Useful (?) remark for HMMs and state space models

Most standard texts on Hidden Markov Models (eg. Rabiner's 1989 tutorial, McDonald & Zucchini's 1997 monograph) ignore a remarkable observation about HMMS:

- The intermediate quantity of the EM (Expectation Maximization) algorithm
- The gradient of the log-likelihood

and more generally any function that can be written as

$$\gamma_t = \sum_{s=1}^{t} \mathsf{E}\left(m_s(X_s) | \, Y_{1:t}\right) + \sum_{s=2}^{t} \mathsf{E}\left(r_s(X_{s-1}, X_s) | \, Y_{1:t}\right)$$

can be computed recursively in t: 1994 book by Elliot, Aggoun & Moore (for experts only) Zeitouni & Dembo (1989) and several references in the control literature (keyword: "exact filter")...

⇢ "Forward-Backward" smoothing is not the only solution

What's the trick?

Consider the example of $\gamma_t = \sum_{s=1}^{t} \mathsf{E}\left(m_s(X_s)|\, Y_{1:t}\right)$ and define

$$\Gamma_t(j) = \sum_{s=1}^{t}\sum_{l=1}^{N} m_s(l)\, \mathsf{P}(X_s = l, X_t = j|Y_{1:t}) \qquad \text{so that } \gamma_t = \sum_{j=1}^{N} \Gamma_t(j)$$

$$\text{Notations: } \begin{cases} X_{t+1}|X_t = x_t \sim k(x_t, \cdot) \\ Y_t|X_t = x_t \sim q(x_t, \cdot) \\ X_t \in \{1, \dots N\} \end{cases}$$

Then (homework...),

$$\Gamma_{t+1}(j) = \frac{\left(\sum_{i=1}^{N} \Gamma_t(i) k(i,j)\right) q(j, Y_{t+1})}{\sum_{l=1}^{N} q(l, Y_{t+1})\ \underbrace{\mathsf{P}(X_{t+1} = l|Y_{1:t})}_{\text{standard predictor}}} + m_{t+1}(j)\, \underbrace{\mathsf{P}(X_{t+1} = j|Y_{1:t+1})}_{\text{standard filter}}$$

Comments

A similar relation holds for the general state space case as well as for continuous-time models (with explicit formulas in the Gaussian linear case).

Warning: Computing Γ_T is $O(N^2 \times T)$ but there are many such statistics of interest: $m_s(x_s) = \mathbb{I}_{\{i\}}(x_s)$ ($N-1$ of them), $r_s(x_{s-1}, x_s) = \mathbb{I}_{\{i\}}(x_{s-1})\mathbb{I}_{\{j\}}(x_s)$ ($N \times (N-1)$ of these)...

This idea can be used for approximating quantities of interest with particle filters, cf. (Cappé, 2001).

Continuous-time jump MCMC and model selection for HMMs

Olivier Cappé
CNRS / ENST, Paris

Joint work with **Christian Robert** (CREST, Paris) and **Tobias Rydén** (Lund University)

Abstract: Discusses a method proposed (again) by Stephens (2000) for model selection based on continuous-time jump chain simulation

Overview

1. Continuous-time jump simulation as an alternative (?) to conventional MCMC

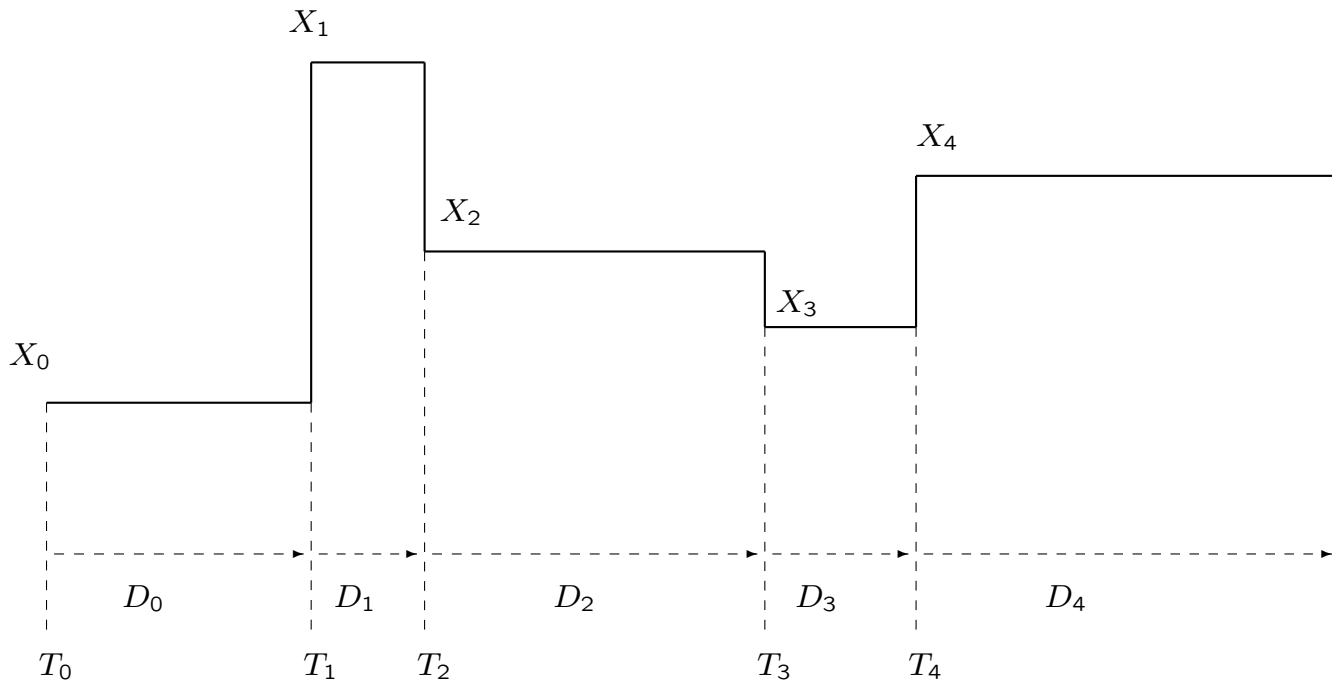
2. Continuous-time jump MCMC for model selection

3. Reversible Jump MCMC samplers converging to continuous time-jump sampler

4. Application to HMMs

5. Continuous-time jump simulation and importance sampling

An alternative to conventional MCMC



where $(X_k)_{k\geq 0}$ is a (discrete-time) Markov chain with kernel Q and $D_k|X_{1:k} \sim$ Exponential $\lambda(X_k)$

The continuous-time process is

$$X(t) = \sum_{k=0}^{+\infty} X_k \mathbb{I}_{[T_k, T_{k+1})} \quad \text{with } T_k = \sum_{j=0}^{k-1} D_j \quad (\text{and } T_0 = 0)$$

A simple example

Detailed balance condition

$$\pi(x)\lambda(x)Q(x,y) = \pi(y)\lambda(y)Q(y,x) \quad \text{(assuming } \pi \text{ and } Q(u,\cdot) \ll \mu)$$

The independent jump sampler

For $Q(x,y) = q(y)$

$$\dashrightarrow \quad \lambda(x) = q(x)/\pi(x)$$

Note: The chain is always *non-explosive* since

$$\mathsf{E}_q[\frac{1}{\lambda(X)}] = \pi(\text{support}(q))$$

but *geometric ergodicity* indeed requires that

$$q(x)/\pi(x) \geq \delta > 0$$

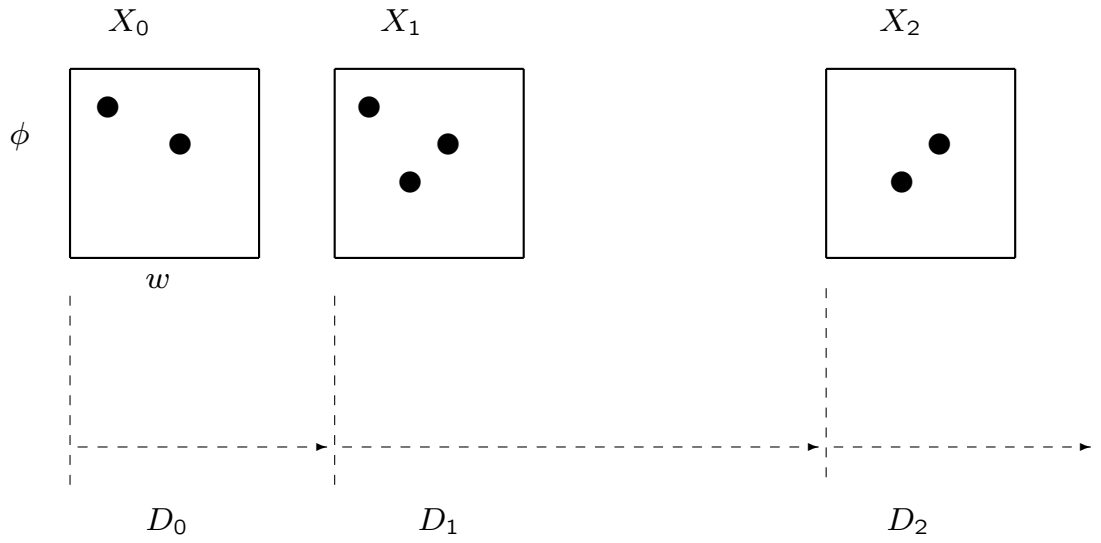
Continuous-time MCMC for model selection

Stephens (*Ann. Statist.*, 2000) – and others before – have proposed using continuous-time simulation for variable dimension MCMC as an alternative to Green's (1995) *Reversible Jump*

The case of mixture (Stephens, 2000) We want to estimate all parameters, including dimension k, of the mixture density

$$\sum_{i=1}^{k} \rho_i f(\cdot; \phi_i) \qquad \underline{\text{Note:}}\ \textit{reparameterization}\ \rho_i := w_i / \sum_{j=1}^{k} w_i$$

⇢ View $(w_i, \phi_i)_{1 \le i \le k}$ as a sample from a spatial point process



Basic move: Propose new component $\sim b(w, \phi)$

RJMCMC (using Birth and Death moves only):

- Births and deaths are proposed with probabilities β_k and δ_k, respectively, when having k components
- Acceptance probability for birth move is $\min(A, 1)$, with

$$A = \frac{\pi(k+1, (\underline{w}_k, \underline{\phi}_k) \cup (w, \phi))}{\pi(k, (\underline{w}_k, \underline{\phi}_k))} \times \frac{\delta_{k+1}}{\beta_k b(w, \phi)}$$

- Acceptance probability for death move is $\min(A^{-1}, 1)$

BDMCMC:

- New components are born according to a Poisson process with rate λ_k when having k components
- Each component (w, ϕ) dies with rate

$$d(w, \phi) = \underbrace{\frac{\pi(k, (\underline{w}_k, \underline{\phi}_k))}{\pi(k+1, (\underline{w}_k, \underline{\phi}_k) \cup (w, \phi))}}_{\text{posterior ratio}} \times \frac{\lambda_k b(w, \phi)}{k+1}$$

Both approaches can (only) be connected by a **time scaling** construction:

- In discrete-time RJMCMC, let the time unit be $1/N$, put $\beta_k = \lambda_k/N$ and $\delta_k = 1 - \lambda_k/N$. Finally consider $X(t) := X^{(N)}_{\lfloor \frac{t}{N} \rfloor}$

- As $N \to \infty$ all birth proposals are accepted, and births occur according to a Poisson process with rate λ_k (when having k components)

- As $N \to \infty$, a component (w, ϕ) of the $k+1$ components configuration dies with rate

$$\begin{aligned} \lim_{N\to\infty} N\delta_{k+1} \times \frac{1}{k+1} \times \min(A^{-1}, 1) & \\ = \lim_{N\to\infty} N\frac{1}{k+1} \times \text{posterior ratio} \times \frac{\beta_k b(w,\phi)}{\delta_{k+1}} & \\ = \text{posterior ratio} \times \frac{\lambda_k b(w,\phi)}{k+1} & \end{aligned}$$

  Hence "RJMCMC→BDMCMC" (This can be shown more formally and holds for general type of moves).

Other comparisons between the two approaches

Stephens (2000) argues that the continuous-time alternative is simpler to implement than reversible jump – But this is only a consequence of the simplicity of the birth-or-death move:

The Jacobian term also appears when using more complex moves. In a split-or-merge implementation where one proposes $(w_j', \phi_j', w_j'', \phi_j'') = T(w_j, \phi_j, \epsilon_w, \epsilon_\phi)$ with $(\epsilon_w, \epsilon_\phi) \sim b$, the death rate becomes

$$\text{posterior ratio} \times \frac{\eta_k}{k(k+1)} \times 2b(\epsilon_w, \epsilon_\phi) \times \left| \frac{\partial T}{\partial(w_j, \phi_j, \epsilon_w, \epsilon_\phi)} \right|^{-1}$$

where η_k is the split rate for a k components configuration.

The continuous-time algorithm is costly to implement for split-or-merge moves since computing the $k(k+1)/2$ merge rates is necessary for simulating the lifetime in a given $k+1$ components configuration.

Application to scalar Gaussian HMM

Parameters

k	number of components
$w_1, \ldots, w_k$	weights
$\mu_1, \ldots, \mu_k$	means
$\sigma_1, \ldots, \sigma_k$	variances

Moves

1. Birth/Death move (rate λ_k), where b is the prior

2. Split/Merge move (rate η_k) with T given by

$$(\mu'_j, \mu''_j) = (\mu_j + \epsilon_\mu, \mu_j + \epsilon_\mu)$$
$$(\sigma'_j, \sigma''_j) = (\sigma_j \epsilon_\sigma, \sigma_j / \epsilon_\sigma)$$
$$(w'_j, w''_j) = (w_j \epsilon_w, w_j / \epsilon_w)$$

   where $\epsilon_\mu \sim \mathcal{N}$, $\epsilon_\sigma, \epsilon_w \sim \log -\mathcal{N}$

3. Conventional fixed k moves (rate ξ_k)

Application to HMM (cont.)

The HMM likelihood is computed exactly (no data augmentation) using forward filtering

In some cases, seems to achieve better mixing than (Robert, Rydén & Titterington, 2000)

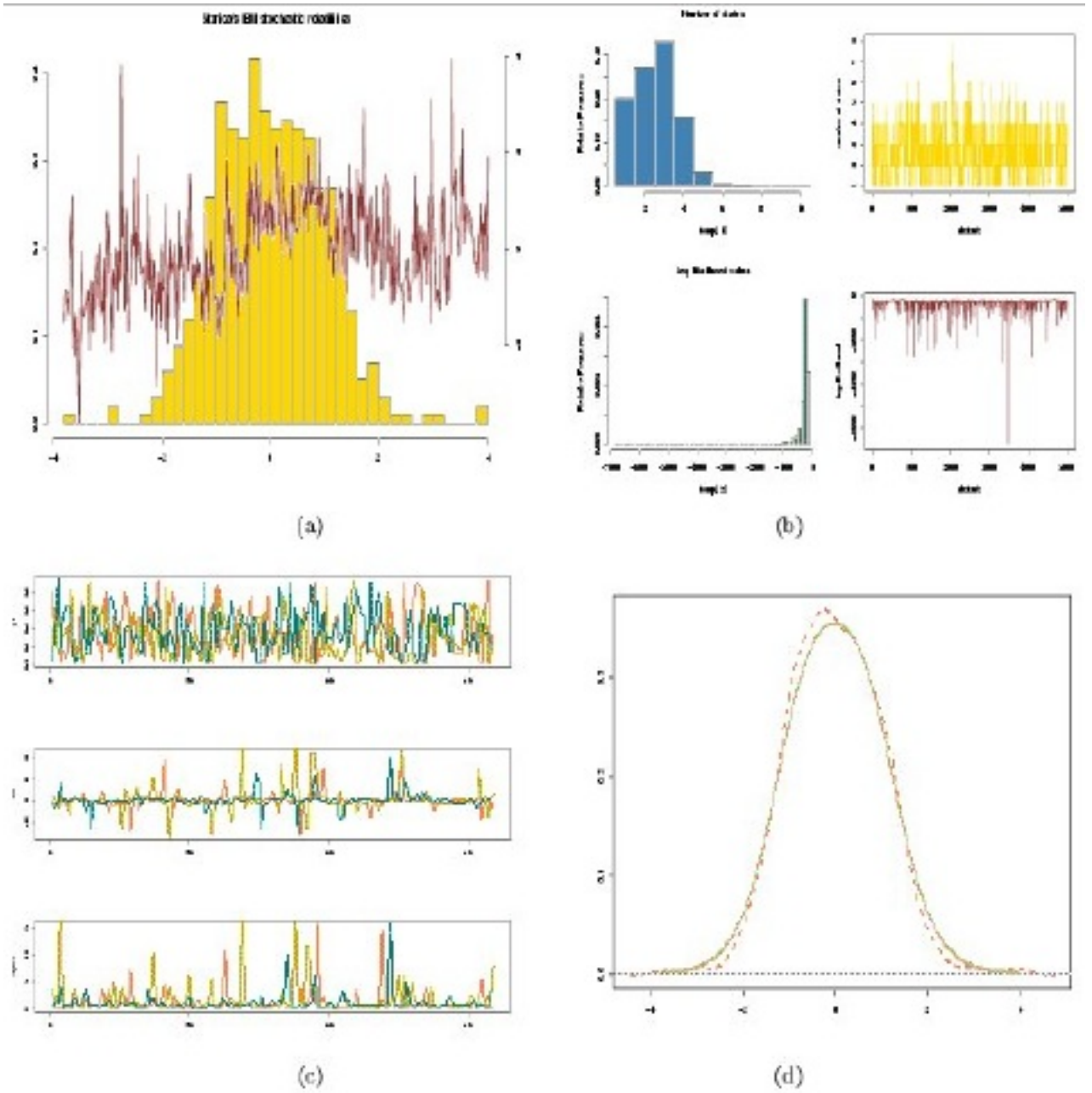


Fig. 2. Continuous time MCMC algorithm output for a transform of 507 IBM stockprices: (a) histogram and rawplot of the dataset; (b) MCMC output on k (histogram and rawplot), number of states, and corresponding likelihood values; (c) MCMC sequence of the parameters of the three components when conditioning on $k = 3$; (d) MCMC evaluation of the marginal density compared with R nonparametric density estimate.

Variance reduction for CT simulation

We typically want to estimate $\mathsf{E}_\pi f(X)$ by $t^{-1}\int_0^t f(X(t))dt$ or $T_k^{-1}\int_0^{T_k} f(X(t))dt$, but

$$\mathsf{E}(\int_0^{T_k} f(X(t))dt|X_{0:k-1}) = \sum_{j=0}^{k-1} f(X_j)\underbrace{\mathsf{E}(D_j|X_j)}_{\lambda^{-1}(X_j)}$$

and computing $\lambda(X_j)$ is required by the method (©Gareth Roberts, 2001).

--→ The "smart" estimate is

$$\frac{\sum_{j=0}^{k-1}\lambda^{-1}(X_j)f(X_j)}{\sum_{j=0}^{k-1}\lambda^{-1}(X_j)}$$

which looks very much like Bayesian importance sampling ($w = \lambda^{-1}$).

For the simple independent CT jump sampler this is exactly B-IS and the gain in asymptotic variance is a factor 2.

Conclusions

The situation is more contrasted than suggested by Stephens (2000)

Some interesting question remains – in particular, the way one actually simulates (approximatively) a random variable $\sim \pi$ with CT simulation is very different from Importance Sampling and Resampling.

See full length version of the paper for details