A sequential particle filter method for static models

NICOLAS CHOPIN

Laboratoire de Statistique, CREST, INSEE, Timbre J120,
75675 Paris cedex 14, France
chopin@ensae.fr

December 12, 2000

Abstract

Particle filter methods are complex inference procedures, which combine importance sampling and Monte Carlo schemes, in order to consistently explore a sequence of multiple distributions of interest. The purpose of this article is to show that such methods can also offer an efficient estimation tool in "static" setups; in this case, $\pi(\theta|y_1, ..., y_N)$ is the only posterior distribution of interest but the preliminary exploration of partial posteriors $\pi(\theta|y_1, ..., y_n)$ $(n < N)$ makes computing time savings possible. A complete "black-box" algorithm is proposed for independent or Markov models. Our method is shown to possibly challenge other common estimation procedures, in terms of robustness and execution time, especially when the sample size is important. Two classes of examples are discussed and illustrated by numerical results: mixture models and discrete generalized linear models.

Key words: Generalized linear models, Hastings-Metropolis, Importance sampling, Importance sub-sampling, MCMC, Mixture models, Parallel processing, Particle filter methods.

1 Introduction

MCMC (Monte Carlo Markov Chains) methods are a common tool nowadays for Bayesian inference, since they can handle a large variety of models. Unfortunately, in dynamic setups, when a sequence of posterior distributions π_t is involved, MCMC techniques fail to offer a quick or even manageable resolution, as they need to generate a different chain run for each posterior π_t, and do not take into account the previous generations from π_{t-1}. Alternatively, some authors have been developing more efficient methods based on importance sampling iterative strategies. In this context, an inference on π_{t-1} can be easily reused to draw an inference on π_t , by a proper "reweighting" operation. These methods are usually referred as "particle filter methods" (Doucet et al., 2001).

The purpose of this paper is to show that such methods can also improve estimation in static setups, when a single posterior distribution $\pi(\theta|y_1, ..., y_N)$ is involved. In fact, the present work was initiated by the following concern: common simulation procedures are hardly implementable on huge datasets, because they consist of numerous iterations, each of them requiring a complete "browsing" of the observations (each iteration needs to access and compute from the whole sample). This internal structure is clearly time-consuming. We propose an alternative structure, which provides significant computational savings by performing preliminary explorations of partial distributions $\pi(\theta|y_1, ..., y_n)$ $(n < N)$: at first, an inference is drawn from the n first observations; at a second time, this inference is "updated" through importance sampling to take into account the p following observations. This strategy can be iterated several times to finally infer from the whole sample, through stages which only required to browse a small part of the observations sample. In a sense, we endow a given static model with artificial dynamics: observations are considered to arrive sequentially at distinct times t, and are smoothly incorporated to the previous inferences. Given this dynamics, a particle filter is applied to a sequence of partial posterior distributions $\pi_t(\theta) = \pi(\theta|y_1, ..., y_{n_t})$, with $n_1 < ... < n_t < ... < n_T = N$.

From now on, the observations y_n's are supposed to be drawn from a parametric family $\{\mathcal{P}_\theta; \theta \in \Theta\}$, where Θ is an open space of $\mathbb{R}^K$ $(K \geq 1)$. The notation $y_{n:n+p}$ will refer to the sequence of the observations $y_n, ..., y_{n+p}$. A partial posterior distribution $\pi(\theta|y_{1:n})$ will be denoted π_n (therefore π_N will stand for the "complete" posterior distribution).

This paper is structured as follows. Section 2 recalls the main properties of particle filter methods. Section 3 presents the *importance sub-sampling* (ISS) scheme, which consists of importing sampling applied to partial posteriors π_n. Section 4 details the ISIS algorithm (Importance Sub-sampling Iterated Scheme), a particle filter method dedicated to estimating expectations $E_{\pi_N}\{h(\theta)\}$, in cases where the observations are either independent or Markov. Section 5 gives two examples of applications: generalized linear models, and mixture models.

2 Particle Filters

2.1 Importance sampling

A *particle system* is a sequence (θ_j, w_j) of weighted random variables in Θ (θ_j is a particle with weight w_j), which *targets* a distribution of interest π over Θ, in the sense that

$$\lim_{H\to+\infty} \frac{\sum_{j=1}^{H} w_j h(\theta_j)}{\sum_{j=1}^{H} w_j} = E_\pi\{h(\theta)\} \text{ almost surely}$$

holds for any measurable h such that $E_\pi\{h(\theta)\}$ exists.

When the particles are drawn from an instrumental distribution g, weights proportional to $\pi(\theta_j)/g(\theta_j)$ give a particle system of target π: this operation is called *importance sampling.* The ratio $\pi(\theta_j)/g(\theta_j)$ is often known only up to a multiplicative constant, which is cancelled by the denominator $\sum_{j=1}^{H} w_j$. The resulting estimate $\hat{\mu}(h) = \sum_{j=1}^{H} w_j h(\theta_j)/\sum_{j=1}^{H} w_j$ is biased but consistent, and we have

$$H^{\frac{1}{2}}\{\hat{\mu}(h) - E_\pi(h)\} \to \mathcal{N}\left(0, V_{\pi/g}(h)\right)$$

where $V_{\pi/g}(h) = \mathrm{V}_g\left[\pi(\theta)/g(\theta)\{h(\theta) - E_\pi(h)\}\right]$. This central limit theorem is valid when the particles are independently drawn from g, but also under more general hypotheses, for instance when obtained through a reversible Markov chain of stationary law g (Madras and Piccioni, 1999).

Thus, it is sensible to consider the quantity $V_{\pi/g}(h)$ as a measure of the efficiency of importance sampling based on the instrumental g. To speak more clearly, it measures how much g is close to π to make a good estimation of $E_\pi\{h(\theta)\}$ possible. This quantity can be "normalized" by dividing by

$\mathrm{V}_\pi\{h(\theta)\}$, in order to remove the variability due to the studied phenomena itself. We define

$$\tau_{\pi/g}(h) = V_{\pi/g}(h)\left[\mathrm{V}_\pi\{h(\theta)\}\right]^{-1}.$$

Note that $\tau_{\pi/g}(h) = I$ when $\pi = g$.

This normalized measure of efficiency is directly related to Carpenter et al. (1999) definition of *effective sample size*, i.e. the sample size required to attain the same precision with particles directly drawn from the target distribution π (the effective sample size is of the same order than the quantity $H\left\|\tau_{\pi/g}(h)\right\|^{-1}$).

We will further use the following properties, as sufficient conditions for $V_{\pi/g}(h)$ and $\tau_{\pi/g}(h)$ finiteness (Geweke, 1989):

$$\sup_{\theta\in\Theta}\left(\frac{\pi(\theta)}{g(\theta)}\right) < +\infty, \text{ and } E_g\{h(\theta)h(\theta)'\} < +\infty.$$

2.2 Iterated applications of importance sampling

A major advantage of particle systems is their flexibility: a simple *reweighting* operation $w_j' = w_j\pi_2(\theta_j)/\pi_1(\theta_j)$ shifts the target of a particle system from π_1 to π_2 (Liu and Chen, 1998; Gilks and Berzuini, 1999). The ratio $\pi_2(\theta_j)/\pi_1(\theta_j)$ is called *incremental weight*.

In a dynamic setting, that is when the distribution of interest π_t is evolving through time, this reweighting scheme can be iterated for each π_t. Unfortunately, each reweighting phase introduces a bit more variability in the estimates, when π_t is moving away from π_1: typically, particles with significant weights get less and less numerous, whereas the mass of the considered distribution π_t can even leave the support of the initial distribution π_0 (π_0 refers here to the distribution from where particles are initially drawn). In others words, $\tau_{\pi_t/\pi_0}(h)$ may dangerously increase, even if $\tau_{\pi_1/\pi_0}(h)$ and the successive $\tau_{\pi_t/\pi_{t-1}}(h)$ are within reasonable values. The particle system suffers from a progressive *impoverishment (*or *degeneracy*).

The reweighting steps can be alternated with *resampling* steps: each particle θ_j is replaced by a number n_j of its replicates (n_j may be equal to 0). The new particles are assigned a weight equal to 1. The n_j's can be determined in various ways, the most common being:

- by multinomial selection (Gordon et al., 1993). A new particle $\theta^r_{j'}$ is drawn from the following multinomial distribution:

$$\mathrm{P}(\theta^r_{j'} = \theta_j) = w_j/\Sigma w_j \text{ for } j = 1, ..., H$$

This random selection keep the estimates unbiased. The n_j's are random.

- by deterministic selection:

$$n_j = [w_j/\Sigma w_j]$$

where [.] stands for the integer part. This selection does not preserve the unbiasness of the estimates, but it seems to avoid the extra variability induced by a random selection, and remains asymptotically valid (Crisan and Doucet, 2000, when $H \to \infty$, see).

It must be stressed that resampling does not protect from degeneracy: it just saves further calculation time by getting rid of particles of insignificant weight. Moreover *resampling* artificially conceals the system impoverishment, by replacing high weights with numerous replicates of an unique particle, thus introducing high correlations between particles.

Liu and Chen (1998) and Crisan and Doucet (2000) are good reviews of methods related to iterations of reweighting and resampling steps.

Gilks and Berzuini (1999) proposed to add to the resampling step a *rejuvenation* step. Resampled particles at stage t are then "moved" according to a Markov chain transition kernel with stationary distribution π_t: $\theta^m_j \sim K_t(\theta^r_j, .)$. This operation does not change the system target, but it may strongly reduces the system impoverishment, since identical replicates of a single particle are replaced by new "fresh" values. Efficiency of the rejuvenation step obviously relies on a sensible choice of K_t (while assessing the efficiency for a given kernel may be a difficult task in practice).

Particle filter methods will refer to algorithms which provide consistent inferences from a sequence of distributions π_t, by iterating the three following steps:

---

1. Reweighting

   Compute $w_j \propto \frac{\pi_{t+1(\theta_j)}}{\pi_t(\theta_j)}$.

2. Resampling

   Resample $(\theta_j, w_j)_{j=1..H} \rightarrow (\theta_j^r, 1)_{j=1..H}$ (according to a given selection scheme).

3. Move

   Draw $\theta_j^m \sim K_{t+1}(\theta_j^r, .)$ with K_{t+1} transition kernel of stationary distribution π_{t+1}.

Recent results on the convergence of such algorithms (when $H \rightarrow \infty$) can be found in Crisan and Doucet (2000).

3 Importance sub-sampling (ISS)

Our aim is to estimate a parameter of fixed dimension $\theta_0 \in \Theta$, from N observations $y_1, ..., y_N$ drawn from the "static" model $\mathcal{P}(\theta_0)$ (in contrast with dynamic models presented before, whose parameters may vary over observations). Therefore, the only posterior distribution of interest is $\pi(\theta|y_{1:N})$.

It is possible to endow this static model with a dynamical structure. Suppose only the first n observations ($n < N$) are available at first. Inference procedures can be managed on the subsample $y_{1:n}$, involving for instance simulations from the posterior $\pi(\theta|y_{1:n})$, or more generally a particle system with target $\pi(\theta|y_{1:n})$. Then assume p new observations are available. Provided p is not too large, $\pi(\theta|y_{1:n})$ and $\pi(\theta|y_{1:n+p})$ are likely to be similar; hence we can take advantage of the first inference results on $\pi(\theta|y_{1:n})$, to shorten the calculation cost of incorporating the new data, by a proper reweighting of the particles by the incremental weight

$$w_{n,p}(\theta) = \frac{\pi(\theta|y_{1:n+p})}{\pi(\theta|y_{1:n})} \propto \frac{\mathrm{P}(y_{1:n+p}|\theta)}{\mathrm{P}(y_{1:n}|\theta)} = \mathrm{P}(y_{n+1:n+p}|y_{1:n}, \theta).$$

We will call this particular case of importance sampling *importance sub-sampling* (ISS). We give now important properties of $w_{n,p}(\theta)$ which highlight the interest of ISS.

First, provided $\mathrm{P}(y_{n+1:n+p}|y_{1:n}, \theta)$ is bounded from above in θ (a clearly weak condition), $w_{n,p}(\theta)$ is bounded too, which ensures that $V_{\pi_{n+p}/\pi_n}(h)$ and $\tau_{\pi_{n+p}/\pi_n}(h)$ are finite, for any h such that $E_{\pi_{n+p}}\{h(\theta)\}$ exists (end of §2.1).

Moreover, the impoverishment of the particle system introduced by importance sub-sampling can be approximately evaluated by this asymptotic result:

Theorem 1 *Under some regularity conditions (listed in appendix), and for any $h \in \mathcal{C}^3(\Theta \to \mathbb{R}^L)$, $L \in \mathbb{N}$, such that the integrals $\int h(\theta)\pi_n(\theta)d\theta$, $\int h(\theta)h(\theta)'\pi_n(\theta)d\theta$, and $\int h(\theta)h(\theta)'\pi_{n+p}(\theta)^2/\pi_n(\theta)d\theta$ exist for all $n, p \in \mathbb{N}$, we have*

$$\tau_{\pi_{n+p}/\pi_n}(h) = O(I_L) \qquad \textit{as } n \to \infty, \frac{p}{n} \to r > 0.$$

A proof is given in the appendix.

We see that, when n is large enough, the relative precision of ISS only depends on the proportion of new data, and given that proportion, any relative precision may be attained with a sufficient amount H of particles, where H does not depend on n.

Finally, in two important cases, the reweighting step only operates on the new data $y_{n+1:n+p}$, thus avoiding a second complete browse of the first part of the data $y_{1:n}$; when the observations are independent:

$$w_{n,p}(\theta) \propto \mathrm{P}(y_{n+1:n+p}|y_{1:n}, \theta) = \prod_{i=1}^{p} \mathrm{P}(y_{n+i}|\theta),$$

and when they are Markov, that is, there exists an integer m such that $\mathrm{P}(y_{n+1}|y_{1:n}, \theta) = \mathrm{P}(y_{n+1}|y_{n-m+1:n}, \theta)$:

$$w_{n,p}(\theta) \propto \mathrm{P}(y_{n+1:n+p}|y_{1:n}, \theta) = \prod_{i=1}^{p} \mathrm{P}(y_{n+i}|y_{n+i-m:n+i-1}, \theta).$$

In both cases, the reweighting step indeed makes a quick update of the particle system possible, provided the related likelihoods $\mathrm{P}(y_{n+i}|\theta)$ or $\mathrm{P}(y_{n+i}|y_{n+i-m:n+i-1}, \theta)$ are computable. In other cases, ISS is less interesting, given that the calculation of $\mathrm{P}(y_{n+1:n+p}|y_{1:n}, \theta)$ may become too intensive as n grows. Therefore, we will assume in the following that the observations are either Markov or independent, and we will adopt a common notation for both cases, that is $\mathrm{P}(y_{n+1}|y_{1:n}, \theta) = \mathrm{P}(y_{n+1}|y_{n-m+1:n}, \theta)$, with $m = 0$ if the y_i's are independent (in that case, $y_{n+1:n}$ will stand for $\varnothing$).

4 The ISIS algorithm

The ISIS algorithm (Importance Sub-sampling Iterative Scheme) is a particle filter method devoted to iterated applications of importance sub-sampling. It consists of iterations of the following steps:

```
0. Initialization
   Generate a particle system $(\theta_j, w_j)_{j=1,...,H}$ which targets the initial distribution $\pi_{n_0}$.
1. Reweighting
   Compute $w(\theta_j) \propto \frac{\pi_{n+p(\theta_j)}}{\pi_n(\theta_j)} \propto P(y_{n+1:n+p}|y_{1:n}, \theta_j)$ for $j = 1, ..., H$.
2. Resampling
   Resample the particle system: $(\theta_j, w_j)_{j=1,...,H} \to (\theta_j^r, 1)_{j=1,...,H}$ (according to a given selection scheme, see §2.2).
3. Move
   Generate $\theta_j^m \sim K_{n+p}(\theta_j^r, .)$ with $K_{n+p}$ transition kernel of stationary distribution $\pi_{n+p}$, for $j = 1, ..., H$.
4. Loop
   $n \leftarrow n + p$, $\theta_j \leftarrow \theta_j^m$ , back to 1.
```

The algorithm stops when $n = N$, i.e. when the particle system targets the distribution of interest $\pi(\theta|y_{1:N})$.

To complete the definition of the algorithm, the following features will be specified in the next subsections: choice of the transition kernel K_{n+p} (how to move?), determination of p (how many new observations can we handle in one step?) and practical choice of H (how many particles are needed to achieve a given precision?).

These specifications will be calibrated upon three important considerations:

1. robustness of the estimates: to avoid too volatile results, the algorithm must check for and efficiently restrict the system degeneracy, at any stage.

2. low execution time: alternatively, we need to avoid extra computational costs, which may induced for instance by too severe settings with regard to degeneracy (such as a high frequency of the move steps).

3. "black box" feature: the ISIS algorithm internal machinery should not depend on the considered model. The user would just have to supply a likelihood calculation routine (needed in the reweighting step) and a particle system initialized to π_{n_0}(with $n_0 = 0$ for instance, if the prior defines a true probability distribution).

Finally, the last subsections (4.4 to 4.6) will be devoted to some important features of the ISIS algorithm, which should appear as clear advantages over other estimation procedures.

4.1 How to move?

Efficiency in the move step is critical, since it is the most computationally demanding step (it requires a complete browsing of the past observations $y_{1:n}$, unlike the two others steps). MCMC techniques usually involve numerous applications of a transition kernel over a single "particle". On the contrary, the move step of particle filter methods consist of a single application of a given kernel, for a large set of particles. Thus the choice of a kernel must rely on different criteria that those usually mentioned in MCMC settings: theoretical convergence for instance does not ensure that the particles should move efficiently in a single application of the kernel. Moreover, assessing correctly how much the system really was rejuvenated is not an easy task in practice: for example, using an Hastings-Metropolis (HM) kernel, one may assess the rejuvenation through the acceptance rate, considering that the "accepted" particles are new particles introduced in the system, and that the greater the number of new particles, the better. Assume the "proposed" value is generated from a random walk $\theta_j^p = \theta_j + \varepsilon_j$ (random walk Hastings-Metropolis), high acceptance rates can artificially be achieved by imposing a very low variance of the random step ε_j: in that case, identical replicates of a single particle will be replaced by an equivalent number of particles taking distinct but very close values, hence being highly correlated between each others. System impoverishment remains high but is not detectable any longer.

Rather, it seems sensible to select a transition kernel which depends weakly on the previous value of the moved particle. Such is the case with an independent Hastings-Metropolis kernel (IHM): the proposed particle is generated independently from an instrumental distribution g, and the moved particle only depends from its previous value through the acceptance probability. Using an IHM kernel, the acceptance rate seems a far more reliable efficiency indicator.

Remark *When the rejuvenation is performed by a independent Hastings-Metropolis kernel, browsing through the whole past subsample is needed in the computation of $\pi_{n+p}(\theta_j^p)$, which appears in the acceptance probability.*

Provided we use an IHM kernel, high acceptance rates will be obtained with an instrumental distribution g close to the target distribution π_{n+p}. However, the sole information we can get on π_{n+p} is given by the particle system itself, which can provide at the current stage an estimate of any expectation $E_{\pi_{n+p}}\{h(\theta)\}$:

$$\hat{\mu}(h) = \frac{\sum_{j=1}^{H} w_j h(\theta_j)}{\sum_{j=1}^{H} w_j}$$

In particular, the estimates

$$\hat{E}_{n+p} = \frac{\sum_{j=1}^{H} w_j \theta_j}{\sum_{j=1}^{H} w_j}, \; \hat{V}_{n+p} = \frac{\sum_{j=1}^{H} w_j \{\theta_j - \hat{E}_{n+p}\}\{\theta_j - \hat{E}_{n+p}\}'}{\sum_{j=1}^{H} w_j},$$

give a rough localization of the mass of π_{n+p}. Thus the instrumental distribution $\mathcal{N}(\hat{E}_{n+p}, \hat{V}_{n+p})$ seems a reasonable choice: it is simple (even when the space dimension is important, we can still easily take into account correlations between components of θ), it is not "model-dependent" (hence it fulfills our black-box requirement) and it is asymptotically justified (as $n \to \infty$, π_n tends towards a normal distribution). The later point is essential: as we said, the rejuvenation step becomes more and more intensive as $n \to \infty$, but fortunately gets more and more efficient at the same time.

However, at finite range, a complex posterior distribution can significantly differ from a gaussian distribution, featuring for instance many local modes, or thick distribution tails. In the later case, note that the posterior distribution tails can be easily reduced by a proper reparametrisation: if for instance the density of π_n decreases towards infinity like $\propto \exp(-K\theta)$ $(\theta > 0)$, we get thinner, gaussian-like tails, by replacing θ by $\theta' = \theta^{\frac{1}{2}}$. And reparametrising the model is likely to take less time to the user that deriving a distinct instrumental distribution with tails of the same order than the considered target distribution. This is another appeal of our black box approach. Nevertheless, the instrumental distribution we chose still could be suspected to hardly move particles from a highly polymodal target distribution. But we will see that, in such settings, our method can lead to satisfactory results all the same (see §5.2). For all these reasons, we see $\mathcal{N}(\hat{E}_{n+p}, \hat{V}_{n+p})$ as a convenient all-purpose instrumental distribution, although a greater efficiency can be expected from more refined, but case-restricted instrumental distributions.

4.2 How many new observations can we handle in one step?

It seems clear that $V_{\pi_{n+p}/\pi_n}(h)$ and $\tau_{\pi_{n+p}/\pi_n}(h)$, for a given h, are increasing functions of p. The support of π_{n+p} progressively shrinks as p increases, comparatively to the initial support of π_n. Thus, each added observation impoverishes a bit more the particle system. Our algorithm can smoothly incorporate new observations, since a reweighting step is a transitive operation: $w_{n,p}(\theta) = \prod_{k=1}^{p} w_{n+k-1,n+k}(\theta)$. Therefore, we can increase p, the amount of added observations, while regularly checking for degeneracy, and stop this process when a given degeneracy level is attained. We derive in this subsection a degeneracy criterion that will be used as a stopping rule for the incorporation of observations. Notice that we cannot afford too cautious a criterion, for this would imply a too high frequency for the move steps (as we already said, these steps are the most intensive).

Due to our particular choice of instrumental distribution (see §4.2), the information we draw from the particle system at an intermediary stage $n+p$ is summarized by the two estimates $\hat{E}_{n+p}$ and $\hat{V}_{n+p}$. These two quantities localize the region of high π_{n+p}-probability, which will be explored more precisely by the adjunction of new points. Therefore, we need to check that the variance of these estimates is sufficiently low to prevent the instrumental distribution of the moving step from missing the target.

As a preliminary remark, note that the Kullback-Leibler distance between two normal distributions $\mathcal{N}_1 = \mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}_2 = \mathcal{N}(\mu + \sigma\eta, \sigma^2(1+\tau)^2)$ is $\Delta = \ln(1+\tau) + \frac{1}{2}[(1+\eta^2)/(1+\tau)^2 - 1] \simeq \eta^2/2 + \tau^2$ (Taylor expansion at order 2 in η and τ). Therefore, if η and τ are small, the quantity $\eta^2/2 + \tau^2$ is an acceptable indicator of similarity between $\mathcal{N}_1$ and $\mathcal{N}_2$; in particular if $\mathcal{N}_1$ is considered as an efficient instrumental law for a given Hastings-Metropolis kernel, this indicator can (roughly) measure to which point the kernel may keep a comparable efficiency if we replace $\mathcal{N}_1$ by $\mathcal{N}_2$. Now suppose $\mathcal{N}_1$ and $\mathcal{N}_2$ stand for, respectively, $\mathcal{N}(E_{\pi_{n+p}}(\theta), \mathrm{V}_{\pi_{n+p}}(\theta))$ and $\mathcal{N}(\hat{E}_{n+p}, \hat{V}_{n+p})$. In that case, and in an unimodal setting, we have

$$\eta^2 = \frac{\{\hat{E}_{n+p} - E_{\pi_{n+p}}(\theta)\}^2}{\mathrm{V}_{\pi_{n+p}}(\theta)}, \qquad (1+\tau)^2 = \frac{\hat{V}_{n+p}}{\mathrm{V}_{\pi_{n+p}}(\theta)}.$$

We do not argue that $\mathcal{N}_1$ is the optimal instrumental distribution for moving π_{n+p}, even in the restricted class of the gaussian distributions; we assume, as in the previous subsection, that it is an acceptable choice. Since its moments are not directly available, we have to replace it by $\mathcal{N}_2$, which may be unable to move efficiently if too far from $\mathcal{N}_1$, and therefore from π_{n+p}.

The quantities η and τ cannot be computed directly. However, since the order of magnitude of $\{\hat{E}_{n+p} - E_{\pi_{n+p}}(\theta)\}^2$ and $\{\hat{V}_{n+p} - \mathrm{V}_{\pi_{n+p}}(\theta)\}^2$ are given, respectively, by $\mathrm{V}(\hat{E}_{n+p})$ and $\mathrm{V}(\hat{V}_{n+p})$, we can write

$$\eta^2 \approx \frac{\mathrm{V}(\hat{E}_{n+p})}{\mathrm{V}_{\pi_{n+p}}(\theta)}, \qquad \tau^2 \approx \frac{\mathrm{V}(\hat{V}_{n+p})}{4\mathrm{V}_{\pi_{n+p}}(\theta)^2},$$

since

$$\tau^2 = (1+\tau-1)^2 = \left[\{1 + \frac{\hat{V}_{n+p} - \mathrm{V}_{\pi_{n+p}}(\theta)}{\mathrm{V}_{\pi_{n+p}}(\theta)}\}^{\frac{1}{2}} - 1\right]^2 \simeq \frac{\{\hat{V}_{n+p} - \mathrm{V}_{\pi_{n+p}}(\theta)\}^2}{4\mathrm{V}_{\pi_{n+p}}(\theta)^2}.$$

We then derive the following degeneracy indicator, from the quantity $\eta^2/2 + \tau^2$:

$$D_{n,p} = \frac{1}{2}\frac{\mathrm{V}(\hat{E}_{n+p})}{\mathrm{V}_{\pi_{n+p}}(\theta)} + \frac{1}{4}\frac{\mathrm{V}(\hat{V}_{n+p})}{\mathrm{V}_{\pi_{n+p}}(\theta)^2},$$

and adopt the corresponding criterion:

$$\text{increase } p \text{ until } D_{n,p} > d,$$

where d is an arbitrary threshold ($d = 10^{-2}$ in our numerical applications).

We defined this indicator first in an unidimensional setting, for the sake of simplicity. Replacing the variances by their norms is a natural generalization in the multidimensional case:

$$D_{n,p} = \frac{1}{2}\left\|\mathrm{V}(\hat{E}_{n+p})\right\| \left\|\mathrm{V}_{\pi_{n+p}}(\theta)\right\|^{-1} + \frac{1}{4}\left\|\mathrm{V}(\hat{V}_{n+p})\right\| \left\|\mathrm{V}_{\pi_{n+p}}(\theta)\right\|^{-2}.$$

The quantity $\mathrm{V}_{\pi_{n+p}}(\theta)$ in $D_{n,p}$ can be directly estimated by $\hat{V}_{n+p}$. On the contrary, the variance terms $\mathrm{V}(\hat{E}_{n+p})$ and $\mathrm{V}(\hat{V}_{n+p})$ cannot be easily evaluated, because the successive reweighting and move steps made the joint distribution of the particles quite intractable. To address this issue, we propose to replace these variances by empirical approximations. Notice first that the particle system $(\theta_j, w_j)_{j=1,\dots,H}$ is made of both unique particles (those generated by an accepted move) and replicated particles (resampled replicates of particles whose move was not accepted). Hence, the particle system can also be represented by a sequence $(\widetilde{\theta}_j, \widetilde{w}_j)_{j=1,\dots,H'}$ where the $\widetilde{\theta}_j$'s are pairwise distinct, and the $\widetilde{w}_j$'s properly weight the particle system, that is for any h:

$$\hat{\mu}(h) = \frac{\sum_{j=1}^{H} w_j h(\theta_j)}{\sum_{j=1}^{H} w_j} = \frac{\sum_{j=1}^{H'} \widetilde{w}_j h(\widetilde{\theta}_j)}{\sum_{j=1}^{H'} \widetilde{w}_j}$$

(in practice, if a $\widetilde{\theta}_{j'}$ in the new system corresponds to a θ_j in the previous one, its weight $\widetilde{w}_{j'}$ will be w_j times the number of replicates of θ_j). Denote $\overline{w}_j = \widetilde{w}_j / \sum \widetilde{w}_j$ the corresponding normalized weights, and assume, at a first approximation, that the couples $(\widetilde{\theta}_j, \overline{w}_j)$ are *independent* random variables. The variance of $\hat{\mu}(h)$ is then estimated by

$$\hat{\Sigma}(h) = \sum_{j=1}^{H'} \overline{w}_j^2 \{h(\widetilde{\theta}_j) - \hat{\mu}(h)\}\{h(\widetilde{\theta}_j) - \hat{\mu}(h)\}'.$$

Using this approximation for both $h(\theta) = \theta$ and $h(\theta) = \{\theta - \hat{\mu}(h)\}\{\theta - \hat{\mu}(h)\}'$, we can deduce $D_{n,p}$.

Obviously, this new indicator is rough and empirical. Yet our assumption of independence is not so far from reality: as mentioned in §4.1, an independent Hastings-Metropolis kernel for the move step strongly limits the dependence between (distinct) particles. For this kernel, a moved particle, conditionally on the fact it was successfully moved, is independent by construction. Therefore, each new value in the particle set has been drawn independently at some stage, and the $\widetilde{\theta}_j$'s are nearly independent (this is not completely true, since the $\widetilde{\theta}_j$'s were drawn from an instrumental distribution which depended itself on moments computed from the particle system, but this dependence is clearly weak). In other words, we assume that the main part of estimates variance is due to the particles replication in the reweighting step, and that the move step, if successful, indeed provides a "brand new" particle.

Finally, Theorem 1 indicates that the level of degeneracy induced by the reweighting step is asymptotically given by the proportion p/n of new points:

thus, if we apply the same criterion at each stage, and assume a constant level of efficiency for the move steps, we expect the number of new points to increase at a geometric speed. Actually, the move steps are likely to be more and more efficient as n increases (see §4.1), hence the incorporation may be even faster.

4.3 How many particles are required?

We must keep in mind that particle filter methods are only justified asymptotically (in H, the number of particles). Therefore, it would be tempting to run the ISIS algorithm with the highest possible values of H, but this would imply a tremendous computational cost.

In practice, we have no choice but determining H by a proper calibration. The aim is to find, for a given problem, a reasonable value of H, in computational terms, that still leads to robust estimates. One can for instance run various times the algorithm to measure the variability of the obtained results for a given number of particles (in the independent case, the observations should be shuffled at each run, to check that the results do not depend on the order of incorporation of the observations).

Note however that, in our numerical applications, we never faced any case of degeneracy, or high estimates variability, as long as H exceeded a very reasonable bound (20000), although we were confronted to very complex posterior distributions, such like mixture posterior distributions over a parameter space of high dimension (up to 20).

4.4 Computational considerations and parallel processing issue

The ISIS algorithm explores the parameter space Θ in a significantly distinct way, compared with others estimation procedures: it tracks numerous particles while smoothly browsing the sample of observations $y_{1:N}$, whereas most estimation algorithms track a single particle through numerous iterations, each of them requiring a complete browsing of the dataset $y_{1:N}$ (particle referring here to a single parameter value varying at each iteration).

This distinct exploration strategy strongly limits the number of accesses to the dataset, and therefore can bring important computational savings, most obviously when the sample size is large. In fact, the ISIS algorithm is able to roughly localize the region of interest, by drawing information from only a small part of the sample, before exploring it more precisely, whereas its competitors would need to perform computations over the whole sample to achieve the same localization.

Moreover, this distinct exploration structure strongly facilitates a parallel processing implementation. Parallel processing here refers to the possibility to share execution time of a given algorithm between various processors, or computers. Computational savings may be significant (ideally in a factor $1/K$ if K processors are used), provided the considered algorithm can be divided in K routines running independently. If this is not the case, too numerous data exchanges between processors may strongly slow the processing, and might even cancel any saving.

Most estimation procedures can hardly be processed in parallel however, since they consist in a sequence of dependent iterations: would an iteration at time t be divided in K independent computations, corresponding results need to be gathered in order to initiate iteration at time $t+1$. Too frequent exchanges between processors would occur. On the contrary, the ISIS algorithm tracks numerous particles through a reasonable amount of iterations. Thus, parallel processing is easily implementable: partition the particle system into K subsets S_k $(k = 1, ..., K)$, and dedicate computations over particles in S_k to processor k. This partitioning must occur during two distinct stages: the reweighting step (the processor k computes the incremental weights of particles in S_k) and the move step (the processor k "moves" the particles in S_k). The resampling step is much less intensive, and does not really need this partitioning. At the end of each of these two steps (reweighting and move), results must be gathered before moving on the next step. Since the ISIS algorithm has the ability to take into account new data at a nearly geometric speed (see end of §4.2), the frequency of move steps quickly decreases as n grows. Therefore the frequency of required information exchanges between processors also decreases at an exponential speed in n, making the parallel processing of the reweighting and the move steps quickly highly efficient.

4.5 Economies of scale

To run properly, the ISIS algorithm need to be supplied a routine computing the partial likelihood $\mathrm{P}(y_{n+1}|y_{n-m+1:n}, \theta)$ for any value of $y_{n-m+1:n+1}$ and θ. In fact, the ISIS algorithm must call intensively this routine, for each particle and each added observation (during the reweighting step in order to compute the corresponding incremental weights, and during the move step, in order to compute the acceptance probabilities of the Hastings-Metropolis kernel). Therefore, the computational feasibility of the ISIS algorithm directly depends on the execution time on this routine. In certain cases, $\mathrm{P}(y_{n+1}|y_{n-m+1:n},\theta)$ may appear at first sight too complex a likelihood, particularly when it is expressed as an integral. Yet economies of scale are in fact easily achievable, and the practitioner should keep in mind that in many cases x evaluations of $\mathrm{P}(y_{n+1}|y_{n-m+1:n},\theta)$ can be cleverly performed in a rather lower time that x times the time needed to evaluate one such quantity. Here are two illustrative examples:

- the likelihood involves a "costly" function $\Psi : \mathbb{R} \rightarrow \mathbb{R}$(in computational terms). One can store preliminary evaluations of Ψ and its first derivatives $\Psi^{(l)}$ over a "grid" of points z_k, and then replace any further computation of $\Psi(z)$ by an appropriate Taylor approximation.
- the likelihood involves an integral $\int \phi(z)\varphi_i(z)dz$; (where the function φ_i differs for each evaluation i). This integral can be evaluated through a given amount of evaluations of its integrand $\phi(.)\varphi_i(.)$. The $\phi(.)$ may be stored the first time they are computed, and reused as often as necessary.

These two numerical recipes can be applied for generalized linear models (see §5.1) and show computational savings far from anecdotal. Economies of scale may be workable in many other cases, which is another appeal of the ISIS algorithm approach.

4.6 ISIS in a latent variable context

Suppose the considered model involves a latent variable structure, that is to say observations y_n may be related to unobserved z_n, such that

$$y_n|\{z_n = z\} \sim \mathcal{P}(\theta_0; z).$$

For such models, an inference can be obtained by methods such as the Gibbs sampler or the EM algorithm, which explore the augmented space of parameters and latent variables, through successive moves along each dimension of this augmented space, while the other dimensions keep the same values. In fact, because of its (much) greater dimension, the augmented space may be rather more difficult to explore efficiently. Moreover, these algorithms imply a greater adaptation cost (in terms of practitioner time), since the intermediary movements along latent variables dimensions they involve are specific to a given model (using a Gibbs sampler for instance, the practitioner must derive for each model a distinct scheme to perform the intermediary latent variable simulations). In other words, these algorithms are not black boxes, unlike the ISIS algorithm. Finally, when the sample size N is important, the complete browsing of the sample at each iteration, already stressed in §4.4, is even more costly than in other cases, since it involves at each iteration N storages and recalls of latent variables values. Thus large databases may be difficult to handle by these methods, even at an age of great computational power availability. In that settings, using our algorithm seems more recommended, provided that the marginal likelihood $\mathrm{P}(y_n|\theta) = \int \mathrm{P}(y_n, z_n|\theta) dz_n$ can be computed.

5 Examples

5.1 Generalized linear models

Generalized linear models (McCullagh and Nelder, 1983) relate independent observations (y_i) to their covariates $(x_i) = (x_i^1, ..., x_i^K)$ through an exponential density of the form

$$\mathrm{P}(y_i|\theta_i) = m(y_i) \exp(y_i \theta_i - \psi(\theta_i)), \qquad y_i \in \mathcal{Y},$$

where the canonical parameter θ_i is determined by the linear relation

$$h(E_{\theta_i}(y_i)) = (h \circ \psi')(\theta_i) = x_i' \beta.$$

The (known) function h is called the "link function". The row vectors of covariates x_i form a full-rank matrix X. $\mathcal{Y}$ may be either discrete or continuous.

Generalized linear models are usually estimated through MCMC techniques: the Gibbs sampler appears at times a natural choice, especially when a latent variable structure can be exhibited (in binary or polytomic regression, for instance). However, simulating from the conditional distributions which constitute the Gibbs sampler is only feasible in some very restrictive cases, such as the probit model (Albert and Chib, 1993). In other cases,

k	1	2	3	4	5
β_0	-1	0.7	-0.5	-0.1	-0.3
$\hat{\beta}=\hat{E}_N$	-0.97 (0.017)	0.75 (0.020)	-0.60 (0.012)	-0.11 (0.011)	-0.36 (0.013)
$\hat{V}_N$	0.0033 ($2.9.10^{-4}$)	0.0032 ($3.2.10^{-4}$)	0.0034 ($4.0.10^{-4}$)	0.0024 ($1.8.10^{-4}$)	0.0029 ($3.1.10^{-4}$)

Table 1: Estimates for a simulated probit model ($H = 2000$, $N = 1000$)

a Gibbs sampler may still offer an inference from an artificial distribution which approximates the distribution of interest (Kolassa, 1999).

On the contrary, the ISIS algorithm can handle directly any generalized linear model such that $\mathrm{P}(y|x,\beta)$ is computable. Moreover, thanks to the exponential nature of the density, such a model often leads to a very regular and unimodal posterior distribution, making the gaussian instrumental distribution of the ISIS algorithm able to move efficiently the particles at early stages (see Figure 1).

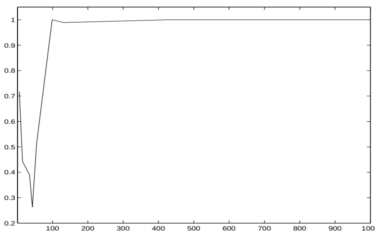
We first consider the probit model:

$$y|x,\beta \sim \mathcal{B}(1,\Phi(x'\beta)), \qquad \mathcal{Y}=\{0,1\}.$$

In that case, $\mathrm{P}(y|x,\beta)=\Phi(x'\beta)^y\Phi(-x'\beta)^{1-y}$, and the algorithm must perform numerous evaluations of the function Φ. In order to get substantial "economies of scales", and therefore reduce the execution time, we implement our Taylor approximation method (§4.5) for $\phi(z)=\log\Phi(z)$ (Computing ϕ instead of Φ is profitable, since the logarithm transforms products into sums).

This method provides an approximate of $\phi(z)$, for any $z\in[-5,5]$, up to an absolute error of 10^{-6}, by performing only two sums and two products, with a few initial evaluations of ϕ, ϕ' and ϕ'' (over a grid of 500 points). Note that an absolute error over ϕ correspond to a relative error over Φ of the same magnitude.

We estimated a probit model with $K=5$ covariates. The observations sample was simulated from the true model, for a given parameter β_0 ($N=1000$). Covariates were simulated as well (from independent standard normal distributions, except the first row, which was constant). Table 1 gives the resulting estimates; the numbers in parenthesis give the standard deviation of these results over S runs ($S=10$). We can see that, with a reasonable number of particles ($H=2000$), estimates are relatively robust. Figures 1 and 2 give the evolution of the acceptance rate and the elapsed execution time



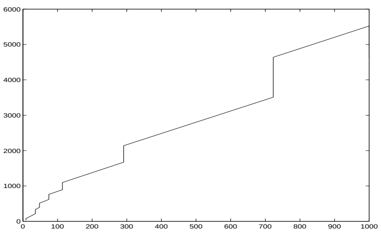


Figure 1: Acceptance rate against n

Figure 2: Execution time against n

given n (the amount of data already incorporated). As we stated in §4.1, the move steps frequency decreases very quickly (the move steps correspond to the vertical parts of the time plot in Figure 2), while their acceptance rates are already close to one at early stages.

Note that adapting our algorithm to the logit model is immediate (Φ is replaced by $L : z \to e^z/(1+e^z)$), while for this model the Gibbs sampler cannot be implemented directly (Albert and Chib, 1993).

We finally consider a generalized two-dimensions probit model:

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(X' \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

$$Y|Z = \begin{pmatrix} 1_{Z_1>0} \\ 1_{Z_2>0} \end{pmatrix}, \qquad \mathcal{Y} = \{0,1\}^2.$$

It can be easily shown that

$$\mathrm{P}(y = (1,1)|x,\beta) = \int_{-x'\beta_1}^{+\infty} f_N(t)\Phi(\frac{x'\beta_2 + \rho t}{1-\rho^2})dt,$$

where f_N stands for the standard gaussian density (for the others possible values of y, the likelihood takes a similar form).

When computing this likelihood, economies of scales can still be achieved (see §4.5). The integral may be evaluated by classical integration methods (such as Romberg integration), which require multiple evaluations of the integrand over a grid of points t_k. Yet the $f_N(t_k)$ only need to be evaluated once, and the evaluations of $\Phi\left((x'\beta_2 + \rho t_k)/(1-\rho^2)\right)$ may again be accelerated by a Taylor approximation.

However, the computation time of $\mathrm{P}(y|x,\beta)$ may remains costly, making parallel processing implementation strongly recommended for this model.

5.2 Mixtures

Mixture models take the following form:

$$\left\{\begin{array}{l} z_i \sim \mathcal{M}(k; p_1, .., p_k) \\ y_i | \{z_i = l\} \sim \mathrm{P}(\zeta_l) \end{array}\right. \qquad i = 1, ..., N,$$

where the z_i's are unobserved and independent. The aim is to estimate $\theta = (p_1, .., p_{k-1}, \zeta_1, .., \zeta_k)$ ($p_k = 1 - p_1 - ... - p_{k-1}$ can be discarded). In this example, we will take $\mathrm{P}(\zeta_l) = \mathcal{N}(\mu_l, \sigma_l^2)$. A constraint over the parameters (such as $p_1 < ... < p_{k-1}$, $\mu_1 < ... < \mu_{k-1}$ or $\sigma_1 < ... < \sigma_k$) is needed to fully identify the model. Celeux et al. (2000) showed that an exploration of the unconstrained posterior distribution was manageable and could lead to satisfactory results, but such an approach is not compatible with our algorithm (since a transition kernel with a gaussian instrumental distribution would hardly "move" a target distribution with $k!$ equivalent modes).

Since latent variables z_i are discrete, the likelihood $\mathrm{P}(y_{n+1:n+p} | y_{1:n}, \theta)$ is easily computable (as a sum over the possible states of the z_i). Difficulties about mixture models are not related to calculation issues, like in the previous example, but to the complexity of the posterior $\pi(\theta | y_{1:N})$, which involves a great number of local modes over a large dimension space, making standard algorithms unable to converge. More refined methods, such as the Gibbs sampler, theoretically converges to the region of highest probability, but can still lead to "trap modes" in practice (Robert and Mengersen, 1999). Hence, the concern here is to see if our algorithm is able to manage such a polymodal target distribution, even when the move steps are performed by a normal instrumental distribution (see §4.1).

The constraint was chosen as $\mu_1 < ... < \mu_k$. The standard errors σ_j were reparametrized: $s_j = \log(\sigma_j)$, in order to lower the posterior distribution tails in σ (see §4.1). Correct choice for a prior distribution is still a sensitive issue in mixture modeling (see for instance citepRobMeng. Since our aim was not to address this problem, we restrained ourselves to a simple yet reasonable prior, that is the uniform distribution over the compact set corresponding to the following constraints:

$$\begin{array}{rl} \forall j < k, p_j \geq 0, & p_1 + ... + p_{k-1} \leq 1, \\ \underline{\mu} < \mu_1 < ... < \mu_k < \overline{\mu}, & \forall j \leq k, s_j \in [\underline{s}, \overline{s}]. \end{array}$$

where $\underline{\mu}$, $\overline{\mu}$, $\underline{s}$, $\overline{s}$ must be adapted to the considered sample (in our examples, we took: $\underline{\mu} = \min(y_i)$, $\overline{\mu} = \max(y_i)$, $\overline{s} = \log[(\overline{\mu} - \underline{\mu})/6]$, $\underline{s} =$

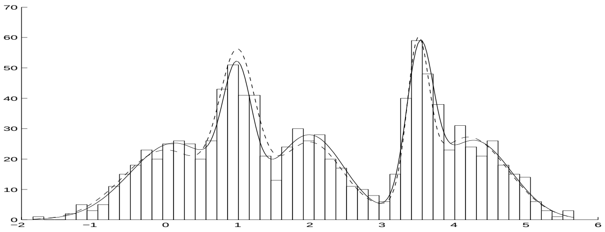


Figure 3: Histogram and density (solid line) for the ISIS estimate against the true density (dashed line) for a simulated sample ($N = 1000$, $k = 5$, $H = 10000$)

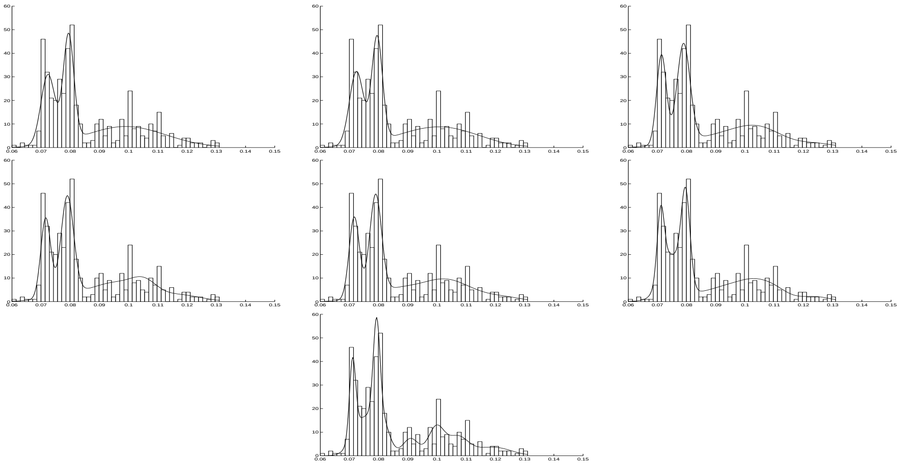


Figure 4: Densities corresponding to the ISIS estimates, for mixture models with resp. 3 to 9 components (Hidalgo stamps dataset, $H = 20000$)

$\log[\min|y_i - y_j|/2]$).

Our algorithm was first applied to simulated data. Figure 3 contrasts the density corresponding to the ISIS estimate from the true density of the simulated data (mixture model with five components). The first density clearly fits well the data.

Our algorithm was also applied to the well known 1872 Hidalgo dataset (first analysed by Izenman and Sommer, 1988, it consists of 485 observed stamp thicknesses for the Mexican 1872 Hidalgo issue), for various numbers of mixture components (3 to 9, see Figure 4). Again, results appear quite satisfactory, and seem in agreement with authors (Izenman and Sommer, 1988; Robert and Mengersen, 1999) who recommend a three-component mixture modeling.

6 Conclusion

The ISIS algorithm is a quick and efficient method for evaluating estimates, in a rather general context. Theoretically, it can handle any model with independent or Markov observations. In practice, we have seen (§5.2) that it could cope with very complex multidimensional posterior distributions, such as the mixture posterior distributions. Moreover, it is indeed a "black box", in the sense that the main adaptation cost to a given problem reduces to supplying a likelihood calculation routine adapted to the considered model (unlike the Gibbs sampler, for instance).

Another advantage of the ISIS algorithm is its computational efficiency: while most rival algorithms need a complete browsing of all observations $y_{1:N}$ for each iteration, the ISIS algorithm strongly limits the number of access to large parts of the sample. Thus the ISIS algorithm may outperform other estimates evaluation methods for large datasets, in computational terms. Finally, the ISIS algorithm allows a parallel processing implementation, which is another clear advantage over its competitors.

Acknowledgements

I have benefited from valuable discussions with Christian P. Robert and Judith Rousseau. This paper is the first part of my Ph.D thesis, and was partially supported by the EU TMR network ERB-FMRX-CT96-0095 and by the HSSS program via the research kitchen workshop of Villard de Lans, January 2000. I dedicate it to my wife of ten months.

Appendix - Proof of theorem 1

The parameter space Θ is on an open space of R^K. $y_1, ..., y_n$ are realizations of a given model $P(\theta_0)$ with $\theta_0 \in \Theta$.

To simplify the notation, we summarize the dependence on the observations $y_1, ..., y_n$ by a single subscript n : for instance, $L_n(\theta)$ will refer to the likelihood $L_n(y_{1:n}|\theta)$, $l_n(\theta)$ to the log-likelihood, $\pi_n(\theta)$ to $\pi(\theta|y_{1:n})$, and so on.

We will need the following conditions to be fulfilled a.s., hence the considered observations $y_1, ..., y_n$ are supposed to belong to the corresponding set of probability 1:

1. existence of the MLE $\widehat{\theta}_n = \underset{\theta\in\Theta}{\arg\max} L_n(\theta)$ for each n and convergence a.s. towards θ_0 as $n \to \infty$,

2. positive-definiteness of the matrix $\Sigma_n = -[\frac{1}{n}\frac{\partial^2 l_n(\widehat{\theta}_n)}{\partial\theta\partial\theta'}]^{-1}$ and convergence towards $I(\theta_0)$ (Fisher information matrix),

3. existence of $\Delta > 0$ such that:
$$0 < \delta < \Delta \Longrightarrow \limsup_{n\to+\infty}\{\frac{1}{n}\sup_{|\theta-\widehat{\theta}_n|>\delta}[l_n(\theta) - l_n(\widehat{\theta}_n)]\} < 0,$$

4. $\sup_{\theta\in\Theta'}(\frac{1}{n}\frac{\partial^3 l_n}{\partial\theta^3}(\theta))$ is bounded from above in n, for any compact set $\Theta' \subset \Theta$ (the bound does not depend on the observations).

Moreover, l_n is assumed to be 3-times continuously-differentiable.

We first give one of the many versions of the Laplace expansion applied to integrands containing a likelihood. A full proof is given, since the result has been customized to our needs, and arguments presented will be reused next. Further references about the Laplace expansion can be found in Johnson (1970), Tierney et al. (1989), and Schervish (1995).

Lemma 2 (Laplace expansion) *For $\varphi \in \mathcal{C}^3(\Theta \to \Re)$ such that $\int_\Theta \varphi(\theta)L_n(\theta)d\theta < +\infty$ and $\varphi(\theta_0) \neq 0$, we have:*
$$\int_\Theta \varphi(\theta)L_n(\theta)d\theta = (2\pi)^{K/2}\sqrt{\det(\Sigma_n)}n^{-K/2}\varphi(\widehat{\theta}_n)L_n(\widehat{\theta}_n)[1 + O(n^{-1})].$$

Proof.

Take δ small enough so that $\Theta' = B(\theta_0, \delta) \subset \Theta$, $\delta < \Delta$ (see condition 3) and $\varphi(\theta) \neq 0$ for $\theta \in \Theta'$. Assume for instance $\varphi(\theta) > 0$ for $\theta \in \Theta'$ (if not, replace φ by $-\varphi$), and define $l_n^*(\theta) \triangleq l_n(\theta) + \log\varphi(\theta)$. We have
$$\int_\Theta \varphi(\theta)L_n(\theta)d\theta = \int_\Theta \exp\{l_n^*(\theta)\}d\theta = \int_{\Theta'} \exp\{l_n^*(\theta)\}d\theta + \int_{\Theta\backslash\Theta'} \exp\{l_n^*(\theta)\}d\theta,$$

The second integral is clearly exponentially negligible, hence $\int_{\Theta} \exp\{l_n^*(\theta)\}d\theta$ will be replaced by $\int_{\Theta'} \exp\{l_n^*(\theta)\}d\theta$ from now on. We perform the following Taylor expansion:

$$\begin{aligned} l_n^*(\theta) = l_n^*(\widehat{\theta}_n) \quad &+ \quad \frac{\partial \log \varphi}{\partial \theta}(\widehat{\theta}_n).(\theta - \widehat{\theta}_n)' \\ &+ \quad (\theta - \widehat{\theta}_n)\frac{1}{2}\frac{\partial^2 l_n^*}{\partial \theta^2}(\widehat{\theta}_n)(\theta - \widehat{\theta}_n)' + \left\|\theta - \widehat{\theta}_n\right\|^3 \overline{r}_n(\theta) \end{aligned}$$

with $|\overline{r}_n(\theta)| \leq \frac{1}{6}\sup\limits_{\theta \in \Theta'}(\frac{\partial^3 l_n^*}{\partial \theta^3}) \leq \frac{1}{6}nM_{\partial^3 l} + \frac{1}{6}M_{\partial^3 \varphi}$.

The bound $M_{\partial^3 l}$ corresponds to condition 4 and $M_{\partial^3 \varphi} = \sup\limits_{\theta \in \Theta'}(\frac{\partial^3 \log \varphi}{\partial \theta^3})$. Thus we can write $\overline{r}_n(\theta) = nr_n(\theta)$ where $r_n(\theta)$ is a bounded quantity (by $M = \frac{1}{6}M_{\partial^3 l} + \frac{1}{6}M_{\partial^3 \varphi}$ for instance).

Let $\Sigma_n^* \triangleq [-\frac{1}{n}\frac{\partial^2 l_n^*}{\partial \theta^2}(\widehat{\theta}_n)]^{-1} = [\Sigma_n^{-1} - \frac{1}{n}\frac{\partial^2 \log \varphi}{\partial \theta^2}]^{-1}$ (Σ_n^* is definite positive for n large enough, since condition 2 holds).

$$\begin{aligned} \exp\{l_n^*(\theta)\} \quad &= \quad \exp\{l_n^*(\widehat{\theta}_n)\}\exp\{\frac{\partial \log \varphi}{\partial \theta}(\widehat{\theta}_n).(\theta - \widehat{\theta}_n)' \\ &\quad +(\theta - \widehat{\theta}_n)\frac{1}{2}\frac{\partial^2 l_n^*}{\partial \theta^2}(\widehat{\theta}_n)(\theta - \widehat{\theta}_n)' + \left\|\theta - \widehat{\theta}_n\right\|^3 nr_n(\theta)\} \\ &= \quad \exp\{l_n^*(\widehat{\theta}_n)\}\exp\{-\frac{n}{2}(\theta - \widehat{\theta}_n)\Sigma_n^{*-1}(\theta - \widehat{\theta}_n)'\}[1 + R_n(\theta)], \end{aligned}$$

with $R_n(\theta) = \exp\{\ \frac{\partial \log \varphi}{\partial \theta}(\widehat{\theta}_n).(\theta - \widehat{\theta}_n)' + \left\|\theta - \widehat{\theta}_n\right\|^3 nr_n(\theta)\ \} - 1$.

The leading term of our expansion is already available:

$$\int_{\Theta'} \exp\{-\frac{n}{2}(\theta - \widehat{\theta}_n)\Sigma_n^{*-1}(\theta - \widehat{\theta}_n)'\}d\theta \underset{n \to \infty}{\sim} (2\pi)^{K/2}\sqrt{\det(\Sigma_n)}n^{-K/2},$$

and since $\Sigma_n^{*-1} = \Sigma_n^{-1} + O(n^{-1})$ we can replace Σ_n^* by Σ_n.

We now show that:

$$\begin{aligned} &\int_{\Theta'} \exp\{-\frac{n}{2}(\theta - \widehat{\theta}_n)\Sigma_n^{*-1}(\theta - \widehat{\theta}_n)'\}R_n(\theta)d\theta = \\ &\qquad O\left(\frac{1}{n}\int_{\Theta'} \exp\{-\frac{n}{2}(\theta - \widehat{\theta}_n)\Sigma_n^{*-1}(\theta - \widehat{\theta}_n)'\}d\theta\right). \end{aligned}$$

Let $t = T_n(\theta) = \sqrt{n}(\theta - \widehat{\theta}_n)$, $\mathcal{T}_n = T_n(\Theta')$ and $\widetilde{r}_n(t) = r_n \circ T_n^{-1}(t)$.

$$\int_{\Theta'} \exp\{-\frac{n}{2}(\theta - \widehat{\theta}_n)\Sigma_n^{*-1}(\theta - \widehat{\theta}_n)'\}R_n(\theta)d\theta =$$
$$n^{-K/2} \int_{\mathcal{T}_n} \exp\{-\frac{1}{2}t\Sigma_n^{*-1}t'\} \left[\exp\{\sqrt{n}(\|t\|^3 \widetilde{r}_n(t) + \frac{\partial \log \varphi}{\partial \theta}(\widehat{\theta}_n).t')\} - 1\right] dt.$$

Since $\|t\|^3 r_n(t) + \frac{\partial \log \varphi}{\partial \theta}(\widehat{\theta}_n).t' \leq M \|t\|^3 + M_{\partial\varphi} \|t\|$, with $M_{\partial\varphi} = \sup\limits_{\theta \in \Theta'} \frac{\partial \log \varphi}{\partial \theta}$, we can write:

$$\int_{\Theta'} \exp\{-\tfrac{n}{2}(\theta - \widehat{\theta}_n)\Sigma_n^{*-1}(\theta - \widehat{\theta}_n)'\}R_n(\theta)d\theta \leq$$

$$n^{-(K+1)/2} \int_{\mathcal{T}_n} \exp\{-\tfrac{1}{2}t\Sigma_n^{*-1}t'\}[M\|t\|^3 + M_{\partial\varphi}\|t\|]\phi(n^{-1/2}[M\|t\|^3 + M_{\partial\varphi}\|t\|])dt,$$

where $\phi(x) = \frac{e^x - 1}{x}$.

Let n_0 such that $n > n_0 \Rightarrow \|\widehat{\theta}_n - \theta_0\| < \delta/2$. Take a given $t \in R^K$. For any integer n such that $n > \max(n_0, \frac{4\|t\|^2}{\delta^2})$, we have at the same time $t \in \mathcal{T}_n$ (since $\mathcal{T}_n = B(\sqrt{n}(\theta_0 - \widehat{\theta}_n), \delta\sqrt{n})$, and $\|t - \sqrt{n}(\theta_0 - \widehat{\theta}_n)\| \leq \|t\| + \|\sqrt{n}(\theta_0 - \widehat{\theta}_n)\| \leq \sqrt{n}\delta$ for such a n) and

$$\begin{aligned} \phi(n^{-1/2}[M\|t\|^3 + M_{\partial\varphi}\|t\|]) &\leq \phi(\frac{\delta}{2\|t\|}[M\|t\|^3 + M_{\partial\varphi}\|t\|]) \\ &\leq \phi(\frac{\delta}{2}[M\|t\|^2 + M_{\partial\varphi}]), \end{aligned}$$

since ϕ is an increasing function for $x > 0$. Therefore:

$$\int_{\mathcal{T}_n} \exp\{-\frac{1}{2}t\Sigma_n^{*-1}t'\}[M\|t\|^3 + M_{\partial\varphi}\|t\|]\phi(n^{-1/2}[M\|t\|^3 + M_{\partial\varphi}\|t\|])dt \leq$$
$$\int_{R^K} \exp\{-\frac{1}{2}t\Sigma_n^{*-1}t'\}[M\|t\|^3 + M_{\partial\varphi}\|t\|]\phi(\delta[M\|t\|^2 + M_{\partial\varphi}])dt.$$

The second integral is finite when δ is small enough (but according to condition 3, we can take δ as small as needed), and is clearly bounded in n, since Σ_n^{*-1} converges a.s. towards $I(\theta_0)$ (condition 1). Hence we have:

$$\int_{\Theta'} \exp\{-\frac{n}{2}(\theta - \widehat{\theta}_n)\Sigma_n^{*-1}(\theta - \widehat{\theta}_n)'\}R_n(\theta)d\theta \leq O(n^{-(K+1)/2}).$$

We can show in the same way that $\int_{\Theta'} \exp\{...\} R_n(\theta) d\theta \geq O(n^{-(K+1)/2})$, starting from $r_n(\theta) \geq -M$ instead of $r_n(\theta) \leq M$. Finally, we get:

$$\int_{\Theta'} \exp\{...\} R_n(\theta) d\theta = O(n^{-(K+1)/2})$$

■

We now apply this result to terms of the form $E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 \varphi(\theta)]$.

Lemma 3 *For $\varphi \in \mathcal{C}^3(\Theta \to \Re)$ such that $\int \varphi(\theta)\pi_n(\theta)d\theta$ and $\int \varphi(\theta)\frac{\pi_{n+p}(\theta)^2}{\pi_n(\theta)}d\theta$ exist,*

$$E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 \varphi(\theta)] = \left(\frac{1+r}{\sqrt{1+2r}}\right)^K \varphi(\theta_0) + O(\frac{1}{n}),$$

as $n \to +\infty$, $\frac{p}{n} \to r > 0$.

Proof.
We have

$$E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 \varphi(\theta)] = \int_\Theta \pi(\theta)\frac{L_{n+p}(\theta)^2}{L_n(\theta)}\varphi(\theta)d\theta \frac{\int_\Theta \pi(\theta)L_n(\theta)d\theta}{\left(\int_\Theta \pi(\theta)L_{n+p}(\theta)d\theta\right)^2}.$$

Assume first $\varphi(\theta_0) \neq 0$. According to lemma 2:

$$\int_\Theta \pi(\theta)L_n(\theta)d\theta = (2\pi)^{K/2}\sqrt{\det(\Sigma_n)}n^{-K/2}\pi(\widehat{\theta}_n)L_n(\widehat{\theta}_n)[1+O(n^{-1})],$$

$$\int_\Theta \pi(\theta)L_{n+p}(\theta)d\theta = \\ (2\pi)^{K/2}\sqrt{\det(\Sigma_{n+p})}(n+p)^{-K/2}\pi(\widehat{\theta}_{n+p})L_{n+p}(\widehat{\theta}_{n+p})[1+O\left((n+p)^{-1}\right)].$$

The asymptotic expansion of the third integral can be performed through similar arguments that those presented in lemma 2. Provided $\varphi > 0$ in the neighborhood of θ_0, let $\widetilde{l}_n(\theta) = \log[\varphi(\theta)\pi(\theta)] + l_n(\theta)$, we can write

$$\int_\Theta \pi(\theta)\frac{L_{n+p}(\theta)^2}{L_n(\theta)}\varphi(\theta)d\theta = \int_\Theta \exp\{2\widetilde{l}_{n+p}(\theta) - \widetilde{l}_n(\theta)\}d\theta,$$

with

$$\begin{aligned}\exp\{2\widetilde{l}_{n+p}(\theta)-\widetilde{l}_n(\theta)\} &= \exp\{2\widetilde{l}_{n+p}(\widehat{\theta}_{n+p})-\widetilde{l}_n(\widehat{\theta}_n)\}\\ &\quad \exp\{-\frac{2(n+p)}{2}(\theta-\widehat{\theta}_{n+p})\Sigma^{*-1}_{n+p}(\theta-\widehat{\theta}_{n+p})'\\ &\qquad +\frac{n}{2}(\theta-\widehat{\theta}_n)\Sigma_n^{*-1}(\theta-\widehat{\theta}_n)'\}[1+R_{n,p}(\theta)],\end{aligned}$$

and

$$\begin{aligned}R_{n,p}(\theta) &= \exp\{2\frac{\partial\log\varphi}{\partial\theta}(\widehat{\theta}_{n+p})(\theta-\widehat{\theta}_{n+p})'+2\left\|\theta-\widehat{\theta}_{n+p}\right\|^3(n+p)\,r_{n+p}(\theta)\\ &\qquad -\frac{\partial\log\varphi}{\partial\theta}(\widehat{\theta}_n)(\theta-\widehat{\theta}_n)'-\left\|\theta-\widehat{\theta}_n\right\|^3 nr_n(\theta)\ \}-1,\end{aligned}$$

where r_n is a bounded function. We can check that

$$\begin{aligned}&\int_\Theta \exp\{-\frac{2(n+p)}{2}(\theta-\widehat{\theta}_{n+p})\Sigma^{*-1}_{n+p}(\theta-\widehat{\theta}_{n+p})'+\frac{n}{2}(\theta-\widehat{\theta}_n)\Sigma_n^{*-1}(\theta-\widehat{\theta}_n)'\}R_{n,p}(\theta)d\theta\\ &\qquad = O\left((n+p)^{-(K+1)/2}\right)\end{aligned}$$

by the same kind of variable change $t=\sqrt{n+p}(\theta-\widehat{\theta}_{n+p})$ used in the previous lemma. Since $\left\|\theta-\widehat{\theta}_n\right\|\leq\left\|\theta-\widehat{\theta}_{n+p}\right\|+\left\|\widehat{\theta}_{n+p}-\widehat{\theta}_n\right\|$ where the second term is clearly bounded in n, we can again exhibit a bound of the kind:

$$\int_{\Re^K}\exp\{-\frac{1}{2}\left\|t\right\|^2\}[\exp\{(n+p)^{-1/2}[\alpha\left\|t\right\|^3+\beta\left\|t\right\|+\gamma\ ]\}-1]dt.$$

So we get:

$$\begin{aligned}&\int_\Theta \pi(\theta)\frac{L_{n+p}(\theta)^2}{L_n(\theta)}\varphi(\theta)d\theta\\ &= \frac{\left[\pi(\widehat{\theta}_{n+p})\varphi(\widehat{\theta}_{n+p})L_{n+p}(\widehat{\theta}_{n+p})\right]^2}{\left[\pi(\widehat{\theta}_n)\varphi(\widehat{\theta}_n)L_n(\widehat{\theta}_n)\right]}\frac{(2\pi)^{K/2}}{\sqrt{\det[\frac{2(n+p)}{\Sigma_{n+p}}-\frac{n}{\Sigma_n}]}}\ [1+O(\frac{1}{n})]\\ &= \frac{\left[\pi(\widehat{\theta}_{n+p})\varphi(\widehat{\theta}_{n+p})L_{n+p}(\widehat{\theta}_{n+p})\right]^2}{\left[\pi(\widehat{\theta}_n)\varphi(\widehat{\theta}_n)L_n(\widehat{\theta}_n)\right]}\frac{(2\pi)^{K/2}\,n^{-K/2}}{\sqrt{\det[\frac{2(1+p/n)}{\Sigma_{n+p}}-\frac{1}{\Sigma_n}]}}[1+O(\frac{1}{n})],\end{aligned}$$

and combining the three integrals:

$$\begin{aligned}
& E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 \varphi(\theta)] \\
= \; & \frac{\varphi(\widehat{\theta}_{n+p})^2}{\varphi(\widehat{\theta}_n)} \left(\frac{n+p}{n}\right)^K \frac{\sqrt{\det(\Sigma_n)}}{\det(\Sigma_{n+p})\sqrt{\det[\frac{2(1+p/n)}{\Sigma_{n+p}} - \frac{1}{\Sigma_n}]}}[1+O(\frac{1}{n})] \\
= \; & \left(\frac{1+r}{\sqrt{1+2r}}\right)^K \varphi(\theta_0) + O(\frac{1}{n}),
\end{aligned}$$

when $n \to +\infty$, $\frac{p}{n} \to r$, since $\widehat{\theta}_{n+p}$, $\widehat{\theta}_n$ converges towards θ_0 and Σ_{n+p}, Σ_n towards $I(\theta_0)$.

In case $\varphi(\theta_0) = 0$, we can write

$$E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 \varphi(\theta)] = E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 (\varphi(\theta)+A)] - AE_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2],$$

with $A > 0$, to get the same result. ■

We now deduce from the later lemma the proof of theorem 1.

Theorem 1 *Under some regularity conditions listed at the beginning of the appendix (conditions 1 to 4), and for any $h \in \mathcal{C}^3(\Theta \to \mathbb{R}^L)$, $L \in \mathbb{N}$, such that the integrals $\int h(\theta)\pi_n(\theta)d\theta$, $\int h(\theta)h(\theta)'\pi_n(\theta)d\theta$, and $\int h(\theta)h(\theta)'\pi_{n+p}(\theta)^2/\pi_n(\theta)d\theta$ exist for all $n, p \in \mathbb{N}$, we have*

$$\tau_{\pi_{n+p}/\pi_n}(h) = O(I_L) \qquad \text{as } n \to \infty, \frac{p}{n} \to r > 0.$$

Proof. It is well known that $\mathrm{V}_{\pi_{n+p}}\{h(\theta)\} \sim I(\theta_0)^{-1}/(n+p)$, and since $\tau_{\pi_{n+p}/\pi_n}(h) = V_{\pi_{n+p}/\pi_n}(h)\left[\mathrm{V}_{\pi_{n+p}}\{h(\theta)\}\right]^{-1}$, we show now that $V_{\pi_{n+p}/\pi_n}(h) = O(I_L/(n+p))$.

Let $\mu_n(h) \triangleq E_n[h(\theta)]$. Assume first $h(\Theta) \subset R$, then:

$$\begin{aligned}
V_{\pi_{n+p}/\pi_n}(h) &= E_{\pi_n}[\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}(h(\theta)-\mu_{n+p}(h))]^2 \\
&= E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 h(\theta)^2] + \mu_{n+p}(h)^2 E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2] \\
&\quad -2\mu_{n+p}(h)E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 h(\theta)]
\end{aligned}$$

In case $h = (h^1, h^2, ...)$ is multidimensional, we notice that:

$$\begin{aligned} &\mathrm{Cov}_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)(h^1(\theta) - \mu^1_{n+p}(h)); \left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)(h^2(\theta) - \mu^2_{n+p}(h))] = \\ &\quad E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 h^1(\theta)h^2(\theta)] + \mu^1_{n+p}(h)\mu^2_{n+p}(h)E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2] \\ &\quad - \mu^1_{n+p}(h)E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 h^2(\theta)] - \mu^2_{n+p}(h)E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 h^1(\theta)] \end{aligned}$$

We see that the result can always be obtained through a smart combination of terms $E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 \varphi(\theta)]$, whose asymptotic expansion was given by lemma 3.

In particular, if $h(\Theta) \subset R$, we have:

$$\begin{aligned} V_{\pi_{n+p}/\pi_n}(h) &= E_{\pi_n}[\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}(h(\theta) - \mu_{n+p}(h))]^2 \\ &= E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 h(\theta)^2] + \mu_{n+p}(h)^2 E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2] \\ &\quad -2\mu_{n+p}(h)E_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)^2 h(\theta)] \\ &= \left(\frac{1+r}{\sqrt{1+2r}}\right)^K [h(\theta_0)^2 + \mu_{n+p}(h)^2 - 2\mu_{n+p}(h)h(\theta_0)] + O(n^{-1}) \\ &= \left(\frac{1+r}{\sqrt{1+2r}}\right)^K [h(\theta_0) - \mu_{n+p}(h)]^2 \ + O(n^{-1}). \end{aligned}$$

Since $\mu_{n+p}(h) = h(\theta_0)[1 + O((n+p)^{-1})]$, we finally get

$$V_{\pi_{n+p}/\pi_n}(h) = O(n^{-1}),$$

and this result is easily generalized to the multidimensional case, by computing the cross terms

$$\mathrm{Cov}_{\pi_n}[\left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)(h^1(\theta) - \mu^1_{n+p}(h)); \left(\frac{\pi_{n+p}(\theta)}{\pi_n(\theta)}\right)(h^2(\theta) - \mu^2_{n+p}(h))].$$

■

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, 88(422):669–679.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proc-Radar, Sonar Navigation*, 146(1):2–7.

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

Crisan, D. and Doucet, A. (2000). Convergence of sequential monte carlo methods. *Technical Report Cambridge University, CUED/F-INFENG/TR381.*

Doucet, A., de Freias, J., and Gordon, N. (2001). *(Editors) Sequential Monte Carlo Methods in Practice.* Springer-Verlag: New York, to appear January 2001.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57:1317–1340.

Gilks, W. and Berzuini, C. (1999). Following a moving target - Monte Carlo inference for dynamic Bayesian models. *J. Roy. Stat. Soc. B (to appear).*

Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-gaussian Bayesian state estimation. *J. Amer. Statist. Assoc.*, 140(2):107–113.

Izenman, A. and Sommer, C. (1988). Philatelic mixtures and multimodal densities. *J. Amer. Statist. Assoc.*, (83):941–953.

Johnson, R. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Stat.*, 41(3):851–864.

Kolassa, J. (1999). Convergence and accuracy of Gibbs sampling for conditional distributions in generalized linear models. *Ann. Stat.*, 27(1):129–142.

Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, 93:1032–1044.

Madras, N. and Piccioni, M. (1999). Importance sampling for families of distributions. *Ann. Applied Prob.*, 9:1202–1225.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models.* Chapman and Hall.

Robert, C. and Mengersen, K. (1999). Reparameterisation issues in mixture modelling and their bearing on MCMC algorithms. *Comp. Stat. Data Anal.*, 29:325–343.

Schervish, M. (1995). *Theory of Statistics.* Springer-Verlag.

Tierney, L., Kass, R., and Kadane, J. (1989). Fully exponential Laplace approximmations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.*, 84(407).