Iterated importance sampling in missing data problems

G. Celeux*, J.M. Marin† and C.P. Robert‡

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. Their special features also allow for an easier calibration of most computational techniques, by virtue of the data completion they naturally offer. This was the case for the EM algorithm of Dempster et al. (1977) and the Gibbs sampler of Gelfand and Smith (1990), and this is also true for the population Monte Carlo algorithm recently studied in Cappé et al. (2002). 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.

Keywords: adaptive algorithms, censored exponential failure time, latent variables, mixture model, particle system, population Monte Carlo, Rao–Blackwellisation, sequential importance sampling, stochastic volatility model.

---

*INRIA celeux@inria.fr
†CEREMADE, Université Paris IX Dauphine, Paris marin@ceremade.dauphine.fr
‡Corresponding author: CEREMADE, Université Paris IX Dauphine and CREST, INSEE, Paris xian@ceremade.dauphine.fr Postal address: CEREMADE, 75775 Paris cedex 16, France
1 Introduction

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

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

where $z \in \mathcal{Z}$ 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 scenarii are however more difficult to come with than completed scenarii 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 scenarii 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., 2002), while completed scenarii may get bogged in terms of convergence because of the large dimension of the missing data.
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 and Rydén, 2004), 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 θ_i’s (i = 1, \ldots, M) from a distribution \rho(\theta), called the importance function, that is not the distribution of interest \pi(\theta|y), by correcting the difference via importance weights
\begin{equation}
\omega_i = \pi(\theta_i|y)/\rho(\theta_i) \left/ \sum_j \pi(\theta_j|y)/\rho(\theta_j) \right.
\end{equation}
to preserve (approximately) unbiasedness, that is,
\begin{equation}
\mathbb{E}[\omega_i h(\theta_i)] \approx \int h(\theta)\pi(\theta|y)d\theta .
\end{equation}
As proposed in Cappé et al. (2002) (see also del Moral and Doucet, 2002), the notion of importance sampling can actually be strongly generalized 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. Indeed, the choice of the importance distribution \rho is paramount to ensure that the convergence to the distribution \pi occurs at the right rate, the minimum requirement being that the variance of the importance weights \omega_i is finite (Robert and Casella, 1999, Chap. 3). 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. (2002): 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 realisation 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 \( q \) or a sequence of importance functions \( q_t \), we can thus propose to use importance functions \( q_{it} \) that depend on both the iteration \( t \) and the sample index \( i \).

The plan of the paper is as follows: Section 2 describes the population Monte Carlo scheme which takes profit of the latent structure of the problem and describes the corresponding Rao–Blackwellisation technique. Sections 3—5 study the behavior of this sampling algorithm on three examples: two toy examples, a censored exponential failure time and a simple mixture of distributions, 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 6 concludes the paper.
2 Population Monte Carlo

2.1 Background

As stated earlier, Population Monte Carlo (PMC) is a Monte Carlo scheme proposed by Cappé et al. (2002), following the denomination of Iba (2000), that encompasses both Monte Carlo and MCMC techniques by producing, at each iteration, say $t$, a valid importance sampling approximation to the distribution of interest, namely

$$
\pi(\theta, z|y) \propto g(y, z|\theta)\pi(\theta)
$$

in the setting of missing data models. The samples of $\left(\theta_t^{(i)}, z_t^{(i)}\right)$'s ($1 \leq i \leq M$) are constructed at each iteration $t$ from almost arbitrary proposals $q_{it}(\theta, z)$ (that depend on both the iteration $t$ and the sample index $i$) and associated with the importance weights

$$
\omega_t^{(i)} \propto \pi \left(\theta_t^{(i)}, z_t^{(i)}|y\right) / q_{it} \left(\theta_t^{(i)}, z_t^{(i)}\right).
$$

Those weights are then used directly in integral approximations like

$$
\int h(\theta, z)\pi(\theta, z|y)d\theta dz \approx \sum_i \omega_t^{(i)} h(\theta_t^{(i)}, z_t^{(i)}) / \sum_i \omega_t^{(i)}
$$

and indirectly in pruning the sample of $\left(\theta_t^{(i)}, z_t^{(i)}\right)$'s from its less likely components, via resampling steps as in, e.g., Rubin (1987). While resampling introduces some purely computational noise in the structure, it is actually necessary to produce unweighted samples that approximate the distribution of interest and can be used exchangeably in the future steps of the algorithm. (Note that there are devices that reduce the multinomial noise to the order $O(1)$, see Carpenter et al., 1999). We stress here that the novelty of the method is that the importance sampling functions $q_{it}$ can almost arbitrarily depend from the samples at earlier iterations, vary over the sample points, and that a whole range of criteria can be used to select the most pertinent points in previous samples while preserving the unbiasedness features of importance sampling.

In fact, a regular importance sampling argument (sees Cappé et al., 2002) shows that the choice of $q_{it}$ is fairly unrestricted and that this proposal distribution can depend on the
previous sample or even on the whole sequence of samples simulated so far, thus taking advantage of the performances of previous proposals. Obviously, this dependence implies that the approach is not designed according to MCMC principles, but the additional complexity of the proposal is more than compensated by the lack of convergence constraints. Indeed, at every iteration $t$, the weighted resulting sample $\left(\theta_t^{(1)}, z_t^{(1)}\right), \ldots, \left(\theta_t^{(M)}, z_t^{(M)}\right)$ can be approximatively regarded as a sample from the posterior distribution $\pi(\theta, z|y)$ in the sense that an unbiased estimator of $E_\pi[h(\theta)]$ is

$$\sum_{j=1}^{M} \omega_t^{(j)} h\left(\theta_t^{(j)}\right) \quad \text{when} \quad \omega_t^{(j)} = \pi\left(\theta_t^{(j)}, z_t^{(j)}|y\right)/q_{jt}\left(\theta_t^{(j)}, z_t^{(j)}\right).$$

Since, in most cases, $\pi(\theta, z|y) \propto g(y, z|\theta)\pi(\theta)$ is unscaled, the weights $\omega_t^{(i)}$ have to sum up to one, which creates some bias in the above approximation. However, the bias induced by using the denominator

$$\mathcal{D}_t = \sum_{i=1}^{M} \frac{\pi\left(\theta_t^{(i)}\right) g\left(y, z_t^{(i)}|\theta_t^{(i)}\right)}{q_{it}\left(\theta_t^{(i)}, z_t^{(i)}\right)}$$

to normalise the $\omega_t^{(i)}$s can be arbitrarily reduced by considering instead cumulative sums over past iterations,

$$\mathcal{D}_T^+ = \sum_{t=1}^{T} \mathcal{D}_t/T,$$

since all the $\mathcal{D}_t$’s evaluate the same normalising constant. (Note that, obviously, we have to assume that the $q_{it}$ are correctly normalized, because of the dependence on $i$ of the importance functions.)

The following pseudo-code summarizes the steps of a PMC algorithm:

**Algorithm 1: General PMC scheme**

- **Step 0:** Choice of $\left(\left(\theta_0^{(1)}, z_0^{(1)}\right), \ldots, \left(\theta_0^{(M)}, z_0^{(M)}\right)\right)$;

- **Step $t$ ($t = 1, \ldots, T$):**
  a) For $i = 1, \ldots, M$:
  
  Generate $\left(\theta_t^{(i)}, z_t^{(i)}\right)$ from $q_{it}(\theta, z)$
Compute $r_t^{(i)} = \frac{g(y, z_t^{(i)} | \theta_t^{(i)}) \pi(\theta_t^{(i)})}{q_t(\theta_t^{(i)}, z_t^{(i)})}$ and take $\omega_t^{(i)} = r_t^{(i)} / \sum_{k=1}^{t} \sum_{s=1}^{M} r_k^{(s)} (1 \leq \ell \leq t)$;

b) Resample the $(\theta_t^{(i)}, z_t^{(i)})$’s using weights proportional to the $w_t^{(i)}$’s.

c) Design the $q_{t(t+1)}$ from the current (and possibly past) sample(s).

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

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

The variance of this estimator obviously decreases when the number $M$ of simulations at each iteration increases. The convergence in $T$ requires further conditions on the choice of the $q_{it}$ and is not the issue here: if we fix a maximal number of iterations $T$, this estimator is validated by the usual argument for importance sampling. Note that, depending on performances, it is also possible to increase $M$ along the iterations, thus giving more importance to the last iterations, or to weight the inner sums differently, depending for instance on the variability of the $\omega_t^{(j)}$. Once again, the asymptotics in $T$ of such a scheme are quite delicate and they are not pursued further in this paper. (If $T$ is chosen by a stopping rule like the stabilisation of some estimates, this also creates a slight bias in the importance sampling.)

Other proposals have been made earlier about iterated importance functions. While the particle system literature (Doucet et al., 1999) is usually more concerned with sequential problems, Chopin (2002) uses particle systems as an iterated importance sampling to handle large datasets. This is a special case of PMC where the $q_{it}$’s only depend on the iteration index $t$ and correspond to the genuine posterior distributions associated with a fraction $k_t$ of the observed dataset. Similarly, the smooth bootstrap of Stravropoulos and Titterington (1999) and the Warnes (2001) of kernel coupler both rely on nonparametric kernels based on the previous importance sample to build an improved importance distribution: again, this is a special case of PMC where the proposal is constant at each iteration $t$, that is, does not
depend on the sample index $i$. See Cappé et al. (2002) and del Moral and Doucet (2002) for further references on related approaches.

## 2.2 Missing data models

We now consider more specifically the case of missing data models, with the target distribution $\pi(\theta, z|y)$. First, if the distribution of $z|y, \theta$ is known, a specific version of the general PMC algorithm can mimic the Gibbs sampler by generating the $z$’s and $\theta$’s from their respective conditional distributions. In fact it can use as proposal $q_{it}$ the distribution that corresponds to generating $z_t^{(i)}$ from $k\left(z_t|y, \theta_{t-1}^{(i)}\right)$ and $\theta_t^{(i)}$ from $\pi\left(\theta|y, z_t^{(i)}\right)$. The corresponding weight is thus

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

The following pseudo-code summarizes these steps:

**Algorithm 2: 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 $k\left(z_t|y, \theta_{t-1}^{(i)}\right)$;

  Generate $\theta_t^{(i)}$ from $\pi\left(\theta|y, z_t^{(i)}\right)$;

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

  and $\omega_t^{(i)} = r_t^{(i)} / \sum_{k=1}^{t} \sum_{s=1}^{M} r_k^{(s)}$ ($1 \leq \ell \leq t$);

  b) Resample the $\left(\theta_t^{(i)}\right)$’s using weights proportional to the $w_t^{(i)}$’s.

Note that, in this special case, the conditional densities $k(z|y, \theta)$ and $\pi(\theta|y, z)$ may be known only up a normalizing constant, given that they appear in every weight.
2.3 Degeneracy and Rao–Blackwellisation

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 sampling: 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 $q_{it}$ are only based on the recently generated values, this may induce serious biases or at least severe impoverishment and 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 of this in the following section on the censored exponential failure time model, with infinite variance on the weights of $(\theta, z)$.

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 $\theta$, rather than using the weight of the joint vector $(\theta, z)$. 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 partly alleviated 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 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 is
then
\[
\frac{1}{M} \sum_{i=1}^{M} k \left( z_t^{(i)} | y, \theta_{t-1}^{(i)} \right) \pi \left( \theta | y, z_t^{(i)} \right) \frac{k \left( z_t^{(i)} | y, \theta_{t-1}^{(i)} \right)}{k \left( z_t^{(i)} | y, \theta_{t-1}^{(i)} \right)}.
\]

The weights used in the PMC are exploiting this Rao–Blackwellisation argument twice, namely to approximate both the true marginal posterior distribution of \( \theta \) and its marginal proposal distribution:

\[
\omega_t^{(i)} \propto \sum_{i=1}^{M} g \left( y, z_t^{(i)} | \theta_t^{(i)} \right) \pi \left( \theta_t^{(i)} \right) \frac{k \left( z_t^{(i)} | y, \theta_{t-1}^{(i)} \right)}{k \left( z_t^{(i)} | y, \theta_{t-1}^{(i)} \right)} /
\sum_{i=1}^{M} k \left( z_t^{(i)} | y, \theta_{t-1}^{(i)} \right) \pi \left( \theta | y, z_t^{(i)} \right),
\]

where the \( \omega_t^{(i)} \) are summing up to 1 (over \( i \) and \( t \)).

**Algorithm 3: Rao–Blackwellised 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 \( k \left( z | y, \theta_{t-1}^{(i)} \right) \);
  
  Generate \( \theta_t^{(i)} \) from \( \pi \left( \theta | y, z_t^{(i)} \right) \);

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

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

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

  Compute \( r_t^{(i)} = \frac{n_t^{(i)}}{d_t^{(i)}} \) and \( \omega_t^{(i)} = r_t^{(i)} / \sum_{k=1}^{t} \sum_{s=1}^{M} r_k^{(s)} (1 \leq \ell \leq t) \);

  c) Resample the \( \left( \theta_t^{(i)} \right) \)'s using weights proportional to the \( w_t^{(i)} \)'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 $\theta$’s. This fact implies that the $z$’s and the $\theta$’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 5), we successfully implemented this strategy, as shown by the non-degeneracy of the samples of $\theta$’s thus obtained.

2.4 Implementation

The previous scheme supposes that the conditional distributions $k(z|y, \theta)$ and $\pi(\theta|y, z)$ are known (up to constants) and it strongly resembles Gibbs sampling since it uses 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 a Metropolis-Hastings approach. In cases where either $k(z|y, \theta)$ or $\pi(\theta|y, z)$ are 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 (and thus preserve the validity of the algorithm). Since this is highly model dependent, we reserve the illustration for the more realistic stochastic volatility model in Section 5.

Although this has not been mentioned till this point, the importance sample 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 $\theta$ 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 setting for which maximum likelihood can provide unreliable estimates. Thus, in such cases it is doubtful that initiating 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 in \( \theta \) 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 \( \theta \) is both proper and available in closed form. Compared with the plug-in 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. (Since the predictive distribution is usually not available in closed form, a Rao-Blackwellisation argument as in Section 2.3 can be used as an approximation.)

As a concluding note, we want to point out that these two 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. (2002). In particular, these specific algorithms only use the previous samples as “stepping stones” for the new importance functions: if a value \( \theta^{(i)}_{t-1} \) is resampled several times, a corresponding number of \( z \)'s will be simulated from \( k(z|y, \theta^{(i)}_{t-1}) \). No further effort is made at analyzing the appropriateness of the resampled set of \( \theta \)'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.

### 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 $\theta \sim G(a, b)$ as prior.

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

The Rao–Blackwellised PMC scheme is evaluated on a simulated data set of $n = 20$ $\exp(1)$ rv’s, with censoring at $c = .4$. The prior is a weakly informative $G(.1,.1)$ distribution.

After 30 iterations of the Rao–Blackwellised PMC algorithm with only $M = 200$ points per sample, we obtain the results summarized in Figure 1. The corresponding evaluation of the posterior mean of $\theta$ is

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

for an 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 $\theta$ (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 $\theta$’s is noticeable but not overwhelming, which means that the importance function at the 30th iteration is well-calibrated for the target distribution.

As discussed in Section 2.1, 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 $q_u$’s, it also appears that the number of iterations till stabilisation is most often situated in the vicinity of 10 iterations.

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 $(\theta_t^{(j)}, z_t^{(j)})$ 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
Figure 1: Rao–Blackwellised PMC sample: (left) sample of $\theta$ 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 $\theta$ (in red).

the weighted sample collapses to a few significant points. In this case, the approximation to the posterior mean of $\theta$ is quite poor for the same number of iterations.

Figure 2: Original PMC sample (same legend as Figure 1).

On this simulated data, we also ran 100,000 iterations of the standard Gibbs sampler. Figure 3 presents the convergence of the Gibbs approximation to the posterior mean of $\theta$. Despite the simplicity of this model, we observe that, contrary to the Rao–Blackwellised PMC, the convergence of the Gibbs sampler to the true value is relatively slow and only appears clearly after 20,000 iterations: as seen in Figure 1, with the Rao–Blackwellised PMC scheme, we obtain some excellent approximations only from 10 iterations and 200 simulations. Obviously, the poor performances of the Gibbs sampler are related to a poor starting value and are not necessarily repeated on the same scale for other starting points. But this clearly illustrates an appealing feature of PMC which is that its parallel structure
naturally eliminates the influence of poor starting points much more faster than the single chain Gibbs samplers.

![Graph](image)

Figure 3: Gibbs approximation to the posterior mean (in red) through iterations.

4 A mixture model

A problem that has induced a lot of work in computational Statistics is the analysis of mixtures of distributions (see, e.g., McLachlan and Peel, 2000). While more challenging versions can be processed, we focus on the case of a mixture of two normal distributions

\[
p \mathcal{N}(\mu_1, \sigma^2) + (1 - p) \mathcal{N}(\mu_2, \sigma^2)
\]

where both \( p \not= 1/2 \) and \( \sigma > 0 \) are known, for illustrative purposes: the parameter space is two-dimensional and the posterior surface can thus be plotted easily. We use a normal prior \( \mathcal{N} (\delta, \sigma^2/\lambda) \) (\( \delta \in \mathbb{R} \) and \( \lambda > 0 \) are known hyper-parameter) on both \( \mu_1 \) and \( \mu_2 \), with \( \delta = \lambda = 1 \) in the simulations. For a simulated dataset of 1000 datapoints, from the distribution \( 0.2 \mathcal{N}(0, 1) + 0.8 \mathcal{N}(2, 1) \), the comparison of the log-likelihood and the log-posterior surfaces [not given here] shows that the effect of the prior modelling is quite limited. As shown on Figure 6, the log-posterior exhibits a bimodal feature of the posterior that appears for every simulated dataset, even with lower sample sizes. This bimodality is not to be confused with the non-identifiability attached to mixture models, when components are exchangeable (Celeux et al., 2000): In the current case, the means \( \mu_1 \) and \( \mu_2 \) are identifiable, but there always exists a spurious mode which may correspond to the allocation of most observations
to one component. Note however that the posterior distribution is mostly concentrated on
the mode of importance, very close to the true values of the parameters.

The standard approach to this problem, used in both the EM algorithm and in the original
Gibbs sampler, is to take advantage of the missing data representation of the mixture, namely
to introduce unobserved indicators $z_i \in \{0, 1\}$ ($1 \leq i \leq n$) such that

$$E[y_i | z_i = 0] = \mu_1 \quad \text{and} \quad E[y_i | z_i = 1] = \mu_2.$$  

The posterior distributions of both $\mu_1$ and $\mu_2$ given $y$ and $z = (z_1, \ldots, z_n)$ are then conjugate,
while the conditional distributions of the $z_i$’s are simple Bernoulli distributions (see Diebolt
and Robert, 1994).

After 30 iterations of both the original and the Rao–Blackwellised PMC algorithms, with
$M = 200$ points in each sample, the results are quite similar. However, it appears that the
Rao–Blackwellised approach slightly reduces the degeneracy by preserving more values from
one iteration to the next, as shown by the weight distributions in Figure 4 which exhibit less

Figure 4: Histograms of the weights of (left) the original PMC scheme; (right) the Rao–
Blackwellised PMC scheme (after 30 iterations).

The evolutions of the Rao–Blackwellised approximation to the posterior means of $\mu_1$ and
$\mu_2$ through iterations are presented in Figure 5, with a very small scale on the $\mu_i$ axis. The
final evaluation of the posterior means of $\mu_1$ and $\mu_2$ are $-0.96$ and $1.98$, respectively.

Figure 6 shows that the sample produced by the PMC algorithm is quite in agreement
with the modal zone of the posterior distribution. The second mode, which is much lower
Figure 5: Evolution of the Rao–Blackwellised PMC approximation to the posterior means of $\mu_1$ (left) and $\mu_2$ (right).

(remember this is a log-posterior level set), is not preserved in the sample after the first iteration. Note also that the weights (represented by circles on the graph) are quite similar, with no overwhelming weight in the sample (as already showed by Figure 5).

For comparison purposes, we also ran 10,000 Gibbs iterations ten times, starting at random from the prior distribution: a rather surprising feature is that the Gibbs sampler is as sensitive as the EM algorithm to the spurious mode. In fact, the phenomenon illustrated in Figure 7, namely that the MCMC sequence may remain trapped in this spurious mode, occurred six times out of ten. This failure never occurred with both PMC schemes, most likely because of the parallel dynamic of PMC that clearly allows for a better exploration of the posterior space.

5 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 modelling 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
Figure 6: Representation of the log-posterior distribution via \texttt{R image} and \texttt{contours} procedures with the Rao–Blackwellised PMC weighted sample after 30 iterations (the weights are proportional to the circles at each point).

Figure 7: Same graph as Figure 6 for the last 1000 points of 10,000 iterations of a Gibbs sampler. \textit{(All points are equally weighted.)}
also an alternative to the Autoregressive Conditional Heteroscedasticity (GARCH) 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:

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

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

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

The set of parameters is thus \( \theta = (\beta, \varphi, \sigma) \), with the usual stationarity condition \( \varphi \in ]-1, 1[ \).

Bayesian inference in this setup is far from easy, 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 \( \beta \) and \( \varphi \). 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 initialized at the true parameter values.

Under a noninformative prior like

\[ \pi(\beta^2, \varphi, \sigma^2) = 1/(\sigma \beta) \mathbb{I}_{[0,1]}(\beta) \mathbb{I}_{[-1,1]}(\varphi), \]

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

\[ \sum_{t=1}^{n} y_t^2 \exp \left( -z_t / 2 \right) \quad \text{and} \quad \sum_{t=2}^{n} \left( z_t - \varphi z_{t-1} \right)^2 / 2 \]
as scales, respectively. The conditional distribution of $\varphi$, $f(\varphi|y, z, \sigma^2)$, is less conventional, since it is proportional to

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

but a standard proposal (Chib et al., 2002) is a truncated normal distribution on $]-1, 1[$ with mean and variance

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

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 $z_t|z_{t-1}, y, \varphi, \sigma^2$. 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 $f(z_t|z_{t-1}, y, \varphi, \sigma^2)$. 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 $z_t|z_{t-1}, \varphi, \sigma^2$ and we will follow suit, since numerous preliminary experiments have shown that it leads to superior performances, compared with the alternatives.

We now describe this proposal distribution in details. First, for the conditional distribution of $z_1$,

$$
f(z_1|y, z_{-1}, \beta^2, \sigma^2, \varphi) \propto \exp \left( -0.5 (z_1 - \varphi z_2)^2 / \sigma^2 - 0.5 \exp(-z_1) y_1^2 / \beta^2 - 0.5 z_1 \right), \quad (2)
$$

we can take advantage of the AR structure since $z_1|z_{-1} \sim \mathcal{N} (\varphi z_2, \sigma^2)$. Therefore, the difficult term $\exp(-z_1)$ in (2) can be overcome by a Taylor expansion around $\varphi z_2$, namely

$$
\exp(-z_1) \approx \exp(-\varphi z_2) - \exp(-\varphi z_2) (z_1 - \varphi z_2) + 0.5 \exp(-\varphi z_2) (z_1 - \varphi z_2)^2.
$$

The corresponding proposal distribution for $z_1$ is thus a normal with mean

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

\[20\]
and variance
\[
1 / \left(1 / \sigma^2 + 0.5 \exp (-\varphi \beta^2) \right).
\]

Similarly, for \(2 \leq t \leq n-1\), the conditional distribution of \(z_t\) given all the other unknowns and \(y\) is proportional to
\[
\exp \left\{ -0.5 \left(1 + \varphi^2 \right) (z_t - \mu_t)^2 / \sigma^2 - 0.5 \exp (-z_t) y_t^2 / \beta^2 - 0.5 z_t \right\}.
\]
where \(\mu_t = \varphi (z_{t-1} + z_{t+1}) / (1 + \varphi^2)\). This quantity \(\mu_t\) being also the conditional expectation of \(z_t\) given \(z_{t-1}, z_{t+1}\), it is natural to expand \(\exp (-z_t)\) in the exponential of (3) by a Taylor expansion around \(\mu_t\). Therefore our proposal distribution for \(z_t\) is a normal distribution with mean
\[
\frac{(1 + \varphi^2) \mu_t / \sigma^2 + 0.5 \exp (-\mu_t) y_t^2 (1 + \mu_t) / \beta^2 - 0.5}{(1 + \varphi^2) / \sigma^2 + 0.5 \exp (-\mu_t) y_t^2 / \beta^2}
\]
and variance
\[
1 / \left\{ (1 + \varphi^2) / \sigma^2 + 0.5 \exp (-\mu_t) y_t^2 / \beta^2 \right\}.
\]

At last, following the same direction, we end up with a normal proposal on \(z_n\) with mean
\[
\varphi z_{n-1} / \sigma^2 + 0.5 \exp (-\varphi z_{n-1}) y_n^2 (1 + \varphi z_{n-1}) / \beta^2 - 0.5
\]
and variance
\[
1 / \left\{ 1 / \sigma^2 + 0.5 \exp (-\varphi z_{n-1}) y_n^2 / \beta^2 \right\}.
\]

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 \(z\).

For the analysis of these performances, we used two simulated datasets of size \(n = 1000\), which reflect typical problems for weekly and daily financial data. In the weekly case, we chose \(\beta^2 = 1\), \(\varphi = 0.9\) and \(\sigma^2 = 0.1\), while in the daily case \(\beta^2 = 1\), \(\varphi = 0.99\) and \(\sigma^2 = 0.01\). These datasets are represented in Figure 8, along with the corresponding simulated volatilities \(z\).
First, the results of the MCMC algorithm (10,000 iterations) applied to these two datasets are presented in Figures 9–12. For the weekly dataset, the estimate of \( \beta^2 \) (over the last 5000 simulated values) is 0.84, while the estimate of \( \varphi \) is equal to 0.87 and the estimate of \( \sigma^2 \) is equal to 0.15. For the daily dataset, the corresponding approximations are 0.98, 0.89 and 0.099, respectively. While the reconstituted volatilities are on average close to the true values, the parameter estimates are rather poor, even though the cumulated averages of Figures 9 and 11 do not exhibit any difficulty with the convergence. Note however the slow mixing on \( \beta \) in Figure 11 (upper left) and, to a lesser degree, on \( \sigma^2 \) in both Figures (middle left).

Then, with the same proposal distributions, we have iterated ten times a Rao–Blackwellised PMC algorithm with \( M = 1000 \). The results are presented in Figures 13–16. For the weekly dataset, the estimate of \( \varphi \) (over the 10 iterations) is equal to 0.88, while the estimate of \( \sigma^2 \) is equal to 0.12 and the estimate of \( \beta^2 \) is equal to 0.94. For the daily dataset, the corresponding approximations are 0.87, 0.012 and 1.04, respectively. These estimations are clearly closer to the true values than the ones obtained with the MCMC algorithm. (Note that the scales

Figure 8: Weekly (upper) and daily (lower) simulated datasets with \( n = 1000 \) observations \( y_t \) (black) and volatilities \( z_t \) (red).
Figure 9: Weekly dataset: evolution of the MCMC samples for the three parameters (left) and convergence of the MCMC estimators (right).

Figure 10: Weekly dataset: estimation of the stochastic volatility (in black the true volatility and in red the MCMC estimation based on the last 5000 iterations).
Figure 11: Daily dataset: same legend as Figure 9.

Figure 12: Daily dataset: same legend as Figure 10.
on Figures 13 (left) and 15 (left) are much smaller than those of Figures 9 (right) and 11 (right). Moreover, Figures 14 and 16 provide an excellent reconstitution of the volatilities.

Figure 13: Weekly dataset: evolution over iterations of the Rao–Blackwellised PMC approximation (left) and 10th iteration weighted PMC sample (right).

6 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
Figure 14: Weekly dataset: estimation of the stochastic volatility (in black the true volatility and in red the PMC estimation based on the 10th iteration weighted PMC sample).

Figure 15: Daily dataset: same legend as Figure 13.
based on the missing data structure but, as illustrated in Cappé et al. (2002), 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 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.

The impact of Rao–Blackwellisation on the quality of the PMC estimation is noticeably superior to the 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 fundamental drawback of importance sampling techniques, that is, the potential degeneracy of infinite variance weights.

As can clearly be seen in Section 5, 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 simplifi-
cation 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.

References

0125, Cahiers du CEREMADE, Université Paris Dauphine.

Mac Fadden, D., editors, The Handbook of Econometrics. Elsevier Sciences, North-Holland Series
in Statistics and Probability.

Computational and Graphical Statistics. (To appear).


evolving data sets. Technical report, Department of Statistics, Oxford University.

94.


28
CUED/F-INFENG, Department of Electrical Engineering, Cambridge University.


normalizing constants for mixture models via mixture importance sampling. Technical Report
398, Department of Statistics, University of Washington, Seattle.


in Statistics and Probability.


Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential MCMC in Practice*. Springer-
Verlag.


method for efficient sampling from multi-mode distributions. Technical Report 395, Department
of Statistics, University of Washington, Seattle.