Incorporating Model Uncertainty into the Sequential Importance Sampling Framework using a Model Averaging Approach

or

Trans-Dimensional Sequential Importance Sampling

CREEM Technical report 2007-06

Christopher Lynam, Ruth King, Len Thomas and Steve Buckland

National Centre for Statistical Ecology

Centre for Research into Ecological and Environmental Modelling

University of St. Andrews, St. Andrews, Scotland KY16 9LZ

ChrisL@mcs.st-amd.ac.uk

Abstract

A sequential Bayesian Monte Carlo approach is proposed in which model space can be explored during the Sequential Importance Sampling (SIS, a.k.a. Particle Filtering) fitting process. The algorithm allows model space to be explored while filtering forwards through time and takes a similar approach to Reversible Jump Markov Chain Monte Carlo (RJMCMC) strategies, whereby parameters jump into and out of the model structure. Possible efficiency gains of the new Trans-Dimensional SIS routine are discussed and the approach is considered most beneficial when the exploration of large model space in the SIS framework is desired.

Keywords: Particle filtering, model space, sequential Monte Carlo, Markov chain

Introduction

Commonly, data analysts seek to choose objectively a single 'best' model and subsequently draw inference using that model. Model selection is thus a fundamental part of many frequentist or Bayesian modelling procedures and a crucial step if inference is to be made with any reliability. Much progress concerning the selection of a model from a set using information theory (such as the Akaike Information Criterion, AIC, and its many variants) has been made in recent years (Seghouane and Amari, 1997; Wager *et al*., 2007; Link and Barker, 2006; Burnham and Anderson, 2002). However, when inference is based upon a single model, predictions may not be robust. A sensible general framework that incorporates model uncertainty into estimation and prediction is that of Bayesian model averaging (Hoeting *et al*., 1999; Wintle *et al*., 2003; Wang *et al*., 2004). In this context, many models are fitted and weights are attached to each so that the whole suite of models can be used to generate a composite forecast (Buckland *et al*., 1997; Kass and Raftery, 1995). Bayesian Monte Carlo approaches, such as Markov Chain Monte Carlo (MCMC) and Sequential Importance Sampling (SIS, a.k.a. Particle Filtering) are able to incorporate model selection and the generation of model weights within the fitting process. MCMC can utilise the Reversible Jump method (RJMCMC, see King and Brooks, 2002ab) to allow parameters to jump into and out of the model structure, while a similar sequential method for use with SIS has been developed by Vermaak *et al*. (2003) and termed Trans-Dimensional Sequential Monte Carlo (TD-SMC). The TD-

SMC method is non-iterative and is based on a generalisation of importance sampling to spaces of variable dimension. Here, we suggest a related procedure, 'Trans-Dimensional Sequential Importance Sampling' (TD-SIS), whereby the dimension of the modelled structure for a sampled particle can be similarly altered during the SIS fitting framework. The major development in the TD-SIS algorithm is the use of a likelihood-based transition probability coupled with the storage of random seeds in order to reduce Monte Carlo error between particles and proposals. This advance should allow for more efficient exploration of model space and permit the investigation of model structures with high dimensionality.

Methods

Sequential Importance Sampling

Consider a system defined by unknown states (for example animal abundances or illegal immigrants) represented by a state vector $\mathbf{n}_t$, and by processes (such as recruitment or economic growth) represented by a single model with parameters $\boldsymbol{\theta}$. Then the joint prior distribution for the state vector and the parameters is

$$g(\boldsymbol{\theta})\times g_0(\mathbf{n}_0\,|\,\boldsymbol{\theta})\times\prod_{t=1}^{T_{\max}} g_t(\mathbf{n}_t\,|\,\mathbf{n}_{t-1},...,\mathbf{n}_0,\boldsymbol{\theta}) \qquad (1)$$

where $g(\boldsymbol{\theta})$ is the prior distribution on the parameters, $g_0(\mathbf{n}_0\,|\,\boldsymbol{\theta})$ is the prior distribution on the initial states $\mathbf{n}_0$ given the parameters at $t = 0$ and $g_t(\mathbf{n}_t\,|\,\mathbf{n}_{t-1},...,\mathbf{n}_0,\boldsymbol{\theta})$ is the prior distribution on the states $\mathbf{n}_t$ at time t given the previous states and the parameters. If $g_t(\mathbf{n}_t\,|\,\mathbf{n}_{t-1},...,\mathbf{n}_0,\boldsymbol{\theta})\equiv g_t(\mathbf{n}_t\,|\,\mathbf{n}_{t-1},\boldsymbol{\theta})$, then the model is first-order Markov, and is termed a state-space model (Buckland *et al.*, 2004; Newman *et al.*, 2006).

Although there are many potential algorithm to implement SIS, typically one draws a large number R of samples from the joint proposal distribution for the parameters and initial states, say $q(\mathbf{n}_0,\boldsymbol{\theta})=q(\boldsymbol{\theta})\times q_0(\mathbf{n}_0\,|\,\boldsymbol{\theta})$, and each independent set of parameters and initial states is termed a 'particle'. So, the r^{th} particle at $t = 0$ represents a single realization $\boldsymbol{\theta}_\mathrm{r}$ of the parameters of the population model, together with a single realization of the population in the initial year, $\mathbf{n}_{0,r}$. So, an empirical measure of the samples from the prior distribution on the parameters is given by

$$\hat{g}_R(\boldsymbol{\theta})\approx\frac{1}{R}\sum_{r=1}^{R}\delta_r(\boldsymbol{\theta}) \qquad (2)$$

where δ_r is the delta-Dirac mass located at $\boldsymbol{\theta}_\mathrm{r}$ (Doucet *et al.* 2001). Each particle can be projected forwards stochastically to time $t=1$ (in the chosen units of time), by simulating the state vector $\mathbf{n}_{t,r}$ from $g_t(\mathbf{n}_{t,r}\,|\,\mathbf{n}_{t-1,r},\boldsymbol{\theta}_r)$. The particles can then be resampled, with weights

$$w_{t,r}\propto L_{t,r}\times g(\boldsymbol{\theta}_r)\times g_{t-1}(\mathbf{n}_{t-1,r}\,|\,\boldsymbol{\theta}_r)\big/\big[q(\boldsymbol{\theta}_r)\times q_{t-1}(\mathbf{n}_{t-1,r}\,|\,\boldsymbol{\theta}_r)\big] \qquad (3)$$

where $L_{t,r} = f_t(\mathbf{y}_t \mid \mathbf{n}_{t,r}, \boldsymbol{\theta}_r)$ is the contribution to the likelihood for particle r from the data $\mathbf{y}_t$ at time point t. It is convenient to normalize these weights, so that $\sum_{r=1}^{R} w_{t,r} = 1$. If the proposal distribution is taken to be the prior distribution $q(\boldsymbol{\theta}_r) \times q_{t-1}(\mathbf{n}_{t-1,r} \mid \boldsymbol{\theta}_r) = g(\boldsymbol{\theta}_r) \times g_{t-1}(\mathbf{n}_{t-1,r} \mid \boldsymbol{\theta}_r)$, then the resampling weights are simply likelihood weights and they can be normalized thus: $w_{t,r} = L_{t,r} \Big/ \sum_{r=1}^{R} L_{t,r}$.

When the resampling scheme above is combined with the use of the posterior distribution of parameters at time t-1 as the prior distribution for time t, the algorithm is known as ‘bootstrap filtering’ (Gordon et al. 1993). For example, given data at time $t = 1$, the surviving particles $(\boldsymbol{\theta}_r, \mathbf{n}_{0,r}, \mathbf{n}_{1,r})$ are approximately a sample from the joint posterior distribution of parameters and states. These particles may be projected forwards to become the joint prior distribution for $t = 2$, with the state vector simulated from $g_2(\mathbf{n}_{2,r} \mid \mathbf{n}_{1,r}, \mathbf{n}_{0,r}, \boldsymbol{\theta}_r)$, and so on, until resampling at the final time point $T_{\max}$ has been carried out. The surviving particles are then an approximate sample from the joint posterior distribution of parameters and states.

Model Averaging by Sequential Importance Sampling: Theory

To incorporate model averaging, it is natural to extend the hierarchical framework of (1) to include prior information for each model considered:

$$g(\mathbf{M}) \times g(\boldsymbol{\theta} \mid \mathbf{M}) \times g_0(\mathbf{n}_0 \mid \boldsymbol{\theta}, \mathbf{M}) \times \prod_{t=1}^{T_{\max}} g_t(\mathbf{n}_t \mid \mathbf{n}_{t-1}, \ldots, \mathbf{n}_0, \boldsymbol{\theta}, \mathbf{M}) \qquad (4)$$

where $g(\mathbf{M})$ represents the prior distribution for the models $\mathbf{M}$, $g(\boldsymbol{\theta} \mid \mathbf{M})$ represents the prior distribution for the parameters $\boldsymbol{\theta}$ given the models and $g(\boldsymbol{\theta}, \mathbf{M}) = g(\mathbf{M}) \times g(\boldsymbol{\theta} \mid \mathbf{M})$ (King and Brooks, 2002a,b). We adopt a slightly different approach to King and Brooks (2002a,b) in that we define a maximal model with K parameters. Other models are obtained by fixing one or more of these parameters to a pre-determined constant, usually 0 (e.g. for a coefficient corresponding to a covariate) or 1 (e.g. for a parameter representing a probability). Thus, the prior distribution for the parameters has two components: a binary component that determines whether each parameter takes a fixed value, and conditional on it not taking a fixed value, a continuous probability density function. Thus, the sampling distribution of the joint prior distribution for the parameters and models is

$$g(\boldsymbol{\theta}, \mathbf{M}) \approx \frac{1}{R} \sum_{r=1}^{R} \prod_{j=1}^{K} [p_j g_j(\theta_{j,r})]^{\delta_{j,r}} (1 - p_j)^{1-\delta_{j,r}} \qquad (5)$$

where $\mathbf{p} = (p_1, \cdots, p_K)'$ is a vector of prior probabilities of whether each parameter is included in the model. $\boldsymbol{\delta} = \begin{pmatrix} \delta_{1,1} & \cdots & \delta_{1,R} \\ \vdots & \ddots & \vdots \\ \delta_{K,1} & \cdots & \delta_{K,R} \end{pmatrix}$ is a matrix of ones and zeros with $\Pr(\delta_{j,r} = 1) = p_j$ such that the number of parameters included in the model structure for particle r is $a_r = \sum_{i=1}^{K} \delta_{i,r}$. If we want to include parameter j in the model space of every particle, we simply set p_j and thus $\delta_{j,r}$ = 1 for all r. Remaining elements of $\mathbf{p}$ might be set equal, subject to the constraint $\sum_{j=1}^{K} p_j = k$, where k is judged *a priori* to be a suitable size of model (in terms of number of parameters) for the system, given available data. For example, if there are K = 20 parameters in the maximal model and we require 5 particular parameters to be in all models and we also judge that k = 10 parameters are appropriate for the model given the size of the dataset available, then we set $p_j = 1$ for the 5 parameters of our minimal model, and $p_j = 1/3$ for each of the remaining 15 parameters. So, the prior probability of model m is $\Pr(m) = \frac{1}{h} \prod_{j=1}^{K} \delta_j^m p_j$ where δ_j^m is the vector of ones and zeros corresponding to the parameters included in model structure m and the normalising constant $h = \sum_{m}^{\mathbf{M}^\mathbf{N}} \Pr(m)$ where $\mathbf{M}^\mathbf{N}$ is the number of model types possible.

Note that we have assumed, as is commonly the case in RJMCMC and SIS, that the individual priors $g_j(\theta_j)$ are independent. However, if there is strong correlation between parameters any transformation of the covariates or reparameterization that reduces these correlations should improve the performance. For regression-type model structures, a simple measure is to ensure that all covariates are centred on their mean, thus removing correlations between regression coefficients and the corresponding intercept.

The joint prior distribution for the models $\mathbf{M}$, parameters $\boldsymbol{\theta}$ and the state vector $\mathbf{n}_t$ is thus

$$g(\boldsymbol{\theta},\mathbf{M}) \times g_0(\mathbf{n}_0 \mid \boldsymbol{\theta},\mathbf{M}) \times \prod_{t=1}^{T_{\max}} g_t(\mathbf{n}_t \mid \mathbf{n}_{t-1},\ldots,\mathbf{n}_0,\boldsymbol{\theta},\mathbf{M}) \tag{6}$$

and given observations $\mathbf{y}_t$, at each time point t the posterior distribution $g(\mathbf{n}_0,\ldots,\mathbf{n}_T,\boldsymbol{\theta},\mathbf{M} \mid \mathbf{y}_1,\ldots,\mathbf{y}_T)$ is

$$\frac{g(\boldsymbol{\theta},\mathbf{M}) \times g_0(\mathbf{n}_0 \mid \boldsymbol{\theta},\mathbf{M}) \times \left(\prod_{t=1}^{T} g_t(\mathbf{n}_t \mid \mathbf{n}_{t-1},\boldsymbol{\theta},\mathbf{M}) f_t(\mathbf{y}_t \mid \mathbf{n}_t,\boldsymbol{\theta},\mathbf{M}) \right)}{f(\mathbf{y}_1,\ldots,\mathbf{y}_T)} \tag{7}$$

The overall likelihood of the particle at time $T \le T_{\max}$ is simply

$$L_r^T = \prod_{t=1}^{T} L_{t,r} = \prod_{t=1}^{T} f_t(\mathbf{y}_t \mid \mathbf{n}_{r,t},\boldsymbol{\theta}_r,\mathbf{M}) \tag{8}$$

and overall likelihood weights for particles simulated from the priors are

$$w_r^T = L_r^T \Big/ \sum_{r=1}^{R} L_r^T \tag{9}$$

In order to assess the performance of the particle set at any time point *t*, the equivalent number of iid samples in the set can be determined from the Effective Sample Size at time point *t* (Kong *et al*. 1994; Liu 1996), while the posterior model probabilities at time *t* indicate the support for each model given the data up to time *t*.

$$\mathrm{ESS}_t = \sum_{r=1}^{R}\left(w_r^T\right)^2 \Big/ \sum_{r=1}^{R}\left(w_r^{T\,2}\right) \qquad (10)$$

Model Averaging through Trans-Dimensional Sequential Importance Sampling

To illustrate simply the implementation of the proposed model selection technique, the SIS framework in the following simulation study is simplified by the exclusion of the resampling scheme of Gordon, Salmond, and Smith (1993). However, it is theoretically possible to conduct the resampling and model jumping methods in tandem. For this model averaging study, particles are generated as above and fit to data to generate overall likelihood weights (w_r^T see above) and then instead of the resampling stage the following dynamic 'model jumping' algorithm is conducted.

For each particle *r*, first simulate a vector of ones and zeros to assign to each row (*j*) in the *r*th column of $\boldsymbol{\delta}$, where each element indicates whether or not a parameter is included. This can be achieved by generating binary deviates with success probabilities given by **p**, although a different proposal distribution could be taken. For each parameter θ_j selected for inclusion in the *r*th particle, i.e. $\delta_{r,j} = 1$, parameter values are simulated from the proposal distribution $q(\boldsymbol{\theta})$. The initial states are then generated by simulating from $q_0(\mathbf{n}_0 \mid \boldsymbol{\theta})$. SIS proceeds as usual and likelihood weights are constructed sequentially. At pre-chosen time points *t*, or when the estimated sample size has reduced past a preset threshold, the particles may then be altered through model jumping, such that some particles will increase their model dimension while others particles will reduce it.

Model jumps are stochastic and for each particle the probability of a transition in model space is computed as follows. At time point *t*, where data are available, for the *j*th parameter or for a randomly chosen subset of $\boldsymbol{\theta}$, if θ_j is currently in the *r*th particle's model space ($\delta_{j,r} = 1$), generate a 'parallel' particle identical to particle *r* (i.e. with the same random numbers used for simulation from each stochastic process and with the identical values of each parameter and initial state) except with θ_j excluded ($\delta_{j,r} = 0$); if θ_j is not currently in the model, generate a parallel particle with θ_j included. Select the particle with θ_j included with transition probability $L_r^{T,\delta_{j,r}=1} \Big/ \left(L_r^{T,\delta_{j,r}=1} + L_r^{T,\delta_{j,r}=0}\right)$ where: $L_r^{T,\delta_{j,r}} = \prod_{t=1}^{T} L_{t,r}^{\delta_{j,r}}$; *t* = 1 is the first time point with data; and $L_{t,r}^{\delta_{j,r}}$ indicates the contribution of time point *t* to the likelihood at time *T* for particle *r* with parameter θ_j included in ($\delta_{j,r} = 1$) or excluded from ($\delta_{j,r} = 0$) the model space. Otherwise select the particle with θ_j excluded, i.e. with 1 minus the transition probability. Discard the parallel particle (or store it for future use, should this model be selected again at a later time point). If the random numbers used to generate deviates for each process of particle *r* are retained, then they can be re-used to generate the parallel particle (except for any associated with a dropped process, if the exclusion of a parameter results in the elimination of a process). If this supply of random numbers is exhausted, more can be generated (and stored) as they are needed, for example when a process is added to a model. Re-using random numbers in this way will reduce Monte Carlo variation between parallel particles.

The definition of the transition probability above will affect the validity of the method and also its efficiency. One should remember that although particles are not independent of their parallels, with which they share random numbers, they are independently generated of other particles. By making the direction of the jumps reversible and by taking the usual importance sampling assumption of very large sample size model space can be explored in the manner set out above with the following potential adjustments. The above probabilities intrinsically favour more complex models because the probability of retaining the simpler model is at most 0.5, while the more complex model is more likely to attain a greater likelihood simply due to its greater flexibility. Hence use of a penalized likelihood is likely to lead to better efficiency. e.g. if for particle *r* we select the model with θ_j included with probability $L_r^{T,\delta_{j,r}=1} \Big/ \left(L_r^{T,\delta_{j,r}=1} + e^1.L_r^{T,\delta_{j,r}=0}\right)$, this equates to using the Akaike Information Criterion to choose between particles and their parallels (Buckland *et al.*, 1997), increasing the chances of selecting the simpler model. In general, denote these transition probabilities by

$$u_r^T = L_r^{T,\delta_{j,r}=1} \Big/ \left(L_r^{T,\delta_{j,r}=1} + c.L_r^{T,\delta_{j,r}=0}\right) \tag{11}$$

where for example: $c=1$ corresponds to likelihood weights; $c=e^1$ to AIC weights $c=\sqrt{n}$ to Bayesian Information Criterion (BIC) weights, where *n* is sample size; or $c = pr$, where *pr* is the ratio of simple over complex model prior probabilities. If the prior probabilities on each model are equal (*pr* = 1) then the transition probabilities are equal to non-penalised transition probabilities. In contrast to RJMCMC, the reversibility condition (eqn 12) is not required by either the TD-SIS method or indeed the TS-SMC method of Vermaak *et al.* (2003).

$$g(\theta_{t,r}^{\delta_{j,r}=0}) \times u_r^t(\theta_{t,r}^{\delta_{j,r}=0}, \theta_{t,r}^{\delta_{j,r}=1}) = g(\theta_{t,r}^{\delta_{j,r}=1}) \times u_r^t(\theta_{t,r}^{\delta_{j,r}=1}, \theta_{t,r}^{\delta_{j,r}=0}) \tag{12}$$

If the above condition were satisfied by TD-SIS then particles would be just as likely to move to a parallel particle that poorly fits the data as they are to stay in a currently satisfactory particle. By formulating the transition probabilities as in equation 11, ‘good’ moves are more likely than ‘poor’ moves and the algorithm is thus relatively efficient. Once the algorithm is complete, the particles should be distributed across models such that the posterior model probabilities can be determined simply by the proportion of the particle set within each model.

Simulation study

A simple model structure utilising two processes was chosen for investigation by the simulation study. The model processes are described by the following equations:

$$\alpha_{y_r} = \frac{\exp(\beta - \gamma X_{y_r})}{1 + \exp(\beta - \gamma X_{y_r})} \tag{13}$$

$$\theta_{y_r} = \frac{\exp(\delta - \lambda X_{y_r})}{1 + \exp(\delta - \lambda X_{y_r})} \qquad (14)$$

where α_{y_r} is the probability of recruitment and θ_{y_r} is the probability of survival in year y_r, X_{y_r} is the state value, and β, γ, δ and λ are parameters to be estimated from the following prior distributions: $\beta \sim$ *normal*(0.5, 0.001); $\gamma \sim$ *normal*(0.00075, 0. 00008); $\delta \sim$ *normal*(0.55, 0.0001); $\lambda \sim$ *normal*(0. 000225, 0. 00005).

Using the above processes and an initial state value of 150 animals, four stochastic models were evaluated: M0, the full model with density dependent recruitment and survival; M1, density independent recruitment (γ set to 0) and density dependent survival; M2, density dependent recruitment but density independent survival (λ set to 0); M3, the simple model density independent recruitment and survival (γ and λ set to 0). Data were generated by a single stochastic realisation from M2 (Fig. 1) but were potentially reproducible by each model. However, deterministic modelling using the expected values of the parameters' prior distributions suggests that density dependence in one process is necessary to fit to the data (Table 1).

Results

With the model jumping algorithm in place, a quarter of a million independent particle simulations were made, from a set of 1 million particles including the non-independent parallel particles, for the two following types of transition probability: prior transition probabilities ($c = pr$ and $pr = 1$ and thus equal to likelihood transition probabilities without penalization in this case), and AIC transition probabilities ($c = e^1$). No difference, above that expected for Monte Carlo error, was found between the posterior support on each model using the two types of transition probability (Table 1). However, differences were evident when compared to 1 million independent simulations from the SIS algorithm without model jumping. M1 and M2 were favoured much more greatly by the standard SIS algorithm (100% of the Akaike model weights, see eqn 16) than by the TD-SIS (75% of particles ended up in these two models). However, the reduced set of particles in M0 and M3 after TD-SIS were much improved over the fully independent set created by the standard SIS method (i.e. fewer poor particles were included). Although, very little difference is visually evident between the model fits (Figs. 2 and 3) fewer independent particles are utilised by the TD-SIS method. In this case, the smaller particle set (¼ M particles) in the posterior distribution created using TD-SIS had an almost equal ESS for each model as the standard SIS algorithm attained using 1M independent particles (Table 1). If the efficiency of the TD-SIS method were improved (i.e. not all parallel particles need be modelled fully as they have been here) then the fewer independent samples required by the TD-SIS algorithm would prove a major benefit, particularly for large model space.

Discussion

The effectiveness of the algorithm in searching through parameter space is a function of the length of the time series and of the priors placed on whether or not parameters are in the model. If there is a trend in average model size with t, which is still evident by the end of the time series, this may indicate a poor prior choice for k, the size of the model. However, large changes in average model size, and model weights, may also occur due to outliers in the data. The algorithm goes some way to address this issue by calculating the transition probabilities (u_r^T) historically, so that the current and all previous time points at which data are available determine the probability of a jump through model space thus damping the effect of the outliers.

A useful and simple tool to investigate the relative model fit at time t may be a penalised likelihood score such as the Akaike Information Criterion (AIC),

$$\mathrm{AIC}_{m,t} = 2\left(a_m - \frac{1}{N_{m,t}} \sum_{r=1}^{N_m} \log L_{\mathrm{r}}^{\mathrm{T}} \right) \tag{15}$$

where $N_{m,t}$ is the number of particles at time t in model m and a_m is the number of parameters in model m, and the Akaike model weight is thus,

$$\mathrm{AW}_{m,t} = \exp\left\{-0.5\,\Delta\mathrm{AIC}_{m,t}\right\} \Big/ \sum_{k=1}^{M} \exp\left\{-0.5\,\Delta\mathrm{AIC}_{k,t}\right\} \tag{16}$$

where M is the number of models considered and $\Delta\mathrm{AIC}_{m,t} = \mathrm{AIC}_{m,t} - \min(\mathrm{AIC}_{all\,m,t})$. The Akaike model weights are determined by particles that are simulated independently for each model structure. The TD-SIS algorithm allows particles to jump through model space such that the posterior sample of parameters for each model is determined by those particles that have jumped into the model type. Although the parameters sets are still independent samples, the choice of which parameters to include in the model is determined sequentially through fitting to the data and altering the model space. In effect, a high-pass filter is created such that the posterior set for each model is under-represented by poor samples from the prior. This is reflected in the difference in the ΔAIC scores for the entire set of particles and their parallels (ΔAIC 1 M) and for the selected particle set only (ΔAIC ¼ M) (Table 1).

Auxiliary Transition Sequential Importance Sampling (AT-SIS)

The TD-SIS routine presented above is computationally expensive and particularly so for long time-series with large model space. We may improve the method through use of an 'auxiliary transition probability'. In order to evaluate the probability of a jump through model space using TD-SIS, one must calculate u_r^T (eqn 11), which is based on the product of each model's likelihood scores $L_r^{T,\delta_{j,r}} = \prod_{t=1}^{T} L_{t,r}^{\delta_{j,r}}$ over the present and all previous time points. To reduce the complexity of the calculation, one could calculate the likelihood scores, and thus the transition probability, approximately: for the particle currently modelled (say, $\delta_{j,r} = 0$), $\hat{L}_r^{T,\delta_{j,r}} = L_r^{T,k,\delta_{j,r}=0} = \prod_{t=k}^{T} L_{t,r}^{\delta_{j,r}=0}$ where $k > 1$; and for the parallel particle, $\hat{L}_r^{T,\delta_{j,r}=1} = L_r^{T,k,\delta_{j,r}=1} = \prod_{t=k}^{T} \hat{L}_{t,r}^{\delta_{j,r}=1}$. For the parallel particle ($\delta_{j,r} = 1$), the estimate of the likelihood is considered here to be based on updates of the modelled states to time T from the previously computed states of the current model ($\delta_{j,r} = 0$) up to time $k - 1$; hence the use of an estimated likelihood $\hat{L}_{t,r}^{\delta_{j,r}=1}$ at time t for the parallel particle.

For example, we might consider model jumping based on the current data year only ($T = k$). If the time-point k is 10 time intervals after the initial state we can estimate $\hat{L}_r^{10,\delta_{j,r}=0}$ and $\hat{L}_r^{10,\delta_{j,r}=1}$ using the previously modelled states (up to t = 9 with, say, $\delta_{j,r}$ = 0) and update to t = 10 using both the current and alternative model type (i.e. the parallel particle with $\delta_{j,r}$ = 1 that we might switch to). However, the actual difference in the likelihoods of the two particles will be under-estimated: i.e. $\left(L_r^{T=10,\delta_{j,r}=0} - L_1^{T=10,\delta_{j,r}=1}\right) > \left(\hat{L}_r^{T=10,k=10,\delta_{j,r}=0} - \hat{L}_1^{T=10,k=10,\delta_{j,r}=1}\right)$. So, making model jumps based on such approximate likelihoods would lead to incorrect posterior model probabilities, particularly if a particular model fits well in the later part of the time-series (and thus the probability of jumping to that model is high in those later years) but not in the earlier years. Therefore we must correct for the inaccuracy of the auxiliary movement; however, this will reduce the efficiency gain of the method. For successful model jumps using the estimate $\hat{u}_r^{T,k} = L_r^{T,k,\delta_{j,r}=1} \Big/ \left(L_r^{T,k,\delta_{j,r}=1} + c.L_r^{T,k,\delta_{j,r}=0}\right)$ we should calculate the entire history of the particle in the new model and determine what u_r^T would have been before jumping through model space. The weight of the particle can then be modified, in order to account for the approximation in the calculation, by

$$w_r^T \propto \frac{u_r^T}{\hat{u}_r^T} L_r^T \tag{15}$$

Altering the weights (after jumping) in this way would correct for the auxiliary calculation of pr(move) if the move is made. However, if a particle does not jump across models no correction is necessary, this is valid if the direction of any model jump is reversible and there are many particles in each model. Auxiliary model jumping is perhaps unwise if all the particles have converged on a single model because the technique might impose an incorrect model fidelity if more than one additional/fewer parameters are required to move to more suitable model space.

The efficiency gains by this auxiliary transition method are due to the removal of the requirement to calculate the complete history of states, likelihoods and weights for a parallel particle under the alternative model. However, once a move has been made and the selected parallel particle is fully specified, if we choose to consider the reverse jump through model space at a subsequent time point (assuming that the original particle data was saved and that the time points are not too widely spaced) we may want to model the original particle forward from the previous states so that the calculation of u_r^T is exact; since there may be little gain in the auxiliary method. However, in practice for very large model space it may be so unlikely to consider a reverse jump, rather than a jump to another model, so that this circumstance does not occur.

In summary, the TD-SIS algorithm manages to perform as well as the SIS routine (e.g. similar ESS by model, Table 1; Figs 2 and 3) but the posterior sample relies upon 1/(number of possible model types) fewer independent samples. The strength of the method relies upon further improvements to its efficiency so that the number of parallel particles that need to be explored do prohibit its successful implementation. However, given the many possibilities to approximate the 'parallel' (non-independent) particles this technique deserves further study.

References

Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997). Model selection: an integral part of inference. *Biometrics* **53**: 603-618.

Buckland, S.T., Newman, K.B., Thomas, L. and Koesters, N.B. (2004). State-space models for the dynamics of wild animal populations. *Ecol. Mod.* **171**: 157–175.

Burnham, K.P. and Anderson, D.R. (2002). Model Selection and Multimodel Inference: a practical information-theoretic approach. 2nd Edition. Springer-Verlag, New York, New York, USA. 488 pp.

Gordon, N.J., Salmond, D.J. and Smith, A.F.M. (1993). A novel approach to non-linear and non-Gaussian Bayesian state estimation. IEE-Proceedings F **140** (2), 107-133.

Doucet, A., de Freitas N. and N. Gordon (eds) (2001). Sequential Monte Carlo Methods in Practice. Springer-Verlag, 2001, ISBN 0-387-95146-6.

Kass, R.E. and Raftery, A.E. (1995). *Bayes Factors*. Journal of the American Statistical Association **90**(430), 773-795

King, R. and Brooks, S.P. (2002a). Model selection for integrated recovery/recapture data. *Biometrics* **58**, 841-851.

King, R. and Brooks, S.P. (2002b). Bayesian model discrimination for multiple strata capture-recapture data. *Biometrika* **89**, 785-806.

Kong, A., Liu, J.S. and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. Journal of the American Statistical Association. **89**, 278–288.

Link, W.A. and Barker, R.J. (2006). Model Weights and the foundations of multimodel Inference. *Ecology* **87**(10), 2626-2635.

Liu, J. S. (1996) Nonparametric hierarchical bayes via sequential imputations. *Annals of Statistics* **24**(3): 911-930.

Newman, K.B., Buckland, S.T., Lindley, S.T., Thomas, L. and Fernández, C. (2006). Hidden process models for animal population dynamics. *Ecological Applications* **16**, 74-86.

Seghouane, A.K. and Amari, S.I. (1997). The AIC criterion and symmetrizing the Kullback-Leibler divergence. *IEEE Transactions On Neural Networks* **18**(1), 97-106.

Tucker, B.C and Anand, M. (2005). On the use of stationary versus hidden Markov models to detect simple versus complex ecological dynamics, *Ecol. Model.* **185**, 177–193

Vermaak, J., Godsill, S.J. and Doucet, A. (2003). Radial basis function regression using trans-dimensional sequential Monte Carlo. In *IEEE Workshop on Statistical Signal Processing*, Pp 545-548. DOI 10.1109/SSP.2003.1289519

Wager, C., Vaida, F. and Kauermann, G. (2007). Model selection for penalized spline smoothing using Akaike Information Criteria. *Australian & New Zealand Journal of Statistics* **49** (2), 173-190

Wang, D.L., Zhang, W.Y. and Bakhai, A. (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression *Statistics In Medicine*. **23** (22), 3451-3467

Wintle, B.A., McCarthy, M.A., Volinsky, C.T. and Kavanagh, R.P. (2003). The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology* **17** (6), 1579–1590.

Tables

		M0	M1	M2	M3
Deterministic simulation All time points. R = 4 particles	ΔAIC 1 M	210.8	0.78	0	45.86
	AW	1 x 10^{-44}	40.3	59.7	7 x 10^{-9}
Deterministic simulation Six time points. R = 4 particles	ΔAIC 1 M	64.3	0.17	0	13.1
	AW	6 x 10^{-15}	47.9	52.0	7 x 10^{-4}
Standard SIS	N	249966	250325	249901	249808
(no resampling)	ESS 1 M	894	2305	1927	286
R = 1 M particles	ΔAIC	290	2.2	0	28
ESS(1 M) 4095	**AW**	**9 x 10^{-64}**	**0.25**	**0.75**	**5 x 10^{-7}**
	$N_{initial}$	62281	62989	62357	62373
TD-SIS ($c = pr$, pr = 1)	N_{final}	21691	95615	92266	40428
R = ¼ M particles	**N_{final}/R**	**0.08**	**0.38**	**0.37**	**0.16**
from 1 M (incl. parallels)	ESS ¼ M	875	2157	1811	257
ESS(¼ M) 3814	ΔAIC ¼ M	0	7.3	9.1	19.4
	ΔAIC 1 M	294	2.5	0	25
	$N_{initial}$	62281	62989	62357	62373
TD-SIS ($c = e^1$, pr =1)	N_{final}	23056	96123	93139	37682
R = ¼ M particles	**N_{final}/R**	**0.09**	**0.38**	**0.37**	**0.15**
from 1 M (incl. parallels)	ESS ¼ M	928	2156	1727	285
ESS (¼ M) 3810	ΔAIC ¼ M	0	7.1	8.9	18.9
	ΔAIC 1 M	294	2.7	0	25

Table 1: Akaike Information Criterion (AIC, see eqn 15) and, for standard SIS, the Akaike Weight (AW, see eqn 16) for each model, given the simulated data and assuming a constant 15% coefficient of variation in the observation process: rows 1 (likelihood calculated from fit at every time point) and 2 (likelihood constructed at y_r = 10, 15, 20, 22, 25, 30) calculated using the expectation (mean) values of the parameters' prior distributions and deterministic model simulation (Fig. 1) and rows 3 – 5 (Figs. 2 and 3) using typical stochastic modelling and 6 time points. EES = Estimated Sample Size from posterior samples, N = number of particles by model, R = total number of particles in study. For TD-SIS, R = ¼ M independent particles (not including parallels) and ¾ M parallels (since there are four models types), the ΔAIC ¼M value incorporates only the independent particles and the ΔAIC 1M is given for comparison. N_{final}/R is the posterior model probability for TD-SIS and are comparable to the AW for standard SIS.

Figures

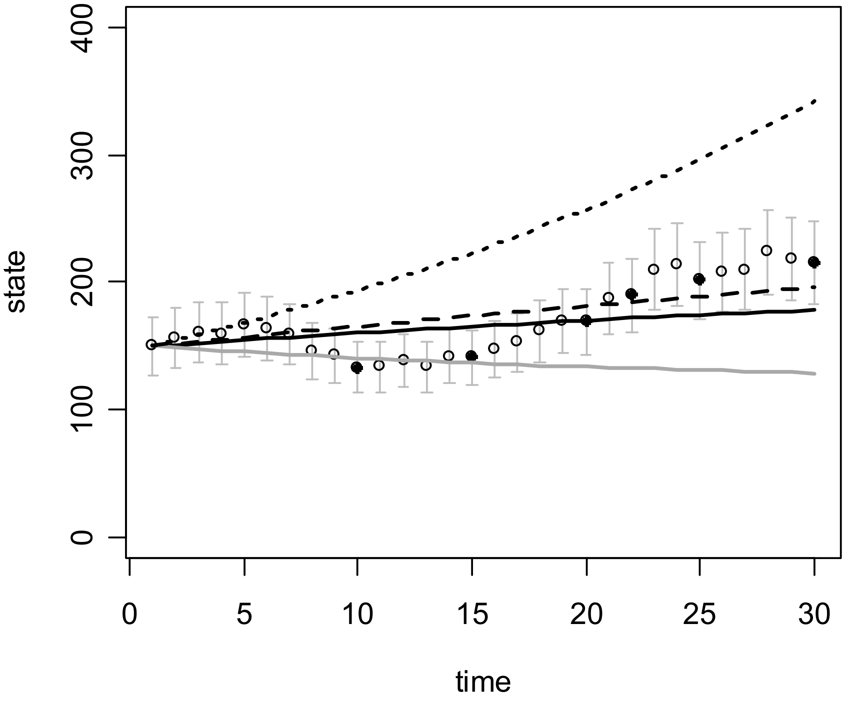

Fig. 1. Deterministic model outputs using the expectations from the prior distributions and selected data (circles; solid circles are the states at the 6 time points used for likelihood evaluation) for the study from a single stochastic simulation using M2. M0, black dotted line; M1, black dashed line; M2, solid black line; M3, solid grey line;

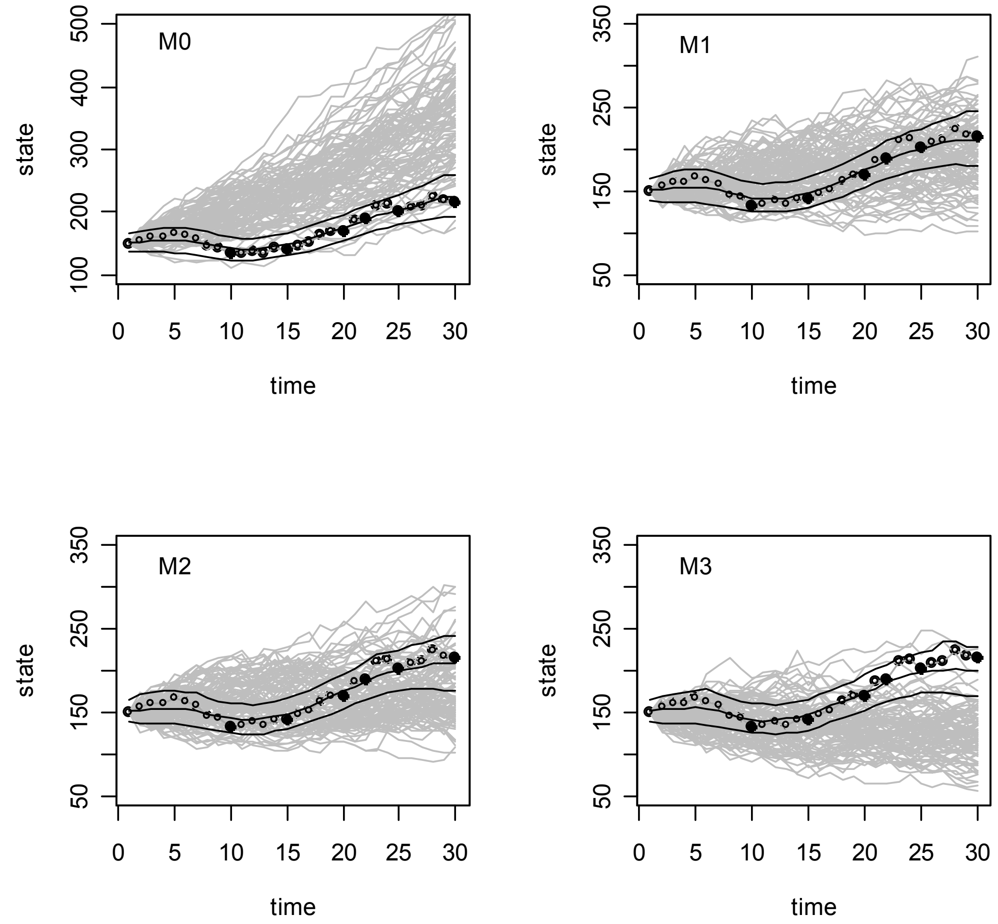

Fig. 2. Fit after standard SIS. Median and 95% credible interval by model (solid black lines) and data (circles) (circles, where solid circles are those used for likelihood calculation). Also shown 100 randomly selected particles from the posterior distribution for each model.

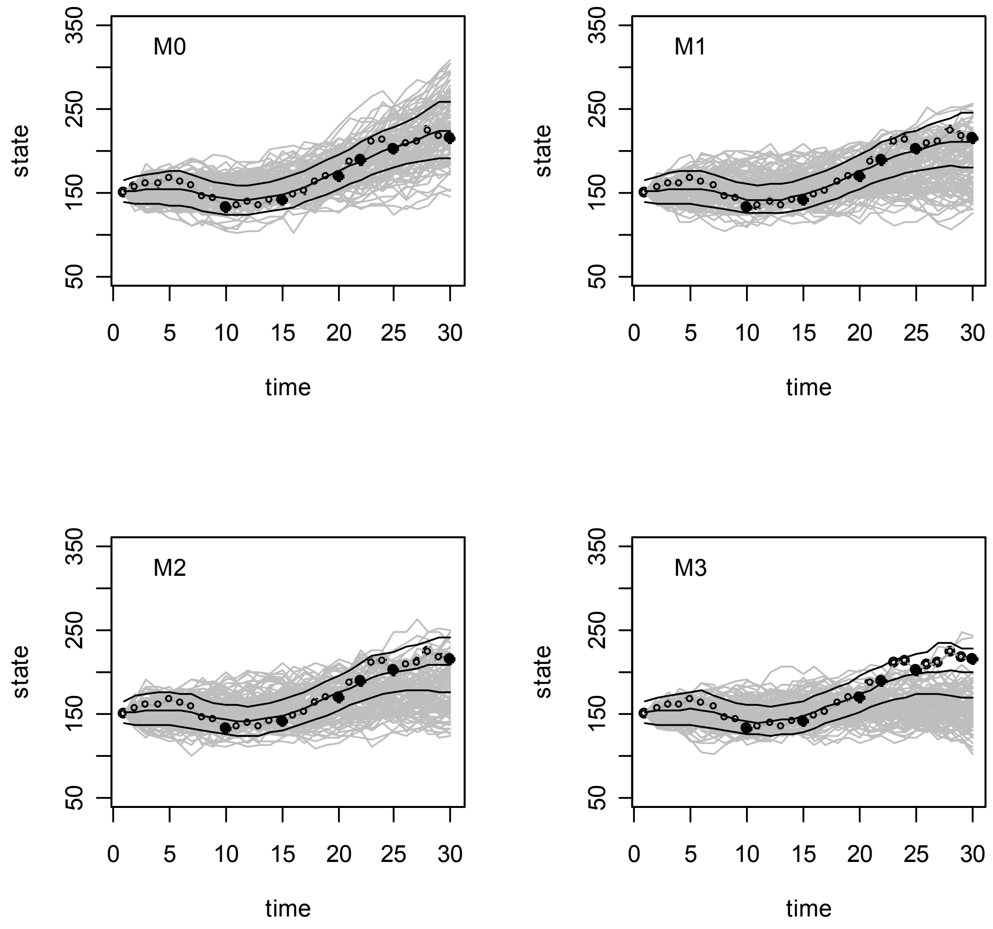

Fig. 3. Fit after TD-SIS using equal model priors and thus non-penalised transition probabilities (*c* = 1). Median and 95% credible interval by model (solid black lines) and data (circles, where solid circles are those used for likelihood calculation). Also shown 100 randomly selected particles from the posterior distribution for each model.