MARGINALIZED POPULATION MONTE CARLO

Mónica F. Bugallo, Mingyi Hong, and Petar M. Djurić

Department of Electrical and Computer Engineering
Stony Brook University, Stony Brook, NY 11794, USA
e-mail: {monica,mhong,djuric}@ece.sunysb.edu

ABSTRACT

Population Monte Carlo is a statistical method that is used for generation of samples approximately from a target distribution. The method is iterative in nature and is based on the principle of importance sampling. In this paper, we show that in problems where some of the parameters are conditionally linear on the remaining parameters, we can improve the computational efficiency of population Monte Carlo by generating samples of the nonlinear parameters only and marginalizing the linear parameters. We demonstrate the marginalized population Monte Carlo on the problem of frequency estimation of closely spaced sinusoids.

Index Terms— Population Monte Carlo, parameter estimation, marginalization.

1. INTRODUCTION

Monte Carlo–based signal processing has gained much steam recently. With the advancement of Markov chain Monte Carlo (MCMC) sampling [1] and particle filtering, sets of very important classes of problems can be addressed with new tools, including problems that require batch processing and ones that are sequential in nature.

MCMC methods are used for generating samples from probability distributions based on constructing Markov chains whose stationary distributions are the target distributions. An alternative approach to sampling from target distributions is the population Monte Carlo (PMC) method, which is also iterative in nature but, unlike the MCMC, does not have burn-in period in which the generated samples cannot be used. The method is based on the principle of importance sampling, which suggests that one should generate samples more frequently from regions of the support that are more important than other regions (in other words, regions with greater probability masses associated to them) [2].

The PMC method was described in [3] where it was presented together with other sample (particle)–based methods. Particle filtering is composed of the same ingredients as PMC, and in [4], an algorithm, in spirit like PMC, was proposed for estimation of static parameters by particle filtering. In [5], a PMC algorithm based on adaptive importance sampling for static models was proposed. PMC was used in [6] for ion channel restoration. A new PMC algorithm that reduces the asymptotic variance for a function of interest was proposed in [7]. In [8], it was shown that PMC algorithms can be progressively adapted to a target distribution with a diminishing Kullback divergence.

In this paper, we address the use of Rao-Blackwellization (RB) and the PMC method [9]. Previous work on this subject can be found in [10], where RB was applied to marginalization of missing data through numerical integration. In our work, we consider the general problem of having conditionally linear parameters, assume certain structure of distributions that allow for analytical integration and apply sampling only for the nonlinear parameters of the model. We argue that with this approach one can construct new proposals much more easily and use the generated samples much more efficiently.

For demonstration of the proposed approach, we work on the classical problem of frequency estimation of closely spaced sinusoids. This is a well studied problem, which perfectly fits the aims of our study. Namely, the nonlinear parameters of the model are the frequencies of the sinusoids and the linear parameters are the amplitudes of the sinusoids.

The paper is organized as follows. First we formulate the problem in Section 2. In Section 3, we explain the proposed method and discuss some of its advantages and disadvantages. We demonstrate the implementation of the marginalized PMC to the problem of frequency estimation of sinusoids in Section 4. The simulation results that demonstrate the method's performance are presented in Section 5. The paper is concluded with final thoughts given in Section 6.

2. PROBLEM FORMULATION

The problem we address can be stated as follows: we have a $d_y \times 1$ vector of observations $\boldsymbol{y}$, which is modeled according to

$$\boldsymbol{y} \quad = \quad h(\boldsymbol{x}_n) + \mathbf{A}(\boldsymbol{x}_n)\,\boldsymbol{x}_l + \boldsymbol{w} \tag{1}$$

where $\boldsymbol{x}$ is a $d_x \times 1$ vector of unknown parameters. This vector is composed of linear parameters $\boldsymbol{x}_l$ of dimension d_{x_l} and nonlinear parameters $\boldsymbol{x}_n$ of dimension d_{x_n}, where $d_x = d_{x_n} + d_{x_l}$. The symbol $h(\cdot)$ denotes a nonlinear function of the parameters $\boldsymbol{x}_n$; $\mathbf{A}(\boldsymbol{x}_n)$ is a $d_y \times d_{x_l}$ matrix whose entries are functions of the nonlinear parameters; and $\boldsymbol{w}$ is a noise vector with a known probability distribution. Let the prior density of the unknown parameters be given by $p(\boldsymbol{x}_n, \boldsymbol{x}_l)$ and the distribution of the noise by $p(\boldsymbol{w})$. The objective is to estimate $\boldsymbol{x}$ from the observation vector $\boldsymbol{y}$ based on the made assumptions. In particular, the objective is to apply the PMC method for estimation.

3. PROPOSED METHOD

PMC is an adaptive importance sampling procedure where the importance function changes with every iteration with the aim to

This work has been supported by the National Science Foundation under Award CCF-0515246 and the Office of Naval Research under Award N00014-06-1-0012.The third author also gratefully acknowledges the support from the award Universidad Carlos III de Madrid-Banco de Santander Chair of Excellence.

produce samples that better represent the target distribution. The advantage of PMC over MCMC methods [11] is that it can be stopped at any time. In the next subsection, we provide a review of the method.

3.1. Brief review of PMC

The underlying principle used in PMC is importance sampling. In the past, importance sampling has primarily been used for numerical integration and more recently in particle filtering [12], [13]. In signal processing, a standard problem is the estimation of unknowns, and to that end we often use point estimates of the unknowns or provide their posterior distributions. For the purpose of a simplified presentation, let the unknown be a scalar denoted by x. One standard point estimator of it is the minimum mean-square estimate of x defined by

$$\eta_x \quad = \quad \int x p(x|\boldsymbol{y}) dx \tag{2}$$

where $\boldsymbol{y}$ is the vector of observations and $p(x|\boldsymbol{y})$ is the posterior of x. If we can draw samples from $p(x|\boldsymbol{y})$, i.e.,

$$x^{(m)} \sim \; p(x|\boldsymbol{y}), \quad m = 1, 2, \cdots, M \tag{3}$$

we can compute the integral (2) according to classical Monte Carlo integration by

$$\hat{\eta}_x \quad \simeq \quad \frac{1}{M} \sum_{m=1}^{M} x^{(m)} \tag{4}$$

where M is the total number of independently drawn samples. By the strong law of large numbers the estimate will converge to the true mean of the posterior. In addition, for M large, one can write

$$\frac{\hat{\eta}_x - \eta_x}{\sigma_{\hat{\eta}_x}} \quad \sim \quad \mathcal{N}(0, 1) \tag{5}$$

where

$$\sigma_{\hat{\eta}_x} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (x^{(m)} - \hat{\eta}_x)^2}.$$

Often, one cannot draw samples from the posterior $p(x|\boldsymbol{y})$. Instead, one can use for drawing $x^{(m)}$ another probability distribution $q(x)$, called an importance function, and proceed as follows:

$$\begin{aligned} \eta_x \quad &= \quad \int x p(x|\boldsymbol{y}) dx \\ &= \quad \int x \frac{p(x|\boldsymbol{y})}{q(x)} q(x) dx \\ &\simeq \quad \frac{1}{M} \sum_{m=1}^{M} \frac{x^{(m)} p(x^{(m)}|\boldsymbol{y})}{q(x^{(m)})}. \end{aligned} \tag{6}$$

When some conditions about $q(x)$ are satisfied, it can be shown that by using the strong law of large numbers, this estimate, too, converges to the true value.

The above estimate is valid if the posterior $p(x|\boldsymbol{y})$ and $q(x)$ are known completely. If they are only known up to their proportionality constants, then η_x can be estimated by

$$\hat{\eta}_x \quad = \quad \frac{\sum_{m=1}^{M} x^{(m)} \frac{p(x^{(m)}|\boldsymbol{y})}{q(x^{(m)})}}{\sum_{m=1}^{M} \frac{p(x^{(m)}|\boldsymbol{y})}{q(x^{(m)})}} \tag{7}$$

where $\hat{\eta}_x$ also converges by the strong law of large numbers.

One interpretation of this result is that the samples $x^{(m)}$ form the support of a discrete random measure $\chi = \{x^{(m)}, w^{(m)}\}$ where $w^{(m)}$ are weights associated to the samples, and

$$w^{(m)} \quad \propto \quad \frac{p(x^{(m)}|\boldsymbol{y})}{q(x^{(m)})} \tag{8}$$

and

$$\sum_{m=1}^{M} w^{(m)} = 1. \tag{9}$$

This random measure "approximates" a target distribution (in our case a posterior) and can be used for computing estimates of integrals under that distribution.

The idea of PMC is to apply the importance sampling iteratively, that is, once the first random measure χ_1 is obtained, based on its support and set of weights one constructs a better importance function followed by the generation of a new set of samples from it and association of new weights to the samples. Thereby, one obtains χ_2 and continues to construct in a similar fashion χ_3, χ_4, and so on.

A generic implementation of a PMC algorithm contains the following steps. Let j denote the iteration number. Then, for $j = 1, 2, \cdots$

1. Choose an importance function $q_{j,m}(x)$
2. Draw the samples $x_j^{(m)}$ from $q_{j,m}(x)$, $m = 1, 2, \cdots, M$.
3. Compute the weights

$$\tilde{w}_j^{(m)} \propto \frac{p(x_j^{(m)}|\boldsymbol{y})}{q_{j,m}(x_j^{(m)})}, \quad m = 1, 2, \cdots, M.$$

4. Normalize the weights by

$$w_j^{(m)} = \frac{\tilde{w}_j^{(m)}}{\sum_{k=1}^{M} \tilde{w}_j^{(k)}}.$$

If another iteration is needed, set $j = j + 1$ and go back to step one.

It should be noted that in the process of selecting an importance function, we can employ resampling according to a multinomial distribution defined by the weights $w_j^{(m)}$.

3.2. Marginalized PMC

The PMC methods are of great interest in high dimensional and nonlinear problems where standard computational methods are difficult to implement. However, when the dimension of the parameter space is high, the use of PMC is also challenging because it requires generation of a large number of samples. In some problems, some of the unknown parameters are conditionally linear given the remaining parameters, which may allow for an implementation of the PMC that requires only generation of samples of the nonlinear parameters. If the distributions permit analytical integrations of the linear parameters (their marginalization), then much improvement in computational efficiency of the PMC can be achieved. This idea is analogous to the one known of RB and used in particle filtering [9].

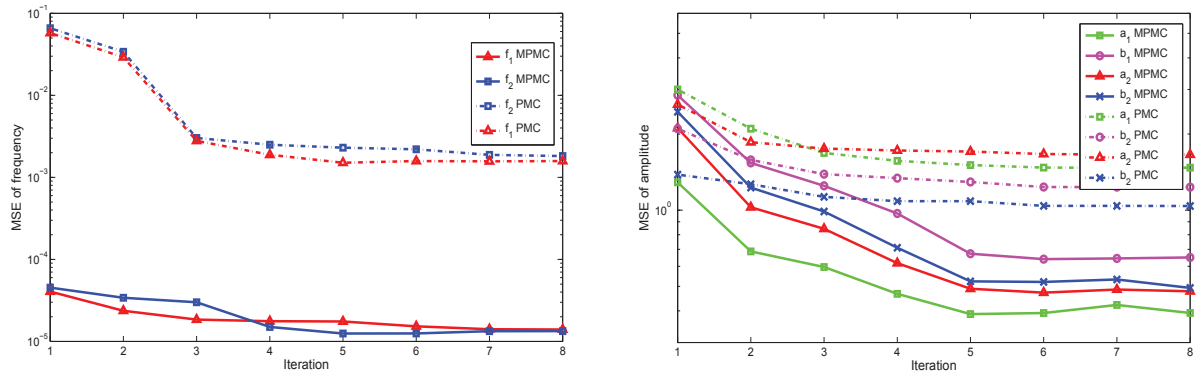

Fig. 1. MSE vs number of iterations. Left: Frequencies. Right: Magnitudes of the amplitudes.

Consider the model given by (1). Suppose that at iteration j, we only generate samples of the nonlinear parameters, $\boldsymbol{x}_{n,j}^{(m)}$. The weights that correspond to these parameters are given by

$$w_j^{(m)} \quad \propto \quad \frac{p(\boldsymbol{x}_{n,j}^{(m)}|\boldsymbol{y})}{q(\boldsymbol{x}_{n,j}^{(m)})} \tag{10}$$

where in the numerator of (10) we have the marginalized posterior of $\boldsymbol{x}_n$ for which we can write

$$p(\boldsymbol{x}_{n,j}^{(m)}|\boldsymbol{y}) \propto \int p(\boldsymbol{y}|\boldsymbol{x}_l, \boldsymbol{x}_{n,j}^{(m)})\, p(\boldsymbol{x}_l, \boldsymbol{x}_{n,j}^{(m)}) d\boldsymbol{x}_l. \tag{11}$$

If we can solve the integral in (11) analytically, the computation of (10) is straightforward. Thus, the implementation of the proposed approach hinges on the ability to solve (11).

4. EXAMPLE

We demonstrate our approach on the problem of frequency estimation of sinusoids in noise. Let the observations $\boldsymbol{y}$ be modeled by

$$\boldsymbol{y} \quad = \quad \mathbf{A}(\boldsymbol{x}_n)\, \boldsymbol{x}_l + \boldsymbol{w} \tag{12}$$

where $\boldsymbol{y}$ is a $d_y \times 1$ vector, $\mathbf{A}(\boldsymbol{x}_n)$ is a $d_y \times 2K$ matrix where K is the number of sinusoids in the data, $\boldsymbol{x}_n$ is a $K \times 1$ vector of frequencies, $\boldsymbol{x}_l$ is a $2K \times 1$ vector of amplitudes and $\boldsymbol{w}$ is a zero-mean Gaussian noise vector with a covariance matrix $\mathbf{C}_w$.

More specifically, an element of $\boldsymbol{y}$ is obtained by

$$y_t = \sum_{k=1}^{K} (a_k \cos(2\pi f_k t) + b_k \sin(2\pi f_k t)) + w_t, \quad t = 1, \cdots, d_y$$

where $\{a_k, b_k\}$ are the amplitudes of the cosine and sine components of the k-th sinusoid, respectively; f_k denotes the frequency of the k-th sinusoid; and w_t is the observation noise. Hence the linear and nonlinear parameters are given by

$$\begin{aligned} \boldsymbol{x}_l &= [a_1\ b_1\ a_2\ b_2\ \cdots\ a_K\ b_K]^\top \\ \boldsymbol{x}_n &= [f_1\ f_2\ \cdots\ f_K]^\top. \end{aligned}$$

The priors of the amplitudes and the frequencies are considered independent, i.e.,

$$p(\boldsymbol{x}_l, \boldsymbol{x}_n) = p(\boldsymbol{x}_l)\ p(\boldsymbol{x}_n).$$

For the prior of the frequencies, we adopt a constant over the region $0 < f_1 < f_2 < \cdots < f_K < 0.5$, where without loss of generality, we identify the sinusoids as first, second, third etc. by their frequencies. As a prior of the amplitudes we use a zero-mean Gaussian with a covariance matrix $\mathbf{C}_{x_l}$.

As mentioned in the previous section, a critical step in the implementation of the marginalized PMC is the ability to compute $p(\boldsymbol{x}_n^{(m)}|\boldsymbol{y})$. For the stated problem, we can readily show that

$$p(\boldsymbol{x}_n^{(m)}|\boldsymbol{y}) \quad \propto \quad \frac{\exp\left(-\frac{1}{2}\boldsymbol{y}^\top \left(\mathbf{C}_w + \mathbf{A}\mathbf{C}_{x_l}\mathbf{A}^\top\right)^{-1} \boldsymbol{y}\right)}{|\mathbf{C}_w + \mathbf{A}\mathbf{C}_{x_l}\mathbf{A}^\top|^{\frac{1}{2}}}$$

where for convenience of presentation, we dropped the argument of $\mathbf{A}$, $\boldsymbol{x}_n^{(m)}$.

5. SIMULATIONS

We present computer simulations that illustrate the validity of our proposed method. We generated data according to the observation model explained in the previous section.

In our experimental setup we generated data sets of $d_y = 30$ samples and we fixed the number of sinusoids to two, i.e., $K = 2$. Therefore we considered a six-dimensional state $\boldsymbol{x}$, composed of a nonlinear part of two components $\boldsymbol{x}_n = [f_1\ f_2]^\top$ and a linear part of four components $\boldsymbol{x}_l = [a_1\ b_1\ a_2\ b_2]^\top$. The prior for the frequencies was uniform over the region $0 \leq f_1 < f_2 < 0.5$ and for the amplitudes was a zero-mean Gaussian with covariance matrix $\mathbf{C}_{x_l} = 5\mathbf{I}$.

We applied the proposed method (labeled as `MPMC`) and for comparison and benchmarking purposes, we also implemented the standard PMC algorithm that does not marginalize out the linear part of the state (labeled as `PMC`). Both algorithms were run with $M = 1000$ particles. For construction of the importance function, we followed the procedure of [5], where we used a mixture density with five mixands with predetermined variance vectors.

The performance of the methods was quantified by computing the mean square error (MSE) of the parameters calculated as

$$\text{MSE} = \frac{1}{J}\sum_{j=1}^{J}(\hat{x}^j - x^j)^2$$

where x^j was the true value of the parameter x (frequency or

amplitude[1]) in the j-th run, and $\hat{x}^j$ was the corresponding estimate obtained by the method. The MSE plots were obtained by averaging over $J = 200$ independent simulations. Figure 1 depicts the average MSE for the parameters of the sinusoids when the true value for the frequency of the first sinusoid was set to $f_1 = 0.24$ and that of the second to $f_2 = f_1 + \frac{1}{2d_y}$ (Note that the differences in the frequencies is two times smaller than the resolution of the classical periodogram). The amplitudes of the sinusoids were $\boldsymbol{x}_l = [a_1 \;\; b_1 \;\; a_2 \;\; b_2]^\top = [2\cos\frac{\pi}{6} \;\; 2\sin\frac{\pi}{6} \;\; 2\cos\frac{-\pi}{5} \;\; 2\sin\frac{-\pi}{5}]^\top$ and the signal-to-noise ratios (SNRs) of both sinusoids were 10 dB. From the results shown in the Figure we see that the proposed method outperforms the PMC which does not marginalize the linear parameters and performs closely to the lower bound imposed by the MMSE for the linear parameters.

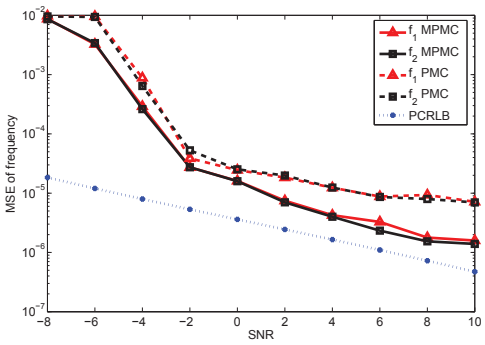

Fig. 2. MSE of the frequencies vs SNR.

Figure 2 shows the performance of the methods in terms of the MSE of the estimated frequencies for various values of SNR[2]. In this case, for each run of the experiment we drew the frequencies and amplitudes of the sinusoids from the corresponding priors. The Figure also includes the posterior Cramér-Rao bound calculated following [14], which serves as a benchmark for the estimates of the unknowns. Similar conclusions as in the previous plot can be drawn.

Finally, Figure 3 displays the histograms of the marginalized posteriors of the nonlinear parameters approximated by our method. In the Figure, we also superimposed the periodogram of the data which clearly cannot discriminate the two sinusoids. The histograms show that the marginalized posterior of the frequencies has most of the probability masses around the true values of the frequencies.

6. CONCLUSIONS

In this paper we present a population Monte Carlo method for parameter estimation in systems with both nonlinear and conditionally linear parameters. Samples are only generated for the nonlinear parameters of the model and the linear parameters are integrated out. We demonstrated the method on estimation of frequencies of sinusoids embedded in Gaussian noise, where samples of the frequencies are generated and the amplitudes of the sinusoids are marginalized. The simulation results show improved performance over the standard population Monte Carlo approach.

[1]Note that, although the proposed algorithm does not explicitly estimate the linear parameters, their estimation can be obtained in a straightforward manner.

[2]The values of SNR in the x-axis of the plot are for each of the sinusoids, and therefore for each experiment both sinusoids were assumed to have the same SNR.

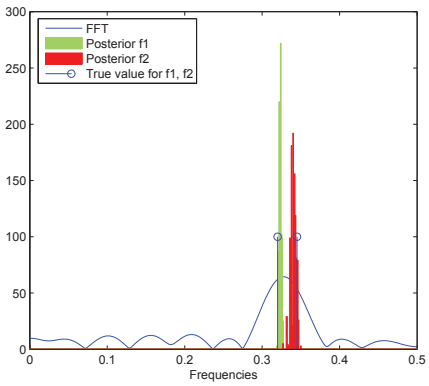

Fig. 3. Periodogram of the data and approximated posterior of the nonlinear parameters by the proposed MPMC.

7. REFERENCES

[1] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer, 2005.

[2] A. Marshall, *Symposium on Monte Carlo Methods*. Wiley. New York, 1956, ch. The use of mulit-stage sampling schemes in Monte Carlo computations, pp. 123–140.

[3] R. Iba, "Population-based Monte Carlo algorithms," *Journal of Computational and Graphical Statistics*, vol. 7, pp. 175–193, 2000.

[4] N. Chopin, "A sequential particle filter for static models," *Biometrika*, vol. 89, no. 3, pp. 539–552, 2002.

[5] O. Cappé, J. M. Marin, and C. P. Robert, "Population Monte Carlo," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 907–929, 2003.

[6] O. Cappé, A. Guillin, J.-M. Marin, and C. P. Robert, "Population Monte Carlo for ion channel restoration," *Journal of Computational & Graphical Statistics*, to appear.

[7] R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert, "Minimum variance importance sampling via population Monte Carlo," Cahiers du CEREMADE, Université Paris Dauphine, Tech. Rep., 2005.

[8] ——, "Convergence of adaptive mixtures of importance sampling schemes," *The Annals of Statistics*, vol. 35, pp. 420–448, 2007.

[9] G. Casella and C. P. Robert, "Rao-Blackwellization of sampling schemes," *Biometrika*, vol. 84, pp. 81–94, 1996.

[10] G. Celeux, J.-M. Marin, and C. P. Robert, "Iterated importance sampling in missing data problems," *Computational Statistics & Data Analysis*, vol. 50, pp. 3386–3404, 2006.

[11] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall, 1996.

[12] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez, "Particle filtering," *IEEE Signal Processing Magazine*, vol. 20, no. 5, pp. 19–38, 2003.

[13] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. New York: Springer, 2001.

[14] H. L. V. Trees, *Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 1968.