} $f(\,|\,)$, the auxiliary point is taken through a sequence of K MCMC moves which successively have target $f_k(\,|\,;\,\hat{}\,;y)$ for $k = K : 1$. The resultant IS estimator is given by

$$\widehat{\frac{1}{Z(\,)}} = \frac{1}{M}\sum_{m=1}^{M}\prod_{k=0}^{K}\ {}_{k+1}(u_k^{(m)}|$$

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

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

where $S(\cdot)$ gives a vector of summary statistics, $|\partial S/\partial u|(u)|$ denotes the Jacobian determinant arising from the change of variable, and $\pi_\epsilon(\cdot|S(y))$ is a density centred at $S(y)$ with bandwidth ϵ. As $\epsilon \to 0$, this distribution becomes more concentrated around $S(y)$, so that in the case where $S(\cdot)$ gives sufficient statistics for estimating θ, as $\epsilon \to 0$ 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

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

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 $f_S(S|\theta)$, 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 $f_S(S|\;)$ 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 are estimated based on the summary statistics of simulations $\{u^{(m)}\}_{m=1}^{M}$ $f(\,|\;)$. Concretely, the estimate of the likelihood is

$\hat{f}_{SL}$ ( ) = Tstrf[(u)]TJ/F2 10 Tf 0.761 0 Td [(s)-9391(m)]TJ 75.113 2.091 Td [(s)-9127 Gaus[(j)] TJ 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,

$$\hat{p}(y) = \frac{f(y|\tilde{\theta})p(\tilde{\theta})}{\hat{b}(\tilde{\theta}|y)}; \tag{13}$$

where $\tilde{\theta}$ can be an arbitrary value of with an estimate of the evidence given by

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

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

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

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

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

This results in weights

$$\tilde{w}^{(p)} = q_u(u \mid \theta^{(p)}) \ldots$$

(p)

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 $q_u(\cdot|\theta, y) = f(\cdot|\hat{\theta})$, as described in section 1.1.1, and γ_k as in equation 6, we obtain

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

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

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

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

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

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.

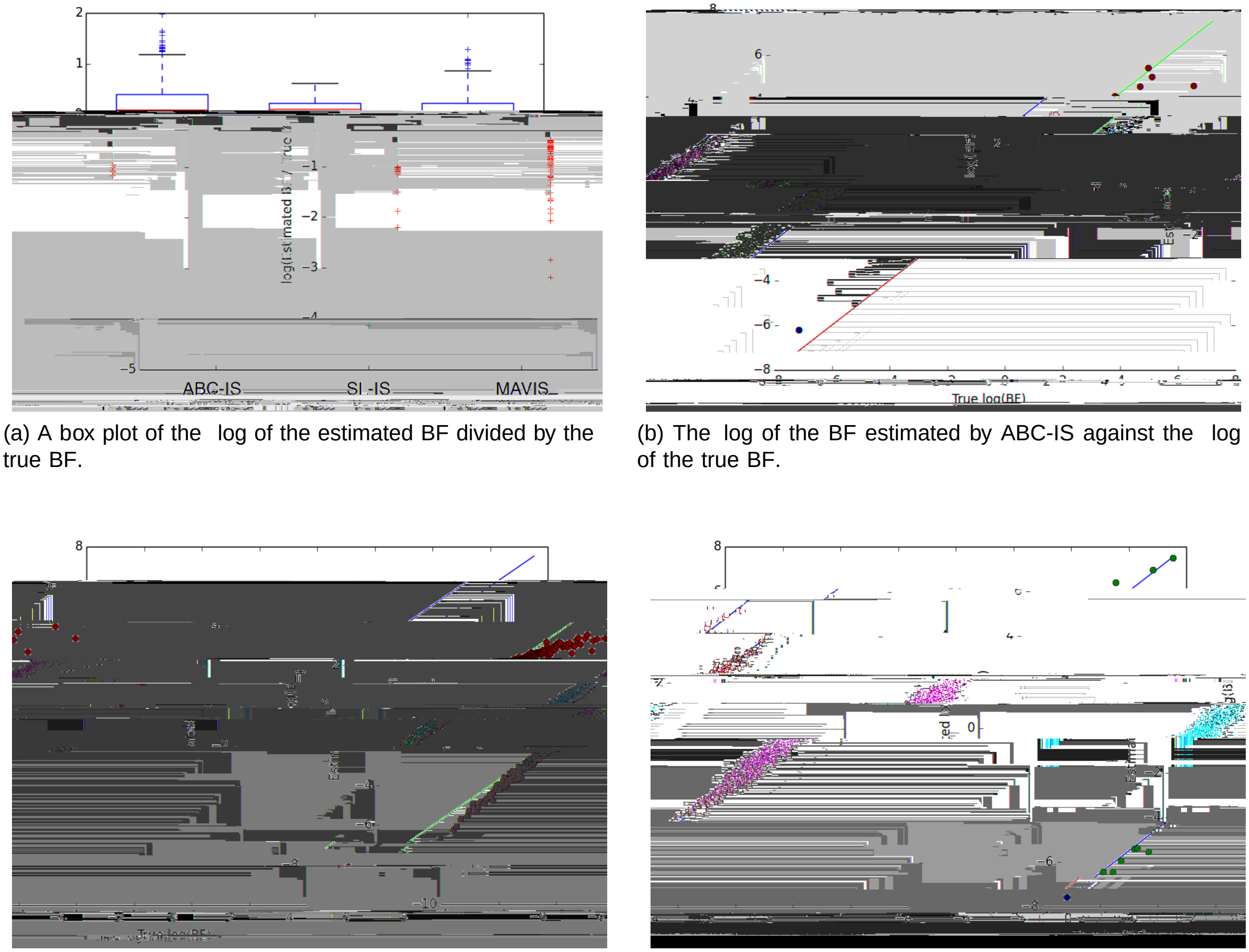
Consider i.i.d. observations $y = \{y_i\}_{i=1}^{n=100}$ of a discrete random variable that takes values in $\mathbb{N}$. For such a dataset, we will find the BF for the models

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

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

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

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



(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 is reduced). We see that the choice of essentially governs a bias-variance trade-o , and that the di culty in using the approach more generally is that it is not easy to evaluate whether a choice of

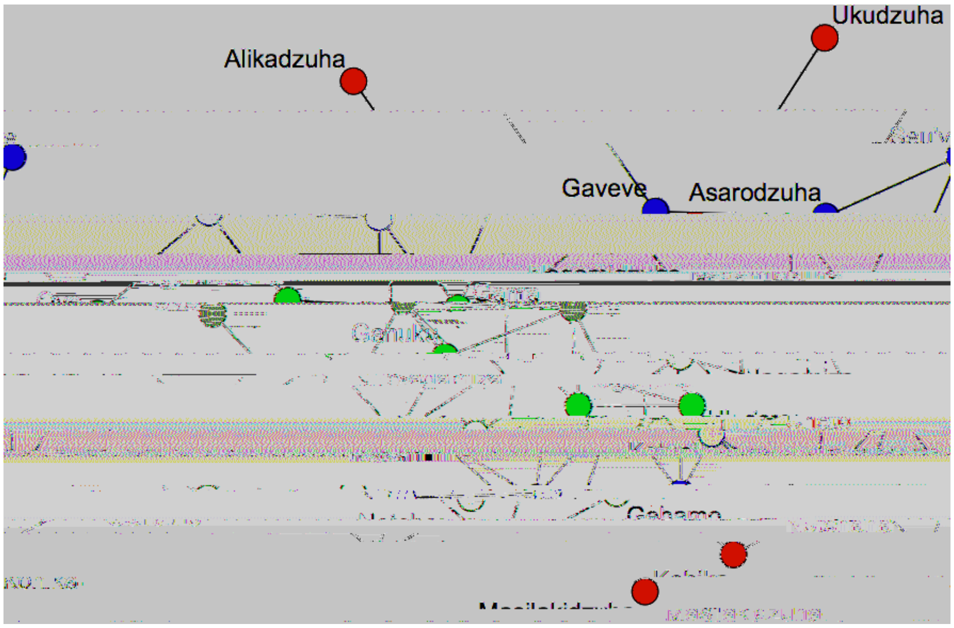


Figure 2: The Gamaneg data.

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

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

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

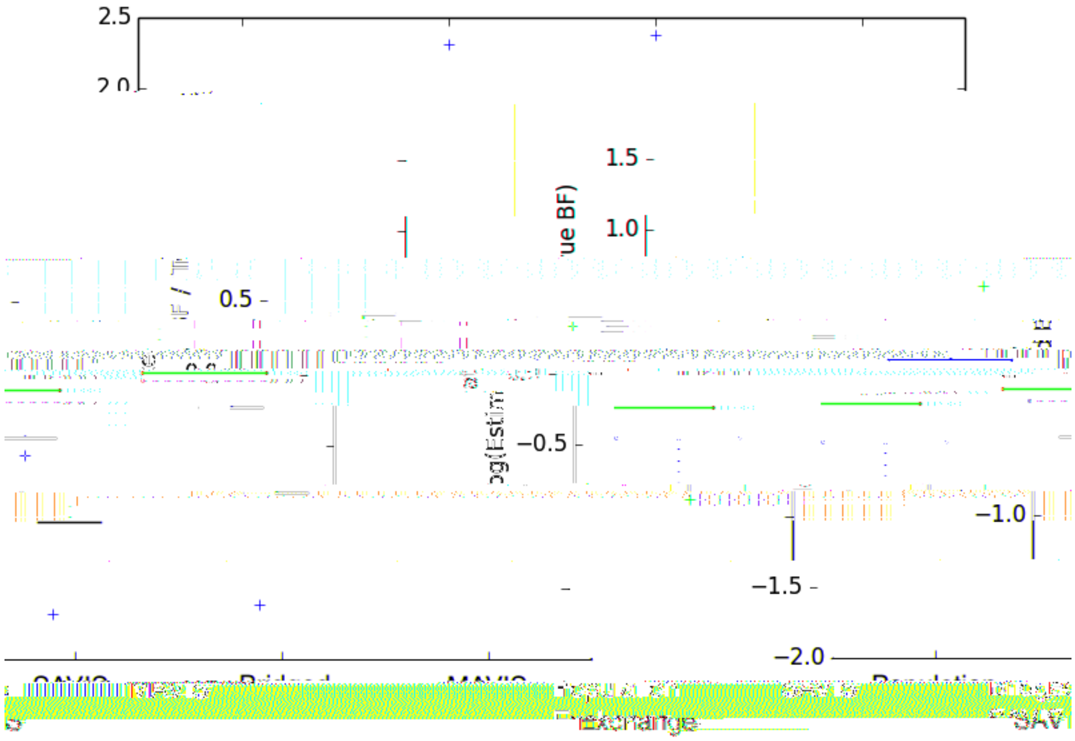


Figure 3: Box plots of the results of population exchange, SAVIS, bridged SAVIS and MAVIS on the Ising data.

We observe that AIS o ers an improvement over IS, but also that the bridged IS method performs better than IS despite the additional bias in estimating $Z(\ \ )=Z(\ )$

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.

3.1.1 Marginal SMC

The rst 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

$$\tilde{w}_t^{(p)} = \frac{p(\ _t^{(p)}) f_t(y|\ _t^{(p)})}{\sum_{r=1}^{P} w_{t\ 1}^{(r)} K_t(\ _t^{(p)}|\ _{t\ 1}^{(r)})} \tag{20}$$

$$= \frac{p(\ _t^{(p)})\ _t(y|\ _t^{(p)})}{Z_t(\ _t^{(p)}) \sum_{r=1}^{P} w_{t\ 1}^{(r)} K_t(\ _t^{(p)}|\ _{t\ 1}^{(r)})} \tag{21}$$

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 de ned on the marginal -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 $1/Z(\ )$ just as does the corresponding expression (equation 14) for standard IS. This leads us to consider the same approach for avoiding the calculation of $Z(\ )$ 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 $u_t^{(p)}\ \ f_t(\ |\ _t^{(p)})$. In this case, the weight update is

$$\tilde{w}_t^{(p)} = \frac{q_u(u_t^{(p)}|\ _t^{(p)}; y) p(\ _t^{(p)}) f_t(y|\ _t^{(p)})}{\sum_{r=1}^{P} K_t(\ _t^{(p)}|\ _{t\ 1}^{(r)}) f_t(u_t^{(p)}|\ _t^{(p)}) w_{t\ 1}^{(r)}}$$

$$= \frac{q_u(u_t^{(p)}|\ _t^{(p)}; y) p(\ _t^{(p)})\ _t(y|\ _t^{(p)})}{\ _t(u_t^{(p)}|\ _t^{(p)}) \sum_{r=1}^{P} K_t(\ _t^{(p)}|\ _{t\ 1}^{(r)}) w_{t\ 1}^{(r)}}.$$

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 $\ _t(\ |y) \propto p(\ ) f_t(y|\ )$, and choose the $L_{t\ 1}$ kernel to be the time reversal of K_t with respect to its invariant

$Z_{t.\ t}$

multivariate Gaussian distribution with zero mean and $p(\Sigma)$ is a Wishart distribution $W(\nu; V)$ 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

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

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, $Z(\Sigma) = \left[(2\pi)^d |\Sigma|\right]^{n/2}$ 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 $y_i \sim MVN(0_d, 0.1 \times I_d)$. The parameter space is thus 55-dimensional, motivating the use of an SMC sampler in place of IS or the population

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 $\hat{G}_t$ 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 $\gamma < 1$ such that:

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

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

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 $y_i \sim N(0; 0.1)$. In this case there is only a single parameter to estimate,

1

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

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 ltering 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 elds. 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 elds. Bayesian Analysis 4(2):317 336

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

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

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

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 di erent sequence of target distributions.

Consider an actual sequence of target distributions $\{\ _t\}_{t\ 0}$. 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 $t > 0$.

De ne the sequence of target distributions:

$$
\begin{aligned}
&e_t\,(\tilde{x}_t = (\ _0;\ _1; u_1; \ldots;\ _t; u_t)) \\
:= &\ _t(\ _t) \prod_{s=0}^{t-1} L_s(\ _{s+1};\ _s) \prod_{s=1}^{t} \quad 1
\end{aligned}
$$

extended space.

C Proof of SMC Sampler Error Bound

A little notation is required. We allow

If A1. holds, then 8 2 P (E) and any t 2 N:

$$jj \quad _{\tilde{G}_t}( ) \qquad _{G_t}( )jj_{TV} \quad 2 :$$

Proof. Let $_t := \tilde{G}_t \quad G_t$ and consider a generic 2 $C_b(E)$:

$$\begin{aligned}
(\quad _{\tilde{G}_t}( ) \qquad _{G_t}( ))(') &= \frac{(\tilde{G}_t')}{(\tilde{G}_t)} \quad \frac{(G_t')}{(G_t)} \\
&= \frac{(G_t) \ (\tilde{G}_t') \qquad (\tilde{G}_t) \ (G_t')}{(\tilde{G}_t) \ (G_t)} \\
&= \frac{(G_t) \ ((G_t + \quad _t)') \qquad ((G_t + \quad _t)) \ (G_t')}{(\tilde{G}_t) \ (G_t)} \\
&= \frac{(G_t) \ (\quad _t') \qquad (\quad _t) \ (G_t')}{(\tilde{G}_t) \ (G_t)}
\end{aligned}$$

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

$$\begin{aligned}
(\quad _{\tilde{G}_t}( ) \qquad _{G_t}( ))(') &= \frac{(G_t) \ (\quad _t') \qquad (\quad _t) \ (G_t')}{(\tilde{G}_t) \ (G_t)} \\
& \frac{(\quad _t')}{(\tilde{G}_t)} + \frac{(\quad _t)}{(\tilde{G}_t)} \quad \frac{(G_t')}{(G_t)}
\end{aligned}$$

Noting that G_t is strictly positive, we can bound j (G_t')j= (G_t) with $(G_t$j' j)= (G_t) and thus with k' k_1 and apply a similar strategy to the rst term:

$$\begin{aligned}
(\quad _{\tilde{G}_t}( ) \qquad _{G_t}( ))(') & \quad \frac{(j \quad _tj) \, k' \ k_1}{(\tilde{G}_t)} + \frac{(\quad _t)}{(\tilde{G}_t)} \quad \frac{(G_tj' \ j)}{(G_t)} \\
& \quad k' \ k_1 \ + \quad k' \ k_1 \ = 2 \quad k' \ k_1 \ :
\end{aligned}$$

(noting that (j $_t$j)= $(\tilde{G}_t)$ <

We thus establish (noting that $e_0 = \ _0$):

$$_t \quad e_t = \sum_{s=1}^{t} \ _{s;t}(\ _s(e_{s \ 1})) \quad _{s;t}(e_s(e_{s \ 1})): \tag{32}$$

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

$$_s(\ )(' \ ) = \ _{G_{s \ 1}}(\ )M_s(' \ ) \qquad e_s(\ )(' \ ) = \ _{e G_{s \ 1}}(\ )M_s(' \ )$$

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

$$\begin{aligned} (\ _s(e_{s \ 1}) \quad e_s(e_{s \ 1}))(' \ ) &= \ _{G_{s \ 1}}(e_{s \ 1})M_s(' \ ) \quad _{e G_{s \ 1}}(e_{s \ 1})M_s(' \ ) \\ _s(e_{s \ 1}) \quad e_s(e_{s \ 1}) \ _{TV} & \quad (1 \quad (M \ )) \quad _{G_{s \ 1}}(e_{s \ 1}) \quad _{e G_{s \ 1}}(e_{s \ 1}) \ _{TV} \\ & \quad 2 \ (1 \quad (M \ )) \end{aligned} \tag{33}$$

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 ; :

$$k \ _{s;s+k}(\ ) \quad _{s;s+k}(\ )k_{TV} \quad (\ _{s;s+k}) \ k \quad k_{TV}$$

where

$$(\ _{s;s+k}) = \frac{2}{(M \ ) \ (G)}(1 \quad ^2(M \ ))^k:$$

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

$$k \ _t \quad e_t k_{TV} \quad \sum_{s=1}^{t} \ _{s;t}(\ _s(e_{s \ 1})) \quad _{s;t}(e_s(e$$

have, by application of a standard Dobrushin contraction argu/F47 10.9091 Tf m 46.1-6.35 -3.272 Td .9091 1 Tf 4.7320.909n5 result:

$$\frac{4\ (1 \quad (M))}{(M)\ (G)} \sum_{s=0}^{1} (1 \quad {}^2(M))^s = \frac{4\ (1 \quad (M))}{(M)\ (G)} \frac{1}{1 \quad (1 \quad {}^2(M))}$$

$$= \frac{4\ (1 \quad (M))}{{}^3(M)\ (G)} :$$

$\square$