Iterated importance sampling in missing data problems

Gilles Celeux

INRIA, FUTURS, Orsay, France

Jean-Michel Marin *

INRIA, FUTURS, Orsay, France and CEREMADE, University Paris Dauphine, Paris, France

Christian P. Robert

CEREMADE, University Paris Dauphine and CREST, INSEE, Paris, France

Abstract

Missing variable models are typical benchmarks for new computational techniques in that the ill-posed nature of missing variable models offer a challenging testing ground for these techniques. This was the case for the EM algorithm and the Gibbs sampler, and this is also true for importance sampling schemes. A population Monte Carlo scheme taking advantage of the latent structure of the problem is proposed. The potential of this approach and its specifics in missing data problems are illustrated in settings of increasing difficulty, in comparison with existing approaches. The improvement brought by a general Rao–Blackwellisation technique is also discussed.

Key words: Adaptive algorithms, Bayesian inference, latent variable models, population Monte Carlo, Rao–Blackwellisation, stochastic volatility model

* Corresponding author: CEREMADE, Place du Maréchal De Lattre de Tassigny, 75775 Paris Cedex 16, France, marin@ceremade.dauphine.fr

Preprint submitted to Computational Statistics and Data Analysis 25 July 2005

1 Introduction

1.1 Missing data models

Missing data models, that is, structures such that the distribution of the data y can be represented via a marginal density

$$f(y|\theta) = \int_{\mathscr{Z}} g(y, z|\theta) dz \,,$$

where $z \in \mathscr{Z}$ denotes the so-called "missing data", have often been at the forefront of computational Statistics, both as a challenge to existing techniques and as a benchmark for incoming techniques. This is for instance the case with the EM algorithm (Dempster et al., 1977), which was purposely designed for missing data problems although it has since then been applied in a much wider setting. Similarly, one of the first occurrences of Gibbs sampling is to be found in the analysis of mixture models by Tanner and Wong (1987). Besides, these models also stand on their own as valuable tools for representing complex phenomena and deserve appropriately efficient computational support; any true advance in statistical computing must thus be able to increase our ability of using and designing new and more elaborate missing data models.

Many different techniques have been proposed and tested on missing data problems (see, e.g., Everitt, 1984, Little and Rubin, 1987, McLachlan and Krishnan, 1997, Robert and Casella, 1999, Chap. 9) and they often take advantage of the specific features of the corresponding models, mostly through completion devices that simulate or approximate the missing part z of the data. This is not always the case, though, as shown for instance in Celeux et al. (2000) where non-completed proposals are advantageously used in a random walk Metropolis–Hastings scheme. Non-completed scenarios are however more difficult to come with than completed scenarios that naturally mimic the

conditional distributions of a full model suggested by the missing data model,

$$z|y,\theta \sim k(z|y,\theta) \propto g(y,z|\theta)\,.$$

Non-completed scenarios may even be impossible to implement because of the explosive nature of the missing part of the data (as in semi-Markov models, see Cappé et al., 2004), while completed scenarios may get bogged down in terms of convergence because of the large dimension of the missing data.

1.2 MCMC and importance sampling

As detailed for instance in McLachlan and Peel (2000) for mixture models or in Robert and Casella (1999) in a more general perspective, Markov Chain Monte Carlo (MCMC) methods have been deeply instrumental in the Bayesian exploration of increasingly complex missing data problems, as further shown by the explosion in the number of papers devoted to specific missing data models since the early 1990's. Besides the processing of mixtures, which stand at the "easy" end of the processing spectrum (even though they offer hard enough challenges!), these years also saw major advances in handling models like hidden Markov models (Cappé et al., 2005), stochastic volatility models (Jacquier et al., 1994, Chib et al., 2002) and networks of hidden Markov models (Jordan, 2004).

Besides, this wealth of advances brought a new vision of the approaches anterior to the MCMC era and in particular to *importance sampling.* Recall (Robert and Casella, 1999, Chap. 3) that importance sampling is based on the simulation of $\theta(i)$'s ($i = 1, \ldots, M$) from a distribution $\varrho(\theta)$, called the importance function, that is not the distribution of interest $\pi(\theta|y)$, by correcting the difference via importance weights

$$\omega^{(i)} = \pi(\theta^{(i)}|y)/\varrho(\theta^{(i)}) \Big/ \sum_j \pi(\theta^{(j)}|y)/\varrho(\theta^{(j)})$$

to preserve (asymptotically) unbiasedness, that is,

$$\mathbb{E}\left[\omega^{(i)} h(\theta^{(i)})\right] \approx \int h(\theta)\pi(\theta|y)d\theta\,, \tag{1}$$

where h is a given function.

1.3 Population Monte Carlo

As proposed in Cappé et al. (2004) (see also del Moral et al., 2002), the notion of importance sampling can actually be strongly generalised to encompass much more adaptive and local schemes than previously thought, and this without relaxing its primary justification that is to provide a correct discrete approximation to the distribution of interest.

As in regular MCMC settings, the missing data structure of the problem can be exploited to produce a simple and feasible importance distribution, but this "natural solution" does not always produce good results. Since an attempt at providing a "universal" importance sampling scheme that would achieve acceptable convergence rates in most settings is doomed to fail, given the multiplicity of situations pertaining to missing data problems, and since specific solutions are bound to work only in a limited vicinity of the models they have been tested on, a logical extension to the regular importance sampling framework is to learn from experience, that is, to build an importance sampling function based on the performances of earlier importance sampling proposals. This is the essence of the *population Monte Carlo* scheme of Cappé et al. (2004): By introducing a temporal dimension in the selection of the importance function, an adaptive perspective can be achieved at little cost, for a potentially large gain in efficiency. Indeed, if iterated importance sampling is considered, with t denoting the index of the iteration, the choice of the importance function at iteration t can be dictated by the importance sample produced at iteration $t-1$, according to criteria that seek improved efficiency

of the sampler. A further advance can be achieved through the realization that importance functions need not be constant over the points in the sample, that is, the $\theta^{(i)}$'s, and, in particular, that they may depend differently on the past samples, while preserving the unbiasedness in (1). Rather than using a constant importance function ϱ or a sequence of importance functions ϱ_t, we can thus propose to use importance functions q_{it} that depend on both the iteration t and the sample index i.

1.4 Plan

The plan of the paper is as follows: Section 2 describes a population Monte Carlo scheme that takes advantage of the latent structure of the problem and describes the corresponding Rao–Blackwellisation technique. Sections 3 and 4 study the behavior of this sampling algorithm on two examples: a toy example, a censored exponential failure time and a model used in the analysis of financial data, the stochastic volatility model. For each model, we compare the population Monte Carlo sampling scheme with classical MCMC approximations. Section 5 concludes the paper.

2 Population Monte Carlo for missing data models

2.1 The basic scheme

If the distribution of $z|y,\theta$, $k\left(z|y,\theta_{t-1}^{(i)}\right)$, is known, a specific version of the general PMC algorithm can mimic the Gibbs sampler by generating the z's and θ's from their respective conditional distributions. In fact its proposal can be the distribution that corresponds to generating $z_t^{(i)}$ (that depends on both the iteration t and the sample index i) from the conditional distribution of z

$\theta_{t-1}^{(i)}$, $k\left(z|y,\theta_{t-1}^{(i)}\right)$, and $\theta_t^{(i)}$ from $\pi\left(\theta|y,z_t^{(i)}\right)$. The corresponding weight is

$$\omega_t^{(i)} \propto g\left(y,z_t^{(i)}|\theta_t^{(i)}\right)\pi\left(\theta_t^{(i)}\right)\Big/ k\left(z_t^{(i)}|y,\theta_{t-1}^{(i)}\right)\pi\left(\theta_t^{(i)}|y,z_t^{(i)}\right),$$

where $\pi(\theta)$ is the prior distribution on θ. The following pseudo-code summarizes these steps:

Alg. 1: Original PMC scheme for missing data models

- Step 0: Choice of $\left(\theta_0^{(1)},\ldots,\theta_0^{(M)}\right)$;
- Step t $(t=1,\ldots,T)$:

a) For $i=1,\ldots,M$:

Generate $z_t^{(i)}$ from $k\left(z|y,\theta_{t-1}^{(i)}\right)$ and $\theta_t^{(i)}$ from $\pi\left(\theta|y,z_t^{(i)}\right)$;

Compute $r_t^{(i)} = g\left(y,z_t^{(i)}|\theta_t^{(i)}\right)\pi\left(\theta_t^{(i)}\right)\Big/ k\left(z_t^{(i)}|y,\theta_{t-1}^{(i)}\right)\pi\left(\theta_t^{(i)}|y,z_t^{(i)}\right)$

and $\omega_t^{(i)} = r_t^{(i)}\Big/\sum_{s=1}^{M} r_t^{(s)}$;

b) Resample the $\left(\theta_t^{(i)}\right)$'s using the weights $\omega_t^{(i)}$'s.

Note that the conditional densities $k(z|y,\theta)$ and $\pi(\theta|y,z)$ may be known only up to a normalizing constant, given that they appear in every weight. Note also that the sequence of $(\theta_t^{(i)})_{1\le i\le M}$'s thus produced is a Markov chain. Contrary to the Gibbs sampler, since we are in the importance sampling setting, it is possible to replace sampling from $k(z|y,\theta)$ and $\pi(\theta|y,z)$ by alternative proposal distributions as long as the weights are modified accordingly.

After T iterations of the previous scheme, an asymptotically unbiased estimator of $\mathbb{E}_\pi(h(\theta))$ is given by the weighted average

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{M}\omega_t^{(j)}h\left(\theta_t^{(j)}\right).$$

The variance of this estimator obviously decreases both in T and in M. Most importantly, for each $t=1,\ldots,T$, as M goes to infinity and under mild

conditions, the average $\sum_{j=1}^{M} \omega_t^{(j)} h\left(\theta_t^{(j)}\right)$ converges in probability to $\mathbb{E}(\theta|y)$ and that a CLT holds, as shown in Douc et al. (2005). From this asymptotic perspective, T is a learning parameter. For this particular PMC scheme with limited adaptivity, if M is very large (typically $M \geq 100,000$), T can be taken equal to 2 or 3. However, for moderate values of M (typically $1,000 \leq M \leq 10,000$), T needs to be increased to compensate for the approximation, for instance $T = 10$.

While natural (as shown by its Gibbs sampler predecessor), the previous scheme has the drawback of being exposed to *degeneracy*, that is, to a strong asymmetry in the importance weights that jeopardizes the appeal of the importance sampling estimate. Indeed, iterated importance sampling encounters this difficulty even more than regular importance sampling because of the repeated resampling: the percentage of resampled particles can be very small between two iterations and the probability that this occurs increases over iterations. The consequence of the degeneracy of the population is that the number of surviving branches of ancestors diminishes very quickly when looking at the samples over generations. If the proposals are only based on the recently generated values, this may induce a serious bias or at least a severe impoverishment and a correlated increase in the variance of the estimators in the final output. As in regular importance sampling, there also is an additional risk that the weights $\omega_t^{(i)}$ misbehave, because of an infinite variance. We will see an illustration in the case of the censored exponential failure time model, with infinite variance on the weights of (θ, z).

2.2 Extensions via Rao–Blackwellisation

An approach that partly alleviates both of the above problems is to recycle the past simulations to estimate by importance sampling the marginal weight of

θ, rather than using the weight of the joint vector (θ, z). This idea is very similar to the Rao–Blackwellisation strategy used from the early days of MCMC algorithms (Gelfand and Smith, 1990, Robert and Casella, 1999): When the $(z_t^{(i)}, \theta_t^{(i)})$'s are generated as in Algorithm 2, the additional randomness due to the simulation of the $z_t^{(i)}$'s can be reduced by considering an importance sampling approximation to the distribution of $\theta_t^{(i)}$ conditional on $\theta_{t-1}^{(i)}$,

$$\int \pi(\theta|z,y)\, k(z|y,\theta_{t-1}^{(i)})\, dz\,,$$

which is the marginal kernel used in the Gibbs sampler. Rather than approximating this integral via costly brute force simulation, that is, by simulating a whole sample of z's from $k(z|y,\theta_{t-1}^{(i)})$ for every i, we can recycle the whole set of pre-simulated $z_t^{(j)}$'s by correcting for their sampling distribution $k\left(z_t^{(j)}|y,\theta_{t-1}^{(j)}\right)$. The corresponding importance sampling approximation of the marginal conditional distribution is then

$$\frac{1}{M}\sum_{l=1}^{M} \frac{k\left(z_t^{(l)}|y,\theta_{t-1}^{(i)}\right)\pi\left(\theta_t^{(i)}|y,z_t^{(l)}\right)}{k\left(z_t^{(l)}|y,\theta_{t-1}^{(l)}\right)}\,.$$

The weights used in the PMC take advantage of this Rao–Blackwellisation argument twice, namely by approximating both the true marginal posterior distribution of θ and its marginal proposal distribution:

$$\omega_t^{(i)} \propto \sum_{l=1}^{M} \frac{g\left(y,z_t^{(l)}|\theta_t^{(i)}\right)\pi\left(\theta_t^{(i)}\right)}{k\left(z_t^{(l)}|y,\theta_{t-1}^{(l)}\right)} \Bigg/ \sum_{l=1}^{M} \frac{k\left(z_t^{(l)}|y,\theta_{t-1}^{(i)}\right)\pi\left(\theta_t^{(i)}|y,z_t^{(l)}\right)}{k\left(z_t^{(l)}|y,\theta_{t-1}^{(l)}\right)} = \frac{n_t^{(i)}}{d_t^{(i)}}\,,$$

where $n_t^{(i)}$ and $d_t^{(i)}$ thus appear as importance sampling estimates of the marginal target and proposal at point $\theta_t^{(i)}$, respectively.

The following pseudo-code summarizes this modification:

Alg. 2: Rao–Blackwellised PMC scheme for missing data models —

- Step t $(t = 1, \ldots, T)$:

a) For $i = 1, \ldots, M$:

Generate $z_t^{(i)}$ from $k\left(z|y,\theta_{t-1}^{(i)}\right)$ and $\theta_t^{(i)}$ from $\pi\left(\theta|y,z_t^{(i)}\right)$;

b) For $i=1,\ldots,M$:

Compute $n_t^{(i)} = \dfrac{1}{M}\displaystyle\sum_{l=1}^{M} \dfrac{g\left(y,z_t^{(l)}|\theta_t^{(i)}\right)\pi\left(\theta_t^{(i)}\right)}{k\left(z_t^{(l)}|y,\theta_{t-1}^{(l)}\right)}$

Compute $d_t^{(i)} = \dfrac{1}{M}\displaystyle\sum_{l=1}^{M} \dfrac{k\left(z_t^{(l)}|y,\theta_{t-1}^{(i)}\right)\pi\left(\theta_t^{(i)}|y,z_t^{(l)}\right)}{k\left(z_t^{(l)}|y,\theta_{t-1}^{(l)}\right)}$

Compute $r_t^{(i)} = \dfrac{n_t^{(i)}}{d_t^{(i)}}$ and $\omega_t^{(i)} = r_t^{(i)} \Big/ \displaystyle\sum_{s=1}^{M} r_t^{(s)}$;

c) Resample the $\left(\theta_t^{(i)}\right)$'s using weights $\omega_t^{(i)}$'s.

In this version, the latent variables are mostly instrumental in that they are used to provide an approximation to the marginal posterior distribution of the θ's. This fact implies that the z's and the θ's can be dissociated in the simulation and, for instance, that a larger number of z's can be simulated to provide more stable evaluations of these marginal posterior distributions and of the corresponding weights. In the case of the stochastic volatility model (Section 4), we successfully implemented this strategy, as shown by the non-degeneracy of the samples of θ's thus obtained. The asymptotic behavior of the resulting PMC estimator is unchanged.

2.3 Implementation

The previous scheme supposes that both conditional distributions $k(z|y,\theta)$ and $\pi(\theta|y,z)$ are known (up to constants) and it strongly resembles Gibbs sampling in that it uses exactly the same kernel. However, as we will see in the stochastic volatility example, the exploration of the parameter space provided by the corresponding PMC scheme is by far superior to the performances of the MCMC approach, simply because it provides a flow of parallel proposals that are compared against the target distribution at each step.

In cases where either $k(z|y,\theta)$ or $\pi(\theta|y,z)$ is unknown, we face the same difficulty as MCMC algorithms, namely we have to select some appropriate proposal distribution to replace the true conditional distribution in both the simulation and the importance weights (which thus preserves the importance sampling validity of the algorithm). Since this is highly model dependent, we postpone the illustration for the more realistic case of stochastic volatility models in Section 4.

Although this has not been mentioned so far, we stress that the importance sample obviously needs to be initialised from some proposal distribution. Just as in MCMC setups, possibilities are numerous, if not always appropriate. A first possibility is to use the maximum likelihood estimator $\hat{\theta}$ of θ as a starting point for the first proposal, as in, for instance, Edmond et al. (2001) where the authors propose to use $\pi(z|y,\hat{\theta})$, instead of the more variable predictive density $\pi(z|y)$. A potential problem with this solution is that, typically, Bayesian inference is most useful in small sample settings for which maximum likelihood can provide unreliable estimates. Thus, in such cases it is doubtful that initialising the sampling scheme at $\hat{\theta}$ is a good choice. A connected criticism is that this choice does not take into account the intrinsic variability of $\hat{\theta}$ and often results in an importance function that is too concentrated around the maximum likelihood estimator. Therefore we propose to initialise the algorithm by simulating directly from the predictive distribution, which is only feasible when the prior on θ is both proper and available in closed form. Compared with the pluggin proposal $\pi(z|y,\hat{\theta})$, this predictive distribution on z has fatter tails and thus better coverage of the latent variable space. Obviously, both proposals, namely

$$\int_\Theta \pi(\theta)\pi(z|y,\theta)\,d\theta$$

and $\pi(z|y,\hat{\theta})$ can be used simultaneously to initialise parts of the sample, provided they are associated with the proper weights (including with a Rao-Blackwell averaging).

We also point out that both PMC structures are very straightforward implementations of the principles behind population Monte Carlo and that more elaborate constructions can be designed, as already illustrated in Cappé et al. (2004). In particular, these specific algorithms only use the previous samples as "stepping stones" for the new importance functions: if a value $\theta_{t-1}^{(i)}$ is resampled several times, a corresponding number of z's will be simulated from $k(z|y,\theta_{t-1}^{(i)})$. No further effort is made at analyzing the appropriateness of the resampled set of θ's against the target distribution. Nonetheless, the following examples will provide enough evidence that this rudimentary adaptive scheme performs satisfactorily even in the more challenging case of the stochastic volatility model. (See also Douc et al. (2005) for more advance adaptive schemes.)

3 Censored exponential failure model

As a first illustration, consider a sample of $n-r$ observed failure times $y_1,\ldots,y_{n-r}$ and r right-censored data points with a constant censoring time $c>0$ from an exponential distribution $\mathcal{E}xp(\theta)$. This is a most obvious missing data problem, $z_{n-r+1},\ldots,z_n$ being the unobserved failure remaining times. We also introduce the sufficient observed and unobserved statistics

$$s=\sum_{i=1}^{n-r} y_i \quad \text{and} \quad z=\sum_{i=n-r+1}^{n} z_i$$

and use $\theta\sim\mathcal{G}(a,b)$ as prior, with expectation a/b and variance a/b^2.
The exact posterior distribution is then a $\mathcal{G}\left(a+n-r,b+s+rc\right)$ distribution, which can be used as a benchmark to evaluate the performances of our PMC scheme. The corresponding conditional distributions are $\theta|y,z\sim\mathcal{G}(n+a,s+rc+z+b)$ and $z|y,\theta\sim\mathcal{G}(r,\theta)$, which means that the Rao–Blackwellised version of PMC can be used (even though the exact marginal posterior of θ is available in this toy example).

The Rao–Blackwellised PMC scheme is evaluated on a simulated data set of $n = 20$ $\mathcal{E}xp(1)$ rv's, with censoring at $c = .4$. The prior is a weakly informative $\mathcal{G}(.1, .1)$ distribution. After 30 iterations of the Rao–Blackwellised PMC algorithm with only $M = 200$ points per sample, we obtain the results summarised in Figure 1. The corresponding evaluation of the posterior mean of θ is

$$\frac{1}{30}\sum_{t=1}^{30}\sum_{j=1}^{200}\omega_t^{(j)}\theta_t^{(j)} = 0.8514 \tag{2}$$

for a exact value of 0.8519. In Figure 1, the third graph gives the evolution of the Rao–Blackwellised PMC approximation to the posterior mean of θ (in red) through iterations. For this simulated dataset, convergence to the true value is ensured after 10 iterations of the PMC scheme. The effect of the weights on the sample of θ's is noticeable but not overwhelming, which means that the importance function at the 30th iteration is well-calibrated enough for the target distribution. As discussed above, there is a wide variety of approaches to the selection of M and T, but it seems that the one based on the stabilisation of the overall average is the most practical. Unless more advanced adaptive schemes are used for the choice of the q_{it}'s, it also appears that the number of iterations till stabilisation is most often situated in the vicinity of $T = 10$ iterations. (Additional computational effort should bear on increasing M rather than T.)

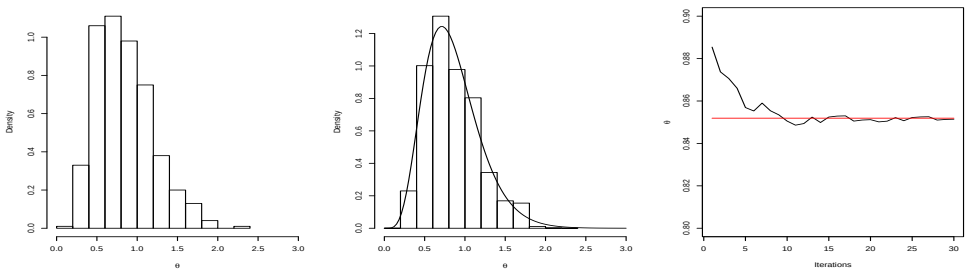

Fig. 1. Rao–Blackwellised PMC sample: *(left)* sample of θ before resampling, at the 30th iteration; *(center)* weighted sample against true posterior distribution; *(right)* evolution over iterations of the Rao–Blackwellised PMC approximation to the posterior mean of θ (straight line in grey).

While straightforward, this example is particularly interesting as a defense of Rao–Blackwellisation. Indeed, at iteration t of the algorithm, the importance sampling weight of $\left(\theta_t^{(j)}, z_t^{(j)}\right)$ in the original PMC algorithm is inversely proportional to

$$\left(\theta_{t-1}^{(j)}\right)^r \exp\left(-\theta_{t-1}^{(j)} z_t^{(j)}\right) \left(b + \sum_{i=1}^{n-r} y_i + rc + z_t^{(j)}\right)^{n+a}$$

and thus has an infinite variance. The consequences of this infinite variance on degeneracy are clearly shown in Figure 2: the weights are much more dispersed than in Figure 1 and the weighted sample collapses to a few significant points. In this case, the approximation to the posterior mean of θ is quite poor for the same number of iterations.

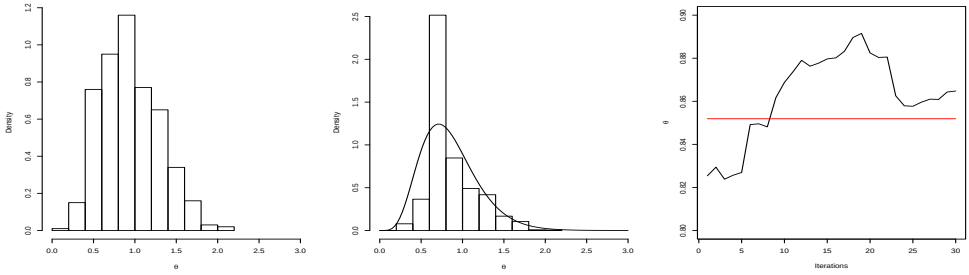

Fig. 2. Original (not Rao–Blackwellised) PMC sample *(same legend as Figure 1).*

Table 1 compares Gibbs estimates for $1,000$ iterations and a burn-in period of 500 (Gibbs estimates are the average over the last 500 iterations) and PMC estimates for $M = 100$ and $T = 10$ (PMC estimates are the average over the last five PMC iterations). Obviously for this simple model both algorithms have a satisfactory behavior. But it is worth noting that the variability of the PMC estimator is smaller, providing a somewhat more precise and less variable estimate of the posterior mean.

c	Gibbs	PMC
0.2	0.0022 (0.071)	-0.0011 (0.058)
0.4	-0.0007 (0.038)	0.0005 (0.029)
0.6	0.0003 (0.028)	0.0001 (0.015)

Table 1

Averages of the differences between Gibbs and PMC estimates and the true posterior mean of θ for $1,000$ simulated datasets of size $n = 20$ and different values of c. The values in parentheses are the standard error estimates of these differences.

4 Stochastic volatility models

Stochastic volatility (SV) models have attracted a lot of attention in the recent years as a way of generalising the Black-Scholes option pricing formula to allow for heterogeneous variations in the scale of time series. These models have gradually emerged as a useful way of modeling time-varying volatility with significant applications, especially in Finance (see for example Taylor (1994), Shephard (1996) and Ghysels et al. (1996) for detailed reviews) and they are also an alternative to the Autoregressive Conditional Heteroscedasticity (ARCH) models of Engle (1982) (see also Bollerslev et al., 1994).

A central feature of stochastic volatility models is that the variance is a latent stochastic process. In the simplest model, the observations are independent conditional on their variance:

$$y_t = \beta \exp\left(z_t/2\right) \epsilon_t \,, \qquad \epsilon_t \sim \mathcal{N}(0,1)$$

and the log-variance process is an AR(1) model $z_{t+1} = \varphi z_t + \sigma u_t$, with $u_t \sim \mathcal{N}(0,1)$ and the stationarity assumption that

$$z_1 \sim \mathcal{N}\left(0, \sigma^2/\left(1-\varphi^2\right)\right) .$$

The set of parameters is thus $\theta = (\beta, \varphi, \sigma)$, with the usual stationarity condi-

tion $\varphi \in]-1,1[$.

Bayesian inference is far from easy in this setup, because this is a missing data model with no closed-form likelihood. Besides, compared with the previous examples, the missing structure z is not countable and much more complex than the censoring structure of Section 3. The only approach to the model is therefore based on its completion by the missing data z, which unfortunately is of the same dimension as the data. MCMC algorithms have been proposed for this model, using different approximations and proposals in the Metropolis-Hastings step, starting with Jacquier et al. (1994) Gamma approximation. See, e.g., Kim et al. (1998) and Chib et al. (2002) for detailed reviews on the MCMC aspects of the problem. Our experience with these algorithms is however that they are not necessarily robust to all types of datasets and may fail to converge for long series or extreme values of the parameters β and φ. In particular, it appears from our experiments that MCMC algorithms are very sensitive to the generation of the missing data and that they may well fail to converge even when initialised at the true parameter values.

Under a noninformative prior like

$$\pi(\beta^2, \varphi, \sigma^2) = 1/(\sigma\beta)\, \mathbb{I}_{]-1,1[}(\varphi)\,,$$

the posterior distributions for β^2 and σ^2 conditional on the completed data are both inverse Gamma distributions with $(n-1)/2$ shape parameters and

$$\sum_{t=1}^{n} y_t^2 \exp(-z_t)/2 \quad \text{and} \quad \sum_{t=2}^{n} \left(z_t - \varphi z_{t-1}\right)^2/2 + z_1^2(1-\varphi^2)/2$$

as scales, respectively. The conditional distribution of φ, $f\left(\varphi|y, z, \sigma^2\right)$, is less conventional, since it is proportional to

$$\sqrt{1-\varphi^2}\exp - \left(\varphi^2 \sum_{t=2}^{n-1} z_t^2 - 2\varphi \sum_{t=2}^{n} z_t z_{t-1}\right) \Big/ 2\sigma^2\, \mathbb{I}_{]-1,1[}(\varphi)\,,$$

but a standard Metropolis–Hastings proposal (Chib et al., 2002) is a truncated

normal distribution on $]-1,1[$ with mean and variance

$$\sum_{t=2}^{n} z_t z_{t-1} \Big/ \sum_{t=2}^{n-1} z_t^2 \quad \text{and} \quad \sigma^2 \Big/ \sum_{t=2}^{n-1} z_t^2 \,.$$

The most challenging and documented part is the simulation from the conditional distribution of $z|y,\varphi,\sigma^2$. Most papers focus on componentwise proposals: First, Shephard (1993) propose to approximate the distribution of $\log(\epsilon_t^2)$ by a normal distribution $\mathcal{N}(-1.27, 4.93)$, to account for both first moments, and this implies the use of a Gaussian proposal for the distribution of $z_t|z_{-t},y,\varphi,\sigma^2$. An alternative is advanced in Jacquier et al. (1994), which approximates the distribution of $\exp(z_t)$ by a Gamma distribution. Independently, Geweke (1994) and Shephard (1994) suggested the use of Gilks and Wild (1992) ARS procedure for sampling from log-concave densities like $f(z_t|z_{-t},y,\varphi,\sigma^2)$. Kim et al. (1998) developed a simple accept/reject procedure, bounding $\exp(-z_t)$ by a function linear in z_t. At last, Shephard and Pitt (1997) used a quadratic (Taylor) expansion of $\exp(z_t)$ around the mean of the distribution of $z_t|z_{-t},\varphi,\sigma^2$ and we use their approximation for illustrating the difference between Gibbs and PMC implementations. Following Shephard and Pitt (1997), the proposal distribution for z_1 is a normal with mean

$$\frac{\varphi z_2/\sigma^2 + 0.5\exp(-\varphi z_2)\, y_1^2\,(1+\varphi z_2)\,/\beta^2 - 0.5}{1/\sigma^2 + 0.5\exp(-\varphi z_2)\, y_1^2/\beta^2}$$

and variance

$$1\Big/\left(1/\sigma^2 + 0.5\exp(-\varphi z_2)\, y_1^2/\beta^2\right) .$$

The proposal for z_t is a normal with mean

$$\frac{\left(1+\varphi^2\right)\mu_t/\sigma^2 + 0.5\exp(-\mu_t)\, y_t^2\,(1+\mu_t)\,/\beta^2 - 0.5}{\left(1+\varphi^2\right)/\sigma^2 + 0.5\exp(-\mu_t)\, y_t^2/\beta^2}$$

and variance

$$1\Big/\left\{\left(1+\varphi^2\right)/\sigma^2 + 0.5\exp(-\mu_t)\, y_t^2/\beta^2\right\}.$$

At last, the proposal for z_n is a normal with mean

$$\frac{\varphi z_{n-1}/\sigma^2 + 0.5\exp\left(-\varphi z_{n-1}\right) y_n^2 \left(1+\varphi z_{n-1}\right)/\beta^2 - 0.5}{1/\sigma^2 + 0.5\exp\left(-\varphi z_{n-1}\right) y_n^2/\beta^2}$$

and variance

$$1\Big/\left\{1/\sigma^2 + 0.5\exp\left(-\varphi z_{n-1}\right) y_n^2/\beta^2\right\} .$$

Note that both Liu et al. (1994) and Shephard and Pitt (1997) suggest *blocking*, that is, a joint simulation of a group of consecutive z_t's, to improve the speed of convergence of simulators. While we did not observe a consistent pattern of improvement in our experiments, the goal here is to compare, for the above proposal distribution, the performances of the PMC approximation algorithm and of the classical hybrid Gibbs Metropolis–Hastings algorithm. We therefore only use the above componentwise proposals for z.

We proceeded to a Monte Carlo numerical experiment on two type of simulated datasets of size $n = 1,000$. Each of then reflects typical problems for weekly and daily financial data and has been replicated teen times. In the weekly case, we chose $\beta^2 = 1$, $\sigma^2 = 0.1$ and $\varphi = 0.9$ while in the daily case $\beta^2 = 1$, $\sigma^2 = 0.01$ and $\varphi = 0.99$. Two datasets of each type are represented in Figure 3, along with the corresponding simulated volatilities z.

For these two particular datasets, the results of the MCMC algorithm, $10,000$ iterations, are presented in Figures 4–7. Figures 4 and 6 do not exhibit any convergence difficulty. Note however the slow mixing on β in Figure 6 (upper left) and, to a lesser degree, on σ^2 in both Figures (middle left).

Moreover, we have iterated ten times, $T = 10$, a Rao–Blackwellised PMC algorithm with $M = 1,000$. For the same datasets, the results are presented in Figures 8–11. Figures 9 and 11 provide an excellent reconstitution of the volatilities.

Tables 2 and 3 summarize the quality of the parameter estimates for both

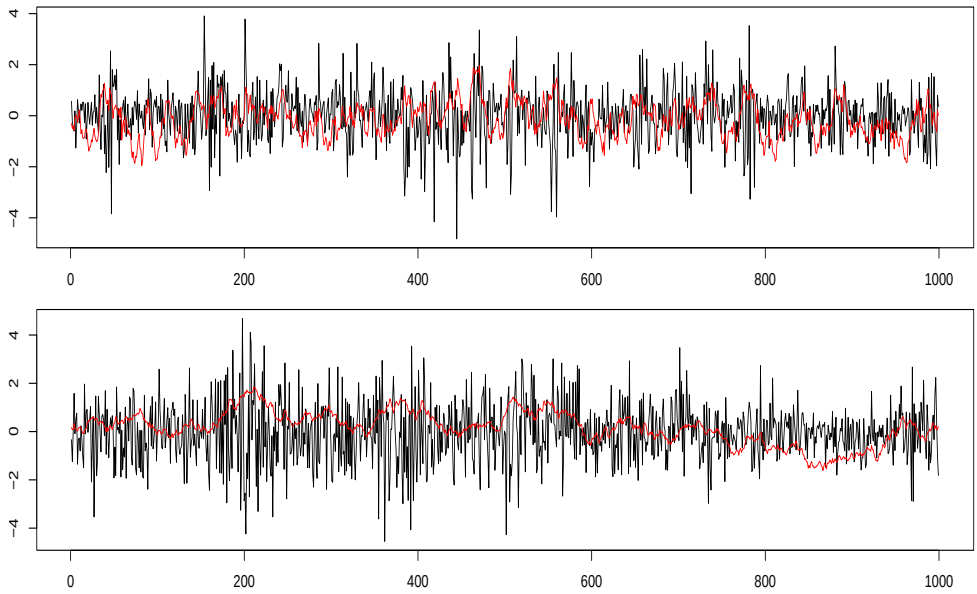

Fig. 3. Weekly *(upper)* and daily *(lower)* simulated datasets with $n = 1,000$ observations y_t *(black)* and volatilities z_t *(grey)*.

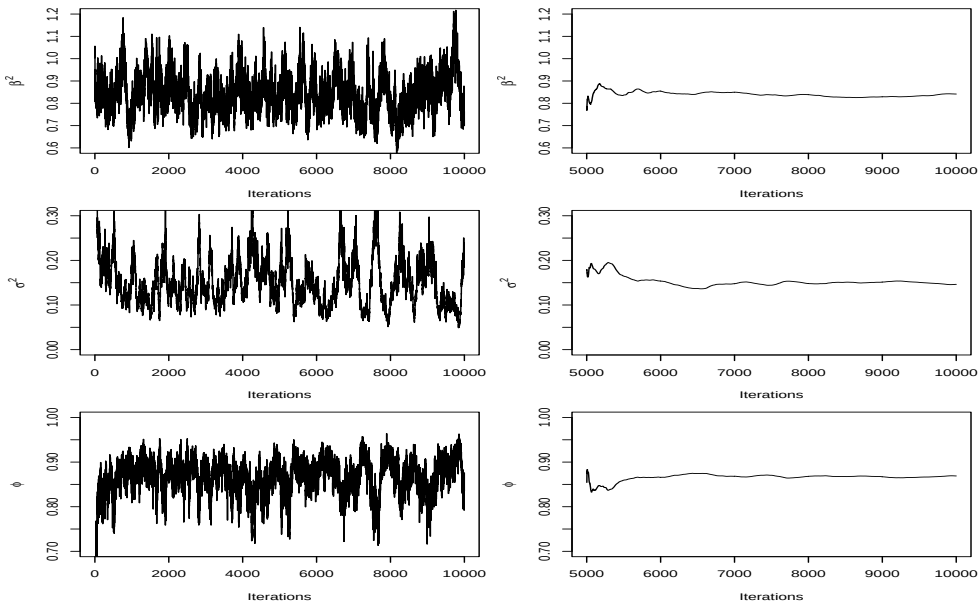

Fig. 4. Weekly dataset: evolution of the MCMC samples for the three parameters *(left)* and convergence of the MCMC estimators *(right)*.

methods. These Tables provide the averages of the differences between MCMC and PMC estimates and the parameters true posterior mean for the 10 replicated datasets. The values in parentheses are the standard error estimates of

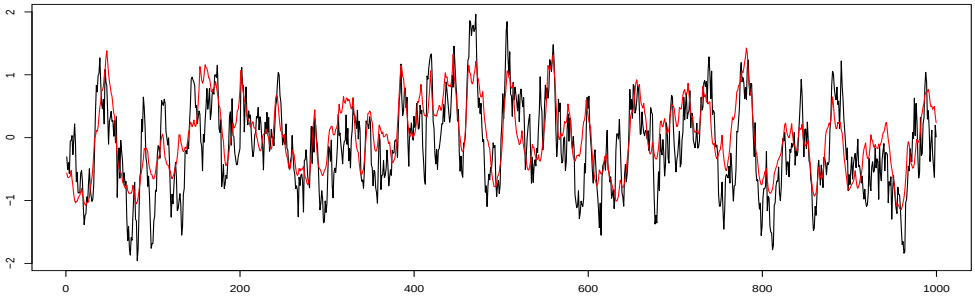

Fig. 5. Weekly dataset: estimation of the stochastic volatility (in black the true volatility and in grey the MCMC estimation based on the last 5000 iterations).

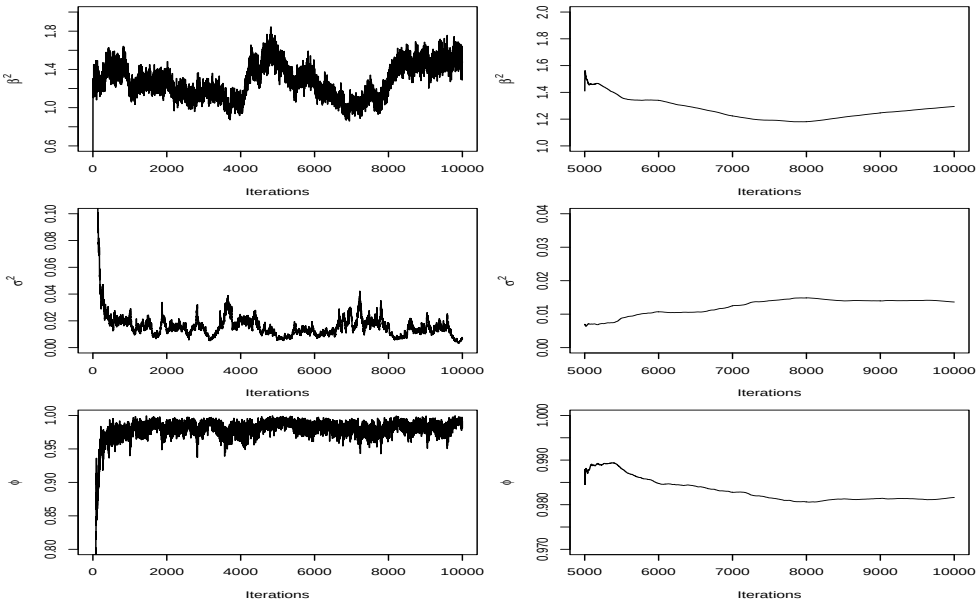

Fig. 6. Daily dataset: same legend as Figure 4.

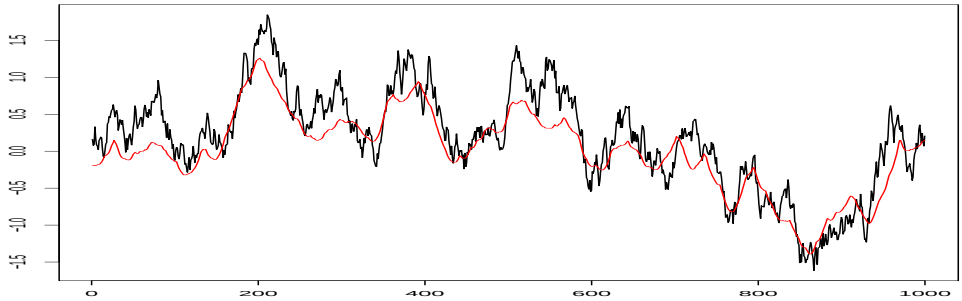

Fig. 7. Daily dataset: same legend as Figure 5.

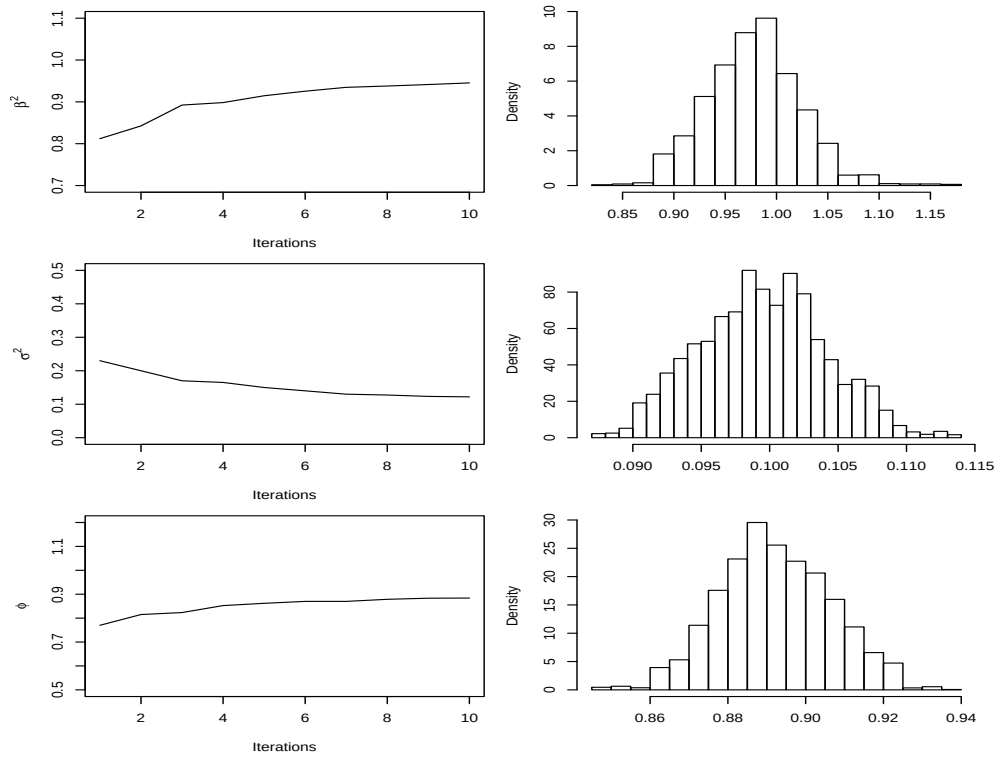

Fig. 8. Weekly dataset: evolution over iterations of the Rao–Blackwellised PMC approximation *(left)* and 10th iteration weighted PMC sample *(right)*.

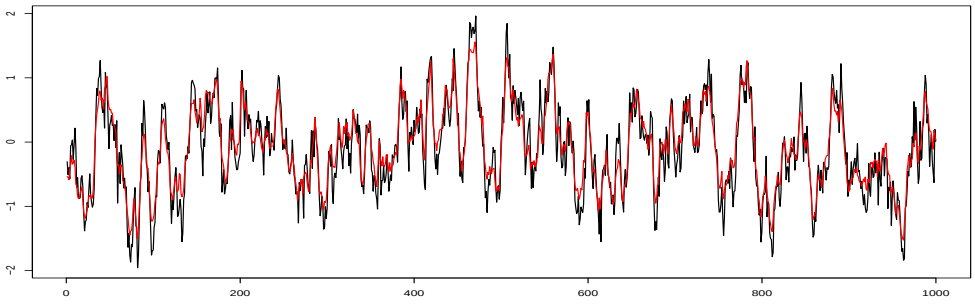

Fig. 9. Weekly dataset: estimation of the stochastic volatility (in black the true volatility and in grey the PMC estimation based on the 10th iteration weighted PMC sample).

these differences. Note that the MCMC estimates are the average over the last $5,000$ iterations and that the PMC parameter estimates are calculated only over the last five iterations. On this small numerical experiment PMC performs slightly better than MCMC. In particular, the standard errors of the estimates of β^2 are significantly reduced. This is not surprising: we have al-

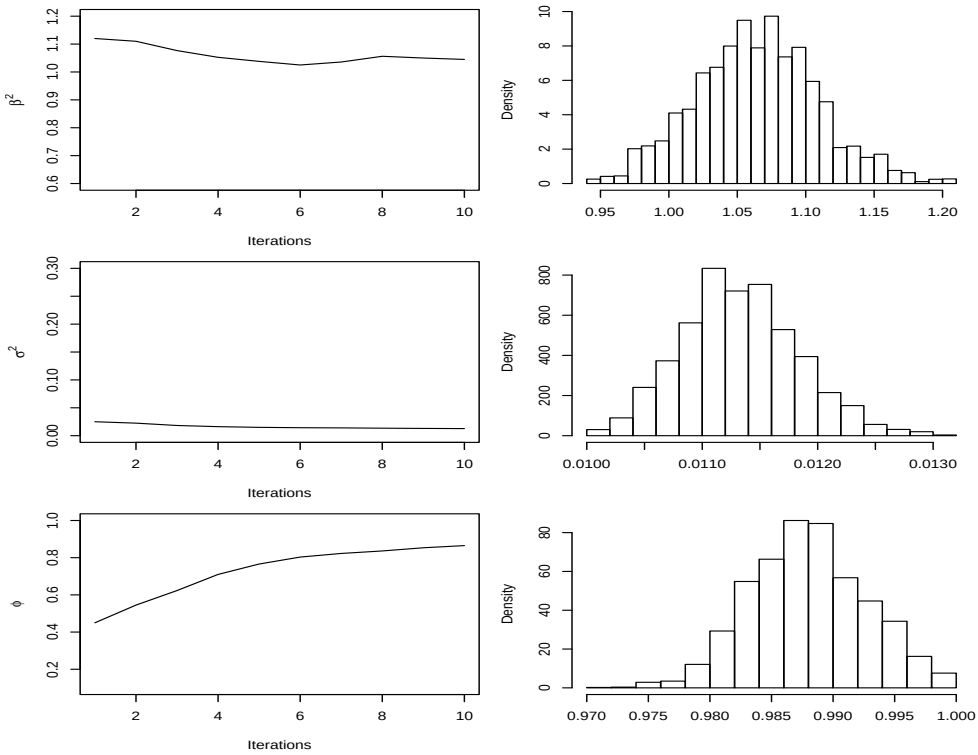

Fig. 10. Daily dataset: same legend as Figure 8.

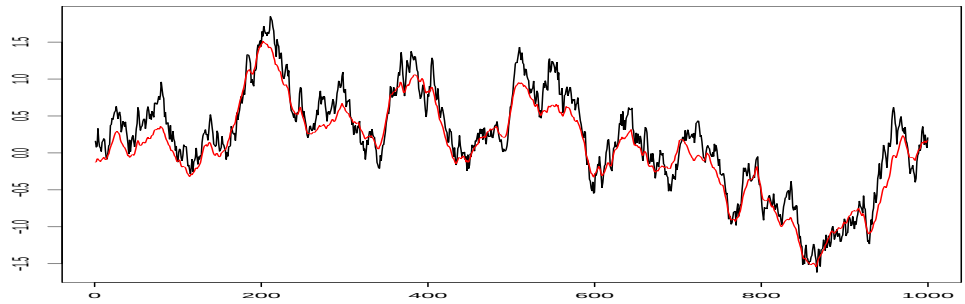

Fig. 11. Daily dataset: same legend as Figure 9.

ready observed, through Figures 4 and 6, the slow mixing of the MCMC chains on β^2. This is clearly related to relative slow mixing of the log-variances themselves, see Figures 5 and 7.

Weekly	MCMC	PMC
β^2	0.057 (0.151)	0.023 (0.091)
σ^2	0.009 (0.031)	0.008 (0.027)
φ	- 0.012 (0.023)	-0.011 (0.020)

Table 2

Weekly dataset: Averages of the differences between MCMC and PMC estimates and the parameters true posterior mean for the 10 simulated datasets. The values in parentheses are the standard error estimates of these differences.

Daily	MCMC	PMC
β^2	-0.069 (0.241)	-0.032 (0.130)
σ^2	0.011 (0.008)	0.011 (0.008)
φ	-0.024 (0.014)	-0.021 (0.012)

Table 3

Daily dataset: Averages of the differences between MCMC and PMC estimates and the parameters true posterior mean for the 10 simulated datasets. The values in parentheses are the standard error estimates of these differences.

5 Conclusion

This paper has shown that the population Monte Carlo scheme is a viable alternative to MCMC schemes in missing data settings. Even with the standard choice of the full conditional distributions, this method provides an accurate representation of the distribution of interest in a few iterations. As in regular importance sampling, the choice of the importance function is paramount, but the iterative nature of PMC erodes the dependence on the importance function by offering a wide range of adaptive kernels that can take advantage of the previously simulated samples. This paper has addressed the most natural proposal kernels based on the missing data structure but, as illustrated in

Cappé et al. (2004) and Douc et al. (2005), multiscale proposals can be added to increase the efficiency of the method and to provide a better approximation to the distribution of interest. In this perspective, a range of proposals can be tested on earlier iterations to improve the approximation of the posterior distribution, even though this may require a larger number T of iterations. In the context of this paper, however, an increase of the number of iterations is unlikely to produce a quantitative improvement, once the algorithm has reached the stationarity region: Indeed, if the $\theta_t^{(i)}$'s are approximately distributed from $\pi(\theta|y)$ and if the proposal distribution is constant, the distribution of the $(\theta_{t+1}^{(i)}, \omega_{t+1}^{(i)})$'s will not change over iterations.

Concerning the computational effort between MCMC and PMC, we can make the following remark: using the same number of overall simulations makes sense in that the computational effort is often decided at the beginning of an experiment. To make the number of PMC double loops equal to the number of MCMC iterations is then meaningful. In addition, the overall CPU times are also comparable because, while PMC requires weight normalisation and resampling, it can be partially parallelised (for instance, in the spirit of parallel algorithms given by Kontoghiorghes (2000), Gatu and Kontoghiorghes (2003) for linear models), compared with the loop used in MCMC algorithms. The PMC Rao–Blackwell step is about between two and four times more expensive than the standard PMC. But the impact of Rao–Blackwellisation on the quality of the PMC estimation is noticeably superior to its impact on MCMC outputs, where Rao–Blackwellised and standard averages most often are not distinguishable (Robert and Casella, 1999, Chap. 8) unless more advanced (and more costly) techniques are used (Casella and Robert, 1996). For instance, in the case of the stochastic volatility model, Rao–Blackwellisation is quintessential in stabilising the estimates, since the original PMC is prone to produce highly variable weights and to degenerate into a single point after resampling. Rao–Blackwellisation thus brings a welcome correction to the fun-

damental drawback of importance sampling techniques, that is, the potential degeneracy of infinite variance weights.

As can clearly be seen in Section 4, the population Monte Carlo approach can benefit from earlier works on MCMC algorithms to select good proposal distributions. It thus does not come as a breakpoint in this area of computational Statistics, but rather as a further advance that exploits dependence on previous iterations without requiring ergodicity and the theoretical apparatus of Markov chain theory. It thus brings a considerable simplification to the development of *adaptive* algorithms, when compared with recent works on adaptive MCMC methods (see, e.g., Haario et al., 1999 2001, Andrieu and Robert, 2001). In particular, the calibration of proposal distributions against explicit performance diagnoses introduced in Andrieu and Robert (2001) can also be reproduced for our algorithm.

Acknowledgments

The authors are grateful to the Associate Editor and all the reviewers for helpful comments and suggestions which helped to improve this work.

References

Andrieu, C. and Robert, C. (2001). Controlled MCMC for optimal sampling. Technical Report 2001-25, Cahiers du CEREMADE, University Paris Dauphine.

Bollerslev, T., Engle, R., and Nelson, D. (1994). ARCH models. In Engle, R. F. and Mac Fadden, D., editors, *The Handbook of Econometrics*. Elsevier Sciences, North-Holland Series in Statistics and Probability.

Cappé, O., Guillin, A., Marin, J., and Robert, C. (2004). Population Monte Carlo. *J. Computational and Graphical Statistics*, 13(4):907–929.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag, New York.

Casella, G. and Robert, C. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94.

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. American Statistical Association*, 95:957–970.

Chib, S., Nadari, F., and Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *J. Econometrics*, 108:281–316.

del Moral, P., Doucet, A., and Peters, G. (2002). Sequential Monte Carlo samplers. Technical report, Department of Electrical Engineering, Cambridge University.

Dempster, A. P., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statistical Society Series B*, 39:1–38.

Douc, R., A., G., Marin, J.-M., and Robert, C. (2005). Convergence of adaptive sampling schemes. Technical Report 2005-6, Cahiers du CEREMADE, University Paris Dauphine.

Edmond, M., Raftery, A., and Russell, J. (2001). Easy computation of Bayes factors and normalizing constants for mixture models via mixture importance sampling. Technical Report 398, Department of Statistics, University of Washington, Seattle.

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica*, 50:987–1007.

Everitt, B. (1984). *An Introduction to Latent Variable Models*. Chapman and Hall, London.

Gatu, C. and Kontoghiorghes, E. (2003). Parallel algorithms for computing all possible subset regression models using the QR decomposition. *Parallel Computing*, 29(4):505–521.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating

marginal densities. *J. American Statistical Association*, 85:398–409.

Geweke, J. (1994). Comment on Bayesian analysis of stochastic volatility models. *J. Business and Economic Statistics*, 12:397–399.

Ghysels, E., Harvey, A., and Renault, E. (1996). Stochastic volatility. In Rao, C. and Maddala, G., editors, *Statistical Methods in Finance*. Elsevier Science, North-Holland Series in Statistics and Probability.

Gilks, W. and Wild, P. (1992). Adaptative rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14:375–395.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Jacquier, R., Polson, N., and Rossi, P. (1994). Bayesian analysis of stochastic volatility models (with discussion). *J. Business and Economic Statistic*, 12:371–417.

Jordan, M. (2004). Graphical models. *Statistical Science*. (to appear).

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies*, 65:361–393.

Kontoghiorghes, E. (2000). *Parallel Algorithms for Linear Models: Numerical Methods and Estimation Problems*. Advances in Computational Economics, Vol. 15, Kluwer Academic Publishers, Boston, MA.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. J. Wiley, New York.

Liu, J., Wong, W., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes. *Biometrika*, 81:27–40.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. J. Wiley, New York.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. J. Wiley, New York.

Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Shephard, N. (1993). Fitting non-linear time series models, with applications to stochastic variance models. *J. Applied Econometrics*, 8:135–152.

Shephard, N. (1994). Partial non-Gaussian state space analysis of non-Gaussian measurement times series. *Biometrika*, 81:115–131.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In Cox, D. R., Barndorff-Nielsen, O. E., and Hinkley, D. V., editors, *Time Series Models in Econometrics, Finance and Other Fieds*. Chapman and Hall.

Shephard, N. and Pitt, M. (1997). Likelihood analysis of non-Gaussian measurement times series. *Biometrika*, 84:653–667.

Tanner, M. and Wong, W. (1987). The calculation of posterior distribution distributions by data augmentation, (with discussion). *J. American Statistical Association*, 82:528–550.

Taylor, S. (1994). Modelling stochastic volatility. *Mathematical Finance*, 4:183–204.