

Improving Population Monte Carlo: Alternative Weighting and Resampling Schemes

Víctor Elvira

Dep. of Signal Theory and Communic., Universidad Carlos III de Madrid, Leganés (Spain).

Luca Martino

Dep. of Mathematics and Statistics, University of Helsinki, Helsinki (Finland).

David Luengo

Dep. of Signal Theory and Communic., Universidad Politécnica de Madrid, Madrid (Spain).

Mónica F. Bugallo

Dep. of Electrical and Computer Eng., Stony Brook University, NY (USA).

Abstract

Population Monte Carlo (PMC) sampling methods are powerful tools for approximating distributions of static unknowns given a set of observations. These methods are iterative in nature: at each step they generate samples from a proposal distribution and assign them weights according to the importance sampling principle. Critical issues in applying PMC methods are the choice of the generating functions for the samples and the avoidance of the sample degeneracy. In this paper, we propose three new schemes that considerably improve the performance of the original PMC formulation by allowing for better exploration of the space of unknowns and by selecting more adequately the surviving samples. A theoretical analysis is performed, proving the superiority of the novel schemes in terms of variance of the associated estimators and preservation of the sample diversity. Furthermore, we show that they outperform other state of the art algorithms (both in terms of mean square error and robustness w.r.t. initialization) through extensive numerical simulations.

Key words: Population Monte Carlo, adaptive importance sampling, resampling.

1. Introduction

Bayesian signal processing, which has become very popular over the last years in statistical signal processing, requires computing distributions of unknowns conditioned on observations (and moments of them). Unfortunately, these distributions are often impossible to obtain analytically in many real-world challenging problems. An alternative is then to resort to Monte Carlo (MC) methods, which approximate the target distributions with random measures composed of samples and associated weights [25].

A well-known class of MC methods are those based on the adaptive importance sampling (AIS) mechanism, such as Population Monte Carlo (PMC) algorithms [19, 7], which have been used in missing data, tracking, and biological applications, among others [8, 5, 4, 2, 3]. In these methods, a population of probability density functions (pdfs) is adapted for approximating a target distribution through an iterative importance sampling procedure. AIS is often preferred to other MC schemes, such as Markov Chain Monte Carlo (MCMC), since they present several advantages. On the one hand, all the generated samples are employed in the estimation (e.g., no “burn-in” period must be checked). On the other hand, the corresponding adaptive schemes are more flexible, since they present less theoretical issues than adaptive MCMC algorithms.

The most characteristic feature in PMC [7] is arguably the use of resampling procedures for adapting the proposal pdfs (see for instance [21] for a review of resampling methods in particle filtering). The resampling step is a fast, often dimensionality-free, and easy way of adapting the proposal pdfs by using information about the target. However, resampling schemes present some important drawbacks, such as the sample impoverishment. At the resampling step, the proposal pdfs with poor performance (i.e., with low associated weights) are likely to be removed, thus yielding a reduction of diversity. Since the publica-

tion of the standard PMC [7], several variants have been considered, partly in an attempt to mitigate this issue. In the D-kernel algorithm [11, 12], the PMC kernel is a mixture of different kernels and the weights of the mixture are iteratively adapted in an implicit expectation-maximization (EM) algorithm. This procedure is refined through a double Rao-Blackwelization in [18]. The mixture population Monte Carlo algorithm (M-PMC) proposed in [6] also adapts a mixture of proposal pdfs (weights and parameters of the kernels). The M-PMC belongs to the family of AIS methods, since it iteratively draws the samples from the mixture that is updated at every iteration without an explicit resampling step. Since drawing from the mixture can be interpreted as an implicit multinomial resampling, this method retains some similarities with the standard PMC scheme. A nonlinear transformation of the importance weights in the PMC framework has also been proposed in [20]. Other sophisticated AIS schemes, such as the AMIS [9] and the APIS [23] algorithms, have been recently proposed in the literature.

In this paper, we study three novel PMC schemes that improve the performance of standard PMC approach by allowing a better exploration of the space of unknowns and by reducing the variance of the estimators. These alternatives can be applied within more sophisticated AIS approaches, as well. For this reason, we mainly compare them with the standard PMC [7], since the novel schemes could be automatically combined with the more sophisticated AIS techniques.

First of all, we introduce an alternative form of the importance weights, using a mixture of the proposal pdfs in the denominator of the weight ratio. We provide an exhaustive theoretical analysis, proving the unbiasedness and consistency of the resulting estimator, and showing the reduction in the variance of the estimator w.r.t. the estimator obtained using the standard weights. We also prove that the use of this mixture decreases the averaged mismatch between the numerator (target) and the function in the denominator of the IS weight in terms of L_2 distance. Moreover, we test this alternative scheme in different numerical simulations, including an illustrative toy example in Section

5.1, showing its practical benefit.

In the second proposed scheme, we generate several samples from every proposal pdf (not only one, as in PMC) and then we resample them jointly (all the samples at once, keeping fixed the total number of proposal pdfs). In the third proposed scheme, we consider again the generation of several samples from every proposal pdf, but the resampling is performed separately on the set of samples coming from each proposal, therefore guaranteeing that there will be exactly one representative from each of the individual mixings in the random measure.

We show, through extensive computer simulations in several different scenarios, that the three newly proposed variants provide a substantial improvement compared to the standard PMC. On the one hand, they yield unbiased estimators with a reduced variance, as also proved theoretically. On the other hand, they outperform the standard PMC in terms of preservation of sample diversity and robustness w.r.t initialization and parameter choice.

2. Problem Statement

Let us consider the variable of interest, $\mathbf{x} \in \mathbb{R}^{D_x}$, and let $\mathbf{y} \in \mathbb{R}^{D_y}$ be the observed data. In a Bayesian framework, the posterior probability density function (pdf), here referred as *target*, contains all the information about the parameters of interest and is defined as

$$\tilde{\pi}(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}{Z(\mathbf{y})} \propto \pi(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}), \quad (1)$$

where $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $g(\mathbf{x})$ is the prior pdf, and $Z(\mathbf{y})$ is the model evidence or partition function (useful in model selection).¹ The goal is to compute some moment of \mathbf{x} , i.e., an integral measure w.r.t. the target pdf,

$$I = \frac{1}{Z} \int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}, \quad (2)$$

¹From now on, we remove the dependence on \mathbf{y} in order to simplify the notation.

where $f(\cdot)$ can be any square integrable function of \mathbf{x} w.r.t. $\pi(\mathbf{x})$, and $Z = \int \pi(\mathbf{x}) d\mathbf{x}$.

In many practical applications, both the integral (2) and Z cannot be obtained in closed form and must be approximated. Importance sampling methods allow for the approximation of both quantities by a set of properly weighted samples.

3. Population Monte Carlo (PMC)

3.1. Description of the original PMC algorithm

The PMC method [7] is a well-known iterative adaptive importance sampling technique. At each iteration it generates a set of N samples $\{\mathbf{x}_i^{(t)}\}_{i=1}^N$, where t denotes the iteration number and i denotes the sample index. In order to obtain the samples, the original PMC algorithm makes use of a collection of proposal densities $\{q_i^{(t)}(\mathbf{x})\}_{i=1}^N$, with each sample being drawn from a different proposal, $\mathbf{x}_i^{(t)} \sim q_i^{(t)}(\mathbf{x})$ for $i = 1, \dots, N$. Then, they are assigned an importance weight, computed as $w_i^{(t)} = \frac{\pi(\mathbf{x}_i^{(t)})}{q_i^{(t)}(\mathbf{x}_i^{(t)})}$, i.e., the weight of a particular sample represents the ratio between the evaluation, at the sample value, of the target distribution and the evaluation at the sample value of the proposal used to generate it. The method proceeds iteratively (up to the maximum iteration step considered, T), building a global importance sampling estimator using different proposals at every iteration. The new proposals are obtained by updating the set of proposals in the previous iteration.

There are two key issues in the application of PMC methods: the adaptation of the proposals from iteration to iteration and the way resampling is applied. The latter is critical to avoid the degeneracy of the random measure, i.e., to avoid a few particles having extremely large weights and the rest negligible ones [25, 22]. Through the resampling procedure one selects the most promising streams of samples from the first iteration up to the current one. Several resampling procedures have been proposed in the literature [13, 10]. In the standard PMC [7], multinomial resampling is the method of choice, and consists of sampling N

Table 1: **Standard PMC algorithm** [7].

1. **[Initialization]**: Select the parameters defining the N proposals:

- The adaptive parameters $\mathcal{P}^{(1)} = \{\boldsymbol{\mu}_1^{(1)}, \dots, \boldsymbol{\mu}_N^{(1)}\}$.
- The set of static parameters, $\{\mathbf{C}_i\}_{i=1}^N$.

E.g., if the proposals were Gaussian distributions one could select the adapting parameters in $\mathcal{P}^{(1)}$ as the means of the proposals (that would be updated through the iterations) and the static parameters $\{\mathbf{C}_i\}_{i=1}^N$ as their covariances [7].

2. **[For $t = 1$ to T]**:

- (a) Draw one sample from each proposal pdf,

$$\mathbf{x}_i^{(t)} \sim q_i^{(t)}(\mathbf{x} | \boldsymbol{\mu}_i^{(t)}, \mathbf{C}_i), \quad i = 1, \dots, N. \quad (3)$$

- (b) Compute the importance weights,

$$w_i^{(t)} = \frac{\pi(\mathbf{x}_i^{(t)})}{q_i^{(t)}(\mathbf{x}_i^{(t)} | \boldsymbol{\mu}_i^{(t)}, \mathbf{C}_i)}, \quad i = 1, \dots, N, \quad (4)$$

and normalize them,

$$\bar{w}_i^{(t)} = \frac{w_i^{(t)}}{\sum_{j=1}^N w_j^{(t)}}, \quad i = 1, \dots, N. \quad (5)$$

- (c) Perform multinomial resampling by drawing N independent parameters $\boldsymbol{\mu}_i^{(t+1)}$ from the discrete probability random measure,

$$\hat{\pi}_t^N(\mathbf{x}) = \sum_{i=1}^N \bar{w}_i^{(t)} \delta(\mathbf{x} - \mathbf{x}_i^{(t)}). \quad (6)$$

The new set of adaptive parameters defining the next population of proposals becomes

$$\mathcal{P}^{(t+1)} = \{\boldsymbol{\mu}_1^{(t+1)}, \dots, \boldsymbol{\mu}_N^{(t+1)}\}. \quad (7)$$

3. **[Output, $t = T$]**: Return the pairs $\{\mathbf{x}_i^{(t)}, \bar{\rho}_i^{(t)}\}$, with $\bar{\rho}_i^{(t)}$ given by Eq. (8), for $i = 1, \dots, N$ and $t = 1, \dots, T$.

times from the discrete probability mass defined by the normalized weights. As a result of this procedure, the new set of parameters used to adapt the proposals for the generation of samples in the next iteration is selected. In summary, the standard PMC technique consists of the steps shown in Table 1.

3.2. Estimators and consistency

All the generated samples can be used to build a global approximation of the target. This can be done by first normalizing all the weights from all the iterations,

$$\bar{\rho}_i^{(t)} = \frac{w_i^{(t)}}{\sum_{\tau=1}^t \sum_{j=1}^N w_j^{(\tau)}}, \quad t = 1, \dots, T, \quad i = 1, \dots, N, \quad (8)$$

and then providing the pairs $\{\mathbf{x}_i^{(t)}, \bar{\rho}_i^{(t)}\}$ for $i = 1, \dots, N$ and $t = 1, \dots, T$. This procedure to compute the weights is equivalent to applying a static importance sampling technique that considers NT different proposal pdfs and all the corresponding samples. If the normalizing constant Z is known, the integral in Eq. (2) is approximated by the unbiased estimator

$$\hat{I}_t = \frac{1}{tN} \frac{1}{Z} \sum_{\tau=1}^t \sum_{j=1}^N w_j^{(\tau)} f(\mathbf{x}_j^{(\tau)}). \quad (9)$$

When the normalizing constant is unknown, the unbiased estimate of Z is substituted in Eq. (9), yielding the self-normalized estimator

$$\tilde{I}_t = \sum_{\tau=1}^t \sum_{j=1}^N \bar{\rho}_j^{(\tau)} f(\mathbf{x}_j^{(\tau)}) = \frac{1}{tN} \frac{1}{\hat{Z}_t} \sum_{\tau=1}^t \sum_{j=1}^N w_j^{(\tau)} f(\mathbf{x}_j^{(\tau)}), \quad (10)$$

where

$$\hat{Z}_t = \frac{1}{tN} \sum_{\tau=1}^t \sum_{j=1}^N w_j^{(\tau)}, \quad (11)$$

is the unbiased estimate of the normalizing constant.

4. Improved PMC schemes

In the following, we introduce several alternative strategies that decrease the variance of the estimators by exploiting the mixture perspective, and improve the diversity of the population w.r.t. to the standard PMC. More specifically, we study three different PMC schemes: one related to the strategy for calculating the weights and the other two based on modifying the way in which the resampling step is performed. Although we concentrate on the standard PMC, we remark that these alternative schemes can be directly applied or combined in other more sophisticated PMC algorithms.

4.1. Scheme 1:Deterministic mixture PMC (DM-PMC)

The underlying idea of PMC is to perform a good adaptation of the parameters $\boldsymbol{\mu}_i^{(t)}$. The location parameters are obtained at each iteration by sampling from $\hat{\pi}_{t-1}^N$ in Eq. (6) (i.e., via resampling), which is a random measure that approximates the target distribution, i.e., $\boldsymbol{\mu}_i^{(t)} \sim \hat{\pi}_{t-1}^N$. Note that the approximation $\hat{\pi}_{t-1}^N$ improves when N grows, namely, we have that $\boldsymbol{\mu}_i^{(t)} \sim \pi(\mathbf{x})$ for $N \rightarrow \infty$ [22]. Therefore, the equally-weighted mixture of the set of proposals at the t -th iteration, given by

$$\psi^{(t)}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N q_i^{(t)}(\mathbf{x} | \boldsymbol{\mu}_i^{(t)}, \mathbf{C}_i), \quad (12)$$

can be seen as a kernel density approximation of the target pdf, where the proposals, $\{q_i^{(t)}(\mathbf{x} | \boldsymbol{\mu}_i^{(t)}, \mathbf{C}_i)\}_{i=1}^N$, play the role of the kernels [26, Chapter 6]. Following kernel density estimation arguments, the resampling procedures naturally leads to a concentration of the proposals around the modes of the target.

Therefore, since the performance of an importance sampling method relies on the discrepancy between the numerator (the target) and the denominator (usually, the proposal pdf), a reasonable choice for calculating the importance weights is

$$w_i^{(t)} = \frac{\pi(\mathbf{x}_i^{(t)})}{\psi^{(t)}(\mathbf{x}_i^{(t)})} = \frac{\pi(\mathbf{x}_i^{(t)})}{\frac{1}{N} \sum_{j=1}^N q_j^{(t)}(\mathbf{x}_i^{(t)} | \boldsymbol{\mu}_j^{(t)}, \mathbf{C}_j)}, \quad (13)$$

where, as opposed to Eq. (4), the complete mixture of proposals $\psi(\mathbf{x})$ is accounted for in the denominator. The first justification for using these *deterministic mixture* (DM) weights is merely mathematical, since the estimator \hat{I}_t of Eq. (9) with these weights is also unbiased (see the proof in A.2). The main advantage of this new scheme is that it yields more efficient estimators, i.e. with less variance, combining the deterministic mixture sampling (as in standard PMC) with the weight calculation that accounts for the whole mixture. Namely, the estimator \hat{I}_t in Eq. (9), computed using the DM approach, has less variance than the estimator obtained by the standard PMC, as proved in A.3 for any target and set of proposal pdfs. The intuition behind the variance reduction

is clear in a multi-modal scenario, where different proposals have been successfully adapted covering the different modes, and therefore, the whole mixture of proposals has less mismatch w.r.t. the target than each proposal separately. Indeed, it can be easily proved that the mismatch of the whole mixture w.r.t to the target is always less than the average mismatch of each proposal. Namely, regarding the integrated squared distances, we need to prove that

$$\int (\tilde{\pi}(\mathbf{x}) - \psi(\mathbf{x}))^2 d\mathbf{x} \leq \frac{1}{N} \sum_{i=1}^N \int (\tilde{\pi}(\mathbf{x}) - q_i(\mathbf{x}))^2 d\mathbf{x}. \quad (14)$$

After some algebraic manipulations, Eq. 14 becomes

$$\int \left(\frac{1}{N} \sum_{i=1}^N q_i(\mathbf{x}) \right)^2 d\mathbf{x} \leq \int \frac{1}{N} \sum_{i=1}^N q_i^2(\mathbf{x}) d\mathbf{x}. \quad (15)$$

Finally, since the integrands of both integrals are non-negative, it is enough to prove that

$$\left(\frac{1}{N} \sum_{i=1}^N q_i(\mathbf{x}) \right)^2 \leq \frac{1}{N} \sum_{i=1}^N q_i^2(\mathbf{x}), \quad (16)$$

which is equivalent to Eq. (42) for $a_i = q_i(\mathbf{x})$, and is proved in Appendix A.3.

Another benefit of the DM-PMC scheme is the improvement in the exploratory behavior of the algorithm. Namely, since the weights in DM-PMC take into account all the proposals (i.e., the complete mixture) for their calculation, they temper the overrepresentation of high probability areas of the target. As a consequence, the effective sample size in DM-PMC is, in average, larger than in standard PMC, which implies that the diversity loss associated to the resampling step is reduced by using with the DM weights. See [15] for a more detailed discussion about effective sample size in static multiple importance sampling schemes.

In this DM-PMC scheme, the performance is improved at the expense of a slight increase in the computational cost (in terms of proposal evaluations) in the calculation of the weights. However, all the proposed schemes keep the same number of evaluations of the target as in the standard PMC. Hence, if the target evaluation is much more costly than the evaluation of the proposal pdfs

(as it often happens in practical applications), the increase in computational cost will be negligible in practice.

In the sequel, we adopt the weights of Eq. (13) for the other two proposed PMC schemes due to their theoretical and practical advantages discussed above.

4.2. Scheme 2: Multiple samples per mixand with global resampling (GR-PMC)

We propose to draw K samples per individual proposal or mixand, instead of only one as done in the standard PMC algorithm. Namely,

$$\mathbf{x}_{i,k}^{(t)} \sim q_i(\mathbf{x} | \boldsymbol{\mu}_i^{(t)}, \mathbf{C}_i) \quad (17)$$

for $i = 1, \dots, N$ and $k = 1, \dots, K$. Then, we compute the corresponding DM weights as in (13),

$$w_{i,k}^{(t)} = \frac{\pi(\mathbf{x}_{i,k}^{(t)})}{\frac{1}{N} \sum_{j=1}^N q_j^{(t)}(\mathbf{x}_{i,k}^{(t)} | \boldsymbol{\mu}_j^{(t)}, \mathbf{C}_j)}. \quad (18)$$

Therefore, at each iteration we have a set of KN generated samples, i.e., $\mathcal{X}^{(t)} = \{\mathbf{x}_{1,1}^{(t)}, \dots, \mathbf{x}_{1,K}^{(t)}, \dots, \mathbf{x}_{N,1}^{(t)}, \dots, \mathbf{x}_{N,K}^{(t)}\}$. Resampling is performed in the same way as in standard PMC, although now the objective is to downsample, from KN samples to N samples, according to the normalized weights,

$$\bar{w}_{i,k}^{(t)} = \frac{w_{i,k}^{(t)}}{\sum_{j=1}^N \sum_{\ell=1}^K w_{j,\ell}^{(t)}}. \quad (19)$$

We refer to this type of resampling as *global* resampling, since all the samples, regardless of the proposal used to generate them, are resampled together. After resampling, a new set of adapted parameters for the next iteration, $\mathcal{P}^{(t+1)} = \{\boldsymbol{\mu}_1^{(t+1)}, \dots, \boldsymbol{\mu}_N^{(t+1)}\}$, is obtained.

The standard PMC algorithm suffers from diversity impoverishment, as discussed previously. In multimodal scenarios, proposals that are exploring areas with negligible probability masses are very likely to be removed before they find unexplored relevant areas. By drawing K samples per proposal, the algorithm increases the local exploration of the proposals, thus increasing their chances of uncovering local relevant features of the target $\tilde{\pi}(\mathbf{x})$.

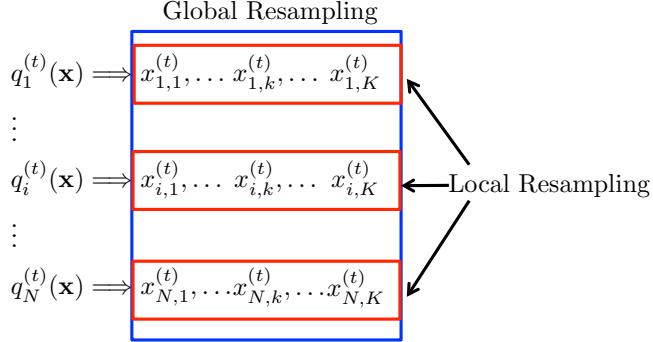


Figure 1: Sketch of the global and local resampling schemes considering N proposal pdfs at the t -th iteration, $q_i^{(t)}$ for $i = 1, \dots, N$ and $t = 1, \dots, T$, and K samples per proposal.

Unfortunately, one of the main issues of increasing K is the impoverishment of the global exploration: since the resampling is performed over the KN samples, most of the surviving samples are likely to come from a reduced set of proposals (even a single one in the worst case) when K is too large. Thus, there is a trade-off between local and global exploration. This suggests the existence of an optimal value of samples per proposal and iteration, K^* , which will depend on the target and cannot be found analytically. This issue can be partially addressed through the use of local resampling, as shown in the following section. Note that using $K > 1$ does not entail an increase in the computational cost w.r.t. the standard PMC or DM-PMC (where $K = 1$) if the number of evaluations of the target is fixed to $L = KNT$. Indeed, since the number of resampling stages is reduced to $T = L/(KN)$, the computational cost decreases, although at the expense of performing less adaptation steps than for $K = 1$.

4.3. Scheme 3: Multiple samples per mixand with local resampling (LR-PMC)

Consider again K samples generated from each proposal pdf. Unlike the previous method, here resampling is performed independently for each proposal. Namely, at the t -th iteration, K samples are drawn from each of the N proposal pdfs, and N parallel resampling procedures are independently performed within

each subset of K samples (see Fig. 1 for a visual comparison of both resampling schemes). More precisely, the adaptive parameter for the next iteration of the i -th proposal, $\mu_i^{(t+1)}$ for $i = 1, \dots, N$, is resampled from the set

$$\mathcal{X}_i^{(t)} = \{\mathbf{x}_{i,1}^{(t)}, \dots, \mathbf{x}_{i,K}^{(t)}\}, \quad (20)$$

using the multinomial probability mass function with probabilities

$$\bar{w}_{i,k}^{(t)} = \frac{w_{i,k}^{(t)}}{\sum_{\ell=1}^N w_{i,\ell}^{(t)}}, \quad k = 1, \dots, K. \quad (21)$$

In this variant, we consider again the use of the weights given by Eq. (18). This scheme guarantees that no loss of diversity occurs in the population of proposals, since each proposal at the current iteration leads to another proposal in the next iteration. Finally, let us remark that a mixed global-local resampling strategy (e.g., performing local resampling on clusters of proposals) could also be devised in order to obtain the advantages of both global and local resampling.

5. Numerical results

5.1. Estimation of the normalizing constant

Let us consider, as a target pdf, a bimodal mixture of Gaussians $\pi(\mathbf{x}) = \frac{1}{2}\mathcal{N}(x; \nu_1, c_1^2) + \frac{1}{2}\mathcal{N}(x; \nu_2, c_2^2)$ with $\nu_1 = -3$ and $\nu_2 = 3$, and $c_1^2 = 1$ and $c_2^2 = 1$. The proposal pdfs are also Gaussians: $q_1(x) = \mathcal{N}(x; \mu_1, \sigma_1^2)$ and $q_2(x) = \mathcal{N}(x; \mu_2, \sigma_2^2)$. At this point, we consider two scenarios:

- Scenario 1: In this case, $\mu_1 = \nu_1$, $\mu_2 = \nu_2$, $\sigma_1^2 = c_1^2$, and $\sigma_2^2 = c_2^2$. Then, both proposal pdfs can be seen as a whole mixture that exactly replicates the target, i.e., $\pi(\mathbf{x}) = \frac{1}{2}q_1(\mathbf{x}) + \frac{1}{2}q_2(\mathbf{x})$. This is the desired situation pursued by an adaptive importance sampling algorithm: each proposal is centered at a different mode of the target, and their scale parameters perfectly match the scales of the modes. Fig. 2(a) shows the target pdf in solid black line, and both proposal pdfs in blue and red dashed lines, respectively. Note that the proposals are scaled (each one integrates up

to $1/2$ so we can see the perfect matching between the target and the mixture of proposal densities).

- Scenario 2: In this case, $\mu_1 = -2.5$, $\mu_2 = 2.5$, $\sigma_1^2 = 1.2$, and $\sigma_2^2 = 1.2$. Therefore, there is a mismatch between the target and the two proposals. Fig. 3(a) shows the target pdf in solid black line, and both proposal pdfs in blue and red dashed lines, respectively.

The goal is estimating the normalizing constant using the estimator \hat{Z} of Eq. (11) with $N = 2$ samples, one from each proposal, and $t = 1$. We use the standard PMC weights of Eq. (4) (estimator \hat{Z}_{IS}) and the DM-PMC weights of Eq. (13) (estimator \hat{Z}_{DM}). In order to characterize the two estimators, we run $2 \cdot 10^5$ simulations for each method. Note that the true value is $Z = 1$.

Figure 2(b) shows a boxplot of the distribution of the estimator \hat{Z} , obtained with both methods for Scenario 1. The blue lower and upper edges of the box correspond to the 25th and 75th percentiles, respectively, while the red line represents the median. The vertical black dashed whiskers extend to the minimum and maximum obtained values. Since the maxima cannot be appreciated in the figure, they are displayed in Table 2, altogether with the variance of the estimators. Note that even in this extremely simple and idealized scenario (perfect adaptation), the estimator obtained using the standard IS weights (i.e., the estimator used in standard PMC) has a poor performance. In most of the realizations, $\hat{Z}_{IS} \approx 0.5$ because each proposal (which integrates up to one) is adapted to one of the two modes (which contain roughly half of the probability

	Estimator	\hat{Z}_{IS}	\hat{Z}_{DM}		Estimator	\hat{Z}_{IS}	\hat{Z}_{DM}
(Sc. 1)	Max.	35864	1	(Sc. 2)	Max.	77238	1.59
	$Var(\hat{Z})$	7891	0		$Var(\hat{Z})$	6874	0.01

Table 2: (**Ex. of Section. 5.1**) Maximum value of the estimator \hat{Z} in $2 \cdot 10^5$ runs for each MIS scheme, in two different scenarios.

mass).² Since $E[\hat{Z}_{IS}] = Z = 1$, in a few runs the value the \hat{Z}_{IS} is extremely high as shown in Table 2. These huge values occur when a sample drawn from the tail of the proposal falls close to the other mode of the target (where actually the other proposal is placed). On the other hand, note that the DM estimator has a perfect performance (i.e., $\hat{Z}_{DM} = 1$ always, thus implying zero variance). Hence, this simple example shows that a substantial variance reduction can be attained by using the mixture at the denominator.

Figure 3(b) shows an equivalent boxplot for Scenario 2. In this case, the mismatch between proposals and target pdfs worsens both schemes. Note that the estimator \hat{Z}_{DM} now does not perfectly approximates Z , but still largely outperform the estimator \hat{Z}_{IS} . In particular, the median is still around the true value, and its variance is smaller.

²In this setup, each proposal approximately covers a different half of the target probability mass, since each one coincides with a different mode of the target. However, in standard PMC, the weight of each sample only accounts for its own proposal, and therefore there is not an exchange of information among the two proposals. Note that, if both proposals were covering the same mode (and therefore missing the other one), the weights would also be $w = 0.5$ in most of the runs; the lack of information exchange between the two samples, makes it impossible to know whether the target mass reported by the weight of each sample is the same and should be accounted “once”, or whether it is from another area and it should be accounted “twice”.

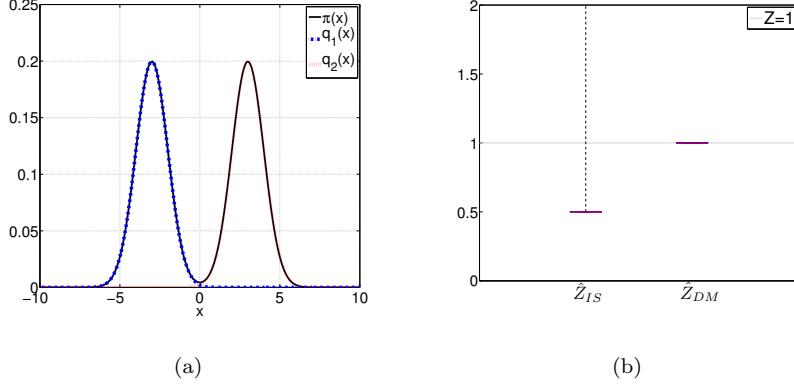


Figure 2: **(Ex. of Section 5.1)** Estimation of the normalizing constant (true value $Z = 1$) in Scenario 1 (perfect matching). **(a)** Target pdf (black solid line) and proposal pdfs (red and blue dashed lines). **(b)** Boxplot showing the 25th and 75th percentiles of the estimators \hat{Z}_{IS} and \hat{Z}_{DM} . The maximum value of \hat{Z}_{IS} is 35864.

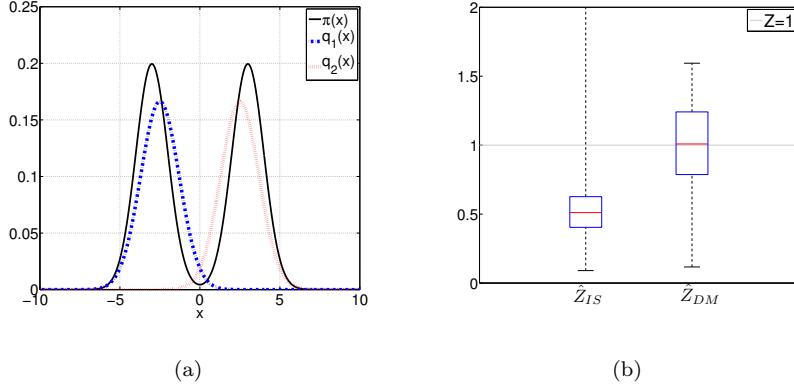


Figure 3: **(Ex. of Section. 5.1)** Estimation of the normalizing constant (true value $Z = 1$) in Scenario 2 (proposal-target mismatch). **(a)** Target pdf (black solid line) and proposal pdfs (red and blue dashed lines). **(b)** Boxplot showing the distribution of the estimators \hat{Z}_{IS} and \hat{Z}_{DM} . The maximum value of \hat{Z}_{IS} is 77238.

5.2. Bi-dimensional example

We first consider a bivariate multimodal target pdf, consisting of a mixture of five Gaussians, i.e.,

$$\pi(\mathbf{x}) = \frac{1}{5} \sum_{i=1}^5 \mathcal{N}(\mathbf{x}; \boldsymbol{\nu}_i, \boldsymbol{\Sigma}_i), \quad \mathbf{x} \in \mathbb{R}^2, \quad (22)$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C})$ denotes a normalized Gaussian pdf with mean vector $\boldsymbol{\mu}$ and covariance matrix \mathbf{C} , $\boldsymbol{\nu}_1 = [-10, -10]^\top$, $\boldsymbol{\nu}_2 = [0, 16]^\top$, $\boldsymbol{\nu}_3 = [13, 8]^\top$, $\boldsymbol{\nu}_4 = [-9, 7]^\top$, $\boldsymbol{\nu}_5 = [14, -14]^\top$, $\boldsymbol{\Sigma}_1 = [2, 0.6; 0.6, 1]$, $\boldsymbol{\Sigma}_2 = [2, -0.4; -0.4, 2]$, $\boldsymbol{\Sigma}_3 = [2, 0.8; 0.8, 2]$, $\boldsymbol{\Sigma}_4 = [3, 0; 0, 0.5]$, and $\boldsymbol{\Sigma}_5 = [2, -0.1; -0.1, 2]$. In this example, we can analytically compute different moments of the target in (22), and therefore we can easily validate the performance of the different techniques. In particular, we consider the computation of the mean of the target, $E[\mathbf{x}] = [1.6, 1.4]^\top$, and the normalizing constant, $Z = 1$, for $\mathbf{x} \sim \frac{1}{Z}\pi(\mathbf{x})$. We use as figure of merit the Mean Squared Error (MSE) in the estimation of $E[\mathbf{x}]$ (averaged over both components) and Z .

For simplicity, we assume Gaussian proposal densities for all of the methods compared, and deliberately choose a “bad” initialization of the means in order to test the robustness and the adaptation capabilities. Specifically, the initial adaptive parameters of the individual proposals are selected uniformly within the $[-4, 4] \times [-4, 4]$ square, i.e., $\boldsymbol{\mu}_i^{(1)} \sim \mathcal{U}([-4, 4] \times [-4, 4])$ for $i = 1, \dots, N$. This initialization is considered “bad”, since none of the modes of the target falls within the initialization square. We test all the alternatives using the same isotropic covariance matrices for all the Gaussian proposals, $\mathbf{C}_i = \sigma^2 \mathbf{I}_2$ with $\sigma \in \{1, 2, 5, 10, 20, 70\}$. All the results have been averaged over 500 independent experiments, where the computational cost of the different techniques (in terms of the total number of evaluations of the target distribution) is fixed to $L = KNT$.³ We compare the following schemes:

³Note that $L = KNT$ also corresponds to the total number of samples generated in all the schemes.

$L = NKT = 2 \cdot 10^5$							
N	Algorithm	$\sigma=1$	$\sigma=2$	$\sigma=5$	$\sigma=10$	$\sigma=20$	$\sigma=70$
5	Standard PMC [7]	92.8031	38.7137	12.6476	0.3832	0.0470	37.4372
100		75.1709	76.9161	14.2442	0.2475	0.0280	0.1803
$5 \cdot 10^4$		68.2938	37.4358	7.0110	0.2474	0.0329	0.1713
100	DM-PMC ($K = 1$)	72.4840	52.5093	5.3388	0.0362	0.0286	0.2126
	GR-PMC ($K = 2$)	69.4076	53.1546	3.0884	0.0225	0.0278	0.1721
	LR-PMC ($K = 2$)	2.6843	0.0076	0.0103	0.0178	0.1023	32.8828
	GR-PMC ($K = 5$)	67.0397	41.7514	0.1075	0.0132	0.0225	0.1471
	LR-PMC ($K = 5$)	8.0366	1.3940	0.0084	0.0158	0.0267	2.0035
	GR-PMC ($K = 20$)	61.5793	34.7170	0.4250	0.0122	0.0244	0.1425
	LR-PMC ($K = 20$)	9.5132	4.5452	0.0111	0.0130	0.0233	0.2158
	GR-PMC ($K = 100$)	64.9364	31.4322	0.0802	0.0147	0.0258	0.1801
	LR-PMC ($K = 100$)	9.6051	3.7193	0.0221	0.0150	0.0265	0.1960
	GR-PMC ($K = 500$)	58.4859	22.2482	0.0769	0.0135	0.0238	0.1639
	LR-PMC ($K = 500$)	15.1310	5.2872	0.0631	0.0103	0.0203	0.3262
100	M-PMC [6]	67.8610	76.3504	17.1312	0.0538	0.0291	0.1908

Table 3: **(Ex. of Section 5.2)** MSE in the estimation of $E[\mathbf{x}]$, for several values of σ and K , keeping the total number of evaluations of the target fixed to $L = KNT = 2 \cdot 10^5$ in all algorithms. The best results for each value of σ are highlighted in bold-face.

- **Standard PMC [7]:** Standard PMC algorithm described in Table 1 with $N = 100$ proposals and $T = 2000$ iterations. The total number of samples drawn is $L = NT = 2 \cdot 10^5$.
- **M-PMC [6]:** M-PMC algorithm proposed in [6] with $D = 100$ proposals, $N = 100$ samples per iteration, and $T = 2000$ iterations. The total number of samples drawn is $L = NT = 2 \cdot 10^5$.
- **K-PMC:** Standard PMC scheme using $N = 100$ proposals, but drawing $K > 1$ samples per proposal at each iteration and performing global resampling (GR). In order to keep the total number of samples constant, the number of iterations of the algorithm is now $T = 2 \cdot 10^5/(KN)$.
- **DM-PMC:** Standard PMC using the weights of Eq. (13), $N = 100$ proposals, $T = 2000$ iterations, and drawing $K = 1$ samples per proposal.
- **GR-PMC:** PMC scheme with multiple samples per mixand (K), weights computed as Eq. (18), and global resampling (GR). We use $N = 100$ proposals and $T = L/(KN)$ iterations with $L = 2 \cdot 10^5$ (as in the three previous schemes). In particular, we test the values $K \in \{2, 5, 20, 100, 500\}$, and thus $T \in \{1000, 400, 100, 20, 4\}$.
- **LR-PMC:** PMC scheme with multiple samples per mixand (K) and local resampling (LR). All the parameters are selected as in the GR-PMC scheme.

Table 3 shows the MSE in the estimation of $E[\mathbf{x}]$ (averaged over both components) for $\mathbf{x} \sim \frac{1}{Z}\pi(\mathbf{x})$. We can see that all the proposed schemes outperform the standard PMC for any value of σ . In general, the local resampling (LR) works better than the global resampling (GR). Moreover, we note that the optimum value of K depends on the value of σ , the scale parameter of the proposals: for small values of σ (e.g., $\sigma = 1$ or $\sigma = 2$) small values of K lead to better performance, whereas a larger value of K (and thus less iterations T) can be used for larger values of σ (e.g., $\sigma = 10$ or $\sigma = 20$). In addition, the proposed

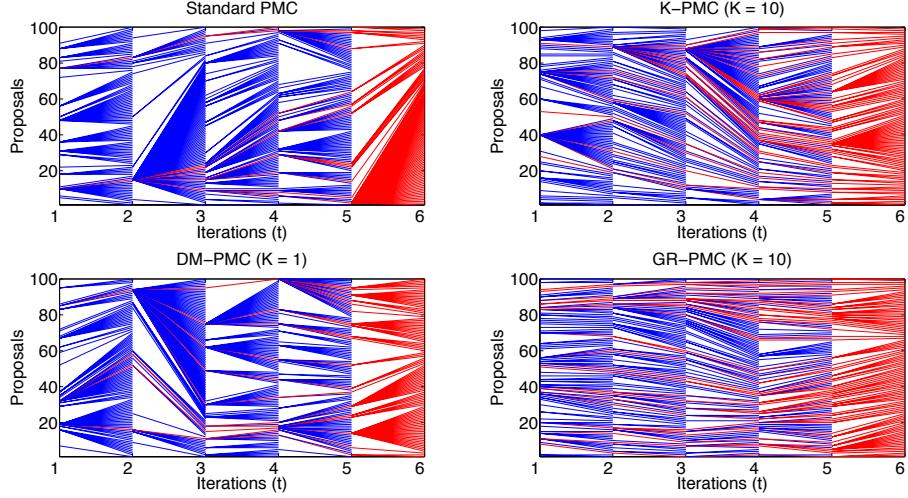


Figure 4: **(Ex. of Section 5.2)** Graphical representation of the indexes of the proposals used to generate the population for the next iteration with different schemes (6 iterations; $N = 100$, $\sigma = 5$). For each pair of iterations, lines link each surviving proposal (“father” proposal) with the next generation. In red, proposals surviving from the 1st to the 6th iteration.

methods also outperform the M-PMC algorithm in this scenario. Note that M-PMC is an adaptive importance sampling algorithm that does not perform the resampling step.

The large MSE values in Table 3 for some schemes and sets of parameters are due to the fact that they fail at discovering all the modes of the target pdf. In order to clarify this issue, Fig. 7 shows the evolution of the population of proposals for the first 4 iterations of the standard PMC ($K = 1$), K-PMC (with $K = 10$), and DM-PMC with global resampling (also for $K = 1$ and $K = 10$). Standard PMC tends to concentrate the whole population on one or two modes, very loosely covering the remaining ones and completely missing the mode in the bottom right corner. This issue is partly solved by using $K = 10$ (after 4 iterations the proposals are evenly distributed around 3 out of the 5 modes) or DM-PMC with $K = 1$ (after 4 iterations the proposals are uniformly distributed among 4 out of the 5 modes). Combining both approaches (DM-

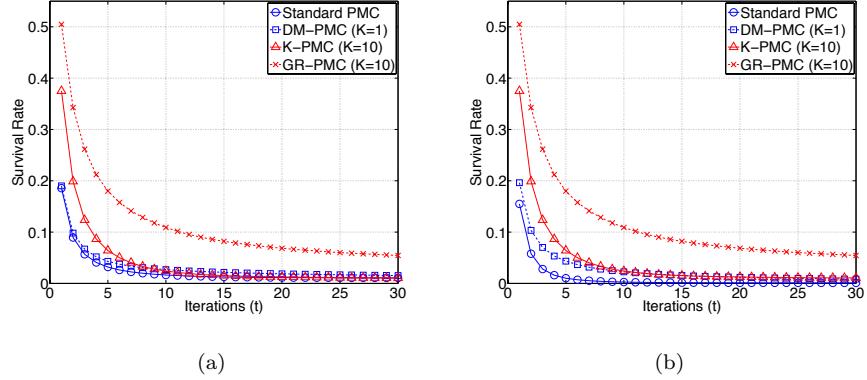


Figure 5: **(Ex. of Section 5.2)** Survival rate (after resampling) of the proposals vs. the distance in iterations among the proposals averaged over 500 runs ($\sigma = 5$). **(a)** $N = 100$ for all methods. **(b)** Different values of N , fixing $NK = 1000$ and thus $N \in \{100, 1000\}$.

Algorithm	$N = 100$			$N = 1000$		
	$\sigma = 1$	$\sigma = 5$	$\sigma = 20$	$\sigma = 1$	$\sigma = 5$	$\sigma = 20$
Standard PMC [7]	12.9178	6.5719	1.3601	8.6151	2.5508	0.4162
GR-PMC ($K = 2$)	14.1444	4.3571	0.5151	13.0727	0.0131	0.0842
LR-PMC ($K = 2$)	8.8212	14.5688	85.4180	3.3308	0.9261	44.1579
GR-PMC ($K = 10$)	12.956	0.8466	0.2469	3.6181	0.0197	0.1370
LR-PMC ($K = 10$)	8.5573	0.2307	2.2471	5.1043	0.0064	1.8587
GR-PMC ($K = 20$)	10.2277	0.9898	0.2655	3.1363	0.0195	0.2437
LR-PMC ($K = 20$)	6.0419	0.2238	0.7999	9.5565	0.0556	0.4444
GR-PMC ($K = 100$)	8.1532	0.2112	0.1521	7.6362	0.7139	0.1702
LR-PMC ($K = 100$)	6.9326	0.4560	0.2996	2.7386	0.1462	0.1745
M-PMC [6]	10.7753	9.0617	0.3460	8.3404	0.1246	0.0753

Table 4: **(Ex. of Section 5.3)** MSE in the estimation of $E[\mathbf{x}]$, for $\sigma \in \{1, 5, 20\}$ and $K \in \{2, 10, 20, 100\}$, keeping the total number of evaluations of the target fixed to $L = 2 \cdot 10^5$. The dimension space of the target is $D_x = 10$. The best results for each value of σ are highlighted in bold-face.

PMC and $K = 10$) an approximately uniform distribution of the proposals around all the modes of the target is attained.

Finally, in Figs. 4 and 5 we explore a well-known problem of PMC: the survival of proposals as the algorithm evolves. On the one hand, Fig. 4 shows which proposals have been used to generate the starting population for the next iteration. After 6 iterations, all of the $N = 100$ proposals in the population have arisen from only 2 of the proposals in the initial population. This situation hardly improves by using the DM-PMC: now 4 initial proposals have generated all the $N = 100$ proposals in the 6-th iteration. However, by drawing multiple samples per mixand ($K = 10$) the situation improves dramatically both when using the standard IS weights (9 proposals survive until the 6-th iteration) and especially when using the DM-PMC (19 surviving proposals). On the other hand, Fig. 5 shows the evolution in the survival rate of proposals w.r.t. the distance in iterations (or generations). In standard PMC, after very few iterations, most of the ancestors do not survive. This rate falls down as t increases in all cases, but the DM weights and especially the use of multiple samples per mixand help in slowing down this decrease. Therefore, we can conclude that the newly proposed schemes can be very useful in preserving the diversity in the population of proposals.

5.3. High-dimensional example

We consider a target corresponding to a mixture of isotropic Gaussians

$$\pi(\mathbf{x}) = \frac{1}{3} \sum_{k=1}^3 \mathcal{N}(\mathbf{x}; \boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k), \quad (23)$$

where $\mathbf{x} \in \mathbb{R}^{10}$, $\boldsymbol{\nu}_k = [\nu_{k,1}, \dots, \nu_{k,10}]^\top$, and $\boldsymbol{\Sigma}_k = \xi_k^2 \mathbf{I}_{10}$ for $k \in \{1, 2, 3\}$, with \mathbf{I}_{10} being the 10×10 identity matrix. We set $\nu_{1,j} = -5$, $\nu_{2,j} = 6$, and $\nu_{3,j} = 3$ for all $j \in \{1, \dots, 10\}$. Moreover, we set $\xi_k = 8$ for all $k \in \{1, 2, 3\}$. The expected value of the target $\pi(\mathbf{x})$ is $E[X_j] = \frac{4}{3}$ for $j = 1, \dots, 10$, and the normalizing constant is $Z = 1$.

We use Gaussian proposal densities for all the compared methods. The initial means (adaptive parameters of the proposals) are selected randomly and

independently in all techniques as $\mu_i^{(1)} \sim \mathcal{U}([-6 \times 6]^{10})$ for $i = 1, \dots, N$. We use the same isotropic covariance matrices for all the methods and proposal pdfs, $\mathbf{C}_i = \sigma^2 \mathbf{I}_{10}$, and we consider $\sigma \in \{1, 5, 20\}$. For every experiment, we run 200 independent simulations and compute the MSE in the estimation of $E[\mathbf{x}]$ (averaging the MSE of each component). We consider the same techniques as in the bi-dimensional example, testing $N \in \{100, 1000\}$ and different values of samples per iteration, $K \in \{2, 10, 20, 100\}$. We have tested different sets of parameters, always keeping the total number of samples fixed to $L = KNT = 2 \cdot 10^5$. Table 4 shows that the proposed PMC schemes outperform the standard PMC in most of the cases. Indeed, a decrease of more than one order of magnitude in the MSE can often be attained by using DM-PMC with an appropriate value of K instead of the standard PMC. Finally, note that, although M-PMC behaves well for most of the parameters tested, overall the proposed methods yield the best performance in terms of MSE and robustness w.r.t. parameter choice.

In order to study the performance of the proposed schemes as the dimension of the state space increases, we change the dimension of the state space in (23). Namely, the target density is still a mixture of three isotropic Gaussians with the same structure for the mean vectors and covariance matrices as before, but now the dimension of \mathbf{x} is $D_x \in [1, 50]$. We have tested all the methods with $\sigma = 5$ and $N = 100$. Fig. 6 shows the evolution of the MSE in the estimation of the normalizing constant as a function of D_x . As expected, the performance of all the methods degrades as the dimension of the problem, D_x , becomes larger. Nonetheless, the performance of the proposed methods decays much more slowly than that of the standard PMC, thus allowing them to still provide a reasonably low MSE in higher dimensions. Note that, since the true normalizing constant of the target is $Z = 1$, when the methods behave poorly in high dimensions and the proposals do not discover the modes, the estimation is $\hat{Z} \approx 0$, and therefore the MSE tends to 1, which is the worst-case situation.

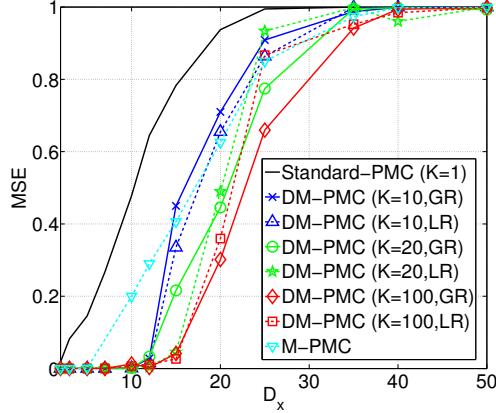


Figure 6: **(Ex. of Section 5.3)** MSE of the normalizing constant Z , using $N = 100$ proposals and a scale parameter $\sigma = 5$, as the dimension of the state space D_x increases.

5.4. Autoregressive filter with non-Gaussian noise

We consider the use of an autoregressive (AR) model contaminated by a non-Gaussian noise. This kind of filters is often used for modeling some financial time series, where the noise is assumed to follow the so-called *generalized hyperbolic distribution* [14]. Namely, we consider the following observation model,

$$y_m = x_1 y_{m-1} + x_2 y_{m-2} + x_3 y_{m-3} + x_4 y_{m-4} + u_m, \quad (24)$$

where $m = 1, \dots, M$ is a time index, and u_m is a heavy-tailed driving noise:

$$u_m \sim p(u) \propto e^{\beta(u-\mu)} \frac{B_{\lambda-\frac{1}{2}} \left(\alpha \sqrt{\delta^2 + (u-\mu)^2} \right)}{\left(\sqrt{\delta^2 + (u-\mu)^2} \right)^{\frac{1}{2}-\lambda}},$$

where B_λ denotes the *modified Bessel function* [1]. The vector of unknowns, $\mathbf{x}^* = [x_1^*, x_2^*, x_3^*, x_4^*]^\top$, contains the coefficients of the AR model.

Given a set of observations $\mathbf{y} = [y_1, \dots, y_M]^\top$, the inference problem consists of obtaining statistical information about \mathbf{x}^* , by studying the corresponding posterior distribution $\tilde{\pi}(\mathbf{x}|\mathbf{y})$. More specifically, we have synthetically generated $M = 200$ observations, $\mathbf{y} = [y_1, \dots, y_M]^\top$, setting $\mathbf{x}^* = [0.5, 0.1, -0.8, 0.1]^\top$,

$\lambda = 0.5$, $\alpha = 2$, $\beta = 1$, $\mu = -1$, and $\delta = 1$.⁴ Assuming improper uniform priors over the unknown coefficients, the objective is computing the expected value $\hat{\mathbf{x}} = \int_{\mathbb{R}^4} \mathbf{x} \tilde{\pi}(\mathbf{x}|\mathbf{y}) d\mathbf{x}$. We note that using $M = 200$ observations, the posterior pdf is quite sharp and concentrated around the true value, $\mathbf{x}^* = [0.5, 0.1, -0.8, 0.1]^\top$.

All the methods use Gaussian proposals, with the initial adaptive parameters of the individual proposals selected uniformly within the $[-6, 6]^4$ square, i.e., $\boldsymbol{\mu}_i^{(1)} \sim \mathcal{U}([-6, 6] \times [-6, 6] \times [-6, 6] \times [-6, 6])$, and the covariance matrices for all the Gaussians selected as $\mathbf{C}_i = \sigma^2 \mathbf{I}_4$, with $\sigma = 5$ for $i = 1, \dots, N$. As in the previous examples, we have tested different combinations of parameters, keeping the total number of evaluations of the target fixed to $L = NKT = 2 \cdot 10^5$. We have evaluated different values of $N \in \{100, 1000, 5000\}$ and $K \in \{5, 10, 50, 100\}$. We ran 500 independent simulations and computed the MSE in the estimation of $\hat{\mathbf{x}}$ w.r.t. the true value \mathbf{x}^* .

The results obtained by the different methods, in terms of MSE averaged over all the components of \mathbf{x} , are shown in Table 5. Note that some combinations of K and N would yield a number of iterations $T < 1$, since we set $T = L/(NK) = 2 \cdot 10^5/(NK)$. Therefore, those simulations cannot be performed and are indicated in the Table with the symbol *. Note that, for any choice of N , the alternative schemes proposed in the paper largely outperform the standard PMC. Furthermore, the advantage of using $K > 1$ can again be clearly seen for the three values of N tested. More specifically, the smallest the value of N the largest the value of K that should be used to attain the best results. Note also that M-PMC behaves particularly well in this scenario for high values of N , but its performance is very poor for $N = 100$ (unlike GR-PMC and LR-PMC, which can still provide a good performance for the right value of K).

⁴For the generation of i.i.d. samples of the generalized hyperbolic noise, we applied a fast and efficient MCMC technique (the FUSS algorithm [24]), drawing samples from univariate distributions. After a few iterations, the resulting samples were virtually independent.

Algorithm	N= 100	N= 1000	N= 5000
Standard PMC [7]	13615.9497	69.9915	0.5553
GR-PMC ($K = 5$)	1597.5728	1.9171	0.0791
LR-PMC ($K = 5$)	31.0446	0.3571	0.2013
GR-PMC ($K = 10$)	520.6153	0.3037	0.0715
LR-PMC ($K = 10$)	14.9870	0.2897	0.2125
GR-PMC ($K = 50$)	16.9080	0.0553	*
LR-PMC ($K = 50$)	2.3072	0.1661	*
GR-PMC ($K = 100$)	2.2322	0.0969	*
LR-PMC ($K = 100$)	0.7727	0.1674	*
M-PMC [6]	182.10	0.0635	0.0464

Table 5: **(Ex. of Section 5.4)** MSE of $E[\mathbf{x}]$ for different values of K and N , keeping the total number of evaluations of the target fixed to $L = KNT = 2 \cdot 10^5$. The symbol * indicates combinations where the number of iterations $T < 1$, and therefore they cannot be performed.

6. Conclusions

The population Monte Carlo (PMC) method is a well-known and widely used scheme for performing statistical inference in many signal processing problems. Three improved PMC algorithms are proposed in this paper. All of them are based on the deterministic mixture (DM) approach, which provides estimators with a reduced variance (as proved in this paper) and increases the exploratory behavior of the resulting algorithms. Additionally, two of the methods draw multiple samples per mixand (both with local and global resampling strategies) to prevent the loss of diversity in the population of proposals. The proposed approaches are shown to substantially outperform the standard PMC on three numerical examples. The proposed improvements can be applied to other existing PMC implementations and other importance sampling techniques, to achieve similar benefits.

Acknowledgment

This work has been supported by the Spanish government's projects AGES (S2010/BMD-2422), ALCIT (TEC2012-38800-C03-01), COMPREHENSION (TEC2012-38883-C02-01), DISSECT (TEC2012-38058-C03-01), and OTOSiS (TEC2013-41718-R); by the BBVA Foundation through project MG-FIAR ("I Convocatoria de Ayudas Fundación BBVA a Investigadores, Innovadores y Creadores Culturales"); by ERC grant 239784 and AoF grant 251170; by the National Science Foundation under Award CCF-0953316; and by the European Union's 7th FP through the Marie Curie ITN MLPM2012 (Grant No. 316861).

References

References

- [1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Number 55. Dover Pub., 1972.
- [2] G. E. Barter, L. K. Purvis, N. P. Teclemariam, and T. H. West. Analysis of detection systems for outdoor chemical or biological attacks. In *Proc. IEEE Conf. on Technologies for Homeland Security*, 2009.
- [3] H. Bhaskar, L. Mihaylova, and S. Maskell. Population based particle filtering. In *IET Seminar on Target Tracking and Data Fusion: Algorithms and Applications*, 2008.
- [4] C. Bi. A Monte Carlo EM algorithm for de novo motif discovery in biomolecular sequences. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:370–386, 2009.
- [5] M. C. A. M. Bink, M. P. Boer, C. J. F. Braak, J. Jansen, R. E. Voorrips, and W. E. van de Weg. Bayesian analysis of complex traits in pedigree plant populations. *Euphytica*, 161:85–96, 2008.

- [6] O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.
- [7] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.
- [8] G. Celeux, J. M. Marin, and C. P. Robert. Iterated importance sampling in missing data problems. *Computational Statistics & Data Analysis*, 50:3386–3404, 2006.
- [9] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, December 2012.
- [10] P. Del Moral, A. Doucet, and A. Jasra. On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, 18(1):252–278, 2012.
- [11] G. R. Douc, J. M. Marin, and C. P. Robert. Convergence of adaptive mixtures of importance sampling schemes. *Annals of Statistics*, 35:420–448, 2007.
- [12] G. R. Douc, J. M. Marin, and C. P. Robert. Minimum variance importance sampling via population Monte Carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.
- [13] R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Proc. 4th Int. Symp. on Image and Signal Processing and Analysis*, pages 64–69, September 2005.
- [14] E. Eberlein. Application of generalized hyperbolic lévy motions to finance. In *Lévy processes*, pages 319–336. Springer, 2001.
- [15] Víctor Elvira, Luca Martino, David Luengo, and Mónica F Bugallo. Generalized multiple importance sampling. *arXiv preprint arXiv:1511.03095*, 2015.

- [16] J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 24:1317–1399, 1989.
- [17] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Univ. Press, 1952.
- [18] Alessandra Iacobucci, Jean-Michel Marin, and Christian Robert. On variance stabilisation in population Monte Carlo by double Rao-Blackwellisation. *Computational Statistics & Data Analysis*, 54(3):698–710, 2010.
- [19] Y. Iba. Population Monte Carlo algorithms. *Transactions of the Japanese Society for Artificial Intelligence*, 16:279–286, 2001.
- [20] Eugenia Koblents and Joaquín Míguez. A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing*, pages 1–19, 2013.
- [21] Tiancheng Li, Miodrag Bolic, and Petar M Djuric. Resampling methods for particle filtering: Classification, implementation, and strategies. *IEEE Signal Processing Magazine*, 32(3):70–86, 2015.
- [22] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.
- [23] L. Martino, V. Elvira, D. Luengo, and J. Corander. An adaptive population importance sampler: Learning from the uncertainty. *IEEE Transactions on Signal Processing*, 63(16):4422–4437, 2015.
- [24] L Martino, H Yang, D Luengo, J Kannainen, and J Corander. A fast universal self-tuned sampler within Gibbs sampling. *Digital Signal Processing, In press, 2015*, 2015.
- [25] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.
- [26] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2009.

A. Standard vs. deterministic mixture importance sampling

In this appendix we review the IS estimators, analyzing the properties (unbiasedness and variance) of the estimator with the DM weights. For the sake of clarity, we remove the temporal indexes.

A.1. Importance sampling estimators

Let us consider the estimator of Eq. (9) when we have a set of N proposal pdfs, $\{q_i(\mathbf{x})\}_{i=1}^N$. We draw exactly $K_i = 1$ sample from each proposal, i.e., $\mathbf{x}_i \sim q_i(\mathbf{x})$ for $i = 1, \dots, N$.⁵ If the normalizing constant Z is known, the IS estimator is then

$$\hat{I} = \frac{1}{NZ} \sum_{i=1}^N w_i f(\mathbf{x}_i). \quad (25)$$

The difference between the standard and deterministic mixture (DM) IS estimators lies in the computation of the unnormalized weights. On the one hand, we recall the standard IS weights are given by

$$w_i = \frac{\pi(\mathbf{x}_i)}{q_i(\mathbf{x}_i)}, \quad (26)$$

where $\pi(\mathbf{x}_i)$ is the target evaluated at the i -th sample (drawn from the i -th proposal). Substituting (26) into (25), we obtain the standard IS estimator,

$$\hat{I}_{IS} = \frac{1}{NZ} \sum_{i=1}^N \frac{f(\mathbf{x}_i)\pi(\mathbf{x}_i)}{q_i(\mathbf{x}_i)}. \quad (27)$$

On the other hand, the weights in the DM approach are given by

$$w_i = \frac{\pi(\mathbf{x}_i)}{\sum_{j=1}^N q_j(\mathbf{x}_i)}. \quad (28)$$

Substituting (28) into (25) we obtain the DM estimator

$$\hat{I}_{DM} = \frac{1}{NZ} \sum_{i=1}^N \frac{f(\mathbf{x}_i)\pi(\mathbf{x}_i)}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x}_i)}. \quad (29)$$

⁵From now on, we use $K_i = 1$, with $i = 1, \dots, N$, for the sake of clarity, but the analysis can be straightforwardly extended to any K_i .

A.2. Unbiasedness of the DM-IS estimator

It is well known that \hat{I}_{IS} in Eq. (27) is an unbiased estimator of the integral I define in Eq. (2) [25, 22]. In this section, we prove that the DM-IS estimator in Eq. (29) is also unbiased. Since $\mathbf{x}_i \sim q_i(\mathbf{x})$, we have

$$E[\hat{I}_{DM}] = \frac{1}{NZ} \sum_{i=1}^N E_{q_i} \left[\frac{f(\mathbf{x}_i)\pi(\mathbf{x}_i)}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x}_i)} \right] \quad (30)$$

$$= \frac{1}{NZ} \sum_{i=1}^N \int \frac{f(\mathbf{x}_i)\pi(\mathbf{x}_i)}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x}_i)} q_i(\mathbf{x}_i) d\mathbf{x}_i \quad (31)$$

$$= \frac{1}{Z} \int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x})} \left[\frac{1}{N} \sum_{i=1}^N q_i(\mathbf{x}) \right] d\mathbf{x} \quad (32)$$

$$= \frac{1}{Z} \int f(\mathbf{x})\pi(\mathbf{x}) d\mathbf{x} = I. \quad (33)$$

□

A.3. Variance of the DM-IS estimator

In this section, we prove that the DM-IS estimator in Eq. (29) always has a lower or equal variance than the standard IS estimator of Eq. (27). We also consider the standard mixture (SM) estimator \hat{I}_{SM} , where N samples are independently drawn from the mixture of proposals, i.e., $\mathbf{z}_i \sim \frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x})$, and

$$\hat{I}_{SM} = \frac{1}{NZ} \sum_{i=1}^N \frac{f(\mathbf{z}_i)\pi(\mathbf{z}_i)}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{z}_i)}. \quad (34)$$

Note that obtaining an IS estimator with finite variance essentially amounts to having a proposal with heavier tails than the target. See [25, 16] for sufficient conditions that guarantee this finite variance.

Theorem 1. *For any target distribution $\pi(\mathbf{x})$, any square integrable function $f(\mathbf{x})$, and any set of proposal densities $\{q_i(\mathbf{x})\}_{i=1}^N$ such that the variance of the corresponding estimators is finite, the variance of the DM estimator is always lower or equal than the variance of the corresponding standard IS and mixture (SM) estimators, i.e.,*

$$Var(\hat{I}_{DM}) \leq Var(\hat{I}_{SM}) \leq Var(\hat{I}_{IS}). \quad (35)$$

Proof: The proof is given by Proposition 2 and Proposition 3. \square

Proposition 2.

$$\text{Var}(\hat{I}_{SM}) \leq \text{Var}(\hat{I}_{IS}). \quad (36)$$

Proof: The variance of the IS estimator is given by

$$\text{Var}(\hat{I}_{IS}) = \sum_{i=1}^N \frac{1}{N^2 Z^2} \int \frac{f^2(\mathbf{x})\pi^2(\mathbf{x})}{q_i(\mathbf{x})} d\mathbf{x} - \frac{I^2}{N}, \quad (37)$$

where $I = \frac{1}{Z} \int f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$ is the true value of the integral that we want to estimate [15]. The variance of the SM estimator is given by

$$\begin{aligned} \text{Var}(\hat{I}_{SM}) &= \frac{1}{N^2} \sum_{i=1}^N \left(\frac{1}{Z^2} \int \frac{f^2(\mathbf{x})\pi^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} - I^2 \right) \\ &= \frac{1}{NZ^2} \int \frac{f^2(\mathbf{x})\pi^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} - \frac{I^2}{N}, \end{aligned} \quad (38)$$

where $\psi(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x})$. Subtracting (38) and (37), we get

$$\text{Var}(\hat{I}_{SM}) - \text{Var}(\hat{I}_{IS}) = \frac{1}{N^2 Z^2} \int \left(\frac{N}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x})} - \sum_{i=1}^N \frac{1}{q_i(\mathbf{x})} \right) f^2(\mathbf{x})\pi^2(\mathbf{x}) d\mathbf{x}.$$

Hence, since $f^2(\mathbf{x})\pi^2(\mathbf{x}) \geq 0 \forall \mathbf{x}$, in order to prove the theorem it is sufficient to show that

$$\frac{1}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x})} \leq \frac{1}{N} \sum_{i=1}^N \frac{1}{q_i(\mathbf{x})}. \quad (39)$$

Now, let us note that the left hand side of (39) is the inverse of the arithmetic mean of $q_1(\mathbf{x}), \dots, q_N(\mathbf{x})$,

$$A_N = \frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x}),$$

whereas the right hand side of (39) is the inverse of the harmonic mean of $q_1(\mathbf{x}), \dots, q_N(\mathbf{x})$,

$$\frac{1}{H_N} = \frac{1}{N} \sum_{i=1}^N \frac{1}{q_i(\mathbf{x})}.$$

Therefore, the inequality in (39) is equivalent to stating that $\frac{1}{A_N} \leq \frac{1}{H_N}$, or equivalently $A_N \geq H_N$, which is the well-known arithmetic mean–harmonic mean inequality for positive real numbers [17, 1]. \square

Proposition 3.

$$\text{Var}(\hat{I}_{DM}) \leq \text{Var}(\hat{I}_{SM}). \quad (40)$$

Proof: The variance of \hat{I}_{DM} is computed

$$\begin{aligned} \text{Var}(\hat{I}_{DM}) &= \frac{1}{N^2 Z^2} \sum_{i=1}^N \left(E_{q_i(\mathbf{x})} \left[\frac{f^2(\mathbf{x})\pi^2(\mathbf{x})}{\psi^2(\mathbf{x})} \right] - E_{q_i(\mathbf{x})}^2 \left[\frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} \right] \right) \\ &= \frac{1}{N^2 Z^2} \sum_{i=1}^N \left(\int \frac{f^2(\mathbf{x})\pi^2(\mathbf{x})}{\psi^2(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right) - \frac{1}{N^2 Z^2} \sum_{i=1}^N \left(\int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right)^2 \\ &= \frac{1}{NZ^2} \left(\int \frac{f^2(\mathbf{x})\pi^2(\mathbf{x})}{\psi^2(\mathbf{x})} \left[\frac{1}{N} \sum_{i=1}^N q_i(\mathbf{x}) \right] d\mathbf{x} \right) - \frac{1}{N^2 Z^2} \sum_{i=1}^N \left(\int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right)^2 \\ &= \frac{1}{NZ^2} \int \frac{f^2(\mathbf{x})\pi^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} - \frac{1}{N^2 Z^2} \sum_{i=1}^N \left(\int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right)^2 \end{aligned} \quad (41)$$

Analyzing Eqs. (38) and (41), we see that proving $\text{Var}(\hat{I}_{DM}) \leq \text{Var}(\hat{I}_{SM})$ is equivalent to proving that

$$\begin{aligned} \frac{1}{Z^2} \sum_{i=1}^N \left(\int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right)^2 &\geq NI^2 \\ \frac{1}{Z^2} \sum_{i=1}^N \left(\int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right)^2 &\geq N \left(\frac{1}{Z} \int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} \psi(\mathbf{x}) d\mathbf{x} \right)^2 \\ \sum_{i=1}^N \left(\int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right)^2 &\geq N \left(\int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} \left(\frac{1}{N} \sum_{i=1}^N q_i(\mathbf{x}) \right) d\mathbf{x} \right)^2 \\ \sum_{i=1}^N \left(\int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right)^2 &\geq \frac{1}{N} \left(\sum_{i=1}^N \int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x} \right)^2 \\ N \sum_{i=1}^N a_i^2 &\geq \left(\sum_{i=1}^N a_i \right)^2 \end{aligned} \quad (42)$$

with $a_i = \int \frac{f(\mathbf{x})\pi(\mathbf{x})}{\psi(\mathbf{x})} q_i(\mathbf{x}) d\mathbf{x}$. The inequality in Eq. (42) holds since it corre-

sponds to the definition of the Cauchy-Schwarz inequality [17],

$$\left(\sum_{i=1}^N a_i^2 \right) \left(\sum_{i=1}^N b_i^2 \right) \geq \left(\sum_{i=1}^N a_i b_i \right)^2, \quad (43)$$

with $b_i = 1$ for $i = 1, \dots, N$. □

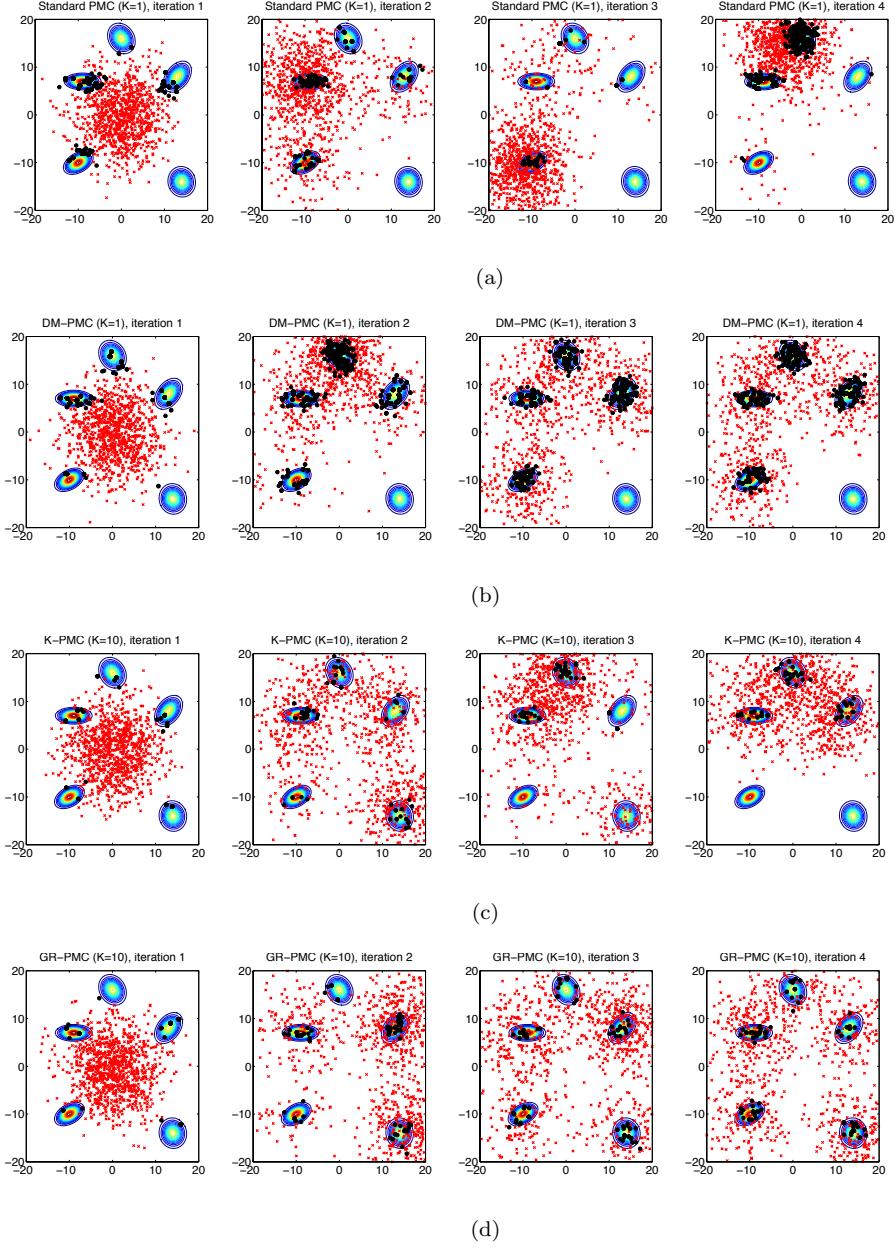


Figure 7: **(Ex1-Section 5.2)** Evolution of the samples before (red crosses) and after resampling (black circles) for different schemes using $N = 100$ and $\sigma = 5$. The contour lines of the target density are also depicted. **(a)** Standard PMC. **(b)** DM-PMC ($K=1$). **(c)** K-PMC ($K=10$) with global resampling. **(d)** GR-PMC ($K=10$).