A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models

Eugenia Koblents · Joaquín Míguez

Abstract This paper addresses the Monte Carlo approximation of posterior probability distributions. In particular, we consider the population Monte Carlo (PMC) technique, which is based on an iterative importance sampling (IS) approach. An important drawback of this methodology is the degeneracy of the importance weights (IWs) when the dimension of either the observations or the variables of interest is high. To alleviate this difficult , we propose a new method that performs a nonlinear transformation of the IWs. This operation reduces the weight variation, hence it avoids degeneracy and increases the efficien y of the IS scheme, specially when drawing from proposal functions which are poorly adapted to the true posterior. For the sake of illustration, we have applied the proposed algorithm to the estimation of the parameters of a Gaussian mixture model. This is a simple problem that enables us to discuss the main features of the proposed technique. As a practical application, we have also considered the challenging problem of estimating the rate parameters of a stochastic kinetic model (SKM). SKMs are multivariate systems that model molecular interactions in biological and chemical problems. We introduce a particularization of the proposed algorithm to SKMs and present numerical results.

Keywords Population Monte Carlo · Importance sampling · Degeneracy of importance weights · Stochastic kinetic models

1 Introduction

The problem of performing inference in multidimensional spaces appears in many practical applications. For example, it is of increasing interest in the biological sciences to develop new techniques that allow for the efficien estimation of the parameters governing the behavior of complex autoregulatory networks (Wilkinson 2011a). The main diffi culty often encountered when tackling this kind of problems is the design of numerical inference algorithms that scale up efficientl with the dimension of the parameter space.

A very common strategy, which has been successfully applied in a broad variety of complex problems, is the Monte Carlo methodology (Robert and Casella 2004). We consider a recently proposed technique known as population Monte Carlo (PMC) (Cappé et al. 2004), which is based on an iterative importance sampling (IS) approach. The aim of this method is the approximation of probability distributions by way of discrete random measures consisting of samples and associated importance weights (IWs). The target distribution is often the posterior distribution of a set of variables of interest, given some observed data.

The main advantages of the PMC scheme, compared to the widely established Markov chain Monte Carlo (MCMC) methodology (Robert and Casella 2004), are the possibility of developing parallel implementations, the sample independence and the fact that an unbiased estimate is provided at

each iteration, which avoids the need of a convergence period.

On the contrary, an important drawback of the IS approach, and particulary of PMC, is that its performance heavily depends on the choice of the proposal distribution (or importance function). When the target probability density function (pdf) is very sharp with respect to the proposal (this occurs when, e.g., the dimension of the variables of interest or the number of observations is high), the vast majority of the IWs become practically zero, leading to an extremely low number of representative samples (Kong et al. 1994; Doucet et al. 2000). This problem is commonly known as weight degeneracy and is closely related to the "curse of dimensionality" (Bengtsson et al. 2008). The issue was already mentioned in the original paper (Cappé et al. 2004). However, to the best of our knowledge, it has not been successfully addressed in the PMC framework.

The effort in the fiel of PMC algorithms has been directed toward the design of efficien proposal functions. The recently proposed mixture PMC technique (Cappé et al. 2008) models the importance functions as mixtures of kernels. The weights and the parameters of each mixture component are adapted along the iterations to minimize the Kullback-Leiber divergence between the target density and the proposal. This scheme also suffers from degeneracy and the authors of Cappé et al. (2008) propose to apply a Rao-Blackwellization scheme in order to mitigate this drawback.

Another recently proposed PMC scheme is based on the Gibbs sampling method (Djuric et al. 2011) and allows to sample efficientl from high-dimensional proposals. However, the IWs still present severe degeneracy due to the extreme values of the likelihood function in high-dimensional spaces. The technique is based on the multiple marginalized PMC algorithm introduced in Bugallo et al. (2009), Shen et al. (2010).

In this paper we propose a novel PMC method, termed nonlinear PMC (NPMC). The emphasis is not placed on the proposal update scheme, which can be very simple.[1] The main feature of the technique is the application of a nonlinear transformation to the IWs in order to reduce their variations. In this way, the efficien y of the sampling scheme is improved (specially when drawing from "poor" proposals) and the degeneracy of the weights is drastically mitigated even when the number of generated samples is relatively small. We provide a simple convergence analysis for two types of nonlinear transformations.

To illustrate the degeneracy problem and evaluate the performance of the proposed method we have used a simple Gaussian mixture model (GMM), already discussed in Cappé et al. (2004). The NPMC scheme outperforms the original PMC of Cappé et al. (2004) in terms of robustness and accuracy.

As a practical application, we have chosen the challenging problem of estimating the parameters in stochastic kinetic models (SKMs) (Wilkinson 2011a, 2011b; Golightly and Wilkinson 2011; Milner et al. 2013). SKMs describe the time evolution of the population of a set of chemical species, which evolve according to a set of constant rate parameters. We introduce a particularization of the NPMC algorithm to SKMs and show numerical results for the Lotka-Volterra model, consisting of two interacting species related by three reaction equations with associated unknown rates (Volterra 1926). In this scenario, the proposed method turns out advantageous compared to state-of-the-art MCMC techniques (Golightly and Wilkinson 2011).

The rest of the paper is organized as follows. A formal problem statement is presented in Sect. 2. In Sect. 3 the PMC algorithm is described and the weight degeneracy problem is discussed. The proposed NPMC method is introduced in Sect. 4, with a convergence analysis in Sect. 5. In Sect. 6 we present numerical results on a GMM that illustrate the effects of degeneracy and the performance of the proposed algorithm. In Sect. 7 we describe the practical application of the proposed algorithm to the estimation of the rate parameters of a SKM, and show numerical results. Section 8 summarizes the main finding of the paper.

2 Problem statement

Let $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]^\top$ be a column vector of K unobserved real random variables with prior density $p(\boldsymbol{\theta})$ and let $\mathbf{y} = [y_1, \ldots, y_N]^\top$ be a vector of N real random observations related to $\boldsymbol{\theta}$ by way of a likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$. In this paper we address the problem of approximating the posterior probability distribution of $\boldsymbol{\theta}$, i.e., the (conditional) distribution with density

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{1}$$

using a random grid of M points, $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{M}$, in the space of the random vector $\boldsymbol{\theta}$. Once the grid is generated, it is simple to approximate any moments of $p(\boldsymbol{\theta}|\mathbf{y})$, i.e., expectations of the form $E_{p(\boldsymbol{\theta}|\mathbf{y})}[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$, where $f : \mathbb{R}^K \to \mathbb{R}$ is some real integrable function of $\boldsymbol{\theta}$. For example, the posterior mean of $\boldsymbol{\theta}$ can be approximated as $E_{p(\boldsymbol{\theta}|\mathbf{y})}[\boldsymbol{\theta}] \approx \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\theta}^{(i)}$.

Unfortunately, the generation of samples that represent the probability measure $p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ adequately when K (or N) is large is normally a very difficul task. The main goal of this work is to devise and assess an efficien computational inference (Monte Carlo) methodology for the approximation of $p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ and its moments.

[1] Here, for instance, we restrict ourselves to multivariate normal densities when choosing the importance functions.

3 Population Monte Carlo

3.1 Importance sampling

One of the main applications of statistical Monte Carlo methods is the approximation of integrals of the form

$$(f,\pi) = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta},$$

where $\pi(\boldsymbol{\theta})$ is some pdf of interest (termed the *target* density). In problems of the type described in Sect. 2, the target density is the posterior pdf of $\boldsymbol{\theta}$, i.e., $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$. If $\pi(\boldsymbol{\theta})$ is some standard pdf, then it is straightforward to draw a random i.i.d. (independent and identically distributed) sample from $\pi(\boldsymbol{\theta})$ and approximate (f,π) by the sample mean. However, in many practical cases it is not possible to draw from $\pi(\boldsymbol{\theta})$ directly. A common approach to overcome this difficult is to apply an IS methodology (Robert and Casella 2004). The key idea is to generate an i.i.d. sample of size M, $\Theta^M = \{\boldsymbol{\theta}^{(i)}\}_{i=1}^M$, from a (simpler) proposal pdf $q(\boldsymbol{\theta})$, and then compute normalized IWs $w^{(i)}$ as

$$w^{(i)*} \propto \frac{\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})}, \qquad w^{(i)} = \frac{w^{(i)*}}{\sum_{j=1}^M w^{(j)*}}, \quad i = 1,\ldots,M.$$

Using Θ^M and the associated weights, we can construct a discrete random measure

$$\pi^M(d\boldsymbol{\theta}) = \sum_{i=1}^M w^{(i)}\delta_{\boldsymbol{\theta}^{(i)}}(d\boldsymbol{\theta}),$$

where $\delta_{\boldsymbol{\theta}^{(i)}}(d\boldsymbol{\theta})$ is the unit delta measure located at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)}$, and approximate (f,π) by the weighted sum

$$\left(f,\pi^M\right) = \sum_{i=1}^M w^{(i)} f\left(\boldsymbol{\theta}^{(i)}\right).$$

The efficien y of an IS algorithm depends heavily on the choice of the proposal, $q(\boldsymbol{\theta})$. However, in order to ensure the asymptotic convergence of the approximation (f,π^M), as $M \to \infty$, it is sufficien to select $q(\boldsymbol{\theta})$ such that $q(\boldsymbol{\theta}) > 0$ whenever $\pi(\boldsymbol{\theta}) > 0$ (Robert and Casella 2004). Finally, note that the computation of the normalized IWs requires that both $\pi(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ can be evaluated up to a proportionality constant.

3.2 Population Monte Carlo algorithm

The population Monte Carlo (PMC) method (Cappé et al. 2004) is an iterative IS scheme that seeks to generate a sequence of proposal pdf's $q_\ell(\boldsymbol{\theta})$, $\ell = 1,\ldots,L$, such that every new proposal is closer (in some adequate sense to be defined to the target density $\pi(\boldsymbol{\theta})$. Such scheme demands, therefore, the ability to learn about the target $\pi(\boldsymbol{\theta})$, given the set of samples and weights at the $(\ell-1)$-th iteration, in order to produce the new proposal $q_\ell(\boldsymbol{\theta})$ for the ℓ-th iteration ($\ell \geq 2$). The PMC algorithm is outlined in Table 1.

Table 1 Generic PMC algorithm (Cappé et al. 2004)

Iteration ($\ell = 1,\ldots,L$)**:**

1. Select a proposal pdf $q_\ell(\boldsymbol{\theta})$, based on $\tilde{\Theta}_{\ell-1}^M$ for $\ell \geq 2$. For $\ell = 1$, choose $q_1(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, the prior density.
2. Draw a set of i.i.d. samples $\Theta_\ell^M = \{\boldsymbol{\theta}_\ell^{(i)}\}_{i=1}^M$ from $q_\ell(\boldsymbol{\theta})$.
3. Compute normalized IWs $w_\ell^{(i)} \propto \pi(\boldsymbol{\theta}_\ell^{(i)})/q_\ell(\boldsymbol{\theta}_\ell^{(i)})$, $i = 1,\ldots,M$.
4. Perform a resampling step according to the weights $w_\ell^{(i)}$ to create an unweighted sample set $\tilde{\Theta}_\ell^M = \{\tilde{\boldsymbol{\theta}}_\ell^{(i)}\}_{i=1}^M$.

At every iteration of the algorithm it is possible to compute an estimate of (f,π) as

$$\left(f,\pi_\ell^M\right) = \sum_{i=1}^M w_\ell^{(i)} f\left(\boldsymbol{\theta}_\ell^{(i)}\right)$$

and, if the proposals $q_\ell(\boldsymbol{\theta})$ are actually improved across iterations, it can be expected that the approximation error $|(f,\pi) - (f,\pi_\ell^M)|$ also decreases with ℓ.

A frequently used index for the performance of Monte Carlo approximations of probability measures is the effective sample size (ESS) $M^{eff} = [\sum_{i=1}^M (w^{(i)})^2]^{-1}$ and its normalized version (NESS) $M^{neff} = M^{eff}/M$ (Kong et al. 1994; Doucet et al. 2000). We expect the ESS to increase along the iterations as the algorithm converges. Thus, it may be used to quantitatively monitor the convergence of the PMC algorithm and to stop the adaptation when the ESS reaches a steady value.

However, unless the proposal pdf is well tailored to the target density, the resulting IWs will often present very large variations, leading to a low number of effective samples. This problem is well known to affect IS schemes and is usually termed degeneracy of the weights (Kong et al. 1994; Doucet et al. 2000).

3.3 Degeneracy of the importance weights

The degeneracy of the IWs is a problem that arises when the normalized IWs $w^{(i)}$, $i = 1,\ldots,M$, of a set of samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^M$ present large fluctuation and their maximum, $\max_i w^{(i)}$, is close to one, leading to an extremely low ESS. This situation occurs when the target and the proposal densities are approximately mutually singular, i.e., they (essentially) have disjoint support.

The degeneracy of the IWs critically increases with K (Bengtsson et al. 2008), which has been widely accepted as one of the main drawbacks of IS. However, it can be easily verifie (numerically) that existing PMC methods can suffer from degeneracy even when applied to low dimensional

systems. Assume that the target pdf is the posterior given by Eq. (1) and consider a set of M samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{M}$ drawn from the prior pdf $p(\boldsymbol{\theta})$, which is the case at the firs iteration of the PMC algorithm. Assuming conditionally independent observations, the IW associated to the i-th sample is given by

$$w^{(i)} \propto p\big(\mathbf{y}|\boldsymbol{\theta}^{(i)}\big) = \prod_{n=1}^{N} p\big(y_n|\boldsymbol{\theta}^{(i)}\big), \quad i = 1, \ldots, M. \tag{2}$$

Thus, the IWs are obtained from a likelihood consisting of the product of a potentially large number of factors. As the number of observations N increases, the posterior probability concentrates in a smaller region (it becomes sharper), leading to a low probability of obtaining representative samples. This shows how in low dimensional systems degeneracy of the IWs can be motivated by a high number of observations N, unless the computational inference method is explicitly designed to account for this difficult . In Sect. 6 we present numerical results to support this claim, which provides a rationale to understand the poor performance of existing PMC methods with certain low dimensional models.

The degeneracy problem was already identifie in Cappé et al. (2004). However, to the best of our knowledge, no systematic solution has been provided so far for this problem. In the next section we introduce a new methodology to tackle the weight degeneracy, either due to large K or to large N. The key feature of the method is the application of a nonlinear transformation to the IWs, in order to reduce their variations and obtain an ESS that is large enough to adequately perform the proposal update and provide consistent estimates of the variables of interest.

4 Algorithms

In this section we describe the proposed algorithm, termed nonlinear PMC (NPMC). We adopt a simple proposal update scheme, where the importance functions are multivariate normal (MVN) pdf's with moments matched to the latest approximation of the posterior distribution. The key feature is the application of a nonlinear transformation of the IWs. Besides the basic version of the algorithm, we propose an adaptive version where this transformation is only applied when the value of the ESS is below a certain threshold. Finally, we explore different forms of the weight transformation.

4.1 Nonlinear PMC

Assume, in the sequel, that the target pdf is the posterior density given by Eq. (1). For simplicity, we select the importance functions in the PMC scheme as MVN densities. The initial proposal is selected as the prior, i.e., $q_1(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$. In the subsequent iterations

$$q_\ell(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell), \quad \ell = 2, \ldots, L,$$

where $\boldsymbol{\mu}_\ell$ is the mean vector and $\boldsymbol{\Sigma}_\ell$ is a positive definit covariance matrix. These parameters are chosen to match the moments of the distribution described by the discrete measure obtained at the previous iteration. In particular, we compute the mean and covariance as

$$\boldsymbol{\mu}_\ell = \frac{1}{M} \sum_{i=1}^{M} \tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)} \tag{3}$$

and

$$\boldsymbol{\Sigma}_\ell = \frac{1}{M} \sum_{i=1}^{M} \big(\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)} - \boldsymbol{\mu}_\ell\big)\big(\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)} - \boldsymbol{\mu}_\ell\big)^\top, \tag{4}$$

where $\{\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)}\}_{i=1}^{M}$ is the set of (unweighted) samples available after the $(\ell-1)$-th iteration. Note that this particular proposal update scheme is not a constraint of the algorithm. The importance functions can be designed as freely as in the standard PMC method.

The key modificatio of the algorithm is the computation of transformed IWs (TIWs). We introduce a sequence of nonlinear, real positive functions φ_ℓ^M, $\ell = 1, \ldots, L$, which depend both on the iteration index ℓ and the size-M sample at the $(\ell-1)$-th iteration. The unnormalized TIWs are computed as $\bar{w}_\ell^{(i)*} = \varphi_\ell^M(w_\ell^{(i)*})$, $i = 1, \ldots, M$, where $w_\ell^{(i)*}$ is the standard unnormalized IW associated to the sample $\boldsymbol{\theta}_\ell^{(i)}$.

The nonlinearity should be chosen so as to reduce the variation of the normalized TIWs, $\bar{w}_\ell^{(i)} = \bar{w}_\ell^{(i)*} / \sum_{j=1}^{M} \bar{w}_\ell^{(j)*}$. Intuitively, it should preserve the ordering of the samples (those with larger IWs should also have the largest TIWs) while reducing the difference $\max_i \bar{w}_\ell^{(i)} - \min_i \bar{w}_\ell^{(i)}$ or some other measure of weight variation. This modificatio of the algorithm mitigates the sensitivity of the conventional IS to the selection of the proposal pdf. The NESS computed from the TIWs $\bar{w}_\ell^{(i)}$ is denoted as $\bar{M}_\ell^{neff} = [M \sum_{i=1}^{M} (\bar{w}_\ell^{(i)})^2]^{-1}$. The proposed generic algorithm is outlined in Table 2.

Step 5 of the NPMC method involves multinomial resampling, which consists in sampling with replacement from the set $\{\boldsymbol{\theta}_\ell^{(i)}\}_{i=1}^{M}$ with probabilities equal to the associated TIWs $\bar{w}_\ell^{(i)}$, to obtain an unweighted set $\{\tilde{\boldsymbol{\theta}}_\ell^{(i)}\}_{i=1}^{M}$. This is not the only choice of resampling algorithm and we use it only for the sake of simplicity. See, e.g., Bain and Crisan (2008), Carpenter et al. (1999), for an overview of resampling techniques.

At each iteration $\ell = 1, \ldots, L$, we obtain two discrete approximations of the posterior distribution with density $\pi(\boldsymbol{\theta})$,

Table 2 Nonlinear PMC with target $\pi(\boldsymbol{\theta}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

Iteration ($\ell = 1, \ldots, L$):

1. Select the proposal pdf $q_\ell(\boldsymbol{\theta})$:
   - At iteration $\ell = 1$, let $q_1(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$.
   - At iterations $\ell = 2, \ldots, L$ the proposal is a MVN pdf $q_\ell(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$, where the mean and covariance are computed according to Eqs. (3) and (4).
2. Draw a set of M samples $\Theta_\ell^M = \{\boldsymbol{\theta}_\ell^{(i)}\}_{i=1}^M$ from $q_\ell(\boldsymbol{\theta})$.
3. Compute the unnormalized IWs

$$w_\ell^{(i)*} \propto \frac{p(\boldsymbol{\theta}_\ell^{(i)}|\mathbf{y})}{q_\ell(\boldsymbol{\theta}_\ell^{(i)})} \propto \frac{p(\mathbf{y}|\boldsymbol{\theta}_\ell^{(i)})p(\boldsymbol{\theta}_\ell^{(i)})}{q_\ell(\boldsymbol{\theta}_\ell^{(i)})}, \quad i = 1, \ldots, M.$$

4. Compute normalized TIWs as

$$\bar{w}_\ell^{(i)*} = \varphi_\ell^M\left(w_\ell^{(i)*}\right), \quad \bar{w}_\ell^{(i)} = \frac{\bar{w}_\ell^{(i)*}}{\sum_{j=1}^M \bar{w}_\ell^{(j)*}}, \quad i = 1, \ldots, M.$$

5. Resample to obtain an unweighted set $\tilde{\Theta}_\ell^M = \{\tilde{\boldsymbol{\theta}}_\ell^{(i)}\}_{i=1}^M$: for $i, j = 1, \ldots, M$, let $\tilde{\boldsymbol{\theta}}_\ell^{(i)} = \boldsymbol{\theta}_\ell^{(j)}$ with probability $\bar{w}_\ell^{(j)}$.

namely the measures

$$\bar{\pi}_\ell^M(d\boldsymbol{\theta}) = \sum_{i=1}^M \bar{w}_\ell^{(i)} \delta_{\boldsymbol{\theta}_\ell^{(i)}}(d\boldsymbol{\theta}) \quad \text{and}$$

$$\tilde{\pi}_\ell^M(d\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^M \delta_{\tilde{\boldsymbol{\theta}}_\ell^{(i)}}(d\boldsymbol{\theta}),$$

and the integral (f, π) can be approximated as either

$$\left(f, \bar{\pi}_\ell^M\right) = \sum_{i=1}^M \bar{w}_\ell^{(i)} f\left(\boldsymbol{\theta}_\ell^{(i)}\right) \quad \text{or}$$

$$\left(f, \tilde{\pi}_\ell^M\right) = \frac{1}{M}\sum_{i=1}^M f\left(\tilde{\boldsymbol{\theta}}_\ell^{(i)}\right).$$

The estimator $(f, \tilde{\pi}_\ell^M)$ involves one extra Monte Carlo step (resampling) and, hence, it has more variance than $(f, \bar{\pi}_\ell^M)$ (Douc et al. 2005). Therefore, we assume in the sequel that estimates are computed by way of the measure $\bar{\pi}_\ell^M$ unless explicitly stated otherwise.

Note as well that, since $\pi(\boldsymbol{\theta}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, any expectation with respect to the posterior distribution is actually an integral with respect to the measure $\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, i.e., $E_{p(\boldsymbol{\theta}|\mathbf{y})}[f(\boldsymbol{\theta})] = (f, \pi)$, and, therefore, it can be approximated using $\bar{\pi}_\ell^M$, namely, $E_{p(\boldsymbol{\theta}|\mathbf{y})}[f(\boldsymbol{\theta})] \approx (f, \bar{\pi}_\ell^M)$.

4.2 Modifie NPMC

The nonlinear transformation φ_ℓ^M is most useful at the firs iterations of the NPMC, when the proposal density is generally much broader than the target density and the standard IWs may display high variability. However, in some applications it may be possible to remove the nonlinear transformation after a few iterations, when the proposal is closer to the target.

Table 3 Modifie NPMC algorithm

Step 4 of the NPMC algorithm is replaced by the following computations:

4. Compute the normalized IWs $w_\ell^{(i)} = w_\ell^{(i)*} / \sum_{j=1}^M w_\ell^{(j)*}$ and the ESS $M_\ell^{eff} = [\sum_{i=1}^M (w_\ell^{(i)})^2]^{-1}$.
   If $M_\ell^{eff} < M_{min}^{eff}$, compute normalized TIWs $\bar{w}_\ell^{(i)*} = \varphi_\ell^M(w_\ell^{(i)*})$, $\bar{w}_\ell^{(i)} = \bar{w}_\ell^{(i)*} / \sum_{j=1}^M \bar{w}_\ell^{(j)*}$, $i = 1, \ldots, M$. Otherwise, set $\bar{w}_\ell^{(i)} = w_\ell^{(i)}$.

Thus, we propose a modificatio of the NPMC algorithm which consists in applying the nonlinear transformation only if the ESS M_ℓ^{eff} computed from the standard normalized IWs $w_\ell^{(i)}$ is below a specifie threshold M_{min}^{eff}. We recommend that the threshold M_{min}^{eff} be a relatively large value (e.g., $\frac{M}{2} \leq M_{min}^{eff} < M$) to ensure that the algorithm is suffi ciently stable before removing the transformation. The modifie algorithm only differs from the NPMC in step 4, which is outlined in Table 3.

4.3 Selecting the transformation of the IWs

The nonlinearity φ_ℓ^M may be constructed in multiple ways. In this section we describe and intuitively justify two specifi functions based on the "tempering" and the "clipping", respectively, of the standard IWs.

4.3.1 Tempering

In this case, the unnormalized TIWs are obtained as

$$\bar{w}_\ell^{(i)*} = \varphi_\ell^M\left(w_\ell^{(i)*}\right) = \left(w_\ell^{(i)*}\right)^{\gamma_\ell}, \quad i = 1, \ldots, M,$$

where $0 < \gamma_\ell \leq 1$. The sequence γ_ℓ, $\ell = 1, \ldots, L$, has to be adapted along the iterations, taking low values at the firs steps and getting closer to 1 as the algorithm converges. The sequence γ_ℓ can be selected a priori, regardless of the values of the IWs. For instance, it may be constructed as a polynomial function $\gamma_\ell \propto \ell^m$, $m \in \mathbb{N}$, or a sigmoid function $\gamma_\ell = \frac{1}{1+e^{-\ell}}$ of the iteration index ℓ. While in simple examples this procedure provides a remarkable reduction of the weight variations and an increase of the ESS, in complex problems it is not enough to guarantee a stable and consistent convergence.

This technique is closely related to the *simulated tempering* of the target density, which has been widely studied in the MCMC literature (Gramacy et al. 2010; Marinari and Parisi 2007). More recently, a class of sequential

Monte Carlo (SMC) samplers that rely on IS and can encompass PMC methods as a particular case have been proposed (Del Moral et al. 2006), and tempering techniques have been specificall considered within this framework (Del Moral et al. 2006; Jasra et al. 2011; Beskos et al. 2012). However, the IWs in the SMC methodology of Del Moral et al. (2006) are computed in the conventional manner, and tempering is only applied to the target density (Beskos et al. 2012). Therefore, these methods depart from the NPMC algorithm, as the same set of samples in the parameter space (even drawn from the same proposal) would be weighted differently. However, it is possible to derive a NPMC algorithm with tempering within the framework of Del Moral et al. (2006), as shown in Appendix A, under some constraints on the choice of the importance functions. Unfortunately, the latter constraints rule out the class of $q_\ell(\boldsymbol{\theta})$ introduced in Sect. 4.1.

4.3.2 Clipping

We now introduce a simple and effective methodology that avoids the fittin of any parameters and guarantees a baseline ESS at all iterations. In particular, we perform a clipping procedure on the $M_T < M$ highest IWs at each iteration of the NPMC algorithm. Since the highest weights $w_\ell^{(i)}$ usually correspond to the most representative samples $\boldsymbol{\theta}_\ell^{(i)}$, we thus obtain fla TIWs in the region of interest of $\boldsymbol{\theta}$. As a consequence, at least M_T samples obtain non negligible weights at all iterations, allowing to consistently update the proposal.

To be specific at each iteration ℓ, consider a permutation $i_1, \ldots, i_M$ of the indices in $\{1, \ldots, M\}$ such that $w_\ell^{(i_1)*} \geq \ldots \geq w_\ell^{(i_M)*}$ and choose $M_T < M$. We select a threshold value $\mathcal{T}_\ell^M = w_\ell^{(i_{M_T})*}$ and apply clipping to the IWs $w_\ell^{(i_k)*} \geq \mathcal{T}_\ell^M$, $k = 1, \ldots, M_T - 1$. Thus, the unnormalized TIWs $\bar{w}_\ell^{(i)*}$, $i = 1, \ldots, M$, are computed from the original IWs $w_\ell^{(i)*}$ as[2]

$$\bar{w}_\ell^{(i)*} = \varphi_\ell^M\left(w_\ell^{(i)*}\right) = \min\left(w_\ell^{(i)*}, \mathcal{T}_\ell^M\right). \tag{5}$$

Note that, since $\mathcal{T}_\ell^M = w_\ell^{(i_{M_T}*)}$, the number of samples with equal TIWs is exactly M_T.

The selection of the parameter M_T in relation to the total number of samples M is not crucial. In practice, we have found that choosing $M_T/M = 0.1$ works well for many examples. If the total number of samples M is very large, it is not necessary that $M_T \propto M$. Indeed, M_T should be simply large enough to "identify" the region where the posterior probability mass is located. Correspondingly, for the asymptotic analysis of Sect. 5 we will assume that $M_T/M \to 0$ as $M \to \infty$.

This technique is a generalization of the one proposed in Koblents and Míguez (2011), which applies clipping to the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ instead of to the complete weights. However, transforming only the likelihood does not guarantee a sufficien ESS, and its performance heavily depends on the selection of the prior.

[2] According to Eq. (5) and the definitio of the threshold $\mathcal{T}_\ell^M$, φ_ℓ^M is a function of both the complete weight set $\{w_\ell^{(j)*}\}_{j=1}^M$ and the index of the weight to be transformed, i.e., $\varphi_\ell^M : \{w_\ell^{(j)*}, j = 1, \ldots, M\} \times \{1, \ldots, M\} \to [1, +\infty)$.

5 Convergence of nonlinear IS

The convergence of the original PMC scheme is easily justifie by the convergence of the standard IS method. Indeed, it can be proved (Geweke 1989) that the discrete measure $\pi_\ell^M(d\boldsymbol{\theta}) = \sum_{i=1}^M w_\ell^{(i)} \delta_{\boldsymbol{\theta}_\ell^{(i)}}(d\boldsymbol{\theta})$ converges to $\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ under mild assumptions, meaning that

$$\lim_{M\to\infty} \left|\left(f, \pi_\ell^M\right) - (f, \pi)\right| = 0 \quad \text{almost surely (a.s.)} \tag{6}$$

for every $\ell \in \{1, \ldots, L\}$ and any $f \in B(\mathbb{R}^K)$, where $B(\mathbb{R}^K)$ is the set of bounded real functions over $\mathbb{R}^K$.

In Sect. 5.1, we provide a result similar to Eq. (6) for the discrete measure $\bar{\pi}_\ell^M$ generated by the NPMC algorithm with a clipping transformation. The analysis, therefore, is concerned with the asymptotic performance of the approximation as the number of samples M grows, but not with the convergence as the iteration index ℓ increases. Hence, we shall drop the latter subscript for convenience in the sequel.

The section is completed with an analysis of the error induced by the tempering transformation. Note that when $\gamma < 1$ (again, we drop the iteration subscript ℓ and focus on a single iteration) the error in the approximation of (f, π) via the NPMC scheme with tempering does not vanish as $M \to \infty$. It is relatively straightforward, however, to fin an upper bound for the approximation error (with fi ed $\gamma < 1$ and $M \to \infty$) and then show that this error vanishes as $\gamma \to 1$. These results are formally obtained in Sect. 5.2.

5.1 Asymptotic convergence of IS estimators with clipping

5.1.1 Notation and basic assumptions

Let π be the pdf associated to the target probability distribution to be approximated, let q be the importance function used to propose samples in an IS scheme (not necessarily normalized) and let $h(\boldsymbol{\theta}) = a\pi(\boldsymbol{\theta})$ be a function proportional to π, with the proportionality constant $a > 0$ independent of $\boldsymbol{\theta}$. The samples drawn from the distribution associated to q are denoted $\boldsymbol{\theta}^{(i)}$, $i = 1, \ldots, M$, and their associated unnormalized IWs are $w^{(i)*} = h(\boldsymbol{\theta}^{(i)})/q(\boldsymbol{\theta}^{(i)})$, $i = 1, \ldots, M$.

Let us defin the weight function $g(\boldsymbol{\theta}) = h(\boldsymbol{\theta})/q(\boldsymbol{\theta})$ and, in particular, $g(\boldsymbol{\theta}^{(i)}) = w^{(i)*}$. The support of g is the same as the support of q, denoted $\mathbb{S} \subseteq \mathbb{R}^K$. If we assume that both $q(\boldsymbol{\theta}) > 0$ and $\pi(\boldsymbol{\theta}) \geq 0$ for any $\boldsymbol{\theta} \in \mathbb{S}$, then $g(\boldsymbol{\theta}) \geq 0$ for every $\boldsymbol{\theta} \in \mathbb{S}$ as well. Also, trivially, $\pi \propto gq$, with the proportionality constant independent of $\boldsymbol{\theta}$. These assumptions are standard for classical IS.

The approximation π^M of the target probability measure generated by the standard IS method is constructed form the normalized IWs $w^{(i)}$, namely

$$\pi^M(d\boldsymbol{\theta}) = \sum_{i=1}^{M} w^{(i)} \delta_{\boldsymbol{\theta}^{(i)}}(d\boldsymbol{\theta}),$$

where $w^{(i)} = \frac{g(\boldsymbol{\theta}^{(i)})}{\sum_{j=1}^{M} g(\boldsymbol{\theta}^{(j)})}$, $i = 1, \ldots, M$.

The nonlinear transformation φ^M of the weights is assumed to be of a clipping class, as described in Sect. 4.3.2. We note that, given an index permutation $i_1, \ldots, i_M$ such that $w^{(i_1)*} \geq \ldots \geq w^{(i_M)*}$, the transformation φ^M can be expressed as

$$\varphi^M\left(w^{(i_k)*}\right) = \begin{cases} w^{(i_{M_T})*}, & \text{for } k = 1, \ldots, M_T, \text{ and} \\ w^{(i_k)*}, & \text{for } k = M_T + 1, \ldots, M. \end{cases} \tag{7}$$

We assume that the weight function $g \in B(\mathbb{R}^K)$ is upper bounded, and thus the TIWs satisfy $\bar{w}^{(i)*} \leq \|g\|_\infty = \sup_{\mathbf{z} \in \mathbb{R}^K} |g(\mathbf{z})| < \infty$.

The approximation $\bar{\pi}^M$ of the target probability measure generated by the nonlinear IS method is constructed form the normalized TIWs $\bar{w}^{(i)}$ as

$$\bar{\pi}^M(d\boldsymbol{\theta}) = \sum_{i=1}^{M} \bar{w}^{(i)} \delta_{\boldsymbol{\theta}^{(i)}}(d\boldsymbol{\theta}),$$

where $\bar{w}^{(i)} = \frac{\varphi^M(g(\boldsymbol{\theta}^{(i)}))}{\sum_{j=1}^{M} \varphi^M(g(\boldsymbol{\theta}^{(j)}))}$, $i = 1, \ldots, M$. Additionally, we introduce an approximation $\check{\pi}^M$ constructed from a set of unnormalized TIWs $\check{w}^{(i)}$ that will be referred to as "bridge weights" in the sequel, namely

$$\check{\pi}^M(d\boldsymbol{\theta}) = \sum_{i=1}^{M} \check{w}^{(i)} \delta_{\boldsymbol{\theta}^{(i)}}(d\boldsymbol{\theta}), \tag{8}$$

where $\check{w}^{(i)} = \frac{\varphi^M(g(\boldsymbol{\theta}^{(i)}))}{\sum_{j=1}^{M} g(\boldsymbol{\theta}^{(j)})}$, $i = 1, \ldots, M$.

5.1.2 Asymptotic convergence

We aim at proving that $\lim_{M\to\infty} |(f, \bar{\pi}^M) - (f, \pi)| = 0$ a.s. for any $f \in B(\mathbb{R}^K)$. To obtain such a result, we split the problem into simpler questions by applying the triangle inequality

$$\left|\left(f, \bar{\pi}^M\right) - (f, \pi)\right| \leq \left|\left(f, \bar{\pi}^M\right) - \left(f, \pi^M\right)\right| + \left|\left(f, \pi^M\right) - (f, \pi)\right|. \tag{9}$$

The second term on the right hand side of (9) is handled easily using standard IS theory. For the firs term, we have to prove that the discrete measure generated by the *nonlinear* IS method ($\bar{\pi}^M$) converges to the discrete measure generated by the standard IS method (π^M). This can be done by resorting to another triangle inequality,

$$\left|\left(f, \bar{\pi}^M\right) - \left(f, \pi^M\right)\right| \leq \left|\left(f, \bar{\pi}^M\right) - \left(f, \check{\pi}^M\right)\right| + \left|\left(f, \check{\pi}^M\right) - \left(f, \pi^M\right)\right|, \tag{10}$$

that reveals the role of the bridge measure in (8).

The following lemma establishes the asymptotic convergence of the term $|(f, \bar{\pi}^M) - (f, \check{\pi}^M)|$ in (10).

Lemma 1 *Assume that* $\lim_{M\to\infty} \frac{M_T}{M} = 0$, $g \in B(\mathbb{R}^K)$, *and the transformation function* φ^M *satisfie* (7). *Then, for every* $f \in B(\mathbb{R}^K)$ *and sufficientl large* M, *there exist positive constants* c_1, c_1' *independent of* M *and* M_T *such that*

$$\mathbb{P}\left\{\left|\left(f, \bar{\pi}^M\right) - \left(f, \check{\pi}^M\right)\right| \leq c_1 \frac{M_T}{M}\right\} \geq 1 - \exp\left(-c_1' M\right).$$

Proof See Appendix B.

Next, we establish the convergence of the bridge measure $\check{\pi}^M$ toward π^M.

Lemma 2 *Assume that* $\lim_{M\to\infty} \frac{M_T}{M} = 0$, $g \in B(\mathbb{R}^K)$ *and the transformation function* φ^M *satisfie* (7). *Then, for every* $f \in B(\mathbb{R}^K)$ *there exist positive constants* c_2, c_2' *independent of* M *and* M_T *such that*

$$\mathbb{P}\left\{\left|\left(f, \check{\pi}^M\right) - \left(f, \pi^M\right)\right| \leq c_2 \frac{M_T}{M}\right\} \geq 1 - \exp\left(-c_2' M\right).$$

Proof See Appendix C.

The combination of Lemmas 1 and 2, together with the triangle inequality (10), yields the convergence of the error $|(f, \bar{\pi}^M) - (f, \pi^M)|$.

Lemma 3 *Assume that* $\lim_{M\to\infty} \frac{M_T}{M} = 0$, $g \in B(\mathbb{R}^K)$, *and the transformation function* φ^M *satisfie* (7). *Then, for every* $f \in B(\mathbb{R}^K)$, *and sufficientl large* M, *there exist positive constants* c, c' *independent of* M *and* M_T *such that*

$$\mathbb{P}\left\{\left|\left(f, \bar{\pi}^M\right) - \left(f, \pi^M\right)\right| \leq c \frac{M_T}{M}\right\} \geq 1 - 2\exp\left(-c' M\right).$$

In particular,

$$\lim_{M\to\infty}\left|\left(f,\bar{\pi}^M\right)-\left(f,\pi^M\right)\right|=0 \quad a.s.$$

Proof See Appendix D.

Finally, Lemma 3 can be combined with inequality (9) to yield the desired result, stated below.

Theorem 1 *Assume that* $\lim_{M\to\infty}\frac{M_T}{M}=0$, $g\in B(\mathbb{R}^K)$ *and the transformation function* φ^M *satisfie* (7). *Then, for every* $f\in B(\mathbb{R}^K)$,

$$\lim_{M\to\infty}\left|\left(f,\bar{\pi}^M\right)-(f,\pi)\right|=0 \quad a.s.$$

Proof It is classical result that (Geweke 1989)

$$\lim_{M\to\infty}\left|\left(f,\pi^M\right)-(f,\pi)\right|=0 \quad \text{a.s.} \tag{11}$$

Combining (11) with the second part of Lemma 3 and the triangle inequality in (9) yields the desired result. □

Remark 1 Lemma 3 shows that the approximation $\bar{\pi}^M$ that uses the transformed weights can be seen as a "distortion" of the conventional IS approximation π^M. Such distortion depends on the ratio M_T/M and, hence, can be controlled by the choice of M_T.

5.2 Asymptotic convergence of IS estimators with tempering

When the tempering transformation is applied, the TIWs can be written as

$$\bar{w}^{(i)}=\frac{g(\boldsymbol{\theta}^{(i)})^\gamma}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})^\gamma}, \quad i=1,\ldots,M. \tag{12}$$

If $\gamma<1$ is fi ed and $f\in B(\mathbb{R}^K)$ is non-constant, it is apparent that the integral $(f,\bar{\pi}^M)$ does not converge to (f,π) as $M\to\infty$. However, it is straightforward to fin an upper bound for the distortion with respect to the conventional IS approximation, (f,π^M), as given by the following proposition.

Proposition 1 *Assume that* $g\in B(\mathbb{R}^K)$, $\varphi^M(w)=w^\gamma$ *and both* $0<\gamma\leq 1$ *and* $M<\infty$ *are fi ed. Then, for every* $f\in B(\mathbb{R}^K)$,

$$\left|\left(f,\pi^M\right)-\left(f,\bar{\pi}^M\right)\right|\leq\left|\left(f\left(1-g^{\gamma-1}\right),\pi^M\right)\right| + \|f\|_\infty\left|\left(1-g^{\gamma-1},\pi^M\right)\right|. \tag{13}$$

Proof See Appendix E.

The inequality (13) is useful because it yields an upper bound for the distortion $|(f,\pi^M)-(f,\bar{\pi}^M)|$, introduced by the tempering nonlinearity, that depends on the standard IS approximating measure π^M alone. Since $1-g^{\gamma-1}\in B(\mathbb{R}^K)$, the standard convergence results for IS (Geweke 1989) can be applied to the integrals on the right hand side of (13) and, as a consequence,

$$\lim_{M\to\infty}\left|\left(f,\pi^M\right)-\left(f,\bar{\pi}^M\right)\right| \leq\left|\left(f\left(1-g^{\gamma-1}\right),\pi\right)\right|+\|f\|_\infty\left|\left(1-g^{\gamma-1},\pi\right)\right| \tag{14}$$

a.s. Moreover, (13) also shows that the difference $(f,\pi^M)-(f,\bar{\pi}^M)$ vanishes when $\gamma\to 1$. Indeed, when $\gamma\to 1$, $(1-g^{\gamma-1},\pi^M)\to 0$ and $(f(1-g^{\gamma-1}),\pi^M)\to 0$, hence

$$\lim_{\gamma\to 1}\left|\left(f,\pi^M\right)-\left(f,\bar{\pi}^M\right)\right|=0.$$

Similarly, from (14) we observe that

$$\lim_{\gamma\to 1}\lim_{M\to\infty}\left|\left(f,\bar{\pi}^M\right)-(f,\pi)\right|=0 \quad \text{a.s.},$$

as intuitively expected.

6 Example 1: a Gaussian mixture model

In this section we provide numerical results that illustrate the degeneracy problem and the performance of the proposed NPMC scheme applied to the Gaussian mixture model (GMM) example of Cappé et al. (2004).

6.1 Model

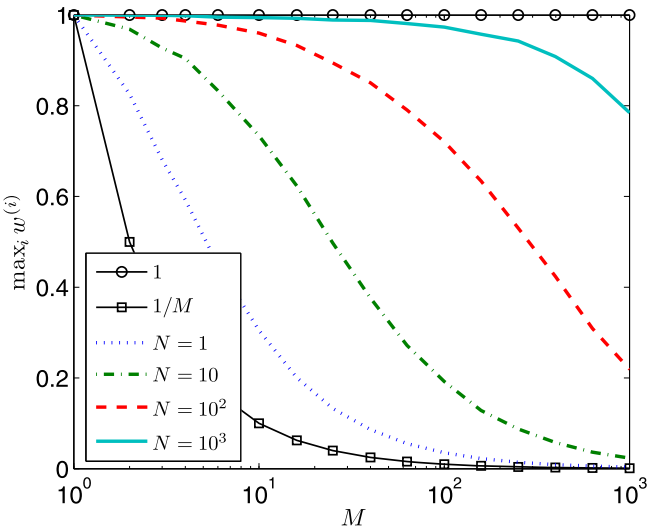
We consider the GMM given by

$$p(y|\boldsymbol{\theta})=\rho\mathcal{N}\left(y;\theta_1,\sigma^2\right)+(1-\rho)\mathcal{N}\left(y;\theta_2,\sigma^2\right) \tag{15}$$

where the variable of interest $\boldsymbol{\theta}=[\theta_1,\theta_2]^\top$ contains the means of the mixture components. The true values of the unknowns are set to $\boldsymbol{\theta}=[0,2]^\top$. The mixture coefficien and the variance of the components are assumed to be known and set to $\rho=0.2$ and $\sigma^2=1$.

We assume a prior pdf $p(\boldsymbol{\theta})=p(\theta_1)p(\theta_2)$ composed of equal independent components for each unknown, given by $p(\theta_k)=\mathcal{N}(\theta_k;\nu,\sigma^2/\lambda)$, for $k=1,2$. The hyperparameters are set to $\nu=1$ and $\lambda=0.1$.

A set $\mathbf{y}$ of N i.i.d. scalar observations are drawn from the mixture model in Eq. (15), and we aim at approximating the posterior pdf $\pi(\boldsymbol{\theta})=p(\boldsymbol{\theta}|\mathbf{y})$.



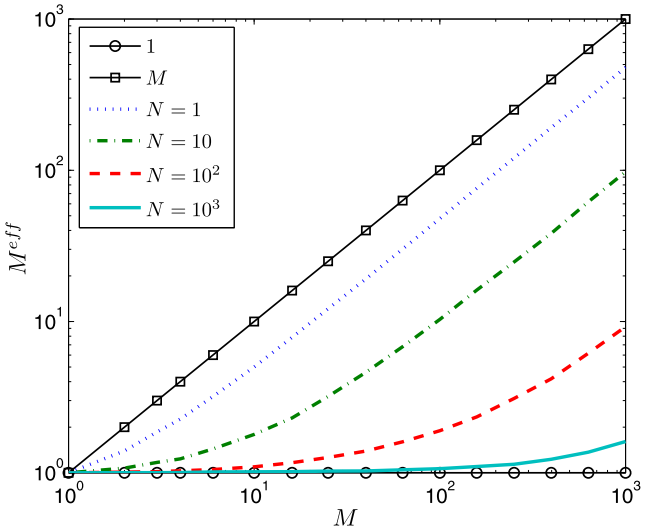


Fig. 1 Evolution of the average maximum IW $\max_i w^{(i)}$ (*left*) and the ESS M^{eff} (*right*) vs the number of observations N and the number of samples M. The curves corresponding to maximum degeneracy ($\max_i w^{(i)} = 1$ and $M^{eff} = 1$) are plotted with *circles*. The curves corresponding to the optimum case with uniform weights ($\max_i w^{(i)} = 1/M$ and $M^{eff} = M$) are depicted with *squares*. All curves are averaged over $P = 10^3$ independent simulation runs

6.2 Degeneracy of the importance weights

The model in Eq. (15) serves to illustrate the effects of the degeneracy problem in a simple and low dimensional IS example. Consider a set of M samples $\Theta^M = \{\boldsymbol{\theta}^{(i)}\}_{i=1}^M$ drawn form the prior pdf $p(\boldsymbol{\theta})$. The IWs are computed from the likelihood function as in Eq. (2). For this model, we have investigated the behavior of the maximum IW, $\max_i w^{(i)}$, and the ESS, M^{eff}, when the number of observations N increases. Let both the number of observations N and the number of samples M vary from 1 to 10^3. For each pair of values of N and M we have performed $P = 10^3$ simulation runs of the standard IS procedure.

In Fig. 1 (*left*) the average maximum IW is represented versus M and N. The curves representing the extreme cases $\max_i w^{(i)} = 1$ (degeneracy) and $\max_i w^{(i)} = 1/M$ (uniform weights) are also plotted on the graph. It can be observed that, for a fi ed M, as the number of observations N increases, $\max_i w^{(i)} \to 1$, leading to severe degeneracy.

Equivalently, in Fig. 1 (*right*) the average ESS is represented versus M and N. The cases $M^{eff} = 1$ and $M^{eff} = M$ are plotted for reference. It can be observed that, as N increases, the ESS is smaller for the same value of M. For example, with $N = 10^3$ observations and $M = 10^3$ samples, the average ESS is only 1.5.

6.3 Comparison of algorithms

In this section we compare, by way of computer simulations, the performance of the GMM-PMC scheme proposed in Cappé et al. (2004), which we reproduce in Table 4, the GMM-PMC with a clipping transformation, and the NPMC scheme of Sect. 4 with tempering and clipping transformations. We have performed $P = 10^4$ independent simulation runs of each algorithm, with $L = 10$ iterations and $M = 200$ samples per iteration.

Table 4 GMM-PMC algorithm (Cappé et al. 2004)

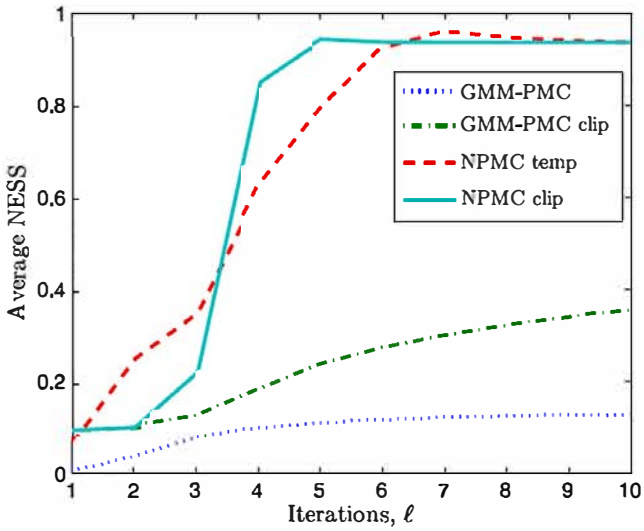
Initialization ($\ell = 0$):

1. Consider a set of p scales (variances) v_j and an initial number $r_j = m$ of samples per scale, $j = 1, \ldots, p$.
2. For $i = 1, \ldots, M = pm$, draw $\{\boldsymbol{\theta}_0^{(i)}\}$ from $q_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$.

Iteration ($\ell = 1, \ldots, L$):

1. For $j = 1, \ldots, p$
   - generate a sample $\{\boldsymbol{\theta}_\ell^{(i)}\}$ of size r_j from $q_\ell(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}_\ell^{(i)}; \boldsymbol{\theta}_{\ell-1}^{(i)}, v_j \mathbf{I}_K)$, where $\mathbf{I}_K$ denotes the identity matrix of size $K \times K$.
   - compute the normalized IWs $w_\ell^{(i)} \propto \frac{p(\mathbf{y}|\boldsymbol{\theta}_\ell^{(i)}) p(\boldsymbol{\theta}_\ell^{(i)})}{q_\ell(\boldsymbol{\theta}_\ell^{(i)})}$.
2. Resample with replacement the set $\{\boldsymbol{\theta}_\ell^{(i)}\}_{i=1}^M$ according to the weights $w_\ell^{(i)}$ to obtain $\{\tilde{\boldsymbol{\theta}}_\ell^{(i)}\}_{i=1}^M$.
3. For $j = 1, \ldots, p$ update r_j as the number of elements generated with variance v_j which have been resampled.

The parameters of the GMM-PMC algorithm have been selected as suggested in Cappé et al. (2004) ($p = 5$ scales, $\mathbf{v} = [5, 2, 0.1, 0.05, 0.01]^\top$, $m = 40$ samples per scale). A minimum of 1 % of samples per scale has been kept as a baseline. The GMM-PMC scheme with TIWs has been simulated simply substituting the standard IWs $w_\ell^{(i)}$ in the resampling step by TIWs $\bar{w}_\ell^{(i)}$ computed via a clipping transformation (with $M_T = 20$).



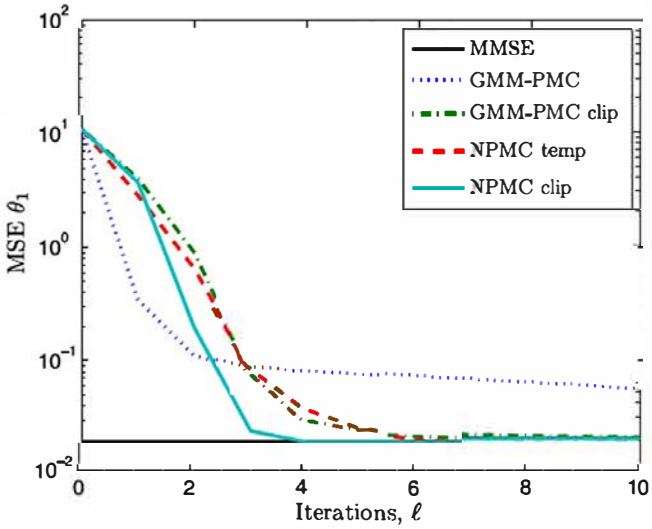


Fig. 2 Evolution along the iterations of the average NESS (*left*) and the average MSE of θ_1 (*right*) for the GMM-PMC, GMM-PMC with clipping, NPMC with tempering and NPMC with clipping in the GMM example of Sect. 6.1. The MMSE of θ_1 is also represented, for reference, as a *solid black line* in the plot on the *right*

In the NPMC algorithm with tempering, the sequence γ_ℓ has been obtained from the sigmoid function of the iteration index as $\gamma_\ell = \frac{1}{1+e^{-(\ell-5)}}$, $\ell = 1, \ldots, L$. With this choice of nonlinearity, the transformation of the weights is practically eliminated after 10 iterations.

The NPMC algorithm with clipping has been simulated in its modified version, i.e., with the nonlinear transformation removed when the ESS M^{eff} reaches a value of $M^{eff}_{min} = 100$. In this problem this occurs on average between the third and fourth iterations. On the contrary, in the GMM-PMC scheme with clipping, the ESS never reaches the threshold value and the nonlinear transformation thus cannot be removed. The clipping parameter has been set to $M_T = 20$ in both algorithms.

In Fig. 2 (*left*) the evolution of the average NESS $\bar{M}^{neff}_\ell$ along the iterations is depicted for the GMM-PMC and the $\bar{M}^{neff}_\ell$ for the rest of schemes. It can be observed that the original GMM-PMC scheme presents a low NESS, converging to a value of 0.13. The GMM-PMC with clipping outperforms the original scheme providing an average final NESS of 0.35. The two NPMC schemes, with tempering and clipping, provide a smooth convergence of the NESS to a value of 0.94.

The degeneracy problem is most critical at the first iterations of the PMC. The GMM-PMC scheme has an initial NESS value close to zero, opposite to the rest of schemes, where $\bar{M}^{neff}_1$ is around 0.1 (it is equal to M_T/M for the clipping schemes and depends on the parameter γ_1 for the tempering scheme). It can be observed from Fig. 2 (*left*) that in the NPMC schemes the average NESS remains constant after convergence, when the nonlinear transformation has been removed.

If we interpret the random vector $\bar{\boldsymbol{\theta}}_\ell$ with distribution $\bar{\pi}^M_\ell(d\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^M \delta_{\bar{\boldsymbol{\theta}}^{(i)}_\ell}(d\boldsymbol{\theta})$ (obtained after the resampling step of the ℓ-th iteration) as an estimator of $\boldsymbol{\theta}$, then the mean square error for the estimator of the k-th log-rate parameter is naturally given by

$$MSE_{\ell,k} = \frac{1}{M}\sum_{i=1}^{M}\left(\tilde{\theta}^{(i)}_{\ell,k} - \theta_k\right)^2 = \left(\hat{\theta}^M_{\ell,k} - \theta_k\right)^2 + \mathrm{Var}(\tilde{\theta}_{\ell,k}),$$

where $\hat{\theta}^M_{\ell,k} = \frac{1}{M}\sum_{i=1}^M \tilde{\theta}^{(i)}_{\ell,k}$ and $\mathrm{Var}(\tilde{\theta}_{\ell,k}) = \frac{1}{M}\sum_{i=1}^M (\tilde{\theta}^{(i)}_{\ell,k} - \hat{\theta}^M_{\ell,k})^2$ are the marginal mean and variance, respectively, of $\tilde{\theta}_{\ell,k}$ given the probability measure $\bar{\pi}^M_\ell$. We have averaged the $MSE_{\ell,k}$ over P independent simulations (with independent sets of observations).

In Fig. 2 (*right*) the evolution of the average MSE for θ_1 ($MSE_{\ell,1}$) is represented for the four algorithms. Similar results have been obtained for θ_2 and have thus been omitted. The minimum MSE (MMSE) of each parameter, which has been approximated numerically, is also shown for reference. It can be observed that the GMM-PMC does not reach the MMSE with the given number of samples $M = 200$. On the other side, the GMM-PMC with clipping and the proposed NPMC schemes outperform the original method in terms of MSE, reaching the MMSE in about 6 iterations.

However, the most outstanding difference in the performance of the analyzed algorithms is observed in the variance of the MSE. The final mean and standard deviation values of the MSE for θ_1 and θ_2 at $\ell = L$ are shown in Table 5. The estimates provided by the GMM-PMC scheme present a very high variance. On the contrary, the modified GMM-PMC and the proposed NPMC schemes reach the MMSE, both in average and in standard deviation.

Table 5 Mean and standard deviation (std) of the MSE of θ_1 and θ_2 at the last iteration $\ell = L$, for the analyzed PMC schemes. The MMSE (mean and std) corresponding to the true posterior $p(\boldsymbol{\theta}|\mathbf{y})$ is also shown for comparison. Note that all entries are multiplied by a factor of 10^3

	MSE θ_1		MSE θ_2	
	mean $\times 10^3$	std $\times 10^3$	mean $\times 10^3$	std $\times 10^3$
GMM-PMC	52.8	498.5	5.6	34.4
GMM-PMC clip	19.7	14.1	3.6	2.4
NPMC temp	19.1	13.8	3.3	2.4
NPMC clip	19.1	13.8	3.3	2.4
True posterior	19.1	13.7	3.2	2.3

Assuming that the computation time for the GMM-PMC method is 1, the GMM-PMC with clipping takes 1.0006 time units (that is, only is, 0.06 % higher) and the NPMC schemes take 0.9565 and 0.9582 time units for the tempering and clipping schemes, respectively. This indicates that the proposed method outperforms the original one also in terms of computational cost.

7 Example 2: A stochastic kinetic model

In this section, the proposed NPMC method is applied to the estimation of the parameters of a simple stochastic kinetic model (SKM), known as the predator-prey model. A SKM is a multivariate continuous-time jump process modeling the interactions among molecules, or species, that take place in chemical reaction networks of biochemical and cellular systems (Wilkinson 2011b).

Several MCMC schemes have been recently proposed to address this problem. In Boys et al. (2008) various MCMC algorithms are evaluated in data-poor scenarios. In Golightly and Wilkinson (2011) a likelihood-free particle-MCMC (pMCMC) scheme is applied to this problem. In Milner et al. (2013) the authors propose an approximation of the likelihood based on the moment closure approximation of the underlying stochastic process.

7.1 Predator-prey model

The Lotka-Volterra, or predator-prey, model is a simple SKM that describes the time evolution of two species $x_1(t)$ (prey) and $x_2(t)$ (predator), $t \in \mathbb{R}$, by means of $K = 3$ reaction equations (Volterra 1926)

$$
\begin{aligned}
x_1 &\xrightarrow{c_1} 2x_1 && \text{prey reproduction} \\
x_1 + x_2 &\xrightarrow{c_2} 2x_2 && \text{predator reproduction} \\
x_2 &\xrightarrow{c_3} \emptyset && \text{predator death}
\end{aligned}
$$

where $\mathbf{c} = [c_1, c_2, c_3]^\top$ is the vector of constant (yet random) rate parameters $c_k > 0$, $k = 1, 2, 3$.

Let $\mathbf{x}_n = [x_{1,n}, x_{2,n}]^\top$ denote the state of the system at time instant $t = n\Delta$, $n = 1, \ldots, R$, where $x_{1,n} = x_1(n\Delta)$, $x_{2,n} = x_2(n\Delta)$ denote the nonnegative, integer population of each species at this time instant and Δ denotes a time-discretization period. We denote by $\mathbf{x}$ the vector containing the population of each species at R discrete time instants, i.e., $\mathbf{x} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_R^\top]^\top$.

Exact stochastic simulation of generic SKMs, and predator-prey models in particular, can be carried out by the Gillespie algorithm (Gillespie 1977), which allows to draw samples from $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{c})$, $n = 1, \ldots, R$.

We consider two different observation scenarios. In the complete observation (CO) scenario we assume that both species x_1 and x_2 are observed at regular time intervals and corrupted by Gaussian noise, i.e., $\mathbf{y}_n = \mathbf{x}_n + \mathbf{u}_n$, where $\mathbf{u}_n \sim \mathcal{N}(\mathbf{u}_n; \mathbf{0}, \sigma^2\mathbf{I})$, $n = 1, \ldots, R$. We denote the complete vector of observations with dimension $2R \times 1$ as $\mathbf{y} = [\mathbf{y}_1^\top, \ldots, \mathbf{y}_R^\top]^\top$.

In the partial observation (PO) scenario only x_1 is observed at discrete time instants and also contaminated by Gaussian noise, i.e., $y_n = x_{1,n} + u_n$, where $u_n \sim \mathcal{N}(u_n; 0, \sigma^2)$, $n = 1, \ldots, R$. In the PO case, the vector of scalar observations with dimension $R \times 1$ is constructed as $\mathbf{y} = [y_1, \ldots, y_R]^\top$.

The goal is to approximate the posterior distribution of the log-rate parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^\top$ (where $\theta_k = \log(c_k)$, $k = 1, 2, 3$), with density $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, given the prior pdf $p(\boldsymbol{\theta})$ and the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, in the CO and PO scenarios.

7.2 NPMC algorithm for SKMs

In this particular problem, the observations $\mathbf{y}$ are related to the parameters $\boldsymbol{\theta}$ through the random vector $\mathbf{x}$. Indeed, the likelihood of $\boldsymbol{\theta}$ has the form

$$
p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x} = E_{p(\mathbf{x}|\boldsymbol{\theta})}\big[p(\mathbf{y}|\mathbf{x})\big],
$$

where $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{x})$, since the observations are independent of the parameter $\boldsymbol{\theta}$ given the population vector $\mathbf{x}$. In practice, the likelihood term $p(\mathbf{y}|\boldsymbol{\theta})$ cannot be evaluated exactly. A set of likelihood-free techniques have been recently proposed to tackle this kind of problems, which avoid the need to evaluate the likelihood function. In Golightly and Wilkinson (2011) a powerful pMCMC (Andrieu et al. 2010) method was proposed for the approximation of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ of the log-rate parameters in SKMs, which uses a particle filte (PF) to estimate the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ required to compute the acceptance ratio.[3]

[3]Note that this approximation becomes hard when the measurement noise variance is very small, as the weights of the PF may degenerate.

We propose to apply the NPMC method to the estimation of the rate parameters in SKMs. Similarly to Golightly and Wilkinson (2011), we resort to a PF to obtain an approximation of the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, required, in our case, to compute the TIWs. We provide details on this approximation in Appendix F.

7.3 Simulation setup

We consider the predator-prey model of Sect. 7.1. Following Golightly and Wilkinson (2011), the true vector of rate parameters which we aim to estimate has been set to $\mathbf{c} = [0.5, 0.0025, 0.3]^\top$, which yields $\boldsymbol{\theta} = [-0.69, -5.99, -1.20]^\top$. The initial populations and the number of observations have been set to $\mathbf{x}_0 = [100, 100]^\top$ and $R = 50$, respectively. The discretization period is $\Delta = 1$ and the noise variance is $\sigma^2 = 100$ (and assumed to be known). Independent uniform priors $\mathcal{U}(\theta_k; -7, 2)$ are taken for each $\theta_k = \log(c_k)$, and independent Poisson priors $p(x_{l,0}) = \mathcal{P}(x_{l,0}; \lambda_l)$ are considered for the initial populations $x_{l,0}$, with parameters set to the true values, that is, $\lambda_l = x_{l,0}$, $l = 1, 2$.

The number of particles of the PF used to compute the likelihood approximation $\hat{p}(\mathbf{y}|\boldsymbol{\theta}^{(i)})$ has been set to $J = 100$. Increasing J improves the performance only slightly, and at the expense of a significan increase of the computational cost (this is coherent with the results, e.g., in Wilkinson 2011b; Golightly and Wilkinson 2011, where the same value of J is selected).

Despite the low dimension of this problem ($K = 3$), the IWs of the PMC scheme present severe degeneracy, partly due to the likelihood approximation, which introduces additional variations to the IWs. Thus, the original PMC scheme without nonlinear transformations of the IWs does not work in this scenario. The NPMC scheme with tempering also performs poorly compared to the method with clipping. Given the extreme variations of the IWs, it is not straightforward to select a priori a tempering sequence γ_ℓ which provides a sufficien ESS at all iterations. For this reason, we have focused on the NPMC scheme with clipping, which computes TIWs at all iterations and guarantees a baseline ESS.

7.4 Results

We have performed $P = 100$ independent simulation runs of the NPMC with clipping in the CO and the PO scenarios, with the same initial populations $\mathbf{x}_0$ and different (independent) population and observation vectors. Both in the CO and the PO cases, the same true population trajectories were used, i.e., only the observations differ. The number of iterations has been set to $L = 10$, the number of samples per iteration is $M = 10^3$ and the clipping parameter is $M_T = 100$.

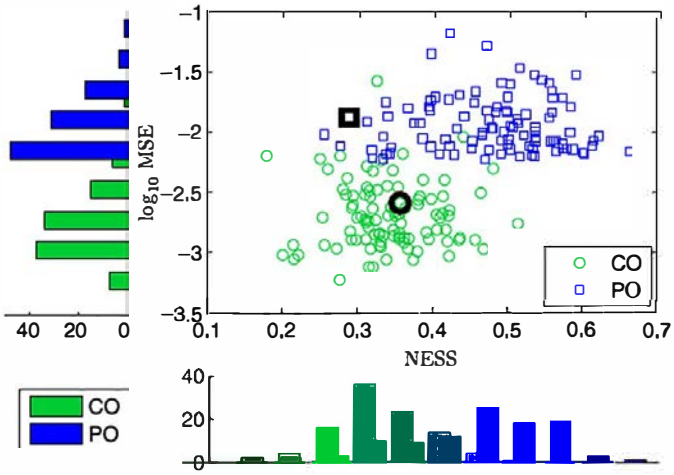
In the CO scenario, 5 simulation runs ended with a numerical error or with a fina NESS value close to M_T/M, and were repeated, for the same observation vectors, with $M = 2000$ and $M_T = 200$. Numerical errors may occur when very few samples $\boldsymbol{\theta}_\ell^{(i)}$ attain a significan likelihood, specially at the firs iteration. The NESS allows to detect whether the algorithm converges properly, when its value increases along the iterations beyond M_T/M. Thus, the average number of samples per iteration required in the CO case was $M = 1050$. On the contrary, in the PO case all the simulation runs ended satisfactorily with $M = 1000$.

In Fig. 3 (*left*) the fina values of the MSE ($MSE_{L,k}$) averaged over the parameters θ_k, $k = 1, 2, 3$, versus the fina NESS $\bar{M}_L^{neff}$ obtained at each simulation run are depicted, in the CO (green circles) and the PO (blue squares) scenarios, together with the histogram of each variable. It can be observed that in the CO scenario a lower MSE is attained compared to the PO scenario, given the larger amount of data available. However, the NESS is also lower in the firs case, which indicates more degeneracy of the IWs, again due to the larger amount of data. The required number of samples is larger in this case, being more computationally demanding and more sensitive to numerical issues. The big circle and square represent two particular simulation runs which attained a fina MSE close to the global average value in the CO and PO scenarios, respectively.

Figure 3 (*right*) depicts the fina estimate of the marginal posteriors $p(\theta_k|\mathbf{y})$ for the simulation runs represented as a big circle (CO) and square (PO) in Fig. 3 (*left*). We have built a Gaussian approximation of the marginal posteriors, namely $\hat{p}(\theta_k|\mathbf{y}) = \mathcal{N}(\theta_k; \mu_k, \sigma_k^2)$, where μ_k and σ_k are the k-th mean and standard deviation components of $\boldsymbol{\mu}_{L+1}$ and $\boldsymbol{\Sigma}_{L+1}$, computed as in Eqs. (3) and (4), respectively. It can be observed that the proposed algorithm successfully identifie the log-rate parameters both in the CO and the PO scenarios, and is robust to degeneracy problems that arise due to a large number of observations (specially in the CO case) and due to the approximation of the likelihood.

Table 6 shows the μ_k and σ_k parameters, $k = 1, 2, 3$, and the MSE, for the average simulation runs represented in Fig. 3 (*left*), and whose estimates $\hat{p}(\theta_k|\mathbf{y})$ are depicted in Fig. 3 (*right*), in both scenarios.

Figure 4 (*left*) shows the evolution of the average NESS in the CO (green lines) and PO (blue lines) case. Both the NESS computed with standard IWs (M_ℓ^{neff}) and TIWs ($\bar{M}_\ell^{neff}$) are represented, with dashed and solid lines, respectively. Both M_ℓ^{neff} and $\bar{M}_\ell^{neff}$ increase beyond the effect of the clipping procedure, which indicates that the algorithm is able to generate more representative samples as it converges. Figure 4 (*right*) shows the evolution of the average MSE in the CO and PO case. The value of the MSE at $\ell = 0$ corresponds to the MSE obtained from the prior pdf.



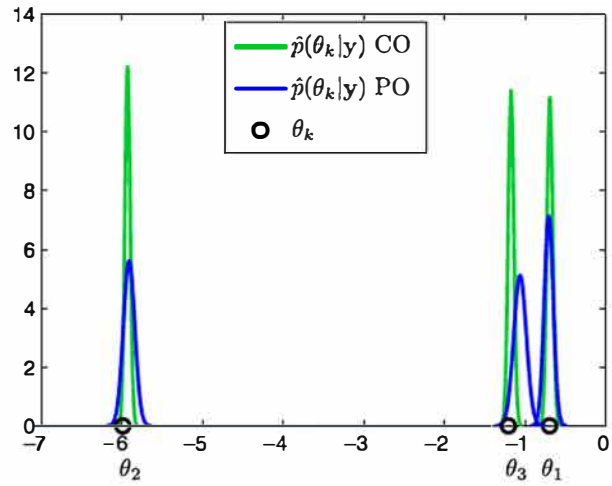
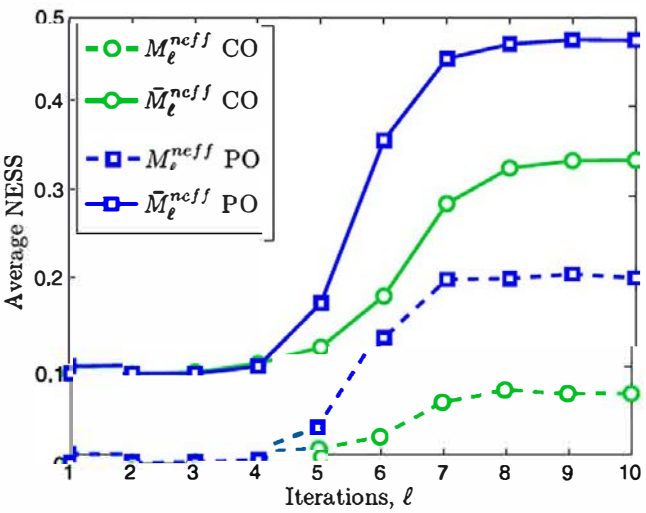


Fig. 3 *Left*: Final MSE in logarithmic scale versus the final NESS in the CO (*green circles*) and the PO (*blue squares*) scenario. Each point in the plots corresponds to an independent simulation run. The histograms of each variable are represented in the corresponding axis. The *big circle* and *square* represent two simulation runs with a final mean MSE close to the global average. *Right*: Marginal estimated posteriors $\hat{p}(\theta_k|\mathbf{y})$ and true values θ_k, $k = 1, 2, 3$, of the simulation runs represented as a *big circle* (CO) and *square* (PO) in the *left* plot, in the CO (*green line*) and PO (*blue line*) scenario (Color figure online)

Table 6 Parameters and MSE of the marginal posteriors $\hat{p}(\theta_k|\mathbf{y})$ for the average simulation run in the CO and PO experiment

	$\hat{p}(\theta_1\|\mathbf{y})$			$\hat{p}(\theta_2\|\mathbf{y})$			$\hat{p}(\theta_3\|\mathbf{y})$		
	μ_k	σ_k	MSE	μ_k	σ_k	MSE	μ_k	σ_k	MSE
CO	−0.690	0.036	1.29×10^{-3}	−5.932	0.033	4.62×10^{-3}	−1.173	0.035	2.19×10^{-3}
PO	−0.704	0.056	3.25×10^{-3}	−5.913	0.071	11.16×10^{-3}	−1.061	0.078	26.65×10^{-3}



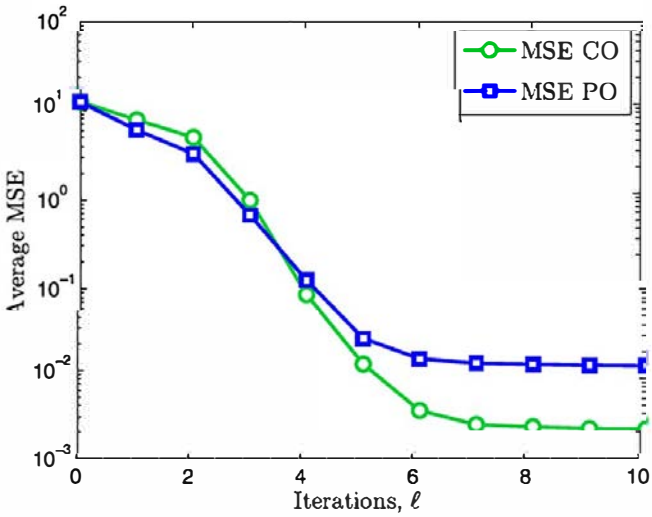


Fig. 4 Average NESS (*left*) and MSE (*right*) in the CO (*green lines with circles*) and PO (*blue lines with squares*) scenarios. In the *left* plot M_ℓ^{neff} (*dashed lines*) are computed from standard IWs and $\bar{M}_\ell^{neff}$ (*solid lines*) are computed from TIWs. The curves are averaged over $P = 100$ simulation runs. Both the NESS and the MSE converge simultaneously after 8 iterations (Color figure online)

It can be seen that the MSE smoothly decreases up to a low fina value, in just a few iterations.

The results presented here for the CO scenario can be compared, with some caution, to those obtained in Golightly and Wilkinson (2011) with a pMCMC scheme. The simulation setup is very similar, but the synthetic datasets employed here ($P = 100$ independent realizations of $\mathbf{y}$) and in Golightly and Wilkinson (2011) are different, as well as the prior describing the initial populations. Our simulations show that nearly equivalent results can been attained with the NPMC method, which involves a considerably lower computational cost. Note that the effort demanded to process one NPMC sample $\boldsymbol{\theta}_\ell^{(i)}$ is approximately equivalent to that of a single pMCMC iteration. In Golightly and Wilkinson (2011) 5×10^5 pMCMC iterations were run to compute solutions for this problem, while the NPMC scheme has only required 10^4 samples overall (taking into account *all* the iterations), reducing the computational cost by a factor of 50 for a similar performance.

8 Summary

We have addressed the problem of approximating posterior probability distributions by means of random samples. A recently proposed approach to tackle this problem is the population Monte Carlo method, that consists in iteratively approximating a target distribution via an IS scheme. The main limitation of this algorithm is that is presents severe degeneracy of the IWs as the dimension of the model, K, and/or the number of observations, N, increase. This leads to a highly varying number of effective samples and inaccurate estimates, unless the number of samples is extremely high (which makes the method computationally prohibitive).

We propose to apply a simple procedure in order to guarantee a prescribed ESS and a smooth and robust convergence. It consists in applying nonlinear transformations to the standard IWs in order to reduce their fluctuation and thus avoid degeneracy. It is straightforward to incorporate the new weight computation scheme into any existing method based on IS. It is possible, for example, to use TIWs within the SMC samplers of Del Moral et al. (2006), leading to a complete family of algorithms, of which the NPMC method introduced in the present paper would be just an instance.

In order to illustrate the application of the proposed technique, we have applied it to two examples of different complexity. The firs example is a simple GMM, which allows to get insight of the performance of the standard PMC scheme, the degeneracy problem and the behavior of the proposed algorithm. We have provided extensive simulation results that show how the proposed NPMC scheme can greatly improve the performance of the standard method.

Additionally, we have tackled the problem of estimating the set of constant (and random) rate parameters of a SKM. Even for the relatively simple predator-prey model that we have studied, this is significantl more complex than the GMM example. The NPMC method yields satisfactory results also in this scenario.

The Matlab code used to generate the presented simulation results is publicly available at http://www.tsc.uc3m.es/~jmiguez/npmc.zip.

The convergence of standard PMC algorithms is often justifie by the asymptotic convergence of IS (with respect to the number of samples). The NPMC scheme modifie the IWs and, hence, the standard theory of IS cannot be applied directly. To address this difficult , we have analyzed the convergence of the approximations of integrals computed using clipped TIWs and proved that they converge a.s., similar to the results available for standard IS. We have also quantifie the distortion introduced when using tempered TIWs.

Appendix A: Connection with SMC samplers

The description of the SMC sampling methodology is adapted from Del Moral et al. (2006). The goal is to approximate a sequence of probability distributions with densities $\pi_\ell(\boldsymbol{\theta})$, $\ell = 1, \ldots, L$. In order to be able to use a sequential importance sampling (SIS) algorithm for this purpose, let us defin the artificia joint target density

$$\alpha_\ell(\boldsymbol{\theta}_{1:\ell}) = \pi_\ell(\boldsymbol{\theta}_\ell) \prod_{r=1}^{\ell-1} b_r(\boldsymbol{\theta}_r|\boldsymbol{\theta}_{r+1}), \tag{16}$$

where $b_r(\boldsymbol{\theta}_r|\boldsymbol{\theta}_{r+1})$ is the density of an arbitrary *backward* kernel. By construction, the joint pdf in (16) has $\pi_\ell(\boldsymbol{\theta}_\ell)$ as a marginal, i.e.,

$$\int \cdots \int \pi_\ell(\boldsymbol{\theta}_\ell) \prod_{r=1}^{\ell-1} b_r(\boldsymbol{\theta}_r|\boldsymbol{\theta}_{r+1}) d\boldsymbol{\theta}_{\ell-1} \cdots d\boldsymbol{\theta}_1 = \pi_\ell(\boldsymbol{\theta}_\ell).$$

If we choose a sequence of *forward* kernels with densities $f_1(\boldsymbol{\theta}_1)$, $f_r(\boldsymbol{\theta}_r|\boldsymbol{\theta}_{r-1})$, $r = 2, \ldots, L$, it is possible to run a standard SIS algorithm (Doucet et al. 2000) to approximate the measure $\alpha_\ell(\boldsymbol{\theta}_{1:\ell})d\boldsymbol{\theta}_{1:\ell}$ (and its marginals). In particular, a simple algorithm would proceed as follows:

- *Initialization*: draw $\boldsymbol{\theta}_1^{(i)}$ from $f_1(\boldsymbol{\theta}_1)$, $i = 1, \ldots, M$, and set the initial (normalized) IWs as $w_1^{(i)} \propto \pi_1(\boldsymbol{\theta}_1^{(i)})/f_1(\boldsymbol{\theta}_1^{(i)})$. Resample to obtain an unweighted set $\{\tilde{\boldsymbol{\theta}}_1^{(i)}\}_{i=1}^M$.
- *Sequential step:* at the ℓ-th iteration,
  - draw $\boldsymbol{\theta}_\ell^{(i)}$ from $f_\ell(\boldsymbol{\theta}_\ell|\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)})$, $i = 1, \ldots, M$;
  - compute (unnormalized) weights

$$w_\ell^{(i)*} = \frac{\pi_\ell(\boldsymbol{\theta}_\ell^{(i)}) b_{\ell-1}(\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)}|\boldsymbol{\theta}_\ell^{(i)})}{\pi_{\ell-1}(\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)}) f_\ell(\boldsymbol{\theta}_\ell^{(i)}|\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)})}, \tag{17}$$

– and resample according to the normalized weights $w_\ell^{(i)} = \frac{w_\ell^{(i)*}}{\sum_{j=1}^M w_\ell^{(j)*}}$ to obtain the set $\{\tilde{\boldsymbol{\theta}}_\ell^{(i)}\}_{i=1}^M$.

In Del Moral et al. (2006) resampling is performed only when the ESS falls below some threshold, but this is not relevant for our discussion (hence we assume that resampling is performed at every sequential step). After the ℓ-th step, the measure $\pi_\ell^M(d\boldsymbol{\theta}_\ell) = \frac{1}{M}\sum_{i=1}^M \delta_{\tilde{\boldsymbol{\theta}}_\ell^{(i)}}(d\boldsymbol{\theta}_\ell)$ is an approximation of $\pi_\ell(\boldsymbol{\theta}_\ell)d\boldsymbol{\theta}_\ell$ (Del Moral et al. 2006).

We address the question of whether a NPMC algorithm with tempering can be obtained as a particular case of the SMC sampler above by a proper choice of the backward and forward kernels. The answer is partially positive. Indeed, consider the generic weight function in Eq. (17). If we select a sequence of exponents $0 < \gamma_1 < \gamma_2 < \cdots < \gamma_L = 1$ and defin $\pi_\ell(\boldsymbol{\theta}_\ell) = \pi(\boldsymbol{\theta}_\ell)^{\gamma_\ell}$, then we can equate

$$\mathrm{IW} \equiv \frac{\pi(\boldsymbol{\theta}_\ell)^{\gamma_\ell} b_{\ell-1}(\boldsymbol{\theta}_{\ell-1}|\boldsymbol{\theta}_\ell)}{\pi(\boldsymbol{\theta}_{\ell-1})^{\gamma_{\ell-1}} f_\ell(\boldsymbol{\theta}_\ell|\boldsymbol{\theta}_{\ell-1})} = \frac{\pi(\boldsymbol{\theta}_\ell)^{\gamma_\ell}}{f_\ell(\boldsymbol{\theta}_\ell|\boldsymbol{\theta}_{\ell-1})^{\gamma_\ell}} \equiv \mathrm{TIW},$$

and solve for the backward kernel density, namely

$$b_{\ell-1}(\boldsymbol{\theta}_{\ell-1}|\boldsymbol{\theta}_\ell) \propto \pi(\boldsymbol{\theta}_{\ell-1})^{\gamma_{\ell-1}} f_\ell(\boldsymbol{\theta}_\ell|\boldsymbol{\theta}_{\ell-1})^{1-\gamma_\ell}. \tag{18}$$

However, it is not possible to make Eq. (18) hold for *any* proposal scheme and, in particular, it cannot hold for the type of proposals introduced in Sect. 4.1. To be precise, the backward kernel density $b_\ell(\boldsymbol{\theta}_{\ell-1}|\boldsymbol{\theta}_\ell)$ can be chosen as in Eq. (18) if the i-th sample in the ℓ-th iteration is drawn conditional on i-th sample from the iteration $\ell - 1$. This is the usual case, e.g., in particle filterin applications where the variables of interest are dynamic and a forward kernel density is actually part of the model. Note that f_ℓ plays the role of the proposal density q_ℓ in Sect. 4. If $f_\ell(\boldsymbol{\theta}_\ell|\boldsymbol{\theta}_{\ell-1}) = f_\ell(\boldsymbol{\theta}_\ell) = q_\ell(\boldsymbol{\theta}_\ell)$ is designed simply from the statistics of the population $\{\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)}\}_{i=1}^M$, then the backward kernel becomes independent of the forward kernel, i.e.,

$$b_{\ell-1}(\boldsymbol{\theta}_{\ell-1}|\boldsymbol{\theta}_\ell) \propto \pi(\boldsymbol{\theta}_{\ell-1})^{\gamma_{\ell-1}} q_\ell(\boldsymbol{\theta}_\ell)^{1-\gamma_\ell} \propto \pi(\boldsymbol{\theta}_{\ell-1})^{\gamma_{\ell-1}}$$

and the weight function of the NPMC algorithm with tempering cannot be reproduced.

Very often, in the PMC framework q_ℓ is selected by matching the empirical moments of the population $\{\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)}\}_{i=1}^M$, and this is actually the case in Sect. 4, where $q_\ell(\boldsymbol{\theta}_\ell)$ is Gaussian with mean $\boldsymbol{\mu}_\ell = \frac{1}{M}\sum_{i=1}^M \tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)}$ and covariance matrix $\boldsymbol{\Sigma}_\ell = \frac{1}{M}\sum_{i=1}^M (\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)} - \boldsymbol{\mu}_\ell)(\tilde{\boldsymbol{\theta}}_{\ell-1}^{(i)} - \boldsymbol{\mu}_\ell)^\top$.

Appendix B: Proof of Lemma 1

As a firs step, we seek a tractable upper bound for the difference

$$\left|\left(f, \bar{\pi}^M\right) - \left(f, \check{\pi}^M\right)\right| = \left|\sum_{i=1}^M f\left(\boldsymbol{\theta}^{(i)}\right)\left(\bar{w}^{(i)} - \check{w}^{(i)}\right)\right|, \tag{19}$$

where

$$\begin{aligned} \bar{w}^{(i)} &= \frac{(\varphi^M \circ g)(\boldsymbol{\theta}^{(i)})}{\sum_{j=1}^M (\varphi^M \circ g)(\boldsymbol{\theta}^{(j)})}, \\ \check{w}^{(i)} &= \frac{(\varphi^M \circ g)(\boldsymbol{\theta}^{(i)})}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})} \end{aligned} \tag{20}$$

and $(\varphi^M \circ g)(\boldsymbol{\theta}) = \varphi^M(g(\boldsymbol{\theta}))$ denotes the composition of the functions φ^M and g. Moreover, the constants in the denominators of the weights can be written as integrals with respect to the random measure

$$q^M(d\boldsymbol{\theta}) = \frac{1}{M}\sum_{j=1}^M \delta_{\boldsymbol{\theta}^{(j)}}(d\boldsymbol{\theta}), \tag{21}$$

namely,

$$\sum_{j=1}^M \left(\varphi^M \circ g\right)\left(\boldsymbol{\theta}^{(j)}\right) = M\left(\varphi^M \circ g, q^M\right) \tag{22}$$

and

$$\sum_{j=1}^M g\left(\boldsymbol{\theta}^{(j)}\right) = M\left(g, q^M\right). \tag{23}$$

Substituting (20), (22) and (23), into (19) yields, after straightforward manipulations,

$$\begin{aligned} &\left|\left(f, \bar{\pi}^M\right) - \left(f, \check{\pi}^M\right)\right| \\ &\quad = \left|\frac{1}{M}\sum_{i=1}^M f\left(\boldsymbol{\theta}^{(i)}\right)\left(\varphi^M \circ g\right)\left(\boldsymbol{\theta}^{(i)}\right)\right. \\ &\qquad \left. \times \frac{(g, q^M) - (\varphi^M \circ g, q^M)}{(\varphi^M \circ g, q^M)(g, q^M)}\right|. \end{aligned} \tag{24}$$

A useful upper bound for the difference of integrals follows quite easily from (24). In particular, note that $|f(\boldsymbol{\theta}^{(i)})(\varphi^M \circ g)| \leq \|f\|_\infty \|g\|_\infty$, since $f, g \in B(\mathbb{R}^K)$ and $\varphi^M \circ g \leq g$, while the latter inequality also implies that $(\varphi^M \circ g, q^M)(g, q^M) \geq (\varphi^M \circ g, q^M)^2$. Also note that, from the definitio of φ^M,

$$\left|\left(g, q^M\right) - \left(\varphi^M \circ g, q^M\right)\right|$$

$$\leq \frac{1}{M}\sum_{k=1}^{M_T}\left|g\left(\boldsymbol{\theta}^{(i_k)}\right)-\left(\varphi^M\circ g\right)\left(\boldsymbol{\theta}^{(i_k)}\right)\right|$$
$$\leq \frac{2M_T\|g\|_\infty}{M}. \tag{25}$$

As a result, we obtain

$$\left|\left(f,\bar{\pi}^M\right)-\left(f,\breve{\pi}^M\right)\right|\leq \frac{2\|f\|_\infty\|g\|_\infty^2 M_T}{M(\varphi^M\circ g,q^M)^2}. \tag{26}$$

Let $c_1>0$ be some arbitrary real constant. From (26),

$$\mathbb{P}\left\{\left|\left(f,\bar{\pi}^M\right)-\left(f,\breve{\pi}^M\right)\right|\leq \frac{M_T}{M}c_1\right\}$$
$$\geq \mathbb{P}\left\{\frac{2\|f\|_\infty\|g\|_\infty^2 M_T}{M(\varphi^M\circ g,q^M)^2}\leq \frac{M_T}{M}c_1\right\}. \tag{27}$$

If we choose

$$c_1=\frac{2\|f\|_\infty\|g\|_\infty^2}{\left(\frac{1}{a}-\frac{1}{\sqrt{2}}\right)^2(g,q)^2}, \tag{28}$$

where $1<a<\sqrt{2}$ and $(g,q)\propto\int\pi(\boldsymbol{\theta})d\boldsymbol{\theta}>0$, then substituting (28) into the right hand side of (27) yields

$$\mathbb{P}\left\{\frac{2\|f\|_\infty\|g\|_\infty^2 M_T}{M(\varphi^M\circ g,q^M)^2}\leq \frac{M_T}{M}c_1\right\}$$
$$=\mathbb{P}\left\{\left(\varphi^M\circ g,q^M\right)^2\geq\left(\frac{1}{a}-\frac{1}{\sqrt{2}}\right)^2(g,q)^2\right\}$$
$$=\mathbb{P}\left\{\left(\varphi^M\circ g,q^M\right)-\frac{1}{a}(g,q)\geq-\frac{1}{\sqrt{2}}(g,q)\right\}$$
$$=\mathbb{P}\left\{M\left(\varphi^M\circ g,q^M\right)-\frac{M}{a}(g,q)\geq-\frac{M}{\sqrt{2}}(g,q)\right\}, \tag{29}$$

where the second equality holds because $\frac{1}{a}-\frac{1}{\sqrt{2}}>0$.

Next, consider the expectations $E_q[(g,q^M)]=(g,q)$ and $E_q[(\varphi^M\circ g,q^M)]$. Since, $|(\varphi^M\circ g,q^M)-(g,q^M)|\leq 2M_T\|g\|_\infty/M$ (see Eq. (25)), it follows that

$$\left|E_q\left[\left(\varphi^M\circ g,q^M\right)\right]-E_q\left[\left(g,q^M\right)\right]\right|$$
$$\leq E_q\left[\left|\left(\varphi^M\circ g,q^M\right)-\left(g,q^M\right)\right|\right]\leq\frac{2M_T\|g\|_\infty}{M}.$$

Therefore, since we have assumed that $\lim_{M\to\infty}\frac{M_T}{M}=0$, there exists M_a such that, for all $M>M_a$,

$$E_q\left[\left(\varphi^M\circ g,q^M\right)\right]>\frac{1}{a}E_q\left[\left(g,q^M\right)\right]=\frac{1}{a}(g,q), \tag{30}$$

and combining Eq. (29) with the inequality (30) we obtain that

$$\mathbb{P}\left\{\frac{2\|f\|_\infty\|g\|_\infty^2 M_T}{M(\varphi^M\circ g,q^M)^2}\leq \frac{M_T}{M}c_1\right\}$$
$$\geq\mathbb{P}\left\{M\left(\varphi^M\circ g,q^M\right)-ME_q\left[\left(\varphi^M\circ g,q^M\right)\right]\right.$$
$$\left.\geq-\frac{M}{\sqrt{2}}(g,q)\right\}. \tag{31}$$

Since $M(\varphi^M\circ g,q^M)=\sum_{i=1}^M\varphi^M(g(\boldsymbol{\theta}^{(i)}))$ is the sum of M independent and bounded random variables, each of them taking values within the interval $(0,\|g\|_\infty)$, it is straightforward to apply Hoeffding's tail inequality (Hoeffding 1963) (see also, e.g., Boucheron et al. 2004) to obtain a lower bound on (31), namely

$$\mathbb{P}\left\{M\left(\varphi^M\circ g,q^M\right)-ME_q\left[\left(\varphi^M\circ g,q^M\right)\right]\geq-\frac{M}{\sqrt{2}}(g,q)\right\}$$
$$\geq 1-\exp\left\{-\frac{(g,q)^2}{\|g\|_\infty^2}M\right\}. \tag{32}$$

Substituting (32) back into (31), (29) and (27) yields the desired result,

$$\mathbb{P}\left\{\left|\left(f,\bar{\pi}^M\right)-\left(f,\breve{\pi}^M\right)\right|\leq\frac{M_T}{M}c_1\right\}\geq 1-\exp\left\{-c_1'M\right\},$$

with $c_1'=\frac{(g,q)^2}{\|g\|_\infty^2}$. □

Appendix C: Proof of Lemma 2

The argument is similar to that of the proof of Lemma 1. Recalling Eqs. (20), (21) and (23) in Appendix B as well as the form of the standard normalized weights, $w^{(i)}=\frac{g(\boldsymbol{\theta}^{(i)})}{M(g,q^M)}$, it is straightforward to show that

$$\left|\left(f,\breve{\pi}^M\right)-\left(f,\pi^M\right)\right|$$
$$=\frac{1}{M(g,q^M)}\left|\sum_{k=1}^{M_T}f\left(\boldsymbol{\theta}^{(i_k)}\right)\left[\left(\varphi^M\circ g\right)\left(\boldsymbol{\theta}^{(i_k)}\right)-g\left(\boldsymbol{\theta}^{(i_k)}\right)\right]\right|,$$

which readily yields the upper bound

$$\left|\left(f,\breve{\pi}^M\right)-\left(f,\pi^M\right)\right|\leq\frac{2\|f\|_\infty\|g\|_\infty M_T}{M(g,q^M)}. \tag{33}$$

Let $c_2>0$ be some arbitrary real constant. From (33),

$$\mathbb{P}\left\{\left|\left(f,\breve{\pi}^M\right)-\left(f,\pi^M\right)\right|\leq\frac{M_T}{M}c_2\right\}$$
$$\geq\mathbb{P}\left\{\frac{2\|f\|_\infty\|g\|_\infty M_T}{M(g,q^M)}\leq\frac{M_T}{M}c_2\right\} \tag{34}$$

and if we choose $c_2=\frac{2\|f\|_\infty\|g\|_\infty}{(1-\frac{1}{\sqrt{2}})(g,q)}$, (recall that $(g,q)\propto\int\pi(\boldsymbol{\theta})d\boldsymbol{\theta}>0$) then

$$\mathbb{P}\left\{\frac{2\|f\|_\infty\|g\|_\infty M_T}{M(g,q^M)}\leq\frac{M_T}{M}c_2\right\}$$

$$= \mathbb{P}\left\{ \left(g, q^M\right) \geq \left(1 - \frac{1}{\sqrt{2}}\right)(g, q) \right\}$$
$$= \mathbb{P}\left\{ M\left(g, q^M\right) - M(g, q) \geq -\frac{M}{\sqrt{2}}(g, q) \right\}. \quad (35)$$

Since $(g, q) = E_q[(g, q^M)]$ and (g, q^M) is the sum of M independent, and bounded, random variables taking values within the interval $[0, \|g\|_\infty]$ (recall that $g \geq 0$), we can readily apply Hoeffding's tail inequality (Hoeffding 1963) on Eq. (35) to obtain

$$\mathbb{P}\left\{ M\left(g, q^M\right) - M(g, q) \geq -\frac{M}{\sqrt{2}}(g, q) \right\}$$
$$\geq 1 - \exp\left\{ -\frac{(g, q)^2}{\|g\|_\infty^2} M \right\}. \quad (36)$$

Substituting (36) back into (35) and (34) yields the desired result,

$$\mathbb{P}\left\{ \left|\left(f, \breve{\pi}^M\right) - \left(f, \pi^M\right)\right| \leq \frac{M_T}{M} c_2 \right\} \geq 1 - \exp\{-c_2' M\},$$

where $c_2' = \frac{(g,q)^2}{\|g\|_\infty^2} > 0$. □

Appendix D: Proof of Lemma 3

The firs part of Lemma 3 follows from the combination of Lemmas 1 and 2. We firs note that, from Lemma 1,

$$\mathbb{P}\left\{ \left|\left(f, \bar{\pi}^M\right) - \left(f, \breve{\pi}^M\right)\right| > c_1 \frac{M_T}{M} \right\} < \exp\{-c_1' M\} \quad (37)$$

for sufficientl large M, while Lemma 2 implies

$$\mathbb{P}\left\{ \left|\left(f, \breve{\pi}^M\right) - \left(f, \pi^M\right)\right| > c_2 \frac{M_T}{M} \right\} < \exp\{-c_2' M\}, \quad (38)$$

where $c_2 = c_2' = \frac{(g,q)^2}{\|g\|_\infty^2}$. Let $c = c_1 + c_2$. Then, since

$$\left|\left(f, \bar{\pi}^M\right) - \left(f, \pi^M\right)\right|$$
$$\leq \left|\left(f, \bar{\pi}^M\right) - \left(f, \breve{\pi}^M\right)\right| + \left|\left(f, \breve{\pi}^M\right) - \left(f, \pi^M\right)\right|,$$

we trivially obtain that

$$\mathbb{P}\left\{ \left|\left(f, \bar{\pi}^M\right) - \left(f, \pi^M\right)\right| > c \frac{M_T}{M} \right\}$$
$$\leq \mathbb{P}\left\{ \left|\left(f, \bar{\pi}^M\right) - \left(f, \breve{\pi}^M\right)\right| \right.$$
$$\left. + \left|\left(f, \breve{\pi}^M\right) - \left(f, \pi^M\right)\right| > c \frac{M_T}{M} \right\}. \quad (39)$$

However, if $|(f, \bar{\pi}^M) - (f, \breve{\pi}^M)| + |(f, \breve{\pi}^M) - (f, \pi^M)| > c\frac{M_T}{M}$ is true, then

$$\left|\left(f, \bar{\pi}^M\right) - \left(f, \breve{\pi}^M\right)\right| > c_1 \frac{M_T}{M} \quad \text{or}$$
$$\left|\left(f, \breve{\pi}^M\right) - \left(f, \pi^M\right)\right| > c_2 \frac{M_T}{M},$$

or both jointly, are true. Therefore,

$$\mathbb{P}\left\{ \left|\left(f, \bar{\pi}^M\right) - \left(f, \breve{\pi}^M\right)\right| + \left|\left(f, \breve{\pi}^M\right) - \left(f, \pi^M\right)\right| > c \frac{M_T}{M} \right\}$$
$$\leq \mathbb{P}\left\{ \left|\left(f, \bar{\pi}^M\right) - \left(f, \breve{\pi}^M\right)\right| > c_1 \frac{M_T}{M} \right\}$$
$$+ \mathbb{P}\left\{ \left|\left(f, \breve{\pi}^M\right) - \left(f, \pi^M\right)\right| > c_2 \frac{M_T}{M} \right\}$$
$$\leq \exp\{-c_1' M\} + \exp\{-c_2' M\}$$
$$= 2\exp\left\{ -\frac{(g, q)^2}{\|g\|_\infty^2} M \right\}, \quad (40)$$

for sufficientl large M, where the second inequality follows from (37) and (38), and the equality is due to the fact that $c_1' = c_2'$.

Combining (39) and (40) yields the firs part of Lemma 3, with $c' = \frac{(g,q)^2}{\|g\|_\infty^2}$.

The second part of Lemma 3 follows from a standard Borel-Cantelli argument. Indeed, let $\mathcal{E}_M$ be the event in which $|(f, \bar{\pi}^M) - (f, \pi^M)| > c\frac{M_T}{M}$ holds true. From the firs part of the Lemma,

$$\mathbb{P}\{\mathcal{E}_M\} < 2\exp\{-c' M\},$$

with $c' > 0$, for sufficientl large M (specificall , for all $M > M_a$, with M_a as in the proof of Lemma 1). Therefore,

$$\sum_{M=1}^{\infty} \mathbb{P}\{\mathcal{E}_M\} \leq M_a + \sum_{M=M_a+1}^{\infty} \exp\{-c' M\} < \infty,$$

because $M_a < \infty$ and $\sum_{M=M_a+1}^{\infty} \exp\{-c' M\} < \infty$. As a consequence (see, e.g., Williams 1991, Theorem 2.7), $\mathbb{P}\{\limsup \mathcal{E}_M\} = 0$, which implies that

$$\lim_{M\to\infty} \left|\left(f, \bar{\pi}^M\right) - \left(f, \pi^M\right)\right| = 0 \quad \text{a.s.}$$ □

Appendix E: Proof of Proposition 1

Let us introduce a new set of (unnormalized) bridge weights of the form

$$\breve{w}^{(i)} = \frac{g(\boldsymbol{\theta}^{(i)})^\gamma}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})}, \quad i = 1, \ldots, M, \quad (41)$$

and the corresponding (unnormalized) measure $\breve{\pi}^M(d\boldsymbol{\theta}) = \sum_{i=1}^M \breve{w}^{(i)}\delta_{\boldsymbol{\theta}^{(i)}}(d\boldsymbol{\theta})$. Using $\breve{\pi}^M$, the absolute difference $|(f,\pi^M) - (f,\bar{\pi}^M)|$ can be upper bounded by way of the triangular inequality

$$\begin{aligned}&\left|\left(f,\pi^M\right) - \left(f,\bar{\pi}^M\right)\right| \\ &\quad\le \left|\left(f,\pi^M\right) - \left(f,\breve{\pi}^M\right)\right| + \left|\left(f,\breve{\pi}^M\right) - \left(f,\bar{\pi}^M\right)\right|. \end{aligned} \tag{42}$$

In the sequel, we manipulate the two terms on the right hand side of (42) to show that (13) holds.

From the definitio of the bridge weights in (41), we obtain that

$$\begin{aligned}\left(f,\pi^M\right) - \left(f,\breve{\pi}^M\right) &= \sum_{i=1}^M f\left(\boldsymbol{\theta}^{(i)}\right)\frac{g(\boldsymbol{\theta}^{(i)}) - g(\boldsymbol{\theta}^{(i)})^\gamma}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})} \\ &= \sum_{i=1}^M f\left(\boldsymbol{\theta}^{(i)}\right)\frac{g(\boldsymbol{\theta}^{(i)})(1 - g(\boldsymbol{\theta}^{(i)})^{\gamma-1})}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})} \\ &= \left(f\left(1 - g^{\gamma-1}\right),\pi^M\right), \end{aligned} \tag{43}$$

where the last equality follows trivially if we consider the standard weight function $w^{(i)} = g(\boldsymbol{\theta}^{(i)})/\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})$.

As for the second term on the right hand side of (42), the definition of $\bar{w}^{(i)}$ and $\breve{w}^{(i)}$ in (12) and (41), respectively, yield

$$\begin{aligned}&\left(f,\breve{\pi}^M\right) - \left(f,\bar{\pi}^M\right) \\ &\quad= \sum_{i=1}^M f\left(\boldsymbol{\theta}^{(i)}\right)g\left(\boldsymbol{\theta}^{(i)}\right)^\gamma \\ &\qquad\times\left(\frac{1}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})} - \frac{1}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})^\gamma}\right). \end{aligned} \tag{44}$$

Some straightforward manipulations show that the difference of fractions above can be rewritten as

$$\begin{aligned}&\frac{1}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})} - \frac{1}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})^\gamma} \\ &\quad= \frac{\sum_{r=1}^M g(\boldsymbol{\theta}^{(r)})(g(\boldsymbol{\theta}^{(r)})^{\gamma-1} - 1)}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})^\gamma \sum_{k=1}^M g(\boldsymbol{\theta}^{(k)})} = \frac{(g^{\gamma-1} - 1,\pi^M)}{\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})^\gamma}, \end{aligned} \tag{45}$$

where we have used, again, the definitio of the standard weights $w^{(i)} = g(\boldsymbol{\theta}^{(i)})/\sum_{j=1}^M g(\boldsymbol{\theta}^{(j)})$. Substituting (45) into (44), and using the definitio of TIWs given by (12), yields

$$\begin{aligned}\left(f,\breve{\pi}^M\right) - \left(f,\bar{\pi}^M\right) &= \sum_{i=1}^M f\left(\boldsymbol{\theta}^{(i)}\right)\bar{w}^{(i)}\left(g^{\gamma-1} - 1,\pi^M\right) \\ &= \left(f,\bar{\pi}^M\right)\left(g^{\gamma-1} - 1,\pi^M\right). \end{aligned} \tag{46}$$

Finally, substituting (46) and (43) into (42) we arrive at

$$\begin{aligned}\left|\left(f,\pi^M\right) - \left(f,\bar{\pi}^M\right)\right| &\le \left|\left(f\left(1 - g^{\gamma-1}\right),\pi^M\right)\right| \\ &\quad + \left|\left(f,\bar{\pi}^M\right)\right|\left|\left(g^{\gamma-1} - 1,\pi^M\right)\right|, \end{aligned}$$

and the proof concludes by simply noting that $|(f,\bar{\pi}^M)| \le \|f\|_\infty$ and $|(g^{\gamma-1} - 1,\pi^M)| = |(1 - g^{\gamma-1},\pi^M)|$.

Appendix F: Estimating the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ via particle filterin

In this appendix we provide details on the approximation of the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$. In order to apply the algorithm below, one should recall that the dimension of the observation vector $\mathbf{y}$ varies from the CO scenario ($2R \times 1$) to the PO scenario ($R \times 1$), since in the latter case the observations are scalars, i.e., $\mathbf{y}_n = y_n \in \mathbb{R}$.

For a given vector of log-rate parameters $\boldsymbol{\theta}$, the following standard PF (see, e.g., Doucet et al. 2001) is run.

Initialization ($n = 0$):

Draw a collection of J samples $\{\mathbf{x}_0^{(j)}\}_{j=1}^J \sim p(\mathbf{x}_0)$.

Recursive step ($n = 1,\ldots,R$):

1. Draw $\{\mathbf{x}_n^{(j)}\}_{j=1}^J \sim p(\mathbf{x}_n|\mathbf{x}_{n-1}^{(j)},\boldsymbol{\theta})$ using the Gillespie algorithm.
2. Compute normalized IWs $\omega_n^{(j)*} = p(\mathbf{y}_n|\mathbf{x}_n^{(j)})$, $\omega_n^{(j)} = \omega_n^{(j)*}/\sum_{l=1}^J \omega_n^{(l)*}$, $j = 1,\ldots,J$.
3. Resample J times with replacement from $\{\mathbf{x}_n^{(j)}\}_{j=1}^J$ according to the weights $\{\omega_n^{(j)}\}_{j=1}^J$.

At every time step, the predictive pdf $p(\mathbf{y}_n|\mathbf{y}_{1:n-1},\boldsymbol{\theta})$ can be approximated as

$$\hat{p}(\mathbf{y}_n|\mathbf{y}_{1:n-1},\boldsymbol{\theta}) = \frac{1}{J}\sum_{j=1}^J p\left(\mathbf{y}_n|\mathbf{x}_n^{(j)}\right),$$

and the full likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ can be approximated in turn as

$$\hat{p}(\mathbf{y}|\boldsymbol{\theta}) = \prod_{n=1}^R \hat{p}(\mathbf{y}_n|\mathbf{y}_{1:n-1},\boldsymbol{\theta}).$$

See, e.g., (Maíz et al. 2012) for an analysis of the convergence of this approximation.

References

Andrieu, C., Doucet, A., Holenstein, R.: Particle Markov chain Monte Carlo methods. J. R. Stat. Soc., Ser. B, Stat. Methodol. **72**(3), 269–342 (2010)

Bain, A., Crisan, D.: Fundamentals of Stochastic Filtering vol. 60. Springer, Berlin (2008)
Bengtsson, T., Bickel, P., Li, B.: Curse of dimensionality revisited: collapse of particle f lter in very large scale systems. In: Probability and Statistics: Essay in Honour of David a Freedman, vol. 2, pp. 316–334 (2008)
Beskos, A., Crisan, D., Jasra, A.: On the stability of sequential Monte Carlo methods in high dimensions (2012). Arxiv preprint arXiv:1103.3965v2 [statCO]
Boucheron, S., Lugosi, G., Bousquet, O.: Concentration inequalities. In: Advanced Lectures on Machine Learning, pp. 208–240. Springer, Berlin (2004)
Boys, R.J., Wilkinson, D.J., Kirkwood, T.B.L.: Bayesian inference for a discretely observed stochastic kinetic model. Stat. Comput. **18**(2), 125–135 (2008)
Bugallo, M.F., Hong, M., Djuric, P.M.: Marginalized population Monte Carlo. In: ICASSP (2009)
Cappé, O., Guillin, A., Marin, J.M., Robert, C.P.: Population Monte Carlo. J. Comput. Graph. Stat. **13**(4), 907–929 (2004)
Cappé, O., Douc, R., Guillin, A., Marin, J.M., Robert, C.P.: Adaptive importance sampling in general mixture classes. Stat. Comput. **18**(4), 447–459 (2008)
Carpenter, J., Clifford, P., Fearnhead, P.: Improved particle filte for nonlinear problems. In: Radar, Sonar and Navigation. IEE Proceedings-, IET, vol. 146, pp. 2–7 (1999)
Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. R. Stat. Soc. B **68**, 411–436 (2006)
Djuric, P., Sven, B., Bugallo, M.: Population Monte Carlo methodology a la Gibbs sampling. In: EUSIPCO (2011)
Douc, R., Cappé, O., Moulines, E.: Comparison of resampling schemes for particle f ltering. In: Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, pp. 64–69 (2005)
Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte Carlo Sampling methods for Bayesian f ltering. Stat. Comput. **10**(3), 197–208 (2000)
Doucet, A., De Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer, Berlin (2001)
Geweke, J.: Bayesian inference in econometric models using Monte Carlo integration. Econometrica 1317–1339 (1989)
Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**(25), 2340–2361 (1977)
Golightly, A., Wilkinson, D.: Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. Interface Focus **1**(6), 807–820 (2011)
Gramacy, R., Samworth, R., King, R.: Importance tempering. Stat. Comput. **20**(1), 1–7 (2010)
Hoeffding, W.: Probability inequalities of sums of bounded random variables. J. Am. Stat. Assoc. **58**(301), 13–30 (1963)
Jasra, A., Stephens, D.A., Doucet, A., Tsagaris, T.: Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. Scand. J. Stat. **38**, 1–22 (2011)
Koblents, E., Míguez, J.: A population Monte Carlo method for Bayesian inference and its application to stochastic kinetic models. In: EUSIPCO (2011)
Kong, A., Liu, J.S., Wong, W.H.: Sequential imputations and Bayesian missing data problems. J. Am. Stat. Assoc. **9**, 278–288 (1994)
Maíz, C.S., Molanes-López, E., Míguez, J., Djurić, P.M.: A particle filterin scheme for processing time series corrupted by outliers. IEEE Trans. Signal Process. **9**(60) (2012)
Marinari, E., Parisi, G.: Simulated tempering: a new Monte Carlo scheme. Europhys. Lett. **19**(6), 451 (2007)
Milner, P., Gillespie, C.S., Wilkinson, D.J.: Moment closure based parameter inference of stochastic kinetic models. Stat. Comput. **23**(2), 287–295 (2013)
Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, Berlin (2004)
Shen, B., Bugallo, M., Djuric, P.: Multiple marginalized population Monte Carlo. In: EUSIPCO (2010)
Volterra, V.: Fluctuations in the abundance of a species considered mathematically. Nature **118**, 558–560 (1926)
Wilkinson, D.: Parameter inference for stochastic kinetic models of bacterial gene regulation: A Bayesian approach to systems biology. (with discussion), in Bayesian Statistics 9 (2011a)
Wilkinson, D.: Stochastic Modelling for Systems Biology vol. 44. CRC Press, Boca Raton (2011b)
Williams, D.: Probability with Martingales. Cambridge University Press, Cambridge (1991)