SEF VRES
Topics in Computational Bayesian Statistics

Ryan Kelly

January 30, 2018

1 Introduction

- ABC? Motivation. (Taken from Topics in Bayesian overview QUT) Bayesian statistics provides a framework for a statistical inference for quantifying the uncertainty of unknowns based on information pre and post data collection. This information is captured in the posterior distribution, which is a probability distribution over the space of unknowns given the observed data. The posterior distribution has the form:

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta) \tag{1}$$

The ability to make inferences based on the posterior essentially amounts to the ability to efficiently simulate from the posterior distribution, which can generally not be done perfectly in practice.

**** Appoximate Bayesian computation (ABC) can be used when the likelihood is unavailabe analytically or computationally but simulation from the model is straightforward.

- Limitations of SMC ABC - useful for single dataset

- Idea: Recycling simulations for SMC ABC

2 Bayesian Statistics Background

-posterior

-Metropolis-Hastings

- What happens when likelihood is intractable?

3 ABC Algorithms

3.1 ABC posterior

ABC targets the 'approximate' posterior:

$$\pi_\epsilon(\theta, x|y) \propto K_\epsilon(\rho(y, x))f(x|\theta)\pi(\theta) \tag{2}$$

(C11 ABC) - discrepancy $\rho(y, x)$ ABC compares the observed data, y, with the simulated data, x. This comparison is usually done on the basis of a summary statistic, $S(y)$. The comparison involves a discrepancy function $\rho(S(y), (x))$ which measures the 'closeness' between the observed

and simulated data.

likelihood $f(x|\theta)$ is the likelihood but now evaluated on the simulated data. In practice this computation does not need to be performed.

$K_\epsilon(\rho(y,x))$ is a weighting function that gives higher weight the closer the simulated summary statistic is the the observed summary statistic. One common choice is to set $K_\epsilon(\rho(y,x)) = 1(\rho(y,x) \leq \epsilon)$ where $1(\cdot)$ is the indicator function.
This gives a joint distribution over θ and the simulated data x. ABC algorithms sample over this joint densty, so as a by-product sample from the marginal θ by ignoring the x samples.

(choice of epsilon): The choice of ϵ in ABC rejection represents a trade-off between accuracy and computational effort. A smaller value of ϵ forces a closer match between observed and simulated data (and leads to higher accuracy) but will decrease the acceptance rate meaning that more computation is required to obtain the desired number of samples.

(sources of error): There are effectively two sources of error in the ABC approximation. The first is from the summary statistic not being sufficient. The second is the necessary introduction of the tolerance ϵ.

weighting function

3.2 ABC Rejection

- ABC rejection -¿ explain how works Importance sampling Because the ABC target is defined over the joint space of the parameter θ and the simulated data x, it is necessary to devise an importance distribution that exists over the same space. $g(\theta, x) = g(\theta)f(x|\theta)$ where g$(\cdot)$ is some general proposal over the θ space. The simulated data x gets drawn from the likelihood conditional on the proposed θ. To implement importance sampling, we generate a large sample of size M from the importance distribution, $\{\theta_i, x_i\}_{i=1}^M \overset{iid}{\sim} g(\theta, x)$. This size M is determined by computational constraints. Then in order to ensure that the samples are from π_ϵ the following weighting must be performed:

$$w_i = \frac{K_\epsilon(\rho(y,x))\cancel{f(x_i|\theta_i)}\pi(\theta_i)}{g(\theta_i)\cancel{f(x_i|\theta_i)}} \tag{3}$$

where we take the target density and divide it by the importance distribution. Note that the likelihood terms cancel and no intractable terms are left. If we choose $g(\theta) = \pi(\theta)$ and let $K_\epsilon(\rho(y,x)) = 1(\rho(y,x) \leq \epsilon)$ then w_i is either 0 or 1 then we ... [Choose method selected for method].

strengths: useful for testing different sets of summary statistics(Nunes and Balding, 2010)-¿[keep referencing?] or analysing multiple datasets from same model since the M prior predictive simulationd vsn be re-used (i.e. they are independent of the actual observed data).
weaknesses: (1) highly inefficient if the posterior is different to the prior (too high ϵ is required to obtain a reasonable size sample from the ABC posterior). (2) Heavy storage requirements

3.3 MCMC ABC

-MCMC ABC explanation Idea of MCMC: construct a Markov chain whose limiting distribution is given by the target distribution of interest, here the ABC posterior, π_ϵ. As in importance sampling, we require a proposal distribution on the joint space θ and x. However, our proposed value is allowed to depend on the state of the Markov chain, $q(\theta^*, x^*|\theta, x)$. Here we select $q(\theta^*, x^*|\theta, x)$

$= f(x^*|\theta^*)q(\theta^*|\theta)$ where $q(\theta^*|\theta)$ is an arbitrary proposal over the θ space and the term $f(x^*|\theta^*)$ tells us we must draw simulated data from the likelihood conditional on the proposed θ^*. This choice leads to the following Metropolis-Hastings ratio:

$$\frac{K_\epsilon(\rho(y,x^*))\cancel{f(x^*|\theta^*)}\pi(\theta^*)q(\theta|\theta^*)\cancel{f(x|\theta)}}{K_\epsilon(\rho(y,x))\cancel{f(x|\theta)}\pi(\theta^*)q(\theta^*|\theta)\cancel{f(x^*|\theta^*)}} \tag{4}$$

Proposal leads to cancellation of all intractable likelihood terms.
stengths: - tends to be more efficient than ABC rejection
weaknesses: -requires tuning of the proposal distribution q (same as MCMC) - method can get 'stuck' in low probability regions (thus long runs are often required to achieve a posterior sample with a reasonable ESS). - must be re-run and tuned for each new dataset or summary statistic selection in contrast to ABC rejection.

[mention Picchini and Forman (2014) early rejection]

3.4 SMC ABC

-SMC ABC (focus Drovandi 2011) (overview) For ABC it is natural to define a sequence of targets in terms of non-increasing sets of tolerances $\epsilon_1 \leq \epsilon_2 \leq \epsilon_T$:

$$\pi_t(\theta, x|y, \epsilon) \propto f(x|\theta)\pi(\theta)1(\rho(x,y) \leq \epsilon_t), \text{for t} = 1, ...T \tag{5}$$

Then SMC traverses a set of N 'particles' through the sequence of targets by iteratively applying re-weighting (importance sampling), re-sampling and mutation steps. (SMC ABC replenishment algorithm of Drovandi)
strengths: - better with multi-modal posteriors and is easy to adapt - tends to be more efficient than ABC rejection and MCMC ABC. This comes from the proposal distribution for θ, which gets determined adaptively. -Very little tuning required - There is natural stopping rules for SMC ABC. We can stop when the MCMC acceptance rate falls below some threshold (e.g. 1%) or if some target tolerance is achieved.
weaknesses: -Method must be re-run and tuned for each new dataset or summary statistic selection. -Duplicated particles can be a problem for very low ϵ (check if relevant -¿ C.Drov. 2011 algorithm). - Not as simple to implement as ABC rejection and MCMC ABC.

-SMC ABC (focus Drovandi 2011): (Drovandi & Pettitt, 2011) Intro bits... Include? ... MCMC used to jitter particles.... weights ....

An example of an SMC ABC algorithm is Drovandi 2011 Macroparasite ... When an MCMC kernel is used, the particle values do not get changed between adjacent targets, only the weights. The incremental weights in this context are:

$$\tilde{w}_t^i \propto \frac{1(\rho(x_{t-1}^i, y) \leq \epsilon_t)}{1(\rho(x_{t-1}^t, y) \leq \epsilon_{t-1})} \tag{6}$$

such that $W_t^i = \tilde{w}_t^i W_{t-1}^i$. Clearly the current particle satisfies the tolerance ϵ_{t-1}, and the incremental weight will be proportional to 1 if it satisfies the tolerance ϵ_t. Otherwise the incremental weight will be zero rendering the particle useless. Therefore after the re-weighting step there will be $\leq N$ particles remaining with non-zero weight, referred to hereafter as 'alive' particles. To overcome degeneracy in this situation, the population can be boosted back to N by resampling from the alive particles proportional to their weight. This has the effect of duplicating the particles with high weight. To ensure particle diversity these resampled particles can be moved according to an MCMC kernel that is invariant for the target involving ϵ_t. (Mention?) This approach differs from Chopin (2002) in that only the resampled particles are moved not all N particles; the particles already satisfying the next tolerance ϵ_t are left alone.

The incremental weighting fomula (MENTION WHAT EQUATION?) would suggest that the next tolerance, ϵ_t, should be chosen such that there is a certain number of alive particles. Setting ϵ_t too low will result in either no or free alive particles. However, choosing ϵ_t too high will result in many targets that require traversing. Nonetheless, the re-weighting step lends itself naturally to the sequence of tolerances being determined dynamically. This can be achieved by sorting the particles by their discrepancies, ρ^i, and dropping a proportion of the particles, α, with the highest discrepancy. That is, the next tolerance, ϵ_t, is dynamically taken as the (1 - α)th empirical quantile of the particles discrepancies.

In an attemp to move the resampled particle with a probability close to 1 - c (with c set small), Drovandi and Pettitt (2011) use the approach:

$$R_t = \frac{log(c)}{log(1 - p_{t-1}^{acc})} \tag{7}$$

Adapting the R_t is crucial in the context of ABC, since as ϵ_t decreases the acceptance rate tends to decrease too. Therefore this intervention is needed to help avoid ending up with too many duplicated particles.

MCMC proposal distribution for the parameter at each iteration, $q_t(\cdot|\cdot)$ can be made adaptive since there are $N - \alpha N$ particles satisfying the current target at each iteration. For example, these particles can be used to estimate the covariance matrix required for a multivariate normal or t random walk proposal. Drovandi and Pettitt (2011) algorithm has the stength of its fully adaptive nature. The only tuning parameters consist of $\epsilon_1, \epsilon_T, \alpha and c$. A reasonable choice for c is 0.01 and a sensible choice for α is 0.5, that is to drop half the particles at each iteration. (CONTINUE?)

4 Recycling in SMC ABC

From (South, Pettitt, & Drovandi, 2016) Ideas from IS and AIS can be applied to SMC in order to weight particles appropriately for the posterior. By reweighting particles in this way, it is possible to achieve a higher ESS from the posterior compared to the final samples $\{\theta_T^i\}_{i=1}^N$ alone. In addition to achieving this improvement is ESS, recycling particles can offer an improved an improved ability to sample from targets with completes landscapes (CITE? Nguyen et al., 2016). In situations where the target is multimodal, recycling past particles may help to recover samples from well sepsrated modes that may gave otherwise not appeared in the posterior approximation due to resampling.

Generally IS is used to weight samples drawn from an importance distribution in order to calculate quantities of interest with respect to a target distribution. Standard IS weights κ^m in a Bayesian setting for samples $\{\theta^m\}_{m=1}^M$ drawn drawn independently from an importance distribution $q^\phi(\cdot)$ with parameter ϕ are

$$\kappa^m = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}^m)\pi(\boldsymbol{\theta}^m)}{q^\phi(\boldsymbol{\theta}^m)} \tag{8}$$

The above weights are based on a single importance distribution. The basic AIS method (CITATION?) involves targeting $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ through a sequence of importance distributions $q^{\phi_t}(\cdot)$ for $t = 0, \ldots, T$ where ϕ_t denotes the parameter of the t-th importance distribution. The parameter ϕ_0 for the initial importance distribution $q^{\phi_0}(\cdot)$ is predetermined. The AIS weights found for M_t samples, $\{\boldsymbol{\theta}\}_{m=1}^{M_t}$, drawn independently from q^{ϕ_t} are

$$\omega_t^m = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_t^m)\pi(\boldsymbol{\theta}_t^m)}{q^{\phi_t}(\boldsymbol{\theta}_t^m)}, \text{ for } m = 1, \ldots, M_t \text{ and } t = 0, \ldots, T. \tag{9}$$

After normalising the weights in (PREVIOS EQ) to $\{\Omega_t^m\}_{m=1}^{M_t}$, the particle set $\{\Omega_t^m, \boldsymbol{\theta}_t^m\}^{M_t} m = 1$ can be used to estimate the parameter ϕ_{t+1} for $q^{\phi_{t+1}}(\cdot)$. This continues until the ESS of the weighted particle set, $ESS_t = 1/\Sigma_{m=1}^{M_t}(\Omega_t^m)^2$, shows little improvement. The result is $T+1$ weighted particle sets, each representing a sample from the posterior. These weighted particle sets could be combined together uniformly with the hope of improving precision, but this does not take account of the fact that early AIS distributions may be a poor approximation of the target and that each particle set has a different ESS.

(INCLUDE?) The ideas behind AIS are similar to SMC likelihood annealing with an independent proposal, though the target is also adapted in SMC. This adapation of the target in SMC makes it a more robust method than AIS which attempts to target the posterior from the start. It may then seem desirable to apply ideas from AIS by using the SMC targets π_t or independent proposal distributions q^{ϕ_t} for $t = 0, \ldots, T$ as importance distributions.

[CITE RIGHT] Nguyen et al. (2014) apply concepts from AIS and from Gramacy et al. (2010) to reuse the accepted particles from π_t from $t = 0, \ldots, T$, that is the particles that remain at the end of (INCLUDE?: line 17 algorithm). To recycle particles $\{\boldsymbol{\theta}_t^i\}_{i=1}^N$ from the t-th power posterior, the target is the posterior π_T and the importance distribution is the t-th power posterior, π_t.S

In the context of our independent SMC method, it is possible to recycle all proposals from all temperatures (not just the final samples from each power posterior) by using the $q^{\phi_t}(\cdot)$ as importance distributions. The simple concept of extending the current recycling methods to allow for the independent proposal q^{ϕ_t} as the importance distribution has the potential to offer substantial advantages in terms of increased ESS from the posterior.

(INCLUDE IN SAME WAY?) The recycling schemes that follow focus on using weighting schemes similar to AIS to reuse past particles and proposals in SMC. Nguyen et al. (2016) explore some methods to efficiently combine samples from different important distributions in the context of power posterior recycling. (HMMMM) These methods are described and extended to the case of reusing all independent proposals below.

COMBINED IMPORTANCE SAMPLING RECYCLING

Whether recycling samples from the power posteriors or all independent proposals, it is relatively simple to obtain $T+1$ weighted particle sets targeting the posterior, each with a different ESS. Gramacy et al. (2010) propose combining these particle sets by weighting set t by its contribution λ_t to the total ESS, where $\lambda_t = ESS_t/\Sigma_{l=0}^t ESS_l$ and $t = 0, \ldots, T$. This simple approach has the benefit that the required quantities have already been computed. (FROM SOURCE) We refer to this general approach as combined importance sampling (CIS) recycling.

INDEPENDENT PROPOSAL RECYCLING

The form of the AIS weights in (EQ W/ NON NORMALISED WEIGHTS) can be applied directly to SMC where $q^{\phi_t}(\cdot)$ is the independent proposal distribution used to make proposals $\{\boldsymbol{\theta}_t^m\}_{m=1}^{R_t N}$ for π_t. Combining these weighted particle sets based on their contribution to the total ESS results in the following normalised weights

$$\tilde{\Omega}_t^m = \lambda_t \Omega_t^m, \text{ for } m = 1, \ldots, R_t N and t = 0, \ldots, T. \tag{10}$$

Using these normalised weights results in an overall ESS targeting the posteriot of $\Sigma_t^T = 0 ESS_t$

5 Examples

5.1 Toy Example

model: Binomial simulated data, prior distributions: parameter interested in: theta, treat as unknown. simulate from binomial distribution

Transform: A reasonable suggestion for the proposal distribution would be a Gaussian random walk (WORDING CORRECT?). However, if applied to theta (which only has a range between 0 and 1) many proposals will be made outside of this range which would be inefficient. Instead we want to transform θ to a parameter that covers the range $-\inf < \phi < \inf$. The transformation used was $\phi\ ln(\frac{\theta}{1-\theta})$

(Mention MCMC approach briefly) SMC approach: Random walk, Independent proposals Random walk:

Independent proposal:

Transform:

5.2 G-and-k example

-model

SMC ABC approach with recycling was applied on the g-and-k distribution described in (Drovandi, Pettitt, & Lee, 2015) Reference Rayner and MacGillivray as well?

This example focuses on the g-and-k distribution. This is a quanitle distribution which can be defined in terms of its quantile function. Such functions can be formulated to create more flexible distributions than other standard distributions. This quantile function, which can also be interpreted as a transformation of a standard normal random variate, has the following form

$$Q(z(p);\boldsymbol{\theta}) = a + b\left(1 + c\frac{1-exp(-gz(p))}{1+exp(-gz(p))}\right)(1+z(p)^2)^k z(p) \tag{11}$$

Here p denotes the quantile of interest while $z(p)$ represents the quantile function of the standard normal distribution. The model parameter is $\boldsymbol{\theta} = (a, b, c, g, k)$, though common practice is to fix c at 0.8 (mentin Rayner?). While the likelihood function can be computed numerically, it is much more expensive than simulating from the model which can be cheaply implemented for quantile distributions via the inversion method.

simulated data, prior distributions The observed dataset consists of (HOW MANY?) independent draws from the g-and-k distribution with $a = 3, b = 1, c = 0.8, g = 2 and k = 0.5$
auxiliary model:
Chris Drovandi suggests... three component normal mixture model with 8 parameters as the auxiliary model. A mixture model is a suitable choice for an auxiliary distribution as it can be made arbitrarily flexible while maintaining a tractable likelihood function.
ABC IS method, based on the score vector, appeared to not have any

- results/ ilustration of proposed method.

6 Discussion

- quick summary of report

- limitations

-future work

References

Drovandi, C. C., & Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate bayesian computation. *Biometrics*, *67*(1), 225-233.

Drovandi, C. C., Pettitt, A. N., & Lee, A. (2015, 02). Bayesian indirect inference using a parametric auxiliary model. *Statist. Sci.*, *30*(1), 72–95. Retrieved from `https://doi.org/10.1214/14-STS498` doi: 10.1214/14-STS498

South, L. F., Pettitt, A. N., & Drovandi, C. C. (2016). *Sequential monte carlo for static bayesian models with independent mcmc proposals.* Retrieved from `https://eprints.qut.edu.au/101729/`