Renewal approach to U-Statistics for Markovian data

Patrice Bertail, Stéphan Clémençon, Jessica Tressou

Renenewal Approach to U-Statistics For Markovian Data

Patrice Bertail*, Stéphan Clémençon, Jessica Tressou
* CREST - INSEE & MODAL'X - Université Paris X
200, ave de la République, 92000 Nanterre

Keywords : Markoc Chain, Nummelin splitting approach, U-statistics, Berry Esseen.

Résumé: Soit une chaîne de Markov X , positive récurrente ergodique de mesure stationnaire μ et une U-statistique basée sur cette chaîne. Les propriétés asymptotiques des U-statistiques basées sur des v.a. indépendantes and identiquement distribuées sont bien connues depuis les années 60 mais font actuellement l'objet de travaux intensifs dans le cadre de données dépendantes. Le but de ce travail est de développer une alternative aux méthodes de couplage, adaptée aux processus regénératifs (ou pseudo-regénératifs).En effet, il est connu que les chaines Harris récurrentes peuvent être découpées en un nombre (aléatoire) de blocs de regénération i.i.d. . Nous développons cette approche dans le cadre des U-statistiques de chaînes Harris récurrentes. Nous établissons une loi forte, des théorèmes centraux limites et des bornes de type Berry-Esséen pour des U-statistiques sous des hypothèses faibles. Etendant les idées de [4], nous proposons également une méthode de bootstrap regénératifs et établissons ses propriétés asymptotiques.

Abstract : Consider a Markov chain X assumed to be positive recurrent with limiting probability distribution μ used to construct a U-statistics. Whereas the asymptotic properties of U-statistics based on independent and identically distributed data are well understood since the sixties the study of this specific class of statistics for dependent data has recently received special attention in the statistical literature.The purpose of this paper is to develop an alternative to the coupling methodology, specifically tailored for regenerative processes or stochastic processes for which a regenerative extension may be built, namely *pseudo-regenerative processes*. Indeed, sample paths of a Harris chain may be classically divided into i.i.d. *regeneration blocks*, namely data segments between random times at which the chain forgets its past. We develop further this view, in order to accurately investigate the asymptotic properties of U-statistics of positive Harris chains. A Strong Law of Large Numbers, Central Limit Theorem and Berry-Esseen bounds are established for markovian U-statistics under weak hypotheses. Following [4], we also propose to bootstrap certain markovian U statistics, using a specific resampling procedure, producing bootstrap data series with a renewal structure mimicking that of the original chain.

1 Introduction

Let $X = (X_n)_{n\in\mathbb{N}}$ be $\psi-$Harris recurrent Markov chain on a countably generated state space $(E,\ \mathcal{E})$, with transition probability Π, and initial probability distribution ν. For any $B \in \mathcal{E}$ and any $n \in \mathbb{N}$, we thus have

$$X_0 \sim \nu \text{ and } \mathbb{P}(X_{n+1} \in B \mid X_0, ...,\ X_n) = \Pi(X_n,\ B) \text{ a.s. .}$$

In what follows, μ will denote the stationary measure , $\mathbb{P}_\nu$ (respectively $\mathbb{P}_A$) will denote the probability measure on the underlying probability space such that $X_0 \sim \nu$ (resp. $X_0 \in A$), $\mathbb{E}_\nu(.)$ the $\mathbb{P}_\nu$-expectation (resp. $\mathbb{E}_A(.)$ the $\mathbb{P}_A$-expectation).

We focus here on the stimation of U-parameters of type

$$\mu(h) = \int_{x_1\in E} \dots \int_{x_k\in E} h(x_1, \dots, x_k)\mu(dx_1)\dots\mu(dx_k), \tag{1}$$

where $k \geq 2$ and $h : E^k \to \mathbb{R}^l$ is a measurable function, $l \geq 1$, such that the quantity (1) is well-defined. For simplicity's sake, we shall restrict ourselves to the case where $k = 2$ and the kernel $h(x, y)$ is symmetric, *i.e.* $\forall (x, y) \in E^2$, $h(x, y) = h(y, x)$. All results of this paper straightforwardly extend to the general case. Like in the i.i.d. setting, a natural counterpart of (1) based on a sample path $X_1,\ \dots,\ X_n$ is given by the U-statistics

$$U_n(h) = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} h(X_i, X_j). \tag{2}$$

Whereas the asymptotic properties of U-statistics based on independent and identically distributed data are well understood since the sixties, the study of this specific class of statistics, generalizing sample means, for dependent data has recently received special attention in the statistical literature. The problem of extending classical limit theorems for U-statistics in the i.i.d. setup to the weakly dependent framework is generally tackled by using *coupling techniques*.

2 The pseudo-regeneration technique

The purpose of this work is to develop an alternative to the coupling methodology, specifically tailored for regenerative processes or stochastic processes for which a regenerative extension may be built, namely *pseudo-regenerative processes*. This includes the case of general Harris Markov chains on which the present study focuses. Within this framework, a Markov chain is said *regenerative* when it possesses an accessible atom, *i.e.*, a measurable set A such that $\psi(A) > 0$ and $\Pi(x, .) = \Pi(y, .)$ for all $(x, y) \in A^2$. Denote then by $\alpha = \tau_A = \tau_A(1) = \inf\{n \geq 1,\ X_n \in A\}$ the hitting time on A, by $\tau_A(j) = \inf\{n > \tau_A(j-1),\ X_n \in A\}$, for $j \geq 2$, the successive return times to A.

In the atomic case, it follows from the *strong Markov property* that the blocks of observations in between consecutive visits to the atom

$$\mathcal{B}_1 = (X_{\tau_A(1)+1}, ..., X_{\tau_A(2)}),\ \ldots,\ \mathcal{B}_j = (X_{\tau_A(j)+1}, ..., X_{\tau_A(j+1)}),\ \ldots \tag{3}$$

form a collection of i.i.d. random variables, valued in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$, and the sequence $\{\tau_A(j)\}_{j\geq 1}$, corresponding to successive times at which the chain forgets its past, is a (possibly delayed) renewal process. We point out that the class of atomic chains is not as restrictive as it seems at first glance. It contains all chains with a countable state space (any recurrent state is an accessible atom), as well as many specific Markov models arising from the field of operations research, see [1] for instance. Moreover any recurrent positive chain may be extended to a "split-chain" which posseses an atom, through the Nummelin *splitting technique*, see [3]. This relies on the notion of *small set* and the following condition referred to as *minorization condition*. A set $S \in \mathcal{E}$ is said to be *small* for the chain if there exist $m \in \mathbb{N}^*$, $\delta > 0$ and a probability measure Φ supported by S such that: $\forall (x, B) \in S \times \mathcal{E}$,

$$\Pi^m(x, B) \geq \delta \Phi(B), \tag{4}$$

denoting by Π^m the m-th iterate of the transition kernel Π. Rather than replacing the original chain X by the chain $\{(X_{nm}, ..., X_{n(m+1)-1})\}_{n\in\mathbb{N}}$, we assume that $m = 1$ here and throughout, with no loss of generality. The Nummelin technique then consists of building a sequence $Y = (Y_n)_{n\in\mathbb{N}}$ of independent Bernoulli r.v.'s with parameter δ such that (X, Y) is a bivariate Markov chain, referred to as the *split chain*, with state space $E \times \{0, 1\}$, the marginal Y indicating whether the chain (X, Y), and consequently the original one, regenerates when X visits the small set S. Precisely, if $X_n \notin S$, the conditional distribution of (X_{n+1}, Y_{n+1}) given (X_n, Y_n) is simply the tensorial product between $\Pi(X_n, dy)$ and the Bernoulli distribution Ber_δ. When $X_n \in S$, if we have in addition $Y_n = +1$, which happens with probability δ, (X_{n+1}, Y_{n+1}) is then drawn from $\Phi \otimes Ber_\delta$, otherwise it is distributed according to $(1-\delta)^{-1}(\Pi(X_n, .) - \delta\Phi) \otimes Ber_\delta$. This way, $A_S = S \times \{1\}$ is an accessible atom for the split chain and the latter inherits all communication and stochastic stability properties from the original chain.

It has been suggested in [5] to extend regeneration-based inference techniques the following way: generate first a sequence $(\widehat{Y}_1,\ \ldots,\ \widehat{Y}_n)$ from the supposedly known parameters (S, δ, Φ) in a way that $((X_1, Y_1),\ \ldots,\ (X_n, Y_n))$ and $((X_1, \widehat{Y}_1),\ \ldots,\ (X_n, \widehat{Y}_n))$ have close distributions in the Mallows sense and then apply adequate statistical procedures to the data blocks thus defined $\widehat{\mathcal{B}}_1,\ \ldots,\ \widehat{\mathcal{B}}_{\widehat{N}_n}$ (corresponding to the successive times when $(X, \widehat{Y})$ visits $S \times \{1\}$), as if they were really regenerative. Here we briefly recall the basic principle underlying this approach.

We now assume that there exists a measure of reference $\lambda(dy)$ on the state space $(E, \mathcal{E})$ that dominates the collection of probability measures $\{\Pi(x, dy);\ x \in E\}$. For all $x \in E$, set $\Pi(x, dy) = \pi(x, y) \cdot \lambda(dy)$ and let (S, ϕ, δ) be the parameters of a minorization condition satisfied by X. Clearly, it is not restrictive

to assume that Φ is also absolutely continuous with respect to λ. We thus set $\Phi(dy) = \phi(y) \cdot \lambda(dy)$ and observe that $\forall (x, y) \in S^2$ we almost-surely have $\pi(x, y) \geq \delta\phi(y)$. Conditioned upon $X^{(n)} = (X_1, \ldots, X_n)$, the random variables $Y_1, \ldots, Y_n$ are mutually independent and, for all $i \in \{1, \ldots, n\}$, Y_i is drawn from a Bernoulli distribution with parameter given by:

$$\delta \mathbb{I}_{\{X_i \notin S\}} + \frac{\delta\phi(X_{i+1})}{\pi(X_i, X_{i+1})} \mathbb{I}_{\{X_i \in S\}}. \tag{5}$$

Suppose that an estimate $\widehat{\pi}(x, y)$ of the transition density over $S \times S$, such that $\forall (x, y) \in S^2$, $\widehat{\pi}(x, y) \geq \delta\phi(y)$, is available. Given $X^{(n)}$, the construction of the $\widehat{Y}_i$'s boils down to draw mutually independent Bernoulli random variables, the parameter of $\widehat{Y}_i$'s conditional distribution being obtained by replacing the unknown quantity $\pi(X_i, X_{i+1})$ by its empirical counterpart $\widehat{\pi}(X_i, X_{i+1})$ in (5). A more detailed description of this plug-in approximation is available in [5] together with a discussion of numerical issues regarding its practical implementation (in particular, special attention is paid to the problem of selecting the parameters (S, δ, ϕ) in a data-driven fashion). The accuracy of the resulting approximation, measured in terms of Mallows distance between the random vectors $(Y_1, \ldots, Y_n)$ and $(\widehat{Y}_1, \ldots, \widehat{Y}_n)$, mainly depends on the quality of the estimate $\widehat{\pi}(x, y)$ over S^2.

3 Regenerative U-statitics

Recall that a chain is *positive recurrent* if and only if the expected return time to the atom is finite, *i.e.* $\mathbb{E}_A[\tau_A] < \infty$, see Theorem 10.2.2 in [2]. Its invariant probability distribution μ is then the occupation measure is given by

$$\mu(B) = \alpha^{-1} \mathbb{E}_A \left[\sum_{i=1}^{\tau_A} \mathbb{I}_{\{X_i \in B\}} \right], \text{ for all } B \in \mathcal{E}. \tag{6}$$

so that the parameter of interest may be rewritten as a functional of blocks.

$$\mu(h) = \alpha^{-2} \mathbb{E}_A [\sum_{i=1}^{\tau_A(1)} \sum_{j=1+\tau_A(1)}^{\tau_A(2)} h(X_i, X_j)].$$

Define the regenerative kernel associated to h is the kernel $\omega_h : \mathbb{T}^2 \to \mathbb{R}$ given by

$$\omega_h \left((x_1, \ldots, x_n), \, (y_1, \ldots, y_m) \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} h(x_i, y_j),$$

for all $x^{(n)} = (x_1, \ldots, x_n)$ and $y^{(m)} = (y_1, \ldots, y_m)$ in the torus $\mathbb{T} = \cup_{n \geq 1} E^n$.

On a strecht of observation, denote by $l_n - 1$ the number of (complete regenerative blocks). A regenerative U-statistic associated to the kernel h is a

U-statistic with kernel $\omega_{\bar{h}}$:

$$R_L(h) = \frac{2}{L(L-1)} \sum_{1 \leq k < l \leq L} \omega_{\bar{h}}(\mathcal{B}_k, \mathcal{B}_l),$$

where $L \geq 1$ and $\mathcal{B}_1, \ldots, \mathcal{B}_L$ are regeneration blocks of the chain X. Hence, we may consider its *Hoeffding decomposition*: $R_L(h) = 2S_L(h) + D_L(h)$, where

$$S_L(h) = \frac{1}{L} \sum_{k=1}^{L} h_1(\mathcal{B}_k) \text{ and } D_L(h) = \frac{2}{L(L-1)} \sum_{1 \leq k < l \leq L} h_2(\mathcal{B}_k, \mathcal{B}_l),$$

with $\forall (b_1, b_2) \in \mathbb{T}^2$,

$$h_1(b_1) = \mathbb{E}[\omega_{\bar{h}}(b_1, \mathcal{B}_1)] \text{ and } h_2(b_1, b_2) = \omega_{\bar{h}}(b_1, b_2) - h_1(b_1) - h_1(b_2).$$

As the approximant $R_{l_n - 1}(h)$ is a U-statistic based on regeneration blocks, classical theorems may be applied to the latter and consequently yield the corresponding results for the original statistic. This way, a Strong Law of Large Numbers, a Central Limit Theorem as well a Berry-Esseen Theorem (taking into account the fact that l_n is random) are established for markovian U-statistics under weak hypotheses. We also examine the question of studentizing markovian U-statistics in connection with the construction of confidence intervals.

4 Main results

Let $L \geq 1$. Consider the empirical counterpart of the conditional expectation based on all the first L regenerative data blocks, except $\mathcal{B}_j$, $j \in \{1, \ldots, L\}$:

$$\hat{h}_{1,-j}(b) = \frac{1}{L-1} \sum_{k=1,\ k \neq j}^{L} \omega_h(b, \mathcal{B}_k) - \frac{2}{L(L-1)} \sum_{1 \leq k < l \leq L} \omega_h(\mathcal{B}_k, \mathcal{B}_l),$$

as well as the *Jacknife estimator* of the asymptotic variance $s^2(h)$

$$\hat{s}_L^2(h) = \frac{1}{L} \sum_{k=1}^{L} \hat{h}_{1,-k}^2(\mathcal{B}_k).$$

and define

$$\hat{\sigma}_n^2(h) = 4 \left(l_n/n\right)^3 \hat{s}_{l_n - 1}^2(h).$$

where $\forall (x, y) \in E^2$, $\bar{h}(x, y) = h(x, y) - \int_{z \in E} h(x, z) \mu(dz)$.

Proposition 1 *(the atomic case) Under some natural moment conditions on the moments of τ_A and the moment on the block of observations. Then, as $n \to \infty$,*

(i) *(Central Limit Theorem) We have the convergence in distribution under $\mathbb{P}_\nu$:*

$$\sqrt{n}\,(U_n(h) - \mu(h)) \Rightarrow \mathcal{N}(0, \sigma^2(h)), \text{ as } n \to \infty,$$

where $\sigma^2(h) = 4\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A} h_1(X_i)\right)^2\right]/\alpha^3$.

(ii) *Moreover with* $q = k = l = 4$

$$\hat{\sigma}_n^2(h) \to \sigma^2(h) \ \ \mathbb{P}_\nu\text{-almost-surely.}$$

and we have the convergence in $\mathbb{P}_\nu$*-distribution*

$$\sqrt{n}\{\hat{\sigma}_n^2(h) - \sigma^2(h)\} \Rightarrow \mathcal{N}(0, \Sigma_h^2),$$

(iii) *(Berry-Esseen bound) There exists a constant* $C(P)$ *depending on the moments of absolute order three of* $|h_1|, |h_2|$ *and* $|\tau_A|$ *under* ν *and starting from* A *such that*

$$\sup_x |\mathbb{P}_\nu\left\{\frac{\sqrt{n}}{\sigma(h)}(U_n(h) - \mu(h)) \le x\right\} - \Phi(x)| \le \frac{C(\nu, A)}{\sqrt{n}}$$

The same results holds under additional technical assumptions in the pseudo regenerative caseas well as for the regenerative bootstrap (see [4]). The main difficulties lies in proving the Berry-Esseen Bounds : for this we use Stein methods instead of the partitioning arguments used in [4]

References

[1] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, 2003.

[2] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1996.

[3] E. Nummelin. A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete*, 43:309–318, 1978.

[4] P.Bertail and S. Clémençon. Edgeworth expansions for suitably normalized sample mean statistics of atomic markov chains. *Prob. Th. Rel. Fields*, 130:388–414, 2004.

[5] P.Bertail and S. Clémençon. Regeneration block bootstrap for markov chains. *Bernoulli*, 12:689–712, 2006.