Bayesian methods for structural break modellings

Arnaud Dufays
Centre de Recherche en Economie et Statistique (CREST)

Course Structure

- **Chapter 1 :**

  Bayesian concepts and inference

- **Chapter 2 :**

  Structural breaks for models *without path dependence*

- **Chapter 3 :**

  Structural breaks for models *with path dependence*

- **Chapter 4 :**

  Introduction to Bayesian econometrics using Matlab

Chapter 1

Bayesian inference : Principles

Bayesian inference

Differences from classic approach :

- Model parameters (random vs fixed)
- Finite sample vs asymptotic theory
- Statistical interpretation (subjective vs objective)

Historical consideration

From effects to causes

- Consider two non-independent events : A and B

- From basic axiom of probability :

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \textbf{and} \quad P(B|A) = \frac{P(A,B)}{P(A)}$$

Bayes' rule : $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

What is the probability of observing A if B has occurred ?

Generalization

- Multiple events :

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

Prior Distribution

Posterior Distribution

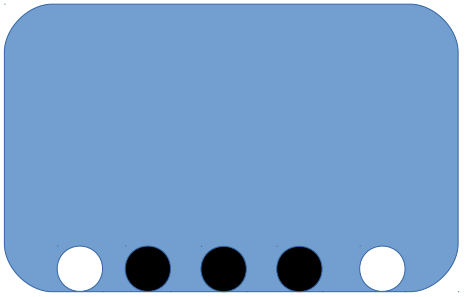
Normalizing constant / Marginal likelihood

Normalizing constant : $P(B) = \sum_{j=1}^{k} P(B|A_j)P(A_j)$

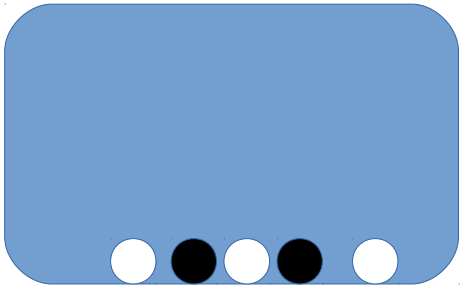
- Continuous case : Sum becomes an integral.

From effects to causes : Examples

Urn 1



Urn 2



One draw : A black ball

What is the probability that the ball comes from urn 1 ?

Prior distribution $P(U_1) = P(U_2) = 0,5$

Post distribution $$P(U_1|B) = \frac{P(B|U_1)P(U_1)}{P(B|U_1)P(U_1) + P(B|U_2)P(U_2)}$$

$$P(U_1|B) = 0.6$$

From effects to causes : Examples

Unwin, S. D. , *'The probability of God : A simple Calculation that proves the ultimate truth'*, 2004

- A priori, God may or may not exist : P(*G* exists) = 0,5
- Observations : Miracles, Wars, Sins, ...

$$P(G \text{ exists}|B) = \frac{P(B|G \text{ exists})P(G \text{ exists})}{P(B|G \text{ exists})P(G \text{ exists}) + P(B|G \text{ not exist})P(G \text{ not exist})}$$

Conclusion... You should bet the Lord exists

Principles of Bayesian inference

- Any unobserved value is random since we shall do a probabilistic statement on it.

  → Distributions on model parameters (prior to observing data)

- Data supposedly generated by the model contain information about the parameters.

  → Likelihood principle

- Through Bayes theorem, the parameter distribution is updated.

  → Posterior distribution of the parameters

Statistical Example

- The model :

$$\begin{cases} y_t = \theta + \epsilon_t \\ \epsilon_t \sim \text{ i.i.d. } N(0,1) \end{cases}$$

- Observations :

$$Y_{1:T} = \{y_1, ..., y_T\}'$$

- Prior distribution :

$$\theta \sim N(\mu_0, n_0^{-1})$$

Example

$$\begin{cases} \mu_0 = 0 \\ n_0^{-1} = 100 \end{cases}$$

- Posterior distribution :

$$\theta | Y_{1:T} \sim N(\mu^*, \sigma^*)$$

$$\mu^* = \frac{\sum_{t=1}^{T} y_t + n_0 \mu_0}{T + n_0}$$

$$\sigma^* = (T + n_0)^{-1}$$

Prior beliefs

Based on another set of observations

- Reflects the uncertainty on the parameter

Calculus

Bayes' rule : $\pi(\theta|Y_{1:T}) = \frac{f(Y_{1:T}|\theta)f(\theta)}{f(Y_{1:T})}$

$\propto f(Y_{1:T}|\theta)f(\theta)$

Likelihood function : $f(Y_{1:T}|\theta) = (2\pi)^{-\frac{T}{2}} e^{-0,5\sum_{t=1}^{T}(y_t-\theta)^2}$

Prior distribution : $f(\theta) = \frac{e^{\frac{-(\theta-\mu_0)^2}{2n_0}}}{\sqrt{2\pi n_0}}$

The posterior must be a proper distribution :
insured by the normalizing constant

⟶ We can drop all the terms that do not depend on θ

and see if the posterior kernel comes from a known distribution

Calculus

Posterior kernel :

$$\begin{aligned}\pi(\theta|Y_{1:T}) &\propto e^{-0,5[n_0(\theta-\mu_0)^2+\sum_{t=1}^{T}(y_t-\theta)^2)]} \\ &\propto e^{-0,5[\theta^2(T+n_0)-2\theta(n_0\mu_0+\sum_{t=1}^{T}y_t)]} \\ &\propto e^{\frac{-(\theta-\mu^*)^2}{2\sigma^*}}\end{aligned}$$

The posterior kernel is a normal one with

$$\mu^* = \frac{\sum_{t=1}^{T} y_t + n_0\mu_0}{T+n_0}$$

$$\sigma^* = (T+n_0)^{-1}$$

Discussion

If $T \to \infty,\ E(\theta|Y_{1:T}) \to \bar{y},\ V(\theta|Y_{1:T}) \to 0$

Data increasingly dominate the prior information.

If $n_0 = 0$ and $\mu_0 = 0$:

Same estimator as in the classical approach.

- Bayesian inference provides an entire distribution only based on the observed data.
- Delivers different statistical interpretations.

Summarizing the posterior

- Posterior means and standard deviations

$$E(\theta|Y_{1:T}) = \int_{-\infty}^{\infty} \theta\pi(\theta|Y_{1:T})d\theta$$

$$V(\theta|Y_{1:T}) = \int_{-\infty}^{\infty} (\theta - E(\theta|Y_{1:T}))^2\pi(\theta|Y_{1:T})d\theta$$

- Credible intervals
- Posterior Covariance matrix
- Quantiles and graphics of the marginal distributions of the parameters

Criteria for statistical procedure

Classical

- Properties (Consistency, efficiency,...) from hypothetical repeated samples/ large sample.

Bayesian

- Only based on the observed sample used 'coherently' through the likelihood principle.

Caution :

- Some Bayesian state that Bayesian inference is 'exact' in finite sample

  → No meaning since what happens in repeated sample is not relevant for Bayesian inference.

  → Based on very restrictive assumptions.

Treatment of parameters

Classical

- Fixed parameters in reality

Bayesian

- Random parameters in reality
- Fixed parameters but subjective probabilistic statement

Eases the interpretation :

Classical

95 % Confidence interval covers the true value of the parameter in nineteen out of twenty trials on average.

Bayesian

95 % credible interval gives the region of the parameter space where the probability of covering θ is equal to 95.

Bayesian learning

The posterior distribution as the prior distribution ?

$$\begin{aligned}\pi(\theta|Y_{1:T}) &= \frac{f(Y_{1:T}|\theta)f(\theta)}{f(Y_{1:T})} \\ &= \frac{f(Y_{t+1:T}|\theta)\pi(\theta|Y_{1:t})}{f(Y_{t+1:T}|Y_{1:t})} \quad \forall \;\; t \in [1, T-1]\end{aligned}$$

New information makes update our belief

Core idea of Sequential Monte Carlo

- Teglas, E. et al., '*Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference*'. 2011.

Complex Bayesian inference

What happens if the posterior distribution is not a known parametric distribution ?

Small dimension

- Deterministic integration
- Direct sampling
- Importance sampling

Medium dimension

- Sequential Monte Carlo (SMC)
- Annealed importance Sampling (SMC sampler)

High dimension

- **Markov-chain Monte Carlo (MCMC)**
- Approximate Bayesian Computation (ABC)
- Variational Bayes methods

Markov-chain Monte Carlo

Markov chain Monte Carlo

Posterior distribution :

1) Not a known parametric distribution.
2) Cannot sample directly from it.
3) Parameter dimension greater than 3.

MCMC principles :

1) Simulation of a Markov-chain exhibiting the posterior distribution of interest as invariant one.

MCMC Outputs :

1) (Correlated) draws of the posterior distribution.
2) LLN theorem allows to approximate any deterministic function of the posterior distribution.

Invariant distribution

Markov-chain :

- Characterizes by a number of reachable states (discrete or continuous), a transition probability matrix (P), a probability vector of being in the initial state.
- Chain that only depends on the current state for moving to the next

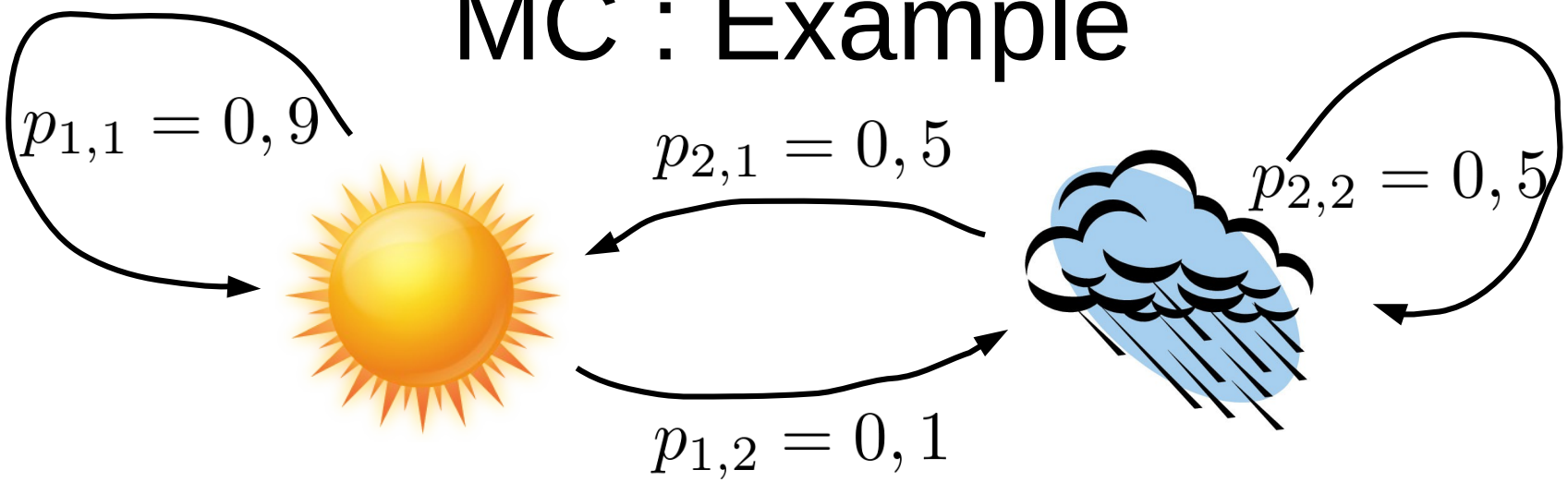
$$s_1 \rightarrow s_2 \rightarrow \ldots \rightarrow s_N$$

Moves according to the transition probability matrix P

Invariant distribution (q) : Independent from the initial state

$$q'P = q' \rightarrow$$ Unchanged when transformed by P

MC : Example



Transition matrix : $P = \begin{pmatrix} 0,9 & 0,1 \\ 0,5 & 0,5 \end{pmatrix}$

Invariant Distribution

$$q'P = q'$$

$$\downarrow$$

$$q' = [0,8333 \quad 0,167]$$

Ergodicity

From any initial point, the MC converges to the invariant distribution.

$$P^{100} = \begin{pmatrix} 0,833 & 0,167 \\ 0,833 & 0,167 \end{pmatrix} \downarrow$$

Based on 10000 simulated draws

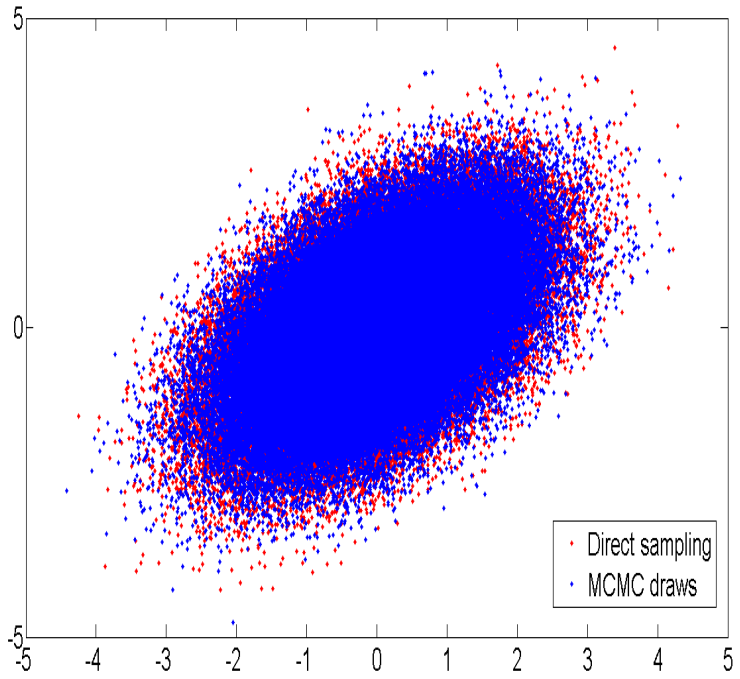
$$q' = [0,8332 \quad 0,1668]$$

MCMC : Example

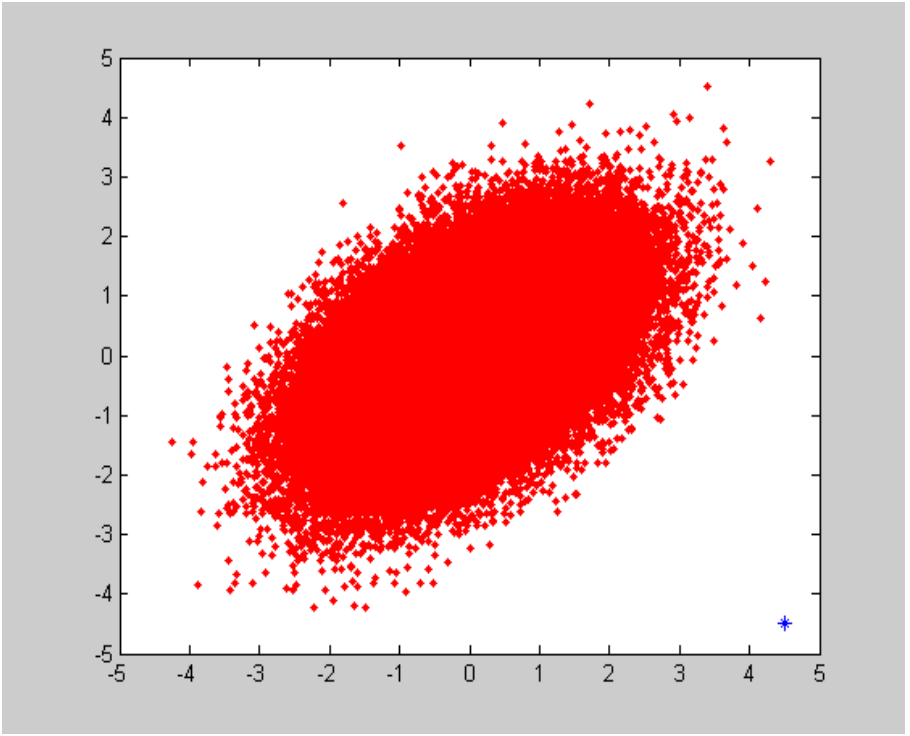
Inversion of the problem

Invariant Dist. : $N(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0,5\\0,5 & 1\end{pmatrix})$ + Ergodicity

Invariant Distribution



Ergodicity



MCMC

- Markov-chain exhibiting the posterior distribution as invariant one :

Sufficient condition : $\pi(\theta|Y_{1:T})K(\theta^*|\theta) = \pi(\theta^*|Y_{1:T})K(\theta|\theta^*)$

Reversibility condition – Detailed balance

- Markov-chain that is ergodic

→ From any initial point, the MC converges to the invariant distribution.

Sufficient condition :

Irreducible → Able to visit all sets A such that $\int_A \pi(\theta|Y_{1:T})d\theta > 0$ from any starting point

Aperiodic → Does not cycle through a finite number of sets

Positive Harris-recurrent

Beyond the scope of the course.

MCMC : pros and cons

- MCMC outputs : correlated draws of the posterior.

For any π-integrable real valued functions h,

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta_i) \rightarrow \int h(\theta)\pi(\theta|Y_{1:T})d\theta \quad as \quad N \rightarrow \infty, \quad a.s.$$

- Posterior expectation : $h(\theta) = \theta$
- Posterior variance : $h(\theta) = (\theta - E(\theta|Y_{1:T}))(\theta - E(\theta|Y_{1:T}))'$

Rate of convergence :

Burn-in period

- Ergodicity implies convergence from any initial point but after how many iterations ?

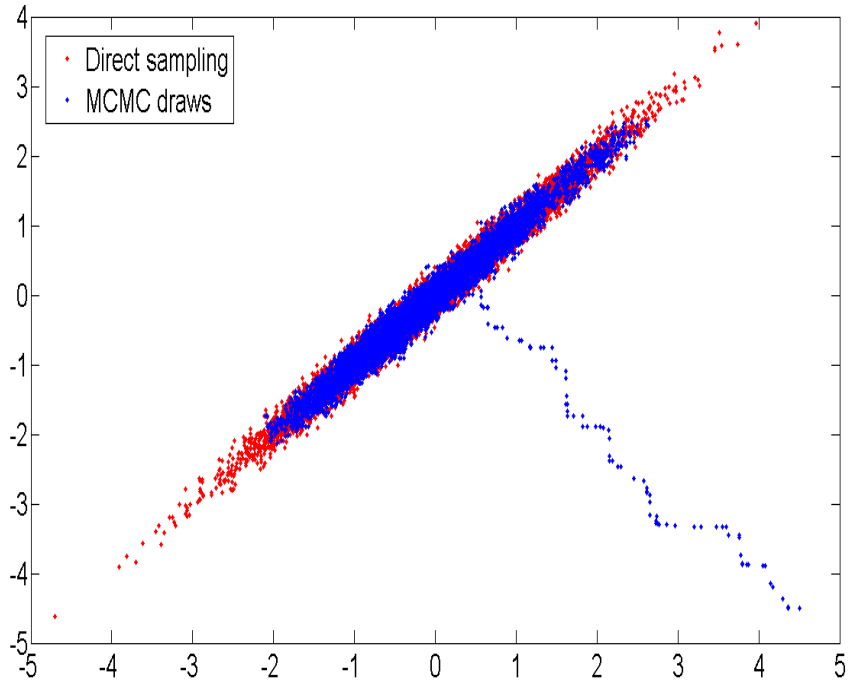
  Criteria : Geweke, Gelman and Rubin, Cusum plot, ...

N ?

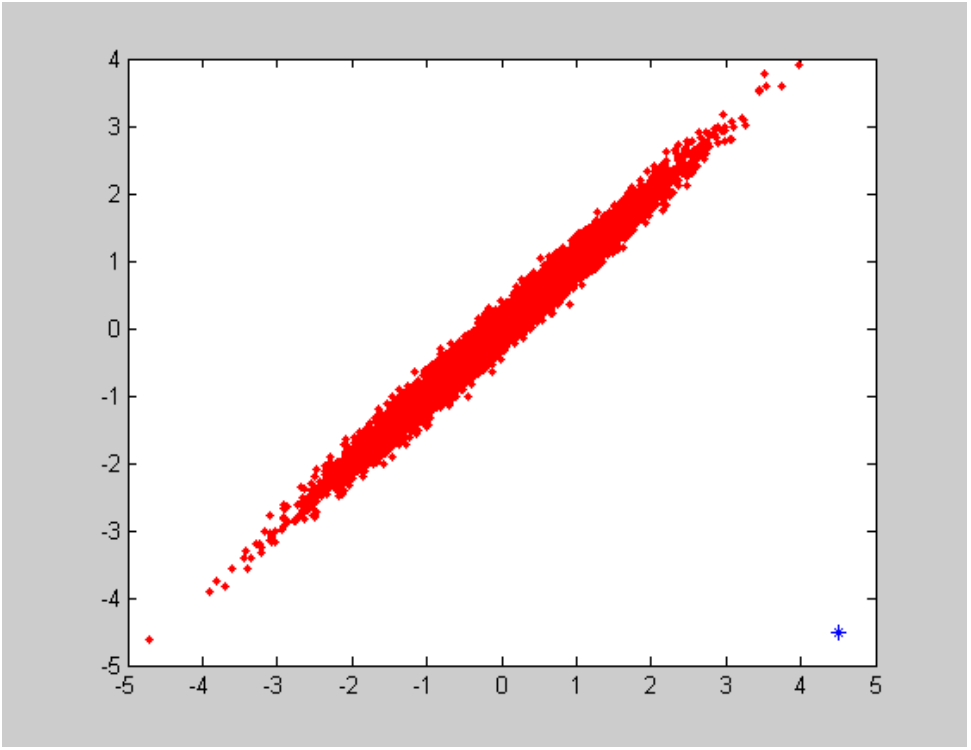
- How many iterations after convergence ? Criterion : Autocorrelation time

MCMC : Burn-in size

Invariant Dist. : $N(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0,99\\0,99 & 1\end{pmatrix})$



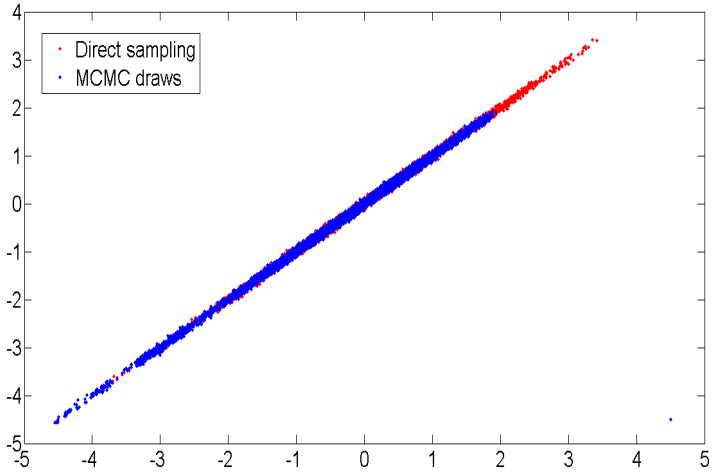
After 10000 draws



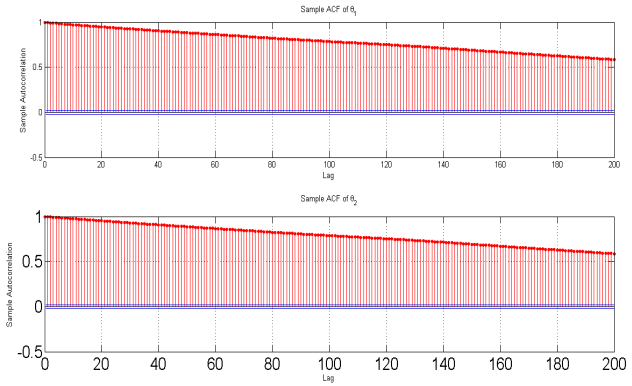
First 200 draws of the MCMC

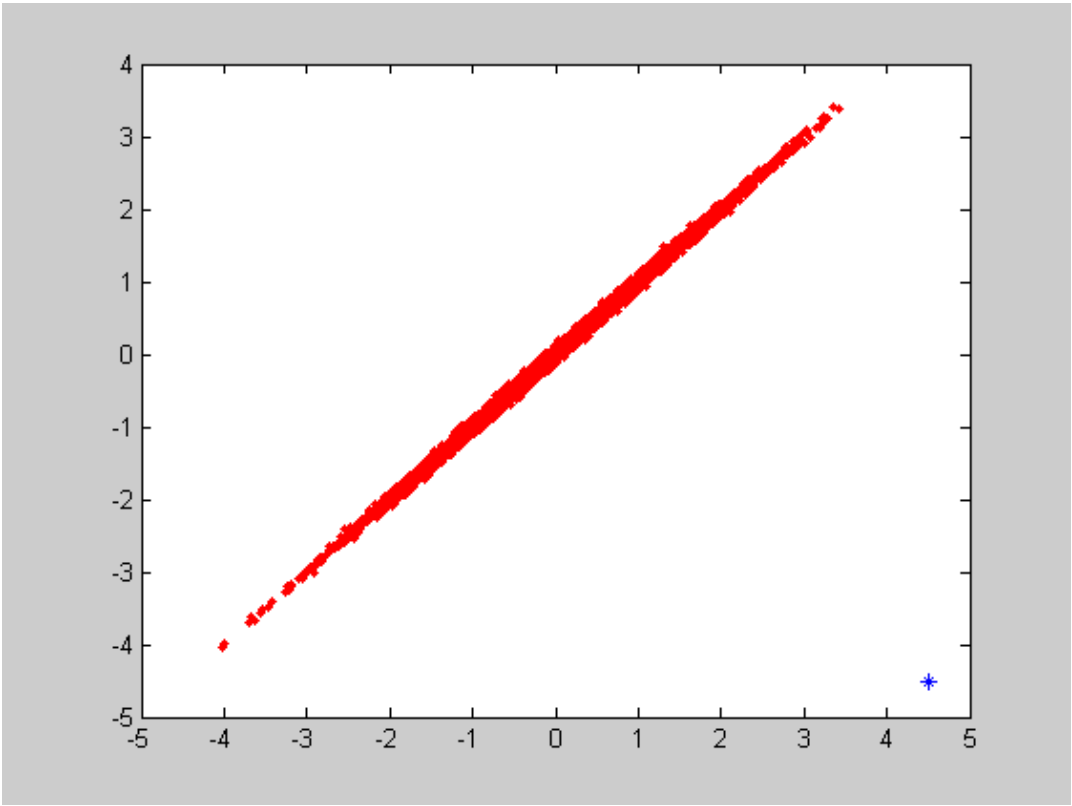
MCMC : Mixing problem

Invariant Dist. : $N(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0,999\\0,999 & 1\end{pmatrix})$



After 10000 draws





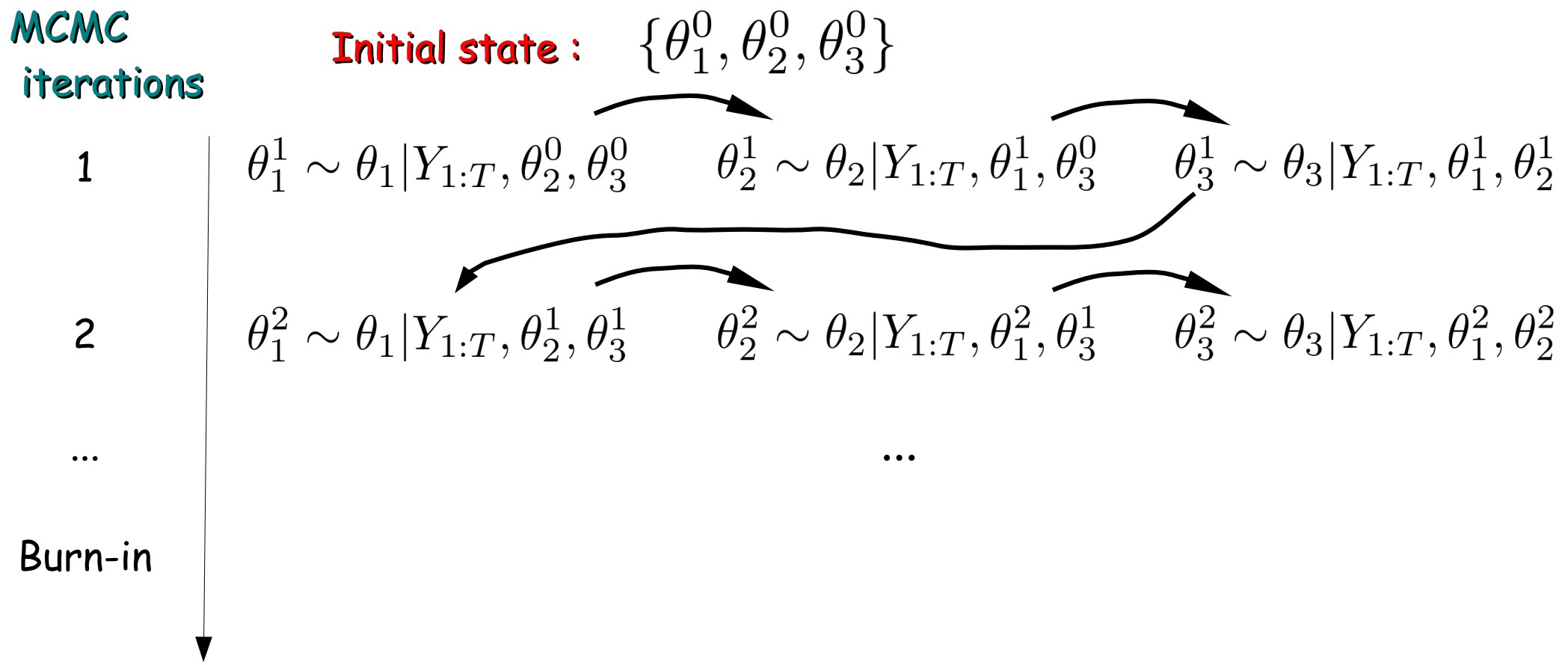
First 200 draws of the MCMC

MCMC : Gibbs sampler

- Iterations on *full conditional distributions*

Inference on $\theta = \{\theta_1', \theta_2', \theta_3'\}'$

Dimension of blocks can be higher than one



Gibbs sampler : Example

- Auto-Regressive process (AR) : $Y_{1:T} = \{y_1, ..., y_T\}'$

$$\begin{aligned} y_t &= \theta_0 + \theta_1 y_{t-1} + \epsilon_t \\ &= \theta' x_t + \epsilon_t \\ \epsilon_t &\sim_{i.i.d.} N(0, \sigma^2) \end{aligned}$$

Prior distributions

$$\theta \sim N(\mu_0, \Sigma_0)$$
$$\sigma^2 \sim IG(\alpha, \beta)$$

- Iterations on *full conditional distributions*

Prior component | Likelihood

$$\begin{aligned} \pi(\theta | Y_{1:T}, \sigma^2) &\propto e^{-0,5(\theta-\mu_0)'\Sigma_0^{-1}(\theta-\mu_0) - 0,5\sum_{t=1}^{T}\frac{(y_t-\theta' x_t)^2}{\sigma^2}} \\ &\sim N(\bar{\mu}, \bar{\Sigma}) \end{aligned}$$

$$\bar{\mu} = \bar{\Sigma}[\sigma^{-2}\sum_{t=1}^{T} x_t y_t + \Sigma_0^{-1}\mu_0] \qquad \bar{\Sigma} = [\sigma^{-2}\sum_{t=1}^{T}(x_t x_t') + \Sigma_0^{-1}]^{-1}$$

Gibbs sampler : Example

- Auto-Regressive process (AR) :

$$\begin{array}{rcl} y_t & = & \theta' x_t + \epsilon_t \\ \epsilon_t & \sim_{i.i.d.} & N(0, \sigma^2) \end{array}$$

Prior distributions

$$\begin{array}{c} \theta \sim N(\mu_0, \Sigma_0) \\ \sigma^2 \sim IG(\alpha, \beta) \end{array}$$

- Iterations on *full conditional distributions*

$$\pi(\theta | Y_{1:T}, \sigma^2) \sim N(\bar{\mu}, \bar{\Sigma})$$

$$\pi(\sigma^2 | Y_{1:T}, \theta) \propto \underbrace{(\sigma^2)^{-\alpha-1} e^{-\beta\sigma^{-2}}}_{\text{Prior component}} \underbrace{(\sigma^2)^{-T/2} e^{-\sigma^{-2}[\sum_{t=1}^{T} \epsilon_t^2/2]}}_{\text{Likelihood}}$$

$$\sim IG(\alpha + T/2, \beta + \sum_{t=1}^{T} \epsilon_t^2/2)$$

- If Z~IG(a,b) then 1/Z~G(a,1/b)

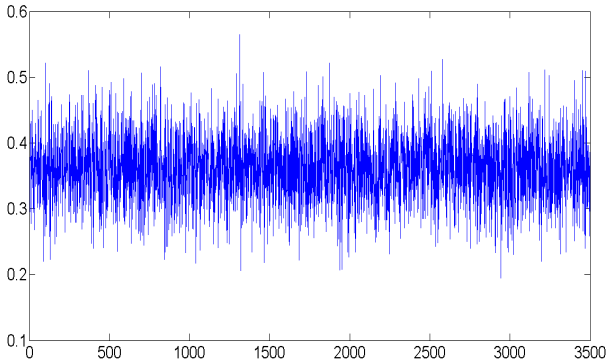
Gibbs sampler : Example

- AR(1) : 1000 Observations

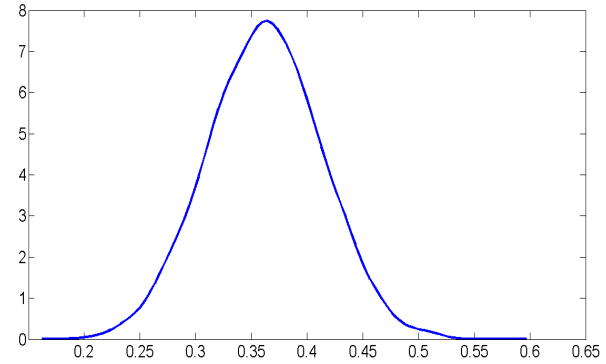
$$y_t = 0,4 + 0,7y_{t-1} + \epsilon_t$$
$$\epsilon_t \sim_{i.i.d.} N(0,2)$$

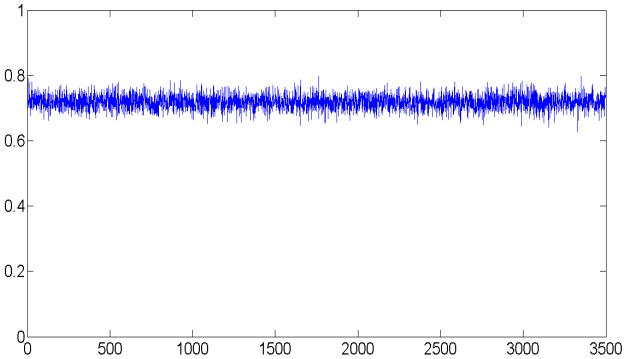
Prior distributions

$$\theta \sim N([0 \;\; 0]', 100I_2)$$
$$\sigma^2 \sim IG(2,2)$$

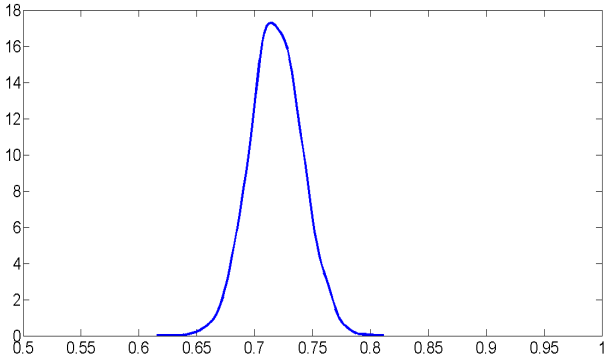


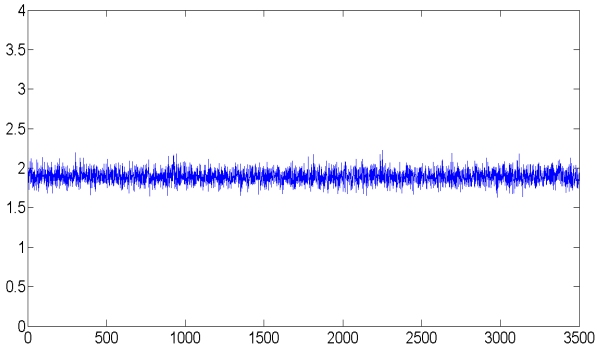
$E(\theta_0|Y_{1:T}) \approx 0,36$



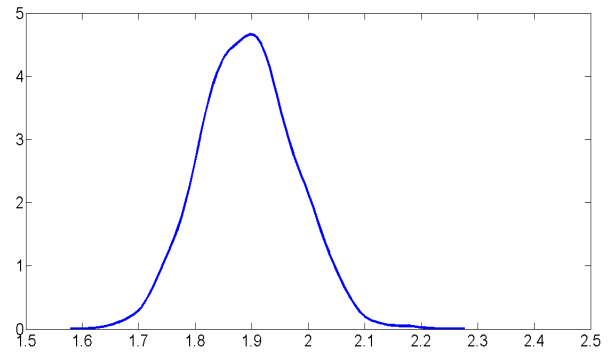


$E(\theta_1|Y_{1:T}) \approx 0,72$





$E(\sigma^2|Y_{1:T}) \approx 1,89$



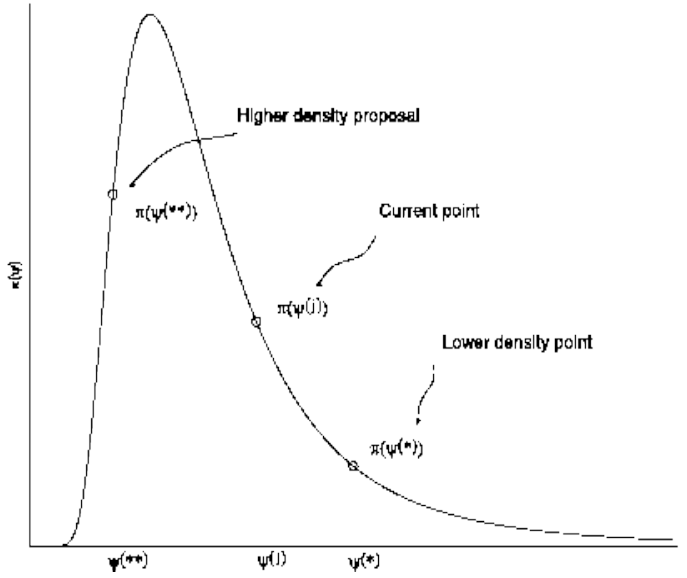
Metropolis-Hastings

- Conditional posterior distributions : too restrictive

- Metropolis-Hastings :

  1- Draw a proposal parameter from any chosen distribution

  2- Accept or reject the draw according to a probability which ensures that the invariant distribution of the MC is the posterior distribution of interest.



Metropolis-Hastings

- Let q be the proposal distribution (e.g. Normal)
- How to determine the probability function ?

Sufficient condition : $\pi(\theta|Y_{1:T})K(\theta^*|\theta) = \pi(\theta^*|Y_{1:T})K(\theta|\theta^*)$

Let assume that $\pi(\theta|Y_{1:T})q(\theta^*|\theta) > \pi(\theta^*|Y_{1:T})q(\theta|\theta^*)$

Move from $\theta \rightarrow \theta^*$ too often

Move from $\theta^* \rightarrow \theta$ too rarely

$$\pi(\theta|Y_{1:T})q(\theta^*|\theta)\alpha(\theta,\theta^*) = \pi(\theta^*|Y_{1:T})q(\theta|\theta^*)$$

$$\alpha(\theta^*,\theta) = \min\left[\frac{\pi(\theta^*|Y_{1:T})q(\theta|\theta^*)}{\pi(\theta|Y_{1:T})q(\theta^*|\theta)}, 1\right]$$

Since it is a probability

Metropolis-Hastings

- Initialize the MCMC
- At each MCMC iteration :
  - Generate a candidate from the proposal distribution

$$\theta^* \sim q(-|\theta)$$

  - Accept or reject the draw according to the probability

$$\phi(\theta^*, \theta) = \min[\frac{\pi(\theta^*|Y_{1:T})q(\theta|\theta^*)}{\pi(\theta|Y_{1:T})q(\theta^*|\theta)}, 1]$$

No need of the normalizing constant !

$$\phi(\theta^*, \theta) = \min[\frac{f(|Y_{1:T}|\theta^*)f(\theta^*)q(\theta|\theta^*)}{f(Y_{1:T}|\theta)f(\theta)q(\theta^*|\theta)}, 1]$$

M-H : Comments

- If q is symmetric : Metropolis algorithm

$$q(\theta|\theta^*) = q(\theta^*|\theta) \quad \text{then} \quad \alpha(\theta^*, \theta) = \min[\frac{f(|Y_{1:T}|\theta^*)f(\theta^*)}{f(Y_{1:T}|\theta)f(\theta)}, 1]$$

- Random Walk Metropolis : $q(\theta^*|\theta) \sim N(\theta, \bar{\Sigma})$

- Independent M-H : $q(\theta^*|\theta) \equiv q(\theta^*) \sim N(\bar{\theta}, \bar{\Sigma})$

Caution : proposal dist : thicker tail than post dist.

- Gibbs sampler : $q(\theta^*|\theta) = \pi(\theta^*|Y_{1:T})$

$$\phi(\theta^*, \theta) = \min[\frac{\pi(\theta^*|Y_{1:T})\pi(\theta|Y_{1:T})}{\pi(\theta|Y_{1:T})\pi(\theta^*|Y_{1:T})}, 1] = 1 \quad \forall\ \theta^*$$

M-H can also be applied to multiple blocks

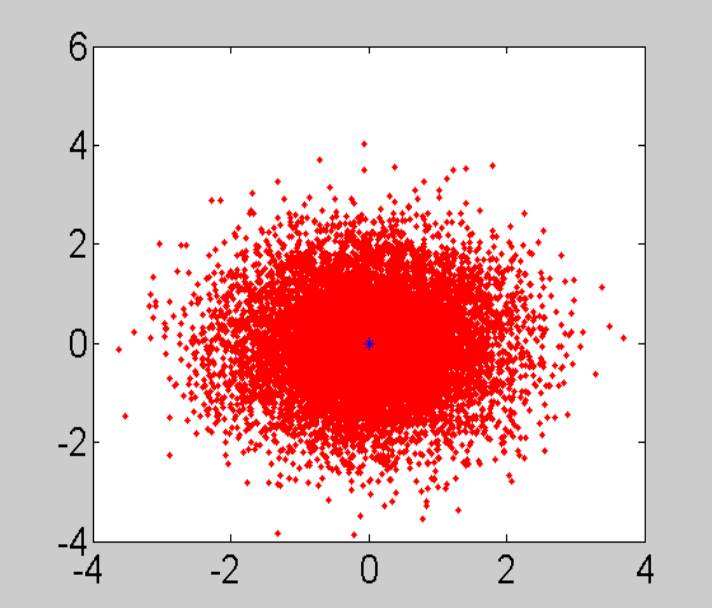
M-H : Comments

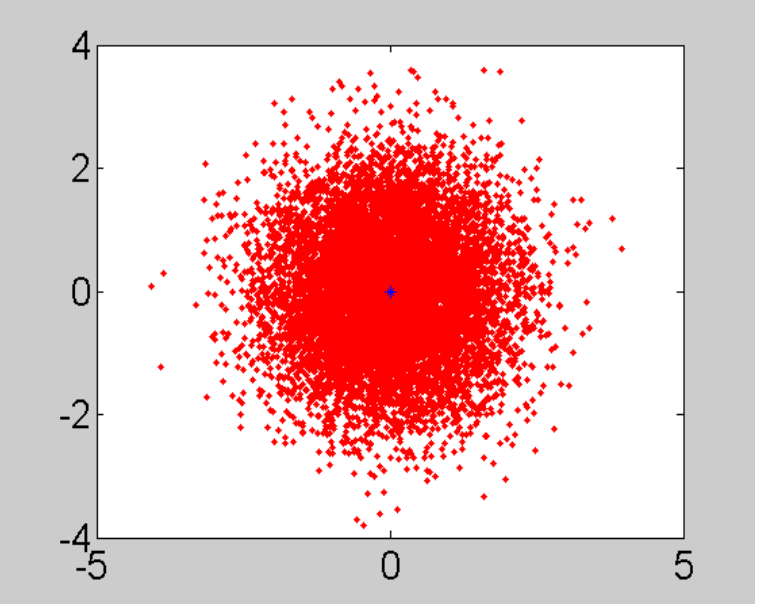
- Most frequent : Random Walk Metropolis - $q(\theta^*|\theta) \sim N(\theta, \bar{\Sigma})$
- How to choose the variance parameter ?

Hot topic in the literature

If $\bar{\Sigma}$ too small ⟶ Too many similar parameter values : Slow exploration of the support

If $\bar{\Sigma}$ too big ⟶ Too many rejected values : Slow exploration of the support





M-H : Comments

- Most frequent : Random Walk Metropolis - $q(\theta^*|\theta) \sim N(\theta, \bar{\Sigma})$
- How to choose the variance parameter ?

Hot topic in the literature

If Posterior distribution = Multivariate Normal distribution

Optimal acceptance rate

Dimension 1	Dimension 5	...	Large Dimension
Accept. Rate = 44 %	Accept. Rate = 28 %		Accept. Rate = 23,4 %

Reference : Roberts, G. O. & Rosenthal, J. S. *'Optimal scaling for various Metropolis-Hastings algorithms'*, Statistical Science, 2001, 16, 351-367

M-H : Comments

- Most frequent : Random Walk Metropolis - $q(\theta^*|\theta) \sim N(\theta, \bar{\Sigma})$
- How to choose the variance parameter ?

Common practice : Find the Optimal acceptance rate

- By trials and errors → Very demanding, Model dependent
- By adaptive RW metropolis

Reference : Atchadé, Y. & Rosenthal, J. *'On adaptive Markov chain Monte Carlo algorithms'*, Bernoulli, 2005, 11(5), 815-828

- At MCMC iteration *i*

$$\bar{\Sigma}_i = \bar{\Sigma}_{i-1} + (\phi_r^{i-1} - \phi_{target})/(i^c) \quad \text{if} \quad \Sigma_{Low} < \bar{\Sigma}_i < \Sigma_{High}$$

With ϕ_r^{i-1} the acceptance rate at iteration *i-1*

ϕ_{target} the targeted acceptance rate

$0.5 < c < 1$

MH sampler : Example

- GARCH process : $Y_{1:T} = \{y_1, ..., y_T\}'$

$$\begin{aligned} y_t &= \epsilon_t \\ \sigma_t^2 &= \omega + \alpha\epsilon_{t-1} + \beta\sigma_{t-1}^2 \\ \epsilon_t | Y_{1:t-1} &\sim_{i.i.d.} N(0, \sigma_t^2) \end{aligned}$$

Prior distributions

$$\omega \sim U(0, 10) \mid \alpha \sim U(0, 1)$$
$$\beta | \alpha \sim U(0, 1 - \alpha)$$

- Adaptive RW Metropolis with block of one dimension

DGP of the simulated time series

$$T = 3000 \qquad \omega = 0.1 \qquad \alpha = 0.15 \qquad \beta = 0.8$$

Choice of the MCMC parameters

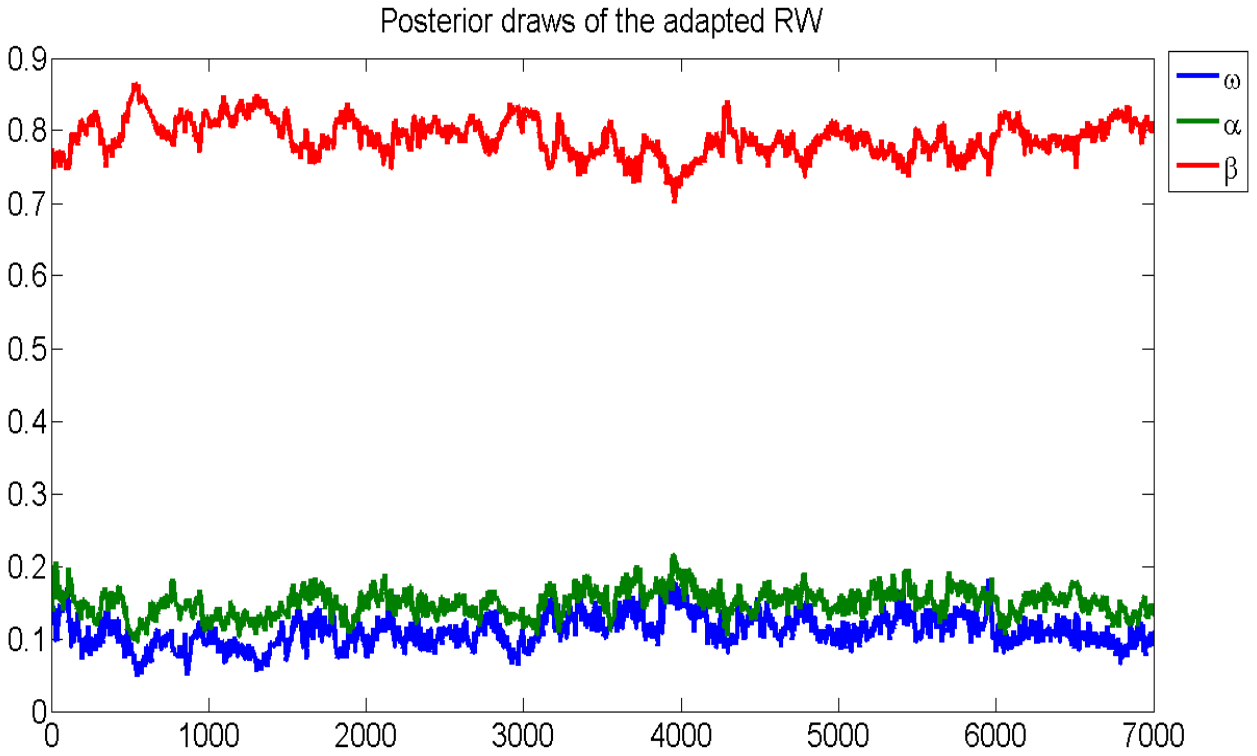
$$\phi_{target} = 0.44 \qquad c = 0.6 \qquad \Sigma_{Low} = 1e-5 \qquad \Sigma_{High} = 10$$

$$N = 10000 \qquad \text{Burn-in} = 3000$$

MH sampler : Example

- GARCH process :

$$T = 3000 \qquad \omega = 0.1 \qquad \alpha = 0.15 \qquad \beta = 0.8$$



- Acc. Rate :

$$\phi_\omega = \phi_\alpha = \phi_\beta = 0,44$$

- Post means :

$$E(\omega|Y_{1:T}) \approx 0.11$$

$$E(\alpha|Y_{1:T}) \approx 0.15$$

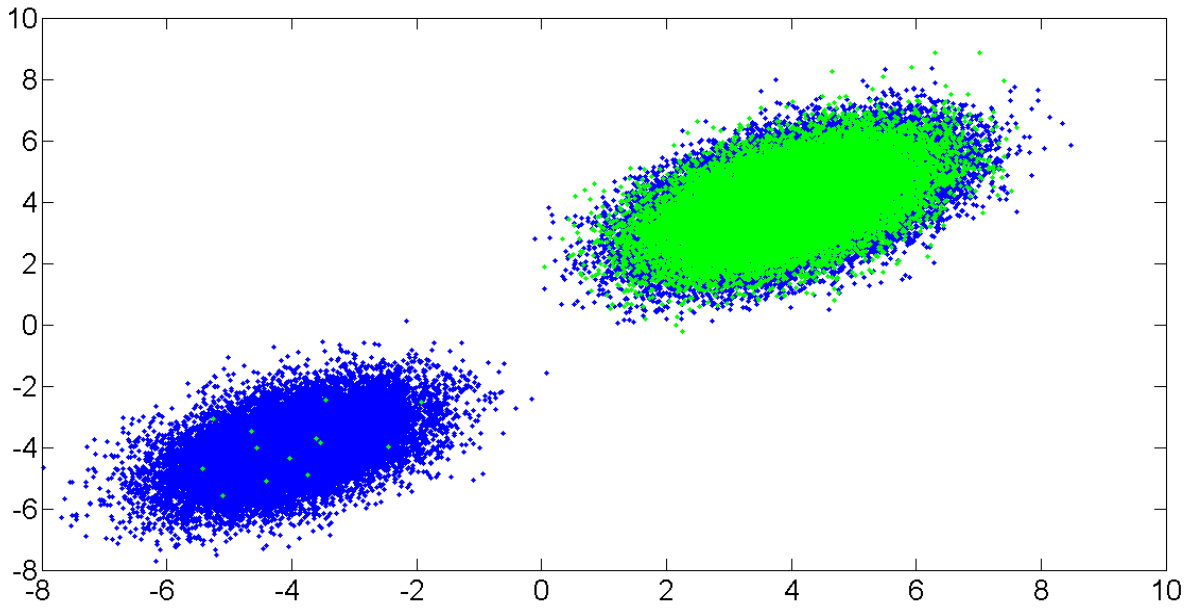
$$E(\beta|Y_{1:T}) \approx 0.79$$

MCMC : limitations

- Difficult to assess if the MC has converged to the post. Dist.
- Difficult to assess how much MCMC draws are required.
- Difficult to infer on multi-modal posterior distributions :

RW Metropolis : Jump from the current MCMC parameter

→ Unlikely to jump from one mode to another.



Questions ?

Model selection

Criteria

- Two popular Criteria :

Marginal likelihood

Intuitive criterion derived from Bayes' rule

Deviance Information Criterion (DIC)

Perfect when Marginal likelihood is out of reach

Spiegelhalter, D.; Best, D.; Carlin, B. & van der Linde, A. *'Bayesian measures of model complexity and fit'*, Journal of Royal Statistical Society, Series B, 2002, 64, 583-639

→ Focus on Marginal likelihood

Marginal likelihood

- Marginal likelihood = Normalizing constant :

How to Choose between two models : M_1 or M_2 ?

Bayes' theorem

$$\pi(M_1|Y_{1:T}) = \frac{f(Y_{1:T}|M_1)f(M_1)}{f(Y_{1:T}|M_1)f(M_1) + f(Y_{1:T}|M_2)f(M_2)}$$

$$\text{Choose } M_1 \text{ if } \pi(M_1|Y_{1:T}) > \pi(M_2|Y_{1:T})$$

If no subjective idea over the two models : $f(M_1) = f(M_2) = 0.5$

$$\pi(M_1|Y_{1:T}) > \pi(M_2|Y_{1:T}) \iff f(Y_{1:T}|M_1) > f(Y_{1:T}|M_2)$$

Marginal likelihood

- Multiple Models :

Bayes' theorem

$$\pi(M_j|Y_{1:T}) = \frac{f(Y_{1:T}|M_j)f(M_j)}{\sum_{i=1}^{k} f(Y_{1:T}|M_i)f(M_i)}$$

$$\text{Choose } M_j \text{ that maximizes } \pi(M_j|Y_{1:T}) \quad \forall\ j \in [1,k]$$

If no subjective idea over the different models : $f(M_j) = \frac{1}{k}$

$$\text{Choose } M_j \text{ that maximizes } f(Y_{1:T}|M_j) \quad \forall\ j \in [1,k]$$

Bayesian Model Averaging (BMA)

- Multiple Models :

Bayes' theorem

$$\pi(M_j|Y_{1:T}) = \frac{f(Y_{1:T}|M_j)f(M_j)}{\sum_{i=1}^{k} f(Y_{1:T}|M_i)f(M_i)}$$

Instead of choosing one model, **keep them all**

⟶ Take into account the model uncertainty

Predictive density : $\pi(y_{T+1}|Y_{1:T}) = \sum_{j=1}^{k} \pi(y_{T+1}|Y_{1:T}, M_j)\pi(M_j|Y_{1:T})$

Point forecast : $E(y_{T+1}|Y_{1:T}) = \sum_{j=1}^{k} [\int y_{T+1}\pi(y_{T+1}|Y_{1:T}, M_j)dy_{T+1}]\pi(M_j|Y_{1:T})$

Marginal likelihood

- Quantity of interest : $f(Y_{1:T}|M_j)$
- Local Formula :

Likelihood Prior

$$f(Y_{1:T}|M_j) = \frac{f(Y_{1:T}|\Theta^*)f(\Theta^*)}{\pi(\Theta^*|Y_{1:T})} \quad \forall\ \Theta^* \text{such that } \pi(\Theta^*|Y_{1:T}) > 0$$

Posterior

- Ockham's razor :

Likelihood	increases as long as the model complexity grows
Prior and Posterior	penalize the model complexity

Marginal likelihood

- Quantity of interest : $f(Y_{1:T}|M_j)$

$$f(Y_{1:T}|M_j) = \int_{\Omega} f(Y_{1:T}|M_j, \Theta) f(\Theta|M_j) d\Theta$$

If complex model with many parameters :

⟶ Highly dimensional integration : difficult to compute.

Dimension < 3	Middle Dimension	High Dimension
• Numerical Integration • Importance sampling	• Bridge sampling	• Path sampling • SMC sampler • MCMC • Variational Bayes

Marginal likelihood by MCMC

- Quantity of interest : $f(Y_{1:T}|M_j)$

$$f(Y_{1:T}|M_j) = \int_{\Omega} f(Y_{1:T}|M_j, \Theta) f(\Theta|M_j) d\Theta$$

- Local Formula :

Likelihood Prior

$$f(Y_{1:T}|M_j) = \frac{f(Y_{1:T}|\Theta^*) f(\Theta^*)}{\pi(\Theta^*|Y_{1:T})} \quad \forall\ \Theta^* \text{such that } \pi(\Theta^*|Y_{1:T}) > 0$$

Posterior

The posterior density is the most tricky part.

Marginal likelihood by MCMC

- Local Formula :

Likelihood Prior

$$f(Y_{1:T}|M_j) = \frac{f(Y_{1:T}|\Theta^*)f(\Theta^*)}{\pi(\Theta^*|Y_{1:T})} \quad \forall\ \Theta^* \text{such that } \pi(\Theta^*|Y_{1:T}) > 0$$

Posterior

- Marginal likelihood from Gibbs sampler :

Chib, S. *'Marginal Likelihood from the Gibbs Output'*, Journal of the American Statistical Association, 1995, 90, 1313-1321

- Marginal likelihood from MH sampler :

Chib, S. & Jeliazkov, I. *'Marginal Likelihood from the Metropolis-Hastings Output'*, Journal of the American Statistical Association, 2001, 96, 270-281

Marginal likelihood : Example

- ML for AR processes : $y_t|Y_{1:t-1} \sim N(\theta' x_t, \sigma^2)$

MCMC iterations

$$\begin{cases} \pi(\theta|Y_{1:T}, \sigma^2) \sim N(\bar{\mu}, \bar{\Sigma}) \\ \pi(\sigma^2|Y_{1:T}, \theta) \sim IG(\alpha + T/2, \beta + \sum_{t=1}^{T} \epsilon_t^2/2) \end{cases}$$

Choose a high density point : $\{\theta^*, \sigma^{2*}\}$

1) **Likelihood :** $f(Y_{1:T}|\theta^*, \sigma^{2*}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^{2*}}} e^{-0,5(y_t - \theta'^* x_t)^2/\sigma^{2*}}$

2) **Prior :** $f(\theta^*) \sim N(\mu_0, \Sigma_0)$ $\quad$ $f(\sigma^{2*}) \sim IG(\alpha, \beta)$

3) **Posterior :** $\pi(\theta^*, \sigma^{2*}|Y_{1:T}) = \pi(\sigma^{2*}|Y_{1:T})\pi(\theta^*|Y_{1:T}, \sigma^{2*})$

NB : $\pi(\sigma^{2*}|Y_{1:T}) = \int \pi(\sigma^{2*}|Y_{1:T}, \theta)\pi(\theta|Y_{1:T})d\theta \approx \frac{1}{N}\sum_{i=1}^{N} \pi(\sigma^{2*}|Y_{1:T}, \theta^i)$

Marginal likelihood : Example

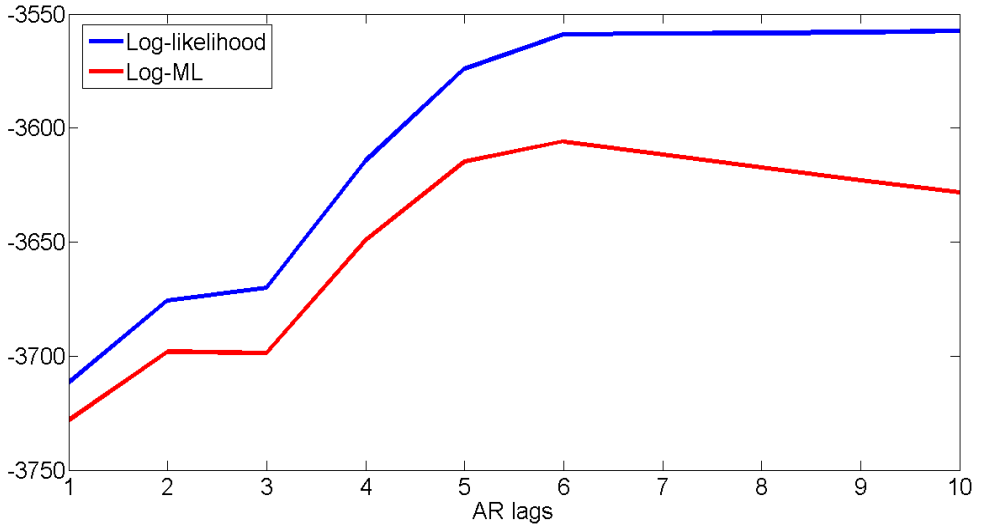
- ML for AR processes : $y_t | Y_{1:t-1} \sim N(\theta' x_t, \sigma^2)$

DGP of the simulated time series

$$T = 2000 \qquad \text{AR lags} = 6$$

- Log Marginal Likelihood :

1	2	3	4	5	6	7	8	9	10
-3728	-3698	-3697	-3649	-3614	-3606	-3611	-3617	-3623	-3628



Questions ?