Dynamic Models with Latent Variables

Jean-Michel Zakoian

CREST-ENSAE

Bayesian and simulated methods

Simulation-based method

- In recent years, inference methods based on simulations have developed tremendously.
- Such methods are particularly appropriate when the likelihood cannot be used directly.
- One popular method is the **Monte Carlo Markov Chain (MCMC)** technique.

Bayesian inference and MCMC

The *Bayesian* approachs combines information brought by the data with *a priori* ideas on the parameters.

Such ideas are represented by probability distributions, called *prior* distributions, where the parameters have the status of random variables.

The goal of the inference is to obtain *posterior distributions* of the parameters, which are deduced, thanks to the Bayes formula, from the *a priori* distribution and the law of the observations conditional on the parameters.

Notations

Parameter : $\theta \in \mathbb{R}^d$.

Observations : $\mathbf{X}$.

A priori density of the parameter : π.

A posteriori density of the parameter : $\pi(\cdot \mid \mathbf{X})$

Likelihood (the density of the observations conditional to θ) : $f(\cdot \mid \theta)$.

By the Bayes formula we have :

$$\pi(\theta \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \theta)\pi(\theta)}{\int f(\mathbf{X} \mid \vartheta)\pi(\vartheta)d\vartheta} \propto f(\mathbf{X} \mid \theta)\pi(\theta).$$

Bayesian estimator of $g(\theta)$

Obtained by minimizing the *a posteriori cost*

$$\int L(\delta, \vartheta)\pi(\vartheta \mid \mathbf{X})d\vartheta,$$

where L is a positive cost function.
For the quadratic cost, $L(\delta,\theta) = \|g(\theta) - \delta\|^2$, the Bayesian estimator is (if the integral exists),

$$\hat{\delta} = E\{g(\theta) \mid \mathbf{X}\} = \int g(\vartheta)\pi(\vartheta \mid \mathbf{X})d\vartheta.$$

This expression has little practical interest because in general,
(i) the *posterior* distribution is not known explicitly,
(ii) the integral is difficult to compute. This problem can be handled by *numerical methods*, or by **simulations based methods**.

Simulations based methods

Suppose that a sequence $\{\theta^{(i)}(\mathbf{X})\}_{i\geq 1}$ of variables with density $\pi(\cdot \mid \mathbf{X})$ is availbale.

Assuming that the ergodic theorem can be applied, conditionally to $\mathbf{X}$,

$$\hat{\delta}_N = \frac{1}{N}\sum_{i=1}^{N} g\{\theta^{(i)}(\mathbf{X})\} \to E\{g(\theta) \mid \mathbf{X}\}, \quad p.s. \text{ as } N \to \infty.$$

MCMC

The MCMC approach is a set of techniques allowing to generate such sequences $\{\theta^{(i)}(\mathbf{X})\}_{i\geq 1}$.

The basic idea of the MCMC method is to approximate a probability law $\mathbb{P}$ using a Markov chain $(\theta^{(i)})$, which is irreducible, aperiodic, and admits $\mathbb{P}$ as invariant law.

The target law is simulated using an initial value $\theta^{(0)}$, and the transition probabilities of the chain to generate $\theta^{(1)},\ldots,\theta^{(N)}$.

For N sufficiently large, $\theta^{(N)}$ can be considered as a realization drawn from the law $\mathbb{P}$.

Two methods for constructing such chains are *the Metropolis-Hastings algorithm* and the *Gibbs sampling.*

Principle

We wish to simulate the density $\pi : \theta \to \pi(\theta)$ of some rv $\boldsymbol{\theta}$.

Finding an explicit formula for $F^{-1}(U)$, where F is the cdf (cumulative distribution function) of $\boldsymbol{\theta}$ and $U \sim \mathscr{U}[0,1]$ is not always possible.

Moreover, even if it is, there may be alternative methods for generating a rv distributed as F that are more efficient than the inverse transform method.

Basic idea : find an alternative probability distribution G, with density function $g(\theta)$, from which we already have an efficient simulation algorithm, but also such that the function $g(\theta)$ is "close" to $f(\theta)$.

Accept-reject algorithm

Suppose that for all θ

$$\pi(\theta) \leq cg(\theta),$$

where c is a constant $(c \geq 1)$ and g is a density which can be easily simulated.

Algorithm :

1. Generate $Z \sim g$ and generate $U \sim \mathcal{U}_{[0,cg(Z)]}$.
2. Accept Z if $U < \pi(Z)$, otherwise reject and go back to 1.

Equivalently :

1. Generate $Z \sim g$ and $U \sim \mathcal{U}_{[0,1]}$.
2. Accept Z if $U < \frac{\pi(Z)}{cg(Z)}$, otherwise reject and go back to 1.

Proof that the algorithm works

Z^* : the variable generated by the algorithm.

For all z^*,

$$P[Z^* < z^*] \quad = \quad P[Z < z^* \mid U < \pi(Z)] = \frac{P[Z < z^*, U < \pi(Z)]}{P[U < \pi(Z)]}.$$

Moreover,

$$P[Z < z^*, U < \pi(Z)] = \int_{z<z^*} g(z) \frac{1}{cg(z)} \int_{u<\pi(z)} du dz = \frac{1}{c} \int_{z<z^*} \pi(z) dz.$$

and, by taking $z^* = \infty$,

$$P[U < \pi(Z)] = \frac{1}{c} \int \pi(z) dz = \frac{1}{c}.$$

Finally,

$$P[Z^* < z^*] \quad = \quad \int_{z<z^*} \pi(z) dz.$$

Remarks

- By iterating the procedure we get an iid sample distributed as π.
- The probability of accepting a simulation of Z is $1/c$: the constant c has to be chosen as small as possible to minimize the computation time.
- c depends on the tails of the laws π and g : in particular, the ratio π/g has to be bounded (it is not possible to simulate a Cauchy from a Gaussian, but the converse is possible).
- If c is chosen too small so that the condition $\pi(\theta) \leq cg(\theta)$ is not satisfied for some values of θ, the simulated density is not π but rather

$$\pi^*(\theta) \propto \min\{\pi(\theta), cg(\theta)\}.$$

Simulation of an a posteriori distribution

We wish to simulate the *a posterior* distribution

$$\pi(\theta \mid \mathbf{X}) = h(\mathbf{X})\pi^*(\theta, \mathbf{X})$$

where π^* is a known function but h has a complicated non explicit form.
If there exits a density $g^*(\cdot \mid \mathbf{X})$ such that

$$\pi^*(\theta, \mathbf{X}) \leq c g^*(\theta \mid \mathbf{X}),$$

the method can be adapted as follows :

1. Generate θ distributed as $g^*(\cdot \mid \mathbf{X})$ and $U \sim \mathcal{U}_{[0,1]}$.
2. Accept θ if $U < \frac{\pi^*(\theta, \mathbf{X})}{c g^*(\theta \mid \mathbf{X})}$, otherwise reject and go back to 1.

Metropolis-Hastings (MH) algorithm

Metropolis et al. (1953) [discrete state space], Hastings (1970) [statistical setting].

The MH algorithm allows to generate a density on which we have little information.

To simulate the *posterior* density $\pi(\cdot \mid \mathbf{X})$, the MH algorithm allows to generate a sequence of variables $\{\theta^{(i)}\}_{i \geq 1} = \{\theta^{(i)}(\mathbf{X})\}_{i \geq 1}$ which, conditionally to $\mathbf{X}$, is a **Markov chain** with stationary distribution $\pi(\cdot \mid \mathbf{X})$.

Markov chains

(X_t) is an homogenous Markov chain on $(E, \mathscr{E})$ if :
$\forall x \in E, \forall B \in \mathscr{E}, \forall s, t \in \mathbb{N}$,

$$P(X_{s+t} \in B \mid X_r, r < s; X_s = x) = P(X_{s+t} \in B \mid X_s = x) := P^t(x, B).$$

The mapping $P : E \times \mathscr{E} \to [0,1]$ is called *transition kernel* and it satisfies :
(i) $\forall B \in \mathscr{E}$, the function $P(\cdot, B)$ is measurable ;
(ii) $\forall x \in E$, the function $P(x, \cdot)$ is a probability over $(E, \mathscr{E})$.

The law of (X_t) is characterized by an initial probability μ and a transition kernel P.

Markov chains

Under certain conditions, there exists a probability π such that $\forall x \in E, \forall B \in \mathscr{E}$,

$$P^t(x, B) \to \pi(B), \quad \text{as } t \to \infty.$$

π is called **invariant probability** and it satisfies :

$$\forall B \in \mathscr{E}, \qquad \pi(B) = \int P(x, B)\pi(dx).$$

Under additional conditions, the chain is **ergodic** and we have, a.s.

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t) \to \int g d\pi, \quad \text{as } n \to \infty$$

for any function g which is with respect to π.

MH algorithm

The algorithm is based on a transition kernel of the form

$$Q(\theta, d\theta') = q(\theta, \theta')\lambda(d\theta'), \quad \theta \in \mathbb{R}^d.$$

The transition densities, $q(\theta, \cdot)$, also called *jump densities*, have to be chosen in order to explore the parameter space. They may depend on $\mathbf{X}$ and are used at each step of the algorithm to simulate a new possible value and to decide whether this value will be kept or rejected.

MH algorithm

1. Choose an initial value $\theta^{(0)}$ such that $\pi(\theta^{(0)} \mid \mathbf{X}) > 0$.
2. For $i = 1, 2, \ldots$
   1. Generate θ^* distributed as $Q(\theta^{(i-1)}, \cdot)$.
   2. Compute the ratio

$$r = \frac{\pi(\theta^* \mid \mathbf{X})}{\pi(\theta^{(i-1)} \mid \mathbf{X})} \frac{q(\theta^*, \theta^{(i-1)})}{q(\theta^{(i-1)}, \theta^*)}.$$

   3. Take

$$\theta^{(i)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{(i-1)} & \text{with probability } 1 - \min(r, 1). \end{cases}$$

Remark : the condition $\pi(\theta^{(0)} \mid \mathbf{X}) > 0$ ensures that the ratio is well defined for all i.

Remarks

- If $q(x,y) = q(y,x)$ for all (x,y), the acceptance rule does not depend on q. The probability of acceptance of θ^* reduces to

$$\min\left(\frac{\pi(\theta^* \mid \mathbf{X})}{\pi(\theta^{(i-1)} \mid \mathbf{X})}, 1\right)$$

  and θ^* will be always accepted if jumping from $\theta^{(i-1)}$ to θ^* increases the *a posteriori* density ; otherwise, θ^* is accepted with a probability equal to the ratio r of the *posterior* densities.
- In the general case, θ^* is accepted with probability 1 only if the ratio of *posterior* densities is greater than the ration of densities of jump, from $\theta^{(i-1)}$ to θ^* and from θ^* to $\theta^{(i-1)}$.
- r **does not depend on normalizing constants** involved in the conditional densities. In particular, it is sufficient to know $\pi(\theta^* \mid \mathbf{X})$ up to a multiplicative constant to use this algorithm. Thus, il suffices to evaluate $f(\mathbf{X} \mid \cdot)\pi(\cdot)$ in θ^* and $\theta^{(i-1)}$.

Examples of transition kernels

(i) *Kernel associated with the random walk :* at step i, the value θ^* is generated using the model

$$\theta^* = \theta^{(i-1)} + \epsilon$$

where ϵ is a centered variable with density f, independent of $\theta^{(i-1)}$. Thus

$$q(\theta, \theta') = f(\theta' - \theta).$$

If f is symmetric around 0, the kernel is symmetric.

Note that the sequence $(\theta^{(i)})$ is not a random walk, due to repetitions $(\theta^{(i)} = \theta^{(i-1)})$ which are non independent from $\theta^{(i-1)}$.

Examples of transition kernels

(ii) *Independent Kernel :* the values θ^* are drawn, independently of $\theta^{(i-1)}$, from a density f :

$$q(\theta, \theta') = f(\theta')$$

and the probability of acceptance of the value θ^* is

$$\min\left(\frac{\omega(\theta^*)}{\omega(\theta^{(i-1)})}, 1\right)$$

where $\omega(\theta) = \pi(\theta \mid \mathbf{X})/f(\theta)$.

The function ω can be interpreted as an importance function used to simulate π from simulations of the law f.
Low-weight candidates will be seldom accepted ; conversely, large-weight candidates will generally be chosen repeatedly.

Examples of transition kernels

(iii) *Kernel associated with an accept-reject algorithm :* particular case of independent kernel where $f = \pi(\cdot \mid \mathbf{X})$ is simulated by the accept-reject method.

If the condition $\pi(\theta \mid \mathbf{X}) \leq cg(\theta)$ is in failure, the density generated for the θ^* will be

$$\pi^*(\theta^* \mid \mathbf{X}) \propto \min\{\pi(\theta^* \mid \mathbf{X}), cg(\theta^*)\}.$$

Letting $C = \{\theta \mid \pi(\theta \mid \mathbf{X}) \leq cg(\theta)\}$, the acceptance probability of the value θ^* is

$$\begin{cases} 1, & \text{if } \theta^{(i-1)} \in C \\ \dfrac{cg(\theta^{(i-1)})}{\pi(\theta^{(i-1)} \mid \mathbf{X})}, & \text{if } \theta^{(i-1)} \notin C, \theta^* \in C, \\ \min\left(\dfrac{\pi(\theta^* \mid \mathbf{X}) g(\theta^{(i-1)})}{\pi(\theta^{(i-1)} \mid \mathbf{X}) g(\theta^*)}, 1\right), & \text{if } \theta^{(i-1)} \notin C, \theta^* \notin C. \end{cases}$$

Examples of transition kernels

- Some proposals are rejected when $\theta^{(i-1)} \notin C$. This value is thus repeated, which compensates the deficiency, due to the failure of the bound cg, in this region.
  Dependance is thus introduced to alleviate this problem.
- If the deficiency never holds, the values accepted by the accept-reject algorithm are also accepted by the global algorithm and the resulting sample is iid.

Examples of transition kernels

(iv) *Kernel associated with an AR :*

$$\theta^* = a + b(\theta^{(i-1)} - a) + \epsilon$$

where ϵ is a variable with density f, independent from $\theta^{(i-1)}$. Thus

$$q(\theta, \theta') = f(\theta' - a - b(\theta - a)).$$

The choice of $b < 0$ allows to introduce negative correlations between the successive value of the algorithm and to explore faster the support of π.

Markov chain property

Conditionally to $\mathbf{X}$, the sequence $(\theta^{(i)}(\mathbf{X}))_{i\geq 0}$ is a Markov chain on $\mathbb{R}^d$.

It can be shown that, under certain conditions on the transition kernel, this MC admits $\pi(\cdot|\mathbf{X})$ as an invariant probability measure and is ergodic :

$$\hat{\delta}_N = \frac{1}{N}\sum_{i=1}^{N} g\{\theta^{(i)}(\mathbf{X})\} \to E_\pi\{g(\theta)\,|\,\mathbf{X}\}, \quad p.s. \text{ as } N \to \infty.$$

Example : STAR model

$$Y_t = a_0 + a_1 Y_{t-1} + \frac{(b_0 - a_0) + (b_1 - a_1) Y_{t-1}}{1 + \exp\{-\gamma(Y_{t-1} - c)\}} + \epsilon_t, \quad \epsilon_t \sim IID(0, \sigma^2), \quad \gamma > 0.$$

▸ Gibbs estimation

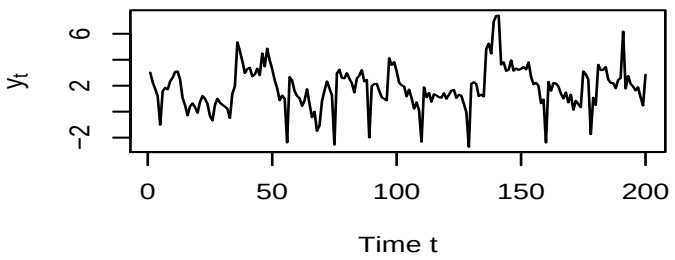
$E(Y_t | Y_{t-1} = y)$ varies continuously from $a_0 + a_1 y$ (when $y \to -\infty$) to $b_0 + b_1 y$ (when $y \to +\infty$).

Simulation of

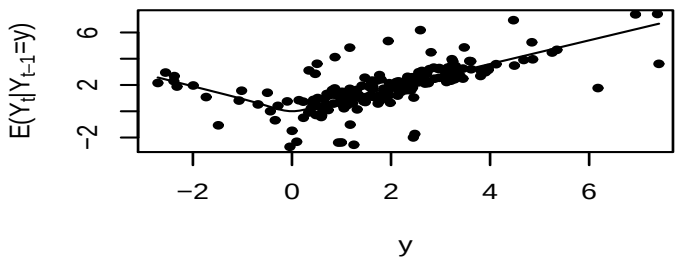
$$Y_t = -0.95 Y_{t-1} + \frac{1.85 Y_{t-1}}{1 + e^{-5 Y_{t-1}}} + \epsilon_t$$

where $\epsilon_t \sim 0.9\mathcal{N}(0, 0.5^2) + 0.05\mathcal{N}(3, 1) + 0.05\mathcal{N}(-3, 1)$.

(a) Simulation (Y_t) of a STAR



(b) Points (Y_{t-1},Y_t) and regression function



Example : STAR model

Independent prior laws : c fixed to 0, a_0, a_1, b_0, and b_1 follow a $\mathcal{N}(0,1)$, $\gamma \sim \mathcal{E}(1)$, and $\sigma^2 \sim IG(1,1)$ *.

We thus have, for $\theta = (a_0, a_1, b_0, b_1, \gamma, \sigma^2)$,

$$\begin{aligned}
\pi(\theta \mid \mathbf{X}) &\propto f(\mathbf{X} \mid \theta)\pi(\theta) \\
&= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(Y_t - m_{t-1})^2\right\} \\
&\quad \times \frac{1}{(2\pi)^2} e^{-(a_0^2+a_1^2+b_0^2+b_1^2)/2} e^{-\gamma} \frac{e^{-1/\sigma^2}}{(\sigma^2)^2}
\end{aligned}$$

∗. the inverse gamma distribution $IG(a,b)$ has density

$$f(x) = \frac{b^a}{\Gamma(a)} \frac{e^{-b/x}}{x^{a+1}} 1_{[0,+\infty[}(x).$$

Example : STAR model

The MH algorithm is used with transition kernel derived from the random walk,

$$Q(x, y) \sim \mathcal{N}(x, \tau I_6).$$

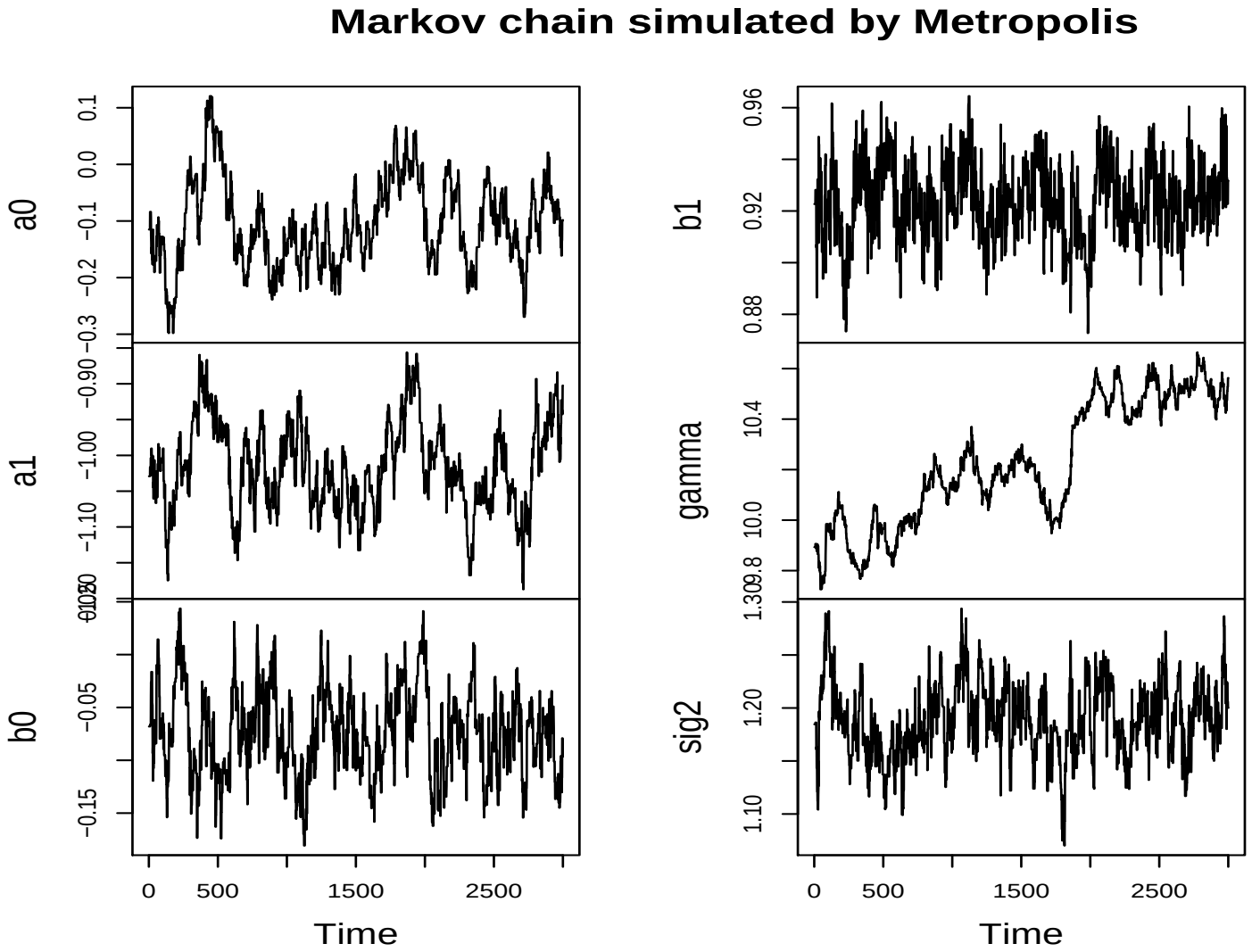
The probability of acceptance of θ^* is thus $\min\left(\frac{\pi(\theta^*|\mathbf{X})}{\pi(\theta^{(i-1)}|\mathbf{X})}, 1\right)$.

In practice the parameter τ is very important for the performance of the algorithm :

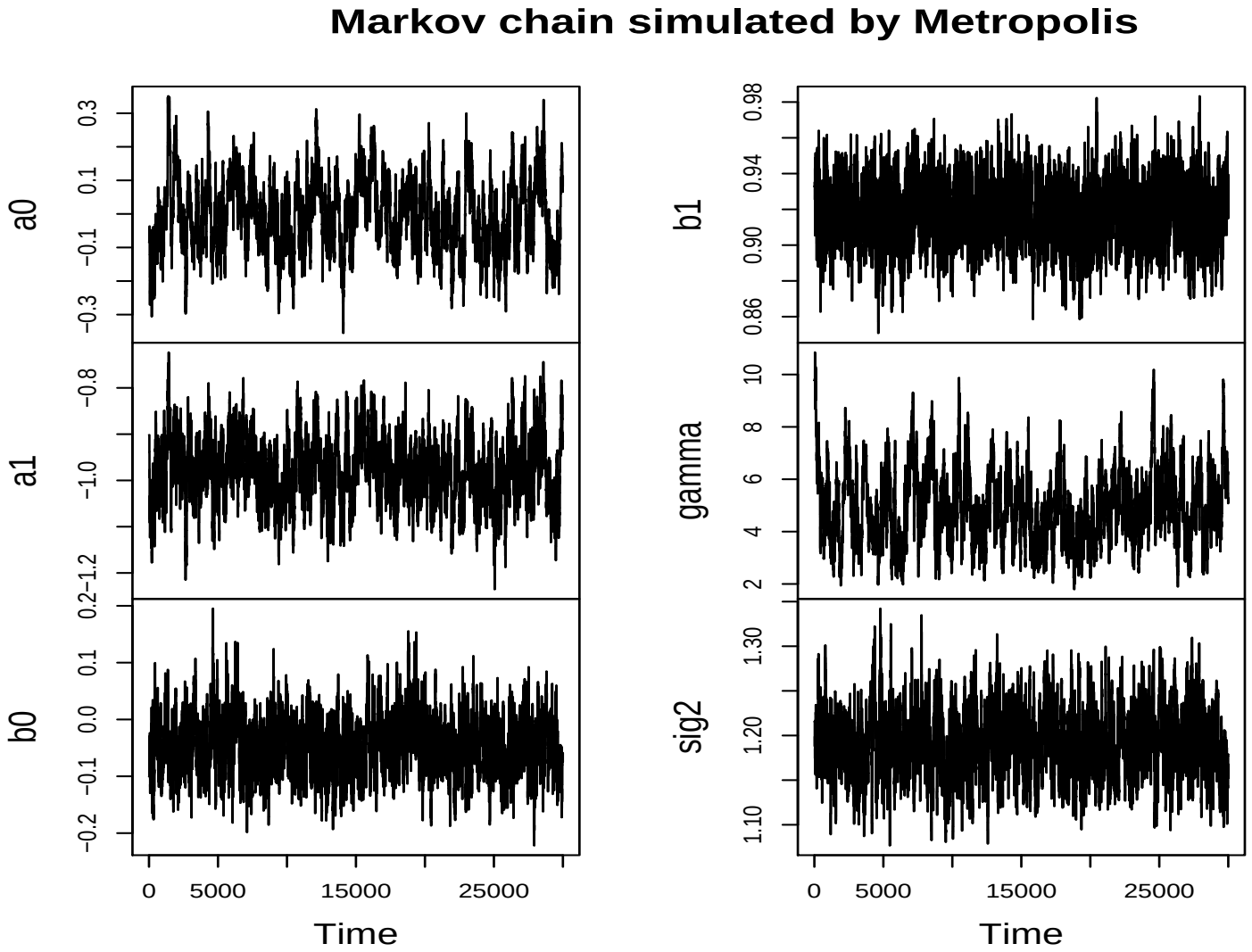
- large values of τ entail small rates of acceptance (θ^* is likely to fall in a region with low density $\pi(B|\mathbf{X})$, and thus is likely to be rejected).
- Small values of τ have high rates of acceptance and correspond to small moves of the MC.

For $\tau = 0.02$ and $n = 3000$ the acceptance rate for θ^* is 0.37 but the MC has not converged.

Example : STAR model



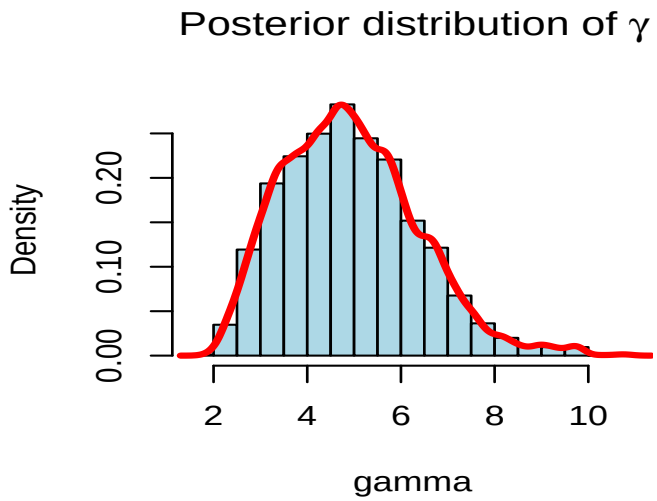
Example : STAR model

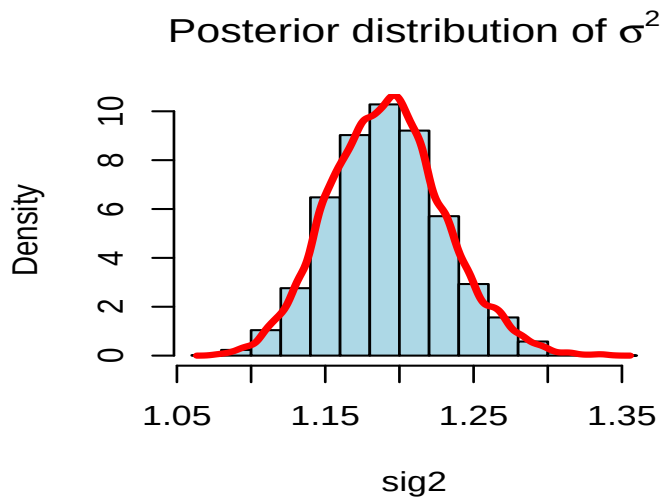


Example : STAR model

True parameter values :

$$a_0 = b_0 = c = 0, \quad a_1 = -0.95, \quad b_1 = 0.9, \quad \gamma = 5, \quad \sigma^2 = 1.225.$$





Gibbs sampling (Geman and Geman (1984), Gelfand and Smith (1990)

An alternative method for generating a MC having a specified invariant distribution.
However, the algorithm

- is not based on an acceptatance/rejection mechanism : all simulated values are accepted ;
- only applies to multivariate distributions ;
- requires more information than MH on the target law.

$\theta_{[-j]} \in \mathbb{R}^{d-1}$: parameter vector without θ_j.
We assume that the conditional densities

$$\pi^j(\theta_j \mid \mathbf{X}, \theta_{[-j]}), \quad j = 1, \ldots, d$$

are available and can be simulated.

Gibbs sampler

1. Choose initial values $\theta_j^{(0)}$, $j = 2,..d$.
2. For $i = 1,2,\ldots$
   1. Generate $\theta_1^{(i)}$ in the law $\pi^1(\cdot \mid \mathbf{X}, \theta_{[-1]}^{(i-1)})$
   2. For $\ell = 2,\ldots,d$, generate $\theta_\ell^{(i)}$ in the law

$$\pi^\ell(\cdot \mid \mathbf{X}, \theta_1^{(i)},\ldots,\theta_{\ell-1}^{(i)},\theta_{\ell+1}^{(i-1)},\ldots,\theta_d^{(i-1)})$$

   3. Take $\theta^{(i)} = (\theta_1^{(i)},\ldots,\theta_d^{(i)})$.

Remarks

- The simulated laws are those of the coordinates of θ, conditional on the other coordinates.
  The components can be multivariate : an extension consists in simulating blocks of θ, provided that their conditional distributions given the other blocks be available.
- It can be interesting to interpret the target law $\pi(\theta \mid \mathbf{X})$, as the marginal law of a higher-dimensional vector whose conditional distributions are easier to simulate than those of θ (data augmentation technique).

Why does it work ?

Let π_i the marginal densities (given $\mathbf{X}$) of the θ_i's. We introduice the *positivity* assumption :

$$\mathbf{H}: \quad \forall \theta, \qquad \prod_{i=1}^{d} \pi_i(\theta_i \,|\, \mathbf{X}) > 0 \quad \Longrightarrow \quad \pi(\theta \,|\, \mathbf{X}) > 0,$$

Theorem (Hammersley and Clifford (1970))

Under Assumption $\mathbf{H}$, *for all* θ, θ' *such that* $\pi(\theta \,|\, \mathbf{X}) > 0$ *and* $\pi(\theta' \,|\, \mathbf{X}) > 0$,

$$\frac{\pi(\theta \,|\, \mathbf{X})}{\pi(\theta' \,|\, \mathbf{X})} = \prod_{i=1}^{d} \frac{\pi^i(\theta_i \,|\, \mathbf{X}, \theta_1, \ldots, \theta_{i-1} \theta'_{i+1} \ldots \theta'_d)}{\pi^i(\theta'_i \,|\, \mathbf{X}, \theta_1, \ldots, \theta_{i-1} \theta'_{i+1} \ldots \theta'_d)}.$$

Proof :

For any θ such that $\pi(\theta \mid \mathbf{X}) > 0$,

$$\pi(\theta \mid \mathbf{X}) = \pi^d(\theta_d \mid \mathbf{X}, \theta_1, \ldots, \theta_{d-1}) \pi_{[-d]}(\theta_1, \ldots, \theta_{d-1} \mid \mathbf{X})$$

where $\pi_{[-i]}(\cdot \mid \mathbf{X})$ is the density of the vector θ without θ_i.
Moreover, for all θ'

$$\pi(\theta_1, \ldots, \theta_{d-1}, \theta'_d \mid \mathbf{X}) = \pi^d(\theta'_d \mid \mathbf{X}, \theta_1, \ldots, \theta_{d-1}) \pi_{[-d]}(\theta_1, \ldots, \theta_{d-1} \mid \mathbf{X}).$$

It follows that

$$\pi(\theta \mid \mathbf{X}) = \frac{\pi^d(\theta_d \mid \mathbf{X}, \theta_1, \ldots, \theta_{d-1})}{\pi^d(\theta'_d \mid \mathbf{X}, \theta_1, \ldots, \theta_{d-1})} \pi(\theta_1, \ldots, \theta_{d-1}, \theta'_d \mid \mathbf{X}).$$

Proof (continued)

Similarly

$$\begin{aligned}
& \pi(\theta_1,\ldots,\theta_{d-1},\theta'_d \,|\, \mathbf{X}) \\
= \;& \pi^{d-1}(\theta_{d-1} \,|\, \mathbf{X},\theta_1,\ldots,\theta_{d-2},\theta'_d)\pi_{[-(d-1)]}(\theta_1,\ldots,\theta_{d-2},\theta'_d \,|\, \mathbf{X}), \\
& \pi(\theta_1,\ldots,\theta'_{d-1},\theta'_d \,|\, \mathbf{X}) \\
= \;& \pi^{d-1}(\theta'_{d-1} \,|\, \mathbf{X},\theta_1,\ldots,\theta_{d-2},\theta'_d)\pi_{[-(d-1)]}(\theta_1,\ldots,\theta_{d-2},\theta'_d \,|\, \mathbf{X}),
\end{aligned}$$

hence

$$\begin{aligned}
\pi(\theta \,|\, \mathbf{X}) = \; & \frac{\pi^d(\theta_d \,|\, \mathbf{X},\theta_1,\ldots,\theta_{d-1})}{\pi^d(\theta'_d \,|\, \mathbf{X},\theta_1,\ldots,\theta_{d-1})} \frac{\pi^{d-1}(\theta_{d-1} \,|\, \mathbf{X},\theta_1,\ldots,\theta_{d-2},\theta'_d)}{\pi^{d-1}(\theta'_{d-1} \,|\, \mathbf{X},\theta_1,\ldots,\theta_{d-2},\theta'_d)} \\
& \times \pi(\theta_1,\ldots,\theta_{d-2},\theta'_{d-1},\theta'_d \,|\, \mathbf{X}).
\end{aligned}$$

Continuing this way the replacement of the coordinates of θ by those of θ', we get the announced formula.

Proof (continued)

Finally, we check the positivity of the denominator in this formula. We have

$$\begin{aligned}
& \pi(\theta \mid \mathbf{X}) > 0, \quad \pi(\theta' \mid \mathbf{X}) > 0 \\
\Longrightarrow \quad & \pi_i(\theta_i \mid \mathbf{X}) > 0, \ i = 1, \ldots, d-1 \ \text{ and } \ \pi_d(\theta'_d \mid \mathbf{X}) > 0 \\
\Longrightarrow \quad & \pi(\theta_1, \ldots, \theta_{d-1}, \theta'_d \mid \mathbf{X}) > 0, \qquad \text{by assumption } \mathbf{H} \\
\Longrightarrow \quad & \pi^d(\theta'_d \mid \mathbf{X}, \theta_1, \ldots, \theta_{d-1}) > 0.
\end{aligned}$$

The other terms of the product are treated similarly.

Remarks

1. This relation allows to get, theoretically, the joint density from the conditional densities (by integrating the inverses of each side of the equality w.r.t. θ') :

$$\pi(\theta \mid \mathbf{X}) = \left(\int \prod_{i=i}^{d} \frac{\pi^i(\theta'_i \mid \mathbf{X}, \theta_1, \ldots, \theta_{i-1} \theta'_{i+1} \ldots \theta'_d)}{\pi^i(\theta_i \mid \mathbf{X}, \theta_1, \ldots, \theta_{i-1} \theta'_{i+1} \ldots \theta'_d)} d\theta' \right)^{-1}.$$

2. It can be shown that under Assumption **H**, $(\theta^{(i)})$ is an ergodic Markov chain admitting $\pi(\cdot|\mathbf{X})$ as invariant probability measure.

Hybrid algorithms

It can be worth combining the Gibbs and MH algorithms :

- the MH algorithm does not take into account possible available information on the conditional laws. Moreover, it is not well suited for hierarchical structures (such as the SV model).
- the Gibbs algorithm cannot be used alone if certain conditional densities are not available in closed form or cannot be easily simulated.
- Even if the Gibbs algorithm can be implemented, the convergence to the stationary distribution of the MC can be very slow because the components are modified one by one.

AR(1) model

$$X_t = \omega + \beta X_{t-1} + \sigma v_t, \quad (v_t)\ iid\ \mathcal{N}(0,1), \quad \sigma > 0.$$

Aim : generate a MC with stationary *a posteriori* distribution, $\pi(\theta \mid \mathbf{X})$, where $\theta = (\omega, \beta, \sigma)'$.

Prior laws : (ω, β) and σ^2 are *a priori* independent with

$$(\omega, \beta) \sim \mathcal{N}\left((\omega^0, \beta^0), \Sigma^0\right), \qquad \sigma^2 \sim IG(a, b)$$

$a > 0, b > 0, \omega^0 \in \mathbb{R}, \beta^0 \in \mathbb{R}, \Sigma^0$ positive definite.

The *posterior* **conditional laws** of (ω, β) and σ are denoted

$$\pi^1((\omega, \beta) \mid \mathbf{X}, \sigma) \quad \text{and} \quad \pi^2(\sigma^2 \mid \mathbf{X}, \omega, \beta).$$

Notations

Likelihood conditional to a fixed initial value X_0 :

$$f(\mathbf{X} \mid \theta) \propto \exp\left\{ \frac{-1}{2\sigma^2} \sum_{t=1}^{n} (X_t - \omega - \beta X_{t-1})^2 \right\}.$$

$(\hat{\omega}, \hat{\beta})$: LS estimator of de (ω, β).

$$\Sigma_n = \Sigma_n(\sigma) = \sigma^2 (\underline{\mathbf{X}}' \underline{\mathbf{X}})^{-1}, \qquad \underline{\mathbf{X}}' = \begin{pmatrix} 1 & \dots & 1 \\ X_0 & \dots & X_{n-1} \end{pmatrix},$$

$$\Sigma^* = \{(\Sigma^0)^{-1} + \Sigma_n^{-1}\}^{-1},$$

$$(\omega^*, \beta^*)' = \Sigma^* \left\{ \Sigma_n^{-1} (\hat{\omega}, \hat{\beta})' + (\Sigma^0)^{-1} (\omega^0, \beta^0)' \right\}.$$

LS estimator

We have

$$\begin{aligned}\sum_{t=1}^{n}(X_t-\omega-\beta X_{t-1})^2 &= \sum_{t=1}^{n}[X_t-\hat{\omega}-\hat{\beta}X_{t-1}-\{\omega-\hat{\omega}+(\beta-\hat{\beta})X_{t-1}\}]^2 \\ &= \sum_{t=1}^{n}(X_t-\hat{\omega}-\hat{\beta}X_{t-1})^2+\sum_{t=1}^{n}\{\omega-\hat{\omega}+(\beta-\hat{\beta})X_{t-1}\}^2,\end{aligned}$$

A posteriori distribution of (ω, β)

$$
\begin{aligned}
\pi^1((\omega,\beta)\,|\,\mathbf{X},\sigma) &\propto \exp\left\{\frac{-1}{2\sigma^2}\sum_{t=1}^{n}(X_t-\omega-\beta X_{t-1})^2\right\} \\
&\quad \times \exp\left\{\frac{-1}{2}(\omega-\omega^0,\beta-\beta^0)(\Sigma^0)^{-1}(\omega-\omega^0,\beta-\beta^0)'\right\} \\
&\propto \exp\left\{\frac{-1}{2\sigma^2}\sum_{t=1}^{n}\{(\omega-\hat{\omega})+(\beta-\hat{\beta})X_{t-1}\}^2\right\} \\
&\quad \times \exp\left\{\frac{-1}{2}(\omega-\omega^0,\beta-\beta^0)(\Sigma^0)^{-1}(\omega-\omega^0,\beta-\beta^0)'\right\} \\
&\propto \exp\left\{\frac{-1}{2}(\omega-\hat{\omega},\beta-\hat{\beta})\Sigma_n^{-1}(\omega-\hat{\omega},\beta-\hat{\beta})'\right\} \\
&\quad \times \exp\left\{\frac{-1}{2}(\omega-\omega^0,\beta-\beta^0)(\Sigma^0)^{-1}(\omega-\omega^0,\beta-\beta^0)'\right\} \\
&\propto \exp\left\{\frac{-1}{2}(\omega-\omega^*,\beta-\beta^*)(\Sigma^*)^{-1}(\omega-\omega^*,\beta-\beta^*)'\right\}.
\end{aligned}
$$

A property of quadratic forms

For the last equality, we used a useful result on quadratic forms :

Let x, a, b some $k \times 1$ vectors, A and B some symmetric $k \times k$ matrices such that $(A+B)^{-1}$ exists. Then

$$\begin{aligned}(x-a)'A(x-a)+(x-b)'B(x-b) &= (x-c)'(A+B)(x-c) \\ &\quad +(a-b)'A(A+B)^{-1}B(a-b)\end{aligned}$$

where $c=(A+B)^{-1}(Aa+Bb)$.

See Box and Tiao (1973) p. 418.

A posteriori distributions

A posteriori distribution of (ω, β) : $\mathcal{N}((\omega^*, \beta^*), \Sigma^*)$.

A posteriori distribution of σ^2 :

$$\pi^2(\sigma^2 \mid \mathbf{X}, \omega, \beta) \quad \propto \quad \frac{e^{\frac{-n\overline{Y}}{2\sigma^2}}}{\sigma^n} \frac{e^{-b/\sigma^2}}{\sigma^{2(a+1)}} = \frac{e^{\frac{-1}{2\sigma^2}(n\overline{Y}+2b)}}{\sigma^{2(a+1+\frac{n}{2})}}.$$

where $\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$ and $Y_t = (X_t - \omega - \beta X_{t-1})^2, \ t = 1, \ldots, n.$

Thus $\pi^2(\sigma^2 \mid \mathbf{X}, \omega, \beta)$ is the density of the law $IG(a + \frac{n}{2}, b + \frac{n\overline{Y}}{2})$.

Gibbs sampler for the AR(1)

1. Specify the hyperparameters $a, b, \omega^0, \beta^0, \Sigma^0$.
2. Choose an initial value $\sigma^{(0)}$.
3. For $i = 1, 2, \ldots$
   1. Generate $(\omega^{(i)}, \beta^{(i)})$ in the law $\mathcal{N}((\omega^*, \beta^*), \Sigma^*)$ obtained for $\Sigma_n = \Sigma_n(\sigma^{(i-1)})$.
   2. Generate $\sigma^{(i)}$ in the law $IG(a + \frac{n}{2}, b + \frac{n\overline{Y}^{(i)}}{2})$ where

$$\overline{Y}^{(i)} = \frac{1}{n} \sum_{t=1}^{n} (X_t - \omega^{(i)} - \beta^{(i)} X_{t-1})^2.$$

   3. Take $\theta^{(i)} = (\omega^{(i)}, \beta^{(i)}, \sigma^{(i)})'$.

Comments

1. The standard laws used in the algorithm, can be simulated efficiently using numerous statistical softwares such as `Gauss`, `Mathematica`, `Matlab`, `R`.
2. This choice of *prior* distributions leads to standard *posterior* laws. It is possible to choose other laws, and to use the MH for the simulation of non standard *posterior* laws.
3. The stationarity condition $|\beta| < 1$ can be taken into account by using, for the *prior* of (ω, β), a troncated bivariate Gaussian law on $\mathbb{R} \times]-1,1[$. The same troncature follows on the *a posteriori* law. Alternatively, one can choose for β a *prior* density with support $[-1,1]$, for instance the law obtained for $\beta = 2X - 1$ when $X \sim \mathcal{B}(\beta_1, \beta_2)$. We get the *a priori* density

$$\pi(\beta) = 0.5 \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} \{0.5(1+\beta)\}^{\beta_1 - 1} \{0.5(1-\beta)\}^{\beta_2 - 1} \mathbf{1}_{[-1,1]}$$

where $\beta_1, \beta_2 > 0.5$ are hyperparameters.

STAR(1) model

For the ◂ STAR model with $c = 0$, we have for a given γ,

$$Y_t = Z'_{t-1}\beta + \epsilon_t,$$

where ϵ_t is iid $\mathcal{N}(0, \sigma^2)$, and

$$Z_t = \begin{pmatrix} 1 - G_t(\gamma) \\ Y_t\{1 - G_t(\gamma)\} \\ G_t(\gamma) \\ Y_t G_t(\gamma) \end{pmatrix}, \quad \beta = \begin{pmatrix} a_0 \\ a_1 \\ b_0 \\ b_1 \end{pmatrix}, \quad G_t(\gamma) = \frac{1}{1 + \exp(-\gamma Y_t)}.$$

STAR(1) model

Calculations similar to those done for the AR(1) show that the conditional laws of β and σ^2 are explicit :

$$P_1(\beta \mid Y, \gamma, \sigma^2) \sim \mathcal{N}\left\{\left(I_4 + \Sigma_n^{-1}\right)^{-1} \Sigma_n^{-1} \hat{\beta},\ \left(I_4 + \Sigma_n^{-1}\right)^{-1}\right\},$$

$$P_2(\sigma^2 \mid Y, \beta, \sigma^2) \sim IG\left(1 + (n-1)/2,\ 1 + \sum_{t=2}^{n} \epsilon_t^2/2\right),$$

The conditional law of γ is not explicit but satisfies

$$P_3(\gamma \mid Y, \beta, \sigma^2) \propto \exp\left\{-\sum_{t=2}^{n} \epsilon_t^2(\gamma)/(2\sigma^2) - \gamma\right\} \mathbf{1}_{\{\gamma>0\}},$$

$$\epsilon_t(\gamma) = Y_t - \beta' Z_{t-1}.$$

By using the MH algorithm to simulate the conditional law P_3, one gets an **hybrid method** combining the MH and Gibbs algorithms.

STAR(1) model : Sample paths and *a posteriori* laws of γ and σ^2, obtained for 1000 iterations of the hybrid algorithm