Department of Mathematics and Statistics

Preprint MPS-2016-01

Statistics and Computing manuscript No.
(will be inserted by the editor)

Bayesian model comparison with un-normalised likelihoods

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

Received: date / Accepted: date

Abstract Models for which the likelihood function can be evaluated only up to a parameter-dependent unknown normalizing constant, such as Markov random field models, are used widely in computer science, statistical physics, spatial statistics, and network analysis. However, Bayesian analysis of these models using stand-

ulation based" methodssuch 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 nd that elements of both approaches may also be more generally applicable. Speci cally:

{ For exact methods we nd 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-J/F8 9.9626 Tf 55.831 0 Td [(appr,-12.454 Td [(in)-388(imp)-28(ort)1(anc)-1(e)-387(sampling)-388((IS))-3

of a sequence of K targets, which in Murray et al (2006) are chosen to be

$$f_k(\theta, u \mid y) \propto \gamma_k(\theta, u \mid y) = \gamma(\theta)^{(K+1-k)/(K+1)} q_u(u \mid \theta, y)^{k/(K+1)} \quad (3)$$

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

$$\frac{\hat{Z}_1}{Z(\theta)} = \frac{1}{M} \sum_{m=1}^{M} \prod_{k=1}^{K} \frac{\gamma_k(u_{k-1}^{(m)} \mid \theta, y)}{\gamma_{k-1}(u_{k-1}^{(m)} \mid \theta, y)}. \quad (4)$$

This estimator has a lower variance (although at a higher computational cost) than the corresponding IS estimator. We note that AIS can be viewed as a particular case of SMC without resampling and one might expect to obtain additional improvements at negligible cost by incorporating resampling steps within such algorithms (see Zhou et al (2015) for an illustration of the potential improvement and some discussion); we do not pursue this here as it is not the focus of this work.

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

$$\min\left(1; \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)} \frac{p(\theta')}{p(\theta)} \frac{\gamma(y \mid \theta')}{\gamma(y \mid \theta)} \frac{\gamma(u \mid \theta)}{\gamma(u \mid \theta')}\right);$$

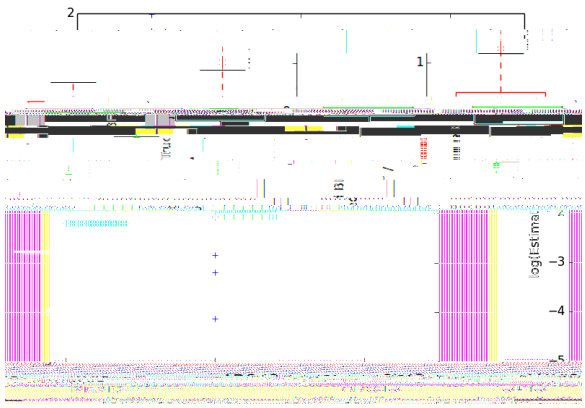
where $u \sim f(u \mid \theta)$, motivated by the intuitive idea that $\gamma(u \mid \theta)/\gamma(u \mid \theta')$ is a single point IS estimator of $Z(\theta)/Z(\theta')$. This method is shown to have the correct invariant distribution, as is the extension in which AIS is used in place of IS. A potential extension might seem to be using multiple importance points $\{u^{(m)}\}_{m=1}^{M} \sim f(u \mid \theta)$ 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 (2015). 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 (2015) 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

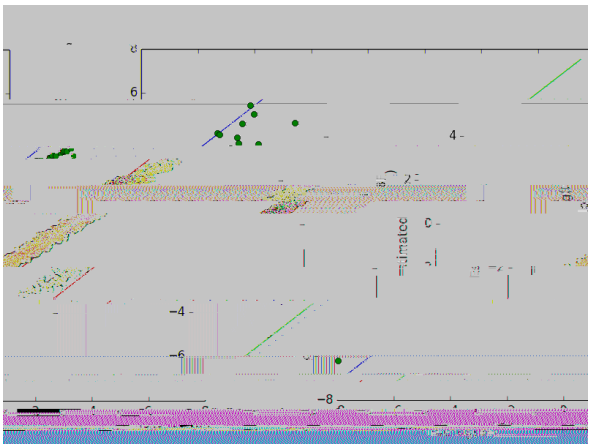
Girolami et al (2013) use series-based approximations to intractable target distributions within the exact-approximation framework, where "Russian Roulette" methods from the physics literature are used to ensure the unbiasedness of truncations of infinite sums. These methods do not require exact simulation from $f(\cdot \mid \theta)$, as 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

in (2); and using AIS, rather than simple IS, for estimating $1/Z(\theta^{(p)})$ as in (4) (giving an algorithm that we refer to as multiple auxiliary variable IS (MAVIS), in common with the terminology in Murray et al (2006)). Using $q_u(\theta, y) = f(\theta|y)$, as described in section 1.1.1, and α_k as in (3), we obtain

$$\frac{d1}{Z(\theta)} = \frac{1}{Z(\theta)}$$



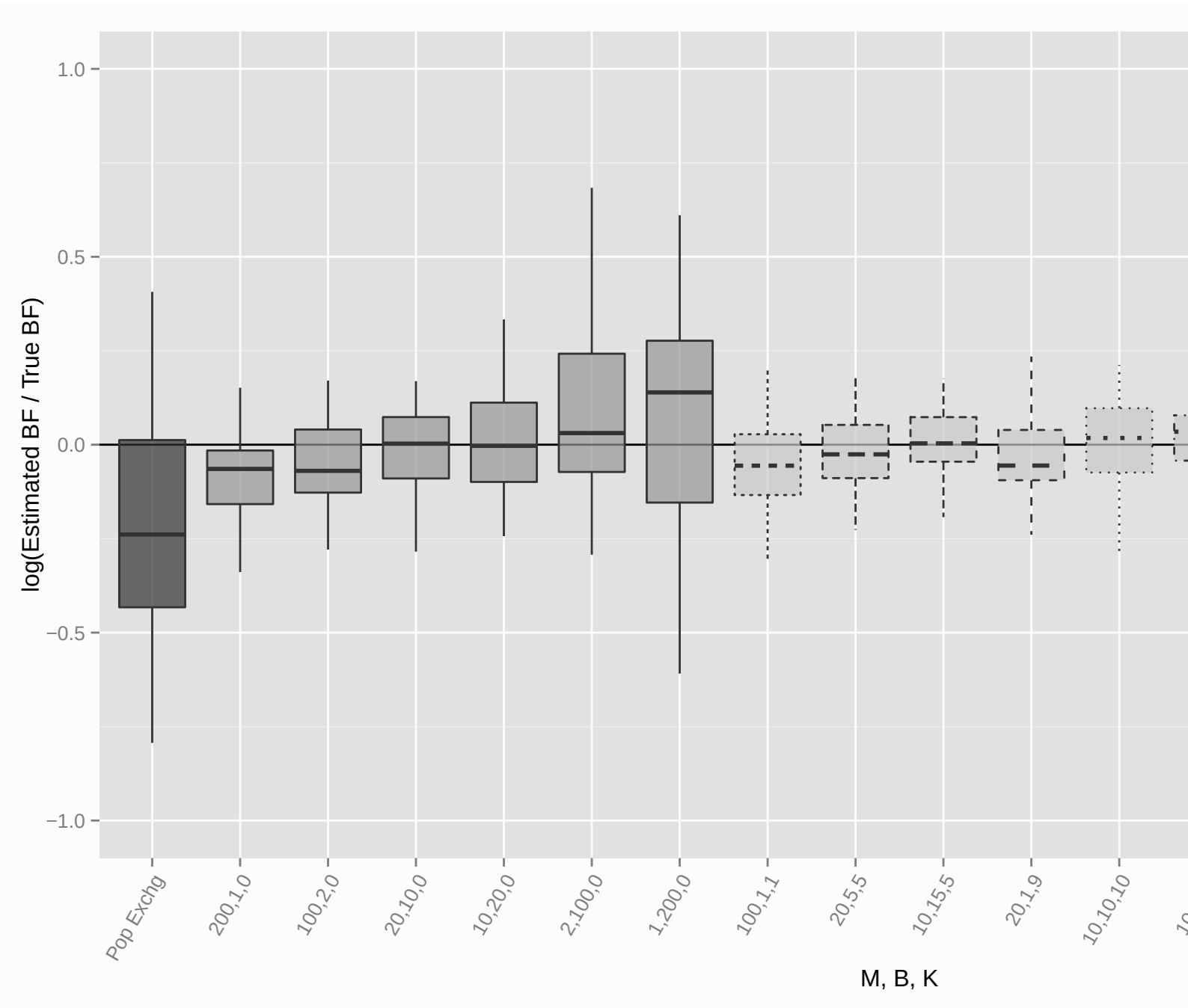
(a) A box plot of the log of the estimated BF divided by the true BF.



Consequently, the mean squared error of this estimate is:

$$\frac{1}{P} \mathrm{Var}_q[w(\theta) + b(\theta)] + E_q[\theta^2] + E_q[b(\theta)]^2.$$

If we compare such a biased estimator with a second es-


1.0
0.5
0.0
−0.5
−1.0
log(Estimated BF / True BF)
Pop Exchg
200,1,0
100,2,0
20,10,0
10,20,0
2,100,0
1,200,0
100,1,1
20,5,5
10,15,5
20,1,9
10,10,10
M, B, K

1=Z used). In section 2.6 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 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 e ect 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 estim-

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

$$\frac{\widehat{Z_{t-1}}(\theta_{t-1}^{(p)})}{Z_t(\theta_{t-1}^{(p)})} = \frac{1}{M}\sum_{m=1}^{M} \frac{\gamma_{t-1}(u_t^{(m;p)} \mid \theta_{t-1}^{(p)})}{\gamma_t(u_t^{(m;p)} \mid \theta_{t-1}^{(p)})}; \qquad (13)$$

where $u_t^{(m;p)} \sim f_t(\cdot \mid \theta_{t-1}^{(p)})$. 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).

Data point tempering For the SMC approach to be efficient we require that the sequence of distributions $\{\pi_t\}$ be chosen such that π_0 is easy to simulate from, π_T is the target of interest and the intermediate distributions 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. exposit686(e)(osit11 s-23Sk14(fou9.9626 Tf -11.695 -8Td 82 Td [(the)-517/F7 6.973

3.2 Application to precision matrices

In this section we examine the performance of the SMC sampler, with MCMC proposal and data-tempered target distributions, for estimating the evidence in an example in which is of moderately high dimension. We consider the case in which = ^{1} is an unknown precision matrix, f (yj) is the d-dimensional multivariate Gaussian distribution with zero mean and p( ) is a Wishart distribution W(; V) with parameters d and V 2 R^{d} ^{d}. Suppose we observe i.i.d. observations y = f y$_i$ g$_{i=1}^{n}$, where y$_i$ 2 R^{d}. The true evidence can be calculated analytically, and is given by

$$p(y) = \frac{1}{nd=2} \frac{_d(\frac{+n}{2})}{_d(\frac{}{2})} \frac{V^{1} + \sum_{i=1}^{n} y_i y_i^{T}}{jVj^{\frac{}{2}}}; \quad (18)$$

where $_d$ denotes the d-dimensional gamma function. For ease of implementation, we parametrise the precision using a Cholesky decomposition ^{1} = LL^{0} with L a lower triangular matrix whose (i; j)'th element is denoted a$_{ij}$.

As in section 2.3, we write f (yj) as (yj)=Z( ) as follows

$$f \quad f y_i g_{i=1}^{n} j \quad ^{1} = j2 \quad j^{n=2} \exp \quad \frac{1}{2} \sum_{i=1}^{n} y_i^{0} \quad ^{1} y_i \quad ;$$

[(n)ep594 Td]TJ/Fst-atrices

where in some of the experiments that follow, Z( ) = j2 j$^{n=2}$ is treated as if it is an INC. In the Wishart prior, we take = 10 + d and V = I$_d$.

Taking d = 10, n = 30 points were simulated using y$_i$ MVN (0$_d$; 0:1 I$_d$). The parameter space is thus 55-dimensional, motivating the use of an SMC sampler TJ/F8 9.9626 Tf 5.016 an SMC sampler TJ33(matrices)]TJ0 -24 TJ354 c

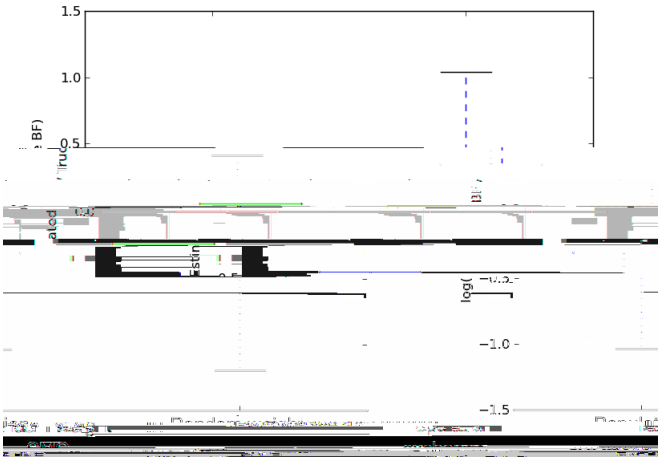


Fig. 3: Box plots of the results of population exchange and random weight SMC.

the results produced by this method in comparison with those from Friel (2013).

We observe that the median of the random weight SMC estimates is more accurate than that of the population exchange estimates - the bias introduced through using an internal Gibbs sampler in place of an exact sampler does not appear to accumulate su ciently to a ect the results (this issue is explored further in the following section). However, it has slightly higher variance than population exchange (much higher than SAVIS and MAVIS). This high variance can be attributed to two factors:

1. Since the SMC sampler begins with points sampled from the prior, larger changes in are considered than in the IS approaches, thus the estimates of the ratio of the normalising constants require more importance points to be accurate - the results suggest that the budget of 200 Gibbs sweeps is insu cient. This is the opposite situation to that encountered in section 2.6.2, where the changes in are small and the estimates of the ratio of the normalising constants are accurate with small numbers of importance points.
2. It's been frequently observed (cf. Lee and Whiteley (2015)) that, as suggested by the asymptotic variance expansion, in some instances the rst few iterations of an SMC sampler contribute substantially to the ultimate error. This issue arises since the forgetting of the sampler doesn't suppress the terms that the initial errors contribute to the asymptotic variance enough to compensate for the fact that they're much larger than the nal ones. This is due, when using data point tempering in the manner we have here, to the much larger relative discrepancy between the rst few distributions in the sequence than between later distributions.

We conclude that the random weight SMC method is a viable approach to estimating Bayes' factors for these models, but that care should be taken in tuning the weight estimates and choosing the sequence of SMC distributions.

3.4 Biased Weights in SMC

3.4.1 Error bounds

We now examine the e ect of using inexact weights on estimates produced by SMC samplers. By way of theoretical motivation of such an approach, we demonstrate that under strong, but standard (cf. Del Moral (2004)), assumptions on the mixing of the sampler, if the approximation error is su ciently 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 eld approximation and its

ployed but this formalism allows for a straightforward statement of the result):

A1 (Bounded Relative Approximation Error) There exists < 1 such that:

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

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

$$\sup_{t \in \mathbb{N}} \inf_{x,y} \frac{dM_t(x; \ )}{dM_t(y; \ )} \quad (M):$$

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

$$\sup_{t \in \mathbb{N}} \inf_{x,y} \frac{G_t(x)}{G_t(y)} \quad (G):$$

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 $_p$ and $\tilde{}_p$ 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, , 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:

$$\sup_{n \in \mathbb{N}} \| \ _n \quad \hat{}_n \|_{TV} \quad \frac{4 \ (1 \quad (M))}{^3(M) \ (G)}:$$

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, together with the empirical results presented below, it does suggest that further investigation of such approximations is warranted particularly in settings in which unbiased estimators are not available.

3.4.2 Empirical results

Fig. 4: The estimated bias in the log evidence estimates of the true (black solid), unbiased random weight (black dashed), biased random weight (grey solid) SMC algorithms using MCMC kernels, and the estimated bias when using the biased random weight algorithm with perfect mixing (grey dashed).

Fig. 5: The estimated MSE in the log evidence estimates of the four SMC samplers (same key as gure 4).

gorithm with true weights, and only a small bias is observed in the unbiased random weight sampler (this bias is likely to be due to the relatively small number of replications). Bias does accumulate in the biased random weight sampler, but we note that the level of bias appears to stabilise. This accumulation of bias means that one should exercise caution in the use of SMC samplers with biased weights. However, we observe that perfect mixing substantially decreases the bias in the evidence estimates from the algorithm. Also, in this case we observe that the bias does not accumulate su ciently to give poor estimates of the evidence. Here the standard deviation of the nal log evidence estimate over the random weight SMC sampler runs is approximately 0.4, so the bias is not large by comparison.

3.5 Discussion

In section 2.6 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 imply 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.

4 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

results for the exact algorithm indicate that the variance of the evidence estimates we use is su ciently small that this e ect is negligible.

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 note that the bias in these biased weight methods may be small compared to errors resulting from commonly accepted approximate techniques such as ABC.

For biased IS, in section 2.5 we show that the error of estimates from low-variance biased methods can be less than those from unbiased methods of higher variance. This is comparable to a result for biased MCMC methods (Johndrow et al 2015), where it is shown that the error of estimates from a computationally cheap biased MCMC can be less than those from an expensive exact MCMC. In both cases, given a nite computational budget, it is not always the case that this budget should be spent on guaranteeing the exactness of the sampler if minimizing approximation error is the objective.

A similar choice concerning the allocation of computational resources arises in SMC. Here, one does need to be especially careful about the use of biased SMC, due to the possible accumulation of bias over SMC iterations. One might expect this accumulated bias to outweigh any bene ts a reduced variance may bring. For this reason we advise caution in the use of biased SMC in general. This paper does, however, indicate that there may exist cases where biased SMC is useful, through: the theoretical result that under strong mixing conditions bias does not accumulate unboundedly; the empirical evidence that fast mixing may reduce the accumulation of bias; and the empirical results where we observe (in a situation where the distance between successive targets decreases) that the rate at which bias accumulates decreases with time.

Acknowledgements The authors would like to thank Nial Friel for useful discussions, and for giving us access to the data and results from Friel (2013).

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 e cient 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)

$f_t(\cdot|\cdot_{t-1})$ 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:

$$e_t(\mathbf{x}_t) = \eta_0(\cdot_0)\prod_{s=1}^{t} K_s(\cdot_{s-1};\cdot_s)\prod_{s=1}^{t}\prod_{m=1}^{M} f_s(u_s^m|\cdot_{s-1})$$

and hence its importance weight, $W_t(\mathbf{x}_t)$, should be:

$$\frac{\gamma_t(\cdot_t)\prod_{s=0}^{t-1} L_s(\cdot_{s+1};\cdot_s)}{\eta_0(\cdot_0)\prod_{s=1}^{t} K_s(\cdot_{s-1};\cdot_s)} \frac{\prod_{s=1}^{t}\frac{1}{M}\sum_{m=1}^{M}\left[f_{s-1}(u_s^m|\cdot_{s-1})\prod_{q\neq m} f_s(u_s^m|\cdot_{s-1})\right]}{\prod_{s=1}^{t}\prod_{m=1}^{M} f_s(u_s^m|\cdot_{s-1})}$$

$$= \frac{\gamma_t(\cdot_t)\prod_{s=0}^{t-1} L_s(\cdot_{s+1};\cdot_s)}{\eta_0(\cdot_0)\prod_{s=1}^{t} K_s(\cdot_{s-1};\cdot_s)}\prod_{s=1}^{t}\frac{1}{M}\sum_{m=1}^{M}\frac{f_{s-1}(u_s^m|\cdot_{s-1})}{f_s(u_s^m|\cdot_{s-1})}$$

$$= W_{t-1}(\mathbf{x}_{t-1})\,\frac{\gamma_t(\cdot_t)L_{t-1}(\cdot_t;\cdot_{t-1})}{\gamma_{t-1}(\cdot_{t-1})K_t(\cdot_{t-1};\cdot_t)}\,\frac{1}{M}\sum_{m=1}^{M}\frac{f_{t-1}(u_t^m;\cdot_{t-1})}{f_t(u_t^m|\cdot_{t-1})},$$

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

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

$$\frac{p(\cdot_t)f_t(y|\cdot_{t-1})}{p(\cdot_{t-1})f_{t-1}(y|\cdot_{t-1})}\,\frac{1}{M}\sum_{m=1}^{M}\frac{f_{t-1}(u_t^m|\cdot_{t-1})}{f_t(u_t^m|\cdot_{t-1})} =$$

$$\frac{p(\cdot_t)\ell_t(y|\cdot_{t-1})}{p(\cdot_{t-1})\ell_{t-1}(y|\cdot_{t-1})}\,\frac{1}{M}\sum_{m=1}^{M}\frac{\ell_{t-1}(u_t^m|\cdot_{t-1})}{\ell_t(u_t^m|\cdot_{t-1})}$$

yielding the algorithm described in section 3.1.1 as an exact SMC algorithm on the described extended space.