Asymptotic behavior of bifurcating autoregressive processes

B. Bercu B. de Saporta A. Gégout-Petit

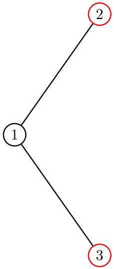
Université de Bordeaux

Mathematical models for cell division
IHP – 3 mars 2009

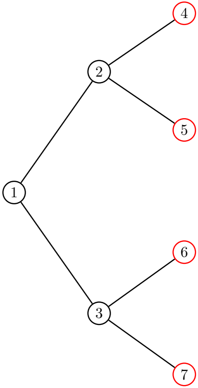
Outline

Cell division

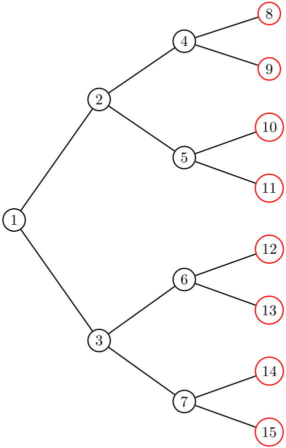
Cell division



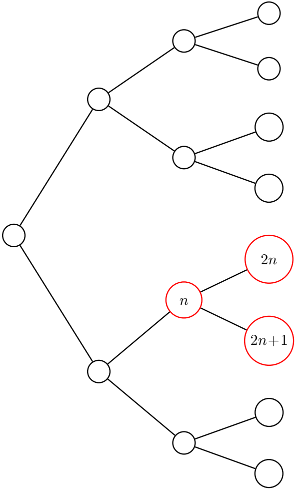
Cell division



Cell division



Offspring



Daughters of cell n:

- $2n$
- $2n+1$

Quantitative characteristic of cell n (diameter, concentration of some molecule, ...)

- X_n

Mathematical definition

Bifurcating auto regressive process BAR

$$\begin{cases} X_{2n} = a + b\,X_n + \varepsilon_{2n}, \\ X_{2n+1} = c + d\,X_n + \varepsilon_{2n+1}. \end{cases}$$

- X_1 initial cell
- $(\varepsilon_{2n}, \varepsilon_{2n+1})$ noise

Aim

Statistical inference on parameters a, b, c and d

Stationary BAR

- Cowan,Staudte, Biometrics, 1986 : Introduction, biological motivation
- Huggins, Annals of Statistic, 1996 : MLE for large trees, order 1
- Huggins, Basawa, Journal of Applied Probability, 1999 and Australian Journal of Statistics, 2000 : MLE for large trees, order > 1
- Basawa, Zhou, Journal of Applied Probability, 2004 : BAR with exponentially distributed noise, and Journal of Time Series Analysis, 2005 : CLT for BAR

Main assumptions

iid noise, symmetry

Non stationary BAR

- Guyon, Annals of Applied Probability, 2007 : Least-square estimation for gaussian BAR using Markov chains
- Delmas, Marsalle, 2008 : previous talk!

Main assumptions

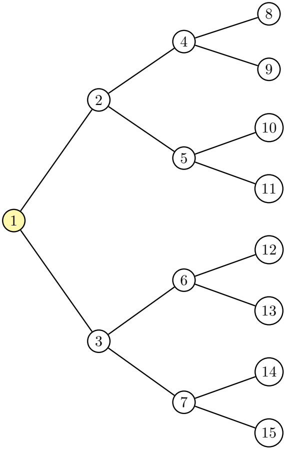
iid (gaussian) noise, no symmetry

New results

Our work

- noise: martingale difference
- rates of convergence
- martingales techniques

Generation

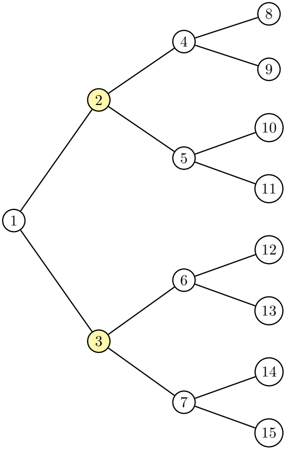


- Generation 0 : $\mathbb{G}_0 = \{1\}$
- Generation 1 : $\mathbb{G}_1 = \{2, 3\}$
- Generation 2 : $\mathbb{G}_2 = \{4, 5, 6, 7\}$
- Generation n : $\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$

Tree up to generation n: $\mathbb{T}_n = \bigcup_{k=0}^n \mathbb{G}_k$

Size of generation n: $|\mathbb{G}_n| = 2^n$
Size of the tree up to generation n: $|\mathbb{T}_n| = 2^{n+1} - 1$

Generation

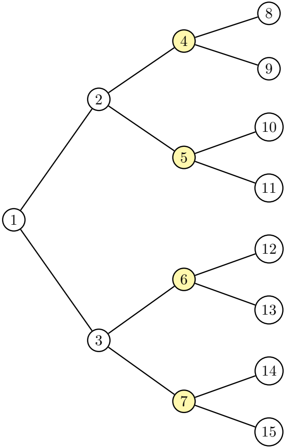


- Generation 0 : $\mathbb{G}_0 = \{1\}$
- Generation 1 : $\mathbb{G}_1 = \{2, 3\}$
- Generation 2 : $\mathbb{G}_2 = \{4, 5, 6, 7\}$
- Generation n : $\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$

Tree up to generation n: $\mathbb{T}_n = \bigcup_{k=0}^{n} \mathbb{G}_k$

Size of generation n: $|\mathbb{G}_n| = 2^n$
Size of the tree up to generation n: $|\mathbb{T}_n| = 2^{n+1} - 1$

Generation

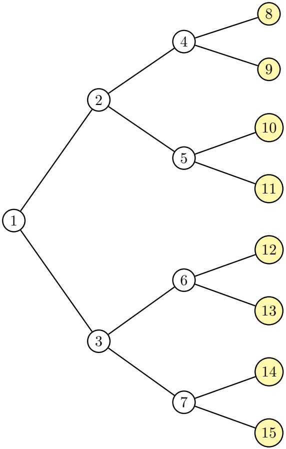


- Generation 0 : $\mathbb{G}_0 = \{1\}$
- Generation 1 : $\mathbb{G}_1 = \{2, 3\}$
- Generation 2 : $\mathbb{G}_2 = \{4, 5, 6, 7\}$
- Generation n : $\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$

Tree up to generation n:
$\mathbb{T}_n = \bigcup_{k=0}^{n} \mathbb{G}_k$

Size of generation n: $|\mathbb{G}_n| = 2^n$
Size of the tree up to generation n:
$|\mathbb{T}_n| = 2^{n+1} - 1$

Generation

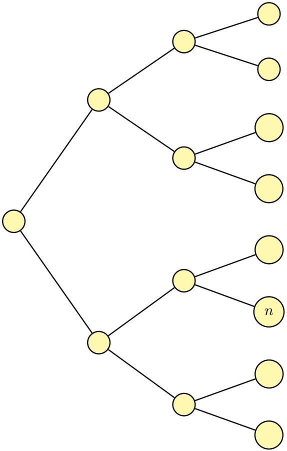


- Generation 0 : $\mathbb{G}_0 = \{1\}$
- Generation 1 : $\mathbb{G}_1 = \{2, 3\}$
- Generation 2 : $\mathbb{G}_2 = \{4, 5, 6, 7\}$
- Generation n :
  $\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$

Tree up to generation n:
$\mathbb{T}_n = \bigcup_{k=0}^{n} \mathbb{G}_k$

Size of generation n: $|\mathbb{G}_n| = 2^n$
Size of the tree up to generation n:
$|\mathbb{T}_n| = 2^{n+1} - 1$

Generation

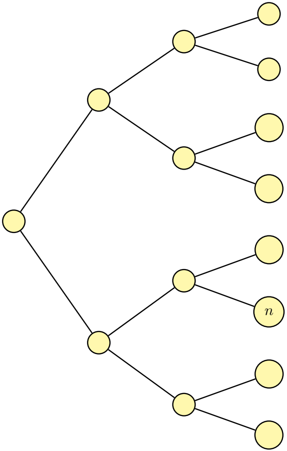


- Generation 0 : $\mathbb{G}_0 = \{1\}$
- Generation 1 : $\mathbb{G}_1 = \{2, 3\}$
- Generation 2 : $\mathbb{G}_2 = \{4, 5, 6, 7\}$
- Generation n :
  $\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$

Tree up to generation n:
$\mathbb{T}_n = \bigcup_{k=0}^{n} \mathbb{G}_k$

Size of generation n: $|\mathbb{G}_n| = 2^n$
Size of the tree up to generation n:
$|\mathbb{T}_n| = 2^{n+1} - 1$

Generation

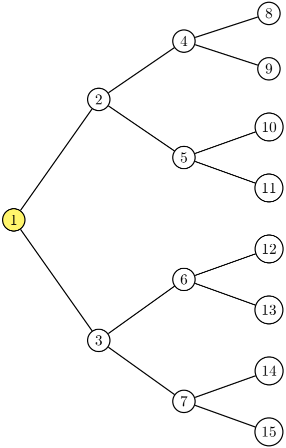


- Generation 0 : $\mathbb{G}_0 = \{1\}$
- Generation 1 : $\mathbb{G}_1 = \{2, 3\}$
- Generation 2 : $\mathbb{G}_2 = \{4, 5, 6, 7\}$
- Generation n : $\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$

Tree up to generation n:
$\mathbb{T}_n = \bigcup_{k=0}^{n} \mathbb{G}_k$

Size of generation n: $|\mathbb{G}_n| = 2^n$
Size of the tree up to generation n:
$|\mathbb{T}_n| = 2^{n+1} - 1$

Filtration

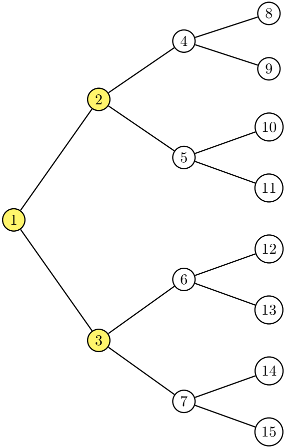


Definition

$$\mathcal{F}_0 = \sigma\{X_1\}$$

Information grows exponentially fast :
$2\times$ more cells in each generation

Filtration

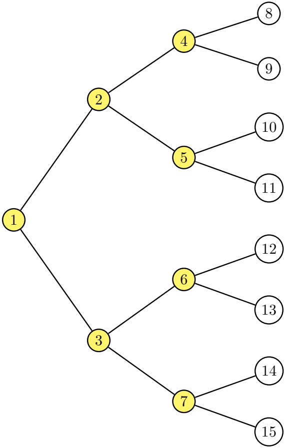


Definition

$$\mathcal{F}_1 = \sigma\{X_1, X_2, X_3\}$$

Information grows exponentially fast :
$2\times$ more cells in each generation

Filtration

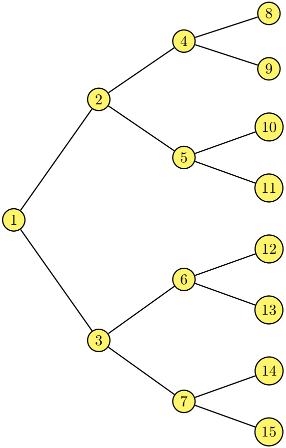


Definition

$$\mathcal{F}_2 = \sigma\{X_k \text{ with } k \in \mathbb{T}_2\}$$

Information grows exponentially fast :
$2\times$ more cells in each generation

Filtration

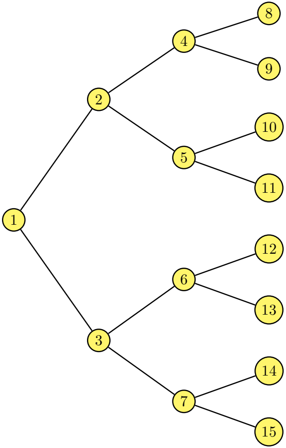


Definition

$$\mathcal{F}_n = \sigma\{X_k \text{ with } k \in \mathbb{T}_n\}$$

Information grows exponentially fast :
$2\times$ more cells in each generation

Filtration



Definition

$$\mathcal{F}_n = \sigma\{X_k \text{ with } k \in \mathbb{T}_n\}$$

Information grows exponentially fast :
2$\times$ more cells in each generation

Our model

BAR process

$$\begin{cases} X_{2n} = a + b\,X_n + \varepsilon_{2n}, \\ X_{2n+1} = c + d\,X_n + \varepsilon_{2n+1}. \end{cases}$$

Assumptions

- $\mathbb{E}[X_1^8] < \infty$
- $0 < \max(|b|, |d|) < 1$
- $|a| + |c| \neq 0$

Assumptions on the noise

(H.1) $\forall n \geq 0$ and $\forall k \in \mathbb{G}_{n+1}$,

$$\mathbb{E}[\varepsilon_k | \mathcal{F}_n] = 0 \qquad \text{and} \qquad \mathbb{E}[\varepsilon_k^2 | \mathcal{F}_n] = \sigma^2 > 0 \qquad \text{a.s.}$$

(H.2) $\forall n \geq 0$ and $\forall k \neq \ell \in \mathbb{G}_{n+1}$,

- if $[k/2] \neq [\ell/2]$, ε_k and ε_ℓ independent conditionally to $\mathcal{F}_n$
- if $[k/2] = [\ell/2]$, then for some $\rho < \sigma^2$

$$\mathbb{E}[\varepsilon_k \varepsilon_\ell | \mathcal{F}_n] = \rho \qquad \text{a.s.}$$

(H.3)

$$\sup_{n \geq 0} \sup_{k \in \mathbb{G}_{n+1}} \mathbb{E}[\varepsilon_k^4 | \mathcal{F}_n] < \infty \qquad \text{a.s.}$$

Assumptions on the noise

(H.1) $\forall n \geq 0$ and $\forall k \in \mathbb{G}_{n+1}$,

$$\mathbb{E}[\varepsilon_k | \mathcal{F}_n] = 0 \qquad \text{and} \qquad \mathbb{E}[\varepsilon_k^2 | \mathcal{F}_n] = \sigma^2 > 0 \qquad \text{a.s.}$$

(H.2) $\forall n \geq 0$ and $\forall k \neq \ell \in \mathbb{G}_{n+1}$,

- if $[k/2] \neq [\ell/2]$, ε_k and ε_ℓ independent conditionally to $\mathcal{F}_n$
- if $[k/2] = [\ell/2]$, then for some $\rho < \sigma^2$

$$\mathbb{E}[\varepsilon_k \varepsilon_\ell | \mathcal{F}_n] = \rho \qquad \text{a.s.}$$

(H.3)

$$\sup_{n \geq 0} \sup_{k \in \mathbb{G}_{n+1}} \mathbb{E}[\varepsilon_k^4 | \mathcal{F}_n] < \infty \qquad \text{a.s.}$$

Assumptions on the noise

(H.1) $\forall n \geq 0$ and $\forall k \in \mathbb{G}_{n+1}$,

$$\mathbb{E}[\varepsilon_k | \mathcal{F}_n] = 0 \qquad \text{and} \qquad \mathbb{E}[\varepsilon_k^2 | \mathcal{F}_n] = \sigma^2 > 0 \qquad \text{a.s.}$$

(H.2) $\forall n \geq 0$ and $\forall k \neq \ell \in \mathbb{G}_{n+1}$,

- if $[k/2] \neq [\ell/2]$, ε_k and ε_ℓ independent conditionally to $\mathcal{F}_n$
- if $[k/2] = [\ell/2]$, then for some $\rho < \sigma^2$

$$\mathbb{E}[\varepsilon_k \varepsilon_\ell | \mathcal{F}_n] = \rho \qquad \text{a.s.}$$

(H.3)

$$\sup_{n \geq 0} \sup_{k \in \mathbb{G}_{n+1}} \mathbb{E}[\varepsilon_k^4 | \mathcal{F}_n] < \infty \qquad \text{a.s.}$$

Least square estimators

Estimator of $\theta = (a, b, c, d)^t$

$$\widehat{\theta}_n = \begin{pmatrix} \widehat{a}_n \\ \widehat{b}_n \\ \widehat{c}_n \\ \widehat{d}_n \end{pmatrix} = (\mathrm{I}_2 \otimes S_{n-1}^{-1}) \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} X_{2k} \\ X_k X_{2k} \\ X_{2k+1} \\ X_k X_{2k+1} \end{pmatrix}$$

with

$$S_n = \sum_{k \in \mathbb{T}_n} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$$

Variance/covariance estimators

Estimator of conditional variance

$$\widehat{\sigma}_n^2 = \frac{1}{2|\mathbb{T}_{n-1}|} \sum_{k \in \mathbb{T}_{n-1}} (\widehat{\varepsilon}_{2k}^2 + \widehat{\varepsilon}_{2k+1}^2)$$

Estimator of conditional covariance

$$\widehat{\rho}_n = \frac{1}{|\mathbb{T}_{n-1}|} \sum_{k \in \mathbb{T}_{n-1}} \widehat{\varepsilon}_{2k} \widehat{\varepsilon}_{2k+1}$$

with, for all $k \in \mathbb{G}_n$

$$\begin{cases} \widehat{\varepsilon}_{2k} = X_{2k} - \widehat{a}_n - \widehat{b}_n X_k \\ \widehat{\varepsilon}_{2k+1} = X_{2k+1} - \widehat{c}_n - \widehat{d}_n X_k \end{cases}$$

Martingales

Definition

(M_n) sequence of square integrable random variables adapted to the filtration $\mathcal{F}$ is a martingale if

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n$$

Examples:

$$\sum_{k \in \mathbb{T}_{n-1}} \varepsilon_{2k}, \quad \sum_{k \in \mathbb{T}_{n-1}} \varepsilon_{2k+1}, \quad \sum_{k \in \mathbb{T}_{n-1}} X_k \varepsilon_{2k}, \quad \sum_{k \in \mathbb{T}_{n-1}} X_k \varepsilon_{2k+1},$$

$$\sum_{k \in \mathbb{T}_n} (\varepsilon_k^2 - \sigma^2), \quad \sum_{k \in \mathbb{T}_{n-1}} (\varepsilon_{2k}\varepsilon_{2k+1} - \rho), \ldots$$

Main martingale

Estimator and martingale

$$(\widehat{\theta}_n - \theta) = (\mathrm{I}_2 \otimes S_{n-1}^{-1}) \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} \varepsilon_{2k} \\ X_k \varepsilon_{2k} \\ \varepsilon_{2k+1} \\ X_k \varepsilon_{2k+1} \end{pmatrix} = \Sigma_{n-1}^{-1} M_n$$

with $\Sigma_n = \mathrm{I}_2 \otimes S_n$ and M_n $\mathcal{F}$-martingale

$$M_n = \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} \varepsilon_{2k} \\ X_k \varepsilon_{2k} \\ \varepsilon_{2k+1} \\ X_k \varepsilon_{2k+1} \end{pmatrix}$$

Martingale convergence results

(M_n) scalar $\mathcal{F}$-martingale bounded in L^2

$\Delta M_{n+1} = M_{n+1} - M_n$

Increasing process $< M >_n = \sum_{k=0}^{n} \mathbb{E}[(\Delta M_{n+1})^2 \mid \mathcal{F}_n]$

Convergence of L^2 martingales

If $\lim_{n\to\infty} < M >_n = +\infty$, then $\frac{M_n}{<M>_n} \to 0$ a.s.

+ moment conditions then $\left(\frac{M_n}{<M>_n}\right)^2 = \mathcal{O}(\frac{log(<M>_n)}{<M>_n})$ a.s.

Similar results for vector-valued martingales.

Here $< M >_n = I_2 \otimes S_n$

LLN for the noise

Laws of large numbers

If (H.1), (H.2) and (H.3) hold

$$\lim_{n\to+\infty} \frac{1}{|\mathbb{T}_n|} \sum_{k\in\mathbb{T}_n} \varepsilon_k = 0 \qquad \text{a.s.}$$

$$\lim_{n\to+\infty} \frac{1}{|\mathbb{T}_n|} \sum_{k\in\mathbb{T}_n} \varepsilon_k^2 = \sigma^2 \qquad \text{a.s.}$$

$$\lim_{n\to+\infty} \frac{1}{|\mathbb{T}_{n-1}|} \sum_{k\in\mathbb{T}_{n-1}} \varepsilon_{2k}\varepsilon_{2k+1} = \rho \qquad \text{a.s.}$$

Direct application of LLN for scalar martingales

Asymptotic behavior of the increasing process

$$S_n = \sum_{k \in \mathbb{T}_n} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$$

Limit of S_n

If (H.1), (H.2) and (H.3) hold

$$\lim_{n \to \infty} \frac{S_n}{|\mathbb{T}_n|} = L \qquad \text{a.s.}$$

where L symmetric positive definite explicit matrix

LLN for the noise and recurrence equation defining the BAR process

Law of large numbers for θ

Theorem

If (H.1), (H.2) and (H.3) hold
$\widehat{\theta}_n$ converges a.s. to θ with rate

$$\| \widehat{\theta}_n - \theta \|^2 = \mathcal{O}\left(\frac{\log |\mathbb{T}_{n-1}|}{|\mathbb{T}_{n-1}|}\right) \qquad \text{a.s.}$$

Quadratic strong law

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} |\mathbb{T}_{k-1}|(\widehat{\theta}_k - \theta)^t \Lambda(\widehat{\theta}_k - \theta) = 4\sigma^2 \qquad \text{a.s.}$$

with $\Lambda = \mathrm{I}_2 \otimes L$

Vector-valued martingales

$$(\widehat{\theta}_n - \theta) = \Sigma_{n-1}^{-1} M_n$$

LLN for vector-valued martingales of the form $M_n = \sum_{k=1}^{n} \Psi_{k-1}\xi_k$, with fixed-sized (Ψ_k) and (ξ_k)

Problem

here the size of (Ψ_k) and (ξ_k) doubles at each generation

Solution

rewrite the LLN proof in this case

Vector-valued martingales

$$(\widehat{\theta}_n - \theta) = \Sigma_{n-1}^{-1} M_n$$

LLN for vector-valued martingales of the form
$M_n = \sum_{k=1}^{n} \Psi_{k-1}\xi_k$, with fixed-sized (Ψ_k) and (ξ_k)

Problem

here the size of (Ψ_k) and (ξ_k) doubles at each generation

Solution

rewrite the LLN proof in this case

LLN for conditional variance

$$\sigma_n^2 = \frac{1}{2|\mathbb{T}_{n-1}|} \sum_{k \in \mathbb{T}_{n-1}} (\varepsilon_{2k}^2 + \varepsilon_{2k+1}^2)$$

Theorem

If (H.1), (H.2) and (H.3) hold, $\widehat{\sigma}_n^2$ converges a.s. to σ^2

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k \in \mathbb{T}_{n-1}} (\widehat{\varepsilon}_{2k} - \varepsilon_{2k})^2 + (\widehat{\varepsilon}_{2k+1} - \varepsilon_{2k+1})^2 = 4\sigma^2 \quad \text{a.s.}$$

$$\lim_{n\to\infty} \frac{|\mathbb{T}_n|}{n} (\widehat{\sigma}_n^2 - \sigma_n^2) = 4\sigma^2 \quad \text{a.s.}$$

LLN for conditional covariance

$$\rho_n = \frac{1}{|\mathbb{T}_{n-1}|} \sum_{k \in \mathbb{T}_{n-1}} \varepsilon_{2k}\varepsilon_{2k+1}$$

Theorem

If (H.1), (H.2) and (H.3) hold, $\widehat{\rho}_n$ converges a.s. to ρ

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k \in \mathbb{T}_{n-1}} (\widehat{\varepsilon}_{2k} - \varepsilon_{2k})(\widehat{\varepsilon}_{2k+1} - \varepsilon_{2k+1}) = 2\rho \qquad \text{a.s.}$$

$$\lim_{n\to\infty} \frac{|\mathbb{T}_n|}{n} (\widehat{\rho}_n - \rho_n) = 4\rho \qquad \text{a.s.}$$

Additional assumptions

(H.4) $\forall n \geq 0$ and $\forall k \in \mathbb{G}_{n+1}$,

$$\mathbb{E}[\varepsilon_k^4 | \mathcal{F}_n] = \tau^4 \qquad \text{a.s.}$$

and $\forall k \neq \ell \in \mathbb{G}_{n+1}$ with $[k/2]=[\ell/2]$ et pour $\nu^2 < \tau^4$

$$\mathbb{E}[\varepsilon_k^2 \varepsilon_\ell^2 | \mathcal{F}_n] = \nu^2 \qquad \text{a.s.}$$

(H.5)

$$\sup_{n \geq 0} \sup_{k \in \mathbb{G}_{n+1}} \mathbb{E}[\varepsilon_k^8 | \mathcal{F}_n] < \infty \qquad \text{a.s.}$$

CLT

Theorem

If (H.1) – (H.5) hold

$$\sqrt{|\mathbb{T}_{n-1}|}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma \otimes L^{-1})$$

$$\sqrt{|\mathbb{T}_{n-1}|}(\widehat{\sigma}_n^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{\tau^4 - 2\sigma^4 + \nu^2}{2}\Big)$$

$$\sqrt{|\mathbb{T}_{n-1}|}(\widehat{\rho}_n - \rho) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nu^2 - \rho^2)$$

Sketch of the proof

- Lindeberg conditions does not hold
- CLT for martingale differences not valid for filtration $\mathcal{F}$

$\Longrightarrow$ New pair-wise filtration $\mathcal{G}$
CLT for martingale differences valid

Sketch of the proof

- Lindeberg conditions does not hold
- CLT for martingale differences not valid for filtration $\mathcal{F}$

$\Longrightarrow$ New pair-wise filtration $\mathcal{G}$
CLT for martingale differences valid

Sketch of the proof

- Lindeberg conditions does not hold
- CLT for martingale differences not valid for filtration $\mathcal{F}$

$\Longrightarrow$ New pair-wise filtration $\mathcal{G}$
CLT for martingale differences valid

Further work

Study BAR with random coefficients b and d

$$\begin{cases} X_{2n} &= a + b X_n + \varepsilon_{2n}, \\ X_{2n+1} &= c + d X_n + \varepsilon_{2n+1}. \end{cases}$$

Questions

Biological interpretation?

The end