Bayesian Analysis (2019) **14**, Number 3, pp. 753–776

Sequential Monte Carlo Samplers with Independent Markov Chain Monte Carlo Proposals

L. F. South[*,†‡], A. N. Pettitt[*,†§], and C. C. Drovandi[*,†¶]

Abstract. Sequential Monte Carlo (SMC) methods for sampling from the posterior of static Bayesian models are flexible, parallelisable and capable of handling complex targets. However, it is common practice to adopt a Markov chain Monte Carlo (MCMC) kernel with a multivariate normal random walk (RW) proposal in the move step, which can be both inefficient and detrimental for exploring challenging posterior distributions. We develop new SMC methods with independent proposals which allow recycling of all candidates generated in the SMC process and are embarrassingly parallelisable. A novel evidence estimator that is easily computed from the output of our independent SMC is proposed. Our independent proposals are constructed via flexible copula-type models calibrated with the population of SMC particles. We demonstrate through several examples that more precise estimates of posterior expectations and the marginal likelihood can be obtained using fewer likelihood evaluations than the more standard RW approach.

Keywords: copula, evidence, importance sampling, independent proposal, Markov chain Monte Carlo, marginal likelihood.

1 Introduction

Sequential Monte Carlo (SMC, Chopin (2002); Del Moral et al. (2006)) methods for static Bayesian models are naturally adaptive, easily parallelisable and are capable of dealing with targets that are multimodal or have complicated landscapes (see e.g. Del Moral et al. (2006) and Cappé et al. (2007)). The basic SMC method involves moving a population of N particles through a sequence of distributions which can be chosen by smoothly introducing either the data (data annealing) or the effect of the likelihood (likelihood annealing). Many of the computations associated with the N particles can be performed in parallel. As a useful by-product, SMC produces an estimate of the normalising constant of the posterior distribution (e.g. Del Moral and Miclo (2000)), the so-called evidence, which is useful for Bayesian model choice.

SMC propagates the particles through the sequence of distributions using three types of steps: reweighting, resampling and moving (or mutation). The reweighting step uses

[*]School of Mathematical Sciences, Queensland University of Technology, Australia
[†]ARC Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS), leah.south@hdr.qut.edu.au
[‡]Supported by an Australian Government Research Training Program Scholarship and by ACEMS.
[§]Supported by Australian Research Council's Discovery Project funding scheme DP110100159.
[¶]Supported by Australian Research Council's Discovery Early Career Researcher Award funding scheme DE160100741.

 https://doi.org/10.1214/18-BA1129

importance sampling (IS) to adjust the weights of particles from the current distribution to obtain a properly weighted set targeting the next distribution. The resampling step is used to focus on promising regions that will then be diversified using a move step, which is the most computationally intensive aspect of SMC. The move step is commonly chosen to be several iterations of a Markov chain Monte Carlo (MCMC, Metropolis et al. (1953)) kernel, often with a random walk proposal. Despite their commonplace use, random walk proposals can be inefficient at exploring the target and this can have a detrimental effect on evidence and posterior expectation estimates.

We develop new and efficient SMC methods using independent MCMC proposals. Independent proposal distributions have the advantage that they can result in uniformly ergodic Markov chains, as opposed to typical geometric ergodicity achieved by random walk proposals (Tierney, 1994). They are highly parallelisable and have the potential to explore complex and multimodal targets in fewer iterations than local proposals. Some early SMC algorithms proposed basic independent proposals like multivariate normal (Chopin, 2002) and more recently an independent proposal for multivariate binary spaces has been used in SMC (Schäfer and Chopin, 2013). However independent proposals are often dismissed in the SMC literature for being too restrictive (see e.g. Del Moral et al. (2006)). Silva et al. (2010) and Schmidl et al. (2013) develop adaptive MCMC methods that use copula (Sklar, 1959) type models to form independent proposals. Copulas provide flexible multivariate distributions as they allow for separate modelling of marginals and dependence structure between components. We extend this idea to the SMC setting, taking advantage of the available population of particles.

We demonstrate that when independent proposals are used, all candidates generated in the SMC process can be used in evidence estimation and posterior inference. Throughout this paper, "candidates" refers to all samples from the prior and all MCMC proposals. We compare several estimators from the IS literature with existing methods for SMC in terms of bias and precision. We also propose a novel evidence estimator which is simple and computationally efficient to calculate from our independent SMC output. These new recycling schemes for SMC can lead to increases in the effective sample size (ESS) targeting the posterior, improved sampling from complex posterior distributions and significant variance reductions when compared to no recycling and the recycling method of Nguyen et al. (2014). In the examples considered, the novel evidence estimator is up to five orders of magnitude more efficient than the standard SMC estimator.

Section 2 of this paper gives a brief review of likelihood annealing SMC, the existing literature on recycling in SMC and the copula models which form the basis of the independent proposals. The main contributions of this paper, developing independent MCMC proposals for SMC and exploiting these proposals, are described in Section 3. In Section 4, we compare our methods with a more standard SMC implementation on applications of varying complexity. A final summary and a discussion of possible limitations and extensions of this work are given in Section 5.

2 Background

The focus of this article is Bayesian inference for a statistical model parameterised by $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ where p is the dimension of vector $\boldsymbol{\theta}$ and the collected data is denoted

$\mathbf{y} \in \mathsf{Y} \subseteq \mathbb{R}^d$. The posterior distribution is defined as

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z(\mathbf{y})}, \tag{1}$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood, $\pi(\boldsymbol{\theta})$ is the prior and $Z(\mathbf{y}) = \int_\Theta f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the normalising constant of the posterior often referred to as the evidence. The evidence is an important quantity in Bayesian model selection because it is used in the calculation of Bayes factors. However, it is a p dimensional integral which makes it difficult to estimate. Here we are interested in estimating both the evidence, Z, and posterior functionals, $\int_\Theta \psi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$, and we use SMC to do so.

2.1 Sequential Monte Carlo

SMC traverses a set of N weighted samples or 'particles', $\{W_t^i, \boldsymbol{\theta}_t^i\}_{i=1}^N$, through a sequence of distributions, $\pi_t(\boldsymbol{\theta}|\mathbf{y})$ for $t = 0, \ldots, T$. SMC consists of reweighting, resampling and move steps, each of which are described in further detail below.

In this paper we focus on the likelihood annealing sequence which is useful for exploration of complex targets (Neal, 2001) and allows for additional benefits in terms of recycling otherwise wasted samples. Interest is in sampling from the sequence of power posteriors, $\pi_t(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})^{\gamma_t}\pi(\boldsymbol{\theta})/Z_t$ where γ_t is referred to as the temperature and $0 = \gamma_0 \leq \cdots \leq \gamma_t \leq \cdots \leq \gamma_T = 1$.

Reweighting Step

The reweighting step uses IS to weight particles targeting the previous distribution appropriately for the next distribution. Following Del Moral et al. (2006), we obtain

$$w_{t+1}^i = W_t^i f(\mathbf{y}|\boldsymbol{\theta}_t^i)^{\gamma_{t+1}-\gamma_t}, \text{ for } i = 1, \ldots, N,$$

where w_{t+1}^i is an unnormalised weight and $W_0^i = 1/N$ for $i = 1, \ldots, N$. The weights are normalised and we set $\boldsymbol{\theta}_{t+1}^i = \boldsymbol{\theta}_t^i$ for $i = 1, \ldots, N$ to obtain the set of particles providing a discrete approximation of $\pi_{t+1}(\boldsymbol{\theta}|\mathbf{y})$.

The temperature γ_{t+1} is chosen adaptively by approximately maintaining a specified effective sample size (ESS), the number of independent samples from the target that would be required to achieve the equivalent variance in the estimator, of ρN (see Jasra et al. (2011) for details). Following Kong et al. (1994), the ESS at target $t+1$ is approximated by $1/\sum_{i=1}^N (W_{t+1}^i)^2$.

Resampling Step

After reweighting, the particles are resampled with probabilities given by their corresponding normalised weights. This resampling has the effect of eliminating particles with negligible weight and replicating the particles with larger weights. See Gerber et al. (2017) for a recent review of the theoretical properties of resampling methods. We use multinomial resampling for simplicity.

Move Step

To diversify the particles and explore π_t, we apply R_t iterations of a Metropolis Hastings (MH) MCMC kernel on each particle. For a single MCMC iteration with proposal distribution q^{ϕ_t} and parameter $\boldsymbol{\phi}_t$, a candidate $\boldsymbol{\theta}_t^{i,*} \sim q^{\phi_t}(\cdot|\boldsymbol{\theta}_t^i)$ is proposed and we set $\boldsymbol{\theta}_t^i = \boldsymbol{\theta}_t^{i,*}$ with probability

$$\alpha(\boldsymbol{\theta}_t^i, \boldsymbol{\theta}_t^{i,*}) = \min\left(1, \frac{f(\mathbf{y}|\boldsymbol{\theta}_t^{i,*})^{\gamma_t}\pi(\boldsymbol{\theta}_t^{i,*})q^{\phi_t}(\boldsymbol{\theta}_t^i|\boldsymbol{\theta}_t^{i,*})}{f(\mathbf{y}|\boldsymbol{\theta}_t^i)^{\gamma_t}\pi(\boldsymbol{\theta}_t^i)q^{\phi_t}(\boldsymbol{\theta}_t^{i,*}|\boldsymbol{\theta}_t^i)}\right),$$

otherwise we retain the current $\boldsymbol{\theta}_t^i$.

A major advantage of SMC is that the parameter $\boldsymbol{\phi}_t$ can be estimated from the population of particles prior to the move step. It is common to use multivariate normal random walk (RW) proposals and to estimate the covariance from the population of particles (Chopin, 2002). The proposal distribution using this scheme is $q^{\phi}(\boldsymbol{\theta}_t^{i,*}|\boldsymbol{\theta}_t^i) = \mathcal{N}(\boldsymbol{\theta}_t^{i,*}; \boldsymbol{\theta}_t^i, h^2\hat{\Sigma}_t)$, where $\boldsymbol{\phi}_t = (\hat{\Sigma}_t, h)$ and $\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal probability density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ evaluated at $\boldsymbol{\theta}$. Here we set $h = 2.38/\sqrt{p}$ (Gelman et al., 1996). Our alternative proposal is described in Section 3 and compared to the RW empirically in Section 4.

We propose to choose R_t adaptively based on the current move step at t so that there is a theoretical probability of $1 - c$ (with c set small) that the particle is moved at least once. The number of repeats is

$$R_t = \left\lceil \frac{\log(c)}{\log(1 - \hat{p}_{\text{acc}}^t)} \right\rceil,$$

where $\lceil\cdot\rceil$ denotes the ceiling function and $\hat{p}_{\text{acc}}^t = \frac{1}{N}\sum_{i=1}^N \alpha(\boldsymbol{\theta}_t^i, \boldsymbol{\theta}_t^{i,*})$ is the acceptance rate based on a trial MCMC iteration on the N particles. This method is similar to Drovandi and Pettitt (2011), who predict p_{acc}^t with the MCMC acceptance rate at $t-1$, p_{acc}^{t-1}. We do not rely on $p_{\text{acc}}^t \approx p_{\text{acc}}^{t-1}$ and we do not require an initial choice for R_1. Only a further $R_t - 1$ MCMC iterations are required per particle in our method since the first iteration is already performed. In the discussion in Section 5, we show how an improved Rao-Blackwellised estimate of R_t and a particle specific R_t^i choice of the MCMC repeats can be obtained.

Estimating Posterior Expectations and the Evidence

The standard SMC estimator for posterior expectations is given by

$$\hat{\psi} = \frac{1}{N}\sum_{i=1}^N \psi(\boldsymbol{\theta}_T^i). \tag{2}$$

The standard SMC estimator of the evidence is

$$\widehat{Z} = \prod_{t=1}^T \sum_{i=1}^N w_t^i. \tag{3}$$

This estimator is based on the identity $Z_T/Z_0 = \prod_{t=1}^{T} Z_t/Z_{t-1}$ where $Z_0 = 1$. It requires little additional implementation or computational effort. Intermediate normalising constants can also be estimated by taking fewer terms in the product, $\widehat{Z_t} = \prod_{s=1}^{t} \sum_{i=1}^{N} w_s^i$.

2.2 Recycling in SMC

Typically the final samples $\{\boldsymbol{\theta}_T^i\}_{i=1}^N$ from an SMC run are used for inference and all other proposed particles are wasted because they do not target the posterior. Particle recycling schemes, such as those presented for general importance tempering in Gramacy et al. (2010) and for SMC in Nguyen et al. (2014) and Finke (2015), use IS to reweight the final particles from π_t for $t = 0, \ldots, T$ to target the posterior.

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, $\pi_t = f(\mathbf{y}|\boldsymbol{\theta})^{\gamma_t}\pi(\boldsymbol{\theta})/\widehat{Z}_t$. The unnormalised IS weights to use the t-th power posterior samples are therefore

$$\kappa_t^i = \frac{f(\mathbf{y}|\boldsymbol{\theta}_t^i)\pi(\boldsymbol{\theta}_t^i)}{f(\mathbf{y}|\boldsymbol{\theta}_t^i)^{\gamma_t}\pi(\boldsymbol{\theta}_t^i)(Z_t)^{-1}} = Z_t f(\mathbf{y}|\boldsymbol{\theta}_t^i)^{1-\gamma_t}, \tag{4}$$

where the term Z_t is unknown. Normalising the weights to $K_t^i = \kappa_t^i / \sum_{i=1}^N \kappa_t^i$ removes the need for Z_t. Posterior expectations can be estimated by normalising the weights and applying standard Monte Carlo integration, $\hat{\psi}_t = \sum_{i=1}^N \psi(\boldsymbol{\theta}_t^i) K_t^i$.

Gramacy et al. (2010) propose a method to combine multiple IS estimators in order to maximise the ESS. Denote the ESS targeting the posterior using $\{\boldsymbol{\theta}_t^i\}_{i=1}^N$ from the t-th power posterior as ESS_t and denote the unbiased estimate of some posterior functional using those samples as $\hat{\psi}_t$. The combined estimate is

$$\hat{\psi} = \sum_{t=0}^{T} \lambda_t \hat{\psi}_t, \tag{5}$$

where $\lambda_t = \text{ESS}_t / \sum_{l=0}^{T} \text{ESS}_l$. By linearity of expectation, this combined estimate is unbiased if the λ values are fixed from a separate run. The motivation behind this combined estimator is that importance distributions which have poor performance (as measured by ESS) should be given less weight in the combined estimator. The ESS targeting the posterior after recycling is $\sum_{t=0}^{T} \text{ESS}_t$ in the adaptive case (see Appendix A of the Supplementary Materials (South et al., 2018)). We refer to the general approach of combining IS estimators based on ESS as combined importance sampling (CIS) and we denote the combination of power posterior samples through CIS as CIS_{PP}. CIS_{PP} has been applied in the SMC context by Nguyen et al. (2014) and Finke (2015).

Another method for combining samples from multiple importance distributions is to treat the samples as coming from a mixture of distributions, with the mixture weights based on the proportion of samples coming from that distribution (Veach and Guibas, 1995; Owen and Zhou, 2000). Following Owen and Zhou (2000), we refer to this method as deterministic mixture sampling or DeMix for short. We refer to the use of DeMix to recycle particles from the power posteriors as DeMix_{PP}. Nguyen et al. (2016) use

DeMix_{PP} in calculating posterior expectations. The unnormalised DeMix weights in this context are

$$\nu_t^i = \frac{f(\mathbf{y}|\boldsymbol{\theta}_t^i)\pi(\boldsymbol{\theta}_t^i)}{\frac{1}{T}\sum_{l=0}^{T} f(\mathbf{y}|\boldsymbol{\theta}_t^i)^{\gamma_l}\pi(\boldsymbol{\theta}_t^i)(Z_l)^{-1}}, \tag{6}$$

for $i = 1, \ldots, N$ and $t = 0, \ldots, T$. Calculating the deterministic multiple mixture weights requires that π_t be normalised so, like Nguyen et al. (2016), we use the SMC estimate $\widehat{Z_t}$ for the normalising constant of π_t. Nguyen et al. (2016) compare DeMix_{PP} and CIS_{PP} for posterior approximation and found empirically that the DeMix_{PP} scheme performs only marginally better.

Normalising constant estimation using power posterior recycling schemes has been restricted since the importance density involves the normalising constants Z_t for π_t. The requirement to estimate intermediate normalising constants in order to estimate the overall normalising constant and the fact that only an additional N particles are used when compared to the standard SMC estimator makes these evidence estimators seem unappealing when compared to the simple SMC estimator in (3). For completeness, we implement CIS_{PP} and DeMix_{PP} estimators of the evidence using the SMC estimates of the intermediate normalising constants. We find empirically in Section 4 that the CIS_{PP} and DeMix_{PP} estimators perform similarly to the SMC estimator of the evidence.

In Section 3.2, we propose evidence estimators that do not require estimates of the intermediate normalising constants and can take advantage of all candidate parameter values generated in the SMC process.

2.3 Copulas

Copulas form the basis of the independent proposals described in Section 3.1. Sklar's theorem (Sklar, 1959) states that a multi-dimensional distribution can be described entirely by its marginal cumulative distribution functions and a copula that captures the dependence between these marginals. Any p-dimensional random vector $\boldsymbol{V} = (V_1, \ldots, V_p)^T$ with cumulative distribution function H and continuous marginal distribution functions $G_j(v) = P(V_j \leq v)$ for $j = 1, \ldots, p$ can be described by a unique copula C such that

$$H(v_1, \ldots, v_p) = C(G_1(v_1), \ldots, G_p(v_p)).$$

Copulas describe the dependence between univariate uniform marginals but this can easily be extended to non-uniform marginals using transformation methods.

A simple and popular family of copulas is the Gaussian copula (see e.g. Fang et al. (2002)). Denote random vectors on the $\mathcal{U}(0,1)$ scale by $\boldsymbol{U}$ and random vectors after transformation to $\mathcal{N}(0,1)$ by $\boldsymbol{X}$. Denote the j-th dimension of $\boldsymbol{X}$ as $\boldsymbol{X}[j]$ for $j = 1, \ldots, p$. The Gaussian copula can be written as

$$H_{\boldsymbol{D}}^{\text{Gauss}} = \Phi_D\left(\Phi^{-1}(\boldsymbol{U}[1]), \ldots, \Phi^{-1}(\boldsymbol{U}[p])\right),$$

where $\Phi_{\boldsymbol{D}}$ is the joint cumulative distribution function of the multivariate Gaussian distribution with correlation matrix $\boldsymbol{D}$. The probability density function is

$$p(\boldsymbol{U}|\boldsymbol{D}) = \frac{\mathcal{N}(\boldsymbol{X}; \mathbf{0}, \boldsymbol{D})}{\prod_{j=1}^{p} \mathcal{N}(\boldsymbol{X}[j]; 0, 1)}.$$

In practice, the correlation matrix $\boldsymbol{D}$ is estimated by the empirical correlation matrix of a set of samples on the $\mathcal{N}(0,1)$ scale.

A more flexible way to model dependence is through a multivariate mixture model, as described in Tran et al. (2014). These models are not strictly copula models because the marginals are not preserved, but they offer a flexible way to model dependence separately from the marginals. We consider the most simple and computationally inexpensive mixture, a multivariate Gaussian mixture model (MGMM, Pearson (1894); Dempster et al. (1977)) but we note that other mixtures considered by Tran et al. (2014), such as mixtures of multivariate t's, may help to improve tail coverage at the cost of computation time.

MGMMs consist of a weighted mixture of multivariate Gaussian distributions. The parameters to be estimated in an MGMM are the mean, $\boldsymbol{\mu}_k$, and covariance, $\boldsymbol{\Sigma}_k$, of the k-th component and a set of weights c_k for $k = 1, \ldots, K$ where K is the number of components and $\sum_{k=1}^{K} c_k = 1$. The density of an MGMM for parameter $\boldsymbol{X}$ is defined as

$$p(\boldsymbol{X}|\boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} c_k \, \mathcal{N}(\boldsymbol{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{7}$$

An independent MCMC proposal using copulas has been explored before in MCMC (Silva et al., 2010; Schmidl et al., 2013), where it is limited by the requirement of finding an approximation of the target distribution to fit the copula and the marginals. In SMC, an approximation from the current target is already available. To our knowledge, independent MCMC proposals using copulas have not yet been developed for general use in an SMC framework. This is the focus of Section 3.1.

3 Independent Proposals in SMC

The main contributions of this work are to develop efficient independent MCMC proposals for SMC and to exploit them in several ways, for example by harnessing all generated candidates to obtain reduced variance estimators of posterior expectations and the evidence. We also present a new evidence estimator that is conveniently computed from the output of independent SMC.

3.1 Copula-Type Independent Proposals

Estimating the Copula-Type Model

Here we describe a method for estimating the parameters of the copula-type independent proposals. A univariate distribution is first chosen for each of the marginals. Different distributions, including univariate beta and Gaussian mixture distributions, are used in this work. When individual parameters are bounded above and below a priori, beta marginals may be a sensible choice as the prior limits can be transformed to $[0, 1]$. The univariate distributions with parameters $\boldsymbol{\eta}_j$ are fitted to $\{\boldsymbol{\theta}_t^i[j]\}_{i=1}^N$ for $j = 1, \ldots, p$

and the cumulative distribution function of the j-th marginal is denoted by $G_j^{\hat{\boldsymbol{\eta}}_j}(\cdot)$. Taking $\boldsymbol{U}_t^i[j] = G_j^{\hat{\boldsymbol{\eta}}_j}(\boldsymbol{\theta}_t^i[j])$, for $i = 1, \ldots, N$ transforms each of the p dimensions of the particle population to be approximately $\mathcal{U}(0,1)$ distributed random variates provided that $G_j^{\hat{\boldsymbol{\eta}}_j}$ fits well. The quantile function of the standard normal distribution, Φ^{-1}, can then be used to transform the marginals to be approximately standard normal through $\boldsymbol{X}_t^i[j] = \Phi^{-1}(\boldsymbol{U}_t^i[j])$.

With approximately $\mathcal{N}(0,1)$ marginals, it is easier to model the dependencies. Here we consider an MGMM as described in Section 2.3. The parameters $\hat{\boldsymbol{c}}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ for the mixture model are obtained using the expectation maximisation (EM) algorithm (Dempster et al., 1977). There are methods to help choose the number of components in a mixture model (McLachlan and Peel, 2000), for example based on the Bayesian information criterion (Schwarz, 1978; Keribin, 2000) or using variational Bayes methods as in Tran et al. (2014), but for simplicity here we fix the number of components.

The process of fitting a copula-type model, including the transformations to approximately $\mathcal{N}(0,1)$ marginals, is given in Algorithm 1. An illustration of this process for the example in Section 4.2 is shown in Figure 1.

Algorithm 1: Fitting the copula-type MGMM from the population of particles at π_t.

Input : A population of particles from the current power posterior $\{\boldsymbol{\theta}_t^i\}_{i=1}^N$, the number K of components in the MGMM and the type of marginals to be fitted (problem dependent).

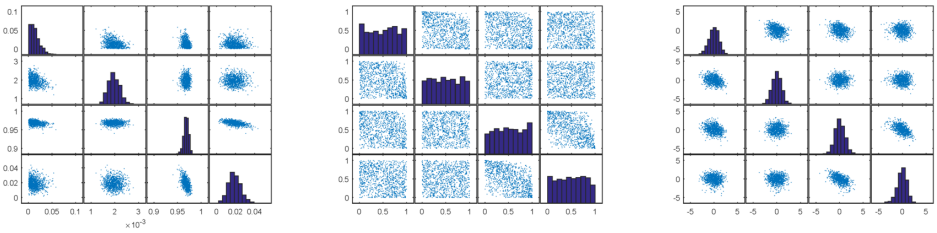
Output: Marginal distribution parameters $\{\hat{\boldsymbol{\eta}}_j\}_{j=1}^p$, mixture model parameters $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and the transformed population on marginal $\mathcal{N}(0,1)$ scale $\{\boldsymbol{X}_t^i\}_{i=1}^N$.

for $j = 1$ *to* p **do**
  Estimate $\hat{\boldsymbol{\eta}}_j$ from $\{\boldsymbol{\theta}_t^i[j]\}_{i=1}^N$
  Compute $\boldsymbol{U}_t^i[j] = G_j^{\boldsymbol{\eta}_j}(\boldsymbol{\theta}_t^i[j])$ for $i = 1, \ldots, N$.
  Compute $\boldsymbol{X}_t^i[j] = \Phi^{-1}(\boldsymbol{U}_t^i[j])$ for $i = 1, \ldots, N$.
end
Estimate $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ from $\{\boldsymbol{X}_t^i\}_{i=1}^N$ using the EM algorithm.

Making Proposals

Once the copula-type model has been fitted, candidates can be drawn from the mixture model which simply involves simulating from the k-th component with probability c_k.

Transformation of candidates $\boldsymbol{X}^*$ on the $\mathcal{N}(0,1)$ scale to the approximately $\mathcal{U}(0,1)$ scale is done through the cumulative distribution function of the standard normal distribution, Φ, such that $\boldsymbol{U}^*[j] = \Phi(\boldsymbol{X}^*[j])$, for $j = 1, \ldots, p$. The candidate on the original scale is $\boldsymbol{\theta}^*[j] = Q_j^{\hat{\boldsymbol{\eta}}_j}(\boldsymbol{U}^*[j])$, for $j = 1, \ldots, p$, where $Q_j \equiv G_j^{-1}$ is the quantile distribution of the j-th fitted marginal distribution.



(a) Original Scale (b) Transformed to approximately $\mathcal{U}(0,1)$ marginals (c) Transformed to approximately $\mathcal{N}(0,1)$ marginals

Figure 1: Scatterplots of the bivariate distributions based on approximate draws from the posterior in Section 4.2. Shown are the first four parameters (a) on the original scale, (b) after transforming to approximately $\mathcal{U}(0,1)$ marginals and (c) after transforming to approximately $\mathcal{N}(0,1)$ marginals. It is simpler to model the dependence based on (c) than (a).

The proposal density for a candidate $\boldsymbol{\theta}_t^*$ with fitted parameters $\boldsymbol{\phi}_t = (\hat{\boldsymbol{\eta}}, \hat{\mathbf{c}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ needs to account for these transformations. The proposal density (7) for $\boldsymbol{X}^*$ is adjusted using transformation methods (see Appendix B of the Online Resources) to obtain

$$q^{\boldsymbol{\phi}_t}(\boldsymbol{\theta}_t^*) = \left\{ \prod_{j=1}^{p} \frac{g_j^{\hat{\boldsymbol{\eta}}_j}(\boldsymbol{\theta}_t^*[j])}{\mathcal{N}(\boldsymbol{X}_t^*[j]; 0, 1)} \right\} \sum_{k=1}^{K} \hat{c}_k \, \mathcal{N}(\boldsymbol{X}_t^*; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k), \tag{8}$$

where $g_j^{\hat{\boldsymbol{\eta}}_j}$ denotes the probability density function of the j-th marginal with estimated parameter $\hat{\boldsymbol{\eta}}_j$. The proposal density in (8) is required for the MH ratio and also for recycling methods.

Algorithm 2 outlines the details of performing a single MCMC step for each particle using our MGMM independent proposals. This step is repeated multiple times depending on R_t. Parallelisation is straightforward as one can draw all candidates from $q^{\boldsymbol{\phi}_t}(\cdot)$ and calculate $\{f(\mathbf{y}|\boldsymbol{\theta}^{i,*})\}_{i=1}^{R_t N}$, $\{\pi(\boldsymbol{\theta}^{i,*})\}_{i=1}^{R_t N}$ and $\{q^{\boldsymbol{\phi}_t}(\boldsymbol{\theta}^{i,*})\}_{i=1}^{R_t N}$ for the candidates in parallel.

3.2 Recycling All Candidates

In the context of our independent SMC method, it is possible to recycle all *candidates* from all temperatures by using the $q^{\boldsymbol{\phi}_t}(\cdot)$ as importance distributions. This means that R_t times as many samples per temperature can be used in estimating posterior expectations and estimating the evidence when compared to the PP recycling methods described previously.

The importance density $q^{\boldsymbol{\phi}_t}$ can be computed pointwise for any candidate when an independent MCMC proposal is used, for example $q^{\boldsymbol{\phi}_t}$ for the MGMM copula-type proposal is given in (8). This means that the SMC estimates of the normalising constants are not required for any calculations, making these recycling methods much more appeal-

Algorithm 2: One iteration of an MCMC kernel on all particles using independent MGMM copula-type proposals.

Input : A population of particles from the current target $\{\boldsymbol{\theta}_t^i\}_{i=1}^N$, the transformed standard normal version of the population of particles $\{\boldsymbol{X}_t^i\}_{i=1}^N$, fitted marginal distribution parameters $\{\hat{\boldsymbol{\eta}}_j\}_{j=1}^p$, fitted mixture model parameters $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, the current temperature γ_t and the prior distribution $\pi(\boldsymbol{\theta})$.

Output: Particles after move step, $\{\boldsymbol{\theta}_t^i\}_{i=1}^N$, and the new transformed population on marginal $\mathcal{N}(0,1)$ scale $\{\boldsymbol{X}_t^i\}_{i=1}^N$.

for $i = 1$ *to* N **do**
  Draw $\boldsymbol{X}^*$ from the MGMM with parameters $(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$
  for $j = 1$ *to* p **do**
    Compute $\boldsymbol{U}^*[j] = \Phi(\boldsymbol{X}^*[j])$
    Compute $\boldsymbol{\theta}^*[j] = Q_j^{\hat{\boldsymbol{\eta}}_j}(\boldsymbol{U}^*[j])$
  end
  Compute $r = \min(1, \frac{f(\mathbf{y}|\boldsymbol{\theta}^*)^{\gamma_t}\pi(\boldsymbol{\theta}^*)q^{\boldsymbol{\phi}_t}(\boldsymbol{\theta}_t^i)}{f(\mathbf{y}|\boldsymbol{\theta}_t^i)^{\gamma_t}\pi(\boldsymbol{\theta}_t^i)q^{\boldsymbol{\phi}_t}(\boldsymbol{\theta}^*)})$
  if $\mathcal{U}(0,1) < r$ **then**
    Set $\boldsymbol{\theta}_t^i \leftarrow \boldsymbol{\theta}^*$
    Set $\boldsymbol{X}_t^i \leftarrow \boldsymbol{X}^*$
  end
end

ing for estimating the evidence when compared to the PP recycling evidence estimators detailed in Section 2.2.

The IS weight for candidate $\boldsymbol{\theta}_t^{i,*}$ drawn from independent proposal $q^{\boldsymbol{\phi}_t}(\cdot)$ targeting π_t is

$$\omega_t^i = \frac{f(\mathbf{y}|\boldsymbol{\theta}_t^{i,*})\pi(\boldsymbol{\theta}_t^{i,*})}{q^{\boldsymbol{\phi}_t}(\boldsymbol{\theta}_t^{i,*})}, \text{ for } i = 1, \ldots, R_t N \text{ and } t = 0, \ldots, T.$$

Posterior expectations using these weights can be combined using the CIS method from (5). We also derive the following novel estimator of the evidence,

$$\widehat{Z} = \sum_{t=0}^{T} \frac{\lambda_t}{R_t N} \sum_{i=1}^{R_t N} \omega_t^i,$$

where $\lambda_t = \mathrm{ESS}_t / \sum_{l=0}^{T} \mathrm{ESS}_l$ and ESS_t is the ESS targeting the posterior using the candidates drawn from independent proposal $q^{\boldsymbol{\phi}_t}(\cdot)$. This estimator is in essence an efficient combination of multiple IS estimators for the evidence which, by linearity of expectation, is unbiased if the λ values are fixed from a separate run. To the best of our knowledge, this is the first time that the ESS-based approach has been used for evidence estimation. We refer to the CIS estimators which reuse all candidates as $\mathrm{CIS}_{\mathrm{IP}}$ where the IP standards for independent proposal recycling.

We also propose the use of the DeMix scheme to reuse all candidates in SMC and we refer to this method as DeMix_{IP}. The DeMix_{IP} weights are

$$\tilde{\nu}_t^i = \frac{f(\mathbf{y}|\boldsymbol{\theta}_t^i)\pi(\boldsymbol{\theta}_t^i)}{\sum_{l=0}^{T} \frac{R_l}{\sum_{m=0}^{T} R_m} q^{\boldsymbol{\phi}_l}(\boldsymbol{\theta}_t^i)}, \tag{9}$$

for $i = 1, \ldots, N$ and $t = 0, \ldots, T$, where $q^{\boldsymbol{\phi}_l}(\cdot)$ represents the independent proposal used to draw the candidates for π_l. The associated evidence estimator is

$$\begin{aligned} \widehat{Z} &= \frac{1}{N \sum_{t=0}^{T} R_t} \sum_{t=0}^{T} \sum_{i=1}^{R_t N} \tilde{\nu}_t^i \\ &= \sum_{t=0}^{T} \sum_{i=1}^{R_t N} \frac{f(\mathbf{y}|\boldsymbol{\theta}_t^{i,*})\pi(\boldsymbol{\theta}_t^{i,*})}{\sum_{s=0}^{T} R_s N q^{\boldsymbol{\phi}_s}(\boldsymbol{\theta}_t^{i,*})}. \end{aligned}$$

This requires an extra T proposal density calculations per particle when compared to CIS but no further target evaluations are required, so this additional cost is generally not prohibitive. When a large number of temperatures are used, methods which limit the computational complexity of DeMix as in Elvira et al. (2015) may be useful. As a result of the mixture term, there tend to be fewer extreme weights. DeMix intuitively seems more natural for candidate recycling (DeMix_{IP}) than power posterior recycling (DeMix_{PP}); in the latter, the terms in the mixture differ only by the likelihood powers and normalising constants which need to be estimated.

4 Simulation Studies

The main comparators which we consider in our simulation studies are summarised in Table 1. The two MCMC kernels being compared are our independent proposal (IND) and the ubiquitous multivariate normal random walk (RW). We compare our methods for recycling all candidates in posterior and evidence estimation (CIS_{IP} and DeMix_{IP}) with the standard practice and with the current state of the art for recycling methods (CIS_{PP} and DeMix_{PP}).

It is well known that the standard SMC estimator of the evidence in (3) is not guaranteed to be unbiased when the sequence of distributions or proposals are adapted online (Beskos et al., 2016). Similarly, the CIS_{PP} and DeMix_{PP} methods are not unbiased if the intermediate normalising constants are estimated from the same SMC run, and the CIS_{PP} and CIS_{IP} methods are not unbiased if the λ (or ESS) values are estimated from the same run. One aim here is to compare the bias of the estimators empirically when adaptive runs are used. If an unbiased estimate is required, the proposal distributions, temperatures, intermediate normalising constant estimates and ESS values from an adaptive run can be used in a fixed run.

MCMC kernels and recycling methods are compared on the basis of estimates of posterior quantiles and estimates of the evidence. We would ideally like to take into account computational effort and statistical efficiency (accuracy and precision) in these

Kernel	Recycling	Samples in $\hat{Z}$	Samples in $\hat{\psi}$	New
RW	–	NT	N	×
IND[a]	–	NT	N	✓
(any)	CIS_{PP}	$N(1+T)$	$N(1+T)$	×
(any)	DeMix_{PP}	$N(1+T)$	$N(1+T)$	×
IND	CIS_{IP}	$N(1+\sum_{t=0}^{T} R_t)$	$N(1+\sum_{t=0}^{T} R_t)$	✓
IND	DeMix_{IP}	$N(1+\sum_{t=0}^{T} R_t)$	$N(1+\sum_{t=0}^{T} R_t)$	✓

[a]IND refers to the MGMM copula proposal described in Section 3.1

Table 1: The kernels and recycling methods considered with information regarding how many samples are used in the estimators and whether they are novel.

comparisons. We measure computational effort by the number of likelihood evaluations (TLL) because this comprises a large proportion of the computational effort in most applications. When a gold standard of approximation is available, for example through Monte Carlo estimates based on a very long MCMC or SMC run, we can measure both accuracy and precision by looking at the *mean square error* (MSE) relative to the gold standard. We resort to measuring the variance of 100 estimates (VAR) when no gold standard is available. Our overall measure of efficiency is MSE $\cdot$ TLL for RW with no recycling divided by MSE $\cdot$ TLL for the method in question. When no gold standard is available, MSE is replaced with VAR. RW with no recycling has an efficiency value of 1 and larger values are preferred.

For likelihood annealing SMC, temperatures are chosen with $\rho = 0.5$ so that the ESS in the next iteration is at least $N/2$. The number of MCMC repeats is chosen so that samples are moved at least once with theoretical probability $1-c = 0.99$. Regularisation is used when estimating the parameters of the Gaussian mixture models, which may assist with tail coverage and helps with the numerical stability of the EM algorithm.

In addition to the applications given in the main paper, three examples are given in the Online Resources. Appendix E contains an 11 dimensional example and Appendix F contains a three dimensional example where an unbiased likelihood estimator is used. In these examples, IND outperforms RW by some efficiency measures before recycling and all efficiency measures after recycling. A challenging application for which our MGMM copula-based IND proposal does not fully cover the tails of the target distribution is given in Appendix G.

4.1 Factor Analysis Example

This model choice example illustrates the potential benefits and limitations of independent SMC for sampling from complex posterior distributions. The model choice aspect of this example also makes the utility of the evidence estimators clear.

Monthly exchange rates of six currencies relative to the British pound were collected from January 1975 to December 1986 (West and Harrison, 1997). Lopes and West (2004)

seek to describe the covariance structure of the standardised (by their sample mean and standard deviation) differences in these exchange rates using factor analysis (FA) models. FA models assume that $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ where the covariance $\boldsymbol{\Omega}$ is restricted to the form $\boldsymbol{\Omega} = \boldsymbol{\beta}\boldsymbol{\beta}^T + \boldsymbol{\Lambda}$, $\boldsymbol{\beta}$ is a $6 \times k$ lower triangular matrix with positive diagonal elements and $\boldsymbol{\Lambda}$ is a 6×6 diagonal matrix with positive elements. Here k is the number of factors and the total number of parameters is $6(k+1) - k(k-1)/2$. We consider $k \leq 3$ to avoid over-parameterising $\boldsymbol{\Omega}$, which leads to models with dimension at most $p = 21$.

Following Lopes and West (2004), we use the priors $\beta_{ij} \sim \mathcal{N}(0,1)$ for $i \neq j$ and $\beta_{ii} \sim \mathcal{N}(0,1)\mathbf{1}(\beta_{ii} > 0)$. The elements of $\boldsymbol{\Lambda}$ have priors $\sigma_i^2 \sim \mathcal{IG}(1.1, 0.05)$ where $\mathcal{IG}(a,b)$ is the inverse Gamma distribution with mean $b/(a-1)$. The positive parameters are log transformed.

For this example, 100 SMC runs with $N = 5,000$ particles are performed. The IND proposals are formed using 5 component Gaussian mixture model marginals and a 6 component MGMM for dependence. Matlab code for this example is available at https://github.com/LeahPrice/Independent-SMC. Given that we are using this example for illustrative purposes and the likelihood is inexpensive, we also do 100 gold standard SMC runs. These gold standard runs use a RW proposal with 50,000 particles and a conservative choice of the tempering adaptation parameter, $\rho = 0.99$, to avoid sudden gaps in modes which discourage mixing.

Posterior Inference

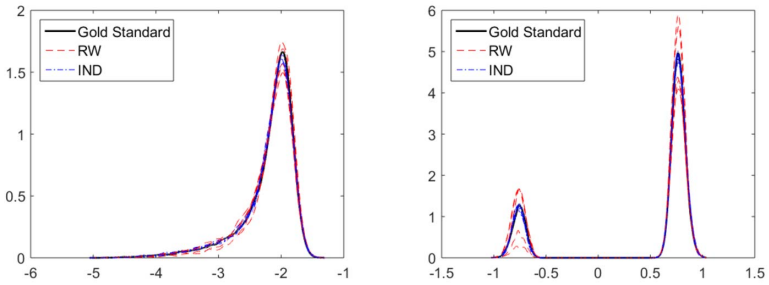
Due to the number of figures required to show the quality of the posterior approximations for the three models, we show some of the more complex marginals here and discuss the results. Appendix C of the Supplementary Materials shows posterior estimates and measures of efficiency for quantiles of the posterior marginals in the three models.

One factor model ($p = 12$)

The marginals in the one factor model are simple and can easily be captured with our independent proposals. In terms of efficiency, IND outperforms RW before recycling and the best results can be achieved by recycling all candidates. The median ESS targeting the posterior is 5,000 before recycling and up to 55,000 with DeMix_{IP} recycling.

Two factor model ($p = 17$)

The introduction of a second factor leads to a more complex posterior. The marginals for β_{32}, β_{42}, β_{52} and β_{62} are all bimodal with well separated modes and some marginals have heavy tails. Posterior approximations of two marginals are shown in Figure 2. The RW proposal fails to achieve consistent proportions in the well separated modes over multiple runs, whereas the IND proposal can easily move between modes. For this reason, IND is significantly more efficient than RW and recycling offers further improvements (see Appendix C of the Supplementary Materials). We note that the well separated mode was not captured in the reversible jump MCMC approach of Lopes and West (2004).

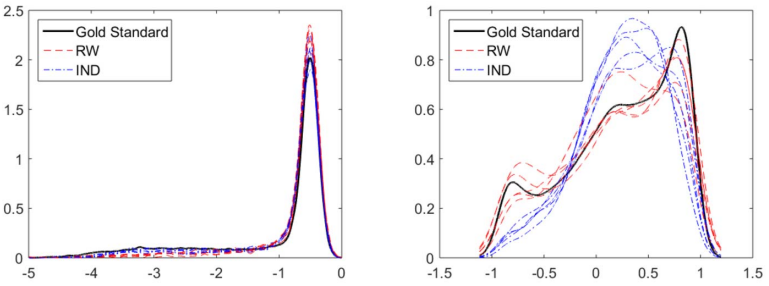


(a) $\log \Lambda_{22}$ (b) β_{62}

Figure 2: FA example: Posterior estimates for (a) $\log \Lambda_{22}$, which is skewed left, and (b) β_{62}, which has two well separated modes. The IND kernel achieves more consistent proportions in each mode than the RW kernel. Results are based on a single gold standard run (solid), five SMC runs with the RW kernel (dash) and five SMC runs with the IND kernel (dot-dash).

Three factor model ($p = 21$)

The three factor model pushes the limits of what can currently be achieved with our IND proposal due to complex and trimodal marginal distributions (Figure 3) and highly complex bivariate distributions (available in Appendix C). The ESS after recycling all candidates is consistently less than N for both CIS_{IP} and DeMix_{IP} recycling, which indicates a lack of tail coverage of the IND proposals with respect to the power posteriors. It is not surprising given the complex dependencies that a 6 component MGMM could not adequately describe the dependence.



(a) $\log \Lambda_{33}$ (b) β_{42}

Figure 3: FA example: Posterior estimates for (a) $\log \Lambda_{33}$, where IND performs well, and (b) β_{62}, where IND fails to capture this marginal well. Results are based on a single gold standard run (solid), five SMC runs with the RW kernel (dash) and five SMC runs with the IND kernel (dot-dash).

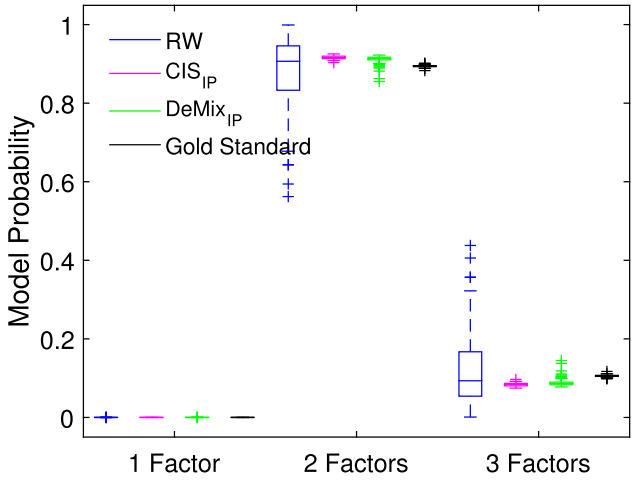


Figure 4: FA example: Boxplots of the model choice probabilities for RW, IND CIS_{IP}, IND DeMix_{IP} and the gold standard.

Evidence

Figure 4 shows boxplots of the estimated model probabilities which are calculated using the model evidence estimates. It is clear that the two factor model is preferred. The three factor model has some known issues with posterior approximation so it is not surprising that there is some discrepancy between the model probabilities estimated from the independent proposal and the probabilities estimated from the gold standard. However, the bias when recycling all independent proposals is small and the estimates are precise. Table 2 shows the evidence estimates and their efficiency, and this supports the claim that the IND proposals with CIS_{IP} or DeMix_{IP} recycling is the most efficient method for estimating the evidence in this example.

4.2 Econometrics Example

In the previous example, our method outperformed RW in terms of the number of likelihood evaluations and the precision of estimators, but it also required more time due to the overhead of constructing the independent proposal, simulating from it and evaluating its density. Here we consider an example more suited for the independent proposal since it has a non-trivial cost in evaluating the likelihood function so we are able to achieve reduced run times.

The "bad environment – good environment" (BEGE) model of Bekaert et al. (2015) is a flexible model used to describe the time-varying, non-Gaussian innovations in financial asset return data. The innovations in returns are modelled using a linear combination of a "bad environment" component and a "good environment" component,

$$r_{t+1} = \mu_t + u_{t+1},$$

Factors	Method	Sampler	$\log \widehat{Z} + 903$	avg. evals	efficiency
One	GOLD RW	SMC	-111.26	2.3×10^8	NA
	RW	SMC	-111.22	2.4×10^6	1.0
		CIS_{PP}	-111.22	$''$	1.0
		DeMix_{PP}	-111.22	$''$	1.0
	IND	SMC	-111.09	5.5×10^5	3.8
		CIS_{PP}	-111.09	$''$	4.0
		DeMix_{PP}	-111.09	$''$	3.8
		CIS_{IP}	-111.27	$''$	2,300
		DeMix_{IP}	-111.28	$''$	530
Two	GOLD RW	SMC	-0.21	4.7×10^8	NA
	RW	SMC	-0.04	5.1×10^6	1.0
		CIS_{PP}	-0.05	$''$	1.0
		DeMix_{PP}	-0.04	$''$	0.99
	IND	SMC	-0.43	9.7×10^5	900
		CIS_{PP}	-0.44	$''$	880
		DeMix_{PP}	-0.43	$''$	910
		CIS_{IP}	-0.22	$''$	240,000
		DeMix_{IP}	-0.24	$''$	39,000
Three	GOLD RW	SMC	-2.34	1.5×10^9	NA
	RW	SMC	-2.59	1.3×10^7	1.0
		CIS_{PP}	-2.57	$''$	0.78
		DeMix_{PP}	-2.59	$''$	0.97
	IND	SMC	-4.17	1.5×10^6	7.2
		CIS_{PP}	-4.17	$''$	7.2
		DeMix_{PP}	-4.17	$''$	7.2
		CIS_{IP}	-2.61	$''$	88
		DeMix_{IP}	-2.57	$''$	99

Table 2: FA example: Log mean of the estimated evidence for each model and efficiency.

$$\begin{aligned} u_{t+1} &= \sigma_p \omega_{p,t+1} - \sigma_n \omega_{n,t+1}, \text{ where} \\ \omega_{p,t+1} &\sim \tilde{\Gamma}(p_t, 1), \\ \omega_{n,t+1} &\sim \tilde{\Gamma}(n_t, 1), \end{aligned}$$

r_t is the return at time t and u_t is the innovation in the return at time t. $\tilde{\Gamma}(k,h)$ is the so-called de-meaned gamma distribution with probability density function

$$\tilde{\Gamma}(x;k,h) = \frac{1}{\Gamma(k)h^k}(x+kh)^{k-1}\exp\left(-\frac{1}{h}(x+kh)\right),$$

for $x > -kh$. The shape parameters for the good and bad environments, p_t and n_t respectively, are modelled using

$$p_t = p_0 + \rho_p p_{t-1} + \frac{\phi_p^+}{2\sigma_p^2} u_t^2 I_{u_t \geq 0} + \frac{\phi_p^-}{2\sigma_p^2} u_t^2 (1 - I_{u_t \geq 0}),$$

$$n_t = n_0 + \rho_n n_{t-1} + \frac{\phi_n^+}{2\sigma_n^2} u_t^2 I_{u_t \geq 0} + \frac{\phi_n^-}{2\sigma_n^2} u_t^2 (1 - I_{u_t \geq 0}).$$

The parameters which we wish to estimate are $\boldsymbol{\theta} = (p_0, n_0, \rho_p, \rho_n, \phi_p^+, \phi_n^+, \phi_p^-, \phi_n^-, \sigma_p, \sigma_n)$ and we assume that $\mu_t = 0$.

We use the Matlab code available from Bekaert et al. (2015) which uses numerical integration and differentiation to estimate the likelihood. This procedure makes the likelihood expensive to evaluate, at 2–4 seconds per single likelihood evaluation on an Intel Core i7-4790 CPU @ 3.60GHz.

Daily closing prices for the S&P 500 Composite Index for the period 31/12/1989 to 29/7/2016 were obtained from Bloomberg (Bloomberg, 2017). Log returns were calculated and we apply the BEGE model to the resulting dataset of 6690 daily return observations.

The priors are

$$\begin{aligned} p_0 &\sim \mathcal{U}(10^{-4}, 0.3), \\ n_0, \phi_p^+, \phi_p^-, \phi_n^- &\sim \mathcal{U}(10^{-4}, 0.5), \\ \rho_p, \rho_n &\sim \mathcal{U}(10^{-4}, 0.99), \\ \phi_n^+ &\sim \mathcal{U}(10^{-4}, 0.005), \\ \sigma_p &\sim \mathcal{U}(-0.2, 0.1), \\ \sigma_n &\sim \mathcal{U}(10^{-4}, 0.1), \end{aligned}$$

which, in addition to the requirements related to stationarity that $\rho_p + \frac{1}{2}\phi_p^+ + \frac{1}{2}\phi_p^- \leq 0.995$ and $\rho_n + \frac{1}{2}\phi_n^+ + \frac{1}{2}\phi_n^- \leq 0.995$, impose constraints on the parameters. Proposals made using a RW which do not satisfy the prior constraints are rejected without any likelihood evaluations which lowers the acceptance probability and therefore inflates the number of MCMC repeats. On the other hand, it is possible with an IND proposal to continue drawing candidates without any likelihood or proposal density evaluations until a candidate which satisfies the constraints is made.

We perform 100 runs using $N = 1,000$ particles. For this example, we use beta marginals and a 2 component MGMM for dependence in the IND proposal. The gold standard in this example is a long MCMC run with 400,000 iterations, taking a 4,000 iteration burn-in and retaining every 40th sample.

Posterior Inference

The posterior marginals in this example are relatively simple and can be explored with a RW or IND kernel (see Appendix D for figures), but the RW SMC runs are significantly more expensive to perform. RW uses on average 3.6 times more log-likelihood evaluations than IND and the run times are 3.2 times longer.

Posterior median estimates and measures of efficiency are given in Table 3. The IND proposal outperforms RW before recycling and IND with IP recycling gives the

best results. This improvement may be related to the increase in the ESS targeting the posterior. The ESS using only the final N samples is 1,000 whereas CIS_{PP}, DeMix_{PP}, CIS_{IP} and DeMix_{IP} result in median ESS values of approximately 2600, 2600, 3700 and 6200, respectively.

	GOLD	RW			IND				
	MCMC	SMC	CIS_{PP}	DeMix_{PP}	SMC	CIS_{PP}	DeMix_{PP}	CIS_{IP}	DeMix_{IP}
p_0	0.01	1	1.1	1.3	3.9	6.4	6.8	12	16
n_0	0.00	1	1.3	1.5	4.2	4.7	6.0	9.5	14
ρ_p	0.97	1	0.9	1.0	3.5	4.5	5.1	9.4	15
ρ_n	0.02	1	1.4	1.4	4.3	9.1	9.2	10	9.8
ϕ_p^+	0.03	1	1.2	1.4	4.9	8.3	9.7	16	20
ϕ_n^+	0.13	1	1.3	1.3	3.1	5.6	6.7	15	20
ϕ_p^-	0.00	1	1.6	1.7	4.1	7.7	8.7	18	29
ϕ_n^-	0.89	1	1.2	1.3	3.9	6.6	7.3	14	17
σ_p	−0.16	1	1.1	1.0	3.7	5.9	5.7	10	9.1
σ_n	0.35	1	1.4	1.4	4.8	7.5	7.3	17	21

Table 3: BEGE example: Gold standard posterior median estimates and efficiency based on 100 SMC runs for all other estimators. The median estimates are the same as the gold standard to two decimal places.

Evidence

Boxplots of the logged evidence estimates based on 100 runs are shown in Figure 5, where it is clear that at least one of the kernel types results in biased SMC and PP recycling evidence estimates. The PP based estimators are not visibly different to the standard SMC estimators, whereas the IP estimators which use all candidates are remarkably precise and are closer to the SMC RW results. Despite the additional density computations required by DeMix_{IP}, the performance of CIS_{IP} and DeMix_{IP} is similar.

Table 4 shows the log estimates of the evidence and their efficiency. The estimators which use all candidates are the most precise by a large margin.

Method	Sampler	$\log \widehat{Z} - 22{,}098$	avg. evals	efficiency
RW	SMC	1.35	4.8×10^5	1.0
	CIS_{PP}	1.35	〃	1.0
	DeMix_{PP}	1.35	〃	0.98
IND	SMC	0.70	1.3×10^5	11
	CIS_{PP}	0.69	〃	11
	DeMix_{PP}	1.70	〃	11
	CIS_{IP}	1.13	〃	835
	DeMix_{IP}	1.14	〃	1,500

Table 4: BEGE example: Log mean of the estimated evidence and efficiency. Efficiency is based on VAR rather than MSE.

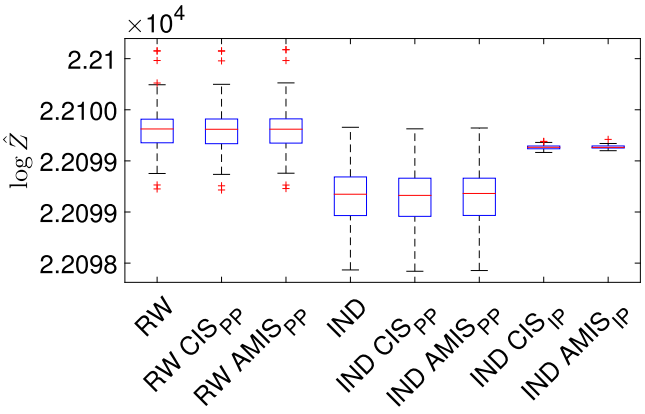


Figure 5: BEGE example: Boxplots of 100 log evidence estimates for all combinations of recycling method and MCMC kernel.

5 Discussion

A flexible independent MCMC proposal for SMC has been developed and the potential advantages of this approach have been demonstrated. The implementation of independent proposals in the SMC context is highly parallelisable and offers the ability to reuse all MCMC candidates in estimating the posterior and evidence. We have described a general method for forming these proposals based on modelling the marginals and the dependence separately, and the specific proposals applied here are based on modelling dependence through MGMMs. Our results are competitive with the multivariate normal random walk kernel in the examples presented here, even before recycling is applied. We find that if the independent proposals cover the tails of the target well, then significant improvements are achieved by recycling the candidates.

The three factor model in Section 4.1 and the example in Appendix G are applications with significant posterior complexity that push the boundaries of what can currently be achieved with our independent proposal. In the example in Appendix G, the MGMM copula-based independent proposal does not provide sufficient tail coverage yet we obtain very precise IS estimates of the evidence and we are able to improve the SE for the lower and upper 95% quantiles using recycling. We found that it is sometimes possible to detect lack of tail coverage through a small ESS after recycling. However, finding a proposal which covers the tails remains challenging for complex targets and a subject which we wish to tackle in further research. Future work may consider a mixture of student's t copulas, which models symmetric tail dependence, or defensive mixture distributions (Hesterberg, 1995), which use a mixture of some approximation q^{ϕ_t} of the target with a sampling distribution that improves tail coverage at the cost of efficiency. We hope that these examples will motivate further research on this topic to increase the capabilities of SMC with an independent proposal.

The concept presented in Section 3.1 for transforming the marginal distributions of a multivariate random variable to approximately $\mathcal{N}(0, 1)$ is a powerful tool to allow for

separate and straightforward modelling of dependence. These transformations may also be relevant to other proposals, including the multivariate normal random walk proposal. However, this method relies on a reasonable fit of the chosen marginals and the ability to simulate from the resulting marginal distributions via the inversion method. If an analytical expression for the quantile function is not available, the bisection method can be used to determine the quantiles, though this may require non-negligible computation time.

The use of independent proposals allows for the novel possibilities of improving acceptance probability estimates through Rao-Blackwellisation or estimating particle-specific acceptance probabilities using a computational effort that is similar to only a single MCMC iteration on all N particles. For particles $\{\boldsymbol{\theta}_t^i\}_{i=1}^N$ and candidates $\{\boldsymbol{\theta}_t^{j,*}\}_{j=1}^N$, we can compute the acceptance probability $\alpha(\boldsymbol{\theta}_t^i, \boldsymbol{\theta}_t^{j,*})$, from which we can obtain a Rao-Blackwellised estimate $\hat{p}_{\text{acc}}^t = \frac{1}{N^2}\sum_{i=1}^N \sum_{j=1}^N \alpha(\boldsymbol{\theta}_t^i, \boldsymbol{\theta}_t^{j,*})$ or a particle specific estimate $\hat{p}_{\text{acc}}^{i,t} = \frac{1}{N}\sum_{j=1}^N \alpha(\boldsymbol{\theta}_t^i, \boldsymbol{\theta}_t^{j,*})$ of the acceptance probabilities. These acceptance probabilities can be applied in choosing the number of MCMC repeats. Using a particle specific number of repeats, R_t^i, removes the restrictive assumption of having a constant probability of moving a particle regardless of its location in the parameter space. It may also be useful to set an upper limit on the number of move steps per particle, $R_t^i \leq R_{\max}$ for all $i = 1, \ldots, N$ to ensure that no one particle is assigned a large amount of computational load.

The candidates from the independent proposals in the empirical studies of Section 4 were drawn using pseudo-random numbers. As a variance reduction technique, the method known as randomised quasi-Monte Carlo (RQMC, e.g. Cranley and Patterson (1976)) could be used to improve the CIS_{IP} and DeMix_{IP} estimators. Using quasi-random samples instead of pseudo-random samples in Monte Carlo integration leads to estimators that have a faster convergence rate, but some randomisation must be introduced to maintain the unbiasedness property of the estimator.

If some initial approximation of the posterior $h(\boldsymbol{\theta})$ is available, it can be used as the initial SMC distribution with targets following the geometric path $\pi_t(\boldsymbol{\theta}|\mathbf{y}) \propto [f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})]^{\gamma_t} h(\boldsymbol{\theta})^{1-\gamma_t}$ in order to reduce computation (Donnet and Robin, 2017).

Independent proposals were used in SMC here via MCMC proposals in the move step but independent proposals could be used in the move step in several ways. We describe an alternative method in Appendix H which is based on skipping the resampling stage and using independent proposals with an IS correction to diversify the particles. This method is similar to adaptive importance sampling (e.g. Oh and Berger (1992)) but with annealing of the targets. This may be a useful alternative but it is harder to implement in practice due to the difficulty in controlling the weight degeneracy and for other reasons which are discussed in Appendix H of the Online Resources.

In general, the proposed methods for applying independent MCMC kernels in SMC and reusing all candidates have performed well in comparison to the multivariate normal random walk kernel. Order of magnitude improvements in the ESS targeting the posterior can be achieved using these independent proposals, and the proposed evidence

estimators can lead to substantial improvements over existing estimators. However, our results also suggest that further research in this area is required.

Supplementary Material

Supplementary Material: Sequential Monte Carlo Samplers with Independent Markov Chain Monte Carlo Proposals (DOI: 10.1214/18-BA1129SUPP; .pdf).

References

Bekaert, G., Engstrom, E., and Ermolov, A. (2015). "Bad environments, good environments: A non-Gaussian asymmetric volatility model." *Journal of Econometrics*, 186(1): 258–275. MR3321537. doi: https://doi.org/10.1016/j.jeconom.2014.06.021. 767, 769

Beskos, A., Jasra, A., Kantas, N., and Thiery, A. (2016). "On the convergence of adaptive sequential Monte Carlo methods." *Annals of Applied Probability*, 26(2): 1111–1146. MR3476634. doi: https://doi.org/10.1214/15-AAP1113. 763

Bloomberg (2017). "S&P 500 Composite Index for the period 31/12/1989 to 29/7/2016." 769

Cappé, O., Godsill, S. J., and Moulines, E. (2007). "An overview of existing methods and recent advances in sequential Monte Carlo." *Proceedings of the IEEE*, 95(5): 899–924. 753

Chopin, N. (2002). "A sequential particle filter method for static models." *Biometrika*, 89(3): 539–552. MR1929161. doi: https://doi.org/10.1093/biomet/89.3.539. 753, 754, 756

Cranley, R. and Patterson, T. N. L. (1976). "Randomization of number theoretic methods for multiple integration." *SIAM Journal on Numerical Analysis*, 13(6): 904–914. MR0494820. doi: https://doi.org/10.1137/0713071. 772

Del Moral, P., Doucet, A., and Jasra, A. (2006). "Sequential Monte Carlo samplers." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68: 411–436. MR2278333. doi: https://doi.org/10.1111/j.1467-9868.2006.00553.x. 753, 754, 755

Del Moral, P. and Miclo, L. (2000). "Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering." *Seminaire de Probabilites XXXIV, Lecture notes in Mathematics*, 1–145. MR1768060. doi: https://doi.org/10.1007/BFb0103798. 753

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1): 1–38. MR0501537. 759, 760

Donnet, S. and Robin, S. (2017). "Using deterministic approximations to accelerate SMC for posterior sampling." https://arxiv.org/abs/1707.07971. 772

Drovandi, C. C. and Pettitt, A. N. (2011). "Estimation of parameters for macroparasite population evolution using approximate Bayesian computation." *Biometrics*, 67(1): 225–233. MR2898834. doi: https://doi.org/10.1111/j.1541-0420.2010.01410.x. 756

Elvira, V., Martino, L., Luengo, D., and Bugallo, M. (2015). "Efficient multiple importance sampling estimators." *IEEE Signal Processing Letters*, 22(10): 1757–1761. 763

Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). "The meta-elliptical distributions with given marginals." *Journal of Multivariate Analysis*, 82(1): 1–16. MR1918612. doi: https://doi.org/10.1006/jmva.2001.2017. 758

Finke, A. (2015). "On extended state-space constructions for Monte Carlo methods." Ph.D. thesis, University of Warwick. 757

Gelman, A., Roberts, G., and Gilks, W. (1996). "Efficient Metropolis jumping rules." *Bayesian Statistics*, 5: 599–607. MR1425429. 756

Gerber, M., Chopin, N., and Whiteley, N. (2017). "Negative association, ordorder and convergence of resampling methods." *arXiv:1707.01845*. 755

Gramacy, R., Samworth, R., and King, R. (2010). "Importance tempering." *Statistics and Computing*, 20(1): 1–7. MR2578072. doi: https://doi.org/10.1007/s11222-008-9108-5. 757

Hesterberg, T. (1995). "Weighted average importance sampling and defensive mixture distributions." *Technometrics*, 37(2): 185–194. 771

Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. (2011). "Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo." *Scandinavian Journal of Statistics*, 38(1): 1–22. MR2760137. doi: https://doi.org/10.1111/j.1467-9469.2010.00723.x. 755

Keribin, C. (2000). "Consistent estimate of the order of mixture models." *Sankhyā: The Indian Journal of Statistics, Series A*, 62(1): 49–66. MR1769735. 760

Kong, A., Liu, J. S., and Wong, W. H. (1994). "Sequential imputations and Bayesian missing data problems." *Journal of the American Statistical Association*, 89(425): 278–288. 755

Lopes, H. F. and West, M. (2004). "Bayesian model assessment in factor analysis." *Statistica Sinica*, 14(1): 41–67. MR2036762. 764, 765

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons. MR1789474. doi: https://doi.org/10.1002/0471721182. 760

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equations of state calculations by fast computing machines." *Journal of Chemical Physics*, 12(6): 1087–1092. 754

Neal, R. M. (2001). "Annealed importance sampling." *Statistics and Computing*, 11: 125–139. MR1837132. doi: https://doi.org/10.1023/A:1008923215028. 755

Nguyen, T. L. T., Septier, F., Peters, G. W., and Delignon, Y. (2014). "Improving SMC sampler estimate by recycling all past simulated particles." In *2014 IEEE Workshop. In Statistical Signal Processing (SSP)*, p. 117–120. 754, 757

Nguyen, T. L. T., Septier, F., Peters, G. W., and Delignon, Y. (2016). "Efficient sequential Monte-Carlo samplers for Bayesian inference." *IEEE Transactions on Signal Processing*, 64(5): 1305–1319. MR3459546. doi: https://doi.org/10.1109/TSP.2015.2504342. 757, 758

Oh, M.-S. and Berger, J. O. (1992). "Adaptive importance sampling in Monte Carlo integration." *Journal of Statistical Computation and Simulation*, 41(3–4): 143–168. MR1276184. doi: https://doi.org/10.1080/00949659208810398. 772

Owen, A. and Zhou, Y. (2000). "Safe and effective importance sampling." *Journal of the American Statistical Association*, 95(449): 135–143. MR1803146. doi: https://doi.org/10.2307/2669533. 757

Pearson, K. (1894). "Contributions to the mathematical theory of evolution." *Philosophical Transactions of the Royal Society of London A*. 759

Schäfer, C. and Chopin, N. (2013). "Sequential Monte Carlo on large binary sampling spaces." *Statistics and Computing*, 23(2): 163–184. MR3016936. doi: https://doi.org/10.1007/s11222-011-9299-z. 754

Schmidl, D., Czado, C., Hug, S., and Theis, F. J. (2013). "A vine-copula based adaptive MCMC sampler for efficient inference of dynamical systems." *Bayesian Analysis*, 8(1): 1–22. MR3036249. doi: https://doi.org/10.1214/13-BA801. 754, 759

Schwarz, G. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6(2): 461–464. MR0468014. 760

Silva, R., Kohn, R., Giordani, P., and Mun, X. (2010). "A copula based approach to adaptive sampling." https://arxiv.org/abs/1002.4775. 754, 759

Sklar, M. (1959). "Fonctions de répartition à n dimensions et leurs marges." *Publications de l'Institut de statistique de l'Université de Paris*, 229–231. MR0125600. 754, 758

South, L. F., Pettitt, A. N., and Drovandi, C. C. (2018). "Supplementary Material: Sequential Monte Carlo Samplers with Independent Markov Chain Monte Carlo Proposals." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1129SUPP. 757

Tierney, L. (1994). "Markov chains for exploring posterior distributions." *The Annals of Statistics*, 22(4): 1701–1728. MR1329166. doi: https://doi.org/10.1214/aos/1176325750. 754

Tran, M.-N., Giordani, P., Mun, X., Kohn, R., and Pitt, M. K. (2014). "Copula-type estimators for flexible multivariate density modeling using mixtures." *Journal of Computational and Graphical Statistics*, 23(4): 1163–1178. MR3270716. doi: https://doi.org/10.1080/10618600.2013.842918. 759, 760

Veach, E. and Guibas, L. (1995). "Optimally combining sampling techniques for Monte Carlo rendering." In *SIGGRAPH 1995 Conference Proceedings*, 419-428. Addison-Wesley. 757

West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*. New York: Springer-Verlag New York. MR1482232. 764

Acknowledgments

The authors would like to thank Víctor Elvira and Khoa Tran for helpful discussions on this work and Adam Clements for discussions about the BEGE example. Computational resources and services used in this work were provided by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia.