Approximate Regenerative Block-Bootstrap for Markov Chains

–

Journées MAS 2008 - Rennes

Stéphan Clémençon

Telecom ParisTech
-
LTCI UMR CNRS/Telecom ParisTech No 5141
-
Institut Telecom

Joint work with Patrice Bertail (CREST-INSEE)

1. A little Markov chain theory
   - Markov with regeneration times
   - General Harris Markov chains
   - Probabilistic study based on renewal theory
   - Sharper results

2. Regeneration-based statistics
   - Asymptotic mean and variance estimation
   - Extension to general Harris chains
   - Regeneration-based U-statistics

3. Regenerative block-bootstrap
   - Bootstrap for dependent data
   - Algorithm/theory for the regenerative block-bootstrap
   - Simulation studies

Markov chains with (pseudo-) regeneration times

- **The Markov property:** $X = (X_n)_{n \in \mathbb{N}}$ is a chain with state space $(E, \mathcal{E})$, trans. prob. $\Pi(x, dy)$ and initial distr. ν iff $X_0 \sim \nu$ and

$$\mathbb{P}(X_{n+1} \in dx \mid X_0, \ldots , X_n) = \Pi(X_n, dx).$$

"Disc.-time proc. that forgets all about its past except its last value".

Markov chains with (pseudo-) regeneration times

- **The Markov property:** $X = (X_n)_{n\in\mathbb{N}}$ is a chain with state space $(E, \mathcal{E})$, trans. prob. $\Pi(x, dy)$ and initial distr. ν iff $X_0 \sim \nu$ and

$$\mathbb{P}(X_{n+1} \in dx \mid X_0, \ldots , X_n) = \Pi(X_n, dx).$$

"Disc.-time proc. that forgets all about its past except its last value".

- Widely used in appl. for model. random phenomena with causality.
- Ubiquity of Markov models: financial/econometric time-series, queuing/storage models, biological systems, epidemic models, *etc.*

Markov chains with (pseudo-) regeneration times

- **The Markov property:** $X = (X_n)_{n\in\mathbb{N}}$ is a chain with state space $(E, \mathcal{E})$, trans. prob. $\Pi(x, dy)$ and initial distr. ν iff $X_0 \sim \nu$ and

$$\mathbb{P}(X_{n+1} \in dx \mid X_0, \ldots, X_n) = \Pi(X_n, dx).$$

"Disc.-time proc. that forgets all about its past except its last value".

- Widely used in appl. for model. random phenomena with causality.
- Ubiquity of Markov models: financial/econometric time-series, queuing/storage models, biological systems, epidemic models, *etc.*
- Strong Markov property:
  - "Shift operator" θ: $X_{n+1} = X_n \circ \theta$,
  - $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$, τ stopp. time, $H = H(x_1, x_2, \ldots)$ bounded,

$$\mathbb{E}_\nu[H \circ \theta^\tau \mid \mathcal{F}_\tau] = \mathbb{E}_{X_\tau}[H] \text{ on } \{\tau < \infty\}.$$

A little Markov chain theory

- **Communication/ Stochastic stability**

- **Communication/ Stochastic stability**
  - Irreducibility: $\Phi/\ \forall B \in \mathcal{E}$,

$$\Phi(B) > 0 \Rightarrow \forall x, \sum_{n \geq 1} \Pi_n(x, B) > 0,$$

$\Rightarrow$ exist. of a maximal irreducibility measure Ψ.

A little Markov chain theory

- **Communication/ Stochastic stability**
  - Irreducibility: $\Phi/\ \forall B \in \mathcal{E}$,

$$\Phi(B) > 0 \Rightarrow \forall x,\ \sum_{n\geq 1} \Pi_n(x, B) > 0,$$

$\Rightarrow$ exist. of a maximal irreducibility measure Ψ.

  - Periodicity: X Ψ-irreduc., $\exists d' \geq 1$, $D_1, \ldots, D_{d'}$, $\Psi(D_i) > 0$, $D_i \cap D_j = \emptyset$ if $i \neq j$ and
    1. $\Pi(x, D_{i+1}) = 1$, $x \in D_i$,
    2. $\Psi(\cup\{D_i\}^c) = 0$,

$\Rightarrow$ period= $GCD\{d' \geq 1/\ 1.\ and\ 2.\}$, aperiodic when $= 1$.

A little Markov chain theory

- **Communication/ Stochastic stability**
  - Irreducibility: $\Phi/\ \forall B \in \mathcal{E}$,

$$\Phi(B) > 0 \Rightarrow \forall x,\ \sum_{n\geq 1} \Pi_n(x, B) > 0,$$

$\Rightarrow$ exist. of a maximal irreducibility measure Ψ.

  - Periodicity: X Ψ-irreduc., $\exists d' \geq 1,\ D_1, \dots, D_{d'},\ \Psi(D_i) > 0$, $D_i \cap D_j = \emptyset$ if $i \neq j$ and
    1. $\Pi(x, D_{i+1}) = 1,\ x \in D_i$,
    2. $\Psi(\cup\{D_i\}^c) = 0$,

  $\Rightarrow$ period$=$ $GCD\{d' \geq 1/\ 1.\ and\ 2.\}$, aperiodic when $= 1$.

  - Positive recurrence: $\exists!$ prob. $\mu/\ \mu(dy) = \int_{x\in E} \mu(dx)\Pi(x, dy)$ and

$$"\Pi_n(x, dy) = \mathbb{P}(X_n \in dy \mid X_0 = x) \to \mu(dy)" \text{ as } n \to \infty.$$

Regenerative Markov chains

- A meas. set $A \subset E$ is an accessible atom if $\Psi(A) > 0$ and

$$\forall (x, y) \in A^2, \ \ \Pi(x, .) = \Pi(y, .).$$

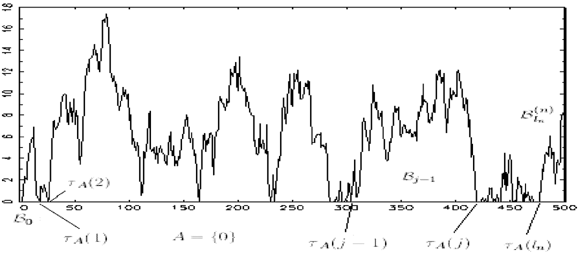
Regenerative Markov chains

- A meas. set $A \subset E$ is an accessible atom if $\Psi(A) > 0$ and

$$\forall (x, y) \in A^2, \quad \Pi(x, .) = \Pi(y, .).$$

- **Examples:** countable chains, queuing systems / storage models, *etc.*

Work-modulated single server queue with the empty file $\{0\}$ *as atom.*



Regenerative Markov chains

Regenerative Markov chains

- "Dividing sample paths into regeneration cycles"
  Supp. A is Harris rec. . Let $\mathbb{E}_A[.] = \mathbb{E}[. \mid X_0 \in A]$ and

$$\begin{aligned}\tau_A(1) &= \tau_A = \inf\{k \geq 1 : \ X_k \in A\}, \\ \tau_A(j) &= \inf\{k \geq 1 + \tau_A(j-1) : \ X_k \in A\}, \ j > 1.\end{aligned}$$

Regenerative Markov chains

- "Dividing sample paths into regeneration cycles"
  Supp. A is Harris rec. . Let $\mathbb{E}_A[.] = \mathbb{E}[. \mid X_0 \in A]$ and

$$\begin{aligned}\tau_A(1) &= \tau_A = \inf\{k \geq 1 : \; X_k \in A\},\\ \tau_A(j) &= \inf\{k \geq 1 + \tau_A(j-1) : \; X_k \in A\},\; j > 1.\end{aligned}$$

Str. Markov ppty $\Rightarrow$ $\{\tau_A(j)\}_{j\geq 1}$ (possibly delayed) renewal process

$$\ldots,\; \underbrace{X_{1+\tau_A(1)}, \ldots, X_{\tau_A(2)}}, \ldots,\; \underbrace{X_{1+\tau_A(j)},\; \ldots,\; X_{\tau_A(j+1)}}, \ldots$$

i.i.d. "regenerative blocks" of random size

$$\mathcal{B}_j = (X_{1+\tau_A(j)},\; \ldots,\; X_{\tau_A(j+1)}) \in \mathbb{T} = \cup_{n\geq 1} E^n.$$

General Harris Markov chains - Pseudo-renewal

- **Lower bounds for the transition probability**

General Harris Markov chains - Pseudo-renewal

- **Lower bounds for the transition probability**
  - A meas. set $S \subset E$ is small if $\exists m \geq 1$, $\delta > 0$, Φ proba. s.t. $\Phi(S) = 1$

$$\forall x \in E, \quad \Pi_m(x, dy) \geq \delta \mathbb{I}_{\{x \in S\}} \Phi(dy). \tag{1}$$

- **Lower bounds for the transition probability**
  - A meas. set $S \subset E$ is small if $\exists m \geq 1$, $\delta > 0$, Φ proba. s.t. $\Phi(S) = 1$

$$\forall x \in E, \quad \Pi_m(x, dy) \geq \delta \mathbb{I}_{\{x \in S\}} \Phi(dy). \tag{1}$$

  - If X Ψ-irred., then $\forall B / \ \Psi(B) > 0$, there exists $S \subset B$ small for X.

General Harris Markov chains - Pseudo-renewal

- **Lower bounds for the transition probability**
  - A meas. set $S \subset E$ is small if $\exists m \geq 1$, $\delta > 0$, Φ proba. s.t. $\Phi(S) = 1$

$$\forall x \in E, \quad \Pi_m(x, dy) \geq \delta \mathbb{I}_{\{x \in S\}} \Phi(dy). \tag{1}$$

  - If X Ψ-irred., then $\forall B / \ \Psi(B) > 0$, there exists $S \subset B$ small for X.
  - A 'small set' is not necessarily physically small !

General Harris Markov chains - Pseudo-renewal

- **Lower bounds for the transition probability**
  - A meas. set $S \subset E$ is small if $\exists m \geq 1$, $\delta > 0$, Φ proba. s.t. $\Phi(S) = 1$

$$\forall x \in E, \quad \Pi_m(x, dy) \geq \delta \mathbb{I}_{\{x \in S\}} \Phi(dy). \tag{1}$$

  - If X Ψ-irred., then $\forall B / \ \Psi(B) > 0$, there exists $S \subset B$ small for X.
  - A 'small set' is not necessarily physically small !

- **Example** - *Nonlinear AR model*:

$$X_{n+1} = m(X_n) + \sigma(X_n)\epsilon_{n+1},$$

$S = [x_0 - \eta, x_0 + \eta]$ with $m = 1$, $\Phi = \mathcal{U}([x_0 - \eta, x_0 + \eta])$,
$\delta = \inf_{(x,y)^2 \in [x_0 - \eta, x_0 + \eta]^2} \frac{1}{\sigma(x)} g_\epsilon(\frac{y - m(x)}{\sigma(x)})$.

General Harris Markov chains - Pseudo-renewal

General Harris Markov chains - Pseudo-renewal

- **The Nummelin technique** ($m = 1$). Construct a bivar. chain (X, Y) with state sp. $E \times \{0, 1\}$ by randomizing each time X_n hits S:
  - if $x \notin S$, $\Pi^*((x, y), B \times \{1\}) = \delta\Pi(x, B)$,
  - if $x \in S$, then

$$\Pi^*((x,0), B \times \{1\}) = \delta \frac{\Pi(x,B) - \delta\Phi(B)}{1-\delta},$$
$$\Pi^*((x,1), B \times \{1\}) = \delta\Phi(B),$$

$\Rightarrow S \times \{1\}$ is an atom for the split chain.

General Harris Markov chains - Pseudo-renewal

- **The Nummelin technique** ($m = 1$). Construct a bivar. chain (X, Y) with state sp. $E \times \{0, 1\}$ by randomizing each time X_n hits S:
  - if $x \notin S$, $\Pi^*((x, y), B \times \{1\}) = \delta \Pi(x, B)$,
  - if $x \in S$, then

$$\Pi^*((x,0), B \times \{1\}) = \delta \frac{\Pi(x, B) - \delta \Phi(B)}{1 - \delta},$$
$$\Pi^*((x,1), B \times \{1\}) = \delta \Phi(B),$$

  $\Rightarrow S \times \{1\}$ is an atom for the split chain.

- **Distribution of Y cond. on X.** Supp. $\Pi(x, dy) = \pi(x, y)\lambda(dy)$ and $\Phi(dy) = \phi(y)\lambda(dy)$. Cond. on X, draw indep^{tly} $Y_1, Y_2, \ldots$ such that

$$Y_i \sim Ber(\delta) \text{ if } X_i \notin S,$$
$$Y_i \sim Ber\left(\frac{\pi(X_i, X_{i+1}) - \delta \phi(X_{i+1})}{1 - \delta}\right) \text{ if } X_i \in S.$$

Stationary/asymptotic behavior

Stationary/asymptotic behavior

- Suppose X is Harris recurrent. Let A be an accessible atom.

The asymptotic behavior is ruled by the renewal process $\{\tau_A(j)\}_{j\geq 1}$.

Stationary/asymptotic behavior

- Suppose X is Harris recurrent. Let A be an accessible atom.

The asymptotic behavior is ruled by the renewal process $\{\tau_A(j)\}_{j\geq 1}$.

Theorem (Kac's theorem)

The chain X is positive recurrent iff $\alpha = \mathbb{E}_A[\tau_A] < \infty$.
If X is positive recurrent, then μ is the Pitman's occupation measure:

$$\mu(B) = \frac{1}{\mathbb{E}_A[\tau_A]}\mathbb{E}_A[\sum_{i=1}^{\tau_A}\mathbb{I}_{\{X_i\in B\}}].$$

Stationary/asymptotic behavior

- Suppose X is Harris recurrent. Let A be an accessible atom.

The asymptotic behavior is ruled by the renewal process $\{\tau_A(j)\}_{j\geq 1}$.

Theorem (Kac's theorem)

The chain X is positive recurrent iff $\alpha = \mathbb{E}_A[\tau_A] < \infty$.
If X is positive recurrent, then μ is the Pitman's occupation measure:

$$\mu(B) = \frac{1}{\mathbb{E}_A[\tau_A]}\mathbb{E}_A[\sum_{i=1}^{\tau_A} \mathbb{I}_{\{X_i \in B\}}].$$

- Let $f : E \to \mathbb{R}$ μ-integrable

$$\mu(f) = \int_{x\in E} f(x)\mu(dx) = \frac{1}{\mathbb{E}_A[\tau_A]}\mathbb{E}_A[\sum_{i=1}^{\tau_A} f(X_i)].$$

Extremal behavior

Extremal behavior

- Consider the cdf of the cycle submaximum

$$G_f(x) = \mathbb{P}_A(\max_{1 \le i \le \tau_A} f(X_i) \le x).$$

Extremal behavior

- Consider the cdf of the cycle submaximum

$$G_f(x) = \mathbb{P}_A(\max_{1\leq i\leq \tau_A} f(X_i) \leq x).$$

Theorem (Rootzen, 1988)

$$\mathbb{P}_\nu(\max_{1\leq i\leq n} f(X_i) \leq x) \sim G_f(x)^{n/\alpha}.$$

The asympt. behavior of $\max_{i\leq n} f(X_i)$ *is ruled by the tail properties of* G_f.

Extremal behavior

- Consider the cdf of the cycle submaximum

$$G_f(x) = \mathbb{P}_A(\max_{1\leq i\leq \tau_A} f(X_i) \leq x).$$

Theorem (Rootzen, 1988)

$$\mathbb{P}_\nu(\max_{1\leq i\leq n} f(X_i) \leq x) \sim G_f(x)^{n/\alpha}.$$

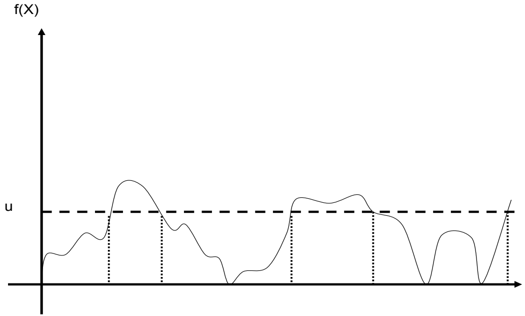
The asympt. behavior of $\max_{i\leq n} f(X_i)$ *is ruled by the tail properties of* G_f.

- Let $H_{\mu,\,f}(x) = \mathbb{P}_\mu(f(X) \leq x) = \alpha^{-1}\mathbb{E}_A[\sum_{i=1}^{\tau_A} f(X_i) \leq x]$. The cdf's G_f and $H_{\mu,\,f}$ are related via the extremal index θ:

$$G_f(x)^{n/\alpha} \sim \mathbb{P}_\nu(\max_{1\leq i\leq n} f(X_i) \leq x) \sim H_{\mu,\,f}(x)^{n\theta}.$$

Extremal behavior

$$G_f \in MDA(\xi) \Leftrightarrow H_{\mu,\, f} \in MDA(\xi).$$



The regenerative method - First order results

The regenerative method - First order results

- **Law of Large Numbers** - Set $\alpha = \mathbb{E}_A[\tau_A]$. Count the nb of renewals up to time n: $l_n = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A\}} \sim n/\alpha$. Write

$$\sum_{i=1}^{n} f(X_i) = \sum_{i=1}^{\tau_A} f(X_i) + \sum_{j=1}^{l_n - 1} S_j(f) + \sum_{i=1+\tau_A(l_n)}^{n} f(X_i),$$

with $S_j(f) = \sum_{i=1+\tau_A(j)}^{\tau_A(j+1)} f(X_i)$. The $S_j(f)$'s are i.i.d. r.v.'s:

$$\Rightarrow \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \to \mu(f) = \alpha^{-1} \mathbb{E}[S_1(f)] \text{ as } n \to \infty.$$

- **Law of Large Numbers** - Set $\alpha = \mathbb{E}_A[\tau_A]$. Count the nb of renewals up to time n: $l_n = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A\}} \sim n/\alpha$. Write

$$\sum_{i=1}^{n} f(X_i) = \sum_{i=1}^{\tau_A} f(X_i) + \sum_{j=1}^{l_n - 1} S_j(f) + \sum_{i=1+\tau_A(l_n)}^{n} f(X_i),$$

with $S_j(f) = \sum_{i=1+\tau_A(j)}^{\tau_A(j+1)} f(X_i)$. The $S_j(f)$'s are i.i.d. r.v.'s:

$$\Rightarrow \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \to \mu(f) = \alpha^{-1}\mathbb{E}[S_1(f)] \text{ as } n \to \infty.$$

- **Central Limit Theorem** - Suppose that $\mathbb{E}_A[(\sum_{i=1}^{\tau_A} f(X_i))^2] < \infty$,

$$\sqrt{n}(\hat{\mu}_n - \mu(f)) \Rightarrow \mathcal{N}(0, \sigma_f^2) \text{ as } n \to \infty,$$

with $\sigma_f^2 = \alpha^{-1}\mathbb{E}_A[(\sum_{i=1}^{\tau_A}\{f(X_i) - \mu(f)\})^2] \neq Var_\mu[f(X)]$.

Non asymptotic bounds and second order results

- Given a traject. of length n, the data blocks correspond. to renewal times

$$B_0 = (X_1, \ldots, X_{\tau_A}),\ B_1 = (X_{1+\tau_A(1)}, \ldots, X_{\tau_A(2)}),\ \ldots$$

$$\mathrm{B}_{l_n - 1} = (X_{1+\tau_A(l_n - 1)}, \ldots, X_{1+\tau_A(l_n)}),\ B_{l_n}^{(n)} = (X_{1+\tau_A(l_n)}, \ldots, X_n)$$

are generally NOT INDEPENDENT! (their lengths sum up to n)

Non asymptotic bounds and second order results

- Given a traject. of length n, the data blocks correspond. to renewal times

$$B_0 = (X_1, \ldots, X_{\tau_A}),\ B_1 = (X_{1+\tau_A(1)}, \ldots, X_{\tau_A(2)}),\ \ldots$$

$$\mathrm{B}_{l_n-1} = (X_{1+\tau_A(l_n-1)}, \ldots, X_{1+\tau_A(l_n)}),\ B_{l_n}^{(n)} = (X_{1+\tau_A(l_n)}, \ldots, X_n)$$

are generally NOT INDEPENDENT! (their lengths sum up to n)

- The regen. method comprises three steps: set $L_j = \tau_A(j+1) - \tau_A(j)$.

Non asymptotic bounds and second order results

- Given a traject. of length n, the data blocks correspond. to renewal times

$$B_0 = (X_1, \ldots, X_{\tau_A}),\ B_1 = (X_{1+\tau_A(1)}, \ldots, X_{\tau_A(2)}),\ \ldots$$

$$\mathrm{B}_{l_n - 1} = (X_{1+\tau_A(l_n-1)}, \ldots, X_{1+\tau_A(l_n)}),\ B_{l_n}^{(n)} = (X_{1+\tau_A(l_n)}, \ldots, X_n)$$

are generally NOT INDEPENDENT! (their lengths sum up to n)

- The regen. method comprises three steps: set $L_j = \tau_A(j+1) - \tau_A(j)$.
  1. **Partition** the underlying prob. space (Ω, $\mathbb{P}_\nu$ according to all fashions for X to regenerate up to time n

$$U_{r,l,m} = \{\tau_A = r,\ \sum_{j=1}^{m-1} L_j = n - r - l,\ \tau_A(m+1) - \tau_A(m) > l\}.$$

Non asymptotic bounds and second order results

- Given a traject. of length n, the data blocks correspond. to renewal times

$$B_0 = (X_1, \ldots,\ X_{\tau_A}),\ B_1 = (X_{1+\tau_A(1)}, \ldots,\ X_{\tau_A(2)}),\ \ldots$$

$$\mathrm{B}_{l_n-1} = (X_{1+\tau_A(l_n-1)}, \ldots,\ X_{1+\tau_A(l_n)}),\ B_{l_n}^{(n)} = (X_{1+\tau_A(l_n)}, \ldots,\ X_n)$$

are generally NOT INDEPENDENT! (their lengths sum up to n)

- The regen. method comprises three steps: set $L_j = \tau_A(j+1) - \tau_A(j)$.
  1. **Partition** the underlying prob. space $(\Omega,\ \mathbb{P}_\nu$ according to all fashions for X to regenerate up to time n

$$U_{r,l,m} = \{\tau_A = r,\ \sum_{j=1}^{m-1} L_j = n - r - l,\ \tau_A(m+1) - \tau_A(m) > l\}.$$

  2. Establish the required result on each subset $U_{r,l,m}$ for the i.i.d. sequ. of 1-**lattice random vectors** $\{(S_j(f), L_j)\}_{j\geq 1}$ (Dubinskaite 82, 84a, 84b).

Non asymptotic bounds and second order results

- Given a traject. of length n, the data blocks correspond. to renewal times

$$B_0 = (X_1, \ldots,\ X_{\tau_A}),\ B_1 = (X_{1+\tau_A(1)}, \ldots,\ X_{\tau_A(2)}),\ \ldots$$

$$\mathrm{B}_{l_n - 1} = (X_{1+\tau_A(l_n - 1)}, \ldots,\ X_{1+\tau_A(l_n)}),\ B_{l_n}^{(n)} = (X_{1+\tau_A(l_n)}, \ldots,\ X_n)$$

  are generally NOT INDEPENDENT! (their lengths sum up to n)
- The regen. method comprises three steps: set $L_j = \tau_A(j+1) - \tau_A(j)$.
  1. **Partition** the underlying prob. space $(\Omega,\ \mathbb{P}_\nu$ according to all fashions for X to regenerate up to time n

$$U_{r,l,m} = \{\tau_A = r,\ \sum_{j=1}^{m-1} L_j = n - r - l,\ \tau_A(m+1) - \tau_A(m) > l\}.$$

  2. Establish the required result on each subset $U_{r,l,m}$ for the i.i.d. sequ. of 1-**lattice random vectors** $\{(S_j(f), L_j)\}_{j\geq 1}$ (Dubinskaite 82, 84a, 84b).
  3. **Sum up** the results obtained so as to **identify** the global bound/limit.

Non asymptotic bounds and second order results

Non asymptotic bounds and second order results

- "Local Berry-Esseen bound" (Bolthausen, 1980)

$$\sup_x |\mathbb{P}_\nu(T_n \leq x) - \Phi(x)| \leq \frac{cst}{n^{1/2}}.$$

Non asymptotic bounds and second order results

- "Local Berry-Esseen bound" (Bolthausen, 1980)

$$\sup_x |\mathbb{P}_\nu(T_n \leq x) - \Phi(x)| \leq \frac{cst}{n^{1/2}}.$$

- Edgeworth expansions
(Malinovskii 1986, Bertail & Clémençon 2004a)

$$\sup_x |\mathbb{P}_\nu(T_n \leq x) - \Phi(x) - \frac{\lambda(2x^2+1)}{6\sqrt{n}}\phi(x)| = O(n^{-1}).$$

Non asymptotic bounds and second order results

- "Local Berry-Esseen bound" (Bolthausen, 1980)

$$\sup_x |\mathbb{P}_\nu(T_n \leq x) - \Phi(x)| \leq \frac{cst}{n^{1/2}}.$$

- Edgeworth expansions
(Malinovskii 1986, Bertail & Clémençon 2004a)

$$\sup_x |\mathbb{P}_\nu(T_n \leq x) - \Phi(x) - \frac{\lambda(2x^2+1)}{6\sqrt{n}}\phi(x)| = O(n^{-1}).$$

- Moment/probability ineq.
(Clémençon 2001, Bertail & Clémençon 2006)

$$\mathbb{P}_\nu(\sum_{i\leq n} f(X_i) \geq x) \leq C \exp\{-\frac{x^2}{4\sigma_f^2}\}.$$

Regeneration-based statistics

Regeneration-based statistics

- **Asymptotic mean and variance estimation**

Regeneration-based statistics

- **Asymptotic mean and variance estimation**
  - Empirical means: $\hat{\mu}_n(f) = \frac{1}{n}\sum_{i \leq n} f(X_i)$, $\mu_n(f) = \frac{\sum_{i=1+\tau_A}^{\tau_A(l_n)} f(X_i)}{\tau_A(l_n) - \tau_A}$.

Regeneration-based statistics

- **Asymptotic mean and variance estimation**
  - Empirical means: $\hat{\mu}_n(f) = \frac{1}{n}\sum_{i\leq n} f(X_i)$, $\mu_n(f) = \frac{\sum_{i=1+\tau_A}^{\tau_A(l_n)} f(X_i)}{\tau_A(l_n)-\tau_A}$.

$$\mathbb{E}_\nu[\hat{\mu}_n(f)] = \mu(f) + (\phi_\nu + \gamma - \beta/\alpha)n^{-1} + O(n^{-3/2}),\ \text{with}$$

1. $\phi_\nu = \mathbb{E}_\nu[\sum_{i\leq\tau_A}\{f(X_i) - \mu(f)\}]$,
2. $\beta = cov(S_j(f), L_j) = \mathbb{E}_A[(\sum_{i\leq\tau_A}\{f(X_i) - \mu(f)\})(\tau_A - \alpha)]$,
3. $\gamma = \alpha^{-1}\mathbb{E}_A[\sum_{i\leq\tau_A}(\tau_A - i)\{f(X_i) - \mu(f)\}]$.

$\Rightarrow$ significantly biased in the nonstationary case.

Regeneration-based statistics

- **Asymptotic mean and variance estimation**
  - Empirical means: $\hat{\mu}_n(f) = \frac{1}{n}\sum_{i\le n} f(X_i)$, $\mu_n(f) = \frac{\sum_{i=1+\tau_A}^{\tau_A(l_n)} f(X_i)}{\tau_A(l_n)-\tau_A}$.

$$\mathbb{E}_\nu[\hat{\mu}_n(f)] = \mu(f) + (\phi_\nu + \gamma - \beta/\alpha)n^{-1} + O(n^{-3/2}),\ \text{with}$$

1. $\phi_\nu = \mathbb{E}_\nu[\sum_{i\le\tau_A}\{f(X_i) - \mu(f)\}]$,
2. $\beta = cov(S_j(f), L_j) = \mathbb{E}_A[(\sum_{i\le\tau_A}\{f(X_i) - \mu(f)\})(\tau_A - \alpha)]$,
3. $\gamma = \alpha^{-1}\mathbb{E}_A[\sum_{i\le\tau_A}(\tau_A - i)\{f(X_i) - \mu(f)\}]$.

$\Rightarrow$ significantly biased in the nonstationary case.

- **Regeneration-based variance estimator**

Definition (Bertail and Clémençon 2004a)

If $l_n > 1$, $\hat{\sigma}_n^2(f) = \frac{1}{n}\sum_{j=1}^{l_n-1}\{S_j(f) - \mu_n(f)L_j\}^2$.

Regeneration-based statistics

Theorem (Bertail and Clémençon, 2004a)

Under adequate 'block'-moment/Cramer conditions,

$$\mathbb{E}_\nu[\hat{\sigma}_n^2(f)] = \sigma_f^2 + O(n^{-1}) \text{ as } n \to \infty.$$

Set $\bar{f} = f - \mu$ *and* $\xi_f^2 = \alpha^{-1} Var_A[(\sum_{i \le \tau_A} \bar{f}(X_i))^2 - 2\alpha^{-1}\beta \sum_{i \le \tau_A} \bar{f}(X_i)]$.

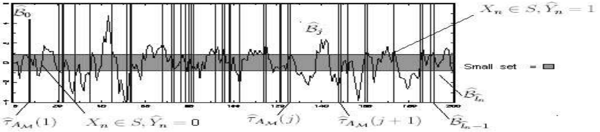
$$\begin{aligned} \sqrt{n}(\hat{\sigma}_n^2(f) - \sigma_f^2) &\Rightarrow \mathcal{N}(0, \xi_f^2), \\ \sqrt{n}\frac{\mu_n(f) - \mu(f)}{\hat{\sigma}_n(f)} &\Rightarrow \mathcal{N}(0, 1). \end{aligned}$$

General (pseudo-regenerative) case

- Consider a small set S and associated parameters ϕ, δ.
- Apply the method to the split chain (X, Y), but... Y is not observable!
- The distribution of $(Y_1, \ldots, Y_n)$ depends on $\pi(x, y)$.

$$Y_i \sim Ber\left(\frac{\pi(X_i, X_{i+1}) - \delta\phi(X_{i+1})}{1 - \delta}\right) \text{ if } X_i \in S.$$

$\Rightarrow$ replace the unknown $\pi(x, y)$ by an estimate $\hat{\pi}_n(x, y)$
Get the **approx.** $(X_1, \hat{Y}_1), \ldots, (X_n, \hat{Y}_n)$
$\Rightarrow$ Construct pseudo-regenerative blocks $\hat{B}_1, \ldots, \hat{B}_{\hat{l}_n - 1}$.



General (pseudo-regenerative) case

- **Accuracy of the approximation**: $P^{(n)} \approx \hat{P}^{(n)}$.

Theorem (Bertail and Clémençon, 2005b)

Suppose that

1. *S is chosen so that* $\inf_{x \in S} \phi(x) > 0$,
2. $\mathbb{E}[\sup_{(x,y)\in S^2} |\pi(x,y) - \hat{\pi}_n(x,y)|^2] \leq \alpha_n$.

Then,

$$l_1(P^{(n)}, \hat{P}^{(n)}) \leq (\delta \inf_{x \in S} \phi(x))^{-1} \alpha_n^{1/2}.$$

Wasserstein/Mallows distance:

$$l_1(P^{(n)}, \hat{P}^{(n)}) = \sum_{k \leq n} 2^{-k} \inf_{\substack{Z \sim Y_k \\ \hat{Z} \sim \hat{Y}_k}} \mathbb{E}[\|Z - \hat{Z}\|]$$

General (pseudo-regenerative) case

- Compute statistics as if $\hat{B}_1, \ldots, \hat{B}_{\hat{l}_n - 1}$ were true regenerative data blocks.

Theorem (Bertail and Clémençon, 2005a)

Under adequate 'block'-moment/Cramer conditions,

$$\begin{aligned} \mathbb{E}_\nu[\mu_n(f)]) &= \mu(f) - \beta/\alpha n^{-1} + O(n^{-1}\alpha_n^{1/2}), \\ \mathbb{E}_\nu[\hat{\sigma}_n^2(f)] &= \sigma_f^2 + O(\alpha_n \vee n^{-1}) \text{ as } n \to \infty. \end{aligned}$$

And as $n \to \infty$,

$$\sqrt{n}\frac{\mu_n(f) - \mu(f)}{\hat{\sigma}_n(f)} \Rightarrow \mathcal{N}(0,1).$$

Rate loss vanishes as α_n gets closer and closer to the parametric rate.

Regeneration-based U-statistics

- **Generalization** of standard means.
- Let $U : E^2 \to \mathbb{R}$ be a sym. kernel. Consider now parameters of type

$$\mu(U) = \mathbb{E}_{(X,X')\sim\mu\otimes\mu}[U(X,X')] = \alpha^{-2}\mathbb{E}_A[\sum_{i=1}^{\tau_A(1)} \sum_{j=1+\tau_A(1)}^{\tau_A(2)} U(X_i, X_j)].$$

- **Example**: Gini index $G(\mu) = \int_x \int_y |x-y|\mu(dx)\mu(dy)$.
- "U-stat. based on (pseudo-) regen. blocks (Bertail and Clémençon, '06a)

$$U_L = \frac{2}{L(L-1)} \sum_{1\leq k<l\leq L} \omega_U(B_k, B_l),$$

with $\omega_U(\mathbf{x},\mathbf{y}) = \sum_{i\leq k}\sum_{j\leq l} U(x_i, y_j)$ sym. ker. on the torus $\mathbb{T} = \cup_{k\geq 1} E^k$.

Regeneration-based U-statistics

- Regeneration-based Hoeffding's decomposition

$$U_L - \mu(U) = \frac{2}{L}\sum_{k=1}^{L}\omega_U^{(1)}(B_k) + \frac{2}{L(L-1)}\sum_{1\le k<l\le L}\omega_U^{(2)}(B_k, B_l),$$

$$\begin{aligned} \text{with } \omega_U^{(1)}(b_1) &= \mathbb{E}[\tilde{\omega}_U(B_1, B_2) \mid B_1 = b_1], \\ \omega_U^{(2)}(b_1, b_2) &= \tilde{\omega}_U(b_1, b_2) - \omega_U^{(1)}(b_1) - \omega_U^{(1)}(b_1), \end{aligned}$$

$$\text{where } \tilde{\omega}_U(B_k, B_l) = \sum_{i=1+\tau_A(k)}^{\tau_A k+1}\sum_{j=1+\tau_A(l)}^{\tau_A l+1}\{U(X_i, X_j) - \mu(U)\}.$$

- $\Rightarrow$ asymptotic properties of $T_n = \frac{2}{\tilde{n}(\tilde{n}-1)}\sum_{1+\tau_A\le i<j\le\tau_A(l_n)} U(X_i, X_j)$.

- If we knew the type of tail behav. of $G_f/H_{\mu,\,f}$ (max. dom. attr.. hyp. MDA(H_ξ)) $\Rightarrow$ apply any standard meth. based on the observed submax. $\zeta_1(f), \ldots, \zeta_{l_n-1}(f)$ for est. the shape par. ξ, norming const. for the max.
- **Example**: Fréchet case - $f(x) = x$ - "Regeneration-based Hill estimator"

Definition (Bertail and Clémençon, 2006a)

Let $\zeta_{(j)}$ be the j-largest submaximum. For $k < l_n - 1$, define

$$RH_k = \left(k^{-1} \sum_{i=1}^{k} \log \frac{\zeta_{(j)}}{\zeta_{(k+1)}} \right)^{-1}.$$

- Est. of the extr. index θ. If $n(1 - G(u_n))/\alpha \to \eta < \infty$, $N_n/n^2 \to \infty$,

$$\hat{\theta}_n = \frac{\sum_{j \le l_{N_n}-1} \mathbb{I}\{\zeta_j > u_n\}}{\sum_{i \le N_n} \mathbb{I}\{X_i > u_n\}} \to \theta.$$

Regenerative-block bootstrap

Regenerative-block bootstrap

"Pulling yourself up by your own bootstraps" - B. Efron (1979)

Regenerative-block bootstrap

"Pulling yourself up by your own bootstraps" - B. Efron (1979)

- Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_\theta$, T_n estimator of θ.
  - The goal: **estimate the accuracy of the estimate**.
  - Est. $H(x) = \mathbb{P}(\frac{T_n - \theta}{S_n} \leq x)$ by resampling $X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} \hat{F}_n$ $\Rightarrow$ recompute stat. T_n^* and S_n^* and consider the bootstrap dist. estimate

$$H_n(x) = \mathbb{P}(\frac{T_n^* - T_n}{S_n^*} \leq x \mid X_1, \ldots, X_n).$$

  - Under regularity assumptions, **2nd order accuracy**

$$\sup_x |H_n(x) - H(x)| = O_{\mathbb{P}}(n^{-1}), \text{ as } n \to \infty,$$

$\Rightarrow$ asympt. more accurate than the Gauss. approx. (*cf* Berry-Esseen)

$$H(x) = \Phi(x) - n^{-1/2} h(x) \frac{d\Phi}{dx}(x) + O(n^{-1})$$

Bootstrap for dependent data

- Statistical challenge for dependent data:

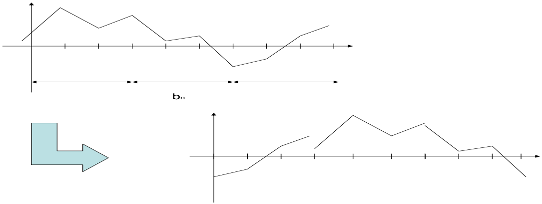
Draw $X_1^*, \ldots, X_n^*$ so as to "mimic" the dependence structure.

- "Model based approach"
  - ex: $X_{n+1} - \hat{a}X_n = \hat{\epsilon}_{n+1}$, $\Rightarrow$ apply the "i.i.d. Bootstrap" to the residuals.
  - "Sieve bootstrap" (*Bühlmann*, 1997) in case of (quasi-) linear structure.
- **Moving-block bootstrap** (*Künsch*, 1989): div. the traj. $X_1, \ldots, X_n$ into data blocks of length $b_n = o(n)$: $B_1, B_2, \ldots$, and resamp. data blocks

$$B_1^*, \ldots, B_n^* \overset{i.i.d.}{\sim} \frac{1}{n - b_n + 1} \sum_{i=1}^{n-b_n+1} \delta_{B_i}.$$

and consider the reconstructed pseudo-trajectory.

Bootstrap for dependent data



- Moving Block Bootstrap "Drawbacks":
  - the bootstrap statistics T_n^* and S_n^* are biased (artificial jumps) $\Rightarrow$ necessary recentering and correction.
  - 2nd order accuracy but with rate $n^{-2/3}$ at best under stationarity/geometrically decr. mixing rate assumptions.
- Blocking techniques may also be used for asympt. variance est. (*Götze* and *Künsch*, 1996), statist. study of extremes.

Algorithm/theory for the regenerative block-bootstrap

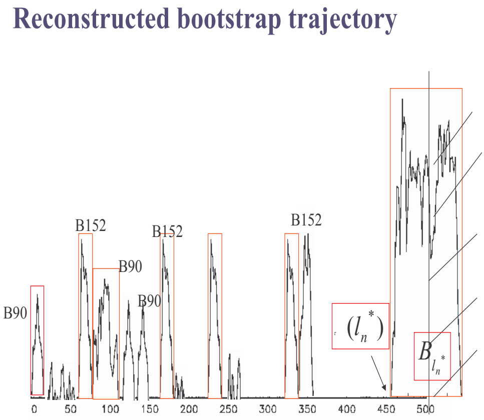
- **Heuristics**: resample (pseudo-) regenerative blocks but... recall that the nb of renewals l_n over a trajectory of length n is random !

$\Rightarrow$ Mimic the renewal structure of the chain by generating a random nb of blocks until the length of the reconstructed bootstrap series is $\geq n$.

- The (approximate) RBB algorithm (Bertail and Clémençon, 05b)

1. Count the nb of visits l_n and form the regen. blocks $B_1, \ldots, B_{l_n - 1}$. Comp. T_n and S_n from the regen. blocks only.
2. Cond. on $X^{(n)} = (X_1, \ldots, X_n)$, draw $B_1^*, \ldots, B_k^* \overset{i.i.d.}{\sim} (l_n - 1)^{-1} \sum_{j=1}^{l_n - 1} \delta_{B_j}$ until $L^*(k) = \sum_{j \leq k} > n$. Set $l_n^* = \inf\{k : L^*(k) > n\}$.
3. Compute the RBB stat. T_n^*, S_n^* from $B_1^*, \ldots, B_{l_n^* - 1}^*$.
4. Estimate $H_\nu(x) = \mathbb{P}_\nu(\frac{T_n - \theta}{S_n} \leq x)$ by $H_{RBB}(x) = \mathbb{P}^*(\frac{T_n^* - T_n}{S_n^*} \leq x \mid X^{(n)})$.

Algorithm/theory for the regenerative block-bootstrap



Algorithm/theory for the regenerative block-bootstrap

- **Second order accuracy** for regeneration-based studentized sample means

Theorem (Bertail and Clémençon, 2005b)

Under adequate "block" moment/Cramer conditions,

$$\sup_x |H_{RBB}(x) - H_\nu(x)| = O_{\mathbb{P}_\nu}(n^{-1}), \text{ as } n \to \infty.$$

Same optimal rate as in the i.i.d. case.

- **Insights**: Edgeworth expansion (Bertail and Clémençon, 2004a)

$$\sup_x |\mathbb{P}_\nu(T_n \leq x) - E_n(x)| = O(n^{-1}),$$

with $E_n(x) = \Phi(x) - n^{-1/2}(\frac{\kappa_3}{6}(x^2 - 1) + b)\frac{d\Phi}{dx}(x)$, $b = -\beta/\alpha$, $\kappa_3 = \alpha^{-1}(M_3 - 3\beta/\sigma) \neq \alpha^{-1} M_3$ and $M_3 = \mathbb{E}_A[(\sum_{i=1}^{\tau_A} \bar{f}(X_i))^3]$.

Algorithm/theory for the regenerative block-bootstrap

- For general pos. rec. chains, the algorithm is performed from appr. blocks.

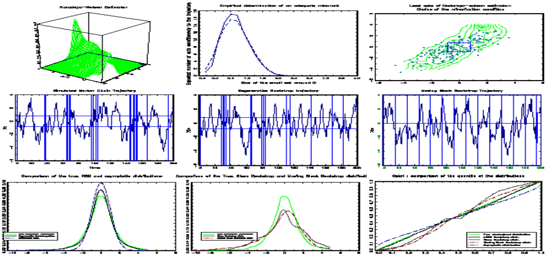
Theorem (Bertail and Clémençon, 2005b)

$$\sup_x |H_{ARBB}(x) - H_\nu(x)| = O_{\mathbb{P}_\nu}(n^{-5/6}\log(n)), \ as\ n \to \infty.$$

- Empirical choice for S and related parameters δ, ϕ. For instance, take ϕ as a uniform df and $\delta(S) = \inf_{(x,y)\in S^2} \pi(x,y)$.
  - As S grows, visits to S occur more frequently but... the splitting probability δ_S decreases.
  - Ideally, choose S so as to max. the expected number of renewals for the split chain $N_n(S) = \frac{\delta(S)}{\lambda(S)} \sum_{i\leq n} \frac{\mathbb{I}\{(X_i,X_{i+1})\in S^2\}}{\pi(X_i,X_{i+1})}$ cond. on the traject. $X_1,\ldots,\ X_n$.
  - Practically opt. the empirical counterpart $N_n(S) = \frac{\delta_n(S)}{\lambda(S)} \sum_{i\leq n} \frac{\mathbb{I}\{(X_i,X_{i+1})\in S^2\}}{\hat{\pi}_n(X_i,X_{i+1})}$.
- Asymptotic validity of the (A)RBB for regeneration-based U-statistics.

Simulation studies

- **Example**:



Simulation studies

- Empirical comparison between MBB, (A)RBB and Sieve Bootstrap
  - (A)RBB and Sieve Boostrap "always" provide better results than the MBB
  - For "quasi-linear" models, the Sieve Bootstrap performs better than the ARBB
  - For significantly non-linear models, the ARBB surpasses its (nonparametric) competitors.

Related bibliography

1. Sharp probability inequalities for Markov chains. (2006), with P. Bertail.
2. Approximate Regenerative Block Bootstrap: some simulation studies. (2006), with P. Bertail. To appear in **Comp. Stat. and Data Analysis**.
3. Regeneration-based statistics for Harris Markov chains. (2006), with P. Bertail. In **Dependence in Probability and Statistics**, LNS, No 187, 1-53. Springer.
4. Regenerative Block Bootstrap for Markov Chains. (2005), with P. Bertail. In **Bernoulli**.
5. Note on the regeneration-based bootstrap for atomic Markov chains. (2005), with P. Bertail. In **Test**.
6. Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. (2004), with P. Bertail. In **PTRF**.
7. Approximate Regenerative Block-Bootstrap for Markov Chains: second-order properties. (2004), with P. Bertail. In **Compstat 2004 Proc.** Physica Verlag.