Estimation of cosmological parameters using adaptive importance sampling

Darren Wraith, Martin Kilbinger, Karim Benabed, Olivier Cappe,
Jean-François Cardoso, Gersende Fort, Simon Prunet, Christian Robert

▶ **To cite this version:**

Darren Wraith, Martin Kilbinger, Karim Benabed, Olivier Cappe, Jean-François Cardoso, et al.. Estimation of cosmological parameters using adaptive importance sampling. Physical Review D, 2009, 80, pp.023507. 10.1103/PhysRevD.80.023507 . hal-00450727

HAL Id: hal-00450727
https://hal.science/hal-00450727v1

Submitted on 1 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

PHYSICAL REVIEW D **80,** 023507 (2009)

Estimation of cosmological parameters using adaptive importance sampling

Darren Wraith,[1,2] Martin Kilbinger,[2] Karim Benabed,[2] Olivier Cappé,[3] Jean-François Cardoso,[3,2] Gersende Fort,[3] Simon Prunet,[2] and Christian P. Robert[1]

[1]*CEREMADE, Université Paris Dauphine, 75775 Paris cedex 16, France*
[2]*Institut d'Astrophysique de Paris, CNRS UMR 7095 & UPMC, 98 bis, boulevard Arago, 75014 Paris, France*
[3]*LTCI, TELECOM ParisTech and CNRS, 46, rue Barrault, 75013 Paris, France*
(Received 11 March 2009; published 10 July 2009)

We present a Bayesian sampling algorithm called adaptive importance sampling or population Monte Carlo (PMC), whose computational workload is easily parallelizable and thus has the potential to considerably reduce the wall-clock time required for sampling, along with providing other benefits. To assess the performance of the approach for cosmological problems, we use simulated and actual data consisting of CMB anisotropies, supernovae of type Ia, and weak cosmological lensing, and provide a comparison of results to those obtained using state-of-the-art Markov chain Monte Carlo (MCMC). For both types of data sets, we find comparable parameter estimates for PMC and MCMC, with the advantage of a significantly lower wall-clock time for PMC. In the case of WMAP5 data, for example, the wall-clock time scale reduces from days for MCMC to hours using PMC on a cluster of processors. Other benefits of the PMC approach, along with potential difficulties in using the approach, are analyzed and discussed.

DOI: 10.1103/PhysRevD.80.023507 PACS numbers: 98.80.Es, 02.50.−r, 02.50.Sk

I. INTRODUCTION

In recent years we have seen spectacular advances in observational cosmology, with the availability of more and more high quality data allowing for the testing of models with higher complexity. Some of these tests have been made possible thanks to the use of Bayesian sampling techniques, and, in particular, Markov chain Monte Carlo (MCMC)—an (iterative) algorithm that produces a Markov chain whose distribution converges to the target posterior π. After a ''burn-in'' period, samples from such a chain can be regarded as samples approximately from π. Proposed values for the chain or the updating scheme of MCMC can be designed to ensure that moves towards regions of higher mass under π are favored, and regions with null probability (under π) are never visited. This way, most of the computational effort can be spent in the region of importance to the posterior distribution, and an MCMC approach is usually much more efficient than traditional grid sampling of the model parameter space.

The MCMC technique is now well known in cosmology, and, in particular, in its most simple form, the Metropolis-Hastings algorithm, thanks to the user-friendly and freely available package COSMOMC [1]. Other forms of the MCMC algorithm, like Gibbs sampling and hybrid Monte Carlo (better known in cosmology as Hamiltonian sampling), have also been proposed and have found some interesting usage in the estimation of the posterior distribution for the cosmic microwave background anisotropy power spectrum at low resolution (see [2], references therein, and also [3,4]).

For all its advantages over grid sampling, the MCMC approach also suffers from problems. One difficulty is to assess the correct convergence of the chain. Another lies in the presence of correlations within the chain which can greatly reduce the efficiency of the sample [5]. A third issue which is particularly relevant for the usage of MCMC in cosmology is the computational time involved. Indeed, whatever the sampling technique, we often need to compute at least one estimate of the posterior for each sampled point. This computation can be slow in cosmology. With the current processing speed of computers, a point of the posterior of, for example, the WMAP5 data set, using CAMB [6][1] and the public WMAP5 likelihood code [4],[2] both with their default precision settings, is computed at the order of several seconds, and can be much slower when exploring nonflat models. Of course, as stated above, for most problems MCMC will require orders of magnitude less samples than a grid for a given target precision, thus providing an important efficiency improvement. However, apart from improving the likelihood codes or waiting for the availability of faster computers there is not much speed improvement to expect from an MCMC approach, while probably needed if one wants to explore yet bigger and more complex models. On the algorithmic side of the problem, some effort has been devoted recently to the improvement of the likelihood codes, mainly by using clever interpolation tricks (segmentation [7] and neural networks [8]) and by looking for improvements in the MCMC algorithm [9–11]. The former [7,8] indeed provide some gain in efficiency, but at the cost of a long precomputation step for each model. The latter improves on the natural inefficiency of the Metropolis algorithm but imposes some other requirements, like the availability of cheap computation of the derivatives of the likelihood [9], or the knowledge of conditional probabilities of some of the parameters [10,11]. Other (non-Markovian)

[1]http://camb.info
[2]http://lambda.gsfc.nasa.gov

 © 2009 The American Physical Society

Monte Carlo methods, such as nested sampling, have also been proposed and applied recently to cosmological problems with some success along with presenting their own problems [12–14].

On the hardware side, however, there is a route to speed improvement that does not lie in quicker CPUs, but on the availability of cheap multi-CPU computers and the standardization of clusters of computers. This opportunity, however, is only partly opened to MCMC. Indeed, there are two ways of parallelizing the parameter exploration: first, by distributing the computation of the likelihood, which is not always possible and does not always lead to speed improvement, and second, by running multiple chains in parallel. This last option is the simplest, but is ''forbidden'' by the iterative nature of the MCMC algorithm. More precisely, running parallel chains and mixing them in the end to build a bigger chain sample is of course possible (and can be advantageous in fully exploring the support of π), but at the condition that each of the individual chains has converged. In the absence of such a condition, significant biases in the sample can be introduced. Determining convergence for each chain is inherently difficult in practice and has largely prevented more widespread use of the approach [15]. Thus, for MCMC any speed improvements through parallelization are difficult to achieve.

In this paper, we propose another sampling algorithm suitable for cosmological applications, that is not based upon MCMC, and can be parallelized. This novel algorithm, called population Monte Carlo (PMC), is an adaptive importance sampling technique, that has been studied recently in the statistics literature [16]. While this algorithm solves some of the issues of MCMC in cosmology, the approach of course has a different set of potential problems that we will analyze and discuss, along with its advantages.

The paper is outlined as follows. In the next section, we provide a brief introduction to the Bayesian approach, which we hope will give the casual reader some important keys for further readings, and we also discuss the challenges and issues involved with using either an MCMC or an importance sampling algorithm for estimation. We then describe details of the PMC approach. In Sec. III, we assess the performance of the PMC approach using a simulated target density with features similar to cosmological parameter posteriors, and provide a comparison to results obtained using an MCMC approach. In Sec. IV, we illustrate the results from the PMC approach using actual data, consisting of CMB anisotropies, supernovae of type Ia and weak cosmological lensing. We conclude in Sec. V with a discussion and an outlook for further work.

II. METHODS

A. Bayesian inference via simulation

A key feature of Bayesian inference is to provide a probabilistic expression for the uncertainty regarding a parameter of interest x by combining prior information along with information brought by the data. Prior information, for example, could take the form of information obtained from previous experiments which cannot readily be incorporated into the current experiment or simply consist of a feasible range. The absence of prior information, however, is not a restriction for the use of Bayesian inference and estimation can still be regarded as valid [17]. Information brought by the data and prior information are entirely subsumed in the posterior probability density function obtained, up to a normalization constant, by

$$\pi(x) \propto \text{likelihood}(\text{data}|x) \times \text{prior}(x). \tag{1}$$

It is however generally difficult to handle the posterior distribution, due to (a) the dimension of the parameter vector x, and (b) the use of nonanalytical likelihood functions. For both of these reasons, the normalizing constant missing from the right-hand side of (1) is usually not explicitly available. A practical solution to this difficulty is to replace the analytical study of the posterior distribution with a simulation from this distribution, since producing a sample from π allows for a straightforward approximation of all integrals related with π, due to the Monte Carlo principle [5]. In short, if $x_1, \ldots, x_N$ is a sample drawn from the distribution π and f denotes a function (with finite expectation under π), the empirical average

$$\frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{2}$$

is a convergent estimator of the integral

$$\pi(f) = \int f(x)\pi(x)dx, \tag{3}$$

in the sense that the empirical mean (2) converges to $\pi(f)$ as N grows to infinity. Quantities of interest in a Bayesian analysis typically include the posterior mean, for which $f(x) = x$; the posterior covariance matrix corresponding to $f(x) = xx^{\mathrm{T}}$; and probability intervals, with $f(x) = \mathbf{1}_S(x)$, where S is a domain of interest, and $\mathbf{1}_S(x)$ denotes the indicator function which is equal to one if $x \in S$ and zero otherwise.

B. Markov chain Monte Carlo methods

For most problems in practice, direct simulation from π is not an option and more sophisticated approximation techniques are necessary. One of the standard approaches [5] to the simulation of complex distributions is the class of MCMC methods that rely on the production of a Markov chain $\{x_n\}$ having the target posterior distribution π as a limiting distribution.

MCMC can be implemented with many Markovian proposal distributions but the standard approach is the random walk Metropolis-Hastings algorithm: given the current value x_n of the chain, a new value $x_\star$ is drawn

from $\psi(x - x_n)$, where the so-called proposal ψ denotes a symmetric probability density function. The point $x_\star$ is then accepted as x_{n+1} with probability (also called the acceptance rate in this context)

$$\min\left\{1, \frac{\pi(x_\star)}{\pi(x_n)}\right\}, \tag{4}$$

and otherwise, $x_{n+1} = x_n$. The algorithm is implemented as follows:

Random walk Metropolis-Hastings algorithm

Do: Choose an arbitrary value of x_1.
For $n \geq 1$:
Generate $x_\star \sim \psi(x - x_n)$ and $u \sim \text{Uniform}(0, 1)$.
Take

$$x_{n+1} = \begin{cases} x_\star & \text{if } u \leq \pi(x_\star)/\pi(x_n), \\ x_n & \text{otherwise.} \end{cases}$$

While this algorithm is universal in that it applies to any choice of posterior distribution π and proposal ψ, its performance highly depends on the choice of the proposal ψ that has to be properly tuned to match some characteristics of π. If the scale of the proposal ψ is too small, that is, if it takes many steps of the random walk to explore the support of π, the algorithm will require many iterations to converge and, in the most extreme cases, will fail to converge in the sense that it will miss some relevant part of the support of π [18]. If, on the other hand, the scale of ψ is too large, the algorithm may also fail to adequately sample from π. This time, the chain may exhibit low acceptance rates and fail to generate a sufficiently diverse sample, even with longer runs. There exist monitors that assess the convergence of such algorithms but they usually are conservative—i.e., require a multiple of the number of necessary iterations—and partial—i.e., only focus on a particular aspect of convergence or on a special class of targets [5]. MCMC algorithms are also notoriously delicate to calibrate online, both from a theoretical point of view and from a practical perspective [19]. For these approaches, often called adaptive MCMC, some recommendations for the optimal scaling and calibration schedule for various proposals in high dimensions have been proposed [20], but this is still at an experimental stage.

C. Population Monte Carlo

PMC [16,21] is an adaptive version of importance sampling [22,23] that produces a sequence of samples (or populations) that are used in a sequential manner to construct improved importance functions and improved estimations of the quantities of interest.

We recall that importance sampling is based on the fundamental identity [5]

$$\pi(f) = \int f(x)\pi(x)dx = \int f(x)\frac{\pi(x)}{q(x)}q(x)dx, \tag{5}$$

which holds for any probability density function q with support including the support of π and any function f for which the expectation $\pi(f)$ is finite. Hence, this approach to approximating integrals linked with complex distributions is also universal in that the above identity always holds. If $x_1, \ldots, x_N$ are drawn *independently* from q,

$$\hat{\pi}(f) = \frac{1}{N}\sum_{n=1}^{N} f(x_n)w_n, \qquad w_n = \pi(x_n)/q(x_n), \tag{6}$$

provides a converging approximation to $\pi(f)$. In this context, q is called the importance function and w_n are commonly referred to as importance weights. For Bayesian inference, one cannot directly use (6) as only the unnormalized version of π [i.e., the right-hand side of Eq. (1)] is available. Conveniently, the self-normalized importance ratio

$$\hat{\pi}_{\mathrm{N}}(f) = \sum_{n=1}^{N} f(x_n)\bar{w}_n, \tag{7}$$

where the normalized importance weights are defined as

$$\bar{w}_n = \frac{w_n}{\sum_{m=1}^{N} w_m}, \tag{8}$$

is also a converging approximation to $\pi(f)$, independent of the normalization of π. For an importance function that is closely matched to the target density, significant reductions in the variance of the Monte Carlo estimates are possible in comparison to estimates obtained using MCMC [5]. However, the importance sampling approach is equally prone to poor performances as MCMC, in that the resulting converging approximation may suffer from a large or even infinite variance if q is not selected in accordance with π. There is no universal importance function and most of the research in this field aims at fitting the most efficient importance functions for the problem at hand.

Population Monte Carlo offers a possible solution to this difficulty through adaptivity: given the target posterior density π up to a constant, PMC produces a sequence q^t of importance functions ($t = 1, \ldots, T$) aimed at approximating this very target. The first sample is produced by a regular importance sampling scheme, $x_1^1, \ldots, x_N^1 \sim q^1$, associated with importance weights

$$w_n^1 = \frac{\pi(x_n^1)}{q^1(x_n^1)}, \qquad n = 1, \ldots, N, \tag{9}$$

and their normalized counterparts $\bar{w}_n^1$ [Eq. (8)], providing a first approximation to a sample from π. Moments of π can then be approximated to construct an updated importance function q^2, etc.

The quality of approximation is measured in terms of the Kullback divergence (also called Kullback-Leibler diver-

gence or relative entropy [24,25]) from the target,

$$K(\pi \parallel q^t) = \int \log\left(\frac{\pi(x)}{q^t(x)}\right)\pi(x)dx, \tag{10}$$

and the density q^t can be adjusted incrementally to minimize this divergence. The importance function should be selected from a family of functions which is sufficiently large to allow for a close match with π but for which the minimization of (10) is computationally feasible. In [16] the authors propose to use mixture densities of the form

$$q^t(x) = q(x; \alpha^t, \theta^t) = \sum_{d=1}^{D} \alpha_d^t \varphi(x; \theta_d^t), \tag{11}$$

where $\alpha^t = (\alpha_1^t, \ldots, \alpha_D^t)$ is a vector of adaptable weights for the D mixture components (with $\alpha_d^t > 0$ and $\sum_{d=1}^{D} \alpha_d^t = 1$), and $\theta^t = (\theta_1^t, \ldots, \theta_D^t)$ is a vector of parameters which specify the components; φ is a parametrized probability density function, usually taken to be multivariate Gaussian or Student-t (where the latter is to be preferred in cases where it is suspected that the tails of the posterior π are indeed heavier than Gaussian tails). Given the vast array of densities that can be approximated by mixtures, such an importance function provides considerable flexibility to efficiently estimate a wide range of posteriors, including in this case those found in cosmological settings. Another benefit of using such mixture models is that their parameters are easily reestimated to minimize (10).

The generic PMC algorithm then consists of the following:

Population Monte Carlo algorithm

Do: Choose an importance function q^1.
Generate an independent sample $x_1^1, \ldots, x_N^1 \sim q^1$.
Compute the importance weights $w_1^1, \ldots, w_N^1$.
For $t \geq 1$:
Update the importance function to q^{t+1}, based on the previous weighted sample $(x_1^t, w_1^t), \ldots, (x_N^t, w_N^t)$.
Generate independently $x_1^{t+1}, \ldots, x_N^{t+1} \sim q^{t+1}$.
Compute the importance weights $w_1^{t+1}, \ldots, w_N^{t+1}$.

Unlike for MCMC, in a PMC approach, the process can be interrupted at any time as the sample produced at each iteration can be validly used to approximate expectations under π using self-normalized importance sampling following (7). Further, sampling outputs from previous iterations can be combined [26,27], and the sample size at each iteration does not necessarily need to be fixed. Both of these properties of PMC can be exploited to improve parameter estimates, either by increasing the coverage of the importance function to the target density or increasing the precision of the approximation for the integral of interest.

Also note that an approximate sample from the *target* density can be obtained by sampling $(x_1^t, \ldots, x_n^t)$ with replacement, using the normalized importance weights $\bar{w}_n^t$. Although this process induces extra Monte Carlo variation, there are a number of methods available which considerably reduce the variation involved (e.g., residual sampling [28] or systematic sampling [29]).

1. Updating the importance function in the Gaussian case

In this section, we particularize the generic PMC algorithm to the case where the importance function consists of a mixture of p-dimensional Gaussian densities with mean μ_d and covariance Σ_d:

$$\varphi(x; \mu_d, \Sigma_d) = (2\pi)^{-p/2}|\Sigma_d|^{-1/2} \times \exp[-\tfrac{1}{2}(x - \mu_d)^{\mathrm{T}}\Sigma_d^{-1}(x - \mu_d)]. \tag{12}$$

Using this importance function for the mixture model (11), we start the PMC algorithm by arbitrarily fixing the mixture parameters $(\alpha^1, \mu^1, \Sigma^1)$, and then sample independently from the resulting importance function q^1 to obtain our initial sample $(x_1^1, \ldots, x_N^1)$. From this stage, updates of the parameters proceed recursively.

At iteration t, the importance weights associated with the sample $(x_1^t, \ldots, x_N^t)$ are given by

$$w_n^t = \frac{\pi(x_n^t)}{\sum_{d=1}^{D} \alpha_d^t \varphi(x_n^t; \mu_d^t, \Sigma_d^t)} \tag{13}$$

with normalized counterparts $\bar{w}_n^t$ given by Eq. (8). The parameters (α^t, μ^t and Σ^t) of the importance function are then updated according to

$$\alpha_d^{t+1} = \sum_{n=1}^{N} \bar{w}_n^t \rho_d(x_n^t; \alpha^t, \mu^t, \Sigma^t), \tag{14}$$

$$\mu_d^{t+1} = \frac{\sum_{n=1}^{N} \bar{w}_n^t x_n^t \rho_d(x_n^t; \alpha^t, \mu^t, \Sigma^t)}{\alpha_d^{t+1}}, \tag{15}$$

$$\Sigma_d^{t+1} = \frac{\sum_{n=1}^{N} \bar{w}_n^t (x_n^t - \mu_d^{t+1})(x_n^t - \mu_d^{t+1})^{\mathrm{T}} \rho_d(x_n^t; \alpha^t, \mu^t, \Sigma^t)}{\alpha_d^{t+1}}, \tag{16}$$

where

$$\rho_d(x; \alpha, \mu, \Sigma) = \frac{\alpha_d \varphi(x; \mu_d, \Sigma_d)}{\sum_{d=1}^{D} \alpha_d \varphi(x; \mu_d, \Sigma_d)}. \tag{17}$$

The appendix provides derivations of these expressions and further details on the general approach, as well as equations pertaining to the (more involved) case of mixtures of multivariate Student-t distributions, which are used in the simulations presented in Sec. III.

As discussed in the appendix, the main theoretical appeal of this particular update rule is that, as N tends to

infinity, the corresponding Kullback divergence $K(\pi \parallel q^{t+1})$ is guaranteed to be less than $K(\pi \parallel q^t)$.

2. Monitoring convergence

The above update process can be repeated a number of times, and although there is no need for a formal stopping rule, some measures of performance against the target density can be used as a guide. As the objective of importance function adaptations is to minimize the Kullback divergence between the target density and the importance function, we can stop the process when further adaptations do not result in significant improvements in $K(\pi \parallel q^t)$. To this end, it can be shown that $\exp[-K(\pi \parallel q^t)]$ may be estimated by the *normalized perplexity*

$$p = \exp(H_{\mathrm{N}}^t)/N, \tag{18}$$

where

$$H_{\mathrm{N}}^t = -\sum_{n=1}^{N} \bar{w}_n^t \log \bar{w}_n^t \tag{19}$$

is the Shannon entropy of the normalized weights, a frequently used measure of the quality of an importance sample. Thus, minimization of the Kullback divergence can be approximately connected with the maximization of the perplexity (18). Values of this criterion close to 1 will therefore indicate good agreement between the importance function and the target density.

Another frequently used criterion for importance sampling is the so-called *effective sample size* (ESS),

$$\mathrm{ESS}_{\mathrm{N}}^t = \left(\sum_{n=1}^{N} \{\bar{w}_n^t\}^2\right)^{-1}, \tag{20}$$

which lies in the interval $[1; N]$ and can be interpreted as the number of sample points with nonzero weight [30]. Both measures (18) and (20) are interconnected, as an importance function which is close to the target density will have both a high normalized perplexity and a relatively large number of points with nonzero weight, compared to an ill-fitting importance function. Given a real-valued function f of interest one can also estimate the asymptotic variance of the self-normalized importance sampling estimator $\hat{\pi}_{\mathrm{N}}^t(f) = \sum_{n=1}^{N} \bar{w}_n^t f(x_n^t)$ [cf. Eq. (7)] using the importance sample itself, as

$$N \sum_{n=1}^{N} \{\bar{w}_n^t (f(x_n^t) - \hat{\pi}_N^t(f))\}^2. \tag{21}$$

Beware that this formula (which is derived from Theorem 2 of [31]) is only valid with normalized weights, and that it is a variance conditional on the current importance proposal q^t; i.e., it does not take into account the adaptation. This measure can be related to the so-called *integrated autocorrelation time* used for Markov chain Monte Carlo simulations, which, in this case, takes into consideration the level of autocorrelation present in the chain [5,20,32].

3. A first illustration

To illustrate the PMC approach, we explore a bananalike target density presented Fig. 1. The same target distribution will be studied in greater depth in the next section. The results of the first 11 iterations of the PMC algorithm using a mixture of Student-t densities are shown Fig. 2 (see also the appendix for details of the update procedure).

While this target density shows slightly more pronounced curvature for an example of a posterior density in practice, it serves to illustrate the process of adaptation of the importance function. The initial importance function q^1 is a mixture of multivariate Student-t's, consisting of nine components placed around the center of the range for the first two variables, each with a relatively large variance (for the first dimension = 200 and the second = 50) and degrees of freedom $\nu = 9$. The different colored circles in Fig. 1 indicate the location of the component means, and the circle size is proportional to the weight α_d associated with the component. At the fourth iteration ($t = 4$), we see that the importance function starts to resemble the shape of the target density, with components becoming more separated and moving into the tails of the target. By the sixth iteration ($t = 6$) the importance function has further adapted to the shape of the target banana density. For this target density and importance function, Fig. 3 presents estimates of the normalized perplexity and normalized effective sample size (ESS/N) for the first 10 iterations over 500 simulation runs. As shown, the estimates of the normalized perplexity improve rapidly from approxi-

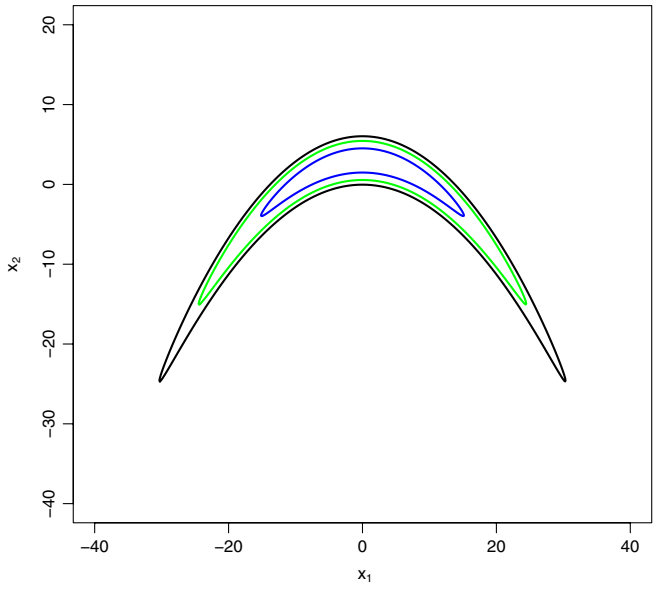

FIG. 1 (color online). Test target density on the (x_1, x_2) plane. Contours represent the 68.3% (blue), 95% (green) and 99% (black) confidence regions.

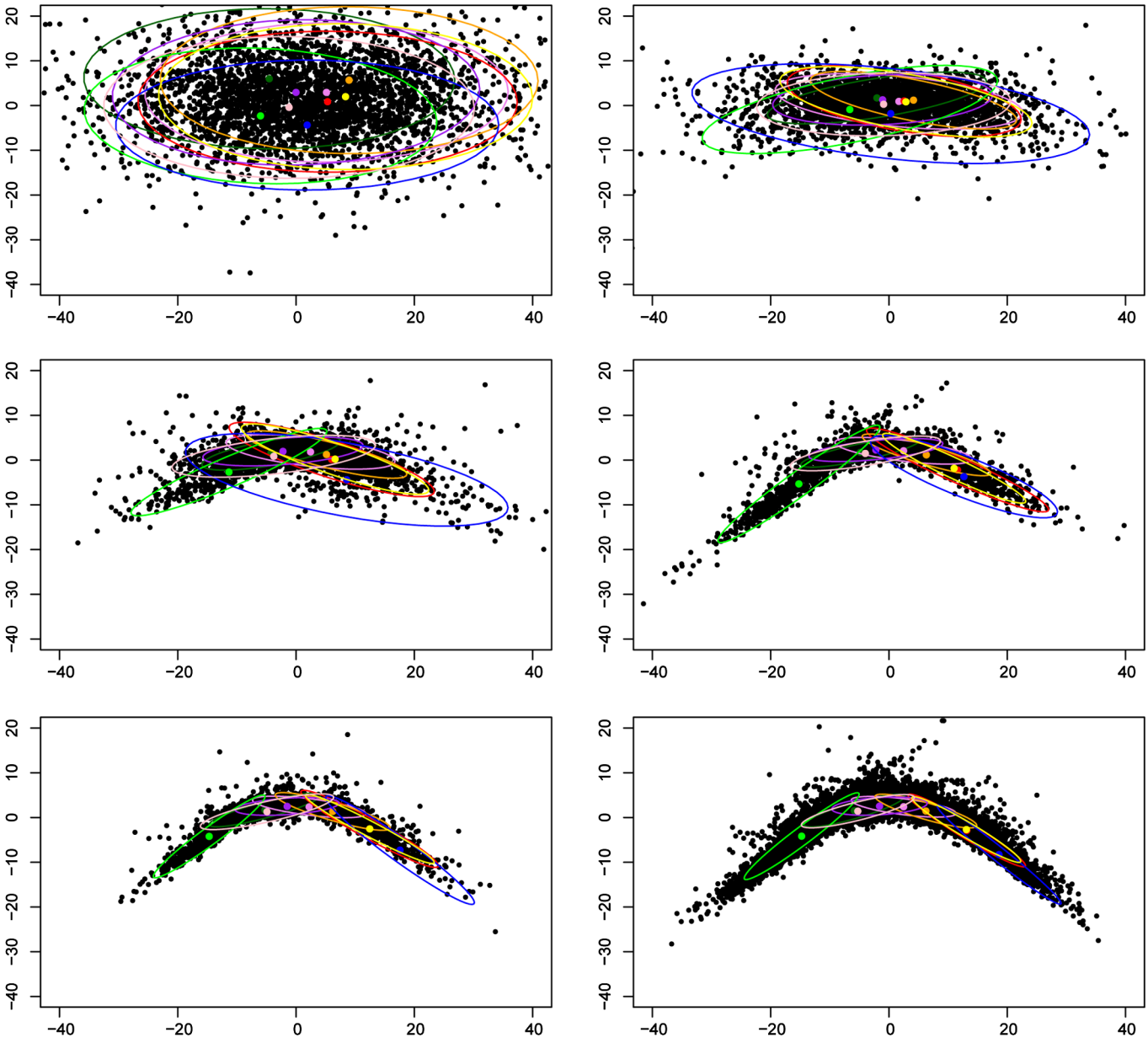

FIG. 2 (color online). Evolution of the importance function for the target density (see Fig. 1) over 11 iterations of 10 000 points for x_1 (horizontal axis) and x_2 (vertical axis), except for the last iteration (11) which is a sample of 100 000 points. Iterations 1 (top left) to 11 (bottom right) from left to right with every second iteration shown (i.e., 1, 3, 5, 7, 9, 11). Colors indicate the mixture components with mean of each component indicated by colored dots and approximate 95% confidence regions for the sample of points from each component by colored ellipses. Every 3rd sample point from the importance function is plotted.

mately 0.14 for the second importance function (iteration 2) to approximately 0.81 for the last importance function (iteration 10), with a similar increase in estimates of the normalized effective sample size (ESS$/N$, increasing from 0.10 to 0.60). For this importance function and target density, the normalized perplexity starts to level off after the 10th iteration (around 0.82), indicating that there is no need for further adaptation of the proposal density. As mentioned previously however, in general, one does not need to observe the convergence of the proposal (as for MCMC) in order to stop the sampling process.

An important consideration, and a choice that needs to be made at the start of the algorithm, are the parameter values α^1, μ^1 and Σ^1 for the initial importance function q^1, including the degrees of freedom ν in the case of the Student-t mixture, and the sample size N. A poor initial importance function, such as one that is tightly centered around only one mode in the case of a multimodal posterior or a narrow importance function with light tails, may take a long time to adapt or may miss important parts of the posterior. For importance sampling the choice of q requires both fat tails and a reasonable match between q and the

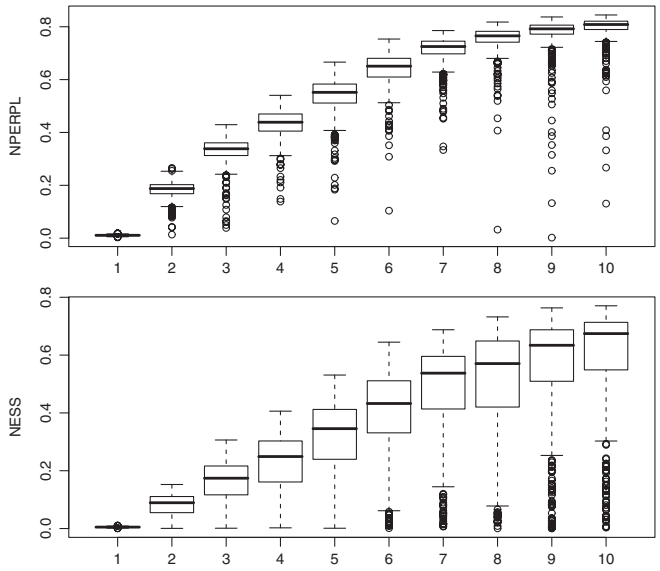

FIG. 3. Normalized perplexity (top panel) and normalized effective sample size (ESS/N) (bottom panel) estimates for the first 10 iterations of PMC (represented in Fig. 2) over 500 simulation runs. The distributions are shown as whisker plots: the thick horizontal line represents the median; the box shows the interquartile range (IQR), containing 50% of the points; the whiskers indicate the interval $1.5 \times$ IQR from either $Q1$ (lower) or $Q3$ (upper); points outside the interval $[Q1, Q3]$ (outliers) are represented as circles.

target π in regions of high density. Such an importance function can be more easily constructed in the presence of a well informed guess about the parameters and possibly the shape of the posterior density. Sample size considerations also play an important role—smaller samples can adapt quite quickly with less computational time but may provide less reliable information about the posterior density relative to larger samples. Such considerations are important as we look at posterior densities of increasingly high dimensions, and thus we can expect to take a larger sample size as the dimension of the problem increases. We will discuss these issues further in the context of simulated and actual data, and also in Sec. V.

III. SIMULATIONS

In this section, we test the performance of PMC using simulated data, and compare the results to an adaptive MCMC procedure.

A. Target density

In order to provide a good test for both approaches we use the target density considered in [19], which is difficult to explore but which also provides a realistic scenario for many problems encountered in cosmology. The target density is based on a centered p-multivariate normal, $x \sim \mathcal{N}_p(0, \Sigma)$ with covariance $\Sigma = \mathrm{diag}(\sigma_1^2, 1, \ldots, 1)$, which is slightly twisted by changing the second coordinate x_2 to $x_2 + b(x_1^2 - \sigma_1^2)$. Other coordinates are unchanged. We obtain a twisted density which is centered with uncorrelated components. Since the Jacobian of twist is equal to 1, the target density is

$$(x_1, x_2 + b(x_1^2 - \sigma_1^2), x_3, \ldots, x_p) \sim \mathcal{N}_p(0, \Sigma). \quad (22)$$

For the target density that we will consider, we set $p = 10$, $\sigma_1^2 = 100$, and $b = 0.03$, which results in a banana-shaped density in the first two dimensions (see Fig. 1).

For the target density considered, interest is in how well PMC and MCMC are able to approximate the tails of the target. While the curvature present in the first two dimensions of this target density is slightly more pronounced than what is typically seen in practice for cosmology it serves to highlight the difficulties faced by both PMC and MCMC in covering the parameter space. In particular, little accurate information is available in order to guide the choice of importance function (for PMC) and proposal distribution (for MCMC) and so both approaches are forced to learn about the parameter space.

B. Test run proposal for PMC

For PMC, and in the absence of any detailed *a priori* information about the target density, except the possible range for each variable, we have chosen the first importance function to be a mixture of multivariate Student-t distributions with components displaced randomly in different directions slightly away from the center of the range for each variable: the mean of the components is drawn from a p-multivariate Gaussian with mean 0 and covariance equal to $\Sigma_0/5$, where Σ_0 is some positive-definite matrix; the variance for components was chosen to be Σ_0. We choose a mixture of 9 components of Student-t distribution with $\nu = 9$ degrees of freedom, and Σ_0 is a diagonal matrix with diagonal entries $(200, 50, 4, \ldots, 4)$. This choice of (ν, Σ_0) ensures adequate coverage, albeit somewhat overdispersed, of the feasible parameter region. In this simulated example, Student-t distributions are preferable to Gaussian distributions because the range of the variables is unbounded (in contrast to the cosmology examples to be discussed in Sec. IV).

A representation of the first importance function for the first two dimensions is shown in the top left-hand box in Fig. 2, with a typical evolution over the next few iterations in the other panels. In pilot runs of various importance functions against the target density, the best fitting importance function required at least seven components in order to adequately represent the coverage of the entire density.

For PMC, an important issue is the sample size for each iteration. A poor initial importance function with a relatively small sample size will take a long time to adapt or it may even be unable to recover sufficiently to provide reasonable parameter estimates. Such problems are more likely to occur as the dimension of the parameter space

increases, the so-called curse of dimensionality. For the simulation exercise each iteration is based on a sample of 10 000 points. To prevent numerical instabilities in the updating of the parameters, components with a very small weight (< 0.002) or containing less than 20 sample points are discarded in the next iteration of the algorithm.

C. Test run proposal for MCMC

As little information is available for the target density, an adaptive MCMC approach is used which can allow for faster learning of the target density than using either independent or nonadaptive random walk proposals [33]. For MCMC, the proposal distribution is a centered Gaussian with covariance matrix which is updated along the iterations. An important choice for adaptive MCMC concerns the scaling of the proposal and the rate of adaptation. There has been much research on this [33,34], and a common choice for the covariance of the Gaussian proposal is to consider $c\Sigma_n$, where Σ_n is an estimate of the covariance of the target density, at update n. The choice $c = 2.38^2/p$ is considered to be optimal when the chain is in its stationary regime, and for target densities that have a product form [33]. However for the target density we consider this does not hold: the first two components are not independent despite being uncorrelated and dependence is not linear but quadratic. However, with no other theoretical results to follow we start with a scaling factor of that form and for the simulation results to follow assess the effect on convergence and results using alternative values. We update the covariance matrix by the recursive formula

$$\Sigma_n = (1 - a_n)\Sigma_{n-1} + a_n S_n, \tag{23}$$

where Σ_{n-1} is the sample covariance of the previous update, and S_n is the covariance of the sampled estimates from the previous update to the current iteration. The value of a_n is $1/n^k$ with k chosen suitably to allow for a cooling of the update, which is a necessary condition to ensure convergence of this adaptive MCMC to the target density as well as convergence of the empirical averages [34,35].

In pilot runs, we explored the effect of this schedule for various values of (k, c) in $(0, 1) \times (0, 2.38^2/p)$ and we observe that the choice of (k, c) plays a role on the time to convergence (for the estimation of the quantities of interest, see below) and on the acceptance rate of the chain.

To ensure a fair comparison with PMC, we start the chain at a random point drawn from the same Gaussian distribution as for PMC [i.e., $\mathcal{N}_d(0, \Sigma_0/5)$, using the same values for Σ_0 as used for PMC]. We also explored in pilot runs the role of the initial value of the chain: despite it being known that MCMC is sensitive to the choice of the initial position of the chain—which has no real counterpart in PMC—this has not been found to have a major impact on performance (for a reasonable choice of the initial value at least) in this particular study. We also fixed the update schedule to be every 10 000 points and we assessed the effect on the results from using less or more points before updating. For the simulation results to follow, (k, c) has been set to $(0.5, 2.38^2/p)$ which ensures convergence after the burn-in period (see Sec. III C 1), and a mean acceptance rate at convergence of about 10%. The proposal distribution is updated for the entire length of the chain and is not stopped after the burn-in period.

1. Test runs

For PMC and the proposal outlined, the perplexity appeared to level off at around the 10th iteration, so for the results to follow for PMC we ran the PMC algorithm for 10 iterations (10 000 points per iteration) and used a final draw of 100 000 points. To assess MCMC for the same number of points we used a chain length of 200 000 points with a burn-in of 100 000 points. Results for both approaches at successive intervals before 200 000 points are also provided. To assess the performance of the approaches, each simulation was replicated 500 times.

2. Results of the simulations

For the results of the simulations, we are interested in both the mean estimates of the parameters (in particular, for x_1 and x_2) and also estimates of the confidence region coverage (68.3% and 95%) which will provide an indication of how well both approaches are covering the tails of the target density. For each run $r = 1, \ldots, 500$, we provide the results for various functions f of interest:

$$f_a(\mathbf{x}) = x_1, \qquad f_b(\mathbf{x}) = x_2,$$
$$f_c(\mathbf{x}) = \mathbf{1}_{68.3}(\mathbf{x}), \qquad f_d(\mathbf{x}) = \mathbf{1}_{95}(\mathbf{x}),$$
$$f_e(\mathbf{x}) = \mathbf{1}_{68.3}(x_1, x_2), \qquad f_g(\mathbf{x}) = \mathbf{1}_{95}(x_1, x_2),$$
$$f_h(\mathbf{x}) = \mathbf{1}_{68.3}(x_1), \qquad f_i(\mathbf{x}) = \mathbf{1}_{95}(x_1).$$

We note here $\mathbf{1}_q$ as the indicator function for the $q\%$ region. f_h and f_i are indicators only for the first dimension, while f_e and f_g are dealing with the first 2. In all cases, the remaining dimensions are marginalized over.

Table I shows the results for estimation of $\pi(f)$ for functions f_a and f_b ($\bar{x}_1$ and $\bar{x}_2$, respectively). The results provided show the mean and standard deviation of estimates calculated over 500 runs. Although the performance is quite similar for both methods, PMC does display a twofold reduction in standard deviation compared to MCMC for both functions. A closer look at the results reveals that for $\pi(f_a)$ the empirical distributions of the estimates (see Fig. 4) are quite similar for both methods, except for the variance which is much reduced for PMC. For $\pi(f_b)$, on the other hand, the empirical distribution of the estimates for PMC are quite skewed, resulting in a slight positive bias for the majority of the runs (second panel of Fig. 4). The difference between $\pi(f_a)$ and $\pi(f_b)$ can be explained from Fig. 1 which shows that failure to visit sufficiently the downward low-probability tails does

indeed imply a positively biased estimate for the mean of the second component. PMC does appear to be more sensitive to this issue than MCMC, despite the fact that the estimates for MCMC display a larger overall variability.

Figure 5 provides the results for the confidence region coverage. To depict the variability of the data, the results are displayed by using whisker plots. The results from both PMC and MCMC against all of the performance measures are similar, with both showing good coverage of the target density. The distribution of this estimator is again more skewed for PMC than it is for MCMC, particularly for the 95% regions in the bottom panel of Fig. 5. Nevertheless, the variability of the estimates obtained with PMC also is significantly reduced compared to MCMC.

Figure 4 shows the evolution of the results for $\pi(f_a)$ and $\pi(f_b)$ from 10 000 points to 100 000 points for both PMC (left panels) and MCMC (right panels). The results from Fig. 4, in general, highlight the reduction in variance of the Monte Carlo estimates for PMC in comparison to MCMC. In particular, it is interesting to note that the variance of the estimates, either for $\pi(f_a)$ or $\pi(f_b)$ for 100 000 posterior evaluations under MCMC, is comparable to estimates obtained using PMC at the second iteration (20 000 points).

Simulating from this target distribution is a challenging problem for both methods. In particular, the use of a vague initial importance function in a multidimensional space represents a challenge to PMC and it has been observed that the importance function takes some time to properly adapt to the target density (about 10 iterations). The choice of the initial importance function in PMC is more crucial than is the choice of the initial proposal distribution in adaptive MCMC. Although different variations for updating the covariance matrix for the MCMC approach are possible we did not see a significant improvement in the results presented from using alternative covariance structures. For most of the simulation results, the proposal covariance matrix was observed to adapt relatively quickly to the true covariance matrix. Changes to the covariance structure considered included changes to the update frequency, the starting proposal Σ_0, the scaling of the proposal (value of c) and adaptation of the covariance (value of k). Hence, the PMC approach may require more precise *a priori* knowledge of the target density than MCMC.

TABLE I. Results of the simulations for the ten-dimensional banana-shaped target density over 500 runs for both PMC and MCMC.

		PMC	MCMC
$\pi(f_a)$	Mean	0.097	−0.028
$\pi(f_a)$	Std	0.218	0.536
$\pi(f_b)$	Mean	0.013	0.002
$\pi(f_b)$	Std	0.163	0.315
Acceptance			0.11
Perplexity		0.80	

In the next section, we apply the PMC approach to typical cosmological examples, and provide results in comparison to MCMC.

IV. APPLICATION TO COSMOLOGY

We apply our new adaptive importance sampling method to the posterior of cosmological parameters. Flat CDM models with either a cosmological constant (ΛCDM) or a constant dark-energy equation-of-state parameter (wCDM) are explored and tested with recent observational data of CMB anisotropies, supernovae type Ia and cosmic shear, as described in the next section.

The three data sets and likelihood functions used here are the same as in [36]; the CMB measurements and likelihood are based on the five-year WMAP data release [37], the SNIa data set is the first-year SNLS survey [38], while the cosmic shear is from the CFHTLS-Wide third release [T0003, [39]]. The results presented in the following sections can be compared to the MCMC analysis in [36].

A. Data sets

1. CMB

To obtain theoretical predictions of the CMB temperature and polarization power and cross spectra we use the publicly available package CAMB [6]. The likelihood is calculated using the public WMAP5 code [4].

The WMAP5 likelihood takes as input the theoretical temperature (TT), polarization (EE and BB) and cross spectra (TE) calculated by CAMB, and returns a likelihood computed from a sum of different parts. It computes a pixel-based Gaussian likelihood based on template-cleaned maps [40] and their associated inverse covariance matrices (see Page *et al.* [41] for details) at large angular scales ($\ell \leq 32$ for TT, $\ell \leq 23$ for TE, EE and BB). At small angular scales, it computes an approximate likelihood based on pseudospectra and their associated covariance for TT and TE [42], based, respectively, on the (Q, V) and (V, W) channel pairs for TE and TT.

In addition, the likelihood computation takes into account analytic marginalizations on nuisance parameters such as the beam transfer function and point-source uncertainties [42,43]. We ignore corrections due to Sunyaev-Zel'dovich and impose a larger (flat) prior on the Hubble constant. Indeed, CMB data alone exhibit a degeneracy between the Hubble constant and, e.g., the cosmological constant [44] which is removed by adding other cosmological probes.

The acoustic oscillation peaks in the CMB anisotropy spectrum are a standard ruler at a redshift of about $z = 1100$. CMB therefore measures the angular diameter distance to that redshift which depends mainly on the total matter-energy density ($\Omega_m + \Omega_{de}$) and weakly on the Hubble constant h. The overall anisotropy amplitude is determined by the large-scale normalization Δ^2_R. The rela-

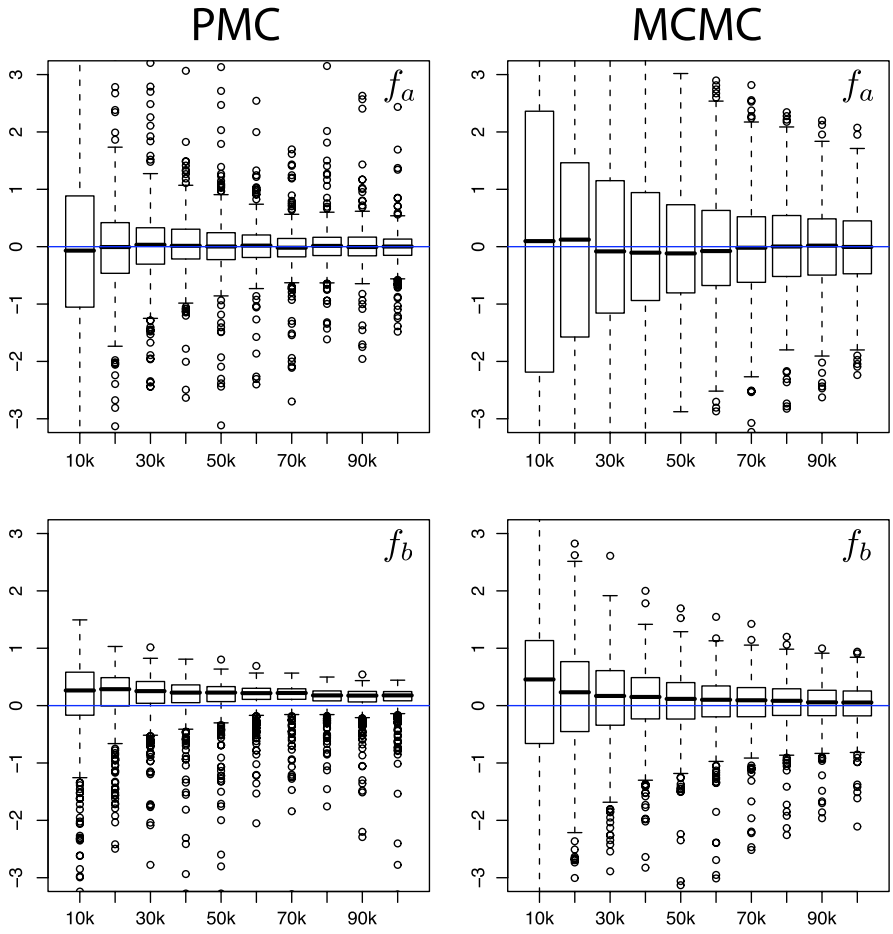

FIG. 4 (color online). Evolution of $\pi(f_a)$ (top panels) and $\pi(f_b)$ (bottom panels) from 10 000 points to 100 000 points for both PMC (left panels) and MCMC (right panels). See Fig. 3 for details about the whisker plot representation.

tive height of the peaks is sensitive to the baryonic and dark matter densities. On large scales, secondary anisotropies are generated at late times ($z \lesssim 20$) due to reionization, which is parametrized by the optical depth τ, and the integrated Sachs-Wolfe effect, which is a probe of $\Omega_{\rm de}$.

2. SNIa

The SNLS data set is described in detail in [38]. We use their results from the SNIa light-curve fits which, for each supernova, provides the rest-frame B-band magnitude m_B^*, the shape or stretch parameter s and the color c. We use the standard likelihood analysis described in [36], adopted from [38].

Under the assumption that supernovae of type Ia are standard candles we can fit the luminosity distance to the SNIa data. The luminosity distance is a function of $\Omega_{\rm m}$, $\Omega_{\rm de}$ and w. Three additional parameters are the universal SNIa magnitude M and the linear response parameters to stretch and color, α and β, respectively. Those three parameters are specific to our choice of distance estimator, and can be regarded as nuisance parameters. The Hubble constant h is integrated into the parameter M, so there is no explicit dependence on h in the SNIa posterior.

3. Cosmic shear

The CFHTLS-Wide 3rd year data release (T0003), the data and weak lensing analysis as well as cosmological results are described in [39]. As in [39] we use the aperture-mass dispersion between 2 and 230 arc minutes as a second-order lensing observable [45]. We assume a multivariate Gaussian likelihood function and take into account the correlation between angular scales. The theoretical aperture-mass dispersion is obtained by nonlinear models of the large-scale structure [46]. This has been calibrated with a ΛCDM cosmology but also provides good fits to wCDM models [47].

The galaxy redshift distribution is obtained by using the CFHTLS-Deep redshift distribution [48] and rescaling it according to the i_{AB} magnitude distribution of CFHTLS-Wide galaxies. We fit the resulting histogram with Eq. (14) from [39], introducing the three fit parameters a, b, and c. The histogram data is modeled as multivariate, uncorrelated Gaussian, the corresponding likelihood is included, independent of the lensing likelihood, in the analysis.

Weak gravitational lensing by the large-scale structure is sensitive to the angular diameter distance and the amount of structure in the Universe. It is an important probe to

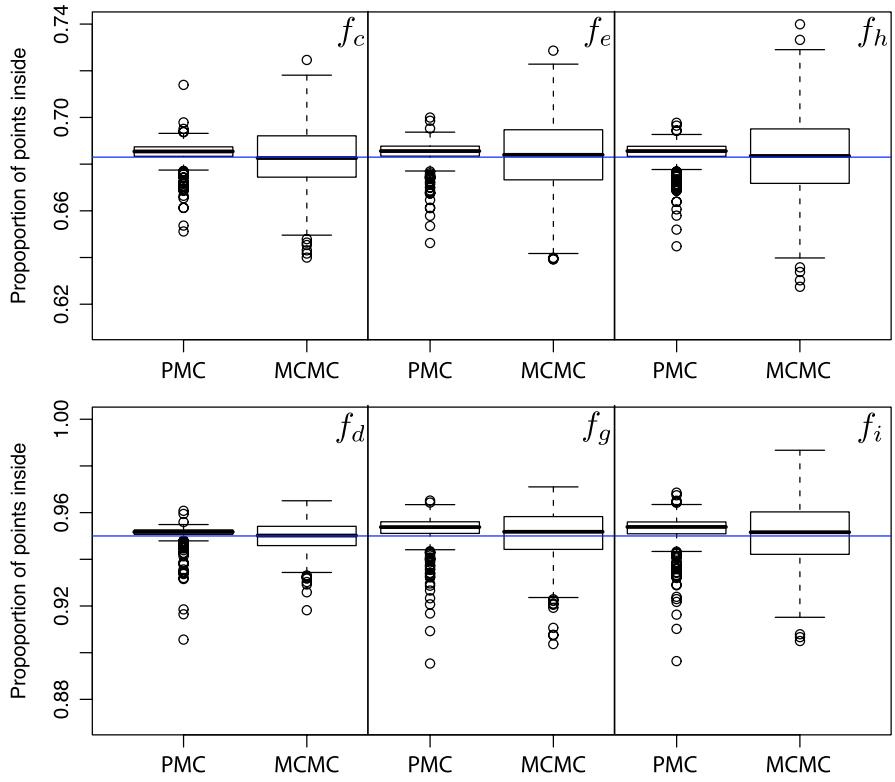

FIG. 5 (color online). Results showing the distributions of the PMC and the MCMC estimates $\pi(f)$ for (top) $f = f_c, f_e, f_h$ and (bottom) $f = f_d, f_g, f_i$ (in this order, left to right). All estimates are based on 500 simulation runs. See Fig. 3 for details about the whisker plot representation.

measure the normalization σ_8 on small scales. With the current data, this parameter is however largely degenerate with Ω_m. This degeneracy is likely to be lifted by future surveys which will include the measurement of higher-order statistics [49,50] and shear tomography [51]. In particular, from the latter a great improvement on the determination of w is to be expected, a parameter which is only weakly constrained by lensing up to now [52,53].

B. Cosmological parameter and priors

We sample a hypercube in parameter space which corresponds to flat priors for all parameters; see Table II for

TABLE II. Parameters for the cosmology likelihood. C = CMB, S = SNIa, L = lensing.

Symbol	Description	Minimum	Maximum	Experiment		
Ω_b	Baryon density	0.01	0.1	C		L
Ω_m	Total matter density	0.01	1.2	C	S	L
w	Dark-energy eq. of state	−3.0	0.5	C	S	L
n_s	Primordial spectral index	0.7	1.4	C		L
Δ_R^2	Normalization (large scales)			C		
σ_8	Normalization (small scales)[a]			C		L
h	Hubble constant			C		L
τ	Optical depth			C		
M	Absolute SNIa magnitude				S	
α	Color response				S	
β	Stretch response				S	
a						L
b	Galaxy z-distribution fit					L
c						L

[a]For WMAP5, σ_8 is a deduced quantity that depends on the other parameters.

more details. Additional priors exist, both in explicit and implicit form, which represent regions of parameter space which are unphysical or where numerical fitting formulas break down. For example, we exclude extremely high baryon fractions ($\Omega_b > 0.75\Omega_m$) because of numerical problems in the computation of the transfer function. Further, for very low values of both Ω_m and σ_8 the pivot scale for the nonlinear power spectrum is outside the allowed range. Very rarely, the calculation of the likelihood for individual points in parameter space is unsuccessful because of numerical errors or limitations of the likelihood code. Since these points cannot be taken into account, a pragmatic solution is to formally modify the prior to exclude those points. Note that these rare cases occur mainly in regions of very low likelihood.

C. Initial choice of the importance function

As described earlier in Sec. II C 3, it is important to have a good guess for the initial importance function. In all cases considered here, we rely on an estimate of the maximum-likelihood point and the Hessian at that point (Fisher matrix) to build our initial proposals. We use the conjugate-gradient approach [54] to find the maximum-likelihood point at which to calculate the Fisher matrix F using the theoretical model. We construct a mixture model consisting of D Gaussian components. Student-t mixtures with small degrees of freedom were tested and turned out to be a poorer approximation to the posterior under study, resulting in lower perplexities. Each mixture component is shifted randomly from the maximum by a small amount. A random scaling is applied to the covariance of each component; i.e., the eigenvectors and ratios between the eigenvalues of the covariance are the same as the ones of the Fisher matrix.

We obtain good results for shifts of about 0.5% to 2% of the box size. Here, a trade-off between too large shifts (resulting in low importance weights) and too small shifts (components stay near the maximum, the posterior tails do not get sampled) has to be found. The stretch factor is chosen randomly between typical values of 1 and 2. In some cases, in particular, with high dimensionality, the derivation of the Fisher matrix is not stable and the matrix is numerically singular. In such cases we set the off-diagonal elements of F to zero.

We found a sample size between 7500 and 10 000 points to be adequate for most cases. The number of components D of the initial importance function was chosen between 5 and 10. For the final iteration we used a sample size 5 times that of the initial sample size.

D. Results

1. General performance

The PMC algorithm is reliable and very efficient in sampling and exploring the parameter space. Both the perplexity as well as the effective sample size increase quickly with each iteration (Fig. 6). The perplexity reaches values of 0.95 or more in many cases, although, in particular, for higher dimensional posteriors the final values are lower. Satisfactory results (i.e., yielding consistent mean and marginals compared to MCMC; see below) are obtained for perplexities larger than about 0.6.

The distribution of importance weights gets narrower from iteration to iteration (Fig. 7). Initially, many sampled points exhibit very low weights. After a few iterations, the importance function has moved towards the posterior increasing the efficiency of the sampling.

Our initial mixture model starts with all mixture components close to the maximum-likelihood point. With consecutive iterations the components spread out to better cover the region where the posterior is significant. This can be seen in Figs. 8 and 9.

Compared to traditional MCMC, our new PMC method is faster by orders of magnitude. The time-consuming calculation of the posterior can be performed in parallel and therefore a speedup by a factor of the number of CPUs is obtained. In times where clusters of multicore processors are readily available, this speedup is easily of the order of 100. In addition, MCMC has a low efficiency with typical acceptance rates of 0.25–0.3. The PMC normalized effective sample size in the WMAP5 case is 0.7 which results in a much larger final sample for the same number of posterior calculations of around 150 000.

We emphasize again that with MCMC one can make only limited use of parallel computing since one has to wait for each Markov chain to converge, and because it is not straightforward to combine chains, as mentioned earlier.

2. Comparison with MCMC

The MCMC results we present here are either obtained using the adaptive MCMC algorithm or a classical one. Indeed, as we show in the following, adaptive MCMC can have some issues that a less efficient classical MCMC algorithm can avoid. Apart from those special cases, the

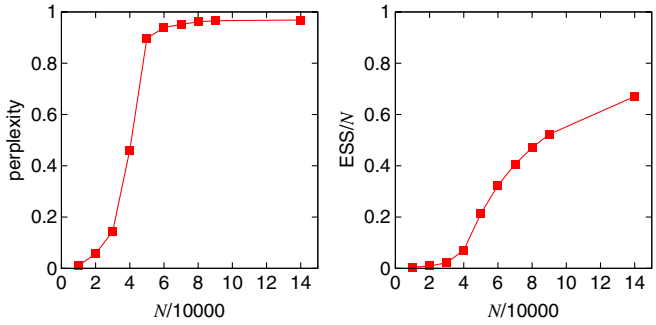

FIG. 6 (color online). Perplexity (left panel) and normalized effective sample size ESS/N (right panel), as a function of the cumulative sample size N. The likelihood is WMAP5 for a flat ΛCDM model with six parameters.

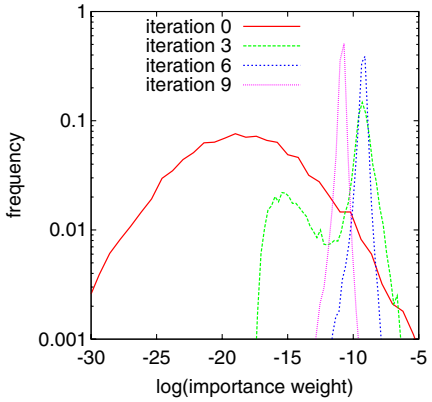

FIG. 7 (color online). Histogram of the normalized importance weights $\bar{w}_n^t$ for four iterations $t = 0, 3, 6, 9$. The posterior is WMAP5, flat ΛCDM model with six parameters.

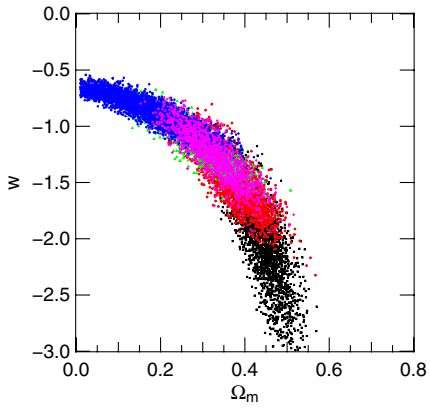

FIG. 9 (color online). The sampled points from the final iteration are plotted; the colors indicate the components of the importance function from which the points are drawn (the colors are the same as in Fig. 8). One out of five points is shown. Note that the density of points does not correspond to the posterior density since the former has to be weighted by the importance weights.

MCMC and adaptive MCMC gave very similar results, the latter usually reaching a better acceptance rate, and thus a better efficiency.

We find excellent agreement between using our respective implementations of MCMC (adaptive or not) and PMC. Mean, confidence intervals and 2D marginals are very similar using both methods. The performance of PMC is superior to MCMC in some cases, which is illustrated by the following examples.

An inherent problem of MCMC is that even for a long run there can be regions in parameter space that are not sampled in an unbiased way. This is illustrated in Fig. 10. The feature at the 99.7% level of MCMC (left panel, for large values of $-M$ and α) is due to an "excursion" of the

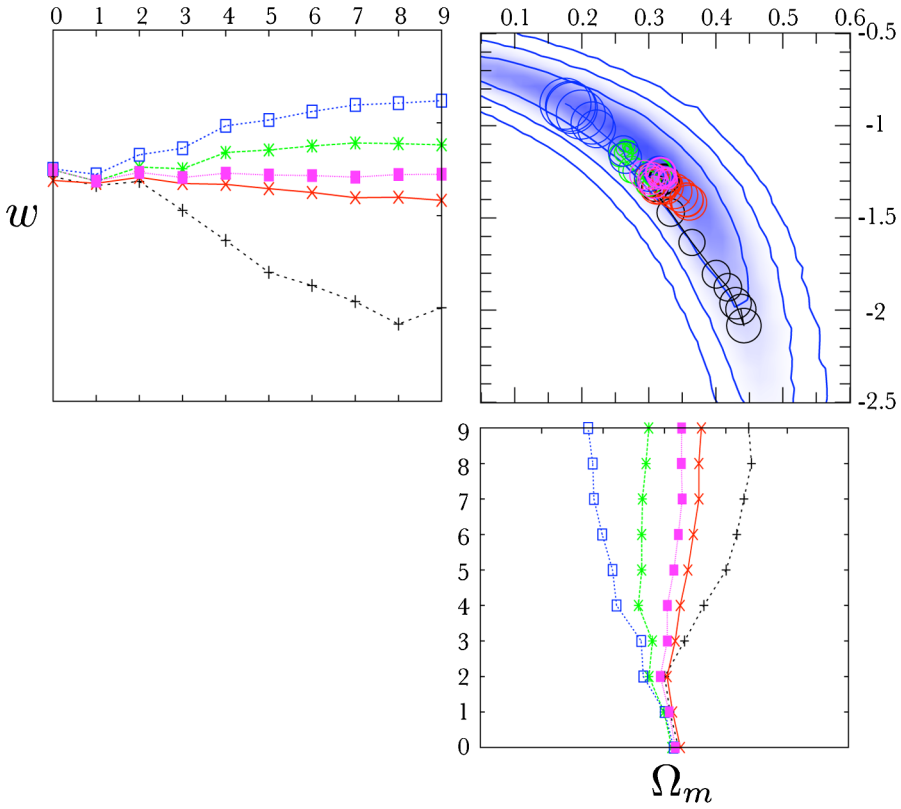

FIG. 8 (color online). Lower left panel: Overlaid to the SNIa confidence contours (68%, 95%, 99.7%) is the movement of the importance function. For each iteration a circle is plotted at the position of the mean of each component, where different colors indicate different components. The circle size indicates the component weight. The starting point [first iteration, at $(0.3, -1.3)$] is marked by a thick circle. The other two panels show the mean positions in projection, fanned out as a function of the iteration.

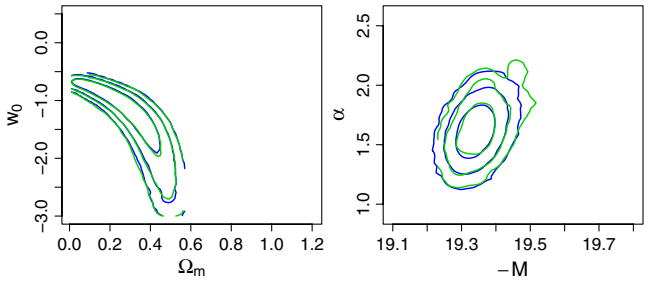

FIG. 10 (color online). Examples of marginalized likelihoods (68%, 95% and 99.7% contours are shown) for PMC (solid blue line) and MCMC (dashed green line) from the SNIa data.

chain into a low-likelihood region at step 130 500, lasting for 300 steps. We ran the chain for 300 000 steps and the feature was still visible. A second run of the chain did not exhibit this anomaly. This kind of sample "noise" can be prevented by running a chain for a very long time or by combining several (converged) chains. Such features are much less likely to occur in an importance sample which consists of uncorrelated points.

A second issue are parameters which are nearly unconstrained by the data with the result that the marginalized posterior in that dimension is flat. To illustrate this we choose weak lensing alone which cannot constrain Ω_b (Fig. 11). Using the Fisher matrix as the initial Gaussian proposal for adaptive MCMC, the chain stays in a small region in the Ω_b direction; the covariance being very flat, most jumps end up out of the prior distribution. This results in an update variance for this parameter which is much too small, and in a bad exploration of the posterior in this flat direction as shown Fig. 11. The classical MCMC algorithm, with the same proposal yields better results, but with a very low acceptance rate and needing 500 000 steps to reach the result presented in Fig. 11. Alternatively, modifying the initial proposal to be smaller and better adapted to the prior, or increasing the covariance stretch factor from the optimal value of $c = 2.38^2/p$ (see Sec. III C) to $c = 3.2^2/p$, helps the chain to explore more of the parameter space in the latter steps of the adaptation. These modifications to the algorithm also result in a very low acceptance rate, and somehow go against the very idea of an adaptive algorithm, since they require very fine-tuning of the initial proposal.

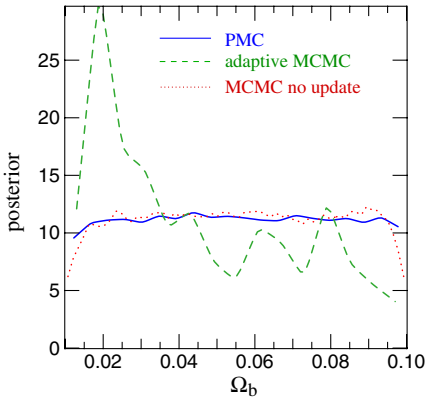

FIG. 11 (color online). Normalized 1D marginals for Ω_b from weak lensing alone for PMC (solid blue line) and MCMC (dashed green line: adaptive; dotted red line: nonadaptive).

TABLE III. Parameter means and 68% confidence intervals for PMC and (nonadaptive) MCMC from the WMAP5 data.

Parameter	PMC	MCMC
Ω_b	$0.04424^{+0.00321}_{-0.00290}$	$0.04418^{+0.00321}_{-0.00294}$
Ω_m	$0.2633^{+0.0340}_{-0.0282}$	$0.2626^{+0.0359}_{-0.0280}$
τ	$0.0878^{+0.0181}_{-0.0160}$	$0.0885^{+0.0181}_{-0.0160}$
n_s	$0.9622^{+0.0145}_{-0.0143}$	$0.9628^{+0.0139}_{-0.0145}$
$10^9\Delta_R^2$	$2.431^{+0.118}_{-0.113}$	$2.429^{+0.123}_{-0.108}$
h	$0.7116^{+0.0271}_{-0.0261}$	$0.7125^{+0.0274}_{-0.0268}$

TABLE IV. Parameter means and 68% confidence intervals for PMC using lensing, SNIa and CMB in combination. The (nonadaptive) MCMC results correspond to the values given in Table 5 from [36].

Parameter	PMC	MCMC
Ω_b	$0.0432^{+0.0027}_{-0.0024}$	$0.0432^{+0.0026}_{-0.0023}$
Ω_m	$0.254^{+0.018}_{-0.017}$	$0.253^{+0.018}_{-0.016}$
τ	$0.088^{+0.018}_{-0.016}$	$0.088^{+0.019}_{-0.015}$
w	-1.011 ± 0.060	$-1.010^{+0.059}_{-0.060}$
n_s	$0.963^{+0.015}_{-0.014}$	$0.963^{+0.015}_{-0.014}$
$10^9\Delta_R^2$	$2.413^{+0.098}_{-0.093}$	$2.414^{+0.098}_{-0.092}$
h	$0.720^{+0.022}_{-0.021}$	$0.720^{+0.023}_{-0.021}$
a	$0.648^{+0.040}_{-0.041}$	$0.649^{+0.043}_{-0.042}$
b	$9.3^{+1.4}_{-0.9}$	$9.3^{+1.7}_{-0.9}$
c	$0.639^{+0.084}_{-0.070}$	$0.639^{+0.082}_{-0.070}$
$-M$	19.331 ± 0.030	$19.332^{+0.029}_{-0.031}$
α	$1.61^{+0.15}_{-0.14}$	$1.62^{+0.16}_{-0.14}$
$-\beta$	$-1.82^{+0.17}_{-0.16}$	-1.82 ± 0.16
σ_8	$0.795^{+0.028}_{-0.030}$	$0.795^{+0.030}_{-0.027}$

With PMC we obtain a much better performance and recover very well the flat posterior.

In Tables III and IV we show the mean and 68% confidence intervals for CMB alone for the ΛCDM model and for lensing + SNIa + CMB using wCDM, respectively. The differences in mean and 68%-confidence intervals is less than a few percent in most cases. Figure 10 shows that the lower-confidence regions and the correlation between parameters agrees very well between (nonadaptive) MCMC and PMC.

V. DISCUSSION

In this paper, we have introduced and assessed an adaptive importance sampling approach, called population Monte Carlo, which aims to overcome the main difficulty in using importance sampling, namely, the reliance on a

single efficient importance function. PMC achieves this goal by iteratively adapting the importance function towards the target density of interest. A significant appeal of the approach, when compared to alternatives such as MCMC, lies in the possibility to use (massive) parallel sampling which considerably reduces the wall-clock time involved in the estimation of parameters for many astrophysical and cosmological problems. Simulated and actual data have been used in this work to assess the performance of PMC for estimation of parameters in a Bayesian inference with features approaching classical cosmological parameter posteriors.

The PMC approach is, in essence, an iterated importance sampling scheme that simultaneously produces, at each iteration, a sample approximately simulated from a target distribution π and an approximation of π in the form of the current proposal distribution. As such, the samples produced by the PMC approach can be exploited as regular importance sampling outputs at any iteration t. Samples from previous iterations can be combined [27], and approximations like $\hat{\pi}(f)$ can be updated dynamically, without necessarily requiring the storage of samples.

Although adaptation of the importance function has the explicit aim of improving the coverage of the posterior density there are instances where this objective may not be met. In some cases, successive updates of the importance function may result in: (a) an importance function which is too peaked and which has light tails (invalid importance function); (b) an importance function which fits only one mode (in the case of a multimodal posterior); (c) numerical problems due to the adaptation procedure (usually involving poor conditioning of some of the covariance matrices). Such cases are likely to produce a poor approximation to the integral of interest, or alternatively lead to highly variable parameter estimates over iterations. These problems can be quickly discovered or signalled by observing a poor ESS, and parameter estimates or normalized perplexity which do not stabilize after a few iterations.

Such cases of poor performance as outlined can be significantly reduced by choosing a reasonably well informed initial importance function with a large enough sample size at each iteration, especially on the initial iteration that requires many points to counterweight a potentially poor importance function. In general, the initial importance function should be chosen to cover a region of the parameter space that has support larger than the posterior. In the absence of reliable prior information, finding such an importance function may be difficult to do. One approach may be to locate the components in the center of the feasible range (if available) for each variable, with reasonably large variances to ensure some coverage of the parameter space. We found this approach to be reasonably successful for the simulated data case discussed in Sec. III. In the presence of some prior information, for example, an estimate of the maximum-likelihood point and an approximation of the covariance matrix (via the Hessian), components can be placed around these points with variance comparable to the approximation. Another approach may be to perform a singular value decomposition of the covariance matrix, and make use of the eigenvectors and eigenvalues to place components along the most likely directions of interest. Alternatively and in the same spirit, components can be placed according to the principal points of the resulting sample, using a k-means clustering approach [55]. Both approaches have been reasonably successful for a range of posterior densities examined, and by placing components in regions of high posterior support in addition to the mode have the potential to further significantly reduce the number of iterations for difficult posterior densities.

The main appeals or advantages of the PMC method are worth re emphasizing at this point:

(1) parallelization of the posterior calculations;
(2) low variance of Monte Carlo estimates;
(3) simple diagnostics of ''convergence'' (perplexity).

We address these three points in more detail now.

(1) The first advantage, namely, the ability to parallelize the computational task, is becoming increasingly useful through the availability of cheap multi-CPU computers and the standardization of clusters of computers. Software to implement the parallelization task, such as MESSAGE PASSING INTERFACE (MPI),[3] are publicly available and relatively straightforward to implement. For the cosmological examples presented (Sec. IV), we used up to 100 CPUs on a computer cluster to explore the cosmology posteriors. In the case of WMAP5, this reduced the wall-clock time from several days for MCMC[4] to a few hours using PMC.

(2) In general, for PMC and an importance function that is closely matched to the target density, significant reductions in the variance of the Monte Carlo estimates are possible in comparison to estimates obtained using MCMC [5]. For example, for the posterior estimates for the WMAP5 data we observed a tenfold reduction in variance for the same number of sample points as observed for MCMC. Such reductions suggest that the wall-clock time savings extend not only to the number of CPUs available but to smaller sample requirements for PMC in total compared to MCMC to achieve similar variability of estimates. For cosmological applications, this observation is valuable as we observed, e.g., in Fig. 6 for CMB data, that the fit between the adapted importance function and the target posterior is sometimes quite good. By combining samples across iterations further time savings are also possible. The absence of construction of a Markov chain

[3]http://www-unix.mcs.anl.gov/mpi/

[4]This represents the time spent by our generic adaptive MCMC code for a similar problem as reported in [36]. A highly cosmology-tuned MCMC code such as COSMOMC can reach better performance by implementing different strategies for each parameters.

for PMC can also have the desirable attribute of reducing sample noise, as observed for the SNIa data in Sec. IV D 2.

(3) As shown in Sec. II C 2, the perplexity [Eq. (18)] is a relatively simple measure of sampling adequacy to the target density of interest. For MCMC and other approaches which rely on formal measures of convergence, assessment of convergence can be very difficult with users facing a potential array of associated diagnostic tests.

In addition to the above points, a further appeal of PMC is the ability to provide a very good approximation to the marginal posterior or evidence, which naturally follows as a by-product of the approach. To demonstrate this appeal, further research is underway to explore the use of PMC in the context of model selection problems in cosmology.

ACKNOWLEDGMENTS

We acknowledge the use of the Legacy Archive for Microwave Background Data Analysis (LAMBDA). Support for LAMBDA is provided by the NASA Office of Space Science. We thank the Planck group at IAP and the TERAPIX group for support and computational facilities. D. W. and M. K. are supported by the CNRS ANR "ECOSSTAT," Contract No. ANR-05-BLAN-0283-04 ANR ECOSSTAT. The authors would like to thank F. Bouchet, S. Bridle, J.-F. Giovannelli, J.-M. Marin, Y. Mellier and I. Tereno for helpful discussions.

APPENDIX: DETAILS OF THE IMPORTANCE FUNCTION UPDATES FOR PMC

The method proposed in [16] to adaptively update the parameters of the importance function q^t is based on a variant of the EM (expectation-maximization) algorithm [56], which is the standard tool for the estimation of the parameters of mixture densities. We describe below the principle underlying the algorithm of [16], showing, in particular, that each iteration decreases, up to the importance sampling approximation errors, the Kullback divergence between the target π and the importance function q^t.

Remember that our goal is to minimize (10), as t increases, by iteratively tuning the parameters α^t and θ^t of the mixture importance function defined in (11). Developing the logarithm in (10), this objective can be equivalently formulated in terms of the maximization of the following quantity:

$$\ell(\alpha, \theta) = \int \log\Big(\sum_{d=1}^{D} \alpha_d \varphi(x; \theta_d)\Big)\pi(x)dx \tag{A1}$$

with respect to α and θ. Using Bayes' rule, we denote by

$$\rho_d(x; \alpha, \theta) = \frac{\alpha_d \varphi(x; \theta_d)}{\sum_{d=1}^{D} \alpha_d \varphi(x; \theta_d)} \tag{A2}$$

the posterior probability that x belongs to the dth component of the mixture (for the mixture parameters α and θ). The EM principle consists of evaluating, at iteration t, the following intermediate quantity:

$$L^t(\alpha, \theta) = \int \sum_{d=1}^{D} \rho_d(x; \alpha^t, \theta^t) \log(\alpha_d \varphi(x; \theta_d))\pi(x)dx. \tag{A3}$$

Using the concavity of the log as well as the expression of ρ_d in (A2), it is easily checked that

$$\sum_{d=1}^{D} \rho_d(x; \alpha^t, \theta^t) \log\Big(\frac{\alpha_d \varphi(x; \theta_d)}{\alpha_d^t \varphi(x; \theta_d^t)}\Big) \le \log\Big(\frac{\sum_{d=1}^{D} \alpha_d \varphi(x; \theta_d)}{\sum_{d'=1}^{D} \alpha_{d'}^t \varphi(x; \theta_{d'}^t)}\Big) \tag{A4}$$

and hence that $L^t(\alpha, \theta) - L^t(\alpha^t, \theta^t) \le \ell(\alpha, \theta) - \ell(\alpha^t, \theta^t)$. Thus, any value of α and θ which increases the intermediate quantity L^t above the level $L^t(\alpha^t, \theta^t)$ also results in, at least, an equivalent increase of the actual objective function ℓ. In the EM algorithm, one sets α^{t+1} and θ^{t+1} to the values where the intermediate quantity $L^t(\alpha, \theta)$ is maximal, thus satisfying the previous requirement. Furthermore, the maximization of $L^t(\alpha, \theta)$ leads to a closed form solution whenever φ belongs to a so-called exponential family of probability densities.

In the example of the multivariate Gaussian density recalled in (12), the parameter θ_d consists of the mean μ_d and the covariance matrix Σ_d and the intermediate quantity may be written as

$$L^t(\alpha, \mu, \Sigma) = \int \sum_{d=1}^{D} \rho_d(x; \alpha^t, \mu^t, \Sigma^t)\Big\{\log(\alpha_d) - \frac{1}{2}[\log|\Sigma_d| + (x - \mu_d)^{\mathrm{T}}\Sigma_d^{-1}(x - \mu_d)]\Big\} \times \pi(x)dx, \tag{A5}$$

up to terms that do not depend on α, μ or Σ. Routine calculations show that the maximum of (A5) is achieved for

$$\alpha_d^{t+1} = \int \rho_d(x; \alpha^t, \mu^t, \Sigma^t)\pi(x)dx, \tag{A6}$$

$$\mu_d^{t+1} = \frac{\int x \rho_d(x; \alpha^t, \mu^t, \Sigma^t)\pi(x)dx}{\alpha_d^{t+1}}, \tag{A7}$$

$$\Sigma_d^{t+1} = \frac{\int (x - \mu_d^{t+1})(x - \mu_d^{t+1})^{\mathrm{T}} \rho_d(x; \alpha^t, \mu^t, \Sigma^t)\pi(x)dx}{\alpha_d^{t+1}}. \tag{A8}$$

In practice, both the numerator and denominator of each of the above expressions are integrals under π which must be approximated. The solution proposed in [16] is based on self-normalized importance sampling using the weighted sample simulated at the previous iteration $(x_1^t, \bar{w}_1^t), \ldots, (x_N^t, \bar{w}_N^t)$. The corresponding empirical update equations are given in Eqs. (14)–(16) of Sec. II C 1.

The Student-t distribution provides a family of multivariate densities with parameters μ and Σ which have the same interpretation as in the Gaussian case [except for the fact that the covariance is equal to $\nu/(\nu-2)\Sigma$ rater than Σ] but with an additional shape factor $\nu \geq 2$ which allows for heavier tails: letting $\nu \to \infty$ yields back the Gaussian but for $\nu = 2$, one obtains a density with polynomially decreasing tails whose only finite moments are the two first ones (note that it is also possible to extend the family to the case where $0 < \nu < 2$). Using mixtures of Student-t distributions will thus be mostly useful in cases where the target posterior distribution π itself has heavy tails. The parameter update corresponding to mixtures of Student-t distributions is a bit more involved but follows the same general pattern. For the sake of completeness, we just recall below the formulas given in [16]:

$$\alpha_d^{t+1} = \sum_{n=1}^{N} \bar{w}_n^t \rho_d(x_n^t; \alpha^t, \theta^t),$$

$$\mu_d^{t+1} = \frac{\sum_{n=1}^{N} \bar{w}_n^t \rho_d(x_n^t; \alpha^t, \theta^t) \gamma_d(x_n^t; \theta^t) x_n^t}{\sum_{n=1}^{N} \bar{w}_n^t \rho_d(x_n^t; \alpha^t, \theta^t) \gamma_d(x_n^t; \theta^t)},$$

$$\Sigma_d^{t+1} = \frac{1}{\sum_{n=1}^{N} \bar{w}_n^t \rho_d(x_n^t; \alpha^t, \theta^t)} \sum_{n=1}^{N} \bar{w}_n^t \rho_d(x_n^t; \alpha^t, \theta^t) \times \gamma_d(x_n^t; \theta^t)(x_n^t - \mu_d^{t+1})(x_n^t - \mu_d^{t+1})^{\mathrm{T}},$$

where

$$\rho_d(x; \alpha, \theta) = \frac{\alpha_d \tau(x; \nu_d, \mu_d, \Sigma_d)}{\sum_{d=1}^{D} \alpha_d \tau(x; \nu_d, \mu_d, \Sigma_d)}, \tag{A9}$$

$$\gamma_d(x_n^t; \theta) = \frac{\nu_d + p}{\nu_d + (x - \mu_d)^{\mathrm{T}} (\Sigma_d)^{-1} (x - \mu_d)}, \tag{A10}$$

and $\tau(\cdot; \mu, \Sigma, \nu)$ denotes the p-dimensional Student-t probability density function

$$\tau(x; \mu, \Sigma, \nu) = \frac{\Gamma((\nu + p)/2)}{\Gamma(\nu/2) \nu^{p/2} \pi^{p/2}} |\Sigma|^{-1/2} \times \left(1 + \frac{1}{\nu}(x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)\right)^{-(\nu+p)/2}. \tag{A11}$$

Sampling from a multivariate Student-t distribution is most easily undertaken by using its derivation in terms of a multivariate Gaussian [$Y \sim N_k(0, \Sigma)$] and chi-squared distribution ($Z \sim \chi^2_\nu$),

$$x = \mu + y\sqrt{\nu/z}$$

and taking advantage of the fact that sampling from Y and Z is straightforward.

[1] A. Lewis and S. Bridle, Phys. Rev. D **66**, 103511 (2002).
[2] Ø. Rudjord, N. E. Groeneboom, H. K. Eriksen, G. Huey, K. M. Górski, and J. B. Jewell, Astrophys. J. **692**, 1669 (2009).
[3] J. F. Taylor, M. A. J. Ashdown, and M. P. Hobson, Mon. Not. R. Astron. Soc. **389**, 1284 (2008).
[4] J. Dunkley *et al.*, Astrophys. J. Suppl. Ser. **180**, 306 (2009).
[5] C. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer-Verlag, New York, 2004), 2nd ed.
[6] A. Lewis, A. Challinor, and A. Lasenby, Astrophys. J. **538**, 473 (2000).
[7] W. A. Fendt and B. D. Wandelt, arXiv:0712.0194.
[8] T. Auld, M. Bridges, M. P. Hobson, and S. F. Gull, Mon. Not. R. Astron. Soc. **376**, L11 (2007).
[9] A. Hajian, Phys. Rev. D **75**, 083525 (2007).
[10] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721 (1984).
[11] D. L. Larson, H. K. Eriksen, B. D. Wandelt, K. M. Gorski, G. Huey, J. B. Jewell, and I. J. O'Dwyer, Astrophys. J. **656**, 653 (2007).
[12] R. Shaw, M. Bridges, and M. P. Hobson, Mon. Not. R. Astron. Soc. **378**, 1365 (2007).
[13] F. Feroz and M. P. Hobson (to be published).
[14] N. Chopin and C. P. Robert, arXiv:0801.3887.
[15] J. S. Rosenthal, Far East J. Theor. Stat. **4**, 207 (2000).
[16] O. Cappé, R. Douc, A. Guillin, J.-M. Marin, and C. Robert, arXiv:0710.4242.
[17] C. Robert, *The Bayesian Choice* (Springer-Verlag, New York, 2001), 2nd ed.
[18] J.-M. Marin, K. Mengersen, and C. Robert, in *Handbook of Statistics*, edited by C. Rao and D. Dey (Springer-Verlag, New York, 2005), Vol. 25.
[19] H. Haario, E. Saksman, and J. Tamminen, Bernoulli **7**, 223 (2001).
[20] G. O. Roberts and J. S. Rosenthal, Stat. Sci. **16**, 351 (2001).
[21] O. Cappé, A. Guillin, J.-M. Marin, and C. Robert, J. Comput. Graph. Stat. **13**, 907 (2004).
[22] J. Von Neumann, Natl. Bur. Stand. Appl. Math. Ser. **12**, 36 (1951).
[23] R. Rubinstein, *Simulation and the Monte Carlo Method* (Wiley, New York, 1981).
[24] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[26] A. Owen and Y. Zhou, J. Am. Stat. Assoc. **95**, 135 (2000).
[27] J.-M. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert (unpublished).
[28] J. Liu and R. Chen, J. Am. Stat. Assoc. **90**, 567 (1995).
[29] D. Whitley, Stat. Comput. **4**, 65 (1994).

[30] J. Liu and R. Chen, J. Am. Stat. Assoc. **90**, 567 (1995).
[31] J. Geweke, Econometrica **57**, 1317 (1989).
[32] J. Dunkley, M. Bucher, P. G. Ferreira, K. Moodley, and C. Skordis, Mon. Not. R. Astron. Soc. **356**, 925 (2005).
[33] G. O. Roberts, A. Gelman, and W. R. Gilks, Ann. Appl. Probab. **7**, 110 (1997).
[34] Y. Atchadé and G. Fort, arXiv:0807.2952.
[35] G. O. Roberts and J. S. Rosenthal, J. Appl. Probab. **44**, 458 (2007).
[36] M. Kilbinger *et al.*, arXiv:0810.5129 [Astron. Astrophys. (to be published)].
[37] G. Hinshaw *et al.*, Astrophys. J. Suppl. Ser. **180**, 225 (2009).
[38] P. Astier *et al.*, Astron. Astrophys. **447**, 31 (2006).
[39] L. Fu, E. Semboloni *et al.*, Astron. Astrophys. **479**, 9 (2008).
[40] B. Gold *et al.*, Astrophys. J. Suppl. Ser. **180**, 265 (2009).
[41] L. Page *et al.*, Astrophys. J. Suppl. Ser. **170**, 335 (2007).
[42] G. Hinshaw *et al.*, Astrophys. J. Suppl. Ser. **170**, 288 (2007).
[43] M. R. Nolta *et al.*, Astrophys. J. Suppl. Ser. **180**, 296 (2009).
[44] G. Efstathiou and J. R. Bond, Mon. Not. R. Astron. Soc. **304**, 75 (1999).
[45] P. Schneider, L. van Waerbeke, B. Jain, and G. Kruse, Mon. Not. R. Astron. Soc. **296**, 873 (1998).
[46] R. E. Smith, J. A. Peacock, A. Jenkins, S. D. M. White, C. S. Frenk, F. R. Pearce, P. A. Thomas, G. Efstathiou, and H. M. P. Couchman, Mon. Not. R. Astron. Soc. **341**, 1311 (2003).
[47] Z. Ma, Astrophys. J. **665**, 887 (2007).
[48] O. Ilbert, S. Arnouts *et al.*, Astron. Astrophys. **457**, 841 (2006).
[49] M. Takada and B. Jain, Mon. Not. R. Astron. Soc. **348**, 897 (2004).
[50] M. Kilbinger and P. Schneider, Astron. Astrophys. **442**, 69 (2005).
[51] W. Hu, Astrophys. J. **522**, L21 (1999).
[52] H. Hoekstra, Y. Mellier, L. van Waerbeke, E. Semboloni, L. Fu, M. Hudson, L. Parker, I. Tereno, and K. Benabed, Astrophys. J. **647**, 116 (2006).
[53] M. Jarvis, B. Jain, G. Bernstein, and D. Dolney, Astrophys. J. **644**, 71 (2006).
[54] W. H. Press, S. A. Teukolsky, B. P. Flannery, and W. T. Vetterling, *Numerical Recipes in C* (Cambridge University Press, Cambridge, England, 1992).
[55] T. Tarpey, Comput. Stat. **22**, 71 (2007).
[56] A. Dempster, N. Laird, and D. Rubin, J. R. Stat. Soc. Ser. B. Methodol. **39**, 1 (1977).