Résumé des Travaux en Statistique et Applications des Statistiques

Stéphan Clémençon

N° attribué par la bibliothèque:

UNIVERSITÉ PARIS X - NANTERRE

Ecole Doctorale Connaissance, Langage et Modélisation

Modélisation Aléatoire de Paris X - MODAL'X

Résumé des travaux en vue de l'habilitation à diriger des recherches

Discipline : Mathématiques Appliquées et Applications des Mathématiques.

Stéphan Clémençon

Le 1er Décembre 2006

M.	Lucien Birgé	Examinateur
M.	Peter Bühlmann	Rapporteur
M.	Eric Moulines	Examinateur
M.	Michael Neumann	Examinateur
Mme	Dominique Picard	Examinateur
M.	Yaacov Ritov	Rapporteur
M.	Philippe Soulier	Rapporteur
M.	Alexandre Tsybakov	Examinateur

Remerciements

Liste de publications et prépublications

- Statistical Inference for an Epidemic Model with Contact-Tracing. (2006), with P. Bertail & V.C. Tran. Submitted.
- Sharp probability inequalities for Markov chains. (2006), with P. Bertail & N. Rhomari. Submitted.
- Statistical analysis of a dynamic model for food contaminant exposure with applications to dietary methyl mercury contamination. (2006), with P. Bertail & J. Tressou. Submitted
- A storage model with random release rate for modelling exposure to food contaminant. (2006), with P. Bertail & J. Tressou. Submitted.
- On Ranking the Best Instances. (2006), with N. Vayatis. Submitted.
- Nonparametric scoring and U-processes. (2006), with G. Lugosi & N. Vayatis. Submitted.
- Portfolio Selection under Extreme Risk Measure: the Heavy-Tailed ICA model. (2006), with S. Slim. To appear in *International Journal of Theoretical and Applied Finance*.
- Approximate Regenerative Block Bootstrap: some simulation studies. (2006), with P. Bertail. To appear in *Computational Statistics and Data Analysis*.
- Some comments on 'Local Rademacher Complexities and Oracle Inequalities in Risk Minimization' by V. Koltchinskii. (2006), with G. Lugosi & N. Vayatis. In *Annals of Statistics*, vol. 34, nř6.
- Regeneration-based statistics for Harris Markov chains. (2006), with P. Bertail. In *Dependence in Probability and Statistics*, eds P. Bertail, P. Soulier & P. Doukhan Lecture Notes in Statistics, No 187, 1-53. Springer-Verlag.
- From classification to ranking: a statistical view. (2006), with G. Lugosi & N. Vayatis. In *Studies in Classification, Data Analysis, and Knowledge Organization*, From Data and Information Analysis to Knowledge Engineering, Vol. 30, eds M. Spiliopoulou, R. Kruse, A. Nürnberger, C. Borgelt & W. Gaul (eds.): Proc. 29th Annual Conference of the GfKl, Otto-von-Guericke-University of Magdeburg, March 9-11, 2005, 214-221. Springer-Verlag, Heidelberg-Berlin, 2006.
- Ranking and Scoring Using Empirical Risk Minimization. (2005), with G. Lugosi & N. Vayatis. In *Lecture Notes in Computer Science*, **3559**, Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005. Proceedings. Editors: Peter Auer, Ron Meir, 1-15. Springer-Verlag.
- Regenerative Block Bootstrap for Markov Chains. (2005), with P. Bertail. To appear in *Bernoulli*.

- Note on the regeneration-based bootstrap for atomic Markov chains. (2005), with P. Bertail. To appear in *Test*.
- Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. (2004), with P. Bertail. *Prob. Th. Rel. Fields*, **130**, 388–414.
- Approximate Regenerative Block-Bootstrap for Markov Chains: second-order properties. (2004), with P. Bertail. In *Compstat 2004 Proc.* Physica Verlag.
- Statistical Analysis of Financial Time Series under the Assumption of Local Stationarity. (2004), with S. Slim. *Quantitative Finance*.
- Nonparametric inference for some class of hidden Markov models. (2002). Rapport technique de l'université Paris X, No 03-9.
- Moment and probability inequalities for sums of bounded additive functionals of regular Markov chains via the Nummelin splitting technique. (2001), *Stat. Prob. Letters,* **55**, 227-238.
- Adaptive estimation of the transition density of a regular Markov chain. (2000), *Math. Meth. Stat.,* **9**, No. 4, 323-357.
- Nonparametric statistics for Markov chains. (2000), PhD thesis, Université Paris 7 Denis Diderot.

Préface

Ce rapport présente brièvement l'essentiel de mon activité de recherche depuis ma thèse de doctorat [53], laquelle visait principalement à étendre l'utilisation des progrès récents de l'*Analyse Harmonique Algorithmique* pour l'estimation non paramétrique adaptative dans le cadre d'observations i.i.d. (tels que l'analyse par ondelettes) à l'estimation statistique pour des données markoviennes. Ainsi qu'il est éxpliqué dans [123], des résultats relatifs aux propriétés de concentration de la mesure (*i.e.* des inégalités de probabilité et de moments sur certaines classes fonctionnelles, adaptées à l'approximation non linéaire) sont indispensables pour exploiter ces outils d'analyse dans un cadre probabiliste et obtenir des procédures d'estimation statistique dont les vitesses de convergence surpassent celles de méthodes antérieures. Dans [53] (voir également [54], [55] et [56]), une méthode d'analyse fondée sur le renouvellement, la méthode dite 'régénérative' (voir [185]), consistant à diviser les trajectoires d'une chaîne de Markov Harris récurrente en segments asymptotiquement i.i.d., a été largement utilisée pour établir les résultats probabilistes requis, le comportement à long terme des processus markoviens étant régi par des processus de renouvellement (définissant de façon aléatoire les segments de la trajectoire). Une fois l'estimateur construit, il importe alors de pouvoir quantifier l'incertitude inhérente à l'estimation fournie (mesurée par des quantiles spécifiques, la variance ou certaines fonctionnelles appropriées de la distribution de la statistique considérée). A cet égard et au delà de l'extrême simplicité de sa mise en oeuvre (puisqu'il s'agit simplement d'effectuer des tirages i.i.d. dans l'échantillon de départ et recalculer la statistique sur le nouvel échantillon, l'*échantillon bootstrap*), le *bootstrap* possède des avantages théoriques majeurs sur l'approximation asymptotique gaussienne (la distribution bootstrap approche automatiquement la structure du second ordre dans le développement d'Edegworth de la distribution de la statistique). Il m'est apparu naturel de considérer le problème de l'extension de la procédure traditionnelle de bootstrap aux données markoviennes. Au travers des travaux réalisés en collaboration avec Patrice Bertail, la méthode régénérative s'est avérée non seulement être un outil d'analyse puissant pour établir des théorèmes limites ou des inégalités, mais aussi pouvoir fournir des méthodes pratiques pour l'estimation statistique: la généralisation du bootstrap proposée consiste à ré-échantillonner un nombre aléatoire de segments de données régénératifs (ou d'approximations de ces derniers) de manière à imiter la structure de renouvellement sous-jacente aux données. Cette approche s'est révélée également pertinente pour de nombreux autres problèmes statistiques. Ainsi la première partie du rapport vise essentiellement à présenter le principe des méthodes statistiques fondées sur le renouvellement pour des chaînes de Markov Harris.

La seconde partie du rapport est consacrée à la construction et à l'étude de méthodes statistiques pour apprendre à ordonner des objets, et non plus seulement à les classer (*i.e.* leur affecter un label), dans un cadre supervisé. Ce problème difficile est d'une importance cruciale dans de nombreux domaines d' application, allant de l'élaboration d'indicateurs pour le diagnostic médical à la recherche d'information (moteurs de recherche) et pose d'ambitieuses questions théoriques et algorithmiques, lesquelles ne sont pas encore résolues de manière satisfaisante. Une approche envisageable consiste à se ramener à la classification de paires d'observations, ainsi que le suggère un critère largement utilisé dans les applications mentionnées ci-dessus (le critère AUC) pour évaluer la

pertinence d'un ordre. Dans un travail mené en collaboration avec Gabor Lugosi et Nicolas Vayatis, plusieurs résultats ont été obtenus dans cette direction, requérant l'étude de U-processus: l'aspect novateur du problème résidant dans le fait que l'estimateur naturel du risque a ici la forme d'une U-statistique. Toutefois, dans de nombreuses applications telles que la recherche d'information, seul l'ordre relatif aux objets les plus pertinents importe véritablement et la recherche de critères correspondant à de tels problèmes (dits *d'ordre localisé*) et d'algorithmes permettant de construire des règles pour obtenir des 'rangements' optimaux à l'égard de ces derniers constitue un enjeu crucial dans ce domaine. Plusieurs développements en ce sens ont été réalisés dans une série de travaux (se poursuivant encore actuellement) en collaboration avec Nicolas Vayatis.

Enfin, la troisième partie du rapport reflète mon intérêt pour les applications des concepts probabilistes et des méthodes statistiques. Du fait de ma formation initiale, j'ai été naturellement conduit à considérer tout d'abord des applications en finance. Et bien que les approches historiques ne suscitent généralement pas d'engouement dans ce domaine, j'ai pu me convaincre progressivement du rôle important que pouvaient jouer les méthodes statistiques non paramétriques pour analyser les données massives (de très grande dimension et de caractère 'haute fréquence') disponibles en finance afin de détecter des structures cachées et en tirer partie pour l'évaluation du risque de marché ou la gestion de portefeuille par exemple. Ce point de vue est illustré par la brève présentation des travaux menés en ce sens en collaboration avec Skander Slim dans cette troisième partie. Ces dernières années, j'ai eu l'opportunité de pouvoir rencontrer des mathématiciens appliqués et des scientifiques travaillant dans d'autres domaines, pouvant également bénéficier des avancées de la modélisation probabiliste et des méthodes statistiques. J'ai pu ainsi aborder des applications relatives à la toxicologie, plus précisément au problème de l'évaluation des risque de contamination par voie alimentaire, lors de mon année de délégation auprès de l'Institut National de la Recherche Agronomique au sein de l'unité Metarisk, unité pluridisciplinaire entièrement consacrée à l'analyse du risque alimentaire. J'ai pu par exemple utiliser mes compétences dans le domaine de la modélisation maarkovienne afin de proposer un modèle stochastique décrivant l'évolution temporelle de la quantité de contaminant présente dans l'organisme (de manère à prendre en compte à la fois le phénomène d'accumulation du aux ingestions successives et la pharmacocinétique propre au contaminant régissant le processus d'élimination) et des méthodes d'inférence statistique adéquates lors de travaux en collaboration avec Patrice Bertail et Jessica Tressou. Cette direction de recherche se poursuit actuellement et l'on peut espérer qu'elle permette à terme de fonder des recommandations dans le domaine de la santé publique. Par ailleurs, j'ai la chance de pouvoir travailler actuellement avec Hector de Arazoza, Bertran Auvert, Patrice Bertail, Rachid Lounes et Viet-Chi Tran sur la modélisation stochastique de l'épidémie du virus VIH à partir des données épidémiologiques recensées sur la population de Cuba, lesquelles constituent l'une des bases de données les mieux renseignées sur l'évolution d'une épidémie de ce type. Et bien que ce projet vise essentiellement à obtenir un modèle numérique (permettant d'effectuer des prévisions quant à l'incidence de l'épidémie à court terme, de manière à pouvoir planifier la fabrication de la quantité d'anti-rétroviraux nécéssaire par exemple), il nous a conduit à aborder des questions théoriques ambitieuses, allant de l'existence d'une mesure quasi-stationnaire décrivant l'évolution à long terme de l'épidémie aux problèmes relatifs au caractère incomplet des données épidémiologiques disponibles. Il m'est malheureusement impossible d'évoquer ces questions ici sans risquer de les dénaturer, la présentation des problèmes mathématiques rencontrés dans ce projet mériterait à elle seule un rapport entier.

Preface

The present report surveys the essentials of my research activity since my PhD thesis [53], which was mainly devoted to extend the use of recent advances in *Computational Harmonic Analysis* (such as wavelet analysis) for adaptive nonparametric estimation methods in the i.i.d. setting to statistical estimation based on Markovian data. As explained at length in [123], certain concentration of measure properties (*i.e.* deviation probability and moment inequalities over functional classes, specifically tailored for nonlinear approximation) are crucially required for taking advantages of these analytical tools in statistical settings and getting estimation procedures with convergence rates surpassing the ones of older methods. In [53] (see also [54], [55] and [56]), the *regenerative method* (refer to [185]), consisting in dividing Harris Markov sample paths into asymptotically i.i.d. blocks, has been crucially exploited for establishing the required probabilistic results, the long term behavior of Markov processes being governed by certain renewal processes (the blocks being actually determined by renewal times). But having constructed an estimator, estimation of the accuracy (measured by the variance, particular quantiles or any functional of the distribution function) of the computed statistic is next of crucial importance. In this respect and beyond its practical simplicity (it consists in resampling data by making i.i.d. draws in the original data sample and recompute the statistic from the bootstrap data sample), the *bootstrap* is known to have major theoretical advantages over asymptotic normal approximation in the i.i.d. setting (it automatically approximates the second order structure in the Edgeworth expansion of the statistic distribution). I then turned naturally to the problem of extending the popular bootstrap procedure to markovian data. Through the works I and Patrice Bertail have jointly carried out, the regenerative method was revealed to be not solely a powerful analytical tool for proving probabilistic limit theorems or inequalities, but also to be of practical use for statistical estimation: our proposed bootstrap generalization is based on the resampling of (a random number of) regeneration data blocks (or of approximation of the latter) so as to mimick the renewal structure of the data. This method has also been shown to be advantageous for many other statistical purposes. And the first part of the report strives to present the principle of regeneration-based statistical methods for Harris Markov chains, as well as some of the various results obtained this way, in a comprehensive manner.

The second part of the report is devoted to the problem of learning how to order instances, instead of classifying them only, in a supervised setting. This difficult problem is of practical importance in many areas, ranging from medical diagnosis to information retrieval (IR) and asks challenging theoretical and algorithmic questions, with no entirely satisfactory answers yet. A possible approach to this subject consists in reducing the problem to a pairwise classification problem, as suggested by a popular criterion (namely, the *AUC criterion*) widely used for evaluating the pertinence of an ordering. In this context some results have been obtained in a joint work with Gabor Lugosi and Nicolas Vayatis, involving the study of U-processes: the major novelty consisting in the fact that here natural estimates of the risk are of the form of a U-statistic. However, in many applications such as IR, only top ranked instances are effectively scanned and a criterion corresponding to such local ranking problems as well as methods for computing optimal ordering rules with respect to the latter are crucially needed. Further developments in this direction have

been considered in a (continuing) series of works in collaboration with Nicolas Vayatis.

Finally, the last part of the report reflects my interest in practical applications of probabilistic concepts and statistical tools. My personal background lead me to consider first applications in finance. Although historical approaches are not preferred in this domain, I have been progressively convinced that nonparametric statistics could play a major role in analyzing the massive (of very large dimension and high-frequency) financial data for detecting hidden structure in the latter and gaining advantage of the latter in risk assesment or portfolio selection for instance. As an illustration, the works I have carried out with Skander Slim in that direction are described in a word in this third part. Recently, I also happened to meet applied mathematicians or scientists working in other fields, which may naturally interface with applied probability ans statistics. Hence, applications to Toxicology, and in particular to toxic chemicals dietary exposure, has also been one of my concern this last year, which I have spent in the pluridisciplinary research unity Metarisk of the National Research Agronomy Institute, entirely dedicated to dietary risk analysis. I could thus make use of my skills in Markov modelling for proposing a stochastic model describing the temporal evolution of the total body burden of chemical (in a way that both the toxicokinetics and the dietary behavior may be taken into account) and adequate inference methods for the latter in a joint work with P. Bertail and J. Tressou. This line of research is still going on and will hopefully provide practical insight and guidance for dietary contamination control in public health practice. It is also briefly presented in this last part. Besides, I have the great opportunity to work currently on the modelling of the AIDS epidemic with H. de Arazoza, B. Auvert, P. Bertail, R. Lounes and C. Tran based on the cuban epidemic data available, which form one of the most informed database on any HIV epidemic. While such a research project (taking place in the framework of the ACI-NIM "Epidemic Modelling") aims at providing a numerical model (for computing incidence predictions on short horizons for instance, so as to plan the quantity of antiretrovirals required), it also poses very challenging probabilistic and statistical problems, ranging from the proof for the existence of a quasi-stationary distribution describing the long term behavior of the epidemic to the difficulties encountered due to the incomplete character of the epidemic data available. Unfortunately, they are not discussed here, presenting the wide variety of mathematical problems arising in this project without denaturing it would have deserved a whole report.

Contents

Part I

Statistical Inference for Markov Chains

Abstract

Harris Markov chains make their appearance in many areas of statistical modeling, in particular in time series analysis. Recent years have seen a rapid growth of statistical techniques adapted to data exhibiting this particular pattern of dependence.

In this first part we endeavoured to present how renewal properties of Harris recurrent Markov chains or of specific extensions of the latter may be practically used for statistical inference in various settings. When the study of probabilistic properties of general Harris Markov chains may be classically carried out by using the regenerative method (see [185] and [182]), via the theoretical construction of regenerative extensions (see [150]), statistical methodologies may also be based on regeneration for general Harris chains. In the regenerative case, such procedures are implemented from data blocks corresponding to consecutive observed regeneration times for the chain. And the main idea for extending the application of these statistical techniques to general Harris chains X consists in generating first a sequence of approximate renewal times for a regenerative extension of X from data $X_1, \ldots, X_n$ and the parameters of a minorization condition satisfied by its transition probability kernel, and then applying the latter techniques to the data blocks determined by these pseudo-regeneration times as if they were exact regeneration blocks.

Numerous applications of this estimation principle may be considered in both the stationary and nonstationary (including the null recurrent case) frameworks. In Chapter 1, key concepts of the Markov chain theory as well as some basic notions about the regenerative method and the Nummelin splitting technique are briefly recalled. This preliminary chapter also presents and discusses how to practically construct (approximate) regeneration data blocks, on which statistical procedures that will be described in the sequel are based. Then Chapters 2 and 3 deal with some important procedures based on (approximate) regeneration data blocks, from both practical and theoretical viewpoints, for the following topics: mean and variance estimation, confidence intervals, Bootstrap, U-statistics, robust estimation and statistical study of extreme values. Finally, some concluding remarks are collected in Chapter 4 and further lines of research are sketched.

Preliminaries

0.1 Markov chain analysis via renewal theory

Renewal theory plays a key role in the analysis of the asymptotic structure of many kinds of stochastic processes, and especially in the development of asymptotic properties of general irreducible Markov chains. The underlying ground consists in the fact that limit theorems proved for sums of independent random vectors may be easily extended to regenerative random processes, that is to say random processes that may be decomposed at random times, called *regeneration times*, into a sequence of mutually independent blocks of observations, namely *regeneration cycles* (see [185] and [182]). The method based on this principle is traditionally called the *regenerative method*. Harris chains that possess an atom, i.e. a Harris set on which the transition probability kernel is constant, are special cases of regenerative processes and so directly fall into the range of application of the regenerative method (Markov chains with discrete state space as well as many markovian models widely used in operational research for modeling storage or queuing systems are remarkable examples of atomic chains). The theory developed in [150] (and in parallel the closely related concepts introduced in [12]) showed that general Markov chains could all be considered as regenerative in a broader sense (*i.e.* in the sense of the existence of a theoretical regenerative extension for the chain, see §1.2.2), as soon as the Harris recurrence property is satisfied. Hence this theory made the regenerative method applicable to the whole class of Harris Markov chains and allowed to carry over many limit theorems to Harris chains such as LLN, CLT, LIL or Edgeworth expansions.

In many cases, parameters of interest for a Harris Markov chain may be thus expressed in terms of regeneration cycles. While, for atomic Markov chains, statistical inference procedures may be then based on a random number of observed regeneration data blocks, in the general Harris recurrent case the regeneration times are theoretical and their occurrence cannot be determined by examination of the data only. Although the *Nummelin splitting technique* for constructing regeneration times has been introduced as a theoretical tool for proving probabilistic results such as limit theorems or probability and moment inequalities in the markovian framework, it is nevertheless possible to make a practical use of the latter for extending regeneration-based statistical tools. Our proposal consists in an empirical method for building approximatively a realization drawn from a Nummelin extension of the chain with a regeneration set and then recovering *approximate regeneration data blocks*. As will be shown in the next two chapters, though the implementation of the latter method requires some prior knowledge about the behaviour of the chain and crucially relies on the computation of a consistent estimate of its transition kernel, this methodology allows for numerous statistical applications.

In section 0.2, notations are set out and key concepts of the Markov chain theory as well as some basic notions about the regenerative method and the Nummelin splitting technique are recalled. Section 0.3 presents how to practically construct (approximate) regeneration data blocks, on which statistical procedures presented in the next two chapters are based. Computational issues related to this construction are discussed in section 0.4.

0.2 Theoretical background

We first set out the notations and recall a few definitions concerning the communication structure and the stochastic stability of Markov chains (for further detail, refer to [167] or [148]). Let $X = (X_n)_{n\in\mathbb{N}}$ be a 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 $n \in \mathbb{N}$, 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, $\mathbb{P}_\nu$ (respectively $\mathbb{P}_x$ for x in E) denotes the probability measure on the underlying probability space such that $X_0 \sim \nu$ (resp. $X_0 = x$), $\mathbb{E}_\nu(.)$ the $\mathbb{P}_\nu$-expectation (resp. $\mathbb{E}_x(.)$ the $\mathbb{P}_x$-expectation), $\mathbb{I}\{\mathcal{A}\}$ denotes the indicator function of the event $\mathcal{A}$ and $\Rightarrow$ the convergence in distribution.

For completeness, recall the following notions. The first one formalizes the idea of communicating structure between specific subsets, while the second one considers the set of time points at which such communication may occur.

- The chain is *irreducible* if there exists a σ-finite measure ψ such that for all set $B \in \mathcal{E}$, when $\psi(B) > 0$, the chain visits B with strictly positive probability, no matter what the starting point.
- Assuming ψ-irreducibility, there is $d' \in \mathbb{N}^*$ and disjoints sets $D_1, ..., D_{d'}$ $(D_{d'+1} = D_1)$ weighted by ψ such that $\psi(E \setminus \cup_{1\leqslant i\leqslant d'} D_i) = 0$ and $\forall x \in D_i$, $\Pi(x, D_{i+1}) = 1$. The g.c.d. d of such integers is the *period* of the chain, which is said *aperiodic* if $d = 1$.

A measurable set B is *Harris recurrent* for the chain if for any $x \in B$, $\mathbb{P}_x(\sum_{n=1}^{\infty} \mathbb{I}\{X_n \in B\} = \infty) = 1$. The chain is said *Harris recurrent* if it is ψ-irreducible and every measurable set B such that $\psi(B) > 0$ is Harris recurrent. When the chain is Harris recurrent, we have the property that $\mathbb{P}_x(\sum_{n=1}^{\infty} \mathbb{I}\{X_n \in B\} = \infty) = 1$ for any $x \in E$ and any $B \in \mathcal{E}$ such that $\psi(B) > 0$.

A probability measure μ on E is said invariant for the chain when $\mu\Pi = \mu$, where $\mu\Pi(dy) = \int_{x\in E} \mu(dx)\Pi(x, dy)$. An irreducible chain is said *positive recurrent* when it admits an invariant probability (it is then unique).

Now we recall some basics concerning the regenerative method and its application to the analysis of the behaviour of general Harris chains via the Nummelin splitting technique (refer to [151] for further detail).

0.2.1 Regenerative Markov chains

Assume that the chain is ψ-irreducible and 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. Denote by $\tau_A = \tau_A(1) = \inf\{n \geq 1,\ X_n \in A\}$ 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 and by $\mathbb{E}_A(.)$ the expectation conditioned on $X_0 \in A$. Assume further that the chain is Harris recurrent, the probability of returning infinitely often to the atom A is thus equal to one, no matter what the starting point. Then, it follows from the *strong Markov property* that, for any initial distribution ν, the sample paths of the chain may be divided into i.i.d. blocks of random length corresponding to consecutive visits to A:

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

taking their values in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$. The sequence $(\tau_A(j))_{j\geqslant 1}$ defines successive times at which the chain forgets its past, called *regeneration times*. We point out that the class of

atomic Markov chains contains not only chains with a countable state space (for the latter, any recurrent state is an accessible atom), but also many specific Markov models arising from the field of operational research (see [8] for regenerative models involved in queuing theory, as well as the examples given in §2.2.3 of Chap. 2). When an accessible atom exists, the *stochastic stability* properties of the chain amount to properties concerning the speed of return time to the atom only. For instance, in this framework, the following result, known as Kac's theorem, holds (*cf* Theorem 10.2.2 in [148]).

Theorem 1 *The chain X is positive recurrent iff $\mathbb{E}_A(\tau_A) < \infty$. The (unique) invariant probability distribution μ is then the Pitman's occupation measure given by*

$$\mu(B) = \mathbb{E}_A(\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i \in B\})/\mathbb{E}_A(\tau_A), \text{ for all } B \in \mathcal{E}.$$

For atomic chains, limit theorems can be derived from the application of the corresponding results to the i.i.d. blocks $(\mathcal{B}_n)_{n\geqslant 1}$ (see [52] and the references therein). One may refer for example to [148] for the LLN, CLT, LIL, [35] for the Berry-Esseen theorem, [137], [138], [139] and [19] for other refinements of the CLT. The same technique can also be applied to establish moment and probability inequalities, which are not asymptotic results (see [55] and [25]). As mentioned above, these results are established from hypotheses related to the distribution of the $\mathcal{B}_n$'s. The following assumptions shall be involved throughout the next two chapters. Let $\kappa > 0$, $f : E \to \mathbb{R}$ be a measurable function and ν be a probability distribution on $(E, \mathcal{E})$.

Regularity conditions:

$$\mathcal{H}_0(\kappa) : \ \mathbb{E}_A(\tau_A^\kappa) < \infty \text{ and } \mathcal{H}_0(\kappa, \ \nu) : \ \mathbb{E}_\nu(\tau_A^\kappa) < \infty.$$

Block-moment conditions:

$$\mathcal{H}_1(\kappa, \ f) : \ \mathbb{E}_A((\sum_{i=1}^{\tau_A} |f(X_i)|)^\kappa) < \infty,$$
$$\mathcal{H}_1(\kappa, \ \nu, \ f) : \ \mathbb{E}_\nu((\sum_{i=1}^{\tau_A} |f(X_i)|)^\kappa) < \infty.$$

We point out that conditions $\mathcal{H}_0(\kappa)$ and $\mathcal{H}_1(\kappa, f)$ do not depend on the accessible atom chosen : if they hold for a given accessible atom A, they are also fulfilled for any other accessible atom (see Chapter 11 in [148]). Besides, the relationship between the "block moment" conditions and the rate of decay of mixing coefficients has been investigated in [36]: for instance, $\mathcal{H}_0(\kappa)$ (as well as $\mathcal{H}_1(\kappa, f)$ when f is bounded) is typically fulfilled as soon as the strong mixing coefficients sequence decreases at an arithmetic rate $n^{-\rho}$, for some $\rho > \kappa - 1$.

0.2.2 General Harris recurrent chains

Regenerative extension. We now recall the *splitting technique* introduced in [150] for extending the probabilistic structure of the chain in order to construct an artificial regeneration set in the general Harris case. It relies on the notion of *small set*. Recall that, for a Markov chain valued in a state space $(E, \mathcal{E})$ with probability Π, a set $S \in \mathcal{E}$ is said to be *small* if there exist $m \in \mathbb{N}^*$, $\delta > 0$ and a probability measure Γ supported by S s.t., for all $x \in S$, $B \in \mathcal{E}$,

$$\Pi^m(x, B) \geq \delta\Gamma(B), \tag{1}$$

denoting by Π^m the m-th iterate of Π. When this holds, we say that the chain satisfies the *minorization condition* $\mathcal{M}(m, S, \delta, \Gamma)$. We emphasize that accessible small sets always exist for ψ-irreducible chains: any set $B \in \mathcal{E}$ such that $\psi(B) > 0$ actually contains such a set (*cf* [118]). Now let us precise how to construct the atomic chain onto which the initial chain X is embedded, from a set on which an iterate Π^m of the transition probability is uniformly bounded below. Suppose that X satisfies $\mathcal{M} = \mathcal{M}(m, S, \delta, \Gamma)$ for $S \in \mathcal{E}$ such that $\psi(S) > 0$. Even if it entails replacing the chain $(X_n)_{n\in\mathbb{N}}$ by the chain $\left((X_{nm}, ..., X_{n(m+1)-1}\right))_{n\in\mathbb{N}}$, we suppose $m = 1$. The sample space is expanded so as to define a sequence $(Y_n)_{n\in\mathbb{N}}$ of independent Bernoulli r.v.'s with parameter δ by defining the joint distribution $\mathbb{P}_{\nu,\mathcal{M}}$ whose construction relies on the following randomization of the transition probability Π each time the chain hits S (note that it happens a.s. since the chain is Harris recurrent and $\psi(S) > 0$). If $X_n \in S$ and

- if $Y_n = 1$ (which happens with probability $\delta \in]0, 1[$), then $X_{n+1} \sim \Gamma$,
- if $Y_n = 0$, (which happens with probability $1 - \delta$), then $X_{n+1} \sim (1 - \delta)^{-1}(\Pi(X_{n+1}, .) - \delta\Gamma(.))$.

Set $Ber_\delta(\beta) = \delta\beta + (1 - \delta)(1 - \beta)$ for $\beta \in \{0, 1\}$. We now have constructed the *split chain*, a bivariate chain $X^{\mathcal{M}} = ((X_n, Y_n))_{n\in\mathbb{N}}$, valued in $E \times \{0, 1\}$ with transition kernel $\Pi_{\mathcal{M}}$ defined by

- for any $x \notin S$, $B \in \mathcal{E}$, β and β' in $\{0, 1\}$,

$$\Pi_{\mathcal{M}}\left((x, \beta), B \times \left\{\beta'\right\}\right) = \Pi(x, B) \times Ber_\delta(\beta'),$$

- for any $x \in S$, $B \in \mathcal{E}$, β' in $\{0, 1\}$,

$$\begin{aligned}\Pi_{\mathcal{M}}\left((x, 1), B \times \left\{\beta'\right\}\right) &= \Gamma(B) \times Ber_\delta(\beta'),\\ \Pi_{\mathcal{M}}\left((x, 0), B \times \left\{\beta'\right\}\right) &= (1 - \delta)^{-1}(\Pi(x, B) - \delta\Gamma(B)) \times Ber_\delta(\beta').\end{aligned}$$

Basic assumptions. The whole point of the construction consists in the fact that $S \times \{1\}$ is an atom for the split chain $X^{\mathcal{M}}$, which inherits all the communication and stochastic stability properties from X (irreducibility, Harris recurrence,...), in particular (for the case $m = 1$ here) the blocks constructed for the split chain are independent. Hence the splitting method enables to extend the regenerative method, and so to establish all of the results known for atomic chains, to general Harris chains. It should be noticed that if the chain X satisfies $\mathcal{M}(m, S, \delta, \Gamma)$ for $m > 1$, the resulting blocks are not independent anymore but 1-dependent, a form of dependence which may be also easily handled. For simplicity 's sake, we suppose in what follows that condition $\mathcal{M}$ is fulfilled with $m = 1$, we shall also omit the subscript $\mathcal{M}$ and abusively denote by $\mathbb{P}_\nu$ the extensions of the underlying probability we consider. The following assumptions, involving the speed of return to the small set S shall be used throughout this part of the report. Let $\kappa > 0$, $f : E \to \mathbb{R}$ be a measurable function and ν be a probability measure on $(E, \mathcal{E})$.

Regularity conditions:

$$\mathcal{H}_0'(\kappa) : \ \sup_{x\in S} \mathbb{E}_x(\tau_S^\kappa) < \infty \text{ and } \mathcal{H}_0'(\kappa, \ \nu) : \mathbb{E}_\nu(\tau_S^\kappa) < \infty.$$

Block-moment conditions:

$$\mathcal{H}_1'(\kappa, \ f) : \sup_{x\in S} \mathbb{E}_x((\sum_{i=1}^{\tau_S} |f(X_i)|)^\kappa) < \infty,$$

$$\mathcal{H}_1'(\kappa,\ f,\ \nu):\ \mathbb{E}_\nu((\sum_{i=1}^{\tau_S}|f(X_i)|)^\kappa)<\infty.$$

It is noteworthy that assumptions $\mathcal{H}_0'(\kappa)$ and $\mathcal{H}_1'(\kappa,\ f)$ do not depend on the choice of the small set S (if they are checked for some accessible small set S, they are fulfilled for all accessible small sets *cf* §11.1 in [148]). Note also that in the case when $\mathcal{H}_0'(\kappa)$ (resp., $\mathcal{H}_0'(\kappa,\ \nu)$) is satisfied, $\mathcal{H}_1'(\kappa,\ f)$ (resp., $\mathcal{H}_1'(\kappa,\ f,\ \nu)$) is fulfilled for any bounded f. Moreover, recall that positive recurrence, conditions $\mathcal{H}_1'(\kappa)$ and $\mathcal{H}_1'(\kappa,\ f)$ may be practically checked by using test functions methods (*cf* [121], [191]). In particular, it is well known that such block moment assumptions may be replaced by drift criteria of Lyapounov's type (refer to Chapter 11 in [148] for further details on such conditions and many illustrating examples).

We recall finally that such assumptions on the initial chain classically imply the desired conditions for the split chain: as soon as X fulfills $\mathcal{H}_0'(\kappa)$ (resp., $\mathcal{H}_0'(\kappa,\ \nu)$, $\mathcal{H}_1'(\kappa,\ f)$, $\mathcal{H}_1'(\kappa,\ f,\ \nu)$), $X^{\mathcal{M}}$ satisfies $\mathcal{H}_0(\kappa)$ (resp., $\mathcal{H}_0(\kappa,\ \nu)$, $\mathcal{H}_1(\kappa,\ f)$, $\mathcal{H}_1(\kappa,\ f,\ \nu)$).

The distribution of $(Y_1,...,Y_n)$ conditioned on $(X_1,...,X_{n+1})$. As will be shown in the next section, the statistical methodology for Harris chains we propose is based on approximating the conditional distribution of the binary sequence $(Y_1,...,Y_n)$ given $X^{(n+1)}=(X_1,...,X_{n+1})$. We thus precise the latter. Let us assume further that the family of the conditional distributions $\{\Pi(x,dy)\}_{x\in E}$ and the initial distribution ν are dominated by a σ-finite measure λ of reference, so that $\nu(dy)=f(y)\lambda(dy)$ and $\Pi(x,dy)=p(x,y)\lambda(dy)$, for all $x\in E$. Notice that the minorization condition entails that Γ is absolutely continuous with respect to λ too, and that

$$p(x,y)\geq\delta\gamma(y),\ \lambda(dy)\text{ a.s.}\tag{2}$$

for any $x\in S$, with $\Gamma(dy)=\gamma(y)dy$. The distribution of $Y^{(n)}=(Y_1,...,\ Y_n)$ conditionally to $X^{(n+1)}=(x_1,...,x_{n+1})$ is then the tensor product of Bernoulli distributions given by: for all $\beta^{(n)}=(\beta_1,...,\beta_n)\in\{0,1\}^n$, $x^{(n+1)}=(x_1,...,x_{n+1})\in E^{n+1}$,

$$\mathbb{P}_\nu\left(Y^{(n)}=\beta^{(n)}\mid X^{(n+1)}=x^{(n+1)}\right)=\prod_{i=1}^n\mathbb{P}_\nu(Y_i=\beta_i\mid X_i=x_i,\ X_{i+1}=x_{i+1}),$$

with, for $1\leqslant i\leqslant n$,

$$\begin{aligned}\mathbb{P}_\nu(Y_i=1\mid X_i=x_i,\ X_{i+1}=x_{i+1})&=\delta,\ \text{if } x_i\notin S,\\ \mathbb{P}_\nu(Y_i=1\mid X_i=x_i,\ X_{i+1}=x_{i+1})&=\frac{\delta\gamma(x_{i+1})}{p(x_i,x_{i+1})},\ \text{if } x_i\in S.\end{aligned}$$

Roughly speaking, conditioned on $X^{(n+1)}$, from $i=1$ to n, Y_i is drawn from the Bernoulli distribution with parameter δ, unless X has hit the small set S at time i: in this case Y_i is drawn from the Bernoulli distribution with parameter $\delta\gamma(X_{i+1})/p(X_i,X_{i+1})$. We denote by $\mathcal{L}^{(n)}(p,S,\delta,\gamma,x^{(n+1)})$ this probability distribution.

0.3 Dividing the trajectory into (pseudo-) regeneration cycles

In the preceding section, we recalled the Nummelin approach for the theoretical construction of regeneration times in the Harris framework. Here we now consider the problem of approximating these random times from data sets in practice and propose a basic preprocessing technique, on which estimation methods we shall discuss further are based.

0.3.1 Regenerative case

Let us suppose we observed a sample path $X_1, ..., X_n$ of length n drawn from the chain X. In the regenerative case, when an atom A for the chain is *a priori* known, regeneration blocks are naturally obtained by simply examining the data, as follows.

Algorithm 1 *(Regeneration blocks construction)*

1. *Count the number of visits* $l_n = \sum_{i=1}^{n} \mathbb{I}\{X_i \in A\}$ *to* A *up to time* n.
2. *Divide the observed trajectory* $X^{(n)} = (X_1, ..., X_n)$ *into* $l_n + 1$ *blocks corresponding to the pieces of the sample path between consecutive visits to the atom* A,

$$\mathcal{B}_0 = (X_1, ..., X_{\tau_A(1)}),\ \mathcal{B}_1 = (X_{\tau_A(1)+1}, ..., X_{\tau_A(2)}), ...,$$
$$\mathcal{B}_{l_n-1} = (X_{\tau_A(l_n-1)+1}, ..., X_{\tau_A(l_n)}),\ \mathcal{B}_{l_n}^{(n)} = (X_{\tau_A(l_n)+1}, ..., X_n),$$

   with the convention $\mathcal{B}_{l_n}^{(n)} = \emptyset$ *when* $\tau_A(l_n) = n$.

3. *Drop the first block* $\mathcal{B}_0$, *as well as the last one* $\mathcal{B}_{l_n}^{(n)}$, *when non-regenerative (i.e. when* $\tau_A(l_n) < n$*).*

The regeneration blocks construction is illustrated by Fig. 1 in the case of a random walk on the half line $\mathbb{R}^+$ with $\{0\}$ as an atom.

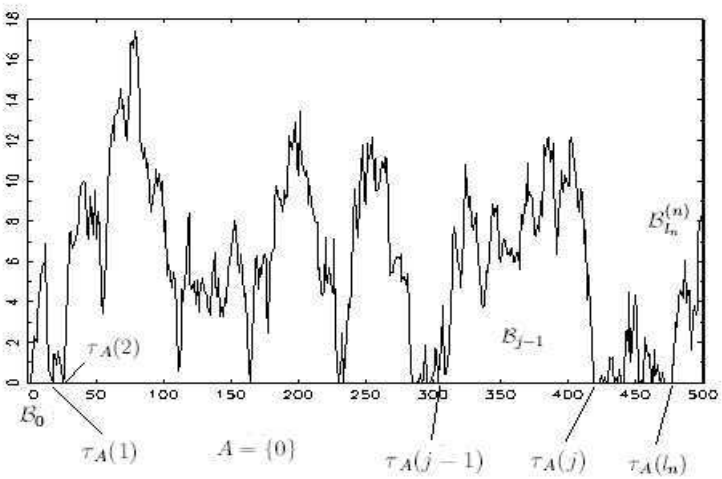

Figure 1: Dividing the trajectory of a random walk on the half line into cycles.

0.3.2 General Harris case

The principle. Suppose now that observations $X_1, ..., X_{n+1}$ are drawn from a Harris chain X satisfying the assumptions of §1.2.2 (refer to the latter paragraph for the notations). If we were able to generate binary data $Y_1, ..., Y_n$, so that $X^{\mathcal{M}\ (n)} = ((X_1, Y_1), ..., (X_n, Y_n))$ be a realization of

the split chain $X^{\mathcal{M}}$ described in §1.2.2, then we could apply the *regeneration blocks construction* procedure to the sample path $X^{\mathcal{M}\ (n)}$. In that case the resulting blocks are still independent since the split chain is atomic. Unfortunately, knowledge of the transition density $p(x,y)$ for $(x,y) \in S^2$ is required to draw practically the Y_i's this way. In [22] a method relying on a preliminary estimation of the "nuisance parameter" $p(x,y)$ is proposed. More precisely, it consists in approximating the splitting construction by computing an estimator $p_n(x,y)$ of $p(x,y)$ using data $X_1, ..., X_{n+1}$, and to generate a random vector $(\widehat{Y}_1, ..., \widehat{Y}_n)$ conditionally to $X^{(n+1)} = (X_1, ..., X_{n+1})$, from distribution $\mathcal{L}^{(n)}(p_n, S, \delta, \gamma, X^{(n+1)})$, which approximates in some sense the conditional distribution $\mathcal{L}^{(n)}(p, S, \delta, \gamma, X^{(n+1)})$ of $(Y_1, ..., Y_n)$ for given $X^{(n+1)}$. Our method, which we call *approximate regeneration blocks construction (ARB construction* in abbreviated form) amounts then to apply the *regeneration blocks construction* procedure to the data $((X_1, \widehat{Y}_1), ..., (X_n, \widehat{Y}_n))$ as if they were drawn from the atomic chain $X^{\mathcal{M}}$. In spite of the necessary consistent transition density estimation step, we shall show in the sequel that many statistical procedures, that would be consistent in the ideal case when they would be based on the regeneration blocks, remain asymptotically valid when implemented from the approximate data blocks. For given parameters $(\delta,\ S,\ \gamma)$ (see §1.4.1 for a data driven choice of these parameters), the approximate regeneration blocks are constructed as follows.

Algorithm 2 *(Approximate regeneration blocks construction)*

1. *From the data* $X^{(n+1)} = (X_1, ..., X_{n+1})$*, compute an estimate* $p_n(x,y)$ *of the transition density such that* $p_n(x,y) \geq \delta\gamma(y)$, $\lambda(dy)$ *a.s., and* $p_n(X_i, X_{i+1}) > 0$, $1 \leqslant i \leqslant n$.

2. *Conditioned on* $X^{(n+1)}$*, draw a binary vector* $(\widehat{Y}_1, ..., \widehat{Y}_n)$ *from the distribution estimate* $\mathcal{L}^{(n)}(p_n, S, \delta, \gamma, X^{(n+1)})$*. It is sufficient in practice to draw the* $\widehat{Y}_i$*'s at time points* i *when the chain visits the set* S *(i.e. when* $X_i \in S$*), since at these times and at these times only the split chain may regenerate. At such a time point* i*, draw* $\widehat{Y}_i$ *according to the Bernoulli distribution with parameter* $\delta\gamma(X_{i+1})/p_n(X_i, X_{i+1}))$*.*

3. *Count the number of visits* $\widehat{l}_n = \sum_{i=1}^n \mathbb{I}\{X_i \in S, \widehat{Y}_i = 1)$ *to the set* $A_{\mathcal{M}} = S \times \{1\}$ *up to time* n *and divide the trajectory* $X^{(n+1)}$ *into* $\widehat{l}_n + 1$ *approximate regeneration blocks corresponding to the successive visits of* $(X, \widehat{Y})$ *to* $A_{\mathcal{M}}$*,*

$$\widehat{\mathcal{B}}_0 = (X_1, ..., X_{\widehat{\tau}_{A_{\mathcal{M}}}(1)}),\ \widehat{\mathcal{B}}_1 = (X_{\widehat{\tau}_{A_{\mathcal{M}}}(1)+1}, ..., X_{\widehat{\tau}_{A_{\mathcal{M}}}(2)}), ...,$$
$$\widehat{\mathcal{B}}_{\widehat{l}_n - 1} = (X_{\widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n - 1)+1}, ..., X_{\widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n)}),\ \widehat{\mathcal{B}}^{(n)}_{\widehat{l}_n} = (X_{\widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n)+1}, ..., X_{n+1}),$$

   where $\widehat{\tau}_{A_{\mathcal{M}}}(1) = \inf\{n \geqslant 1, X_n \in S, \widehat{Y}_n = 1\}$ *and* $\widehat{\tau}_{A_{\mathcal{M}}}(j+1) = \inf\{n > \widehat{\tau}_{A_{\mathcal{M}}}(j), X_n \in S, \widehat{Y}_n = 1\}$ *for* $j \geqslant 1$*.*

4. *Drop the first block* $\widehat{\mathcal{B}}_0$ *and the last one* $\widehat{\mathcal{B}}^{(n)}_{\widehat{l}_n}$ *when* $\widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n) < n$*.*

Such a division of the sample path is illustrated by Fig. 2 below: from a practical viewpoint the trajectory may only be cut when hitting the small set. At such a point, drawing a Bernoulli r.v. with the estimated parameter indicates whether one should cut here the time series trajectory or not. Of course, due to the dependence induced by the estimated transition density, the resulting blocks are not i.i.d. but, as will be shown later, are close (in some sense) to the true regeneration blocks which are i.i.d.

Next, the accuracy of this approximation in the Mallows distance's sense (which metric is a crucial tool for proving asymptotic validity of bootstrap methods, see [30]) is shown to depend mainly on the rate of the uniform convergence of $p_n(x,y)$ to $p(x,y)$ over $S \times S$.

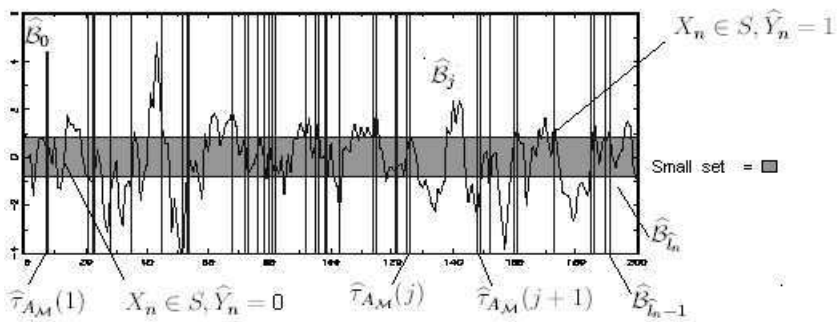

Figure 2: ARB construction for an AR(1) simulated time-series.

0.3.3 A coupling result for $(X_i, \widehat{Y}_i)_{1 \leqslant i \leqslant n}$ and $(X_i, Y_i)_{1 \leqslant i \leqslant n}$

We now state a result claiming that the distribution of $(X_i, \widehat{Y}_i)_{1 \leqslant i \leqslant n}$ gets closer and closer to the distribution of $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ in the sense of the Mallows distance (also known as the Kantorovich or Wasserstein metric in the probability literature) as $n \to \infty$. Hence, we express here the distance between the distributions P^Z and $P^{Z'}$ of two random sequences $Z = (Z_n)_{n \in \mathbb{N}}$ and $Z' = (Z'_n)_{n \in \mathbb{N}}$, taking their values in $\mathbb{R}^k$, by (see [160], p 76)

$$l_p(Z, Z') = l_p(P^Z, P^{Z'}) = \min \left\{ L_p(W, W') \, ; \; W \sim P^Z, \; W' \sim P^{Z'} \right\},$$

with $(L_p(W, W'))^{1/q} = \mathbb{E}\left[D^p(W, W')\right]$, where D denotes the metric on the space $\chi(\mathbb{R}^k) = (\mathbb{R}^k)^\infty$ defined by $D(w, w') = \sum_{k=0}^{\infty} 2^{-k} \|w_k - w'_k\|_{\mathbb{R}^k}$. For any w, w' in $\chi(\mathbb{R}^k)$ ($\|.\|_{\mathbb{R}^k}$ denoting the usual euclidian norm of $\mathbb{R}^k$). Thus, viewing the sequences $Z^{(n)} = (X_k, Y_k)_{1 \leqslant k \leqslant n}$ and $\widehat{Z}^{(n)} = (X_k, \widehat{Y}_k)_{1 \leqslant i \leqslant n}$ as the beginning segments of infinite series, we evaluate the deviation between the distribution $P^{(n)}$ of $Z^{(n)}$ and the distribution $\widehat{P}^{(n)}$ of $\widehat{Z}^{(n)}$ using $l_1(P^{(n)}, \widehat{P}^{(n)})$.

Theorem 2 *(Bertail & Clémençon, 2005b) Assume that*

- S *is chosen so that* $\inf_{x \in S} \phi(x) > 0$,
- p *is estimated by* p_n *at the rate* α_n *for the MSE when error is measured by the* L^∞ *loss over* S^2,

then

$$l_1(P^{(n)}, \widehat{P}^{(n)}) \leqslant (\delta \inf_{x \in S} \phi(x))^{-1} \alpha_n^{1/2}. \tag{3}$$

This theorem is established in [22] by exhibiting a specific coupling of $(X_i, \widehat{Y}_i)_{1 \leqslant i \leqslant n}$ and $(X_i, Y_i)_{1 \leqslant i \leqslant n}$. It is a crucial tool for deriving the results stated in the next two chapters. It also clearly shows that the closeness between the two distributions is tightly connected to the rate of convergence of the estimator $p_n(x, y)$ but also to the minorization condition parameters. This gives us some hints on how to choose the small set with a *data driven method* to obtain better finite sample results, as shall be shown in the following section.

0.4 Practical issues

0.4.1 Choosing the minorization condition parameters

Because the construction above is highly dependent on the minorization condition parameters chosen, we now discuss how to select the latter with a data-driven technique so as to construct enough blocks for computing meaningful statistics. As a matter of fact, the rates of convergence of the statistics we shall study in the sequel increase as the mean number of regenerative (or pseudo-regenerative) blocks, which depends on the size of the small set chosen (or more exactly, on how often the chain visits the latter in a trajectory of finite length) and how sharp is the lower bound in the minorization condition: the larger the size of the small set is, the smaller the uniform lower bound for the transition density. This leads us to the following trade-off. Roughly speaking, for a given realization of the trajectory, as one increases the size of the small set S used for the data blocks construction, one naturally increases the number of points of the trajectory that are candidates for determining a block (*i.e.* a cut in the trajectory), but one also decreases the probability of cutting the trajectory (since the uniform lower bound for $\{p(x,y)\}_{(x,y)\in S^2}$ then decreases). This gives an insight into the fact that better numerical results for statistical procedures based on the ARB construction may be obtained in practice for some specific choices of the small set, likely for choices corresponding to a maximum expected number of data blocks given the trajectory, that is

$$N_n(S) = \mathbb{E}_\nu(\sum_{i=1}^{n} \mathbb{I}\{X_i \in S, Y_i = 1\} \,|X^{(n+1)}).$$

Hence, when no prior information about the structure of the chain is available, here is a practical data-driven method for selecting the minorization condition parameters in the case when the chain takes real values. Consider a collection $\mathcal{S}$ of borelian sets S (typically compact intervals) and denote by $\mathcal{U}_S(dy) = \gamma_S(y).\lambda(dy)$ the uniform distribution on S, where $\gamma_S(y) = \mathbb{I}\{y \in S\}/\lambda(S)$ and λ is the Lebesgue measure on $\mathbb{R}$. Now, for any $S \in \mathcal{S}$, set $\delta(S) = \lambda(S).\inf_{(x,y)\in S^2} p(x,y)$. We have for any x, y in S, $p(x,y) \geq \delta(S)\gamma_S(y)$. In the case when $\delta(S) > 0$, the ideal criterion to optimize may be then expressed as

$$N_n(S) = \frac{\delta(S)}{\lambda(S)} \sum_{i=1}^{n} \frac{\mathbb{I}\{(X_i, X_{i+1}) \in S^2\}}{p(X_i, X_{i+1})}. \tag{4}$$

However, as the transition kernel $p(x,y)$ and its minimum over S^2 are unknown, a practical empirical criterion is obtained by replacing $p(x,y)$ by an estimate $p_n(x,y)$ and $\delta(S)$ by a lower bound $\delta_n(S)$ for $\lambda(S).p_n(x,y)$ over S^2 in expression (4). Once $p_n(x,y)$ is computed, calculate $\delta_n(S) = \lambda(S).\inf_{(x,y)\in S^2} p_n(x,y)$ and maximize thus the empirical criterion over $S \in \mathcal{S}$

$$\widehat{N}_n(S) = \frac{\delta_n(S)}{\lambda(S)} \sum_{i=1}^{n} \frac{\mathbb{I}\{(X_i, X_{i+1}) \in S^2\}}{p_n(X_i, X_{i+1})}. \tag{5}$$

More specifically, one may easily check at hand on many examples of real valued chains (see §2.2.3 for instance), that any compact interval $V_{x_0}(\varepsilon) = [x_0 - \varepsilon, x_0 + \varepsilon]$ for some well chosen $x_0 \in \mathbb{R}$ and $\varepsilon > 0$ small enough, is a small set, choosing γ as the density of the uniform distribution on $V_{x_0}(\varepsilon)$. For practical purpose, one may fix x_0 and perform the optimization over $\varepsilon > 0$ only (see [22]) but both x_0 and ε may be considered as tuning parameters. A possible numerically feasible selection rule could rely then on searching for (x_0, ε) on a given pre-selected grid $\mathcal{G} = \{(x_0(k), \varepsilon(l)),$ $1 \leqslant k \leqslant K$, $1 \leqslant l \leqslant L\}$ s.t. $\inf_{(x,y)\in V_{x_0}(\varepsilon)^2} p_n(x,y) > 0$ for any $(x_0, \varepsilon) \in \mathcal{G}$.

Algorithm 3 *(ARB construction with empirical choice of the small set)*

1. *Compute an estimator* $p_n(x, y)$ *of* $p(x, y)$.

2. *For any* $(x_0, \varepsilon) \in \mathcal{G}$*, compute the estimated expected number of pseudo-regenerations:*

$$\widehat{N}_n(x_0, \varepsilon) = \frac{\delta_n(x_0, \varepsilon)}{2\varepsilon} \sum_{i=1}^{n} \frac{\mathbb{I}\{(X_i, X_{i+1}) \in V_{x_0}(\varepsilon)^2\}}{p_n(X_i, X_{i+1})},$$

   with $\delta_n(x_0, \varepsilon) = 2\varepsilon . \inf_{(x,y)\in V_{x_0}(\varepsilon)^2} p_n(x, y)$.

3. *Pick* (x_0^*, ε^*) *in* $\mathcal{G}$ *maximizing* $\widehat{N}_n(x_0, \varepsilon)$ *over* $\mathcal{G}$*, corresponding to the set* $S^* = [x_0^* - \varepsilon^*,\ x_0^* + \varepsilon^*]$ *and the minorization constant* $\delta_n^* = \delta_n(x_0^*, \varepsilon^*)$.

4. *Apply Algorithm 2 for ARB construction using* S^*, δ_n^* *and* p_n.

Remark 1 Numerous consistent estimators of the transition density of Harris chains have been proposed in the literature. Refer to [172], [173], [174] [170], [33], [76], [158], [10] or [54] for instance in positive recurrent cases, [122] in specific null recurrent cases.

This method is illustrated by Fig. 3 in the case of an AR(1) model: $X_{i+1} = \alpha X_i + \varepsilon_{i+1}$, $i \in \mathbb{N}$, with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, $\alpha = 0.95$ and $X_0 = 0$, for a trajectory of length $n = 200$. Taking $x_0 = 0$ and letting ε grow, the expected number regeneration blocks is maximum for ε^* close to 0.9. The true minimum value of $p(x, y)$ over the corresponding square is actually $\delta = 0.118$. The first graphic in this panel shows the *Nadaraya-Watson estimator*

$$p_n(x, y) = \frac{\sum_{i=1}^{n} K(h^{-1}(x - X_i))K(h^{-1}(y - X_{i+1}))}{\sum_{i=1}^{n} K(h^{-1}(x - X_i))}, \tag{6}$$

computed from the gaussian kernel $K(x) = (2\pi)^{-1} \exp(-x^2/2)$ with an optimal bandwidth $h \sim n^{-1/5}$. The second one plots $\widehat{N}_n(\varepsilon)$ as a function of ε. The next one indicates the set S^* corresponding to our empirical rule, while the last one displays the *optimal* ARB construction.

Note finally that other approaches may be considered for determining practically small sets and establishing accurate minorization conditions, which conditions do not necessarily involve uniform distributions besides. Refer for instance to [168] for Markov diffusion processes.

0.4.2 A two-split version of the ARB construction

When carrying out the theoretical study of statistical methods based on the ARB construction, one must deal with difficult problems arising from the dependence structure in the set of the resulting data blocks, due to the preliminary estimation step. Such difficulties are somehow similar as the ones that one traditionally faces in a semiparametric framework, even in the i.i.d. setting. The first step of semiparametric methodologies usually consists in a preliminary estimation of some infinite dimensional nuisance parameter (typically a density function or a nonparametric curve), on which the remaining (parametric) steps of the procedure are based. For handling theoretical difficulties related to this dependence problem, a well known method, called the *splitting trick*, amounts to split the data set into two parts, the first subset being used for estimating the nuisance parameter, while the parameter of interest is then estimated from the other subset (using the preliminary estimate). An analogous principle may be implemented in our framework using an additional split of the data in the "middle of the trajectory", for ensuring that a regeneration at least occurs in between with an overwhelming probability (so as to get two independent data subsets, see step 2 in

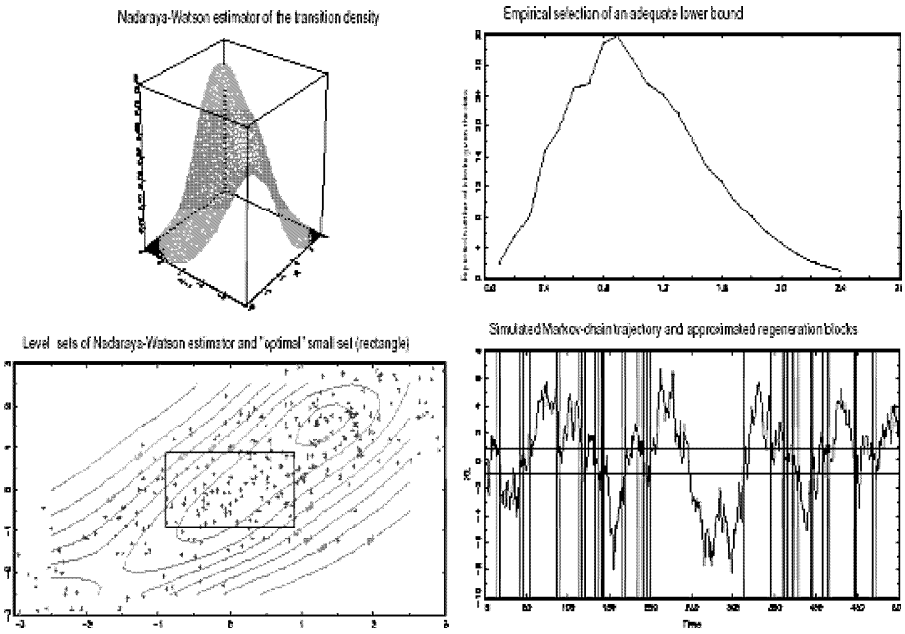

Figure 3: Illustration of Algorithm 3 - ARB construction with empirical choice of the small set.

the algorithm below). For this reason, we consider the following variant of the ARB construction. Let $1 < m < n$, $1 \leqslant p < n - m$.

Algorithm 4 *(two-split ARB construction)*

1. *From the data* $X^{(n+1)} = (X_1, ..., X_{n+1})$, *keep only the first* m *observations* $X^{(m)}$ *for computing an estimate* $p_m(x, y)$ *of* $p(x, y)$ *such that* $p_m(x, y) \geq \delta\gamma(y)$, $\lambda(dy)$ *a.s. and* $p_m(X_i, X_{i+1}) > 0$, $1 \leqslant i \leqslant n - 1$.

2. *Drop the observations between times* $m + 1$ *and* $m^* = m + p$ *(under standard assumptions, the split chain regenerates once at least in between with large probability).*

3. *From remaining observations* $X^{(m^*,n)} = (X_{m^*+1}, ..., X_n)$ *and estimate* p_m, *apply steps 2-4 of* ***Algorithm 2*** *(respectively of* ***Algorithm 3****).*

This procedure is similar to the *2-split method* proposed in [177], except that here the number of deleted observations is arbitrary and easier to interpret in terms of regeneration. Of course, the more often the split chain regenerates, the smaller p may be chosen. And the main problem consists in picking $m = m_n$ so that $m_n \to \infty$ as $n \to \infty$ for the estimate of the transition kernel to be accurate enough, while keeping enough observation $n - m^*$ for the block construction step: one typically chooses $m = o(n)$ as $n \to \infty$. Further assumptions are required for investigating precisely how to select m. In [21], a choice based on the rate of convergence α_m of the estimator $p_m(x, y)$ (for the MSE when error is measured by the sup-norm over $S \times S$) is proposed: when considering smooth markovian models for instance, estimators with rate $\alpha_m = m^{-1} \log(m)$ may be

exhibited and one shows that $m = n^{2/3}$ is then an optimal choice (up to a $\log(n)$). However, one may argue, as in the semiparametric case, that this methodology is motivated by our limitations in the analysis of asymptotic properties of the estimators only, whereas from a practical viewpoint it may deteriorate the finite sample performance of the initial algorithm. To our own experience, it is actually better to construct the estimate $p(x, y)$ from the whole trajectory and the interest of **Algorithm 4** is mainly theoretical.

Regeneration-based statistics for Harris Markov chains

0.5 Introduction

The present chapter mainly surveys results established at length in [19], [20] and [23]. More precisely, the problem of estimating additive functionals of the stationary distribution in the Harris positive recurrent case is considered in section 0.6. Estimators based on the (pseudo) regenerative blocks, as well as estimates of their asymptotic variance are exhibited, and limit theorems describing the asymptotic behaviour of their bias and their sampling distribution are also displayed. A specific notion of robustness for statistics based on the (approximate) regenerative blocks is introduced and investigated in section 0.7. And asymptotic properties of some regeneration-based statistics related to the extremal behaviour of Markov chains are studied in section 0.8 in the regenerative case only.

0.6 Asymptotic mean and variance estimation

In this section, we suppose that the chain X is positive recurrent with unknown stationary probability μ and consider the problem of estimating an additive functional of type $\mu(f) = \int f(x)\mu(dx) = \mathbb{E}_\mu(f(X_1))$, where f is a μ-integrable real valued function defined on the state space $(E, \mathcal{E})$. Estimation of additive functionals of type $\mathbb{E}_\mu(F(X_1, ..., X_k))$, for fixed $k \geqslant 1$, may be investigated in a similar fashion. We set $\bar{f}(x) = f(x) - \mu(f)$.

0.6.1 Regenerative case

Here we assume further that X admits an *a priori* known accessible atom A. As in the i.i.d. setting, a natural estimator of $\mu(f)$ is the sample mean statistic,

$$\mu'_n(f) = n^{-1} \sum_{i=1}^{n} f(X_i). \tag{7}$$

When the chain is stationary (*i.e.* when $\nu = \mu$), the estimator $\mu'_n(f)$ is zero-bias. However, its bias is significant in all other cases, mainly because of the presence of the first and last (non-regenerative) data blocks $\mathcal{B}_0$ and $\mathcal{B}_{l_n}^{(n)}$ (see Proposition 3 below). Besides, by virtue of Theorem 1, $\mu(f)$ may be expressed as the mean of the $f(X_i)$'s over a regeneration cycle (renormalized by the mean length of a regeneration cycle)

$$\mu(f) = \mathbb{E}_A(\tau_A)^{-1}\mathbb{E}_A(\sum_{i=1}^{\tau_A} f(X_i)).$$

Because the bias due to the first block depends on the unknown initial distribution (see Proposition 3 below) and thus can not be consistently estimated, we suggest to introduce the following estimators of the mean $\mu(f)$. Define the sample mean based on the observations (eventually) collected

after the first regeneration time only by

$$\tilde{\mu}_n(f) = (n - \tau_A)^{-1} \sum_{i=1+\tau_A}^{n} f(X_i)$$

with the convention $\tilde{\mu}_n(f) = 0$, when $\tau_A > n$, as well as the sample mean based on the observations collected between the first and last regeneration times before n by

$$\overline{\mu}_n(f) = (\tau_A(l_n) - \tau_A)^{-1} \sum_{i=1+\tau_A}^{\tau_A(l_n)} f(X_i)$$

with $l_n = \sum_{i=1}^{n} \mathbb{I}\{X_i \in A\}$ and the convention $\overline{\mu}_n(f) = 0$, when $l_n \leqslant 1$ (observe that, by Markov's inequality, $\mathbb{P}_\nu(l_n \leqslant 1) = O(n^{-1})$ as $n \to \infty$, as soon as $\mathcal{H}_0(1, \nu)$ and $\mathcal{H}_0(2)$ are fulfilled).

Let us introduce some additional notation for the block sums (resp. the block lengths), that shall be used here and throughout. For $j \geqslant 1$, $n \geqslant 1$, set

$$L_0 = \tau_A,\ L_j = \tau_A(j+1) - \tau_A(j),\ L_{l_n}^{(n)} = n - \tau_A(l_n)$$
$$f(\mathcal{B}_0) = \sum_{i=1}^{\tau_A} f(X_i),\ f(\mathcal{B}_j) = \sum_{i=1+\tau_A(j)}^{\tau_A(j+1)} f(X_i),\ f(\mathcal{B}_{l_n}^{(n)}) = \sum_{i=1+\tau_A(l_n)}^{n} f(X_i).$$

With these notations, the estimators above may be rewritten as

$$\mu_n'(f) = \frac{f(\mathcal{B}_0) + \sum_{j=1}^{l_n} f(\mathcal{B}_j) + f(\mathcal{B}_{l_n}^{(n)})}{L_0 + \sum_{j=1}^{l_n} L_j + L_{l_n}^{(n)}},$$
$$\tilde{\mu}_n(f) = \frac{\sum_{j=1}^{l_n} f(\mathcal{B}_j) + f(\mathcal{B}_{l_n}^{(n)})}{\sum_{j=1}^{l_n} L_j + L_{l_n}^{(n)}},\ \overline{\mu}_n(f) = \frac{\sum_{j=1}^{l_n} f(\mathcal{B}_j)}{\sum_{j=1}^{l_n} L_j}.$$

Let $\mu_n(f)$ design any of the three estimators $\mu_n'(f)$, $\tilde{\mu}_n(f)$ or $\overline{\mu}_n(f)$. If X fulfills conditions $\mathcal{H}_0(2)$, $\mathcal{H}_0(2,\nu)$, $\mathcal{H}_1(f,2,A)$, $\mathcal{H}_1(f,2,\nu)$ then the following CLT holds under $\mathbb{P}_\nu$ (*cf* Theorem 17.2.2 in [148])

$$n^{1/2}\sigma^{-1}(f)(\mu_n(f) - \mu(f)) \Rightarrow \mathcal{N}(0,1), \text{ as } n \to \infty,$$

with a normalizing constant

$$\sigma^2(f) = \mu(A)\,\mathbb{E}_A((\sum_{i=1}^{\tau_A} f(X_i) - \mu(f)\tau_A)^2). \tag{8}$$

From this expression, the following estimator of the asymptotic variance has been proposed in [19], adopting the usual convention regarding to empty summation,

$$\sigma_n^2(f) = n^{-1} \sum_{j=1}^{l_n - 1} (f(\mathcal{B}_j) - \overline{\mu}_n(f) L_j)^2. \tag{9}$$

Notice that the first and last data blocks are not involved in its construction. We could have proposed estimators involving different estimates of $\mu(f)$, but as will be seen later, it is preferable to consider an estimator based on regeneration blocks only. The following quantities shall be involved in the statistical analysis below. Define

$$\alpha = \mathbb{E}_A(\tau_A),\ \beta = \mathbb{E}_A(\tau_A \sum_{i=1}^{\tau_A} \overline{f}(X_i)) = \mathrm{Cov}_A(\tau_A, \sum_{i=1}^{\tau_A} \overline{f}(X_i)),$$

$$\varphi_\nu = \mathbb{E}_\nu(\sum_{i=1}^{\tau_A} \bar{f}(X_i)),\ \gamma = \alpha^{-1}\mathbb{E}_A(\sum_{i=1}^{\tau_A}(\tau_A - i)\bar{f}(X_i)).$$

We also introduce the following technical conditions.

(C1) (*Cramer condition*)

$$\overline{\lim_{t\to\infty}} \mid \mathbb{E}_A(\exp(it\sum_{i=1}^{\tau_A} \bar{f}(X_i))) \mid < 1.$$

(C2) (*Cramer condition*)

$$\overline{\lim_{t\to\infty}} \mid \mathbb{E}_A(\exp(it(\sum_{i=1}^{\tau_A} \bar{f}(X_i))^2)) \mid < 1.$$

(C3) *There exists* $N \geqslant 1$ *such that the* N*-fold convoluted density* g^{*N} *is bounded, denoting by* g *the density of the* $(\sum_{i=1+\tau_A(1)}^{\tau_A(2)} \bar{f}(X_i) - \alpha^{-1}\beta)^2$*'s.*

(C4) *There exists* $N \geqslant 1$ *such that the* N*-fold convoluted density* G^{*N} *is bounded, denoting by* G *the density of the* $(\sum_{i=1+\tau_A(1)}^{\tau_A(2)} \bar{f}(X_i))^2$*'s.*

These two conditions are automatically satisfied if $\sum_{i=1+\tau_A(1)}^{\tau_A(2)} \bar{f}(X_i)$ has a bounded density. The result below is a straightforward extension of Theorem 1 in [137] (see also Prop. 3.1 in [19]).

Proposition 3 *(Bertail & Clémençon, 2004a) Suppose that* $\mathcal{H}_0(4)$, $\mathcal{H}_0(2, \nu)$, $\mathcal{H}_1(4, f)$, $\mathcal{H}_1(2, \nu, f)$ *and Cramer condition (C1) are satisfied by the chain. Then, as* $n \to \infty$*, we have*

$$\mathbb{E}_\nu(\mu'_n(f)) = \mu(f) + (\varphi_\nu + \gamma - \beta/\alpha)n^{-1} + O(n^{-3/2}), \tag{10}$$

$$\mathbb{E}_\nu(\tilde{\mu}_n(f)) = \mu(f) + (\gamma - \beta/\alpha)n^{-1} + O(n^{-3/2}), \tag{11}$$

$$\mathbb{E}_\nu(\overline{\mu}_n(f)) = \mu(f) - (\beta/\alpha)n^{-1} + O(n^{-3/2}). \tag{12}$$

If the Cramer condition (C2) is also fulfilled, then

$$\mathbb{E}_\nu(\sigma_n^2(f)) = \sigma^2(f) + O(n^{-1}),\ \ as\ n \to \infty, \tag{13}$$

and we have the following CLT under $\mathbb{P}_\nu$*,*

$$n^{1/2}(\sigma_n^2(f) - \sigma^2(f)) \Rightarrow \mathcal{N}(0, \xi^2(f)),\ \ as\ n \to \infty, \tag{14}$$

with $\xi^2(f) = \mu(A)\mathrm{Var}_A((\sum_{i=1}^{\tau_A} \bar{f}(X_i))^2 - 2\alpha^{-1}\beta \sum_{i=1}^{\tau_A} \bar{f}(X_i))$.

We emphasize that in a non i.i.d. setting, it is generally difficult to construct an accurate (positive) estimator of the asymptotic variance. When no structural assumption, except stationarity and square integrability, is made on the underlying process X, a possible method, currently used in practice, is based on so-called *blocking techniques*. Indeed under some appropriate mixing conditions (which ensure that the following series converge), it can be shown that the variance of $n^{-1/2}\mu'_n(f)$ may be written

$$\mathrm{Var}(n^{-1/2}\mu'_n(f)) = \Gamma(0) + 2\sum_{t=1}^{n}(1 - t/n)\Gamma(t)$$

and converges to

$$\sigma^2(f) = \sum_{t=\infty}^{\infty} \Gamma(t) = 2\pi g(0),$$

where $g(w) = (2\pi)^{-1}\sum_{t=-\infty}^{\infty}\Gamma(t)\cos(wt)$ and $(\Gamma(t))_{t\geqslant 0}$ denote respectively the spectral density and the autocovariance sequence of the discrete-time stationary process X. Most of the estimators of $\sigma^2(f)$ that have been proposed in the literature (such as the Bartlett spectral density estimator, the moving-block jackknife/subsampling variance estimator, the overlapping or non-overlapping batch means estimator) may be seen as variants of the basic *moving-block bootstrap estimator*(see [129], [133])

$$\hat{\sigma}^2_{M,n} = \frac{M}{Q}\sum_{i=1}^{Q}(\overline{\mu}_{i,M,L} - \mu_n(f))^2, \tag{15}$$

where $\overline{\mu}_{i,M,L} = M^{-1}\sum_{t=L(i-1)+1}^{L(i-1)+M} f(X_t)$ is the mean of f on the i-th data block $(X_{L(i-1)+1},\ldots,$ $X_{L(i-1)+M})$. Here, the size M of the blocks and the amount L of 'lag' or overlap between each block are deterministic (eventually depending on n) and $Q = [\frac{n-M}{L}]+1$, denoting by $[\cdot]$ the integer part, is the number of blocks that may be constructed from the sample $X_1,\ldots, X_n$. In the case when $L = M$, there is no overlap between block i and block $i+1$ (as the original solution considered by [101], [50]), whereas the case $L = 1$ corresponds to maximum overlap (see [154], [156] for a survey). Under suitable regularity conditions (mixing and moments conditions), it can be shown that if $M \to \infty$ with $M/n \to 0$ and $L/M \to a \in [0,1]$ as $n \to \infty$, then we have

$$\mathbb{E}(\hat{\sigma}^2_{M,n}) - \sigma^2(f) = O(1/M) + O(\sqrt{M/n}), \tag{16}$$

$$Var(\hat{\sigma}^2_{M,n}) = 2c\frac{M}{n}\sigma^4(f) + o(M/n), \tag{17}$$

as $n \to \infty$, where c is a constant depending on a, taking its smallest value (namely $c = 2/3$) for $a = 0$. This result shows that the bias of such estimators may be very large. Indeed, by optimizing in M we find the optimal choice $M \sim n^{1/3}$, for which we have $\mathbb{E}(\hat{\sigma}^2_{M,n}) - \sigma^2(f) = O(n^{-1/3})$. Various extrapolation and jackknife techniques or kernel smoothing methods have been suggested to get rid of this large bias (refer to [154] [96], [18] and [28]). The latter somehow amount to make use of Rosenblatt smoothing kernels of order higher than two (taking some negative values) for estimating the spectral density at 0. However, the main drawback in using these estimators is that they take negative values for some n, and lead consequently to face problems, when dealing with studentized statistics. In our specific Markovian framework, the estimate $\sigma^2_n(f)$ in the atomic case (or latter $\hat{\sigma}^2_n(f)$ in the general case) is much more natural and allows to avoid these problems. This is particularly important when the matter is to establish Edgeworth expansions at orders higher than two in such a non i.i.d. setting. As a matter of fact, the bias of the variance may completely cancel the accuracy provided by higher order Edgeworth expansions (but also the one of its Bootstrap approximation) in the studentized case, given its explicit role in such expansions (see [96]).

From Proposition 3, we immediately derive that

$$t_n = n^{1/2}\sigma_n^{-1}(f)(\mu_n(f) - \mu(f)) \Rightarrow \mathcal{N}(0,1), \text{ as } n \to \infty,$$

so that asymptotic confidence intervals for $\mu(f)$ are immediately available in the atomic case. This result also shows that using estimators $\tilde{\mu}_n(f)$ or $\overline{\mu}_n(f)$ instead of $\mu'_n(f)$ allows to eliminate the only quantity depending on the initial distribution ν in the first order term of the bias, which may be interesting for estimation purpose and is crucial when the matter is to deal with an estimator of which variance or sampling distribution may be approximated by a resampling procedure in a nonstationary setting (given the impossibility to approximate the distribution of the "first block sum" $\sum_{i=1}^{\tau_A} f(X_i)$ from one single realization of X starting from ν). For these estimators, it is actually possible to implement specific Bootstrap methodologies, for constructing second order correct confidence intervals for instance. Regarding to this, it should be noticed that Edgeworth

expansions (E.E. in abbreviated form) may be obtained using the regenerative method by partitioning the state space according to all possible values for the number l_n regeneration times before n and for the sizes of the first and last block as in [138]. [19] proved the validity of an E.E. in the studentized case, of which form is recalled below. Notice that actually (C3) corresponding to their v) in Proposition 3.1 in [19] is not needed in the unstudentized case. Let $\Phi(x)$ denote the distribution function of the standard normal distribution and set $\phi(x) = d\Phi(x)/dx$.

Theorem 4 *(Bertail & Clémençon, 2004a) Let* $b(f) = \lim_{n\to\infty} n(\mu_n(f) - \mu(f))$ *be the asymptotic bias of* $\mu_n(f)$*. Under conditions* $\mathcal{H}_0(4)$, $\mathcal{H}_0(2, \nu)$ $\mathcal{H}_1(4, f)$, $\mathcal{H}_1(2, \nu, f)$, *(C1), we have the following E.E.,*

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_\nu \left(n^{1/2}\sigma(f)^{-1}(\mu_n(f) - \mu(f)) \leq x\right) - E_n^{(2)}(x)| = O(n^{-1}), \text{ as } n \to \infty,$$

with

$$E_n^{(2)}(x) = \Phi(x) - n^{-1/2}\frac{k_3(f)}{6}(x^2 - 1)\phi(x) - n^{-1/2}b(f)\phi(x), \tag{18}$$

$$k_3(f) = \alpha^{-1}(M_{3,A} - \frac{3\beta}{\sigma(f)}), \ M_{3,A} = \frac{\mathbb{E}_A((\sum_{i=1}^{\tau_A} \bar{f}(X_i))^3)}{\sigma(f)^3}. \tag{19}$$

A similar limit result holds for the studentized statistic under the further hypothesis that (C2), (C3), $\mathcal{H}_0(s)$ *and* $\mathcal{H}_1(s, f)$ *are fulfilled with* $s = 8 + \varepsilon$ *for some* $\varepsilon > 0$*:*

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_\nu(n^{1/2}\sigma_n^{-1}(f)(\mu_n(f) - \mu(f)) \leq x) - F_n^{(2)}(x)| = O(n^{-1}\log(n)), \tag{20}$$

as $n \to \infty$*, with*

$$F_n^{(2)}(x) = \Phi(x) + n^{-1/2}\frac{1}{6}k_3(f)(2x^2 + 1)\phi(x) - n^{-1/2}b(f)\phi(x).$$

When $\mu_n(f) = \bar{\mu}_n(f)$*, under C4),* $O(n^{-1}\log(n))$ *may be replaced by* $O(n^{-1})$*.*

This theorem may serve for building accurate confidence intervals for $\mu(f)$ (by E.E. inversion as in [1] or [100]). It also paves the way for studying precisely specific bootstrap methods, as in [22]. It should be noted that the skewness $k_3(f)$ is the sum of two terms: the third moment of the recentered block sums and a correlation term between the block sums and the block lengths. The coefficients involved in the E.E. may be directly estimated from the regenerative blocks. The next result follows from straightforward CLT arguments.

Proposition 5 *(Bertail & Clémençon, 2006a) For* $s \geqslant 1$*, under* $\mathcal{H}_1(f, 2s)$, $\mathcal{H}_1(f, 2, \nu)$, $\mathcal{H}_0(2s)$ *and* $\mathcal{H}_0(2, \nu)$*, then* $M_{s,A} = \mathbb{E}_A((\sum_{i=1}^{\tau_A} \bar{f}(X_i))^s)$ *is well-defined and we have*

$$\widehat{\mu}_{s,n} = n^{-1}\sum_{i=1}^{l_n-1}(f(\mathcal{B}_j) - \bar{\mu}_n(f)L_j)^s = \alpha^{-1}M_{s,A} + O_{\mathbb{P}_\nu}(n^{-1/2}), \text{ as } n \to \infty.$$

0.6.2 Positive recurrent case

We now turn to the general positive recurrent case. It is noteworthy that, though they may be expressed using the parameters of the minorization condition $\mathcal{M}$, the constants involved in the

CLT are independent from the latter. In particular the mean and the asymptotic variance may be written as

$$\mu(f) = \mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-1}\mathbb{E}_{A_{\mathcal{M}}}(\sum_{i=1}^{\tau_{A_{\mathcal{M}}}} f(X_i)),$$
$$\sigma^2(f) = \mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-1}\mathbb{E}_{A_{\mathcal{M}}}((\sum_{i=1}^{\tau_{A_{\mathcal{M}}}} \bar{f}(X_i))^2),$$

where $\tau_{A_{\mathcal{M}}} = \inf\{n \geqslant 1,\ (X_n,\ Y_n) \in S \times \{1\}\}$ and $\mathbb{E}_{A_{\mathcal{M}}}(.)$ denotes the expectation conditionally to $(X_0,\ Y_0) \in A_{\mathcal{M}} = S \times \{1\}$. However, one cannot use the estimators of $\mu(f)$ and $\sigma^2(f)$ defined in the atomic setting, applied to the split chain, since the times when the latter regenerates are unobserved. We thus consider the following estimators based on the *approximate regeneration times* (*i.e.* times i when $(X_i, \widehat{Y}_i) \in S \times \{1\}$), as constructed in §1.3.2,

$$\widehat{\mu}_n(f) = \widehat{n}_{A_{\mathcal{M}}}^{-1} \sum_{j=1}^{\widehat{l}_n - 1} f(\widehat{\mathcal{B}}_j) \text{ and } \widehat{\sigma}_n^2(f) = \widehat{n}_{A_{\mathcal{M}}}^{-1} \sum_{j=1}^{\widehat{l}_n - 1} \{f(\widehat{\mathcal{B}}_j) - \widehat{\mu}_n(f)\widehat{L}_j\}^2,$$

with, for $j \geqslant 1$,

$$f(\widehat{\mathcal{B}}_j) = \sum_{i=1+\widehat{\tau}_{A_{\mathcal{M}}}(j)}^{\widehat{\tau}_{A_{\mathcal{M}}}(j+1)} f(X_i),\ \widehat{L}_j = \widehat{\tau}_{A_{\mathcal{M}}}(j+1) - \widehat{\tau}_{A_{\mathcal{M}}}(j),$$
$$\widehat{n}_{A_{\mathcal{M}}} = \widehat{\tau}_{A_{\mathcal{M}}}(\widehat{l}_n) - \widehat{\tau}_{A_{\mathcal{M}}}(1) = \sum_{j=1}^{\widehat{l}_n - 1} \widehat{L}_j.$$

By convention, $\widehat{\mu}_n(f) = 0$ and $\widehat{\sigma}_n^2(f) = 0$ (resp. $\widehat{n}_{A_{\mathcal{M}}} = 0$), when $\widehat{l}_n \leqslant 1$ (resp., when $\widehat{l}_n = 0$). Since the ARB construction involves the use of an estimate $p_n(x, y)$ of the transition kernel $p(x, y)$, we consider conditions on the rate of convergence of this estimator. For a sequence of nonnegative real numbers $(\alpha_n)_{n\in\mathbb{N}}$ converging to 0 as $n \to \infty$,

$\mathcal{H}_2$: *$p(x,\ y)$ is estimated by $p_n(x,\ y)$ at the rate α_n for the MSE when error is measured by the L^∞ loss over $S \times S$:*

$$\mathbb{E}_\nu(\sup_{(x,y)\in S\times S} |p_n(x, y) - p(x, y)|^2) = O(\alpha_n),\ \text{as } n \to \infty.$$

See Remark 1 for references concerning the construction and the study of transition density estimators for positive recurrent chains, estimation rates are usually established under various smoothness assumptions on the density of the joint distribution $\mu(dx)\Pi(x, dy)$ and the one of $\mu(dx)$. For instance, under classical Hölder constraints of order s, the typical rate for the risk in this setup is $\alpha_n \sim (\ln n/n)^{s/(s+1)}$ (refer to [54]).

$\mathcal{H}_3$: *The "minorizing" density γ is such that* $\inf_{x\in S} \gamma(x) > 0$.

$\mathcal{H}_4$: *The transition density $p(x, y)$ and its estimate $p_n(x, y)$ are bounded by a constant $R < \infty$ over S^2.*

Some asymptotic properties of these statistics based on the approximate regeneration data blocks are stated in the following theorem (their proof immediately follows from the argument of Theorem 3.2 and Lemma 5.3 in [20]).

Theorem 6 *(Bertail & Clémençon, 2006a) If assumptions $\mathcal{H}_0'(2,\nu)$, $\mathcal{H}_0'(8)$, $\mathcal{H}_1'(f,2,\nu)$, $\mathcal{H}_1'(f,8)$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$ are satisfied by X, as well as conditions (C1) and (C2) by the split chain, we have, as $n \to \infty$,*

$$\mathbb{E}_\nu(\widehat{\mu}_n(f)) = \mu(f) - \beta/\alpha\, n^{-1} + O(n^{-1}\alpha_n^{1/2}),$$
$$\mathbb{E}_\nu(\widehat{\sigma}_n^2(f)) = \sigma^2(f) + O(\alpha_n \vee n^{-1}),$$

and if $\alpha_n = o(n^{-1/2})$, then

$$n^{1/2}(\widehat{\sigma}_n^2(f) - \sigma^2(f)) \Rightarrow \mathcal{N}(0, \xi^2(f)),$$

where α, β and $\xi^2(f)$ are the quantities related to the split chain defined in Prop. 3.

Remark 2 The condition $\alpha_n = o(n^{-1/2})$ as $n \to \infty$ may be ensured by smoothness conditions satisfied by the transition kernel $p(x,y)$: under Hölder constraints of order s such rates are achieved as soon as $s > 1$, that is a rather weak assumption.

Define also the *pseudo-regeneration based standardized* (resp., *studentized*) *sample mean* by

$$\widehat{\sigma}_n = n^{1/2}\sigma^{-1}(f)(\widehat{\mu}_n(f) - \mu(f)),$$
$$\widehat{t}_n = \widehat{n}_{A_{\mathcal{M}}}^{1/2}\,\widehat{\sigma}_n(f)^{-1}(\widehat{\mu}_n(f) - \mu(f)).$$

The following theorem straightforwardly results from Theorem 6.

Theorem 7 *(Bertail & Clémençon, 2006a) Under the assumptions of Theorem 6, we have*

$$\widehat{\sigma}_n \Rightarrow \mathcal{N}(0,1) \text{ and } \widehat{t}_n \Rightarrow \mathcal{N}(0,1), \text{ as } n \to \infty.$$

This shows that from pseudo-regeneration blocks one may easily construct a consistent estimator of the asymptotic variance $\sigma^2(f)$ and asymptotic confidence intervals for $\mu(f)$ in the general positive recurrent case (see Chapter 3 for more accurate confidence intervals based on a regenerative bootstrap method). In [19], an E.E. is proved for the studentized statistic $\widehat{t}_n$. The main problem consists in handling computational difficulties induced by the dependence structure, that results from the preliminary estimation of the transition density. For partly solving this problem, one may use **Algorithm 4**, involving the *2-split trick*. Under smoothness assumptions for the transition kernel (which are often fulfilled in practice), [22] established the validity of the E.E. up to $O(n^{-5/6}\log(n))$, stated in the result below.

Theorem 8 *(Bertail & Clémençon, 2005b) Suppose that (C1) is satisfied by the split chain, and that $\mathcal{H}_0'(\kappa,\nu)$, $\mathcal{H}_1'(\kappa,f,\nu)$, $\mathcal{H}_0'(\kappa)$, $\mathcal{H}_1'(\kappa,f)$ with $\kappa > 6$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$ are fulfilled. Let m_n and p_n be integer sequences tending to ∞ as $n \to \infty$, such that $n^{1/\gamma} \le p_n \le m_n$ and $m_n = o(n)$ as $n \to \infty$. Then, the following limit result holds for the pseudo-regeneration based standardized sample mean obtained via Algorithm 4*

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_\nu(\widehat{\sigma}_n \le x) - E_n^{(2)}(x)| = O(n^{-1/2}\alpha_{m_n}^{1/2} \vee n^{-3/2}m_n), \text{ as } n \to \infty,$$

and if in addition the preceding assumptions with $\kappa > 8$ and C4) are satisfied, we also have

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_\nu(\widehat{t}_n \le x) - F_n^{(2)}(x)| = O(n^{-1/2}\alpha_{m_n}^{1/2} \vee n^{-3/2}m_n), \text{ as } n \to \infty,$$

where $E_n^{(2)}(x)$ and $F_n^{(2)}(x)$ are the expansions defined in Theorem 4 related to the split chain. In particular, if $\alpha_{m_n} = m_n\log(m_n)$, by picking $m_n = n^{2/3}$, these E.E. hold up to $O(n^{-5/6}\log(n))$.

The conditions stipulated in this result are weaker than the conditions ensuring that the Moving Block Bootstrap is second order correct. More precisely, they are satisfied for a wide range of Markov chains, including nonstationary cases and chains with polynomial decay of α−mixing coefficients that do not fall into the validity framework of the MBB methodology. In particular it is worth noticing that these conditions are weaker than [95]'s conditions (in a strong mixing setting).

As stated in the following proposition, the coefficients involved in the E.E.'s above may be estimated from the approximate regeneration blocks.

Proposition 9 *(Bertail & Clémençon, 2006a) Under $\mathcal{H}_0'(2s,\ \nu)$, $\mathcal{H}_1'(2s,\ \nu,\ f)$, $\mathcal{H}_0'(2s \vee 8)$, $\mathcal{H}_1'(2s \vee 8,\ f)$ with $s \geq 2$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$, the expectation $M_{s,\mathcal{A}_\mathcal{M}} = \mathbb{E}_{\mathcal{A}_\mathcal{M}}((\sum_{i=1}^{\tau_{\mathcal{A}_\mathcal{M}}} \bar{f}(X_i))^s)$ is well-defined and we have, as $n \to \infty$,*

$$\widehat{\mu}_{s,n} = n^{-1} \sum_{i=1}^{l_n-1} (f(\widehat{\mathcal{B}}_j) - \widehat{\mu}_n(f)\widehat{L}_j)^s = \mathbb{E}_{\mathcal{A}_\mathcal{M}}(\tau_{\mathcal{A}_\mathcal{M}})^{-1} M_{s,\mathcal{A}_\mathcal{M}} + O_{\mathbb{P}_\nu}(\alpha_{m_n}^{1/2}).$$

0.6.3 Some illustrative examples

Here we give some examples with the aim to illustrate the wide range of applications of the results previously stated.

Example 1 : countable Markov chains. Let X be a general irreducible chain with a countable state space E. For such a chain, any recurrent state $a \in E$ is naturally an accessible atom and conditions involved in the limit results presented in §2.2.1 may be easily checked at hand. Consider for instance Cramer condition (C1). Denote by Π the transition matrix and set $A = \{a\}$. Assuming that f is μ-centered. We have, for any $k \in \mathbb{N}^*$:

$$\begin{aligned}\left|\mathbb{E}_A(e^{it\sum_{j=1}^{\tau_A} f(X_j)})\right| &= \left|\sum_{l=1}^{\infty} \mathbb{E}_A(e^{it\sum_{j=1}^{l} f(X_j)}|\tau_A = l)\mathbb{P}_A(\tau_A = l)\right| \\ &\leqslant \left|\mathbb{E}_A(e^{it\sum_{j=1}^{k} f(X_j)}|\tau_A = k)\right| \mathbb{P}_A(\tau_A = k) + 1 - \mathbb{P}_A(\tau_A = k).\end{aligned}$$

It follows that checking (C1) boils down to showing the partial conditional Cramer condition

$$\overline{\lim_{t\to\infty}} \left|\mathbb{E}_A(e^{it\sum_{j=1}^{k} f(X_j)}|\tau_A = k)\right| < 1,$$

for some $k > 0$ such that $\mathbb{P}_A(\tau_A = k) > 0$. In particular, similarly to the i.i.d. case, this condition then holds, as soon as the set $\{f(x)\}_{x\in E}$ is not a point lattice (*i.e.* a regular grid).

Example 2 : modulated random walk on $\mathbb{R}_+$. Consider the model

$$X_0 = 0 \text{ and } X_{n+1} = (X_n + W_n)_+ \text{ for } n \in \mathbb{N}, \tag{21}$$

where $x_+ = \max(x,\ 0)$, (X_n) and (W_n) are sequences of r.v.'s such that, for all $n \in \mathbb{N}$, the distribution of W_n conditionally to $X_0, ..., X_n$ is given by $U(X_n, .)$ where $U(x, w)$ is a transition kernel from $\mathbb{R}_+$ to $\mathbb{R}$. Then, X_n is a Markov chain on $\mathbb{R}_+$ with transition probability kernel:

$$\begin{aligned}\Pi(x,\ \{0\}) &= U(x,\ ]-\infty,\ -x]), \\ \Pi(x,\ ]y,\ \infty[) &= U(x,\ ]y-x,\ \infty[),\end{aligned}$$

for all $x \geqslant 0$. Observe that the chain Π is δ_0-irreducible when $U(x, .)$ has infinite left tail for all $x \geqslant 0$ and that $\{0\}$ is then an accessible atom for X. The chain is shown to be positive recurrent iff

there exists $b > 0$ and a test function $V : \mathbb{R}_+ \to [0, \infty]$ such that $V(0) < \infty$ and the drift condition below holds for all $x \geqslant 0$

$$\int \Pi(x, dy)V(y) - V(x) \leqslant -1 + b\mathbb{I}\{x = 0\},$$

(see [148]. The times at which X reaches the value 0 are thus regeneration times, and allow to define regeneration blocks dividing the sample path, as shown in Fig. 1. Such a modulated random walk (for which, at each step n, the increasing W_n depends on the actual state $X_n = x$), provides a model for various systems, such as the popular *content-dependent storage process* studied in [107] (see also [44]) or the *work-modulated single server queue* in the context of queuing systems (*cf* [45]). For such atomic chains with continuous state space (refer to [148], [83], [84] and [8] for other examples of such chains), one may easily check conditions used in §2.2.1 in many cases. One may show for instance that (C1) is fulfilled as soon as there exists $k \geqslant 1$ such that $0 < \mathbb{P}_A(\tau_A = k) < 1$ and the distribution of $\sum_{i=1}^{k} f(X_i)$ conditioned on $X_0 \in A$ and $\tau_A = k$ is absolutely continuous. For the regenerative model described above, this sufficient condition is fulfilled with $k = 2$, $f(x) = x$ and $A = \{0\}$, when it is assumed for instance that $U(x, dy)$ is absolutely continuous for all $x \geqslant 0$ and $\emptyset \neq \text{supp}U(0, dy) \cap \mathbb{R}_+^* \neq \mathbb{R}_+^*$.

Example 3: nonlinear time series. Consider the heteroskedastic autoregressive model

$$X_{n+1} = m(X_n) + \sigma(X_n)\varepsilon_{n+1}, \ n \in \mathbb{N},$$

where $m : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}_+^*$ are measurable functions, $(\varepsilon_n)_{n\in\mathbb{N}}$ is a i.i.d. sequence of r.v.'s drawn from $g(x)dx$ such that, for all $n \in \mathbb{N}$, ε_{n+1} is independent from the X_k's, $k \leqslant n$ with $\mathbb{E}(\varepsilon_{n+1}) = 0$ and $\mathbb{E}(\varepsilon_{n+1}^2) = 1$. The transition kernel density of the chain is given by $p(x, y) = \sigma(x)^{-1}g((y - m(x))/\sigma(x))$, $(x, y) \in \mathbb{R}^2$. Assume further that g, m and σ are continuous functions and there exists $x_0 \in \mathbb{R}$ such that $p(x_0, x_0) > 0$. Then, the transition density is uniformly bounded from below over some neighborhood $V_{x_0}(\varepsilon)^2 = [x_0 - \varepsilon, x_0 + \varepsilon]^2$ of (x_0, x_0) in $\mathbb{R}^2$: there exists $\delta = \delta(\varepsilon) \in]0, 1[$ such that,

$$\inf_{(x,y)\in V_{x_0}^2} p(x, y) \geqslant \delta(2\varepsilon)^{-1}. \tag{22}$$

We thus showed that the chain X satisfies the minorization condition $\mathcal{M}(1, V_{x_0}(\varepsilon), \delta, \mathcal{U}_{V_{x_0}(\varepsilon)})$.

0.7 Robust functional parameter estimation

Extending the notion of *influence function* and/or *robustness* to the framework of general time series is a difficult task (see [128] or [143]). Such concepts are important not only to detect *"outliers"* among the data or influential observations but also to generalize the important notion of *efficient estimation* in semiparametric frameworks (see the recent discussion in [31] for instance). In the markovian setting, a recent proposal based on martingale approximation has been made by [149]. In [23] an alternative definition of the influence function based on the (approximate) regeneration blocks construction is proposed, which is easier to manipulate and immediately leads to central limit and convolution theorems.

0.7.1 Defining the influence function on the torus

The leitmotiv of this chapter is that most parameters of interest related to Harris chains are functionals of the distribution $\mathcal{L}$ of the regenerative blocks (observe that $\mathcal{L}$ is a distribution on the torus $\mathbb{T} = \cup_{n\geqslant 1}E^n$), namely the distribution of $(X_1, ..., X_{\tau_A})$ conditioned on $X_0 \in A$ when the

chain possesses an atom A, or the distribution of $(X_1, ..., X_{\tau_{A_{\mathcal{M}}}})$ conditioned on $(X_0, Y_0) \in A_{\mathcal{M}}$ in the general case when one considers the split chain (refer to §1.2.2 for assumptions and notation, here we shall omit the subscript A and $\mathcal{M}$ to make the notation simpler). In view of Theorem 1, this is obviously true in the positive recurrent case for any functional of the stationary law μ. But, more generally, the probability distribution $\mathbb{P}_\nu$ of the Markov chain X starting from ν may be decomposed as follows :

$$\mathbb{P}_\nu((X_n)_{n\geqslant 1}) = \mathcal{L}_\nu((X_1, ..., X_{\tau_{A(1)}})) \prod_{k=1}^{\infty} \mathcal{L}((X_{1+\tau_A(k)}, ..., X_{\tau_A(k+1)})),$$

denoting by $\mathcal{L}_\nu$ the distribution of $(X_1, ..., X_{\tau_A})$ conditioned on $X_0 \sim \nu$. Thus any functional of the law of $(X_n)_{n\geqslant 1}$ may be seen as a functional of $(\mathcal{L}_\nu, \mathcal{L})$. However, pointing out that the distribution of $\mathcal{L}_\nu$ cannot be estimated in most cases encountered in practice, only functionals of $\mathcal{L}$ are of practical interest. The object of this subsection is to propose the following definition of the influence function for such functionals. Let $\mathcal{P}_{\mathbb{T}}$ denote the set of all probability measures on the torus $\mathbb{T}$ and for any $b \in \mathbb{T}$, set $L(b) = k$ if $b \in E^k$, $k \geqslant 1$. We then have the following natural definition, that straightforwardly extends the classical notion of influence function in the i.i.d. case, with the important novelty that distributions on the torus are considered here.

Definition 10 *Let $T : \mathcal{P}_{\mathbb{T}} \to \mathbb{R}$ be a functional on $\mathcal{P}_{\mathbb{T}}$. If for $\mathcal{L}$ in $\mathcal{P}_{\mathbb{T}}$, $t^{-1}(T((1-t)\mathcal{L}+t\delta_b)-T(\mathcal{L}))$ has a finite limit as $t \to 0$ for any $b \in \mathbb{T}$, then the influence function $T^{(1)}$ of the functional T is well defined, and by definition one has for all b in $\mathbb{T}$,*

$$T^{(1)}(b,\ \mathcal{L}) = \lim_{t\to 0} \frac{T((1-t)\mathcal{L} + t\delta_b) - T(\mathcal{L})}{t}. \tag{23}$$

0.7.2 Some examples

The relevance of this definition is illustrated through the following examples, which aim to show how easy it is to adapt known calculations of influence function on $\mathbb{R}$ to this framework.

a) Suppose that X is positive recurrent with stationary distribution μ. Let $f : E \to \mathbb{R}$ be μ-integrable and consider the parameter $\mu_0(f) = \mathbb{E}_\mu(f(X))$. Denote by $\mathcal{B}$ a r.v. valued in $\mathbb{T}$ with distribution $\mathcal{L}$ and observe that $\mu_0(f) = \mathbb{E}_{\mathcal{L}}(f(\mathcal{B}))/\mathbb{E}_{\mathcal{L}}(L(\mathcal{B})) = T(\mathcal{L})$ (recall the notation $f(b) = \sum_{i=1}^{L(b)} f(b_i)$ for any $b \in \mathbb{T}$). A classical calculation for the influence function of ratios yields then

$$T^{(1)}(b, \mathcal{L}) = \frac{d}{dt}(T((1-t)\mathcal{L} + tb)|_{t=0} = \frac{f(b) - \mu(f)L(b)}{\mathbb{E}_{\mathcal{L}}(L(\mathcal{B}))}$$

Notice that $\mathbb{E}_{\mathcal{L}}(T^{(1)}(\mathcal{B}, \mathcal{L})) = 0$.

b) Let θ be the unique solution of the equation: $\mathbb{E}_\mu(\psi(X,\ \theta)) = 0$, where $\psi :\ \mathbb{R}^2 \to \mathbb{R}$ is $\mathcal{C}^2$. Observing that it may be rewritten as $\mathbb{E}_{\mathcal{L}}(\psi(\mathcal{B},\ \theta)) = 0$, a similar calculation to the one used in the i.i.d. setting (if differentiating inside the expectation is authorized) gives in this case

$$T^{(1)}_\psi(b, \mathcal{L}) = -\frac{\psi(b,\ \theta)}{\mathbb{E}_A(\sum_{i=1}^{\tau_A} \frac{\partial\psi(X_i,\theta)}{\partial\theta})}.$$

By definition of θ, we naturally have $\mathbb{E}_{\mathcal{L}}(T^{(1)}_\psi(B, \mathcal{L})) = 0$.

c) Assuming that the chain takes real values and its stationary law μ has zero mean and finite variance, let ρ be the correlation coefficient between consecutive observations under the stationary distribution:

$$\rho = \frac{\mathbb{E}_\mu(X_nX_{n+1})}{\mathbb{E}_\mu(X_n^2)} = \frac{\mathbb{E}_A(\sum_{n=1}^{\tau_A} X_nX_{n+1})}{\mathbb{E}_A(\sum_{n=1}^{\tau_A} X_n^2)}.$$

For all b in $\mathbb{T}$, the influence function is

$$T_\rho^{(1)}(b,\mathcal{L}) = \frac{\sum_{i=1}^{L(b)} b_i(b_{i+1}-\rho b_i)}{\mathbb{E}_A(\sum_{t=1}^{\tau_A} X_t^2)},$$

and one may check that $\mathbb{E}_{\mathcal{L}}(T_\rho^{(1)}(\mathcal{B},\mathcal{L})) = 0$.

d) It is now possible to reinterpret the results obtained for U-statistics in section 6. With the notation above, the parameter of interest may be rewritten

$$\mu(U) = \mathbb{E}_{\mathcal{L}}\left(L(\mathcal{B})\right)^{-2}\mathbb{E}_{\mathcal{L}\times\mathcal{L}}(U(\mathcal{B}_1,\mathcal{B}_2)),$$

yielding the influence function: $\forall b \in \mathbb{T}$,

$$\mu^{(1)}(b,\mathcal{L}) = 2\mathbb{E}_{\mathcal{L}}\left(L(\mathcal{B})\right)^{-2}\mathbb{E}_{\mathcal{L}}(\tilde{\omega}_U(\mathcal{B}_1,\mathcal{B}_2)|\mathcal{B}_1 = b).$$

0.7.3 Main results

In order to lighten the notation, the study is restricted to the case when X takes real values, *i.e.* $E \subset \mathbb{R}$, but straightforwardly extends to a more general framework. Given an observed trajectory of length n, natural empirical estimates of parameters $T(\mathcal{L})$ are of course the *plug-in estimators* $T(\mathcal{L}_n)$ based on the empirical distribution of the observed regeneration blocks $\mathcal{L}_n = (l_n-1)^{-1}\sum_{j=1}^{l_n-1}\delta_{\mathcal{B}_j} \in \mathcal{P}_{\mathbb{T}}$ in the atomic case, which is defined as soon as $l_n \geqslant 2$ (notice that $\mathbb{P}_\nu(l_n \leqslant 1) = O(n^{-1})$ as $n \to \infty$, if $\mathcal{H}_0(1,\nu)$ and $\mathcal{H}_0(2)$ are satisfied). For measuring the closeness between $\mathcal{L}_n$ and $\mathcal{L}$, consider the bounded Lipschitz type metric on $\mathcal{P}_{\mathbb{T}}$

$$d_{BL}(\mathcal{L},\mathcal{L}') = \sup_{f\in \mathrm{Lip}_{\mathbb{T}}^1}\{\int f(b)\mathcal{L}(db) - \int f(b)\mathcal{L}'(db)\}, \tag{24}$$

for any $\mathcal{L}$, $\mathcal{L}'$ in $\mathcal{P}_{\mathbb{T}}$, denoting by $\mathrm{Lip}_{\mathbb{T}}^1$ the set of functions $F : \mathbb{T} \to \mathbb{R}$ of type $F(b) = \sum_{i=1}^{L(b)} f(b_i)$, $b \in \mathbb{T}$, where $f : E \to \mathbb{R}$ is such that $\sup_{x\in E}|f(x)| \leqslant 1$ and is 1-Lipschitz. Other metrics (of Zolotarev type for instance, *cf* [160]) may be considered. In the general Harris case, the influence function based on the atom of the split chain, as well as the empirical distribution of the (unobserved) regeneration blocks have to be approximated to be of practical interest. Once again, we shall use the approximate regeneration blocks $\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n-1}$ (using *Algorithm 2, 3*) in the general case and consider

$$\widehat{\mathcal{L}}_n = (\widehat{l}_n - 1)\sum_{j=1}^{\widehat{l}_n-1}\delta_{\widehat{\mathcal{B}}_j},$$

when $\widehat{l}_n \geqslant 2$. The next theorem gives an asymptotic bound for the error committed by replacing the empirical distribution $\mathcal{L}_n$ of the true regeneration blocks by $\widehat{\mathcal{L}}_n$, when measured by d_{BL}.

Theorem 11 *(Bertail & Clémençon, 2006a) Under $\mathcal{H}_0'(4), \mathcal{H}_0'(4,\nu), \mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$, we have*

$$d_{BL}(\mathcal{L}_n,\widehat{\mathcal{L}}_n) = O(\alpha_n^{1/2}), \text{ as } n \to \infty.$$

And if in addition $d_{BL}(\mathcal{L}_n,\mathcal{L}) = O(n^{-1/2})$ as $n \to \infty$, then

$$d_{BL}(\mathcal{L}_n,\widehat{\mathcal{L}}_n) = O(\alpha_n^{1/2}n^{-1/2}), \text{ as } n \to \infty.$$

Given the metric on $\mathcal{P}_{\mathbb{T}}$ defined by d_{BL}, we consider now the *Fréchet differentiability* for functionals $T : \mathcal{P}_{\mathbb{T}} \to \mathbb{R}$.

Definition 12 *We say that* T *is Fréchet-differentiable at* $\mathcal{L}_0 \in \mathcal{P}_{\mathbb{T}}$*, if there exists a linear operator* $DT^{(1)}_{\mathcal{L}_0}$ *and a function* $\epsilon^{(1)}(.,\mathcal{L}_0)$*:* $\mathbb{R} \to \mathbb{R}$*, continuous at* 0 *with* $\epsilon^{(1)}(0,\mathcal{L}_0) = 0$*, s.t.:*

$$\forall \mathcal{L} \in \mathcal{P}_{\mathbb{T}},\ T(\mathcal{L}) - T(\mathcal{L}_0) = D^{(1)}T_{\mathcal{L}_0}(\mathcal{L} - \mathcal{L}_0) + R^{(1)}(\mathcal{L},\mathcal{L}_0),$$

with $R^{(1)}(\mathcal{L},\mathcal{L}_0) = d_{BL}(\mathcal{L},\mathcal{L}_0)\epsilon^{(1)}(d_{BL}(\mathcal{L},\mathcal{L}_0),\mathcal{L}_0)$*. Moreover,* T *is said to have a canonical gradient (or influence function)* $T^{(1)}(.,\mathcal{L}_0)$*, if the following representation for* $DT^{(1)}_{\mathcal{L}_0}$ *holds:*

$$\forall \mathcal{L} \in \mathcal{P}_{\mathbb{T}},\ DT^{(1)}_{\mathcal{L}_0}(\mathcal{L} - \mathcal{L}_0) = \int_{\mathbb{T}} T^{(1)}(b,\mathcal{L}_0)\mathcal{L}(db).$$

Now it is easy to see that from this notion of differentiability on the torus one may directly derive CLT's, provided the distance $d(\mathcal{L}_n,\mathcal{L})$ may be controlled.

Theorem 13 *(Bertail & Clémençon, 2006a) In the regenerative case, if* $T : \mathcal{P}_{\mathbb{T}} \to \mathbb{R}$ *is Fréchet differentiable at* $\mathcal{L}$ *and* $d_{BL}(\mathcal{L}_n,\mathcal{L}) = O_{\mathbb{P}_\nu}(n^{-1/2})$ *(or* $R^{(1)}(\mathcal{L}_n,\mathcal{L}) = o_{\mathbb{P}_\nu}(n^{-1/2})$*) as* $n \to \infty$*, and if* $\mathbb{E}_A(\tau_A) < \infty$ *and* $0 < Var_A(T^{(1)}(\mathcal{B}_1,\mathcal{L})) < \infty$ *then under* $\mathbb{P}_\nu$*,*

$$n^{1/2}(T(\mathcal{L}_n) - T(\mathcal{L})) \Rightarrow \mathcal{N}(0,\mathbb{E}_A(\tau_A)Var_A(T^{(1)}(\mathcal{B}_1,\mathcal{L})),\ \text{as } n \to \infty.$$

In the general Harris case, if the split chain satisfies the assumptions above (with A *replaced by* $A_{\mathcal{M}}$*), under the assumptions of Theorem 11, as* $n \to \infty$ *we have under* $\mathbb{P}_\nu$*,*

$$n^{1/2}(T(\widehat{\mathcal{L}}_n) - T(\mathcal{L})) \Rightarrow \mathcal{N}(0,\mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})Var_{A_{\mathcal{M}}}(T^{(1)}(\mathcal{B}_1,\mathcal{L})).$$

Observe that if one renormalizes by $l_n^{1/2}$ instead of renormalizing by $n^{1/2}$ in the atomic case (resp., by $\widehat{l}_n^{1/2}$ in the general case), one would simply get $\mathcal{N}(0, Var_A(T^{(1)}(\mathcal{B}_1,\mathcal{L}))$ (respectively, $Var_{A_{\mathcal{M}}}(T^{(1)}(\mathcal{B}_1,\mathcal{L}))$ as asymptotic distribution, which depends on the atom chosen (resp. on the parameters of condition $\mathcal{M}$).

Going back to the preceding examples, we immediately deduce the results stated below.

a) Since $n^{1/2}/l_n^{1/2} \to \mathbb{E}_A(\tau_A)^{1/2}$ $\mathbb{P}_\nu$- a.s. as $n \to \infty$, we get that under $\mathbb{P}_\nu$,

$$n^{1/2}(\mu_n(f) - \mu(f)) \Rightarrow \mathcal{N}(0,\mathbb{E}_A(\tau_A)^{-1}Var_A(\sum_{i=1}^{\tau_A}(f(X_i) - \mu(f)) \text{ as } n \to \infty.$$

b) In a similar fashion, under smoothness assumptions ensuring Fréchet differentiability, the M-estimator $\widehat{\theta}_n$ being the (unique) solution of the block-estimating equation

$$\sum_{i=\tau_A+1}^{\tau_A(l_n)} \psi(X_i,\theta) = \sum_{j=1}^{l_n} \sum_{i=\tau_A(j)+1}^{\tau_A(j+1)} \psi(X_i,\theta) = 0,$$

we formally obtain that, if $\mathbb{E}_A(\sum_{i=1}^{\tau_A} \frac{\partial\psi(X_i,\theta)}{\partial\theta}) \neq 0$ and θ is the true value of the parameter, then under $\mathbb{P}_\nu$, as $n \to \infty$,

$$n^{1/2}(\widehat{\theta}_n - \theta) \Rightarrow \mathcal{N}(0,\ [\frac{\mathbb{E}_A(\sum_{i=1}^{\tau_A} \frac{\partial\psi(X_i,\theta)}{\partial\theta})}{\mathbb{E}_A(\tau_A)}]^{-2}\frac{Var_A(\sum_{i=1}^{\tau_A}\psi(X_i,\theta))}{\mathbb{E}_A(\tau_A)}).$$

Observe that both factors in the variance are independent from the atom A chosen. It is worth noticing that, by writing the asymptotic variance in this way, as a function of the distribution of the blocks, a consistent estimator for the latter is readily available, from the (approximate) regeneration blocks. Examples c) and d) may be treated similarly.

The concepts developed in [23] may also serve as a tool for robustness purpose, for deciding whether a specific data block has an important influence on the value of some given estimate or not, and/or whether it may be considered as *outlier*. The concept of robustness introduced here is related to blocks of observations, instead of individual observations. Heuristically, one may consider that, given the regenerative dependency structure of the process, a single suspiciously outlying value at some time point n may have a strong impact on the trajectory, until the (split) chain regenerates again, so that not only this particular observation but the whole "contaminated" segment of observations should be eventually removed. Roughly stated, it turns out that examining (approximate) regeneration blocks as proposed before, allows to identify more accurately outlying data in the sample path, as well as their nature (in the time series context, different type of outliers may occur, such as additive or innovative outliers). By comparing the data blocks (their length, as well as the values of the functional of interest on these blocks) this way, one may detect the ones to remove eventually from further computations.

0.8 Some Extreme Values Statistics

We now turn to statistics related to the extremal behaviour of functionals of type $f(X_n)$ in the atomic positive Harris recurrent case, where $f : (E, \mathcal{E}) \to \mathbb{R}$ is a given measurable function. More precisely, we shall focus on the limiting distribution of the maximum $M_n(f) = \max_{1 \leqslant i \leqslant n} f(X_i)$ over a trajectory of length n, in the case when the chain X possesses an accessible atom A (see [9] and the references therein for various examples of such processes X in the area of queuing systems and a theoretical study of the tail properties of $M_n(f)$ in this setting).

0.8.1 Submaxima over regeneration blocks

For $j \geqslant 1$, we define the *submaximum* over the j-th cycle of the sample path:

$$\zeta_j(f) = \max_{1+\tau_A(j) \leqslant i \leqslant \tau_A(j+1)} f(X_i).$$

The $\zeta_j(f)$'s are i.i.d. r.v.'s with common d.f. $G_f(x) = \mathbb{P}(\zeta_1(f) \leqslant x)$. The following result established by [171] shows that the limiting distribution of the sample maximum of $f(X)$ is entirely determined by the tail behaviour of the df G_f and relies on the crucial observation that the maximum value $M_n(f) = \max_{1 \leqslant i \leqslant n} f(X_i)$ over a trajectory of length n, may be expressed in terms of submaxima over regeneration blocks as follows

$$M_n(f) = \max(\zeta_0(f), \max_{1 \leqslant j \leqslant l_n - 1} \zeta_j(f), \zeta_{l_n}^{(n)}(f)),$$

where $\zeta_0(f) = \max_{1 \leqslant i \leqslant \tau_A} f(X_i)$ and $\zeta_{l_n}^{(n)}(f) = \max_{1+\tau_A(l_n) \leqslant i \leqslant n} f(X_i)$ denote the maxima over the non regenerative data blocks, and with the usual convention that the maximum over an empty set equals $-\infty$.

Proposition 14 *(Rootzén, 1988) Let $\alpha = \mathbb{E}_A(\tau_A)$ be the mean return time to the atom A. Under the assumption (A1) that the first (non-regenerative) block does not affect the extremal behaviour, i.e. $\mathbb{P}_\nu(\zeta_0(f) > \max_{1 \leqslant k \leqslant l} \zeta_k(f)) \to 0$ as $l \to \infty$, we have*

$$\sup_{x \in \mathbb{R}} | \mathbb{P}_\nu(M_n(f) \leqslant x) - G_f(x)^{n/\alpha} | \to 0, \text{ as } n \to \infty. \tag{25}$$

Hence, as soon as condition (A1) is fulfilled, the asymptotic behaviour of the sample maximum may be deduced from the tail properties of G_f. In particular, the limiting distribution of $M_n(f)$

(for a suitable normalization) is the extreme df $H_\xi(x)$ of shape parameter $\xi \in \mathbb{R}$ (with $H_\xi(x) = \exp(-x^{-1/\xi})\mathbb{I}\{x > 0\}$ when $\xi > 0$, $H_0(x) = \exp(-\exp(-x))$ and $H_\xi(x) = \exp(-(-x)^{-1/\xi})\mathbb{I}\{x < 0\}$ if $\xi < 0$) iff G_f belongs to the maximum domain of attraction $MDA(H_\xi)$ of the latter df (refer to Resnick (1987) for basics in extreme value theory). Thus, when $G_f \in MDA(H_\xi)$, there are sequences of norming constants a_n and b_n such that $G_f(a_n x + b_n)^n \to H_\xi(x)$ as $n \to \infty$, we then have $\mathbb{P}_\nu(M_n(f) \leqslant a'_n x + b_n) \to H_\xi(x)$ as $n \to \infty$, with $a'_n = a_n/\alpha^\xi$.

0.8.2 Tail estimation based on submaxima over cycles

In the case when assumption (A1) holds, one may straightforwardly derive from (25) estimates of $H_{f,\,n}(x) = \mathbb{P}_\nu(M_n(f) \leqslant x)$ as $n \to \infty$ based on the observation of a random number of submaxima $\zeta_j(f)$ over a sample path, as proposed in [92]:

$$\widehat{H}_{f,\,n,\,l}(x) = (\widehat{G}_{f,\,n}(x))^l,$$

with $1 \leqslant l \leqslant l_n$ and denoting by $\widehat{G}_{f,\,n}(x) = \frac{1}{l_n - 1}\sum_{i=1}^{l_n-1}\mathbb{I}\{\zeta_j(f) \leqslant x\}$ the empirical df of the $\zeta_j(f)$'s (with $\widehat{G}_{f,\,n}(x) = 0$ by convention when $l_n \leqslant 1$). We have the following limit result (see also Proposition 3.6 in [92] for a different formulation, stipulating the observation of a deterministic number of regeneration cycles).

Proposition 15 *(Bertail & Clémençon, 2006a) Let (u_n) be such that $n(1 - G_f(u_n))/\alpha \to \eta < \infty$ as $n \to \infty$. Suppose that assumptions $\mathcal{H}_0(1, \nu)$ and (A1) holds, then $H_{f,\,n}(u_n) \to \exp(-\eta)$ as $\eta \to \infty$. And let $N_n \in \mathbb{N}$ such that $N_n/n^2 \to 0$ as $n \to \infty$, then we have*

$$\widehat{H}_{f,\,N_n,\,l_n}(u_n)/H_{f,\,n}(u_n) \to 1 \text{ in } \mathbb{P}_\nu\text{- probability, as } n \to \infty. \tag{26}$$

Moreover if $N_n/n^{2+\rho} \to \infty$ as $n \to \infty$ for some $\rho > 0$, this limit result also holds $\mathbb{P}_\nu$- a.s. .

This result indicates that observation of a trajectory of length N_n, with $n^2 = o(N_n)$ as $n \to \infty$, is required for estimating consistently the extremal behaviour of the chain over a trajectory of length n. As shall be shown below, it is nevertheless possible to estimate the tail of the sample maximum $M_n(f)$ from the observation of a sample path of length n only, when assuming some type of behaviour for the latter, namely under maximum domain of attraction hypotheses. As a matter of fact, if one assume that $G_f \in MDA(H_\xi)$ for some $\xi \in \mathbb{R}$, of which sign is *a priori* known, one may implement classical inference procedures (refer to §6.4 in [80] for instance) from the observed submaxima $\zeta_1(f), \ldots, \zeta_{l_n-1}(f)$ for estimating the shape parameter ξ of the extremal distribution, as well as the norming constants a_n and b_n. We now illustrate this point in the Fréchet case (*i.e.* when $\xi > 0$), through the example of the Hill method.

0.8.3 Heavy-tailed stationary distribution

As shown in [171], when the chain takes real values, assumption (A1) is checked for $f(x) = x$ (for this specific choice, we write $M_n(f) = M_n$, $G_f = G$, and $\zeta_j(f) = \zeta_j$ in what follows) in the particular case when the chain is stationary, i.e. when $\nu = \mu$. Moreover, it is known that when the chain is positive recurrent there exists some index θ, namely the *extremal index* of the sequence $X = (X_n)_{n\in\mathbb{N}}$ (see [131] for instance), such that

$$\mathbb{P}_\mu(M_n \leqslant x) \underset{n\to\infty}{\sim} F_\mu(x)^{n\theta}, \tag{27}$$

denoting by $F_\mu(x) = \mu(]-\infty, x]) = \alpha\mathbb{E}_A(\sum_{i=1}^{\tau_A}\mathbb{I}\{X_i \leqslant x\})$ the stationary df. In this case, as remarked in [171], if (u_n) is such that $n(1 - G(u_n))/\alpha \to \eta < \infty$, we deduce from Proposition 15

and (27) that

$$\theta = \lim_{n\to\infty} \frac{\mathbb{P}_A(\max_{1\leqslant i\leqslant \tau_A} X_i > u_n)}{\mathbb{E}_A(\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i > u_n\})}.$$

In [23] a natural estimate of the extremal index θ based on the observation of a trajectory of length N has been proposed,

$$\widehat{\theta}_N = \frac{\sum_{j=1}^{l_N-1} \mathbb{I}\{\zeta_j > u_n\}}{\sum_{i=1}^{N} \mathbb{I}\{X_i > u_n\}},$$

which may be shown to be consistent (resp., strongly consistent) under $\mathbb{P}_\mu$ when $N = N_n$ is such that $N_n/n^2 \to \infty$ (resp. $N_n/n^{2+\rho} \to \infty$ for some $\rho > 0$) as $n \to \infty$ and $\mathcal{H}_0(2)$ is fulfilled. And Proposition 14 combined with (27) also entails that for all ξ in $\mathbb{R}$,

$$G \in \mathrm{MDA}(H_\xi) \;\Leftrightarrow F_\mu \in \mathrm{MDA}(H_\xi).$$

0.8.4 Regeneration-based Hill estimator

This crucial equivalence holds in particular in the Fréchet case, *i.e.* for $\xi > 0$. Recall that assuming that a df F belongs to $\mathrm{MDA}(H_\xi)$ classically amounts then to suppose that it satisfies the tail regularity condition

$$1 - F(x) = L(x)x^{-a},$$

where $a = \xi^{-1}$ and L is a slowly varying function, *i.e.* a function L such that $L(tx)/L(x) \to 1$ as $x \to \infty$ for any $t > 0$ (*cf* Theorem 8.13.2 in [32]). Since the seminal contribution of [110], numerous papers have been devoted to the development and the study of statistical methods in the i.i.d. setting for estimating the tail index $a > 0$ of a regularly varying df. Various inference methods, mainly based on an increasing sequence of upper order statistics, have been proposed for dealing with this estimation problem, among which the popular *Hill estimator*, relying on a conditional maximum likelihood approach. More precisely, based on i.i.d. observations $X_1, \ldots, X_n$ drawn from F, the Hill estimator is given by

$$H^X_{k,\,n} = (k^{-1} \sum_{i=1}^{k} \ln \frac{X_{(i)}}{X_{(k+1)}})^{-1}, \tag{28}$$

where $X_{(i)}$ denotes the i-th largest order statistic of the sample $X^{(n)} = (X_1, \ldots, X_n)$, $1 \leqslant i \leqslant n$, $1 \leqslant k < n$. Strong consistency (*cf* [71]) of this estimate has been established when $k = k_n \to \infty$ at a suitable rate, namely for $k_n = o(n)$ and $\ln\ln n = o(k_n)$ as $n \to \infty$, as well as asymptotic normality (see [94]) under further conditions on F and k_n, $\sqrt{k_n}(H^X_{k_n,n} - a) \Rightarrow \mathcal{N}(0,\, a^2)$, as $n \to \infty$. Now let us define the *regeneration-based Hill estimator* from the observation of the $l_n - 1$ submaxima $\zeta_1, \ldots, \zeta_{l_n-1}$, denoting by $\xi_{(j)}$ the j-th largest submaximum,

$$\widehat{a}_{n,\,k} = H^\zeta_{k,\,l_n-1} = (k^{-1} \sum_{i=1}^{k} \ln \frac{\zeta_{(i)}}{\zeta_{(k+1)}})^{-1}.$$

Given that $l_n \to \infty$, $\mathbb{P}_\nu$- a.s. as $n \to \infty$, results established in the case of i.i.d. observations straightforwardly extend to our setting (for comparison purpose, see [166] for properties of the classical Hill estimate in dependent settings).

Proposition 16 *(Bertail & Clémençon, 2006a) Suppose that* $F_\mu \in \mathrm{MDA}(H_{a^{-1}})$ *with* $a > 0$. *Let* (k_n) *be an increasing sequence of integers such that* $k_n \leqslant n$ *for all* n, $k_n = o(n)$ *and* $\ln\ln n = o(k_n)$ *as* $n \to \infty$. *Then the regeneration-based Hill estimator is strongly consistent*

$$\widehat{a}_{n,\,k_{l_n-1}} \to a,\ \mathbb{P}_\nu\text{- a.s., as } n \to \infty.$$

Under the further assumption that F_μ *satisfies the Von Mises condition and that* k_n *is chosen accordingly (cf [94]), it is moreover asymptotically normal in the sense that*

$$\sqrt{k_{l_n-1}}(\widehat{a}_{n,\,k_{l_n-1}} - a) \Rightarrow \mathcal{N}(0,\ a^2) \text{ under } \mathbb{P}_\nu, \text{ as } n \to \infty.$$

Regenerative-Block Bootstrap for Harris Markov chains

0.9 Introduction

In the statistical literature there has been substantial interest in transposing the naive bootstrap method (ses [78]) introduced in the i.i.d. setting to dependent settings. The now well known idea of the *moving-block bootstrap* (MBB) is to resample (overlapping or disjoint) blocks of observations to capture the dependence structure of the observations (see [130] for a recent survey and exhaustive references). However, as noticed by many authors, the results obtained by using this method are not completely satisfactory for the following reasons. First, the MBB approach usually requires stationarity for the observations and generally fails in a general nonstationary framework. Secondly, the asymptotic behaviour of the MBB distribution crucially depends on the estimation of the bias and of the asymptotic variance of the statistic of interest, which makes it difficult to apply in practice (see [96], [130]). From a theoretical viewpoint, the rate of convergence of the MBB distribution is much slower than the one of the bootstrap in the i.i.d. case: at best it is of order $O_{\mathbb{P}}(n^{-3/4})$ under restrictive conditions, stipulating the finiteness of moments at any order and an exponential rate for the decay of the strong mixing coefficients, while the bootstrap achieves $O_{\mathbb{P}}(n^{-1})$ in the i.i.d. setting. Finally, the choice of the size of the blocks is a key point to get an accurate estimation: this practical problem still remains open in the general case.

Recently, various authors have been interested in bootstrapping some particular types of Markov chain (see [130], [85] and references therein). However, second order results in this framework are scarcely available, except in [113] for discrete Markov chains. Unfortunately, these results are weakened by the unrealistic technical assumptions (m-dependence) made on the Markovian models considered. Most bootstrap methods proposed in the literature are all asymptotically equivalent at the first order. And obtaining their exact rate of convergence is thus of prime importance for helping pratictioners to choose a particular bootstrap technique. In [22] a specific bootstrap method based on the renewal properties of Markov chains has been proposed, which almost achieves the same rate (up to $\log$ factor) as the one in the i.i.d. case in a general (eventually nonstationary) framework.

This method originates from [11] and [67], which exploits the regeneration properties of Markov chains when a (recurrent) state is infinitely often visited. The main idea underlying this method consists in resampling a deterministic number of data blocks corresponding to regeneration cycles. However, because of some inadequate standardization, the *regeneration-based bootstrap* method proposed in [67] is not second order correct (its rate is $O_{\mathbb{P}}(n^{-1/2})$ only). In [21] a modification of the procedure introduced by [67] is proposed, which is second order correct up to $O_{\mathbb{P}}(n^{-1}\log(n))$ in the unstudentized case (*i.e.* when the variance is known) when the chain is stationary. However, this method fails to be second order correct in the nonstationary case, as a careful look at the Edgeworth expansions (E.E.) of the statistic of interest shows (see Theorem 4, refer also to [19], [20]). The proposal in [22] consists in imitating further the renewal structure of the chain by

sampling regeneration data blocks, until the length of the reconstructed bootstrap series is larger than the length n of the original data series. In this way, we approximate the distribution of the (random) number of regeneration blocks in a series of length n and remove significant bias terms. This resampling method, which we call the *regenerative block-bootstrap* (RBB), has an uniform rate of convergence of order $O_{\mathbb{P}}(n^{-1})$, that is the optimal rate in the i.i.d case. Unlike the MMB, there is no need in the RBB procedure to choose the size of the blocks, which are entirely determined by the data. Besides, the second order accuracy of the RBB holds under weak conditions (stipulating a polynomial rate for the decay of the strong mixing coefficients only). These results may be extended to the much broader class of Harris Markov chains by using the empirical method to build approximatively a realization drawn from an extension of the chain with a regeneration set described in Chapter 1. This procedure is shown to be asymptotically valid, even in a nonstationary framework, that is clearly more suitable for many applications. Its second order validity is only established in the unstudentized stationary case, up to a rate close to the one in the i.i.d setting. The technical study of the second order properties of this method and of the optimal rate that may be attained in the studentized case will be carried out at length in a forthcoming article. Here we mainly focus on the case of the sample mean in the positive recurrent case, but the ideas set out in this chapter may be straightforwardly extended to much more general functionals (some extensions to V and U statistics are presented in section 0.12) and even to the null recurrent case, when specific models are considered.

0.10 The (approximate) regenerative block-bootstrap algorithm

Although our higher order results are stated in the case of the sample mean only in this chapter, we present here a valid algorithm, applicable to general statistics T_n for which there exists an adequate standardisation S_n : this covers the case of nondegenerate U-statistics (see section 0.12), as well as the case of differentiable functionals. For the reasons mentioned in section 0.6, both the statistic T_n and the estimate of its asymptotic variance we consider are constructed from the true or approximate regeneration blocks $\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n-1}$, obtained by implementing *Algorithm 1, 2, 3* or *4*. The *(approximate) regenerative block-bootstrap* algorithm for estimating accurately the corresponding sampling distribution under $\mathbb{P}_\nu$, say $H^{(n)}_{\mathbb{P}_\nu}(x) = \mathbb{P}_\nu(S_n^{-1}(T_n - \theta) \leqslant x)$, is performed in 3 steps as follows.

Algorithm 17 *(the (A)RBB algorithm, Bertail & Clémençon, 2005b)*

1. *Draw sequentially bootstrap data blocks $\mathcal{B}^*_1, ..., \mathcal{B}^*_k$ independently from the empirical distribution $\widehat{\mathcal{L}}_n = (\widehat{l}_n - 1)^{-1} \sum_{j=1}^{\widehat{l}_n-1} \delta_{\widehat{\mathcal{B}}_j}$ of the initial blocks $\widehat{\mathcal{B}}_1, ..., \widehat{\mathcal{B}}_{\widehat{l}_n-1}$, until the length of the bootstrap data series $l^*(k) = \sum_{j=1}^{k} l(\mathcal{B}^*_j)$ is larger than n. Let $l^*_n = \inf\{k \geqslant 1, l^*(k) > n\}$.*

2. *From the bootstrap data blocks generated at step 1, reconstruct a pseudo-trajectory by binding the blocks together, getting the reconstructed (A)RBB sample path*

$$X^{*(n)} = (\mathcal{B}^*_1, ..., \mathcal{B}^*_{l^*_n-1}).$$

Then compute the (A)RBB statistic and its (A)RBB standardization

$$T^*_n = T(X^{*(n)}) \text{ and } S^*_n = S(X^{*(n)}).$$

3. *The (A)RBB distribution is then given by*

$$H_{(A)RBB}(x) = \mathbb{P}^*(S_n^{*-1}(T_n^* - T_n) \leqslant x),$$

where $\mathbb{P}^*$ denotes the conditional probability given the original data.

One may naturally compute a Monte-Carlo approximation to $H_{RBB}(x)$ by repeating independently the procedure above B times.

We point out that the RBB differs from the regeneration-based bootstrap proposed by [67], which is not second order correct up to $O_{\mathbb{P}}(n^{-1/2})$, (and from its modified version in [21] too) in which the number of resampled blocks is held fixed to $l_n - 1$, conditionally to the sample. By generating this way a random number $l_n^* - 1$ of bootstrap regenerative blocks, the bootstrap data series mimics the renewal properties of the chain, although it is not markovian. Consequently, the usual properties of the i.i.d. bootstrap cannot be directly used for studying the RBB method, contrary to the regeneration-based bootstrap studied in [21].

We also emphasize that the principles underlying the RBB may be applied to any (eventually continuous time) regenerative process (and not necessarily markovian). In [89] for instance, an extension of the ARBB procedure to diffusion processes has been investigated.

0.11 Main asymptotic results

0.11.1 Second order accuracy of the RBB

Here we consider the asymptotic validity of the RBB for the sample mean standardized by an adequate estimator of the asymptotic variance. This is the useful version for confidence intervals but also for practical use of the bootstrap (see [102]). The accuracy reached by the RBB is similar to the optimal rate of the bootstrap distribution in the i.i.d. case, contrary to the MBB (see [96]). The proof relies on the E.E. stated in Theorem 4 for the studentized sample mean established in [19], which result mainly derives from the methods used in [138] to obtain the E.E. for the unstandardized sample mean (see also [138], [139] and [36]).

In the case of the sample mean, the bootstrap counterparts of the estimators $\overline{\mu}_n(f)$ and $\sigma_n^2(f)$ considered in §2.2.1 are

$$\mu_n^*(f) = n_A^{*-1} \sum_{j=1}^{l_n^*-1} f(\mathcal{B}_j^*) \text{ and } \sigma_n^{*2}(f) = n_A^{*-1} \sum_{j=1}^{l_n^*-1} \left\{ f(\mathcal{B}_j^*) - \mu_n^*(f) l(\mathcal{B}_j^*) \right\}^2, \tag{29}$$

with $n_A^* = \sum_{j=1}^{l_n^*-1} l(\mathcal{B}_j^*)$.

As shall be shown below, this standardization does not deteriorate the performance of the RBB, while the standardization of the MBB distribution in the strong mixing case is the main barrier to achieve good performance (as shown by [96]). Moreover, in opposition to the MBB, the bootstrap counterparts in the studentized case may be defined straightforwardly in our regenerative setting. Let us consider the RBB distribution estimates of the unstandardized and studentized sample means

$$H_{RBB}^U(x) = \mathbb{P}^*(n_A^{1/2}\sigma_n(f)^{-1}\{\mu_n^*(f) - \overline{\mu}_n(f)\} \leq x),$$
$$H_{RBB}^S(x) = \mathbb{P}^*(n_A^{*-1/2}\sigma_n^{*-1}(f)\{\mu_n^*(f) - \overline{\mu}_n(f)\} \leq x).$$

The following theorem established in [22] shows the RBB is asymptotically valid for the sample mean. Moreover it ensures that the RBB attains the optimal rate of the i.i.d. Bootstrap.

This is noteworthy, since the RBB method applies to countable chains (for which any recurrent state is an atom) but also to many specific Markov chains widely used in practice for modelling queuing/storage systems (see §2.4 in [148] and [8] for a detailed account of such models).

Theorem 18 *(Bertail & Clémençon, 2005b) Suppose that (C1) is satisfied. Under $\mathcal{H}_0'(2, \nu)$, $\mathcal{H}_1'(2, f, \nu)$, $\mathcal{H}_0'(\kappa)$ and $\mathcal{H}_1(\kappa, f)$ with $\kappa > 6$, the RBB distribution estimate for the unstandardized sample mean is second order accurate in the sense that*

$$\Delta_n^U = \sup_{x\in\mathbb{R}} |H_{RBB}^U(x) - H_\nu^U(x)| = O_{\mathbb{P}_\nu}(n^{-1}), \text{ as } n \to \infty,$$

with $H_\nu^U(x) = \mathbb{P}_\nu(n_A^{1/2}\sigma_f^{-1}\{\overline{\mu}_n(f) - \mu(f)\} \leq x)$. And if in addition (C4), $\mathcal{H}_0'(\kappa)$ and $\mathcal{H}_1(\kappa, f)$ are checked with $\kappa > 8$, the RBB distribution estimate for the standardized sample mean is also 2nd order correct

$$\Delta_n^S = \sup_{x\in\mathbb{R}} |H_{RBB}^S(x) - H_\nu^S(x)| = O_{\mathbb{P}_\nu}(n^{-1}), \text{ as } n \to \infty,$$

with $H_\nu^S(x) = \mathbb{P}_\nu(n_A^{1/2}\sigma_n^{-1}(f)\{\overline{\mu}_n(f) - \mu(f)\} \leq x)$.

The same results holds *a.s.* up to $O_{\mathbb{P}_\nu}(n^{-1}\log\log(n)^{1/2})$, like in the i.i.d. case under the same moment conditions. This results from the LIL applied to the empirical moments of the blocks appearing in the E.E. of the RBB distribution. And if one is interested in getting the second order validity up to $o_{\mathbb{P}_\nu}(n^{-1/2})$ only, then a careful examination of [138] and [19] shows that $\kappa > 3$ (resp. $\kappa > 4$) is sufficient in the standardized case (resp. the studentized case).

0.11.2 Asymptotic validity of the ARBB for general chains

The ARBB counterparts of statistics $\widehat{\mu}_n(f)$ and $\widehat{\sigma}_n^2(f)$ considered in §2.2.2 may be expressed as

$$\mu_n^*(f) = n_{A_\mathcal{M}}^{*-1} \sum_{j=1}^{l_n^*-1} f(\mathcal{B}_j^*)$$

and

$$\sigma_n^{*2}(f) = n_{A_\mathcal{M}}^{*-1} \sum_{j=1}^{l_n^*-1} \left\{f(\mathcal{B}_j^*) - \mu_n^*(f)l(\mathcal{B}_j^*)\right\}^2,$$

denoting by $n_{A_\mathcal{M}}^* = \sum_{j=1}^{l_n^*-1} l(\mathcal{B}_j^*)$ the length of the ARBB data series. Define the ARBB versions of the pseudo-regeneration based unstudentized and studentized sample means (*cf* §2.2.2) by

$$\widehat{\sigma}_n^* = n_{A_\mathcal{M}}^{1/2} \frac{\mu_n^*(f) - \widehat{\mu}_n(f)}{\widehat{\sigma}_n(f)} \text{ and } \widehat{t}_n^* = n_{A_\mathcal{M}}^{*1/2} \frac{\mu_n^*(f) - \widehat{\mu}_n(f)}{\sigma_n^*(f)}.$$

The unstandardized and studentized version of the ARBB distribution estimates are then given by

$$H_{ARBB}^U(x) = \mathbb{P}^*(\widehat{\sigma}_n^* \leq x \mid X^{(n+1)}) \text{ and } H_{ARBB}^S(x) = \mathbb{P}^*(\widehat{t}_n^* \leq x \mid X^{(n+1)}).$$

This is the same construction as in the atomic case, except that one uses the approximate regeneration blocks instead of the exact regenerative ones (*cf* Theorem 3.3 in [22]).

Theorem 19 *(Bertail & Clémençon, 2005b) Under the hypotheses of Theorem 4.2, we have the following convergence results in distribution under $\mathbb{P}_\nu$*

$$\Delta_n^U = \sup_{x\in\mathbb{R}} |H_{ARBB}^U(x) - H_\nu^U(x)| \to 0, \text{ as } n \to \infty,$$
$$\Delta_n^S = \sup_{x\in\mathbb{R}} |H_{ARBB}^S(x) - H_\nu^S(x)| \to 0, \text{ as } n \to \infty.$$

Second order properties of the ARBB using the 2-split trick. To bypass the technical difficulties related to the dependence problem induced by the preliminary step estimation, assume now that the pseudo regenerative blocks are constructed according to Algorithm 4 (possibly including the selection rule for the small set of Algorithm 3). It is then easier (at the price of a small loss in the 2nd order term) to get second order results both in the case of standardized and studentized statistics, as stated below.

Theorem 20 *(Bertail & Clémençon, 2004b) Suppose that (C1) and (C4) are satisfied by the split chain. Under assumptions $\mathcal{H}_0'(\kappa, \nu)$, $\mathcal{H}_1'(\kappa, f, \nu)$, $\mathcal{H}_0'(f, \kappa)$, $\mathcal{H}_1'(f, \kappa)$ with $\kappa > 6$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$, we have the second order validity of the ARBB distribution both in the standardized and unstandardized case up to order*

$$\Delta_n^U = O_{\mathbb{P}_\nu}(n^{-1/2}\alpha_{m_n}^{1/2} \vee n^{-1/2}n^{-1}m_n\}), \text{ as } n \to \infty.$$

And if in addition these assumptions hold with $k > 8$, we have

$$\Delta_n^S = O_{\mathbb{P}_\nu}(n^{-1/2}\alpha_{m_n}^{1/2} \vee n^{-1/2}n^{-1}m_n), \text{ as } n \to \infty$$

In particular if $\alpha_m = m\log(m)$, by choosing $m_n = n^{2/3}$, the ARBB is second order correct up to $O(n^{-5/6}\log(n))$.

It is worth noticing that the rate that can be attained by the 2-split trick variant of the ARBB for such chains is faster than the optimal rate the MBB may achieve, which is typically of order $O(n^{-3/4})$ under very strong assumptions (see [96], [130]). Other variants of the bootstrap (sieve bootstrap) for time-series may also yield (at least practically) very accurate approximation (see [47], [46]). When some specific non-linear structure is assumed for the chain (see our example 3 in §2.2.3), nonparametric method estimation and residual based resampling methods may also be used : see for instance [85]. However to our knowledge, no rate of convergence is explicitly available for these bootstrap techniques. An empirical comparison of all these recent methods may be found in [24] (see also section 3.5 below).

0.12 Some extensions to U-statistics

We now turn to extend some of the asymptotic results stated in section 0.11 for sample mean statistics to a wider class of functionals and shall consider statistics of the form $\sum_{1\leqslant i\neq j\leqslant n} U(X_i, X_j)$. For the sake of simplicity, we confined the study to U-statistics of degree 2, in the real case only. As will be shown below, asymptotic validity of inference procedures based on such statistics does not straightforwardly follow from results established in the previous sections, even for atomic chains. Furthermore, whereas asymptotic validity of the (approximate) regenerative block-bootstrap for these functionals may be easily obtained, establishing its second order validity and give precise rate is much more difficult from a technical viewpoint and is left to a further study. Besides, arguments presented in the sequel may be easily adapted to V-statistics $\sum_{1\leqslant i,\, j\leqslant n} U(X_i, X_j)$.

0.12.1 Regenerative case

Given a trajectory $X^{(n)} = (X_1, ..., X_n)$ of a Harris positive atomic Markov chain with stationary probability law μ, we shall consider in the following U-statistics of the form

$$T_n = \frac{1}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} U(X_i,\ X_j), \tag{30}$$

where $U : E^2 \to \mathbb{R}$ is a kernel of degree 2. Even if it entails introducing the symmetrized version of T_n, it is assumed throughout the section that the kernel $U(x, y)$ is symmetric. Although such statistics have been mainly used and studied in the case of i.i.d. observations, in dependent settings such as ours, these statistics are also of interest, as shown by the following examples.

• In the case when the chain takes real values and is positive recurrent with stationary distribution μ, the variance of the stationary distribution $s^2 = \mathbb{E}_\mu((X - \mathbb{E}_\mu(X))^2)$, if well defined (note that it differs in general from the asymptotic variance of the mean statistic studied in section 2.2), may be consistently estimated under adequate block moment conditions by

$$\widehat{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_n)^2 = \frac{1}{n(n-1)}\sum_{1\leqslant i\neq j\leqslant n}(X_i - X_j)^2/2,$$

where $\mu_n = n^{-1}\sum_{i=1}^{n} X_i$, which is a U-statistic of degree 2 with symmetric kernel $U(x, y) = (x - y)^2/2$.

• In the case when the chain takes its values in the multidimensional space $\mathbb{R}^p$, endowed with some norm $\|.\|$, many statistics of interest may be written as a U-statistic of the form

$$U_n = \frac{1}{n(n-1)}\sum_{1\leqslant i\neq j\leqslant n} H(\|X_i - X_j\|),$$

where $H : \mathbb{R} \to \mathbb{R}$ is some measurable function. And in the particular case when $p = 2$, for some fixed t in $\mathbb{R}^2$ and some smooth function h, statistics of type

$$U_n = \frac{1}{n(n-1)}\sum_{1\leqslant i\neq j\leqslant n} h(t,\ X_i,\ X_j)$$

arise in the study of the *correlation dimension* for dynamic systems (see [37]). *Depth statistical functions* for spatial data are also particular examples of such statistics (*cf* [181]).

In what follows, the parameter of interest is

$$\mu(U) = \int_{(x,y)\in E^2} U(x, y)\mu(dx)\mu(dy), \tag{31}$$

which quantity we assume to be finite. As in the case of i.i.d. observations, a natural estimator of $\mu(U)$ in our markovian setting is T_n. Its consistency properties are now detailed and an adequate sequence of renormalizing constants for the latter is exhibited, by using the *regeneration blocks construction* once again. For later use, define $\omega_U : \mathbb{T}^2 \to \mathbb{R}$ by

$$\omega_U(x^{(k)}, y^{(l)}) = \sum_{i=1}^{k}\sum_{j=1}^{l} U(x_i, y_j),$$

for any $x^{(k)} = (x_1, ..., x_k)$, $y^{(l)} = (y_1, ..., y_l)$ in the torus $\mathbb{T} = \cup_{n=1}^{\infty} E^n$ and observe that ω_U is symmetric, as U.

Regeneration-based Hoeffding's decomposition. By the representation of μ as a Pitman's occupation measure (*cf* Theorem 1), we have

$$\mu(U) = \alpha^{-2}\mathbb{E}_A\Big(\sum_{i=1}^{\tau_A(1)}\sum_{l=\tau_A(1)+1}^{\tau_A(2)} U(X_i, X_j)\Big)$$

$$= \alpha^{-2}\mathbb{E}(\omega_U(\mathcal{B}_l, \mathcal{B}_k)),$$

for any integers k, l such that $k \neq l$. In the case of U-statistics based on dependent data, the classical (orthogonal) Hoeffding decomposition (*cf* [180]) does not hold anymore. Nevertheless, we may apply the underlying projection principle for establishing the asymptotic normality of T_n by approximatively rewriting it as a U-statistic of degree 2 computed on the regenerative blocks only, in a fashion very similar to the *Bernstein blocks technique* for strongly mixing random fields (*cf* [75]), as follows. As a matter of fact, the estimator T_n may be decomposed as

$$T_n = \frac{(l_n - 1)(l_n - 2)}{n(n-1)} U_{l_n - 1} + T_n^{(0)} + T_n^{(n)} + \Delta_n, \tag{32}$$

where,

$$U_L = \frac{2}{L(L-1)} \sum_{1 \leqslant k < l \leqslant L} \omega_U(\mathcal{B}_k, \mathcal{B}_l),$$

$$T_n^{(0)} = \frac{2}{n(n-1)} \sum_{1 \leqslant k \leqslant l_n - 1} \omega_U(\mathcal{B}_k, \mathcal{B}_0), \quad T_n^{(n)} = \frac{2}{n(n-1)} \sum_{0 \leqslant k \leqslant l_n - 1} \omega_U(\mathcal{B}_k, \mathcal{B}_{l_n}^{(n)}),$$

$$\Delta_n = \frac{1}{n(n-1)} \{ \sum_{k=0}^{l_n - 1} \omega_U(\mathcal{B}_k, \mathcal{B}_k) + \omega_U(\mathcal{B}_{l_n}^{(n)}, \mathcal{B}_{l_n}^{(n)}) - \sum_{i=1}^{n} U(X_i, X_i) \}.$$

Observe that the "block diagonal part" of T_n, namely Δ_n, may be straightforwardly shown to converge $\mathbb{P}_\nu$- a.s. to 0 as $n \to \infty$, as well as $T_n^{(0)}$ and $T_n^{(1)}$ under obvious block moment conditions (see conditions *(ii)-(iii)* below). And, since $l_n/n \to \alpha^{-1}$ $\mathbb{P}_\nu$- a.s. as $n \to \infty$, asymptotic properties of T_n may be derived from the ones of $U_{l_n - 1}$, which statistic depends on the regeneration blocks only. The key point relies in the fact that the theory of U-statistics based on i.i.d. data may be straightforwardly adapted to functionals of the i.i.d. regeneration blocks of the form $\sum_{k<l} \omega_U(\mathcal{B}_k, \mathcal{B}_l)$. Hence, the asymptotic behaviour of the U-statistic U_L as $L \to \infty$ essentially depends on the properties of the linear and quadratic terms appearing in the following variant of *Hoeffding's decomposition*. For k, $l \geqslant 1$, define

$$\tilde{\omega}_U(\mathcal{B}_k, \mathcal{B}_l) = \sum_{i=\tau_A(k)+1}^{\tau_A(k+1)} \sum_{j=\tau_A(l)+1}^{\tau_A(l+1)} \{U(X_i, X_j) - \mu(U)\}.$$

(notice that $\mathbb{E}(\tilde{\omega}_U(\mathcal{B}_k, \mathcal{B}_l)) = 0$ when $k \neq l$) and for $L \geqslant 1$ write the expansion

$$U_L - \mu(U) = \frac{2}{L} \sum_{k=1}^{L} \omega_U^{(1)}(\mathcal{B}_k) + \frac{2}{L(L-1)} \sum_{1 \leqslant k < l \leqslant L} \omega_U^{(2)}(\mathcal{B}_k, \mathcal{B}_l), \tag{33}$$

where, for any $b_1 = (x_1, ..., x_l) \in \mathbb{T}$,

$$\omega_U^{(1)}(b_1) = \mathbb{E}(\tilde{\omega}_U(\mathcal{B}_1, \mathcal{B}_2) | \mathcal{B}_1 = b_1) = \mathbb{E}_A(\sum_{i=1}^{l} \sum_{j=1}^{\tau_A} \tilde{\omega}_U(x_i, X_j))$$

is the linear term (see also our definition of the *influence function* of the parameter $\mathbb{E}(\omega(\mathcal{B}_1, \mathcal{B}_2))$ in section 0.7) and for all b_1, b_2 in $\mathbb{T}$,

$$\omega_U^{(2)}(b_1, b_2) = \tilde{\omega}_U(b_1, b_2) - \tilde{\omega}_U^{(1)}(b_1) - \tilde{\omega}_U^{(1)}(b_2)$$

is the quadratic degenerate term (gradient of order 2). Notice that by using the Pitman's occupation measure representation of μ, we have as well, for any $b_1 = (x_1, ..., x_l) \in \mathbb{T}$,

$$(\mathbb{E}_A \tau_A)^{-1} \omega_U^{(1)}(b_1) = \sum_{i=1}^{l} \mathbb{E}_\mu(\tilde{\omega}_U(x_i, X_1)).$$

For resampling purposes, consider also the U-statistic based on the data between the first regeneration time and the last one only:

$$\tilde{T}_n = \frac{2}{\tilde{n}(\tilde{n}-1)} \sum_{1+\tau_A \leqslant i < j \leqslant \tau_A(l_n)} U(X_i, X_j),$$

with $\tilde{n} = \tau_A(l_n) - \tau_A$ and $\tilde{T}_n = 0$ when $l_n \leqslant 1$ by convention.

Asymptotic normality and asymptotic validity of the RBB. Suppose that the following conditions, which are involved in the next result, are fulfilled by the chain.

(i) (Non degeneracy of the U-statistic)

$$0 < \sigma_U^2 = \mathbb{E}(\omega_U^{(1)}(\mathcal{B}_1)^2) < \infty.$$

(ii) (Block-moment conditions: linear part) For some $s \geqslant 2$,

$$\mathbb{E}(\omega_{|U|}^{(1)}(\mathcal{B}_1)^s) < \infty \text{ and } \mathbb{E}_\nu(\omega_{|U|}^{(1)}(\mathcal{B}_0)^2) < \infty.$$

(iii) (Block-moment conditions: quadratic part) For some $s \geqslant 2$,

$$\mathbb{E}|\omega_{|U|}(\mathcal{B}_1, \mathcal{B}_2)|^s < \infty \text{ and } \mathbb{E}|\omega_{|U|}(\mathcal{B}_1, \mathcal{B}_1)|^s < \infty,$$
$$\mathbb{E}_\nu|\omega_{|U|}(\mathcal{B}_0, \mathcal{B}_1)|^2 < \infty \text{ and } \mathbb{E}_\nu|\omega_{|U|}(\mathcal{B}_0, \mathcal{B}_0)|^2 < \infty.$$

By construction, under *(ii)*-*(iii)* we have the crucial orthogonality property:

$$\mathrm{Cov}(\omega_U^{(1)}(\mathcal{B}_1),\ \omega_U^{(2)}(\mathcal{B}_1, \mathcal{B}_2)) = 0. \tag{34}$$

A slight modification of the argument given in [111] allows to prove straightforwardly that $\sqrt{L}(U_L - \mu(U))$ is asymptotically normal with zero mean and variance $4\sigma_U^2$. Furthermore, by adapting the classical CLT argument for sample means of Markov chains and using (34) and $l_n/n \to \alpha^{-1}$ $\mathbb{P}_\nu$-a.s. as $n \to \infty$, one deduces that $\sqrt{n}(T_n - \mu(U)) \Rightarrow \mathcal{N}(0, \Sigma^2)$ as $n \to \infty$ under $\mathbb{P}_\nu$, with $\Sigma^2 = 4\alpha^{-3}\sigma_U^2$.

Besides, estimating the normalizing constant is important (for constructing confidence intervals or bootstrap counterparts for instance). So we define the natural estimator $\sigma_{U,\, l_n-1}^2$ of σ_U^2 based on the (asymptotically i.i.d.) $l_n - 1$ regeneration data blocks by

$$\sigma_{U,\, L}^2 = (L-1)(L-2)^{-2} \sum_{k=1}^{L} [(L-1)^{-1} \sum_{l=1, k \neq l}^{L} \omega_U(\mathcal{B}_k, \mathcal{B}_l) - U_L]^2,$$

for $L \geqslant 1$. The estimate $\sigma_{U,\, L}^2$ is a simple transposition of the *jackknife estimator* considered in [48] to our setting and may be easily shown to be strongly consistent (by adapting the SLLN for U-statistics to this specific functional of the i.i.d regeneration blocks). Furthermore, we derive that $\Sigma_n^2 \to \Sigma^2$ $\mathbb{P}_\nu$-a.s., as $n \to \infty$, where

$$\Sigma_n^2 = 4(l_n/n)^3 \sigma_{U,\, l_n-1}^2.$$

We also consider the regenerative block-bootstrap counterparts T_n^* and Σ_n^{*2} of $\tilde{T}_n$ and Σ_n^2 respectively, constructed via *Algorithm 5*:

$$T_n^* = \frac{2}{n^*(n^*-1)} \sum_{1 \leqslant i < j \leqslant n^*} U(X_i^*,\ X_j^*),$$

$$\Sigma_n^{*2} = 4(l_n^*/n^*)^3 \sigma^{*2}_{U,\, l_n^*-1},$$

where n^* denotes the length of the RBB data series $X^{*(n)} = (X_1, ..., X_{n^*})$ constructed from the $l_n^* - 1$ bootstrap data blocks, and

$$\sigma^{*2}_{U,\, l_n^*-1} = (l_n^*-2)(l_n^*-3)^{-2} \sum_{k=1}^{l_n^*-1} [(l_n^*-2)^{-1} \sum_{l=1, k\neq l}^{l_n^*-1} \omega_U(\mathcal{B}_k^*, \mathcal{B}_l^*) - U^*_{l_n^*-1}]^2, \tag{35}$$

$$U^*_{l_n^*-1} = \frac{2}{(l_n^*-1)(l_n^*-2)} \sum_{1\leqslant k<l\leqslant l_n^*-1} \omega_U(\mathcal{B}_k^*, \mathcal{B}_l^*).$$

We may then state the following result.

Theorem 21 *(Bertail & Clémençon, 2006a) If conditions (i)-(iii) are checked with $s = 4$, we have the CLT under $\mathbb{P}_\nu$*

$$\sqrt{n}(T_n - \mu(U))/\Sigma_n \Rightarrow \mathcal{N}(0,1), \text{ as } n \to \infty.$$

This limit result also holds for $\tilde{T}_n$, as well as the asymptotic validity of the RBB distribution: as $n \to \infty$,

$$\sup_{x\in\mathbb{R}} |\mathbb{P}^*(\sqrt{n^*}(T_n^* - \tilde{T}_n))/\Sigma_n^* \leq x) - \mathbb{P}_\nu(\sqrt{n}(\tilde{T}_n - \mu(U))/\Sigma_n \leq x)| \overset{\mathbb{P}_\nu}{\to} 0.$$

Whereas proving the asymptotic validity of the RBB for U-statistics under these assumptions is straightforward (its second order accuracy up to $o(n^{-1/2})$ seems also quite easy to prove by simply adapting the argument used by [109] under appropriate Cramer condition on $\omega_U^{(1)}(\mathcal{B}_1)$ and block-moment assumptions), establishing an exact rate, $O(n^{-1})$ for instance as in the case of sample mean statistics, is much more difficult. Even if one tries to reproduce the argument in [19] consisting in partitioning the underlying probability space according to every possible realization of the regeneration times sequence between 0 and n, the problem boils down to control the asymptotic behaviour of the distribution $\mathbb{P}(\sum_{1\leqslant i\neq j\leqslant m} \omega_U^{(2)}(\mathcal{B}_i, \mathcal{B}_j)/\sigma^2_{U,\, m} \leqslant y,\ \sum_{j=1}^m L_j = l)$ as $m \to \infty$, which is a highly difficult technical task (due to the lattice component).

We point out that the approach developed here to deal with the statistic U_L naturally applies to more general functionals of the regeneration blocks $\sum_{k<l} \omega(\mathcal{B}_k, \mathcal{B}_l)$, with $\omega : \mathbb{T}^2 \to \mathbb{R}$ being some measurable function. For instance, the estimator of the asymptotic variance $\widehat{\sigma}_n^2(f)$ proposed in §2.2.1 could be derived from such a functional, that may be seen as a U-statistic based on observation blocks with kernel $\omega(\mathcal{B}_k, \mathcal{B}_l) = (f(\mathcal{B}_k) - f(\mathcal{B}_l))^2/2$.

0.12.2 General case

Suppose now that the observed trajectory $X^{(n+1)} = (X_1, ..., X_{n+1})$ is drawn from a general Harris positive chain with stationary probability μ. Using the split chain, we have the representation of the parameter $\mu(U)$:

$$\mu(U) = \mathbb{E}_{A_\mathcal{M}}(\tau_{A_\mathcal{M}})^{-2} \mathbb{E}_{A_\mathcal{M}}(\omega_U(\mathcal{B}_1, \mathcal{B}_2)).$$

Using the pseudo-blocks $\widehat{\mathcal{B}}_l$, $1 \leqslant l \leqslant \widehat{l}_n - 1$, as constructed in §1.3.2, we consider the sequence of renormalizing constants for T_n :

$$\widehat{\Sigma}_n^2 = 4(\widehat{l}_n/n)^3 \widehat{\sigma}^2_{U,\, \widehat{l}_n-1}, \tag{36}$$

with

$$\widehat{\sigma}^2_{U,\, \widehat{l}_n-1} = (\widehat{l}_n-2)(\widehat{l}_n-3)^{-2} \sum_{k=1}^{\widehat{l}_n-1} [(\widehat{l}_n-2)^{-1} \sum_{l=1, k\neq l}^{\widehat{l}_n-1} \omega_U(\widehat{\mathcal{B}}_k, \widehat{\mathcal{B}}_l) - \widehat{U}_{\widehat{l}_n-1}]^2,$$

$$\widehat{U}_{\widehat{l}_n-1} = \frac{2}{(\widehat{l}_n-1)(\widehat{l}_n-2)} \sum_{1\leqslant k<l\leqslant \widehat{l}_n-1} \omega_U(\widehat{\mathcal{B}}_k, \widehat{\mathcal{B}}_l).$$

We also introduce the U-statistic computed from the first approximate regeneration time and the last one:

$$\widehat{T}_n = \frac{2}{\widehat{n}(\widehat{n}-1)} \sum_{1+\widehat{\tau}_A(1)\leqslant i<j\leqslant \widehat{\tau}_A(l_n)} U(X_i, X_j),$$

with $\widehat{n} = \widehat{\tau}_A(\widehat{l}_n) - \widehat{\tau}_A(1)$. Let us define the bootstrap counterparts T_n^* and Σ_n^* of $\widehat{T}_n$ and $\widehat{\Sigma}_n^2$ constructed from the pseudo-blocks via *Algorithm 5*. Although approximate blocks are used here instead of the (unknown) regenerative ones $\mathcal{B}_l$, $1 \leqslant l \leqslant l_n - 1$, asymptotic normality still holds under appropriate assumptions, as shown by the theorem below, which we state in the only case when the kernel U is bounded (with the aim to make its formulation simpler).

Theorem 22 *(Bertail & Clémençon, 2006a) Suppose that the kernel $U(x, y)$ is bounded and that $\mathcal{H}_2$, $\mathcal{H}_3$, $\mathcal{H}_4$ are fulfilled, as well as (i)-(iii) for $s = 4$. Then we have as $n \to \infty$,*

$$\widehat{\Sigma}_n^2 \to \Sigma^2 = 4\mathbb{E}_{A_{\mathcal{M}}}(\tau_{A_{\mathcal{M}}})^{-3}\mathbb{E}_{A_{\mathcal{M}}}(\omega_U^{(1)}(\mathcal{B}_1)^2), \text{ in } \mathbb{P}_\nu\text{-pr.}$$

Moreover as $n \to \infty$, under $\mathbb{P}_\nu$ we have the convergence in distribution

$$n^{1/2}\widehat{\Sigma}_n^{-1}(\widehat{T}_n - \mu(U)) \Rightarrow \mathcal{N}(0, 1),$$

as well as the asymptotic validity of the ARBB counterpart

$$\sup_{x\in\mathbb{R}} |\mathbb{P}^*(\sqrt{n^*}(T_n^* - \widehat{T}_n))/\Sigma_n^* \leq x) - \mathbb{P}_\nu(\sqrt{n}(\widehat{T}_n - \mu(U))/\widehat{\Sigma}_n \leq x)| \underset{n\to\infty}{\overset{\mathbb{P}_\nu}{\to}} 0.$$

0.13 Some simulation studies

We now give two examples, with a view to illustrate the scope of applications of our methodology. The first example presents a regenerative Markov chain described and studied at greater length in [107] (see also [44] and [45]) for modeling storage systems. In consideration of the recent emphasis on nonlinear models in the time series literature, our second example shows to what extent the ARBB method may apply to a general nonlinear AR model. Further, we point out that the principles exposed in this paper are by no means restricted to the markovian setting, but may apply to any process for which a regenerative extension can be constructed and simulated from the data available (see chapter 10 in [193]).

0.13.1 Example 1 : content-dependent storage systems

We consider a general model for storage, evolving through a sequence of *input times* $(T_n)_{n\in\mathbb{N}}$ (with $T_0 = 0$ by convention), at which the storage system is replenished. Let S_n be the amount of input into the storage system at the nth input time T_n and C_t be the amount of contents of the storage system at time t. When possible, there is withdrawal from the storage system between these input times at the constant rate r and the amount of stored contents that drops in a time period $[T, T+\Delta T]$ since the latter input time is equal to $C_T - C_{T+\Delta T} = r\Delta T$, and when the amount of contents reaches zero, it continues to take the value zero until it is replenished at the next input time. If X_n denotes the amount of contents immediately before the input time T_n (*i.e.* $X_n = C_{T_n} - S_n$), we have for all $n \in \mathbb{N}$,

$$X_{n+1} = (X_n + S_n - r\Delta T_{n+1})_+,$$

with $(x)_+ = \sup(x, 0)$, $X_0 = 0$ by convention and $\Delta T_n = T_n - T_{n-1}$ for all $n \geq 1$. Let $K(x, ds)$ be a transition probability kernel on $\mathbb{R}_+$. Assume that, conditionally to $X_1, ..., X_n$, the amounts of input $S_1, ..., S_n$ are independent from each other and independent from the inter-arrival times $\Delta T_1, ..., \Delta T_n$ and that the distribution of S_i is given by $K(X_i, .)$, for $0 \leqslant i \leqslant n$. Under the further assumption that $(\Delta T_n)_{n \geqslant 1}$ is an i.i.d. sequence with common distribution G, independent from $X = (X_n)_{n \in \mathbb{N}}$, the storage process X is a Markov chain with transition probability kernel Π:

$$\Pi(X_n, \ \{0\}) = \Gamma(X_n, \ [X_n, \ \infty[),$$
$$\Pi(X_n, \ ]x, \ \infty[) = \Gamma(X_n, \ ]-\infty, \ X_n - x[)$$

for all $x > 0$, where the transition probability Γ is given by the convolution product $\Gamma(x,]-\infty, \ y[) = \int_{t=0}^{\infty} \int_{z=0}^{\infty} G(dt) K(x, \ dz) \mathbb{I}\{rt - z < y\}$.

One may check that the chain Π is δ_0-irreducible as soon as $K(x, .)$ has infinite tail for all $x \geqslant 0$. In this case, $\{0\}$ is an accessible atom for X and it can be shown that it is positive recurrent if and only if there exists $b > 0$ and a test function $V : \mathbb{R}_+ \to [0, \infty]$ such that $V(0) < \infty$ and for all $x \geqslant 0$:

$$\int \Pi(x, dy) V(y) - V(x) \leqslant -1 + b\mathbb{I}\{x = 0\}.$$

The times at which the storage process X reaches the value 0 are thus regeneration times, and allow to define regeneration blocks dividing the sample path, as shown in Figure 1. Figure 4 below shows a reconstructed RBB data series, generated by a sequential sampling of the regeneration blocks (as described in section 0.3), on which RBB statistics may be based.

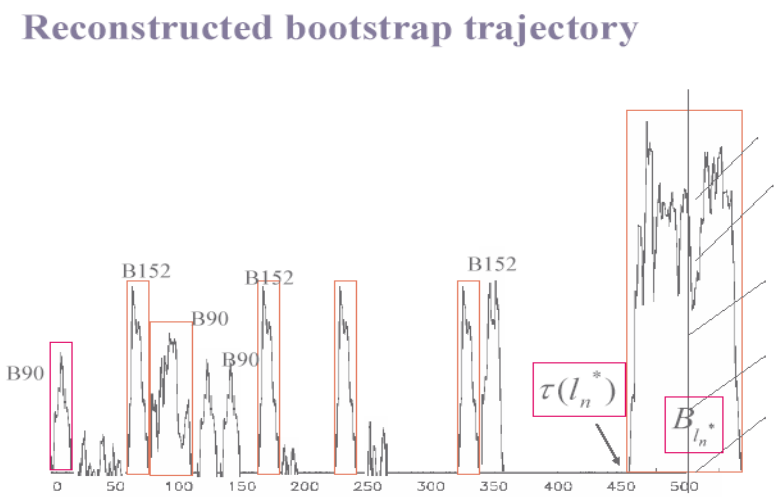

Figure 4: Reconstructed RBB data series

Simulation results We simulated two trajectories of respective length $n = 100$ and $n = 200$ drawn from this Markov chain with $r = 1$, $K(x, dy) = Exp_3(dy)$ and $G(dy) = Exp_1(dy)$, denoting by $Exp_\lambda(dy)$ the exponential distribution with mean $1/\lambda > 0$, which is a standard M/M/1 model (see [8]) for instance). In Fig. 5 below, a Monte-Carlo estimate of the true distribution of the

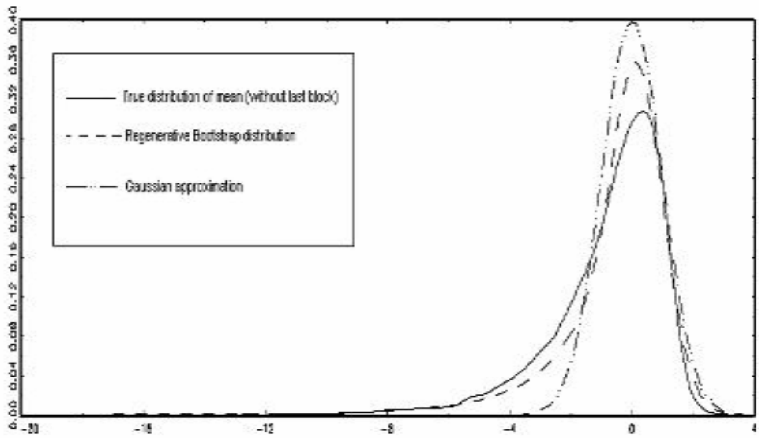

Figure 5: True and estimated mean's distributions

sample mean standardized by its estimated standard error (as defined in (9)) computed with 10000 simulated trajectories is compared to the RBB distribution (in both cases, Monte-Carlo approximations of RBB estimates are computed from $B = 2000$ repetitions of the RBB procedure) and to the gaussian approximation. Note also that in the ideal case where one *a priori* knows the exact form of the markovian data generating process, one may naturally construct a bootstrap distribution in a parametric fashion by estimating first the parameters of the M/M/1 model, and then simulating bootstrap trajectories based on these estimates. Such an ideal procedure naturally performs very well in practice. Of course, in most applications practicioners have generally no knowledge of the exact form of the underlying Markov model, since this is oftenly one of the major goals of statistical inference.

With the aim of constructing accurate confidence intervals, Table 1 compares the quantile of order γ of the true distribution, the one of the gaussian approximation (both estimated with 10000 simulated trajectories) and the mean of the quantile of order γ of the RBB distribution over 100 repetitions of the RBB procedure in the tail regions.

The left tail is clearly very well estimated, whereas the right tail gives a better approximation than the asymptotic distribution. The gain in term of coverage accuracy is quite enormous in comparison to the asymptotic distribution. For instance at the level 95%, for $n = 200$, the asymptotic distribution yields a bilateral coverage interval of level 71% only, whereas the RBB distribution yields a level of 92% in our simulation.

n=	100		200		∞	n=	100		200		∞
γ%	TD	RBB	TD	RBB	ASY	γ%	TD	RBB	TD	RBB	ASY
1	-7.733	-7.044	-5.492	-5.588	-2.326	90	1.041	1.032	1.029	1.047	1.282
2	-6.179	-5.734	-4.607	-4.695	-2.054	91	1.078	1.085	1.083	1.095	1.341
3	-5.302	-5.014	-4.170	-4.165	-1.881	92	1.125	1.145	1.122	1.150	1.405
4	-4.816	-4.473	-3.708	-3.757	-1.751	93	1.168	1.207	1.177	1.209	1.476
5	-4.374	-4.134	-3.430	-3.477	-1.645	94	1.220	1.276	1.236	1.277	1.555
6	-4.086	-3.853	-3.153	-3.243	-1.555	95	1.287	1.360	1.299	1.356	1.645
7	-3.795	-3.607	-2.966	-3.045	-1.476	96	1.366	1.453	1.380	1.442	1.751
8	-3.576	-3.374	-2.771	-2.866	-1.405	97	1.433	1.568	1.479	1.549	1.881
9	-3.370	-3.157	-2.606	-2.709	-1.341	98	1.540	1.722	1.646	1.685	2.054
10	-3.184	-2.950	-2.472	-2.560	-1.282	99	1.762	1.970	1.839	1.916	2.326

Table 1 : Comparison of the tails of the true distribution (TD), RBB and gaussian distributions.

0.13.2 Example 2 : general autoregressive models

Consider now the general heteroskedastic autoregressive model

$$X_{n+1} = m(X_n) + \sigma(X_n)\varepsilon_{n+1}, \ n \in \mathbb{N},$$

where $m : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}_+^*$ are measurable functions, $(\varepsilon_n)_{n\in\mathbb{N}}$ is a i.i.d. sequence of r.v.'s drawn from $g(x)dx$ such that, for all $n \in \mathbb{N}$, ε_{n+1} is independent from the X_k's, $k \leqslant n$ with $E(\varepsilon_{n+1}) = 0$ and $var(\varepsilon_{n+1}) = 1$. The transition kernel density of the chain is given by $p(x, y) = g((y - m(x))/\sigma(x))$, $(x, y) \in \mathbb{R}^2$. Assume further that g, m and σ are continuous functions and there exists $x_0 \in \mathbb{R}$ such that $p(x_0, x_0) > 0$. Then, the transition density is uniformly bounded from below over some neighborhood $V_{x_0}(\varepsilon)^2 = [x_0 - \varepsilon, x_0 + \varepsilon]^2$ of (x_0, x_0) in $\mathbb{R}^2$: there exists $\delta = \delta(\varepsilon) \in]0, 1[$ such that,

$$\inf_{(x,y)\in V_{x_0}^2} p(x, y) \geqslant \delta(2\varepsilon)^{-1}. \tag{37}$$

Any compact interval $V_{x_0}(\varepsilon)$ is thus a small set for the chain X, which satisfies the minorization condition $\mathcal{M}(1, V_{x_0}(\varepsilon), \delta, \mathcal{U}_{V_{x_0}(\varepsilon)})$, where $\mathcal{U}_{V_{x_0}(\varepsilon)}$ denotes the uniform distribution on $V_{x_0}(\varepsilon)$ (see example 3 in §2.2.3). Hence, in the case when one knows x_0, ε and δ such that (37) holds (this simply amounts to know a uniform lower bound estimate for the probability to return to $V_{x_0}(\varepsilon)$ in one step), one may effectively apply the ARBB methodology to X. In the following, we use the pratical criterion $\widehat{N}_n(x_0, \varepsilon)$ with $x_0 = 0$. The choice $x_0 = 0$ is simply motivated by observing that our temporal simulated data fluctuate around 0. Actually, to our own practical experience, optimizing over x_0 does not really improve the performance of the procedure in this case.

In what follows, we compare the performance of the ARBB to the one of some reference competitors for bootstrapping time series.

The *sieve bootstrap* is specifically tailored for linear time series (see [46], [47]). The fact that it fully exploits the underlying linear structure explains why it performs very well in this framework. When simulating linear time series, we use it as a benchmark for evaluating the pertinence of the ARBB distribution. Recall also that this method requires a preliminary estimation of the order q of the sieve : for this purpose we choose an AIC criterion of the type $AIC(q) = n\log(\widehat{MSE}) + 2q$. As will be seen, in the linear AR(q) model below, this information criterion (almost) always enables us to pick the right order of the model. And the resulting sieve bootstrap behaves like a parametric bootstrap method in these cases (see [38]), leading to very good numerical results, as soon as the roots of the AR(q) model are far from the unit circle. In contradistinction, we actually experienced

problems in our simulations, when dealing with an AR(1) model with a root close to 1: in such cases, it may happen with high probability that one gets an estimate of the root larger than one, yielding to explosive bootstrap trajectories.

We also compared the ARBB method to the usual MBB. The difficulty for applying the latter method essentially relies in the choice of the block size for estimating the variance and in the choice of the block size for the resampling procedure. As there is actually no reason for these two sizes to be equal, they should be picked separately and the estimator of the variance should be correctly debiased (see [96]). To our knowledge, the problem of simultaneously calibrating these two quantities has not been treated yet and leads to extremely volatile results. For comparing directly the MBB distribution to the true standardized distribution, we have chosen here to standardize all the distributions by the estimator (9), so as to avoid a deteriorating preliminary variance estimation step. The MBB distribution is also correctly recentered (by the bootstrap mean). And the block size is chosen according to the method of [103]). It consists in estimating first the MSE of the MBB distribution corresponding to blocks of size l with a subsampling technique for various size values l and then picking the size corresponding to a minimum MSE estimate. This unfortunately requires to select a subsampling size and a plausible pilot size, which are in their turn also very difficult to calibrate (see the discussion in Section 7.3 of [130]): here we have chosen $n^{1/4}$ as pilot size and $b_n = n^{10/21}$ as subsampling size (which is close to $n^{1/2}$ in our simulations and satisfies the conditions needed for the MBB to be asymptotically valid). When standardized this way, the MBB has performed quite well in most simulations, except notably when data exhibit significant nonlinear features and/or nonstationarity. The reason of this misbehavior arises from the fact that, for some drawing of the fixed size blocks, the jumps between the blocks were so important, that the reconstructed series could not be splitted according to our randomized procedure leading to an invalid estimator of the variance. Thus the MBB considered here can be considered as a MBB with a Markovian control ensuring that the MBB reconstructed series has some regereration properties. Such procedure clearly improved the resulting estimated distributions.

Simulation results Here are empirical evidences for three specific autoregressive models.

The AR(1) model :

$$X_{i+1} = \alpha X_i + \varepsilon_{i+1},\ i \in \mathbf{N},$$

with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\alpha = 0.8$, $X_0 = 0$ and for a trajectory of length $n = 200$.

The AR(1) model with ARCH(1) residuals called *AR-ARCH model*:

$$X_{i+1} = \alpha X_i + (1 + \beta X_i^2)^{1/2} \varepsilon_{i+1},\ i \in \mathbf{N},$$

with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\alpha = 0.6$, $\beta = 0.35$, $X_0 = 0$ and for a trajectory of length $n = 200$.

The so called *ExpAR(1) model*

$$X_{i+1} = (\alpha_1 + \alpha_2 e^{-|X_i|^2}) X_{i+1} + \varepsilon_{i+1},\ i \in \mathbf{N},$$

with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\alpha_1 = 0.6$, $\alpha_2 = 0.1$, $X_0 = 0$ and for a trajectory of length $n = 200$. Such a chain is recurrent positive under the sole assumption that $|\alpha_1| < 1$, (see [191]). This highly nonlinear model behaves like a threshold model: when the chain takes large values, this is almost an AR(1) model with coefficient α_1, whereas for small values, it behaves as an AR(1) model with a larger autoregressive coefficient $\alpha_1 + \alpha_2$.

Here the true distribution of the sample mean is estimated with 10000 simulations. And for a given trajectory, the ARBB distribution is approximated with $B = 1000$ resamplings of the pseudo-blocks. In a previous simulation work, we experienced that the ARBB distribution obtained may

strongly fluctuate, depending on the randomisation steps. For a given trajectory, this problem may be avoided by repeating the ARBB procedure several times (50 times in our simulations) and averaging the resulting ARBB distribution estimates. According to our experiments, only a small number of repetitions (leading to different ways of dividing the same trajectory) suffices for smoothing the ARBB distribution.

For the ARBB, the sieve and the MBB methods, the whole procedure has been repeated 1000 times. Averaging over the results thus obtained, mean quantiles at several orders are displayed in Table 2 for each bootstrap methodology.

The small set is selected by maximizing over $\varepsilon > 0$ the empirical criterion $\widehat{N}_n(0, \varepsilon)$ described above. The main steps of the procedure are summarized in the graph panels shown below.

The first figure in Graph panel 1 shows the Nadaraya-Watson (NW) estimator (6), the second one represents $\widehat{N}_n(0, \varepsilon)$ as ε grows (as well as the smoother empirical criterion (5), see the dotted line). It clearly allows to identify an optimal value for the size of the small set. In the case of the AR model for instance, this selection rule leads to pick in mean $\widehat{\varepsilon} = 0.83$ and $\widehat{\delta} = 0.123$. Our empirical criterion tends to overestimate very sligthly the size of the "optimal" small set (a phenomenon that we have noticed on several occasions in our simulations). The level sets of the NW estimator, the data points (X_i, X_{i+1}) and the estimated small set are represented in the next graphic. This also shows that the small set chosen may be not that "small" if the transition density is flat around $(x_0, x_0) = (0, 0)$ (in some cases it may be thus preferable to choose $x_0 \neq 0$ so as to be in this situation). In the second line of the panel, the figure on the left hand side represents a sample path of the chain and indicates the pseudo-regenerative blocks obtained by applying the randomization rule with $Ber(1 - \widehat{\delta}(2\varepsilon)^{-1}/\ p_n(X_i, X_{i+1}))$ at times i when $(X_i, X_{i+1}) \in V_0(\varepsilon)^2$. The next figure shows how binded blocks form a typical ARBB trajectory. It is noteworthy that such a trajectory presents less artificial "jumps" than a trajectory reconstructed from a classical MBB procedure: by construction, blocks are joined end to end at values belonging to the small set. For comparison purpose, the figure on the right hand side displays a typical realization of a MBB trajectory. Finally, on the last line of the panel, the true distribution (green), the ARBB distribution (black), the sieve boostrap distribution (gray), the MBB distribution (red dotted line) and the asymptotic gaussian distribution (blue dotted line) are compared. And the last figure shows the QQ-plots $\alpha \in [0, 1] \longmapsto G_n(H^{-1}(\alpha))$, where H is the true distribution and G_n denotes one of the approximations: this enables us to discriminate between the various approximations in a sharper fashion, especially in the tail regions. Furthermore, Table 2 below gives the median of the quantiles at several orders γ of the bootstrap distributions over the 1000 replications for each of the three AR models, compared to the true and asymptotic corresponding quantiles.

n=200	AR(1), $\alpha = 0.80$, Gauss. err.				AR-ARCH(1), $\alpha = 0.60$ $\beta = 0.35$				EXP-AR(1), $\alpha_1 = 0.8$ $\alpha_2 = 0.5$				
$\gamma\%$	TD	ARBB	Sieve	MBB	TD	ARBB	Sieve	MBB	TD	ARBB	Sieve	MBB	ASY
1	-3.511	-3.608	-3.413	-3.423	-4.480	-5.228	-5.593	-9.610	-4.480	-5.227	-5.593	-9.610	-2.326
2..5	-2.837	-2.784	-2.814	-2.715	-3.350	-3.873	-4.758	-6.437	-3.349	-3.873	-4.758	-6.437	-1.960
5	-2.225	-2.133	-2.112	-2.104	-2.576	-2.789	-3.741	-4.995	-2.576	-2.789	-3.741	-4.995	-1.645
10	-1.621	-1.565	-1.648	-1.554	-1.825	-1.975	-2.931	-3.505	-1.825	-1.975	-2.931	-3.505	-1.282

n=200	AR(1), $\alpha = 0.90$, Gaussian error				AR-ARCH(1),$\alpha = 0.60$ $\beta = 0.35$				EXP-AR(1),$\alpha_1 = 0.8$ $\alpha_2 = 0.5$				
$\gamma\%$	TD	ARBB	Sieve	MBB	TD	ARBB	Sieve	MBB	TD	ARBB	Sieve	MBB	ASY
90	1.621	1.496	1.608	1.611	1.803	1.890	2.737	2.256	1.803	1.890	2.737	2.256	1.282
95	2.214	2.078	2.193	2.144	2.576	2.678	3.888	3.067	2.576	2.678	3.888	3.067	1.645
97	2,792	2.706	2.727	2.693	3.535	3.670	4.993	4.222	3.235	3.470	4.793	4.022	1.960
99	3.461	3.731	3.855	3.477	4.371	5.359	5.923	6.251	4.371	5.359	5.923	6.251	2.326

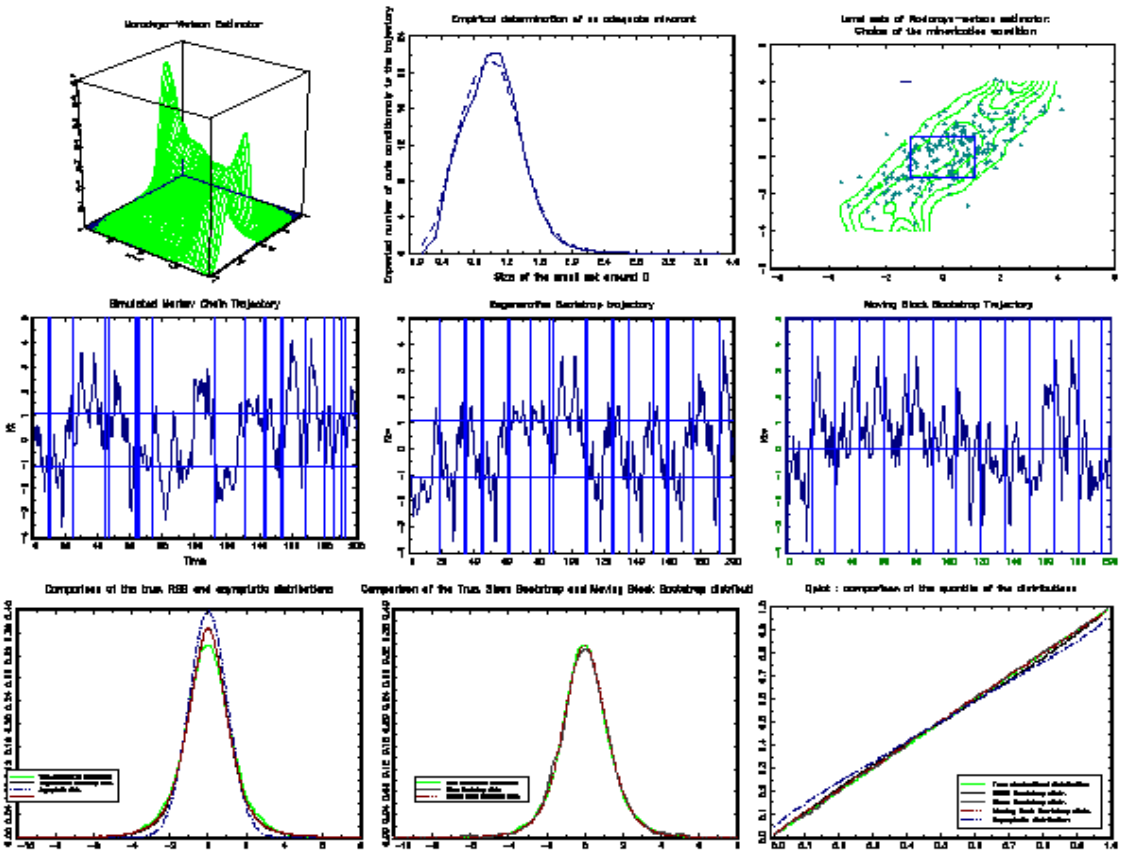

Figure 6: Graph panel 1: AR(1) time-series with $\alpha = 0.8$

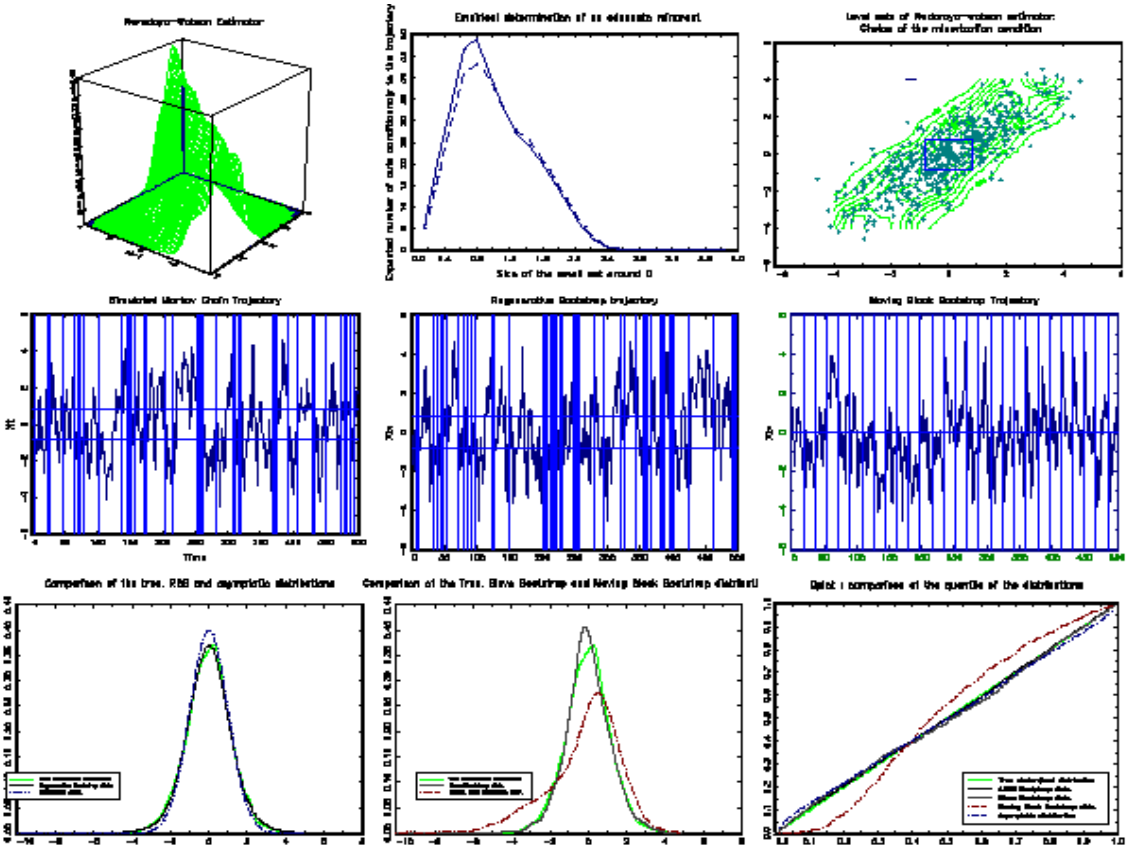

Figure 7: Graph panel 2: AR-ARCH(1) model with $\alpha = 0.6$ and $= 0.35$, $n = 200$

Table 2: Comparing the tails of the true, ARBB and gaussian distributions for the three models

These results clearly indicate that both the sieve and MBB methods perform very well for linear time series. In this case, the ARBB distribution tends to have larger tails. However, when considering nonlinear models, the advantage of the ARBB method over its rivals plainly come into sight: for moderate sample sizes n, the sieve bootstrap tends to choose a too large value $\widehat{q}_n$ for the lag order of the approximate sieve $AR(\widehat{q}_n)$. This problem is less serious for larger sample sizes. In these situations, the MBB behaves very poorly : we conjecture that it could be possibly improved by investigating further how to tune optimally the block size, especially for standardized distributions.

Pictures in Graph panels 2 and 3 speak volumes: for both nonlinear models, the true distribution is accurately approximated by the ARBB distribution. Note nevertheless the difference in the size of the "optimal small set" and in the number of pseudo-regenerations between these models. We point out that, though remarkable when compared to the gaussian approximation, the gain in accuracy obtained by applying the ARBB methodology to the EXP-AR model is higher than the one obtained for the AR-ARCH type model. As may be confirmed by other simulations, the ARBB method provides less accurate results for a given (moderate) sample size, as one gets closer to a unit root model (*i.e.* as α tends to 1): one may get an insight into this phenomenon by simply noticing that the rate of the number of regenerations (respectively, of the number of visits to the small set) then drastically decreases.

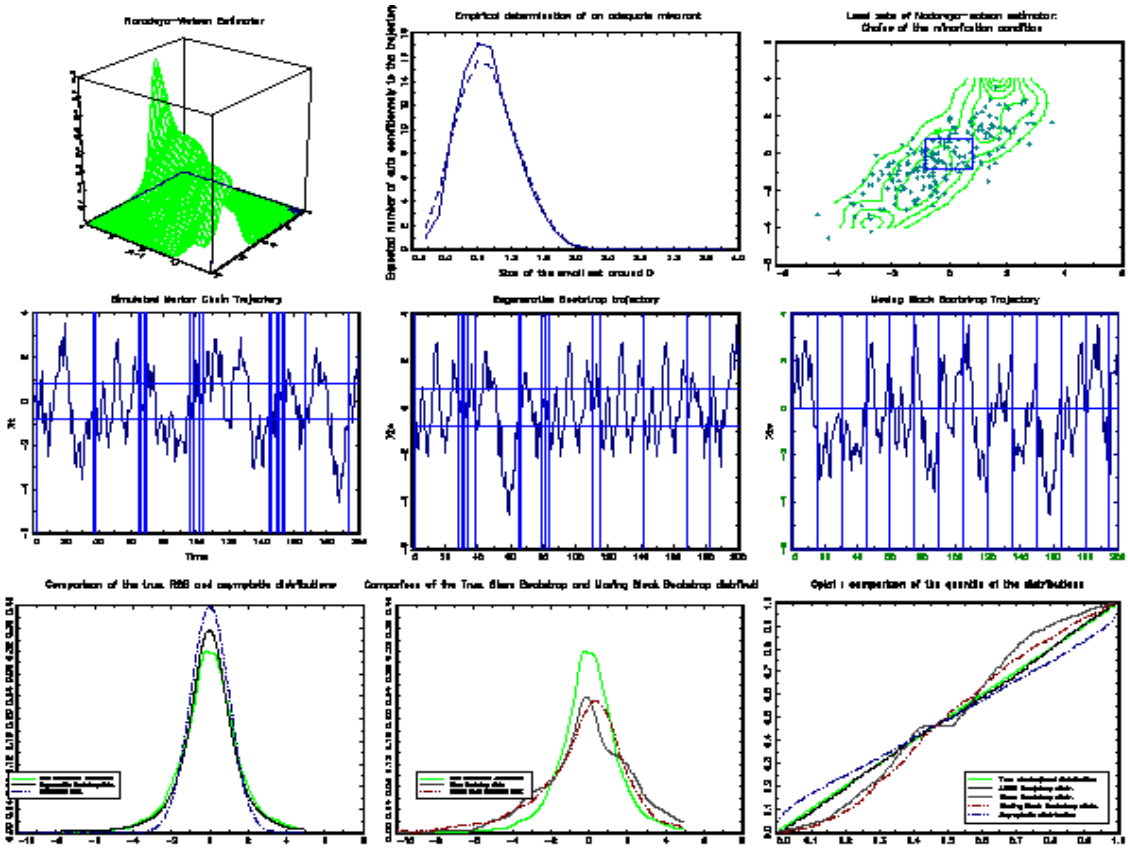

Figure 8: Graph panel 3: EXP-AR(1) model with $\alpha_1 = 0.8$ and $\alpha_2 = 0.5$, $n = 200$

0.13.3 Further remarks

We finally summarize our empirical findings. We first point out that, in the linear case when roots are much less than 1 in amplitude, the sieve bootstrap clearly surpasses its competitors. But it is noteworthy that both the ARBB and the MBB also provides very good numerical results in this case. Besides, all these methods seem to break down from a practical viewpoint for an AR(1) model with an autoregressive coefficient α tending to 1 and with a fixed (moderate) sample size: in such a case, too few pseudo-regeneration blocks may be constructed for the ARBB methodology to be practically performant (although it is asymptotically valid). In this respect, the graph of the estimated number of pseudo-regenerations (see Graph panels 1-3) provides a crucial help for diagnosing the success or the failure of the ARBB method. It is also remarkable that the sieve bootstrap can lead to very bad results in this case, due to the fact that the estimated AR model may have a root larger than 1 (generating then explosive sieve bootstrap trajectories). This strongly advocates the use of preliminary tests or contrained estimation procedures (ensuring that the resulting reconstructed series is asymptotically stationary).

And as may be reported from our simulation results, the advantage of the ARBB over the sieve bootstrap, the MBB and the asymptotic distributions, clearly appears when dealing with nonlinear models. Even if the lag is chosen very large (in mean 85 for the AR-ARCH model and 21 for the exp-AR model), the linear sieve method is unable to capture the non-linearities and performs very badly (the distribution tends to be too much concentrated in these cases). The MBB also performs poorly in such a nonlinear setting for moderate sample sizes but nevertheless tends to surpass the sieve bootstrap for larger sample sizes, whereas the ARBB provides very accurate approximations of the tail distributions in these examples.

Concluding remarks

Although we are far from having covered the unifying theme of statistics based on (pseudo) regeneration for Harris Markov chains, an exhaustive treatment of the possible applications of this methodology being naturally beyond the scope of the present survey, we endeavoured to present here enough material to illustrate the power of this method. Most of the results reviewed in this part of the report are very recent and this line of research is still in development. Now we conclude by making a few remarks raising several open questions among the topics we focused on, and emphasizing the potential gain that the regeneration-based statistical method could provide in further applications.

• We point out that establishing sharper rates for the 2nd order accuracy of the ARBB when applied to sample mean statistics in the general Harris case presents considerable technical difficulties (at least to us). However, one might expect that this problem could be successfully addressed by refining some of the (rather loose) bounds put forward in the proof. Furthermore, as previously indicated, extending the argument to U-statistics requires to prove preliminary non-uniform limit theorems for U-statistics of random vectors with a lattice component.

• In numerous applications it is relevant to consider null recurrent (eventually regenerative) chains: such chains frequently arise in queuing/network systems, related to teletraffic data for instance (see [165]) or [93] for example), with heavy-tailed cycle lengths. Hence, exploring the theoretical properties of the (A)RBB for these specific time series provides thus another subject of further research: as shown by [122], consistent estimates of the transition kernel, as well as rates of convergence for the latter, may still be exhibited for β-recurrent null chains (*i.e.* chains for which the return time to an atom is in the domain of attraction of a stable law with $\beta \in]0,1[$ being the stable index), so that extending the asymptotic validity of the (A)RBB distribution in this case seems conceivable.

• In [89], the pseudo-regeneration approach has been successfully extended to certain continuous Markov processes (namely, diffusion processes) for bootstrap purpose.

• Turning to the statistical study of extremes now (which matters in insurance and finance applications for instance), a thorough investigation of the asymptotic behaviour of extreme value statistics based on the approximate regeneration blocks remains to be carried out in the general Harris case.

• We finally mention ongoing work on empirical likelihood estimation in the markovian setting (see [105]), for which methods based on (pseudo-) regeneration blocks are expected to provide significant results.

Part II

Supervised Learning Methods for Ranking Problems

Abstract

Motivated by various applications, the problem of ranking/ordering instances, instead of classifying them solely, has recently gained much attention in machine learning. For example, a challenging task in document retrieval applications, consists in comparing documents by degree of relevance for a particular request, rather than simply classifying them as relevant or not. Similarly, credit establishments collect and manage large databases containing the socio-demographic and credit-history characteristics of their clients to build a ranking rule which aims at indicating reliability. In this part of the report, the ranking problem is formulated in a rigorous statistical framework.

In Chapter 5, it is reduced to the problem of learning a *ranking rule* for deciding, among two instances, which one is "better," with minimum *ranking risk* (*i.e.* the probability of misclassifying a pair of instances). Since the natural estimates of the risk are of the form of a U-statistic, results of the theory of U-processes are required for investigating the consistency of empirical risk minimizers. Results established in [59] (see also [57] and [58]) are surveyed, laying emphasis on a tail inequality for degenerate U-processes, and on its application for establishing that fast rates of convergence may be achieved under specific noise assumptions, just like in classification. Results related to convex risk minimization methods are also displayed.

Chapter 6 mainly focuses on certain aspects of the so-called *bipartite ranking problem*: the goal is to learn how to order best all the instances x of a set $\mathcal{X}$ by degree of relevance from i.i.d. observations of a pair (X, Y), when Y is some *binary* r.v. indicating relevancy and X is a $\mathcal{X}$-valued r.v. modelling some observation for predicting Y. This practically amounts to find a *scoring function* $s : \mathcal{X} \to \mathbb{R}$ for ranking all input values x according to the order of magnitude of $\mathbb{P}(Y = 1 \mid X = x)$. The problem of ranking a given proportion of instances x among the "most relevant instances" only is considered. Such a *local ranking problem* is of practical importance, since in most ranking applications, in particular in the field of information retrieval, only top ranked instances are effectively scanned. This chapter recapitulates the results in [62], in which paper criteria specifically tailored for selecting scoring functions accomplishing this task in an optimal fashion and extending the *ranking risk* (pairwise classification error) studied in Chapter 5 in this context have been proposed and bounds for learning rates of nonparametric scoring methods based on minimization of such empirical criteria over specific sets of scoring functions have also been established.

Ranking Methods and U-processes

0.14 Introduction and preliminaries

Motivated by various applications including problems related to document retrieval or credit-risk screening, the ranking problem has received increasing attention both in the statistical and machine learning literature. For example, in information retrieval applications, one may be concerned with comparing documents by degree of relevance for a particular request, rather than simply classifying them as relevant or not. In a similar fashion, credit establishments collect and manage large databases containing the socio-demographic and credit-history characteristics of their clients to build a ranking rule which aims at indicating reliability. Such special cases of ranking/ordering problems may be formulated in the following framework.

0.14.1 The bipartite ranking problem

In the so-called *bipartite ranking problem*, the matter is to order all the elements x of a set $\mathcal{X}$ by degree of relevance, when relevancy may be observed through some *binary* indicator variable Y: one has a system consisting of a binary random *output* (response) variable Y, taking its values in $\{-1, 1\}$ say, and a random *input* (predictor) variable X, taking its values in the space $\mathcal{X}$. In documents retrieval applications for instance, one is concerned by ordering all the documents x of a list $\mathcal{X}$ by degree of relevance for a particular request, rather than simply classifying them as relevant or not. Hence, this amounts to assign to each document x in $\mathcal{X}$ a *score* $s(x)$ indicating its degree of relevance for this specific query. In this context, the challenge is to build a *scoring function* $s : \mathcal{X} \to \mathbb{R}$ from sampling data, so as to rank the observations x by increasing order of their score $s(x)$ as accurately as possible: the higher the score $s(X)$ is, the more likely one should observe $Y = 1$.

The ROC curve The accuracy of the ranking induced by s is classically measured by the so-called *ROC curve* (ROC standing for *Receiving Operator Characteristic*, refer to [98] and see also [115]), that consists in plotting the *true positive rate* against the *false positive rate* (see Fig. 9), namely the curve

$$z \in \mathbb{R} \mapsto (1 - F_S^{(-)}(z), 1 - F_S^{(+)}(z)),$$

denoting by $F_S^{(-)}(z) = \mathbb{P}(s(X) \leq z \mid Y = -1)$ (respectively, by $F_s^{(+)}(z) = \mathbb{P}(s(X) \leq z \mid Y = 1)$) the cdf of $s(X)$ conditioned on $Y = -1$ (resp., conditioned on $Y = 1$), or equivalently the graph of the power function $\beta_s : \alpha \in (0, 1) \mapsto 1 - F_s^{(+)}(q_\alpha)$, where $q_\alpha = F_s^{(-)-1}(1 - \alpha) = \inf\{z \in \mathbb{R} / \ F_s^{(-)}(z) \geq 1 - \alpha\}$, $\beta_s(\alpha)$ being the power of the test of level α for testing the null hypothesis "$Y = -1$" based on the test statistic $s(X)$.

This measure of accuracy induces a partial order on the set $\mathcal{S} = \{s : \mathcal{X} \to \mathbb{R} \text{ measurable}\}$ of all scoring functions: for any s_1, s_2 in $\mathcal{S}$, we shall say that s_1 is more accurate than s_2 iff its ROC curve is above the one of s_2 everywhere, that is to say iff $\beta_{s_2}(\alpha) \leq \beta_{s_1}(\alpha)$ for all α in $(0, 1)$ (or equivalently, if the test defined by s_1 for testing the hypothesis that $Y = -1$ is uniformly more

powerful than the one defined by the test function s_2). With respect to this criterion, the optimal ranking is naturally the one induced by the regression function $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$, since one may straightforwardly check that the test it defines is of Neyman-Pearson's type. More precisely, the class of optimal scoring functions corresponds to measurable and strictly increasing transforms of the regression function $\eta(x)$, as stated in the theorem below.

Theorem 23 *A scoring function s in $\mathcal{S}$ is optimal w.r.t. the ROC criterion iff there exists $\Psi : \mathbb{R} \to \mathbb{R}$ measurable and strictly increasing on the support of $\eta(X)$ s.t.: $s = \Psi \circ \eta$.*

Another criterion for evaluating the accuracy of a scoring function that may be found in the literature devoted to information retrieval is the *precision-recall curve* $x \in \mathbb{R} \mapsto (\mathbb{P}(Y = 1 \mid s(X) > x), \mathbb{P}(s(X) > x \mid Y = 1))$.

As emphasized in [63] (see also [13]), the ranking problem is by nature very different from the classification problem. Whereas the goal in the ranking problem consists in finding $s \in \mathcal{S}$ such that for all α in $(0, 1)$, the power $\beta_s(\alpha)$ be as large as possible, the error rate of the classifier $C_{s,q}(X) = 2 \cdot \{s(X) > q\} - 1$ obtained by thresholding a scoring function s at level q is given by:

$$\begin{aligned} M(C_{s,q}) &= (1-p)(1 - F_s^{(-)}(q)) + pF_s^{(+)}(q) \\ &= (1-p)\alpha + p(1 - \beta_s(\alpha)), \end{aligned}$$

with $p = \mathbb{P}(Y = 1)$ and $\alpha = \alpha(q) = 1 - F_S^{(-)}(q)$. In the case when β_s is continuously differentiable, the minimum error rate is thus obtained by tuning the threshold at level $q = q_\alpha$ with α such that $\beta_s'(\alpha) = (1-p)/p$. Hence, constructing an optimal classifier of this type consists in finding a scoring function with a ROC curve having a tangent line with slope $(1-p)/p$ in a point with a first coordinate as small as possible.

The Area Under the ROC curve: a standard summary ranking criterion As previously recalled, the optimal ordering of $\mathcal{X}$ is obtained for scoring functions with a ROC curve above the one of any other scoring function, which are scoring functions s such that for all $\alpha \in [0, 1]$:

$$\beta_s(\alpha) = \beta_\eta(\alpha).$$

Hence, optimizing the ROC curve is a difficult problem, which amounts to recover such a transform of the regression function. In applications, this highly complex optimization problem may be reduced to optimizing a specific summary criterion, known as the *AUC criterion* (AUC standing for Area Under a ROC Curve) for selecting scoring functions (see [104]). The latter is based on the simple observation that the optimal ROC curve is also the one under which the area is maximum. The problem boils down then to searching for a scoring function s that maximizes the area under its ROC curve (see Fig. 1), namely

$$AUC(s) = \int_0^1 \beta_s(\alpha) d\alpha.$$

Note that two scoring functions with the same AUC might lead to quite different rankings (different ROC curves with same integral).

This theoretical summary quantity may be easily interpreted in a probabilistic fashion, since by a simple change of variable it can be written as follows:

$$AUC(s) = \mathbb{P}(s(X) > s(X') \mid Y = 1,\ Y' = -1),$$

where (X', Y') denotes an input/output pair distributed as (X, Y) and independent from the latter. The AUC criterion amounts thus to choose a scoring function s such that, given two independent

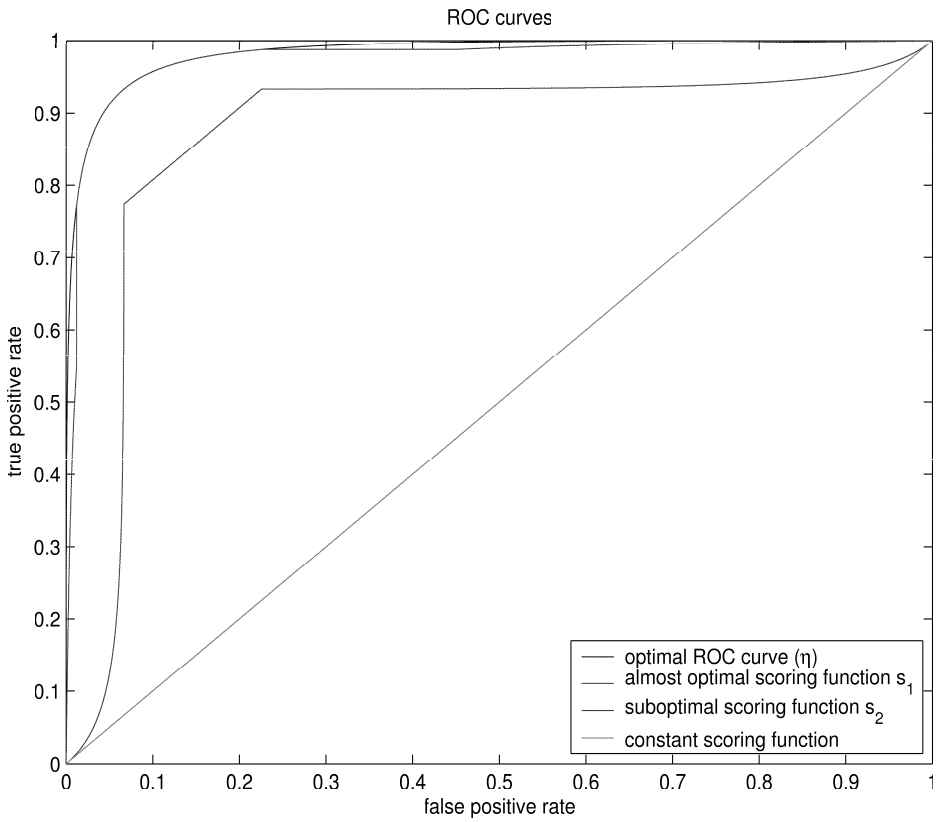

Figure 9: ROC curves.

input observations X and X' such that $Y = 1$ and $Y' = -1$ respectively, the probability that s ranks the instance X' higher than X is minimum.

Let $p = \mathbb{P}(Y = 1)$. From the expression above, one may write

$$\mathrm{AUC}(s) = 1 - \frac{1}{2p(1-p)}\mathbb{P}((s(X) - s(X'))(Y - Y') < 0). \tag{38}$$

Hence, maximizing $\mathrm{AUC}(s)$ amounts to minimize $\mathbb{P}((s(X) - s(X'))(Y - Y') < 0)$, which may be interpreted as a pairwise classification error: the aim being to predict $(y - y')/2$ from (x, x'). Under this form, the ranking problem is now being reduced to a classification problem with the particular error measure given above involving a pair of independently drawn observations.

0.14.2 Outline

In the ranking problem as formulated in this chapter, one has to compare two different observations and decide which one is "better". In certain settings, finding a good ranking rule amounts to constructing a scoring function s. As previously seen, an important special case is the bipartite ranking problem in which the available instances in the data are labelled by binary labels (good and bad): the ranking criterion is then closely related to the AUC.

When viewed this way, the ranking problem is tightly connected to Stute's *conditional* U-*statistics* [187, 188]. Whereas Stute's results imply that certain nonparametric estimates based on local U-statistics give universally consistent ranking rules, the approach in [57] is different: empirical minimizers of U-statistics are considered instead of local averages, looking at things from the point of view of empirical risk minimization, popular in statistical learning theory (see, e.g. [198], [17], [40], [125] or [145] for instance). The crucial point is that natural estimates of the ranking risk involve U-statistics and the methodology is thus based on the theory of U-processes, and calls in particular on maximal and concentration inequalities, symmetrization tricks, and a "contraction principle" for U-processes (see [70] for a comprehensive account of the theory of U-statistics and U-processes).

In [59] (see also [57] and [58]), a theoretical analysis of certain nonparametric ranking methods inspired by boosting-, and support vector machine-type algorithms for classification and which are based on empirical minimization of convex cost functionals over convex sets of scoring functions has been carried out. More precisely, universal consistency of properly regularized versions of these methods has been established and it has been shown that fast rates of convergence may be achieved for empirical risk minimizers under suitable noise conditions, based on a novel tail inequality for degenerate U-processes.

The basic statistical model is introduced in section 0.15, as well as the two main special cases of the ranking problem we consider. Then basic uniform convergence and consistency results for empirical risk minimizers are stated in section 0.16. A new exponential concentration inequality for U-processes is stated in section 0.17, which serves as the main tool in establishing the performance bounds for empirical ranking risk minimization. The noise assumptions guaranteeing fast rates of convergence in particular cases are described in section 0.18, while section 0.19 is devoted to convex risk minimization for ranking problems, providing this way a theoretical framework for studying boosting and support vector machine-type ranking methods (such as the ones developped in [86]).

0.15 The ranking problem as a pairwise classification problem

Let (X, Y) be a pair of r.v.'s taking values in $\mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is a measurable space. The random object X models some observation and Y its real-valued label. Let (X', Y') denote a pair of r.v.'s

identically distributed with (X, Y), and independent from the latter. Set

$$Z = \frac{Y - Y'}{2} .$$

In the ranking problem, X and X' are supposed to observed, but not their labels Y and Y'. We think about X being "better" than X' if $Y > Y'$, that is, if $Z > 0$. (The normalization factor $1/2$ in the definition of Z above is arbitrary and plays no role in the analysis) The goal is to rank X and X' such that the probability that *the better ranked of them has a smaller label* is as small as possible. We thus define a *ranking rule* as a function $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. If $r(x, x') = 1$ then the rule r ranks x higher than x'. Its performance is measured by the *ranking risk*

$$L(r) = \mathbb{P}\{Z \cdot r(X, X') < 0\} ,$$

that is, the probability that r ranks two randomly drawn instances incorrectly. In this set-up, the ranking problem boils down to a binary classification problem in which the sign of the random variable Z has to be guessed based upon the pair of observations (X, X'). The ranking rule with minimal risk can be now easily determined. Set

$$\rho_+(X, X') = \mathbb{P}\{Z > 0 \mid X, X'\}$$
$$\rho_-(X, X') = \mathbb{P}\{Z < 0 \mid X, X'\}.$$

Proposition 24 *(Clémençon, Lugosi & Vayatis, 2005b) Define*

$$r^*(x, x') = 2\mathbb{I}_{[\rho_+(x,x') \geq \rho_-(x,x')]} - 1$$

and denote $L^* = L(r^*) = \mathbb{E}\{\min(\rho_+(X, X'), \rho_-(X, X'))\}$. *Then for any ranking rule* r,

$$L^* \leq L(r) .$$

Let us now consider the problem of constructing ranking rules of low risk based on training data. Supposed that n independent, identically distributed copies of (X, Y), have been observed: $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$. Given a ranking rule r, these training data may be used for estimating its risk $L(r) = \mathbb{P}\{Z \cdot r(X, X') < 0\}$, the most natural estimate being the U-*statistic*

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{[Z_{i,j} \cdot r(X_i, X_j) < 0]}.$$

The statistical challenge amounts then to study the performance of minimizers of the empirical estimate $L_n(r)$ over a class $\mathcal{R}$ of ranking rules.

• In the definition of r^* ties are broken in favor of ρ_+ but obviously if $\rho_+(x, x') = \rho_-(x, x')$, an arbitrary value can be chosen for r^* without altering its risk.

• The actual values of the Y_i's are never used in the ranking rules discussed in this paper. It is sufficient to know the values of the $Z_{i,j}$, or, equivalently, the ordering of the Y_i's.

• Instead of ranking two observations X, X' only, one may be interested in ranking m independent observations $X^{(1)}, \ldots, X^{(m)}$. The value of a ranking function $r(X^{(1)}, \ldots, X^{(m)})$ is then a permutation π of $\{1, \ldots, m\}$ and the goal is that π should coincide with (or at least resemble to) the permutation $\overline{\pi}$ for which $Y^{(\overline{\pi}(1))} \geq \cdots \geq Y^{(\overline{\pi}(m))}$. Given a loss function ℓ that assigns a number in $[0, 1]$ to a pair of permutations, the ranking risk is defined as

$$L(r) = \mathbb{E}\ell(r(X^{(1)}, \ldots, X^{(m)}), \overline{\pi}) .$$

In this general case, natural estimates of $L(r)$ involve m-th order U-statistics. Many of the results stated in the sequel may be straightforwardly extended to this general setup.

Ranking and Scoring. In many interesting cases the ranking problem may be reduced to finding an appropriate *scoring function*. These are the cases when the joint distribution of X and Y is such that there exists a function $s^* : \mathcal{X} \to \mathbb{R}$ such that

$$r^*(x, x') = 1 \quad \text{if and only if} \quad s^*(x) \geq s^*(x') \ .$$

We call such a function s^* an *optimal scoring function* (any strictly increasing transformation of an optimal scoring function being still optimal). Here are some important special cases when the ranking problem may be reduced to scoring.

Example 1 (THE BIPARTITE RANKING PROBLEM.) As recalled above, in the bipartite ranking problem the label Y is binary, taking its values in $\{-1, 1\}$. Writing $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$, the Bayes ranking risk equals

$$\begin{aligned} L^* &= \mathbb{E} \min\{\eta(X)(1 - \eta(X')), \eta(X')(1 - \eta(X))\} \\ &= \mathbb{E} \min\{\eta(X), \eta(X')\} - (\mathbb{E}\eta(X))^2 \end{aligned}$$

and also,

$$L^* = \mathrm{Var}\left(\frac{Y+1}{2}\right) - \frac{1}{2}\mathbb{E}\left|\eta(X) - \eta(X')\right| \ .$$

In particular, $L^* \leq \mathrm{Var}\left(\frac{Y+1}{2}\right) \leq 1/4$, where the equality $L^* = \mathrm{Var}\left(\frac{Y+1}{2}\right)$ holds when X and Y are independent and the maximum is attained when $\eta \equiv 1/2$. The difficulty of the bipartite ranking problem depends on the concentration properties of the distribution of $\eta(X) = \mathbb{P}(Y = 1 \mid X)$ through the *Gini's mean difference* $\mathbb{E}(|\eta(X) - \eta(X')|)$, which is a classical measure of concentration. For given $p = \mathbb{E}(\eta(X))$, Gini's mean difference ranges from a minimum value of zero, when $\eta(X) \equiv p$, to a maximum value of $\frac{1}{2}p(1-p)$, when $\eta(X) = (Y+1)/2$. The optimal ranking rule is given by a scoring function s^* where s^* is any strictly increasing transformation of η. Then one may restrict the search to ranking rules defined by scoring functions s, that is, ranking rules of form $r(x, x') = 2\mathbb{I}_{[s(x) \geq s(x')]} - 1$. Writing $L(s) \stackrel{\text{def}}{=} L(r)$, one has

$$L(s) - L^* = \mathbb{E}\left(\left|\eta(X') - \eta(X)\right| \mathbb{I}_{[(s(X)-s(X'))(\eta(X)-\eta(X'))<0]}\right) \ .$$

As already mentioned in section 0.14, the ranking risk is related to the AUC criterion which is a standard performance measure in this setup (see [86]). Set $p = \mathbb{P}(Y = 1)$, we have

$$\mathrm{AUC}(s) = 1 - \frac{1}{2p(1-p)} L(s), \tag{39}$$

so that maximizing the AUC criterion boils down to minimizing the ranking error.

Example 2 (A REGRESSION MODEL). Let Y be real-valued and the joint distribution of X and Y be s.t. $Y = m(X) + \epsilon$, where $m(x) = \mathbb{E}(Y|X = x)$ is the regression function, ϵ is independent of X and has a symmetric distribution around zero. The optimal ranking rule r^* may be then defined by a scoring function s^*, where s^* may be any strictly increasing transformation of m.

0.16 Empirical ranking risk minimization

Based on the empirical estimate $L_n(r)$ of the risk $L(r)$ defined above, one may select a ranking rule by minimizing the empirical risk over a class $\mathcal{R}$ of ranking rules $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. Consider the empirical risk minimizer, over $\mathcal{R}$, by

$$r_n = \arg\min_{r \in \mathcal{R}} L_n(r) \ .$$

In a *first-order* approach, the performance $L(r_n) = \mathbb{P}\{Z \cdot r_n(X, X') < 0 | D_n\}$ of the empirical risk minimizer may be investigated by using the standard bound (see [72])

$$L(r_n) - \inf_{r \in \mathcal{R}} L(r) \leq 2 \sup_{r \in \mathcal{R}} |L_n(r) - L(r)| \ . \tag{40}$$

This shows that bounding the performance of an empirical minimizer of the ranking risk boils down to investigating the properties of U-*processes*, that is, suprema of U-statistics indexed by a class of ranking rules. In a first-order approach, the next inequality (based on the standard results in [111]) permits to reduce the problem to the study of ordinary empirical processes.

Lemma 25 *(Clémençon, Lugosi & Vayatis, 2005b) Let $q_\tau : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be real-valued functions indexed by $\tau \in T$ where T is some set. If $X_1, \ldots, X_n$ are i.i.d. then for any convex nondecreasing function ψ,*

$$\mathbb{E}\psi\left(\sup_{\tau \in T} \frac{1}{n(n-1)} \sum_{i \neq j} q_\tau(X_i, X_j)\right) \leq \mathbb{E}\psi\left(\sup_{\tau \in T} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q_\tau(X_i, X_{\lfloor n/2 \rfloor + i})\right) \ ,$$

assuming the suprema are measurable and the expected values exist.

For each $\tau \in T$, $\sum_{i=1}^{\lfloor n/2 \rfloor} q_\tau(X_i, X_{\lfloor n/2 \rfloor + i})$ is a sum of $\lfloor n/2 \rfloor$ i.i.d. r.v.'s, the moment generating function of the U-process may be thus bounded by the moment generating function of an ordinary empirical process: $\forall \lambda > 0$,

$$\mathbb{E}\exp\left(\lambda \sup_{\tau \in T} \frac{1}{n(n-1)} \sum_{i \neq j} q_\tau(X_i, X_j)\right) \leq \mathbb{E}\exp\left(\lambda \sup_{\tau \in T} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q_\tau(X_i, X_{\lfloor n/2 \rfloor + i})\right) \ .$$

Hence, basics methods for handling empirical processes can be applied directly. Using the bounded differences inequality (see [147]) for instance, one gets that

$$\begin{aligned} &\mathbb{E}\exp\left(\lambda \sup_{\tau \in T} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q_\tau(X_i, X_{\lfloor n/2 \rfloor + i})\right) \\ &\quad \leq \exp\left(\lambda \mathbb{E} \sup_{\tau \in T} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q_\tau(X_i, X_{\lfloor n/2 \rfloor + i}) + \frac{\lambda^2}{4(n-1)}\right) \ , \end{aligned}$$

if each q_τ takes its values in an interval of length 1. This leads to

$$\begin{aligned} &\log \mathbb{E}\exp\left(\lambda \sup_{r \in \mathcal{R}} |L_n(r) - L(r)|\right) \\ &\quad \leq \lambda \mathbb{E} \sup_{r \in \mathcal{R}} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} |\mathbb{I}_{[Z_{i,\lfloor n/2 \rfloor + i} \cdot r(X_i, X_{\lfloor n/2 \rfloor + i}) < 0]} - L(r)| + \frac{\lambda^2}{4(n-1)} \ . \end{aligned}$$

Standard inequalities may be applied for bounding the expected value on the right-hand side. For example, if the class $\mathcal{R}$ of indicator functions has finite VC dimension V (see [134]), then

$$\mathbb{E}\sup_{r\in\mathcal{R}}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}|\mathbb{I}_{[Z_{i,\lfloor n/2\rfloor+i}\cdot r(X_i,X_{\lfloor n/2\rfloor+i})<0]}-L(r)|\leq c\sqrt{\frac{V}{n}}$$

for a universal constant c. The next result immediately derives from the Chernoff bound.

Proposition 26 *(Clémençon, Lugosi & Vayatis, 2005b) Let $\mathcal{R}$ be a class of ranking rules of* VC *dimension V. Then for any $t>0$,*

$$\mathbb{P}\left\{\sup_{r\in\mathcal{R}}|L_n(r)-L(r)|>c\sqrt{\frac{V}{n}}+t\right\}\leq e^{-(n-1)t^2}\ .$$

It is well known from the theory of empirical risk minimization for classification that the bound (40) is often quite loose, due to the fact that the variance of the estimators of the risk is ignored and bounded uniformly by a constant. Therefore, the main interest in considering U-statistics precisely consists in the fact that they have minimal variance among all unbiased estimators. However, the reduced-variance property of U-statistics plays no role in the previous analysis: all upper bounds obtained above remain true for an empirical risk minimizer that, instead of using estimates based on U-statistics, estimates the risk of a ranking rule by splitting the data set into two halves as follows

$$\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbb{I}_{[Z_{i,\lfloor n/2\rfloor+i}\cdot r(X_i,X_{\lfloor n/2\rfloor+i})<0]}\ .$$

Hence, one looses the advantage of calling on U-statistics in the previous study. Let us now give a more precise insight into how one may benefit from using U-statistics.

Hoeffding's decomposition (sharper bounds). U-statistics have been studied in depth and their behavior is well understood. The following classical probability inequality concerning U-statistics is due to Hoeffding [112]: $\forall t>0$,

$$\mathbb{P}\{|L_n(r)-L(r)|>t\}\leq 2\exp\left(-\frac{\lfloor(n/2)\rfloor t^2}{2\sigma^2+2t/3}\right)\ , \tag{41}$$

with $\sigma^2=\mathrm{Var}(\mathbb{I}_{[Z\cdot r(X,X')<0]})=L(r)(1-L(r))$. Therefore, the latter inequality may be improved by replacing σ^2 by a smaller term. This results from the so-called Hoeffding's decomposition as recalled below. Hoeffding's decomposition (see [180] for more details) is a basic tool for studying U-statistics. Let $X,X_1,\ldots,X_n$ be i.i.d. r.v.'s and denote by

$$U_n(X_1,\ldots,X_n)=\frac{1}{n(n-1)}\sum_{i\neq j}q(X_i,X_j)$$

a U-statistic of order 2 where the kernel q is a symmetric real-valued function.

Assuming that $q(X_1,X_2)$ is square integrable, $U_n-\mathbb{E}U_n$ may be decomposed as a sum T_n of i.i.d. r.v's plus a *degenerate* U-statistic W_n. In order to write this decomposition, consider the following function of one variable

$$h(X_i)=\mathbb{E}(q(X_i,X)\mid X_i)-\mathbb{E}U_n\,,$$

and the function of two variables

$$\tilde{h}(X_i,X_j)=q(X_i,X_j)-\mathbb{E}U_n-h(X_i)-h(X_j).$$

Then we have the orthogonal expansion

$$U_n = \mathbb{E}U_n + 2T_n + W_n \ ,$$

where

$$T_n(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} h(X_i),$$
$$W_n(X_1, ..., X_n) = \frac{1}{n(n-1)}\sum_{i\neq j} \tilde{h}(X_i, X_j) \, .$$

W_n is called a degenerate U-statistic because its kernel $\tilde{h}$ satisfies

$$\mathbb{E}\left(\tilde{h}(X_i, X) \mid X_i\right) = 0 \ .$$

The variance of T_n is

$$\mathrm{Var}(T_n) = \frac{\mathrm{Var}(\mathbb{E}(q(X_1, X) \mid X_1))}{n} \, .$$

Therefore, $\mathrm{Var}(\mathbb{E}(q(X_1, X) \mid X_1))$ is less than $\mathrm{Var}(q(X_1, X))$ (unless q is already degenerate) and the variance of the degenerate U-statistic W_n is $O(\frac{1}{n^2})$: T_n is thus leading term in this orthogonal decomposition. Indeed, the limit distribution of $\sqrt{n}(U_n - \mathbb{E}U_n)$ is the normal distribution $\mathcal{N}(0, 4\mathrm{Var}(\mathbb{E}(q(X_1, X) \mid X_1))$ (see [111]). This suggests that inequality (41) may be quite loose.

Indeed, exploiting further Hoeffding's decomposition (combined with arguments related to decoupling, randomization and hypercontractivity of Raŋdeŋmaŋcher chaos) de la Peña and Giné [?] established a Bernstein's type inequality of the form (41) but with σ^2 replaced by the variance of the conditional expectation (see Theorem 4.1.13 in [70]). When specialized to our setting (*i.e.* with $q(X_i, X_j) = \mathbb{I}_{[Z_{i,j}\cdot r(X_i,X_j)<0]}$), this yields

$$\mathbb{P}\{|L_n(r) - L(r)| > t\} \leq 4\exp\left(-\frac{nt^2}{8s^2 + ct}\right) \ ,$$

where $s^2 = \mathrm{Var}(\mathbb{P}\{Z \cdot r(X, X') < 0|X\})$ is the variance of the conditional expectation and c is some constant. This remarkable improvement is not exploited in the first-order analysis above but shall become crucial when establishing fast rates of convergence. As a matter of fact, it is shown in section 0.17 that under certain, quite general, conditions significantly smaller risk bounds are achievable. There it will have an essential importance to use sharp exponential bounds for U-processes, involving their reduced variance.

0.17 Fast rates

It is well known (refer to §5.2 in [40] and the references therein) that faster rate bounds for the excess risk in the context of binary classification may be achieved when the variance of the excess risk can be controlled by its expected value. This is guaranteed under certain "low-noise" conditions (see [192], [146] or [125]). As shown in [59], significantly sharper bounds may also be established in the ranking problem under some conditions that are somehow analogous to the low-noise conditions in the classification problem, which permit to benefit from the small variance of the U-statistic (as opposed to splitting the sample) to estimate the ranking risk. The analysis is based on the Hoeffding decomposition recalled above. Consider the following estimate of the *excess risk* $\Lambda(r) = L(r) - L^* = \mathbb{E}q_r((X, Y), (X', Y'))$:

$$\Lambda_n(r) = \frac{1}{n(n-1)}\sum_{i\neq j} q_r((X_i, Y_i), (X_j, Y_j)),$$

which is a U-statistic of degree 2 with symmetric kernel:

$$q_r((x,y),(x',y')) = \mathbb{I}_{[(y-y')\cdot r(x,x')<0]} - \mathbb{I}_{[(y-y')\cdot r^*(x,x')<0]}$$

The minimizer r_n of the empirical ranking risk $L_n(r)$ over $\mathcal{R}$ also minimizes the empirical excess risk $\Lambda_n(r)$. The Hoeffding decomposition of $\Lambda_n(r)$ may be written as follows:

$$\Lambda_n(r) - \Lambda(r) = 2T_n(r) + W_n(r)\,,$$

where

$$T_n(r) = \frac{1}{n}\sum_{i=1}^{n} h_r(X_i, Y_i)$$

is a sum of i.i.d. r.v.'s with

$$h_r(x,y) = \mathbb{E}q_r((x,y),(X',Y')) - \Lambda(r)$$

and

$$W_n(r) = \frac{1}{n(n-1)}\sum_{i\neq j} \widehat{h}_r((X_i,Y_i),(X_j,Y_j))$$

is a *degenerate* U-statistic with symmetric kernel

$$\widehat{h}_r((x,y),(x',y')) = q_r((x,y),(x',y')) - \Lambda(r) - h_r(x,y) - h_r(x',y')\,.$$

The argument of the analysis relies on the fact that the contribution of the degenerate part $W_n(r)$ is negligible compared to that of $T_n(r)$. Minimization of Λ_n is thus approximately equivalent to minimizing $T_n(r)$, which is a simple average of i.i.d. r.v.'s. Hence, known techniques used in empirical risk minimization can be invoked for studying the minimization of $T_n(r)$.

The main tool for handling the degenerate part is a new general moment inequality for U-processes proved in [59], stated below. It is based on moment inequalities obtained for empirical processes and Rademacher chaoses in [41]. However, it is noteworthy that it generalizes the inequality established in [5], which would be actually sufficient for dealing with VC classes (see also the results in [3], [90] and [114]).

Theorem 27 *Let $X, X_1, ..., X_n$ be i.i.d. random variables and let $\mathcal{F}$ be a class of kernels. Consider a degenerate U-process Z of order 2 indexed by $\mathcal{F}$,*

$$Z = \sup_{f\in\mathcal{F}} \left| \sum_{i,j} f(X_i, X_j) \right|$$

where $\mathbb{E}f(X,x) = 0$, $\forall x, f$. Assume also $f(x,x) = 0$, $\forall x$ and $\sup_{f\in\mathcal{F}} \|f\|_\infty = F$. Let $\epsilon_1, ..., \epsilon_n$ be i.i.d. Rademacher random variables and introduce the random variables

$$Z_\epsilon = \sup_{f\in\mathcal{F}} \left| \sum_{i,j} \epsilon_i\epsilon_j f(X_i, X_j) \right| \;,$$

$$U_\epsilon = \sup_{f\in\mathcal{F}} \sup_{\alpha:\|\alpha\|_2\leq 1} \sum_{i,j} \epsilon_i\alpha_j f(X_i, X_j) \;,$$

$$M = \sup_{f\in\mathcal{F}, k=1...n} \left| \sum_{i=1}^{n} \epsilon_i f(X_i, X_k) \right| \;.$$

Then there exists a universal constant $C > 0$ *such that for all* n *and* $q \geq 2$,

$$(\mathbb{E}Z^q)^{1/q} \leq C\left(\mathbb{E}Z_\epsilon + q^{1/2}\mathbb{E}U_\epsilon + q(\mathbb{E}M + Fn) + q^{3/2}Fn^{1/2} + q^2F\right).$$

Also, there exists a universal constant C *such that for all* n *and* $t > 0$,

$$\mathbb{P}\{Z > C\mathbb{E}Z_\epsilon + t\} \leq \exp\left(-\frac{1}{C}\min\left(\left(\frac{t}{\mathbb{E}U_\epsilon}\right)^2, \frac{t}{\mathbb{E}M + Fn}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right).$$

In a fashion very similar to the conditions required for obtaining faster rates of convergence in the context of binary classification (see [192], [17], [125] and [145]), our key assumption on the joint distribution of (X, Y) takes the following form:

Assumption 28 *There exist constants* $c > 0$ *and* $\alpha \in [0, 1]$ *such that for all* $r \in \mathcal{R}$,

$$\mathrm{Var}(h_r(X, Y)) \leq c\,\Lambda(r)^\alpha.$$

For $\alpha = 0$ the assumption is always fulfilled and the corresponding performance bound does not yield any improvement over those of Section 0.16. However, in many natural examples Assumption 28 is satisfied with values of α close to one, guaranteeing significant improvements in the rates of convergence.

The next theorem provides a performance bound in terms of expected values of certain Rademacher chaoses indexed by $\mathcal{R}$ and local properties of an ordinary empirical process. These quantities have been thoroughly studied and are well understood. They may be easily bounded in many interesting cases (see the corollary below, where it is applied to the case when $\mathcal{R}$ is a VC class of indicator functions). In order to state the result, we introduce some quantities related to the class $\mathcal{R}$. Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher r.v.'s independent of the (X_i, Y_i). Let

$$Z_\epsilon = \sup_{r\in\mathcal{R}}\left|\sum_{i,j}\epsilon_i\epsilon_j\widehat{h}_r((X_i, Y_i), (X_j, Y_j))\right|,$$

$$U_\epsilon = \sup_{r\in\mathcal{R}}\sup_{\alpha:\|\alpha\|_2\leq 1}\sum_{i,j}\epsilon_i\alpha_j\widehat{h}_r((X_i, Y_i), (X_j, Y_j)),$$

$$M = \sup_{r\in\mathcal{R}, k=1,\ldots,n}\left|\sum_{i=1}^n\epsilon_i\widehat{h}_r((X_i, Y_i), (X_k, Y_k))\right|.$$

Introduce the loss function

$$\ell(r, (x, y)) = 2\mathbb{E}\mathbb{I}_{[(y-Y)\cdot r(x,X)<0]} - L(r)$$

and define

$$\nu_n(r) = \frac{1}{n}\sum_{i=1}^n \ell(r, (X_i, Y_i)) - L(r).$$

And define the pseudo-distance

$$d(r, r') = \left(\mathbb{E}\left(\mathbb{E}[\mathbb{I}_{[r(X,X')\neq r'(X,X')]}|X]\right)^2\right)^{1/2}.$$

Let $\phi : [0, \infty) \to [0, \infty)$ be a nondecreasing function such that $\phi(x)/x$ is nonincreasing and $\phi(1) \geq 1$ such that for all $r \in \mathcal{R}$,

$$\sqrt{n}\mathbb{E}\sup_{r'\in\mathbb{R}, d(r,r')\leq\sigma}|\nu_n(r) - \nu_n(r')| \leq \phi(\sigma).$$

Theorem 29 *(Clémençon, Lugosi & Vayatis, 2006) Consider a minimizer r_n of the empirical ranking risk $L_n(r)$ over a class $\mathcal{R}$ of ranking rules and assume Assumption 28. Then there exists a universal constant C such that, with probability at least $1-\delta$, the ranking risk of r_n satisfies*

$$\begin{aligned} L(r_n) - L^* &\leq 2\left(\inf_{r\in\mathcal{R}} L(r) - L^*\right) \\ &\quad + C\left(\frac{\mathbb{E}Z_\epsilon}{n^2} + \frac{\mathbb{E}U_\epsilon\sqrt{\log(1/\delta)}}{n^2} + \frac{\mathbb{E}M\log(1/\delta)}{n^2} + \frac{\log(1/\delta)}{n} + \rho^2\log(1/\delta)\right) \end{aligned}$$

where $\rho > 0$ is the unique solution of the equation $\sqrt{n}\rho^2 = \phi(\rho^\alpha)$.

Examples in which Assumption 28 is satisfied with $\alpha > 0$ are displayed in the next section. We will see below that the value of α in this assumption determines the magnitude of the last term which, in turn, dominates the right-hand side (apart from the approximation error term). The improvement of Theorem 29 is naturally meaningful only if $\inf_{r\in\mathcal{R}} L(r) - L^*$ does not dominate the other terms in the bound. Ideally, the class $\mathcal{R}$ should be chosen such that the approximation error and the other terms in the bound are balanced. The theorem would then guarantee faster rates of convergence. Hence, this bound enables us to design penalized empirical minimizers of the ranking risk that select the class $\mathcal{R}$ from a collection of classes achieving this objective, as explained in [145] or [125].

The next result illustrates Theorem 29 in the case when $\mathcal{R}$ is a VC class.

Corollary 30 *(Clémençon, Lugosi & Vayatis, 2006) Consider the minimizer r_n of the empirical ranking risk $L_n(r)$ over a class $\mathcal{R}$ of ranking rules of finite VC dimension V and assume Assumption 28. Then there exists a universal constant C such that, with probability at least $1-\delta$, the ranking risk of r_n satisfies*

$$L(r_n) - L^* \leq 2\left(\inf_{r\in\mathcal{R}} L(r) - L^*\right) + C\left(\frac{V\log(n/\delta)}{n}\right)^{1/(2-\alpha)}.$$

0.18 Examples

The bipartite ranking problem. Recall that here it suffices to consider ranking rules of the form $r(x,x') = 2\mathbb{I}_{[s(x)\geq s(x')]} - 1$ where s is a scoring function. We shall abusively write h_s for h_r.

Noise assumption: there exist constants $c > 0$ and $\alpha \in [0,1]$ such that for all $x \in \mathcal{X}$,

$$\mathbb{E}_{X'}\left(\left|\eta(x) - \eta(X')\right|^{-\alpha}\right) \leq c\,. \tag{42}$$

As shown by the next result, this assumption ensures that Assumption 28 is satisfied.

Proposition 31 *(Clémençon, Lugosi & Vayatis, 2005b) Under (42), we have: $\forall s \in \mathcal{F}$*

$$\mathrm{Var}(h_s(X,Y)) \leq c\,\Lambda(s)^\alpha\,.$$

If $\alpha = 0$ then condition (42) poses no restriction, but also no improvement is achieved. At the other extreme, when $\alpha = 1$, the condition is quite restrictive as it excludes η to be differentiable, for example, if X has a uniform distribution over $[0,1]$. However, interestingly, for any $\alpha < 1$, it poses quite mild restrictions as highlighted in the following example:

Corollary 32 *(Clémençon, Lugosi & Vayatis, 2005b) Consider the bipartite ranking problem and assume that* $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$ *is such that the r.v.* $\eta(X)$ *has an absolutely continuous distribution on* $[0, 1]$ *with a density bounded by* B*. Then for any* $\epsilon > 0$*,*

$$\forall x \in \mathcal{X}, \quad \mathbb{E}_{X'}(|\eta(x) - \eta(X')|^{-1+\epsilon}) \leq \frac{2B}{\epsilon}$$

and therefore, by Propositions 0.17 and 31, there is a constant C *such that for every* $\delta, \epsilon \in (0, 1)$*, the excess ranking risk of the empirical minimizer* r_n *satisfies, with probability at least* $1 - \delta$*,*

$$L(r_n) - L^* \leq 2\left(\inf_{r \in \mathcal{R}} L(r) - L^*\right) + CB\epsilon^{-1}\left(\frac{V \log(n/\delta)}{n}\right)^{1/(1+\epsilon)} .$$

Condition (42) stipulates that the distribution of $\eta(X)$ is sufficiently spread out, its density cannot have atoms or infinite peaks for instance. Under this condition a rate of convergence of the order of $n^{-1+\epsilon}$ is achievable for any $\epsilon > 0$. The reduced variance of the U-statistic $L(r_n)$ has been crucially exploited to derive fast rates from the rather weak condition (42). Applying a similar reasoning for the variance of $q_s((X, Y), (X', Y'))$ (which would be the case if one considered a risk estimate based on independent pairs by splitting the training data into two halves, see section 0.16), would have led to the condition:

$$|\eta(x) - \eta(x')| \geq c, \tag{43}$$

for some constant c, and $x \neq x'$, which is satisfied only when $\eta(X)$ has a lattice distribution.

Noiseless regression model. Consider the *noise-free regression model* in which $Y = m(X)$ for some (unknown) function $m : \mathcal{X} \to \mathbb{R}$. Here we have $L^* = 0$ and the Bayes ranking rule is given by the scoring function $s^* = m$. In this case

$$q_r(x, x') = \mathbb{I}_{[(m(x)-m(x'))\cdot r(x,x')<0]}$$

and

$$\mathrm{Var}(h_r(X, Y)) \leq \mathbb{E}q_r^2(X, X') = L(r) .$$

Hence, the condition of Proposition 0.17 is satisfied with $c = 1$ and $\alpha = 1$. If $\mathcal{R}$ has finite VC dimension V, the empirical risk minimizer r_n is thus such that, with probability at least $1 - \delta$,

$$L(r_n) \leq 2 \inf_{r \in \mathcal{R}} L(r) + C\frac{V \log(n/\delta)}{n}.$$

Regression model with noise. Consider now the *general regression model with heteroskedastic errors* in which $Y = m(X) + \sigma(X)\epsilon$ for some (unknown) functions $m : \mathcal{X} \to \mathbb{R}$ and $\sigma : \mathcal{X} \to \mathbb{R}$, where ϵ is a standard gaussian r.v., independent of X. Set

$$\Delta(X, X') = \frac{m(X) - m(X')}{\sqrt{\sigma^2(X) + \sigma^2(X')}} .$$

We have again $s^* = m$ and the optimal risk is

$$L^* = \mathbb{E}\Phi\left(-|\Delta(X, X')|\right)$$

where Φ is the distribution function of the standard gaussian random variable. The maximal value of L^* is attained when the regression function $m(x)$ is constant. Furthermore, we have

$$L(s) - L^* = \mathbb{E}\left(\left|2\Phi\left(\Delta(X, X')\right) - 1\right| \cdot \mathbb{I}_{[(m(x)-m(x'))\cdot(s(x)-s(x'))<0]}\right) .$$

Noise assumption: there exist constants $c > 0$ and $\alpha \in [0,1]$ such that for all $x \in \mathcal{X}$,

$$\mathbb{E}_{X'}(|\Delta(x, X')|^{-\alpha}) \leq c\,. \tag{44}$$

Proposition 33 *(Clémençon, Lugosi & Vayatis, 2005b) Under (44), we have:* $\forall s \in \mathcal{F}$

$$\mathrm{Var}(h_s(X,Y)) \leq (2\Phi(c) - 1)\,\Lambda(s)^{\alpha}\,.$$

The preceding noise condition is fulfilled in many cases, as illustrated by the example below.

Corollary 34 *(Clémençon, Lugosi & Vayatis, 2005b) Suppose that* $m(X)$ *has a bounded density and the conditional variance* $\sigma(x)$ *is bounded over* $\mathcal{X}$*. Then the noise condition (44) is satisfied for any* $\alpha < 1$*.*

The argument above still holds even if the gaussian noise assumption is dropped. Indeed the assumption that the r.v. ϵ has a symmetric density decreasing over $\mathbb{R}_+$ is only required.

0.19 Further remarks on convex risk minimization

Many popular classification algorithms, including various versions of *boosting* and *support vector machines* and performing very well in practice, consist in minimizing a convex version of the empirical risk over a certain class of functions (typically over a ball of an appropriately chosen Hilbert or Banach space of functions) obtained by replacing the loss function by a convex function. Minimizing empirical convex functionals being numerically feasible by gradient descent algorithms, this approach as important computational advantages. The statistical behavior of such methods have been recently thoroughly investigated (see [16], [43], [120], [135] or [200]). By adapting the arguments of [135] developed in the simple binary classification setup to the ranking problem, the principle of convex risk minimization has been extended to the ranking problem in [59], which paper provides a theoretical framework for the analysis of successful ranking algorithms such as the RankBoost algorithm of [86].

The basic idea is to consider ranking rules induced by real-valued functions, that is, ranking rules of the form

$$r(x,x') = \begin{cases} 1 & \text{if } f(x,x') > 0 \\ -1 & \text{otherwise} \end{cases}$$

where $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is some measurable real-valued function. We abusively denote by $L(f) = \mathbb{P}\{\mathrm{sgn}(Z) \cdot f(X,X') < 0\} = L(r)$ the risk of the ranking rule induced by f, where $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = -1$ if $x < 0$, and $\mathrm{sgn}(x) = 0$ if $x = 0$. Let $\phi : \mathbb{R} \to [0,\infty)$ be a convex *cost function* such that $\phi(0) = 1$ and $\phi(x) \geq \mathbb{I}_{[x \geq 0]}$. Typical choices of ϕ include the exponential cost function $\phi(x) = e^x$, the "logit" function $\phi(x) = \log_2(1 + e^x)$, or the "hinge loss" $\phi(x) = (1+x)_+$. Define the *cost functional* associated to the cost function ϕ by

$$A(f) = \mathbb{E}\phi(-\mathrm{sgn}(Z) \cdot f(X,X'))\ .$$

We clearly have that $L(f) \leq A(f)$. Let $A^* = \inf_f A(f)$ be the optimal value of the cost functional where the infimum is taken over all measurable functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The most natural estimate of the cost functional $A(f)$ is the *empirical cost functional* defined by the U-statistic

$$A_n(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \phi(-\mathrm{sgn}(Z_{i,j}) \cdot f(X_i, X_j))\ .$$

Consider the ranking rule based on minimizing the empirical cost functional A_n over a set $\mathcal{F}$ of real-valued functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, *i.e.* $f_n = \arg\min_{f\in\mathcal{F}} A_n(f)$:

$$r_n(x, x') = \begin{cases} 1 & \text{if } f_n(x, x') > 0 \\ -1 & \text{otherwise.} \end{cases}$$

As shown in Theorem 3 of [16], minimizing convex risk functionals is meaningful for ranking since the excess convex risk may be related to the excess ranking risk $L(f_n) - L^*$ as follows. Set $H(\rho) = \inf_{\alpha\in\mathbb{R}} (\rho\phi(-\alpha) + (1-\rho)\phi(\alpha))$ and $H^-(\rho) = \inf_{\alpha:\alpha(2\rho-1)\le 0} (\rho\phi(-\alpha) + (1-\rho)\phi(\alpha))$. And define $\psi(x) = H^-\left(\frac{1+x}{2}\right) - H^-\left(\frac{1-x}{2}\right)$. Then for all functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

$$L(f) - L^* \le \psi^{-1}(A(f) - A^*)$$

where ψ^{-1} is the inverse of ψ. If ϕ is chosen convex, it is shown in [16] that $\lim_{x\to 0} \psi^{-1}(x) = 0$: if the excess convex risk converges to zero, so does the excess ranking risk. Moreover, in most interesting cases $\psi^{-1}(x)$ may be bounded, for $x > 0$, by a constant multiple of $\sqrt{x}$ (such as in the case of exponential or logit cost functions) or even by x if $\phi(x) = (1 + x)_+$.

Hence, bounding the excess ranking risk $L(f) - L^*$ for convex risk minimization follows from analyzing the excess convex risk. This may be done by decomposing it into estimation and approximation errors:

$$A(f_n) - A^*(f) \le \left(A(f_n) - \inf_{f\in\mathcal{F}} A(f)\right) + \left(\inf_{f\in\mathcal{F}} A(f) - A^*\right).$$

One may naturally bound the excess convex risk over the class $\mathcal{F}$ as follows

$$A(f_n) - \inf_{f\in\mathcal{F}} A(f) \le 2\sup_{f\in\mathcal{F}} |A_n(f) - A(f)| .$$

For simplicity's sake, suppose that the class $\mathcal{F}$ is uniformly bounded, say $\sup_{f\in\mathcal{F},x\in\mathcal{X}} |f(x)| \le B$. Then Lemma 25 combined with the bounded differences inequality entails that for any $\lambda > 0$,

$$\begin{aligned} &\mathbb{E}\exp\left(\lambda \sup_{f\in\mathcal{F}} |A_n(f) - A(f)|\right) \\ &\quad\le \exp\left(\lambda\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} \phi\left(-\operatorname{sgn}(Z_{i,\lfloor n/2\rfloor+i})\cdot f(X_i, X_{\lfloor n/2\rfloor+i})\right) - A(f)\right) + \frac{\lambda^2 B^2}{2n}\right) . \end{aligned}$$

Now standard symmetrization and contraction inequalities may provide an upper bound for the expected supremum appearing in the exponent. In fact, by mimicking [126], we get

$$\begin{aligned} &\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} \phi\left(-\operatorname{sgn}(Z_{i,\lfloor n/2\rfloor+i})\cdot f(X_i, X_{\lfloor n/2\rfloor+i})\right) - A(f)\right) \\ &\qquad\le 4B\phi'(B)\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} \sigma_i\cdot f(X_i, X_{\lfloor n/2\rfloor+i})\right) \end{aligned}$$

where $\sigma_1, \ldots, \sigma_{\lfloor n/2\rfloor}$ i.i.d. Rademacher random variables independent of D_n, that is, symmetric sign variables with $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$.

The Rademacher average $R_n(\mathcal{F}) = \mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} \sigma_i\cdot f(X_i, X_{\lfloor n/2\rfloor+i})\right)$ may be easily bounded for various classes of functions. For example, consider the convex class $\mathcal{F}_B = \left\{f(x, x') = \sum_{j=1}^{N} w_j g_j(x, x\right.$ related to boosting-type classification procedures, where $\mathcal{R}$ is a class of ranking rules. In this case,

$$R_n(\mathcal{F}_B) \le BR_n(\mathcal{R}) \le \text{const.}\frac{BV}{\sqrt{n}}$$

where V is the VC dimension of the base class $\mathcal{R}$.

This can also be used for establishing performance bounds for kernel methods such as support vector machines. Let $k : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ be a symmetric positive definite function, that is,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(w_i, w_j) \geq 0 \,,$$

for all choices of n, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $w_1, \ldots, w_n \in \mathcal{X}^2$.

A kernel-type ranking algorithm may be defined as one that performs minimization of the empirical convex risk $A_n(f)$ (typically based on the hinge loss $\phi(x) = (1+x)_+$) over the class $\mathcal{F}_B$ of functions defined by a ball of the associated reproducing kernel Hilbert space of the form (where $w = (x, x')$)

$$\mathcal{F}_B = \left\{ f(w) = \sum_{j=1}^{N} c_j k(w_j, w) : \; N \in \mathbb{N}, \; \sum_{i,j=1}^{N} c_i c_j k(w_i, w_j) \leq B^2, w_1, \ldots, w_N \in \mathcal{W} \right\} .$$

See [65], [179], [184] or [186] for the approximation properties of such classes. We have

$$R_n(\mathcal{F}_B) \leq \frac{2B}{n} \mathbb{E} \sqrt{\sum_{i=1}^{\lfloor n/2 \rfloor} k((X_i, X_{\lfloor n/2 \rfloor + i}), (X_i, X_{\lfloor n/2 \rfloor + i}))} \,,$$

see for example [40].

In conclusion, universal consistency of such ranking rules may be derived in a straightforward way if the approximation error $\inf_{f \in \mathcal{F}_B} A(f) - A^* \to 0$ as $B \to \infty$.

Ranking the Best Instances

0.20 Introduction

In [62], an important variant of the bipartite ranking problem has been considered, namely the problem of ranking a given proportion of instances x among the "most relevant instances" only. Such a *local ranking problem* is of crucial practical importance, since in most ranking applications, in particular in the field of *Information Retrieval*, only top ranked instances are effectively scanned. In a similar fashion, scoring rules for *credit-risk screening* elaborated by credit establishment aim at indicating reliability, laying much more emphasis on most risky prospects. A novel criterion specifically tailored for selecting scoring functions accomplishing this task in an optimal fashion is required. As a matter of fact, recall that the AUC of a scoring function s is

$$\mathrm{AUC}(s) = \int_{\alpha=0}^{1} \beta_s(\alpha) d\alpha. \tag{45}$$

It is thus simply the mean power of the test function s over all possible levels α: the criterion $\mathrm{AUC}(s)$ weights in an uniform fashion all test errors, independently from their level. This may be not appropriate when one seeks for scoring functions s that induce an accurate ordering for the most relevant instances x only. More precisely, let $u_0 \in [0, 1]$ be fixed. The local ranking problem described above is somehow a double issue, consisting in simultaneously determining and properly ordering the $u_0\%$ the most relevant instances. Denoting by F_η the cdf of $\eta(X)$, the matter is to 1. determine the set $G^*_{u_0}$ of instances that are considered relevant enough and 2. recover the order induced by η on $G^*_{u_0}$.

By way of preliminary, we first consider the problem of finding the best $u_0\%$ instances, which we formulate as a binary classification problem with mass constraint. Then a new theoretical criterion, wich we call the *generalized AUC*, is proposed for solving precisely the problem mentioned above. It boils down to the standard AUC criterion for $u_0 = 1$. And (basic and fast) bound rates for the local ranking risk of empirical risk minimizers established in [62] under specific conditions are briefly recalled.

0.21 On Finding the Best Instances

We first tackle the problem of determining a given proportion $u_0 \in (0, 1)$ among the most relevant instances from training data. In other words, the matter is here to find an estimate of the set

$$G^*_{u_0} = \{x \in \mathcal{X} / \eta(x) > F_\eta^{-1}(1 - u_0)\}. \tag{46}$$

0.21.1 A mass-constrained classification problem

As shown by the result stated below, the set $G^*_{u_0}$ corresponds to the classifier $C : \mathcal{X} \to \{-1, +1\}$ with minimum misclassification probability $M(C) = \mathbb{P}(Y \neq C(X))$ among classifiers that assign positive label only to a proportion u_0 of the instances.

Theorem 35 *(Clémençon & Vayatis, 2006) Define* $C^*_{u_0}(X) = 2\mathbb{I}_{[X \in G^*_{u_0}]} - 1$ *and denote* $M^*_{u_0} = M(C^*_{u_0})$. *Then for any classifier* C *such that* $\mathbb{P}(C(X) = 1) = u_0$,

$$M^*_{u_0} \leq M(C).$$

Furthermore, we have

$$M(C) - M^*_{u_0} = 2\mathbb{E}(\left|\eta(X) - F_\eta^{-1}(1 - u_0)\right| \mathbb{I}_{[X \in G^*_{u_0} \Delta G]}), \tag{47}$$

with $G = \{x \in \mathcal{X} / C(x) = 1\}$ *and denoting by* Δ *the symmetric difference between two subsets of* $\mathcal{X}$.

Remark 3 Note that if a classifier C satisfies the mass constraint $\mathbb{P}(C(X) = 1) = u_0$, its type I error, *i.e* the probability that C leads to an incorrect prediction on a negative labeled example, $\alpha(C) = \mathbb{P}(C(X) = +1 \mid Y = -1)$, is related to its misclassification error by

$$M(C) = 2(1 - p)\alpha(C) + p - u_0.$$

Similarly, we have $M(C) = 2p(1 - \beta(C)) + p - u_0$ with $\beta(C) = \mathbb{P}(C(X) = 1 \mid Y = +1)$. Hence, $C^*_{u_0}$ is also the classifier with minimum type I error (respectively, type II error) among mass-constrained classifiers.

Although the point of view adopted in this chapter is very different (since we are mainly concerned here with building scoring functions), the problem described above may be formulated in the framework of *minimum volume set* learning as considered in [178]. As a matter of fact, the set $G^*_{u_0}$ may be viewed as the solution of the constrained optimization problem:

$$\min_{G \text{ measurable}} \mathbb{P}(X \in G \mid Y = -1)$$

subject to

$$\mathbb{P}(X \in G) \geq u_0.$$

Although the *volume set* is generally computed from a known measure of reference (contrary to the probability involved in the constraint, which must be estimated) in such problems (applications of the latter being mainly related to anomaly detection), learning methods for MV-set estimation may hopefully be extended to our setting. A natural way to do it would consist in replacing X's distribution conditioned on $Y = -1$ by its empirical counterpart. This point has not been treated in [62] but will be the subject of further investigation.

0.21.2 Empirical risk minimization

The goal is to investigate how to build estimates of the set $G^*_{u_0}$ based on training data. Suppose that we are given n i.i.d. copies of the pair (X, Y): $(X_1, Y_1), \cdots, (X_n, Y_n)$. Since we are mainly motivated here by scoring applications, it is natural to consider candidate sets obtained by thresholding a scoring function s at level $F_s^{-1}(1 - u_0)$, as in (46), for solving the mass-constrained classification problem described above, selection being ideally based on minimizing the quantity

$$M_n(s) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i \neq 2\mathbb{I}_{[s(X_i) \geq F_s^{-1}(1 - u_0)]} - 1\}.$$

Unfortunately, the adequate threshold level depends on X's distribution and is generally unknown. In practice, one has to replace it by its empirical counterpart $\hat{F}_s^{-1}(1 - u_0)$, denoting by $\hat{F}_s(x) = n^{-1} \sum_i \mathbb{I}_{[s(X_i) \leq x]}$ the empirical cdf of the $s(X_i)$'s, so as to consider

$$\hat{M}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \neq 2\mathbb{I}_{[s(X_i)\geq \hat{F}_s^{-1}(1-u_0)]} - 1\}.$$

Note that $\hat{M}_n(s)$ is a biased estimate of the misclassification risk $M(s) = M(C_s)$ of the classifier $C_s(X) = 2\mathbb{I}_{[s(X)\geq F_s^{-1}(1-u_0)]} - 1$. And the interesting part of the analysis consists in showing that the bias induced by plugging the empirical quantile estimate in the risk functional does not deteriorate the rate bound, as stated in the next proposition which provides a standard bound for the performance of minimizers of $\hat{M}_n(s)$ over a class $\mathcal{S}_0$ of scoring functions satisfying mild conditions. With no restriction, we may assume that all scoring functions $s \in \mathcal{S}_0$ take their values in $(0,1)$. Define the empirical risk minimizer over $\mathcal{S}_0$ by

$$\hat{s}_n = \arg\inf_{s\in\mathcal{S}_0} \hat{M}_n(s).$$

The following assumptions are required in the next result.

- (i) The class of functions $x \mapsto \mathbb{I}_{[s(x)\geq t]}$, indexed by $(s,t) \in \mathcal{S}_0 \times (0,1)$, has finite VC dimension V.
- (ii) For all $s \in \mathcal{S}_0$, $F_{s,+}$ and $F_{s,-}$ are differentiable on $(0,1)$ with derivatives uniformly bounded: there exist strictly positive constants b_-, b_+, B_- and B_+ s.t. $\forall (s,v) \in \mathcal{S}_0 \times (0,1)$,

$$b_+ \leq F'_{s,+}(v) \leq B_+ \text{ and } b_- \leq F'_{s,-}(v) \leq B_-. \tag{48}$$

Proposition 36 *(Clémençon & Vayatis, 2006) Suppose that conditions (i)-(ii) hold. Let $\delta > 0$. Then with probability at least $1-\delta$,*

$$M(\hat{s}_n) - \inf_{s\in\mathcal{S}_0} M(s) \leq c_1\sqrt{\frac{V}{n}} + c_2\sqrt{\frac{\ln(1/\delta)}{n-1}}, \tag{49}$$

for some constants c_1 and c_2.

0.21.3 Fast Rates

As in section 0.17, it is possible to improve significantly the rate (49) in some cases. Consider the following conditions.

- (iii) There exist constants $\alpha \in]1/2, 1[$ and $B > 0$ s.t. for all $t \geq 0$

$$\mathbb{P}(\left|\eta(X) - F_\eta^{-1}(1-u_0)\right| \leq t) \leq Bt^{\frac{\alpha}{1-\alpha}}.$$

- (iv) The class of df's F_s with $s \in \mathcal{S}_0$ is a subset of $\mathcal{H}(\beta, L)$, the Hölder class of functions $F : (0,1) \to \mathbb{R}$ satisfying:

$$\sup_{(x,y)\in(0,1)^2} \frac{|F(x) - F(y)|}{|x-y|^\beta} \leq L,$$

  with $L < \infty$ and $\beta \geq 1$.
- (v) The class $\mathcal{S}_0$ is finite with cardinal N.
- (vi) The class $\mathcal{S}_0$ contains an optimal scoring function s^*.

Observe first that for all $t \geq 0$,

$$\mathbb{P}(\left|\eta(X) - F_\eta^{-1}(1-u_0)\right| \leq t) = F_\eta(t + F_\eta^{-1}(1-u_0)) - F_\eta(-t + F_\eta^{-1}(1-u_0)).$$

Hence, if conditions (iv) and (vi) are both fulfilled, so is (iii) with $\alpha = \beta/(1+\beta)$ and $B = L$. These conditions are by no means the weakest conditions under which faster rates may be derived, but since our main goal here is to give an insight into the problem, we use them to make the formulation of the results simpler. Hence, as shall be seen, the bound (49) may be significantly improved under the conditions above. Define

$$s_n = \arg\inf_{s \in \mathcal{S}_0} M_n(s).$$

Besides, from the relation (47), one may easily derive that

$$Var(\mathbb{I}_{[C_s(X) \neq Y]} - \mathbb{I}_{[C^*_{u_0}(X) \neq Y]}) \leq (M(s) - M^*_{u_0})^\alpha$$

under (iii). And by slightly adapting the now standard argument in [192] based on Bernstein's inequality (see also references in section 0.17), a sharper estimate of the convergence rate for the excess risk of the minimizer s_n of the 'ideal' empirical criterion M_n may be established if one assumes further that (v)-(vi) hold: with probability $1-\delta$,

$$M(s_n) - M^*_{u_0} \leq c(\log(N/\delta)/n)^{\frac{1}{2-\alpha}}. \tag{50}$$

This bound may be preserved to some extent when minimizing the biased empirical criterion, as claimed in the next theorem. The proof relies on assumption (iv), allowing us to control the deviation between $M_n(s)$ and $\hat{M}_n(s)$ uniformly over $s \in \mathcal{S}_0$, combined with Theorem 8.3 in [145], which permits to evaluate the performance of the $\hat{M}_n$-minimizer in terms of M_n-risk.

Theorem 37 *(Clémençon & Vayatis, 2006) Under assumptions (ii)-(vi), we have with probability* $1-\delta$

$$M(\hat{s}_n) - M^*_{u_0} \leq c_1(\log(2N/\delta)/n)^{\frac{1}{2-\alpha}} + c_2(\log(2N/\delta)/n)^{\beta/2}. \tag{51}$$

This bound calls for some comments. As a matter of fact, taking $\beta = \alpha/(1-\alpha)$ the first term corresponds to the rate bound when quantile estimation poses no problem, namely for $\alpha \in (1/2, 2-\sqrt{2})$, while the second term is governed by the dificulty of approximating (uniformly) the component of the risk involving quantile, and is faster than the first one for $\alpha \in (2-\sqrt{2}, 1)$ only. But truth should be said, we do not know at present whether this 'elbow' phenomenon in the rate is simply due to our method of proof or to a deeper aspect of the problem considered (establishing lower bounds would be then required).

0.22 The Local Ranking Problem

The *local ranking problem* as considered in section 0.20 may be stated in very simple terms: the matter is to order the instances of the input space $\mathcal{X}$ so that the best $u_0\%$ be ranked as accurately as possible. And the class $\mathcal{S}^*$ of scoring functions solving it in an optimal manner is clearly the subset of $\mathcal{S}$ made of functions s such that for all $u \in (0, u_0)$

$$G_{s,u} = G^*_u,$$

with $G_{s,u} = \{x \in \mathcal{X}/s(x) > F_s^{-1}(1-u))\}$, denoting by F_s the (unconditional) cdf of $s(X)$. In other terms, $\mathcal{S}^*$ is the collection of scoring functions s^* such that:

$$s^*(x) = \Phi(\eta(x)), \text{ on } G^*_{u_0}$$

and

$$s^*(x) < \inf_{x' \in G^*_{u_0}} s^*(x'), \text{ for all } x \notin G^*_{u_0},$$

where $\Phi : [0,1] \to \mathbb{R}$ is any Borel function of which restriction on $]F_\eta(1-u_0), 1]$ is strictly increasing. Next we propose a performance measure for evaluating the pertinence of scoring functions regarding to the local ranking problem.

0.22.1 Tailoring a criterion for the local ranking problem

By way of preliminary, we first state the following result.

Lemma 38 *(Clémençon & Vayatis, 2006) For any $s \in \mathcal{S}$, we have for all $u \in (0,1)$*

$$\begin{aligned} F_{\eta,+}(F_\eta^{-1}(u)) &\leq F_{s,+}(F_s^{-1}(u)), \\ F_{s,-}(F_s^{-1}(u)) &\leq F_{\eta,-}(F_\eta^{-1}(u)), \end{aligned}$$

and there is equality in the sole case when $G_{s,1-u} = G^*_{1-u}$.

In view of this result, a wide collection of criteria with $\mathcal{S}^*$ as set of optimal values could be naturally considered, depending on how one weights the type II error $1-\beta(s,u) = F_{s,+}(F_s^{-1}(1-u))$ (resp. the type I error $\alpha(s,u) = 1-F_{s,-}(F_s^{-1}(1-u))$) according to the value of the mass $u \in [0, u_0]$. However, not all criteria obtained this manner can be easily interpreted.

AUC generalization

A possible way of generalizing the standard AUC would be to consider

$$AUC_{u_0}(s) = \mathbb{P}(\{s(X) > s(X')\} \cap \{s(X) \geq F_s^{-1}(1-u_0)\} \mid Y = 1, Y' = -1). \tag{52}$$

It obviously boils down to the standard AUC criterion (45) for $u_0 = 1$. And as claimed in the following theorem resulting from lemma 38, $\mathcal{S}^*$ corresponds to the set of *generalized AUC* maximizers. Furthermore, this criterion may be expressed in a simple fashion in terms of the type II error $\beta(s,u_0)$ measuring how accurate is the estimation G_{s,u_0} of $G^*_{u_0}$ (see section 0.21) and a quantity measuring the performance of the ranking induced by s on the set G_{s,u_0}, as described in chapter 5.

Theorem 39 *(Clémençon & Vayatis, 2006) Let $u_0 \in (0,1)$. We have*

$$\mathcal{S}^* = \arg\min_{s \in \mathcal{S}} AUC_{u_0}(s^*).$$

Furthermore, we have

$$AUC_{u_0}(s) = \beta(s,u_0) - \frac{1}{2p(1-p)} L(s, G_{s,u_0}), \tag{53}$$

with $L(s,G) = \mathbb{P}(\{(s(X)-s(X'))(Y-Y') > 0\} \cap \{(X,X') \in G^2\})$ for any measurable $G \subset \mathcal{X}$.

Observe that (53) reduces to (39) when $u_0 = 1$. Besides one may relate the generalized AUC criterion to the truncated AUC. As a matter of fact, one may write

$$AUC_{u_0}(s) = \int_{\alpha=0}^{\alpha(s,u_0)} \beta_s(\alpha) d\alpha + \beta(s,u_0) - \alpha(s,u_0)\beta(s,u_0). \tag{54}$$

The values $\alpha(s,u_0)$ and $\beta(s,u_0)$ are the coordinates of the intersecting point between the ROC curve $\alpha \mapsto \beta_s(\alpha)$ and the line $D : \beta = -\frac{1-p}{p}\alpha + \frac{u_0}{p}$. And the integral $\int_{\alpha=0}^{\alpha(s,u_0)} \beta_s(\alpha) d\alpha$ represents

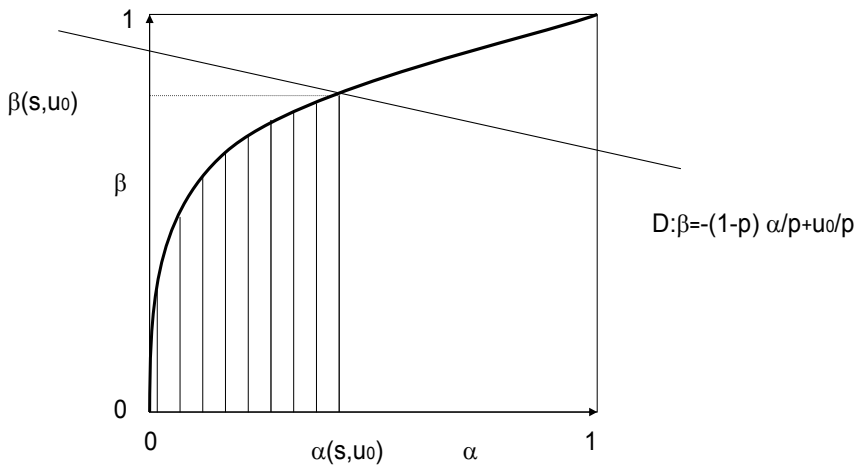

Figure 10: Truncated AUC.

the area of the surface delimited by the ROC curve, the α-axis and the line $\alpha = \alpha(s, u_0)$ (see Fig. ??). It is worth mentioning that, contrary to what is claimed in [153], it is not necessarily pertinent to evaluate the local performance of a scoring statistic $s(X)$ by the truncated AUC, since in general $\arg\min_{s\in\mathcal{S}} \int_{\alpha=0}^{\alpha(s,u_0)} \beta_s(\alpha)d\alpha \neq \mathcal{S}^*$.

Generalized Mann-Whitney Wilcoxon statistic

Here is another possible fashion of generalizing the AUC criterion. A natural empirical estimate of the AUC is the *rate of concording pairs*:

$$\hat{AUC}(s) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_-} \sum_{j=1}^{n_+} \mathbb{I}_{[s(X_i^-)<s(X_j^+)]},$$

with $n_+ = n - n_- = \sum_{i=1}^{N} \mathbb{I}_{[Y_i=+1]}$ and denoting by $\{X_i^+\}_{1\leq i\leq n_+}$ (resp. by $\{X_i^-\}_{1\leq i\leq n_-}$) the set of instances with positive labels (resp. with negative labels) among the sample data $\{(X_i), Y_i)\}_{1\leq i\leq n}$. And we have the classical relation

$$\frac{n_+ n_-}{n+1} \hat{AUC}(s) + \frac{n_+(n_+ + 1)}{2} = W_{n,1}(s), \tag{55}$$

where $W_{1,n}(s)$ is the standard Wilcoxon statistic. Recall that it is the *two-sample linear rank statistic* associated to the *score generating function* $\Phi_1(u) = u$, $u \in (0, 1)$, obtained by summing the ranks corresponding to positive labels, namely

$$W_{1,n}(s) = \sum_{i=1}^{n_+} \frac{\mathrm{rank}(s(X_i^+))}{n+1},$$

denoting by $\mathrm{rank}(s(X_i^+))$ the rank of $s(X_i^+)$ in the pooled sample $\{s(X_j), 1 \leq j \leq n\}$. Refer to [197] for basic results related to linear rank statistics. In particular, the statistic $\hat{\mu}_1(s) = W_{1,n}(s)/n_+$ is an asymptotically normal estimate of

$$\begin{aligned} \mu_1(s) &= \mathbb{E}_X[\mathbb{P}_{X'}(s(X') < s(X)) \mid Y = 1] \\ &= \mathbb{E}[F_s(s(X)) \mid Y = 1], \end{aligned}$$

where X' denotes an independent copy of X and $\mathbb{P}_{X'}$ is the probability taken with respect to the variable X' (conditional probability given (X, Y)). Note the theoretical counterpart of (55) may be written as

$$\mu_1(s) = (1-p)AUC(s) + p/2, \tag{56}$$

and this quantity is related to the ranking risk studied in Chapter 5 by

$$L_1(s) = -2p\mu_1(s) + 2p(1-p) + p^2. \tag{57}$$

Now, in order to take into account the highest $u_0\%$ ranks only, one may consider the criterion related to the score generating function $\Phi_{u_0}(u) = u\mathbb{I}_{[u>1-u_0]}$

$$\begin{aligned} \mu_{u_0}(s) &= \mathbb{E}_X[\Phi_{u_0}(\mathbb{P}_{X'}(s(X') < s(X))) \mid Y = 1] \\ &= \mathbb{E}[\Phi_{u_0}(F_s(s(X))) \mid Y = 1]. \end{aligned}$$

It has empirical counterpart $\hat{\mu}_{u_0} = W_{u_0,n}(s)/n_+$, with

$$W_{u_0,n}(s) = \sum_{i=1}^{n} \mathbb{I}_{[Y_i=1]}\Phi_{u_0}\left(\frac{\mathrm{rank}(s(X_i))}{n+1}\right).$$

Define the *local ranking risk* at level u_0 of any $s \in \mathcal{S}$ by

$$L_{u_0}(s) = -2p\mu_{u_0}(s) + 2p(1-p) + p^2,$$

generalizing this way (57). The next result, also based on (38), claims that $\mathcal{S}^*$ coincides with the set of scoring functions with minimum local ranking risk.

Theorem 40 *(Clémençon & Vayatis, 2006) Let $u_0 \in [0,1]$ be fixed. We have*

$$\mathcal{S}^* = \arg\min_{s \in \mathcal{S}} L_{u_0}(s).$$

Furthermore,

$$2p\mu_{u_0}(s) - 1 + (1-u_0)^2 = -L(s, G_{s,u_0}) - 2(1-p)(1-u_0)\alpha(s,u_0) - (1-p)^2\alpha(s,u_0)^2.$$

It is worth noticing that not all combinations of $\alpha(s,u_0)$ and $L(s, G_{s,u_0})$ lead to a criterion with $\mathcal{S}^*$ as optimal set. This prevents from considering naive 'divide and conqueer' strategies for solving the local ranking problem (*i.e.* strategies consisting in simply computing first an estimate $\hat{G}$ of the set containing the best instances and secondly solving the ranking problem over $\hat{G}$ as described in Chapter 5) and stresses the importance of making use of a global criterion, synthesizing the double goal one would like to achieve: finding and ranking the best instances.

0.22.2 Empirical risk minimization

Finally we collect here some remarks about the method considered in [62] for studying the performance of empirical risk minimizers, when risk is measured either by the generalized AUC or else by the local ranking risk. As an illustration, we deal with the case when accuracy of the scoring rule is measured by the generalized AUC.

The empirical generalized AUC criterion may be naturally decomposed as follows:

$$\widehat{AUC}_n(s) = \widetilde{AUC}_n(s) + (\widehat{AUC}_n(s) - \widetilde{AUC}_n(s)),$$

denoting by $\widetilde{AUC}_n(s) = \frac{n(n-1)}{n_+ n_-}\mathcal{U}_n(s)$ the empirical AUC in the ideal case when the quantile value is known and with

$$\mathcal{U}_n(s) = \frac{1}{n(n-1)} \sum_{i,j} \mathbb{I}_{[Y_i=-1, Y_j=1, s(X_i)<s(X_j), s(X_j)>F_s^{-1}(1-u_0)]},$$

which is the (non symmetric) U-statistic of degree 2 with kernel

$$q_s((x,y),(x',y')) = \mathbb{I}_{[y=-1, y'=1, s(x)<s(x'), s(x')>F_s^{-1}(1-u_0)]}.$$

Hence, even symmetrization of the U-statistic $\mathcal{U}_n(s)$ is needed first, exactly the same tools of the theory of U-processes as those used in chapter 5 permit to control $\sup_{s\in\mathcal{S}_0} \left|\widetilde{AUC}_n(s) - AUC(s)\right|$ under suitable conditions. And besides, the difference $\sup_{s\in\mathcal{S}_0} \left|\widehat{AUC}_n(s) - \widetilde{AUC}_n(s)\right|$ is controlled using the same tricks as those invoked for the mass-constrained classification problem in the previous section, but for establishing here that minimizing $\widehat{AUC}_n(s)$ is almost aquivalent to minimizing $\widetilde{AUC}_n(s)$ (the classicial results required may be found mainly in section 8 of [145] and chapter 5 in [196]). Hence, similar rate bounds as those stated for ranking risk in chapter 5 may be then easily derived for the local ranking problem. We refer to [62] for the formulation of the latter results.

Part III

Probabilistic Modelling and Applied Statistics

Abstract

In this last part, several applications of recent advances in applied probability and nonparametric statistics are presented. Among the numerous disciplines with which applied mathematics may successfully interface, finance is a natural field of application of probability and statistics. Beyond the ubiquity of data in this domain, risk and uncertainty lie indeed at the core of the financial market activity.

In [60] a nonparametric methodology developped by [74] for estimating an autocovariance sequence is applied to the statistical analysis of financial returns and advantages offered by this approach over other existing methods like fixed-window-length segmentation procedures are argumented. Theoretical properties of adaptivity of this method have been proved for a specific class of time series, the class of *locally stationary processes*, with an autocovariance structure which varies slowly over time in most cases but might exhibit abrupt changes of regime. It is based on an algorithm that selects empirically from the data the tiling of the time-frequency plane which exposes best in the least squares sense the underlying second-order time-varying structure of the time series, and so may properly describe the time-inhomogeneous variations of speculative prices. The applications considered in [60] mainly concern the analysis of structural changes occuring in stock market returns, VaR estimation and the comparison between the variation structure of stock indexes returns in developed markets and in developing markets.

In [61], the Independent Component Analysis (ICA) methodology is applied to the problem of selecting portfolio strategies, so as to provide against extremal movements in financial markets. A specific semi-parametric ICA model for describing the extreme fluctuations of asset prices is introduced, stipulating that the distributions of the IC's are heavy tailed. An inference method based on conditional maximum likelihood estimation has been proposed, which permits to determine practically optimal investment strategies with respect to extreme risk.

In Biosciences, statistical analysis, embracing probabilistic modeling, estimation, hypothesis testing and design of experiments is also of crucial importance. In epidemiology, it may actually play an essential role in public health practice, as illustrated by the topic treated in Chapter 8: dietary contamination in toxicology.

In [26] a specific piecewise-deterministic Markov process for describing the temporal evolution of exposure to a given food contaminant. The quantity X of contaminant present in the body evolves through its accumulation after repeated dietary intakes on the one hand and the kinetics behavior of the chemical on the other hand. The accumulation phenomenon is viewed as a marked point process and elimination in between intakes occurs at a random linear rate θX, randomness of the coefficient θ accounting for the variability of the elimination process due to metabolic factors. Ergodic properties of this process have been investigated, the latter being of crucial importance for describing the long-term behavior of the exposure process and assessing values of quantities such as the proportion of time the body burden in contaminant is over a certain threshold. The process being not directly observable, simulation-based statistical methods for estimating steady-state or time-dependent quantities have also been studied.

Applications in Finance

0.23 Time-Frequency Analysis of Financial Time Series

The modeling of the temporal variations of stock market prices has been the subject of intense research for a long time now, starting with the famous *Random Walk Hypothesis* introduced in [14], claiming that the successive stock price variations $(X_{t+1} - X_t)_{t \geqslant 0}$ are i.i.d. Gaussian r.v.'s. As numerous statistical studies showed, even if X_t is replaced by $\log(X_t)$, this classic model cannot explain some prominent features, such as the number of large price changes observed, that is much larger than predicted by the Gaussian. As emphasized by many statistical works, far too numerous to mention, the following features of stock price series came into sight.

- Spells of small amplitude for the price variations alternates with spells of large amplitude. This phenomenon is traditionally called *volatility persistence*.
- The "efficient markets assumption", which claims, roughly speaking, that financial returns are unforecastable, seems to be contradicted by the existence of very localized periods when return sequences exhibit strong positive autocorrelation.
- The magnitude of the variations evolves in the long run so as to reach an "equilibrium" level, one calls this feature *mean-reversion* in a stylized manner.

Although the classic *Random Walk model* provides explicit formula for asset pricing and the economic doctrine is able to interpret it, the limitations mentioned above motivated the emergence of an abundant econometric literature, with the object to model structure in financial data. Portfolio selection/optimization, Value at Risk estimation, hedging strategies are the main stakes of this research activity, still developing.

Even if the seminal contribution of [82], which introduced the *ARCH model*, has been followed by a large number of variants, the whole complexity underlying these data has not been captured yet by any parsimonious model and let the field of statistical analysis of financial time series open to further investigation. Selecting a statistical procedure, which allows to handle properly the time-inhomogeneous character of return series, is not an easy task, as Mandelbrot (1963) emphasized: "Price records *do not* "look" stationary, and statistical expressions such as the sample variance take very different values at different times; this nonstationarity seems to put a precise statistical model of price change out of the question". According to the estimation method chosen, one can either enhance specific patterns in the data or else make them disappear. This strongly advocates for the application of recent adaptive nonparametric procedures to the statistical analysis of financial series, which, by selecting adaptively from the data a "best" representation among a large (non-parametric) class of models and including the type of structure that contributes significantly to the fit of the model only, allow to achieve more flexibility. This alternative approach has been followed by some authors for several years now (see [161], [163]). Among these attempts to deal with non stationarities in financial data, one may mention the following works. [97] considered the use of fast

algorithms such as the *Matching Pursuit*, the *Method of Frames* and the *Basis Pursuit* to select adaptively, from a dictionary, the superposition of "atoms", that is to say elementary functions localized both in time and frequency (wavelets for instance), that best exhibit structure in a financial time series, viewed as a noisy signal. This methodology is applied to obtain sparse/parsimonious representations of exchange rate data in order to analyze the evolution of the frequency content of the underlying data generating process. In [162] the Matching Pursuit algorithm also applied over a larger dictionary, namely a waveform dictionary, to decompose more efficiently exchange rates using tick-by-tick data, allowing to detect the presence of significant low frequency components. In [49], the presence of pronounced GARCH effects in high frequency financial time series is investigated after a preliminary denoising of the data using the *wavelet shrinkage* procedure. Several statistical procedures have been based on an explicit functional modeling of the nonstationarities occuring in financial time-series. [106] introduced a method consisting in a sequence of nonparametric tests to identify periods of second-order homogeneity for each moment in time. The general formalism defining *locally stationary wavelet processes* is developped in [88] and applied to prediction and the time-varying second order structure estimation of the DJIA index.

In [60] the use of an adaptive nonparametric methodology developped by [74] for estimating the covariance of specific second-order nonstationary processes is promoted in the field of financial time-series statistical analysis. It amounts to analyze the data to find which out of a specific massive library of bases, *local cosine packets bases*, comes closest to diagonalizing the empirical covariance and make use of the latter to perform estimation. These bases have localization properties both in the time domain and in the frequency domain, and hopefully the selected basis may conveniently exhibit the time-varying character of the second-order structure of the time-series in some cases, as argumented by the numerical applications performed.

0.23.1 Statistical analysis of financial returns as locally stationary series

Heuristics A significant part of the information carried by economic and financial time series consists in temporal inhomogeneities: beginning or end of certain phenomena, ruptures due to shocks or structural changes, drifts reflecting economical trends, business cycles, etc. Stationarity is a concept introduced to mean the independence of statistical properties from time. Hence, nonstationarity simply expresses the need for reintroducing time as a necessary description parameter, so as to be able to speak about the evolution through time of some properties of the time series and compute meaningful statistics. A constructive fashion to deal with nonstationary time-series consists in restricting oneself to a class of time series, for which one is able to specify precisely how they diverge from stationarity, while keeping a certain level of generality. On grounds of parsimony, statistical analysis of stock prices variations mainly focused on the second order properties (that is not restrictive in the gaussian case), which amount to the covariance structure, since the assumption that financial returns are zero mean is beared out by both empirical evidence and theoretical economic arguments. It may be thus relevant to start with making assumptions on the autocovariance function. Consider a zero mean second order time series $X = (X_n)_{n\geqslant 0}$ with autocovariance $\Gamma_X(n,m) = E(X_n X_m) = C_X((n+m)/2, m-n)$. It seems natural to call X a *locally stationary* time series, when it is "approximately stationary" on time intervals of varying size and the variables are uncorrelated outside these quasi-stationarity intervals. As this class is supposed to describe random phenomena, which mechanism may evolve through time, it is legitimate to assume that the size $l(n)$ of the quasi-stationarity interval may depend on the time n on which it is centered. A qualitative characterization of *locally stationary processes* could be as follows: on each time interval $[n - l(n)/2, n + l(n)/2]$, the covariance between X_m and $X_{m'}$ at times m and m' may be well approximated by a function depending only on $m' - m$ as soon these time points

are close enough

$$\Gamma_X(m, m') \simeq C_X(n, m' - m) \text{ if } |m' - m| \leqslant l(n)/2 \tag{58}$$

and is approximately zero when the length between the time points considered is larger than a certain threshold $d(n)$ measuring somehow the "decorrelation rate" of the time series

$$\Gamma_X(m, m') \simeq 0 \text{ if } |m - m'| > d(n)/2. \tag{59}$$

Under these assumptions, for any time points $m \in [n - l(n)/2,\ n + l(n)/2]$ and $m' \geqslant 0$

$$C_X((m + m')/2, m' - m) \simeq C_X(n, m' - m). \tag{60}$$

Set out in such general terms, the concept of local stationarity seems to be relevant for modeling financial data and account for the features previously recalled. The returns of a security are known to decorrelate rapidly when the market behaves in an "efficient way", on equilibrium, but when the latter is "evolving", when a change of business cycle occur for instance, the autocorrelation structure may evolve too and then one may attend changes of regime.

There are many concepts of local stationarity, depending on the sense given to approximation (3). Following [159], most of them consist in generalizing the Cramer representation for stationary processes and defining a reasonable notion of time-varying spectral density. In [66] for instance the definition of locally stationary processes (including ARMA processes with time-varying coefficients) is based on a Karhunen representation and a specific transfer function. This viewpoint is also adopted in [127], where a wavelet basis is used to define a spectral representation and a "wavelet-periodogram". [140] showed that using local cosine packet bases also makes good sense to define locally stationary processes. Even if their goal is almost the same, namely to allow to extend the statistical tools and concepts (mainly stemmed from Fourier spectral analysis) available in the stationary framework, not all the approaches yield a tractable statistical procedure for which a precise study of its error may be carried out. In this respect, the one worked out in [74] combines several advantages. It refines the statistical method introduced in [73], who developped a full machinery to process the data and provides theoretical arguments to support it (see also [141] and [176]). This methodology applies to Gaussian triangular arrays of second-order processes $X^{(T)} = (X_{t,T})_{0 \leqslant t \leqslant T}$ for $T = 2^\tau$, $\tau = \tau_1, \tau_1 + 1, \ldots$ obeying the assumption of *uniform decay of the autocovariance*

$$\sum_{n=-t}^{T-t} \left(1 + 2\,|n|^{\delta_1}\right)^2 \Gamma^2_{X^{(T)}}(t, t + n) \leqslant c_1, \tag{61}$$

and the assumption of *quasi-stationarity of the covariance*

$$\frac{1}{T} \sum_{t=0}^{T} \|\Gamma_{X^{(T)}}(t, t + .) - \Gamma_{X^{(T)}}(t + h, t + h + .)\|_{l^2} \leqslant c_2 \left(\frac{|h|}{T}\right)^{\delta_2}, \tag{62}$$

for any h, where $\delta_1 > 1/2$, $0 < \delta_2 \leqslant 1$, c_1 and c_2 are constants. Beyond the scaling character of these assumptions and whereas (4) is a simple functional formulation of the heuristic condition (2) (expressing that the decay of the decorrelation rate $d(n)$ is faster than $n^{-\delta_1}$), their main attraction is due to the averaging component in (5): a stochastic process $X^{(T)}$ obeying this constraint has a covariance matrix, which nearby rows $\Gamma_{X^{(T)}}(t, t + .)$ are, *on average*, very similar, but might occasionnally be very different, allowing for sudden changes of regime.

The library of cosine packets bases The crucial point in the statistical analysis of a stationary time series $(X_t)_{t \in \aleph}$ consists in viewing it as a linear superposition of uncorrelated periodic elementary time series $A_r e_r(t)$, where the e_r's denote the functions of the Fourier basis and the

weights A_r are square integrable r.v.'s. The estimation of the variances of the A_r's from the record of the past observations of the time series yields both a low bias estimate of the covariance function and a spectral tool to analyze the structure of the time series: a current "physical" interpretation consists in measuring the relative importance of each periodic component $A_r e_r$ in the mechanism ruling the fluctuations of the time series by the variance of A_r. Since the 60's, spectral analysis has been in current use in econometrics for investigating the structure of economic series and computing predictions for instance). The idea underlying the use of the Coifman & Wickerhaüser system to describe locally stationary processes is to keep the notion of an expansion of the time series in a basis made of mutually orthogonal cosine functions, while introducing the point of view of *temporal localization*. The construction of this system amounts to concatenate adequately the sequences

$$\xi_{M,m}(t) = \sqrt{\frac{2}{M}} \cos\left(\omega_m(t+1/2)\right), \ 0 \leqslant t < M, \tag{63}$$

where $M = 2^j$ is a dyadic integer, $0 \leqslant m < M$ and $\omega_m = \pi(m+1/2)/M$. For reasons of a computational nature, the concatenations are induced by *recursive dyadic partitions* (RDP). A RDP of an inetravl $I_{0,0} = \{0, 1, ..., T-1\}$ of dyadic length $T = 2^\tau$ is any partition reachable from the trivial partition $\mathcal{P}_0 = \{I_{0,0}\}$ by successive application of the following rule: choose a dyadic subinterval $I_{j,k} = \left\{kT/2^j, ..., (k+1)T/2^j - 1\right\}$ in the current partition and split it into two (dyadic) subintervals $I_{j+1,2k}$ and $I_{j+1,2k+1}$ of same size, creating then a new (finer) partition. RDP may generate a very inhomogeneous segmentation of the time interval, with both very short subsegments and much longer ones for instance, permitting to describe the successive regimes of a nonstationary time series (see Fig. 11. Given a RDP $\mathcal{P}$ of the time axis $\{0, 1, ..., T-1\}$, one defines a *local cosine packets* basis $\mathcal{B}_\mathcal{P}$ of $\mathfrak{R}^T$ by setting

$$\varphi_{I_{j,k},m}(t) = \left\{ \begin{array}{ll} \xi_{2^{\tau-j},m}(t - kT/2^j) & \text{if } t \in I_{j,k} \\ 0 & \text{if } t \notin I_{j,k} \end{array} \right., \tag{64}$$

for all $I_{j,k}$ in $\mathcal{P}$, and $0 \leqslant m < 2^{\tau-j}$. Beyond their orthonormal character, the vectors of such a basis have the crucial property of being localized both in time and in frequency: $\varphi_{I_{j,k},m}$ is supported on the subinterval $I_{j,k}$, on which it oscillates at the frequency ω_m.

Every random sequence $X^{(T)} = (X_0, ..., X_{T-1})$ may be expanded in the local cosine packets basis $\mathcal{B}_\mathcal{P}$. Let $I_1, ..., I_{n_\mathcal{P}}$ be the subintervals forming $\mathcal{P}$. Then, one may write for $0 \leqslant t < T$

$$X_t^{(T)} = \sum_{u=1}^{n_\mathcal{P}} \sum_{m=0}^{2^{j_u}-1} \left\langle X^{(T)}, \varphi_{I_u,m} \right\rangle \varphi_{I_u,m}(t), \tag{65}$$

where $\langle .,. \rangle$ denotes the usual scalar product in $\mathfrak{R}^T$ and 2^{j_u} the length of the subinterval I_u. Thus, on each subsegment I_u of the time interval, one has a "Fourier type" decomposition of the time series into periodic time series. If, for each subinterval I_u, the components $\left\langle X^{(T)}, \varphi_{I_u,m} \right\rangle$ were almost uncorrelated, or if $\mathcal{B}_\mathcal{P}$ almost "diagonalizes" $\Gamma_{X^{(T)}}$ in an equivalent way (since $\mathcal{B}_\mathcal{P}\Gamma_{X^{(T)}}\mathcal{B}_\mathcal{P}'$ is the covariance matrix of the $\left\langle X^{(T)}, \varphi_{I_u,m} \right\rangle$'s), then one could interpret the segments $I_1, ..., I_{n_\mathcal{P}}$ as successive regimes of quasi-stationarity for the time series $X^{(T)}$. In [73] it is proved that for a *locally stationary* time series, there always exists such an "almost" diagonalizing basis and a *data-driven* method by complexity penalization has been introduced in [74] to select such a basis and yield a consistent covariance estimation procedure, the resulting covariance estimate being obtained by averaging, roughly speaking, contiguous time-frequency components with small variance, getting what is called *macrotiles*.

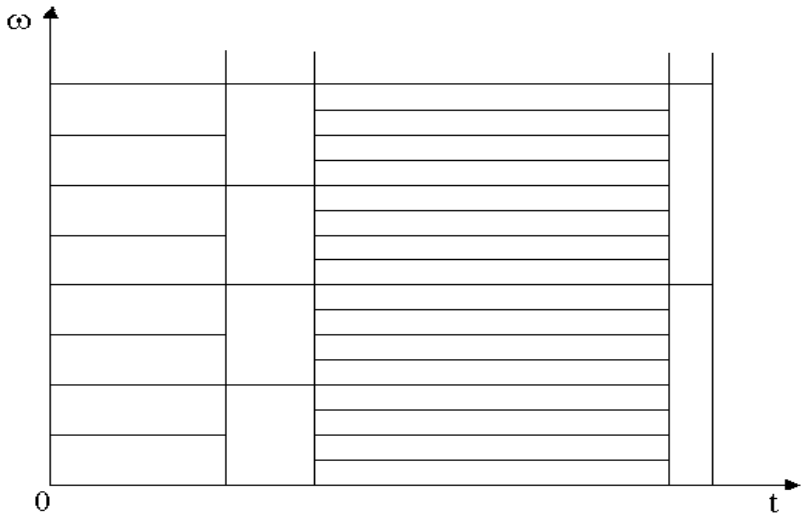

Figure 11: Recursive Dyadic Partition of the Time-Frequency Plane.

0.23.2 Empirical results

In [60] (see also [183]), several empirical studies based on the estimation procedure in [74] have been carried out. From the estimates thus obtained, the temporal inhomogeneities in the fluctuation of many financial returns have been investigated. These empirical studies provided in particular empirical evidence to support that the temporal behaviour of stock market returns usually diverges more from efficiency in emerging markets than in developed markets. To this concern, see the time-frequency representation of the return series of the IGPA index (Chile) for the period 1986 - 2002 (see Fig. 12) as estimated by the macrotile methodology, in comparison with the covariance estimate of the DJIA (Dow Jones) return series (refer to Fig. 13). As may be confirmed by application to many other series, quasi-stationary time intervals, on which the return series exhibits a strong autocorrelation, are larger and much more frequent in developing markets (that suggests more forecastability from past observations for these time series).

The estimation procedure has been also applied via a simple *plug-in* approach to Value at Risk forecasting (conditional mean and variance estimates being available as byproducts of the autocovariance estimate) using a forward data rolling history. It has been shown to have advantages over less flexible methods based on moving averages. Recall that VaR techniques intend to quantify the risk for an asset (respectively, a stock index, a portfolio), by measuring the level of loss $\mathrm{VaR}_{t,h}(\alpha_0)$ that the asset price I_t could loose over a given time horizon h with a given degree of confidence $1-\alpha_0$ at time t conditionally on the information available $\mathcal{I}_t$:

$$P\left((I_{t+h}-I_t)/I_t \geqslant \mathrm{VaR}_{t,h}(\alpha_0) \mid \mathcal{I}_t\right) = 1-\alpha_0.$$

In Fig. 14, the VaR plug-in estimate computed from the macrotile method is compared to three classical approaches: the VaR estimate based on the simple moving average (MA), the VaR estimate based on the Riskmetrics variance-covariance model built by an exponential weighted moving average (EMA) with a decay factor $\lambda = 0.94$, as Riskmetrics^{TM}, using for both a moving window with fixed length of 250 observation days, and the VaR estimate based on gaussian GARCH(1, 1) modeling. On may refer to [183] for a thorough comparison between the macrotile method and a wide range of VaR models. These encouraging empirical results suggest that locally stationary processes, as defined above, are relevant representations of financial data in many cases and clearly

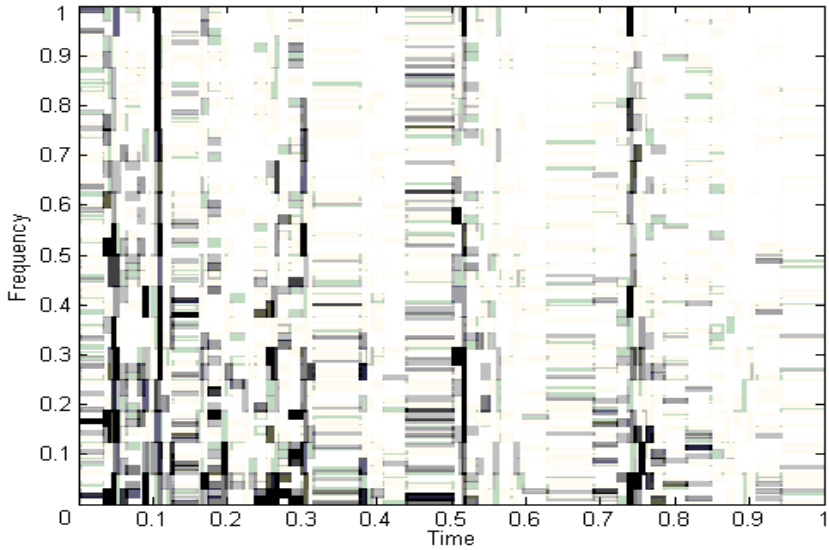

Figure 12: Time-Frequency Representation for the IGPA index.

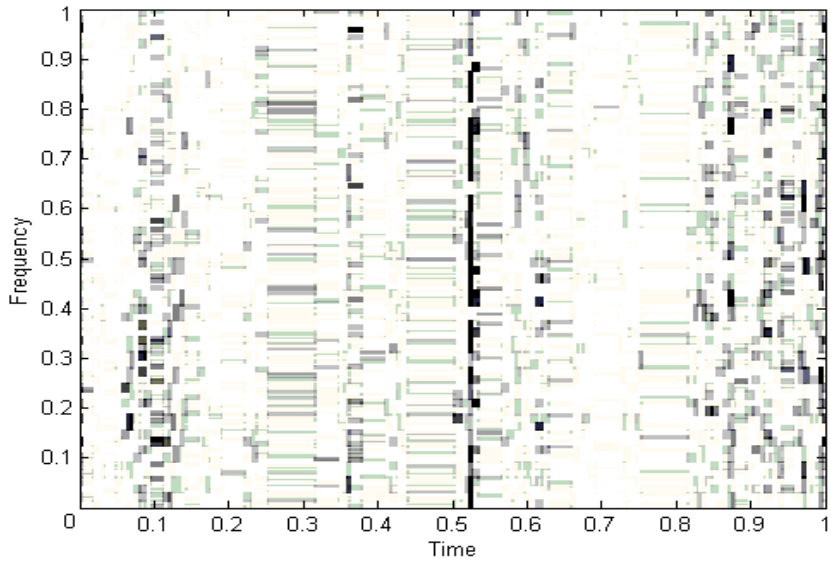

Figure 13: Time-Frequency Representation for the DJIA index.

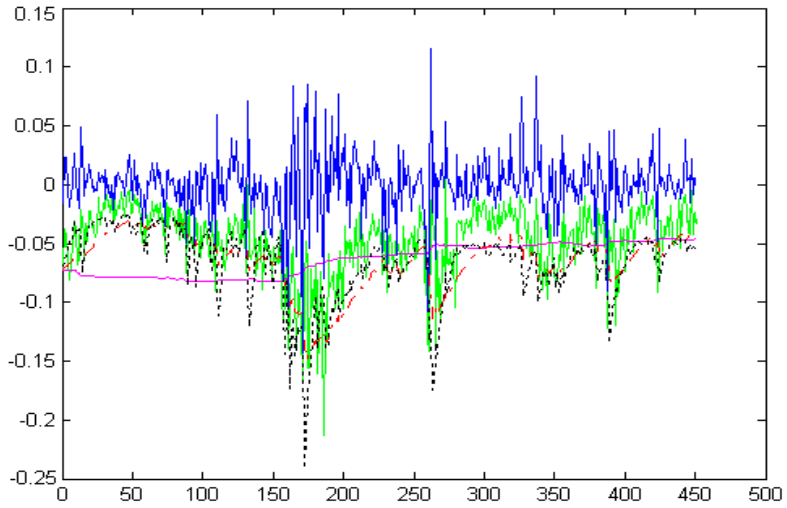

Figure 14: Time-plots of the 1-day ahead VaR forecasts of the Argentina market index from January 1998 to october 1999 using the best-basis method (green line), the EMA method (red dash-dot line), the GARCH method (black dotted line), the MA method (magenta dashed line) and the blue line denotes the daily return series.

indicate that the macrotile method, by successfully capturing some nonlinear features of the time-varying second order structure of financial series, may then provide reliable estimates for interval forecast. In particular, as illustrated by Fig. 14, the VaR computed from the macrotile method does not tend to overestimate the risk, contrary to many standard VaR methods of reference.

0.24 ICA Modelling for Safety-First Portfolio Selection

Borrowing tools and statistical techniques commonly used in the field of insurance for risk assessment, many finance experts have to face questions related to extremal events for handling problems concerning the probable maximal loss of investment strategies (see [81], [79] for instance and refer to [80] for a comprehensive overview of the applications of extreme values methodology to insurance and finance). In this specific context, namely when risk aversion takes precedence of potential gain in an overwhelming fashion (see [175] or [7]), standard methodologies for portfolio allocation, such as approaches based on mean-variance optimization (*cf* [142]), are not convenient any more. Hence, various methods have been recently proposed for addressing the portfolio construction problem in specific nongaussian frameworks. For instance, in [39] return distributions are modelled by power-laws with the same index tail and a statistical methodology based on the scale parameters is proposed for determining an optimal portfolio regarding to the probability of large losses. In [136] gaussian copulas combined with a family of modified Weibull distributions are used for modelling the tail of the flow of returns and obtaining as a byproduct the tail behaviour of the return of any portfolio. And in [42], the problem of how to allocate assets for minimizing particular quantile based measures of risk is considered using the tools of univariate extreme values theory for modelling the tail of the portfolio (see also [119]).

In [61] the problem of selecting a portfolio so as to provide against extreme loss is tackled by introducing a specific modelling of extremal lower fluctuations of asset returns based on *Independent Component Analysis* (*ICA*). As will be shown below, this permits to quantify the risk of

extreme loss of any investment strategy and to determine how diversification should be carried for minimizing this particular extreme risk measure both at the same time.

0.24.1 On measuring extreme risks of portfolio strategies

Risk quantification for financial strategies have been the object of intense research, still developping. Given the nongaussian character of financial returns distributions and in consequence the limitation of the variance as an indicator for describing the amount of uncertainty in their fluctuations, various risk measures have been proposed (see [189] for a recent survey on this subject), such as Value at Risk (*VaR*, see [77] for instance) or Expected Shortfall (*ES*, see [2]), which are both quantile-based risk measures. Risk measures may be considered in particular for guiding investment behaviour. Once a risk measure is chosen, the goal is to select an optimal portfolio with respect to this latter. Suppose that there are D (risky) assets available, indexed by $i \in \{1, ..., D\}$. Let a certain (discrete) time scale for observing the price fluctuations be fixed and denote by $X_i(t)$ the price of the i-th asset at time t. Let $r_i(t) = (X_i(t) - X_i(t-1))/X_i(t-1)$ be the return at time $t \geq 1$ and denote by $r(t) = (r_1(t), ..., r_D(t))'$ the flow of returns. Consider the portfolio strategy consisting in investing a fixed relative amount $w_i \in [0, 1]$ of the capital in the i-th asset (short sales being excluded), so that $\sum_{i=1}^{D} w_i = 1$ (the portfolio is fully invested). The return of the corresponding portfolio at time t is then

$$R_w(t) = \sum_{i=1}^{D} w_i r_i(t). \tag{66}$$

Hence, if the $r(t)$'s are assumed to be i.i.d., so are the $R_w(t)$'s, with common distribution function F_w. Although in the case when one does not consider investment strategies involving short sales $R_w(t)$ is bounded below by -1 (like the $r_i(t)$'s), here we classically use an infinite lower tail approximation for modelling extreme lower values of portfolio returns (*i.e.* the left tail of the df F_w). Any risk measure is classically defined as a functional of the portfolio return distribution F_w. [61] focused on the maximal relative loss of the portfolio over a large period of time T is, $m_T = \min_{t=1,...,T} R_w(t)$, of which fluctuations may be characterized by an asymptotic extreme value distribution H in some cases, *i.e.* in the cases when F_w is in the domain of attraction of an extreme value distribution H with respect to its lower tail. It is a well-known result in extreme values theory that there are only three types of possible limit distributions for the minimum of i.i.d. r.v.'s under positive affine transformations, depending on the tail behavior of their common density (refer to [164] for further details on extreme values theory). It is pertinent to consider investment strategies w with distribution functions F_w in the maximum domain of attraction of Fréchet distribution functions Φ_α, $\alpha > 0$ ($MDA(\Phi_\alpha)$ in abbreviated form), since such dfs form the prime examples for modelling heavy-tailed phenomena (see Chap. 6 in [80] for instance). Recall that $\Phi_\alpha(x) = \exp(-x^{-\alpha})\mathbb{I}_{x>0}$ and that a df $F \in MDA(\Phi_\alpha)$ iff $F(-x)$ is regularly varying with index $-\alpha$, that is $F(-x) = x^{-\alpha}L(x)$, for some measurable function L slowly varying at ∞ (i.e. for some function L such that $L(tx)/L(x) \to 1$ as $x \to \infty$ for all $t > 0$). In the case when the df F behaves as a power law distribution at $-\infty$, the *tail index* α characterizes the extreme lower behaviour of an i.i.d. sequence $(R(t))_{t\geqslant 0}$ drawn from F regarding to its minimum value:

$$\mathbb{P}(c_T^{-1}(\min_{t=1,...,T} R(t)) \leqslant -x) \to 1 - \Phi_\alpha(x),$$

as $T \to \infty$ for any $x > 0$, where $c_T = \sup\{y \in \mathbb{R} : F(-y) \leqslant n^{-1}\}$.

As shown in several empirical studies, this class of dfs F contains left heavy-tailed distributions, usually called *power law-like* dfs, that may be appropriate for modelling large lower fluctuations of returns. Let us observe that the smaller the tail index α is, the heavier the left tail of F is. Hence, when modelling the lower tail behaviour of the distribution of portfolio returns this way, the tail

index α may appear as a legitimate measure of extreme risk for the portfolio strategy (refer to [116] for a discussion about the relevance of this specific safety first criterion, when managing the downside risk of portfolios is the matter). Moreover, the tail index α rules the asymptotic behavior of the *excess-of-loss df* $F^{(u)}(x) = \mathbb{P}(R < -u - x \mid R < -u)$ below the threshold $-u$ related to $F(x) = \mathbb{P}(R \leqslant x) \in MDA(\Phi_\alpha)$ for large $u > 0$, as shown by the following limit distributional approximation (see [99] for details on the convergence of normalized excesses over large thresholds to *Generalised Pareto Distributions*): for $1 + x/\alpha > 0$,

$$\lim_{u\to\infty} F^{(u)}(xa(u)) = (1 + \frac{x}{\alpha})^{-\alpha},$$

where $a(u)$ is a measurable positive function such that $a(u)/u \to \alpha^{-1}$ as $u \to \infty$. The set of strategies w of which returns have dfs F_w in $MDA(\Phi_\alpha)$ is thus naturally equipped with a complete preference relation, as follows.

Definition 41 *Suppose that w_1 and w_2 are two portfolio strategies s.t. $F_{w_i} \in MDA(\Phi_{\alpha_i})$ with $\alpha_i > 0$ for $i = 1, 2$. We shall say that strategy w_1 is riskier (respectively, strictly riskier) than strategy w_2 regarding to extreme relative loss iff $\alpha_1 \leqslant \alpha_2$ (resp., $\alpha_1 < \alpha_2$).*

0.24.2 The Heavy-Tailed ICA Model

Multivariate data are often viewed as multiple indirect measurement of underlying factors or components, which cannot be directly observed. In some cases, one may hope that a few of these latent factors are responsible for the essential structure we see in the observed data and correspond to interpretable causes. *Latent Variables Analysis* aims to provide tractable theoretical framework and develop statistical methods for identifying these underlying components. The general setting of latent variables modelling stipulates that

$$X = AY, \tag{67}$$

where X is an observable D-dimensional r.v. and Y is a vector of d unobserved latent variables converted to X by the linear transform A, classically called the *mixing matrix.* When $d \leq D$ and the mixing matrix is of full rank d, there is no loss (but some redundancy on the contrary, when $d < D$) in the information carried by X. By inverting, one may write

$$Y = \Omega X, \tag{68}$$

where the *de-mixing matrix* Ω is any generalized pseudo-inverse of A: whereas in the case $d = D$, Ω is uniquely determined by (68) and is simply the inverse A^{-1} of the mixing matrix, additional identifiability constraints are necessary for guaranteeing unicity when $d < D$. *Principal Component Analysis* (*PCA*) and *Factor Analysis* form part of a family of statistical techniques for recovering Y and A on the basis of an observed sample drawn from X, that are typically designed for normal distributions, which assumption clearly limited their practical application. In spite of the considerable evidence for non-gaussianity of financial returns or log-returns, and on the basis of theoretical market modelling such as CAPM or APT, Factor Analysis and PCA have been nevertheless extensively used by practicioners for gaining insight to the explanatory structure in observed returns. In the mid-90's, the methodology of *Independent Component Analysis* (refer to [117] for a comprehensive presentation of ICA), comprising highly successfull new algorithms (mainly introduced in the field of signal processing for *Blind Source Separation* (BSS)) has emerged as a serious competitor to PCA and Factor Analysis, and is based, on the contrary to these latter techniques, on the non-normal nature of the latent components. Indeed, ICA only relies on the assumption that the underlying factors Y_1, $Y_2, \ldots$ are statistically independent (and are naturally

called *independent components* for this reason), which hypothesis is much stronger than uncorrelation when Y is nongaussian (heuristically, uncorrelation determines second-order cross moments, whereas statistical independence determines all of the cross-moments). Stipulating further identifiability conditions apart from independence and non-normality of the IC's, many valid statistical methods have been proposed for estimating the ICA model, based on entropy, maximum likelihood, mutual information or tensorial methods. The nongaussian character of financial returns or log-returns being now carried unanimously, several contributions to the application of ICA to finance have been recently made. The modelling of the fluctuations of financial returns and the search for independent factors through ICA thus gave rise to several works, among which [15], [124] and [51].

Modelling extremal events via ICA The specific ICA model proposed in [61] for describing the extreme lower fluctuations of asset returns, which they called the *heavy-tailed ICA model*. Suppose that there are D securities indexed by $i \in \{1, ..., D\}$ and let $X_i(t)$ be the price of the i-th security at time t. The returns of the i-th security are defined by

$$r_i(t) = (X_i(t) - X_i(t-1))/X_i(t-1), \ t \geqslant 1. \tag{69}$$

Suppose further that the flow of daily returns of the D assets are i.i.d. realizations of a r.v. $r = (r_1, ..., r_D)'$ with components that are linear combinations of D independent *elementary portfolios* returns $R_1, ..., R_D$, so that

$$r = AR, \tag{70}$$

where $R = (R_1, ..., R_D)'$ and $A = (a_{ij})$ is a D by D matrix of full rank, of which inverse Ω belongs to the (compact and convex) set of parameters

$$\mathcal{B} = \{\Omega = (\omega_{ij})_{1 \leqslant i \leqslant, j \leqslant D} / \ \omega_{ij} \geqslant 0, \sum_{k=1}^{D} \omega_{ik} = 1 \text{ for } 1 \leqslant i, \ j \leqslant D\}. \tag{71}$$

We thus have

$$R = \Omega r. \tag{72}$$

The R_i's are assumed to have heavy-tailed distributions, furthermore lower-tails are supposed to be Pareto-like below some (unknown) thresholds:

$$G_i(y) = \mathbb{P}(R_i < -y) = C_i y^{-\alpha_i}, \ \text{for } y \geqslant s_i, \tag{73}$$

with strictly positive constants α_i, C_i and s_i, $1 \leqslant i \leqslant D$. In addition, the α_i's are supposed to be distinct and, with no loss of generality, the IC's are indexed so that $\alpha_1 > ... > \alpha_D$ (the elementary portfolios are thus sorted by increasing order of their riskiness regarding to Definition 41), so as to ensure that the statistical model is identifiable.

In this framework, the next result shows that an optimal strategy with respect to extreme relative loss is straightforwardly available.

Theorem 42 *(Clémençon & Slim, 2006) Under the assumptions above, the elementary portfolio strategy $\omega_1 = (\omega_{11}, ..., \omega_{1D})$ is optimal with respect to the extreme risk measure.*

Semi-parametric inference Several objects must be estimated, the tail indexes $\alpha_1, ..., \alpha_D$, the constants $C_1, ..., C_D$, as well as the matrix Ω of elementary strategies. A feasible method for estimating the semi-parametric ICA model described above is based on conditional MLE, as the classical Hill inference procedure for tail index estimation (refer to [110]). Let us first introduce

some additional notation. For $i \in \{1, ..., D\}$, denote by $R_i(1), ..., R_i(N)$ an i.i.d. sample drawn from R_i and by $R_i(\sigma_i(1)) \leqslant ... \leqslant R_i(\sigma_i(N))$ the corresponding order statistics. Recall that the basic *Hill estimator* for the tail index α_i based on this sample is:

$$\widehat{\alpha}_{i,k}^{H} = (\frac{1}{k}\sum_{l=1}^{k} \ln(\frac{R_i(\sigma_i(l))}{R_i(\sigma_i(k))}))^{-1},$$

with $1 \leqslant k \leqslant N$ such that $R_i(\sigma_i(k)) < 0$, while C_i is estimated by $\widehat{C}_i = \frac{k}{N}(-R_i(\sigma_i(k)))^{\widehat{\alpha}_i}$. These estimates are (weakly) consistent, as soon as $k = k(N)$ is picked such that $k(N) \to \infty$ and $N/k(N) \to \infty$ as $N \to \infty$ (*cf* [144]) and are strongly consistent if furthermore $k(N)/\ln\ln N \to \infty$ as $N \to \infty$ (see [71]). They are classically interpreted as a conditional maximum likelihood estimators based on maximization of the joint density $f_{i,k}(y_1, ..., y_k)$ of $(-R_i(\sigma_i(1)), ..., -R_i(\sigma_i(k)))$ conditioned on the event $\{R_i(\sigma_i(k)) \leqslant -s_i\}$:

$$f_{i,k}(y_1, ..., y_k) = \frac{N!}{(N-k)!}(1 - C_i y_k^{-\alpha_i})^{N-k} C_i^k \alpha_i^k \prod_{l=1}^{k} y_l^{-(\alpha_i+1)},$$

for $0 < s_i \leqslant y_1 \leqslant ... \leqslant y_k$ and $f_{i,k}(y_1, ..., y_k) = 0$ otherwise. Hence the conditional likelihood based on R is

$$\prod_{i=1}^{D} f_{i,k}(-R_i(\sigma_i(1)), ..., -R_i(\sigma_i(k))).$$

One may now derive the conditional likelihood of the model from the observation of a sample of length N of asset returns $r_{(N)} = (r(1), ..., r(N)) = ((r_i(1))_{1\leqslant i\leqslant D}, ..., (r_i(N))_{1\leqslant i\leqslant D})$. For all $1 \leqslant i \leqslant D$, sort the return vector observations $r(l)$, $1 \leqslant l \leqslant N$, so that $\omega_i r(\sigma_i(1)) \leqslant ... \leqslant \omega_i r(\sigma_i(N))$ (observe that the permutation σ_i depends on ω_i: $\omega_i r(\sigma_i(l)) = R(\sigma_i(l))$, $1 \leqslant l \leqslant N$). Hence, the likelihood function based on the observations $\{r(\sigma_i(l)), 1 \leqslant l \leqslant k, 1 \leqslant i \leqslant D\}$ and conditioned on the event $\{\omega_i r(\sigma_i(k)) \leqslant -s_i, 1 \leqslant i \leqslant D\}$ is

$$L_k(r_{(N)}, \Omega, \alpha, C) = |\det \Omega|^k \prod_{i=1}^{D} f_{i,k}(-\omega_i r(\sigma_i(1)), ..., -\omega_i r(\sigma_i(k))).$$

For any given $r_{(N)}$, the functional $L_k(r_{(N)}, . \ , . \ , .)$ is continuous and piecewise differentiable on $\mathcal{B} \times \mathbb{R}_+^* \times \mathbb{R}_+^*$. Furthermore, as previously recalled, for any fixed $\Omega \in \mathcal{B}$ and for all $i \in \{1, ..., D\}$, $f_{i,k}(-\omega_i r(\sigma_i(1)), ..., -\omega_i r(\sigma_i(k))$ is maximum for $\alpha_i = \widehat{\alpha}_i$ and $C_i = \widehat{C}_i$ with

$$\widehat{\alpha}_i = (\frac{1}{k}\sum_{l=1}^{k} \ln(\frac{\omega_i r(\sigma_i(l))}{\omega_i r(\sigma_i(k))}))^{-1} \text{ and } \widehat{C}_i = \frac{k}{N}(-\omega_i r(\sigma_i(k)))^{\widehat{\alpha}_i}.$$

For any $\Omega \in \mathcal{B}$, $L(r^{(N)}, \Omega, \alpha, C)$ is thus maximum for $\alpha = \widehat{\alpha} = (\widehat{\alpha}_i)_{1\leqslant i\leqslant D}$ and $C = \widehat{C} = (\widehat{C}_i)_{1\leqslant i\leqslant D}$ and we denote this maximal value by $\widetilde{L}_k(\Omega) = L_k(r^{(N)}, \ \Omega, \ \widehat{\alpha}, \ \widehat{C})$. Here, conditional MLE reduces then to maximizing the multivariate scalar function $\widetilde{L}_k(\Omega)$ over $\Omega \in \mathcal{B}$, which may be easily shown as equivalent to maximizing over $\Omega \in \mathcal{B}$:

$$l_k(\Omega) = |\det \Omega|^k \exp\left(-\sum_{i=1}^{D}\{k\ln(\sum_{l=1}^{k}\ln(\frac{\omega_i r(\sigma_i(l))}{\omega_i r(\sigma_i(k))})) + \sum_{l=1}^{k}\ln(-\omega_i r(\sigma_i(l))\}\right).$$

In our setting, estimating the ICA model (70) thus boils down to the task of finding $\widehat{\Omega}$ in the closed convex set $\mathcal{B}$ such that $l_k(\widehat{\Omega}) = \max_{\Omega\in\mathcal{B}} l_k(\Omega)$. In addition to the theoretical estimation principle described above, a numerical method for maximizing the objective function $l_k(\Omega)$ (or

$\widetilde{L}_k(\Omega)$) subject to the linear matrix constraint $\Omega \in \mathcal{B}$ is required. Various optimization methods, among which the popular *subgradient-type learning algorithms,* have been introduced for practically solving such a constrained optimization problem approximatively. The objective function $l_k(\Omega)$ is continuous and piecewise differentiable on $\mathcal{M}_D(\mathbb{R})$: its gradient $\nabla l_k(\Omega)$ is well-defined at each point $\Omega = (\omega_{i,j})_{1 \leqslant i,j \leqslant D}$ such that $\det \Omega \neq 0$ and $\omega_i r(\sigma_i(k-1)) < \omega_i r(\sigma_i(k)) < \omega_i r(\sigma_i(k+1))$ for all $i \in \{1, ..., D\}$, we have

$$\frac{\partial l_k}{\partial \omega_{i,j}}(\Omega)/l_k(\Omega) = k\gamma_{i,j}(\Omega)/\det \Omega + \sum_{l=1}^{k} r_j(\sigma_i(l))/\omega_i r(\sigma_i(l)) - k\frac{\sum_{l=1}^{k}\{r_j(\sigma_i(l))/\omega_i r(\sigma_i(l)) - r_j(\sigma_i(k))/\omega_i r(\sigma_i(k))\}}{\sum_{l=1}^{k} \ln(\omega_i r(\sigma_i(l))/\omega_i r(\sigma_i(k)))}, \tag{74}$$

where $\gamma_{i,j}(\Omega)$ denotes the cofactor of $\omega_{i,j}$ in Ω, $1 \leqslant i, j \leqslant D$ (the subdifferential $\partial l_k(\Omega)$ is then easily determined through equation (74) at any point $\Omega \in \mathcal{M}_D(\mathbb{R})$). In the applications mentioned below, the classical *projected subgradient method* is used for estimating the Heavy-tailed ICA model and selecting an optimal portfolio knowing the sample $r_{(N)}$.

0.24.3 Some empirical results

As explained above, the statistical method proposed in [61] aims at searching for independent portfolio strategies and at estimating their Pareto left tail indexes as well. It also permits to recover as a byproduct a portfolio strategy with a maximum left tail index).

As an illustration of the Heavy-tailed ICA model, here the latter is applied for analyzing the daily return series of $D = 11$ international equity indexes of (developed or developing) financial markets over the period running from 02-January-1987 to 22-October-2002 listed in Table 0.24.3. The results indicate the allocations corresponding to the 11 *independent elementary portfolios*, as well as their left tail index estimate, sorted by increasing order of their extreme risk measure, obtained by implementing the statistical procedure described in §7.2.2 with the $k = 200$ lowest values (representing roughly the lowest 5% values). Descriptive statistics related to the lower tail behaviour of each elementary portfolio are also displayed in Table 0.24.3: minimum return values, empirical estimates of the probability of excess (*EPE* in short), $\mathbb{P}(R_i < -u)$, that the i-th elementary portfolio looses more than u% of its value (at a one day horizon) are calculated at various threshold levels u over the time period considered, as well as the empirical counterpart of the mean excess function (*ME* in abbreviated form), $e_i(u) = -\mathbb{E}(R_i + u \mid R_i < -u)$, traditionally referred to as the *expected shortfall* in the financial risk management context. In a general fashion, the lower tails of the single assets are globally much heavier than the ones of the least risky elementary portfolios we obtained. For instance, the maximum relative loss suffered by the optimal elementary portfolio (PF1) over the period of interest is 3.86% , while the minimum values of single market indexes range from -10.29% to -40.54%. As expected, except for the FTSE100 index, zero or small weights in PF1 correspond to the market indexes with the most heavier left tails (namely, the Hong Kong, Malaisia and Singapore indexes, which correspond to emerging financial markets that are presumably very interdependent, whereas on the contrary the latter indexes have the largest weights for PF11. These results clearly show the benefit of the diversification induced by our specific modelling of the dependence structure between the assets regarding to extreme risk. This phenomenon is also illustrated by Table 0.24.3. It shows lower tail statistics of the optimal portfolio obtained by applying our modelling as a function of the number D of market indexes involved in the ICA model (financial indexes being progressively added, by decreasing order of their tail index estimates) and plainly indicates that the lower tail becomes thinner as D grows.

Table 1: Lower tail characteristics of the optimal portfolio obtained by using the Heavy-tailed ICA model with D market indexes, as D grows.

Number of assets D	3	5	7	9
Pareto Index	2.57	2.90	3.04	3.30
%Min	-36.54	-14.60	-7.25	-6.71
EME at u = 1	1.27	0.92	0.58	0.56
EME at u = 2	1.53	1.09	0.73	0.88
EME at u = 3	1.98	1.39	1.39	1.19

Table 2: Weights of the elementary portfolios are given under columns in percentages, together with their tail index estimates, the minimum return values over the time period considered, the standard deviation, and the EPE and EME computed at levels 1%, 2% and 3%.

	PF1	PF2	PF3	PF4	PF5	PF6	PF7	PF8	PF9	PF10	PF11
Markets						Weights					
Canada	7.57	20.64	0.00	4.66	21.58	19.35	23.92	12.88	0.00	6.34	0.00
Chile	24.44	0.00	13.61	8.61	9.20	0.00	20.47	0.00	7.45	7.42	12.52
Germany	17.12	22.96	3.89	28.28	3.53	0.00	0.00	14.43	13.20	16.12	0.00
HongKong	0.00	18.09	16.22	0.00	0.00	0.00	3.78	6.37	0.00	0.00	19.40
Korea	1.92	0.12	12.46	17.43	0.00	21.78	0.00	0.00	0.00	6.54	3.64
Japan	13.43	16.82	9.01	0.00	12.27	8.21	2.76	0.00	16.97	33.01	0.94
Malaisia	5.87	6.89	5.48	8.57	0.00	11.68	20.42	22.83	4.78	19.49	18.68
Singapore	0.00	3.26	0.00	18.77	22.36	9.46	7.38	0.00	8.14	11.08	24.21
Taiwan	11.54	0.00	21.67	0.00	18.28	3.81	0.00	23.35	2.79	0.00	0.78
U.S.	18.11	9.56	4.68	5.54	0.00	23.17	0.00	17.63	8.18	0.00	19.83
U.K.	0.00	1.66	12.98	8.14	12.78	2.54	21.27	2.51	38.49	0.00	0.00
Tail Index	3.93	3.82	3.64	3.46	3.39	3.35	3.27	3.19	3.03	2.92	2.84
Min (%)	-3.86	-5.89	-8.34	-6.13	-6.35	-9.80	-9.38	-8.47	-7.93	12.31	-9.72
Std (%)	0.58	0.67	0.80	0.75	0.72	0.68	0.66	0.74	0.65	0.78	0.80
						EME(%)					
u=1%	0.44	0.58	0.58	0.54	0.62	0.58	0.58	0.57	0.59	0.60	0.71
u=2%	0.73	0.75	0.77	0.70	0.79	1.06	0.88	0.82	1.12	0.87	1.08
u=3%	0.46	1.13	1.38	1.05	0.86	1.51	2.07	1.30	1.91	2.03	1.49
						EPE(%)					
u=1%	0.083	0.105	0.177	0.157	0.139	0.112	0.101	0.138	0.100	0.157	0.154
u=2%	0.008	0.017	0.026	0.021	0.022	0.015	0.015	0.021	0.015	0.025	0.031
u=3%	0.034	0.004	0.005	0.048	0.005	0.053	0.003	0.005	0.004	0.005	0.010

Applications in Biosciences

0.25 Stochastic Toxicologic Models for Dietary Risk Analysis

Certain foods may contain varying amounts of chemicals such as methyl mercury (present in sea food), dioxins (in poultry, meat) or mycotoxins (in cereals, dried fruits, etc.), which may cause major health problems when accumulating inside the body in excessive doses. Food safety is now a crucial stake as regards public health in many countries. This topic naturally interfaces with various disciplines, such as biology, nutritional medicine, toxicology and of course applied mathematics with the aim to develop rigorous methods for quantitative risk assessment. A scientific literature devoted to probabilistic and statistical methods for the study of dietary exposure to food contaminants is progressively carving out a place in applied probability and statistics journals (see [195] or [29] for instance).
Static viewpoints for the probabilistic modeling of the quantity X of a given food contaminant ingested on a short period have been considered in recent works, mainly focussing on the tail behavior of X and allowing for computation of the probability that X exceeds a maximum tolerable dose (see [194]). However, such approaches for food risk analysis do not take into account the accumulating and eliminating processes occurring in the body, which naturally requires to introduce time as a crucial description parameter of a comprehensive model.

0.26 Modeling the exposure to a food contaminant

Dietary behavior Suppose that an exhaustive list of P types of food, indexed by $p = 1, \cdots, P$, involved in the alimentation of a given population and possibly contaminated by a certain chemical, is drawn up. Each type of food $p \in \{1, \cdots, P\}$ is contaminated in random ratio $K^{(p)}$, with probability distribution $F_{\mathcal{K}^{(p)}}$, regarding to the chemical of interest. Concerning this specific contaminant exposure, a meal may be viewed as a realization of a r.v. $Q = (Q^{(1)}, \cdots, Q^{(P)})$ indicating the quantity of food of each type consumed, renormalized by the body weight. For a meal Q drawn from a distribution $F_{\mathcal{Q}}$ on $(\mathbb{R}_+^P, \mathcal{B}_{\mathbb{R}_+^P})$, cooked from foods of which toxicity is described by a contamination ratio vector $K = (K^{(1)}, \cdots, K^{(P)})$ drawn from $F_{\mathcal{K}} = \otimes_{p=1}^P F_{\mathcal{K}^{(p)}}$, the global contaminant intake is $U = \langle K, Q \rangle$, denoting by $\langle ., . \rangle$ the standard inner product on $\mathbb{R}^P$. Its probability distribution $F_{\mathcal{U}}$ is the image of $F_{\mathcal{K}} \otimes F_{\mathcal{Q}}$ by the inner product $\langle ., . \rangle$, assuming that the quantities of food consumed are independent from the contamination levels. Here and throughout, we suppose that the contaminant intake distribution F_U has a density f_U with respect to λ, the Lebesgue measure on $\mathbb{R}_+$.

The food contamination phenomenon through time for an individual of the population of interest may be classically modeled by a marked point process $\{(T_n, Q_n, K_n)\}_{n \geq 1}$ on $\mathbb{R}_+ \times \mathbb{R}_+^P \times \mathbb{R}_+^P$, the T_n's being the successive times at which the individual consumes foods among the list $\{1, \cdots, P\}$ ($T_0 = 0$ being chosen as time origin) and the marks (Q_n, K_n) being respectively the vector of food quantities and the vector of contamination ratios related to the meal had at time T_n. The process $\{(T_n, Q_n)\}_{n \geq 1}$ describing dietary behavior is assumed independent from the sequence $(K_n)_{n \geq 1}$ of chemical contamination ratios. Although the modeling of dietary behaviors could certainly give

rise to a huge variety of models, depending on the dependence structure between (T_n, Q_n) and past values $\{(T_m, Q_m)\}_{m<n}$ that one stipulates, we make here the simplifying assumption that the marks Q_n, $n \geq 1$, form an i.i.d. sequence with common distribution F_Q, independent from the location times $(T_n)_{n\geq 1}$. This assumption is acceptable for chemicals present in a few types of food, such as methyl mercury, but certainly not for all contaminants. For chemicals present in many foods of everyday consumption, it would be necessary to introduce additional autoregressive structure in the model for capturing important features of any realistic diet (the consumption of certain food being typically alternated for reasons related to taste or nutritional aspects). Such a modeling task is left for further investigation. Finally, suppose that the inter-intake times $\Delta T_{n+1} = T_{n+1} - T_n$, $n \geq 1$, form a sequence of i.i.d. r.v.'s with common probability distribution $G(dt) = g(t)dt$ and finite expectation $m_G = \int_{t=0}^{\infty} tG(dt) < \infty$, the sequence $(T_n)_{n\geq 1}$ of intake times being thus a pure renewal process.

Toxicokinetics Contamination sources other than dietary are neglected in the present study and denote by $X(t)$ the total body burden in contaminant at time $t \geq 0$. In between intakes, assume that the *contamination exposure process* $X(t)$ is governed by the differential equation

$$\dot{x}(t) = -r(x(t), \theta), \tag{75}$$

θ being a random parameter, taking its values in a set $\Theta \subset \mathbb{R}^d$ with $d \geq 1$ say, and accounting in the modeling for fluctuations of the (content dependent) elimination rate due to metabolic factors at the intake times (the successive values θ_n, $n \in \mathbb{N}$, of θ are thus fixed at times T_0, T_1, $\cdots$). And the function $r(x, \theta)$ is assumed to be strictly positive and continuous on $\mathbb{R}_+^* \times \Theta$, such that for all $\theta \in \Theta$, $r(0, \theta) = 0$ (so that when $X(t)$ eventually reaches the level 0, the process stays at this level until the next intake) and for all $(\epsilon, \theta) \in (0, 1) \times \Theta$:

$$\inf_{\epsilon<x<\epsilon^{-1}} r(x, \theta) > 0 \text{ and } \sup_{0<x<\epsilon^{-1}} r(x, \theta) < \infty. \tag{76}$$

Under these conditions, for any initial value $x(0) \geq 0$ and metabolic parameter value $\theta \in \Theta$, Eq. (75) has clearly a unique solution.

Remark 4 In toxicology, Eq. (75) is widely used with $r(x, \theta) = \theta x$ for modelling the kinetics in man of certain contaminants following intakes. As shown by many pharmacokinetics studies, there is considerable empirical evidence that it properly describes the way the elimination rate depends on the total body burden of the chemical in numerous cases. In this context, the release parameter $\log 2/\theta$ is known as the *half-life* of the contaminant in the body (the time required for X to decrease by half in absence of new contaminant intake).

We assume that $(\theta_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence with common distribution $H(d\theta)$. For a given value of the metabolic parameter $\theta \in \Theta$, the time necessary for the body burden (without further intake) to decrease from $x_0 > 0$ to $x \in (0, x_0)$ is given by

$$\tau_\theta(x_0, x) = \int_x^{x_0} \frac{1}{r(y, \theta)} dy.$$

Under the assumptions stated above, we clearly have that $H(\{\tau_\theta(x_0, x) < \infty\}) = 1$ for all $0 < x \leq x_0$. The contaminant may be thus entirely eliminated from the body (the amount x reaching then the level 0) with probability one in the sole case when the following condition holds.

Condition ($\mathcal{C}_1$): $H(\{\tau_\theta(x_0, 0) < \infty\}) = 1$ for some $x_0 > 0$.

In such a case, $H(\{\tau_\theta(x, 0) < \infty\}) = 1$ for all $x \geq 0$. In this respect, it is noteworthy that, in the linear case mentioned in Remark 2, $\tau_\theta(x, 0) = \infty$ for all $\theta > 0$ and $x > 0$.

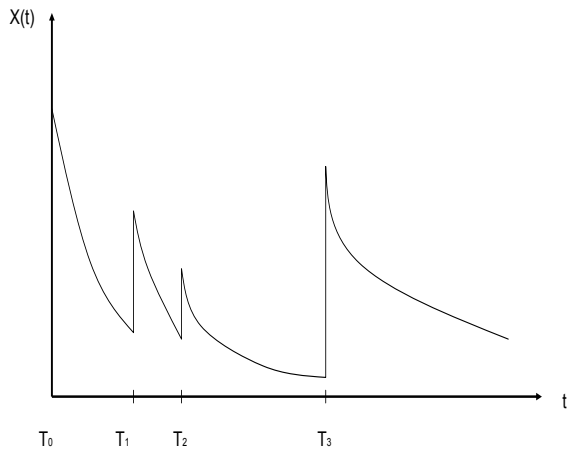

Figure 15: Sample path of the exposure process X.

Hence, in between intake times and given the current value of the metabolic parameter θ, the process moves in a deterministic fashion according to (75), and has the same (upward) jumps as the process of cumulative intakes

$$B(t) = \sum_{n=1}^{N(t)} U_n, \tag{77}$$

with $U_n = \langle K_n, Q_n \rangle$, $n \in \mathbb{N}$, and denoting by $N(t) = \sum_{n\in\mathbb{N}} \mathbb{I}\{T_n \leq t\}$ the number of intakes until time t. The process X is piecewise-deterministic with càd-làg trajectories (see a typical sample path in Fig. 15) and satisfies the equation

$$X(t) = X(0) + B(t) - \sum_{n=1}^{N(t)+1} \int_{T_{n-1}}^{T_n \wedge t} r(X(s), \theta_n) ds, \tag{78}$$

$X(0)$ denoting the total body burden in contaminant at initial time $T_0 = 0$. For an account of such piecewise deterministic processes, one may refer to [68] (see also [69] and some ergodic results may be found in [64]).

For the continuous-time process thus defined to be markovian, one has to record the current value $\theta(t) = \sum_{n\in\mathbb{N}} \theta_n \mathbb{I}_{\{t\in[T_n, T_{n+1}[\}}$ of the metabolic parameter as well as the backward recurrence time $A(t) = t - T_{N(t)}$ (the time since the last intake). By construction, the process $(X(t), \theta(t), A(t))_{t\geq 0}$ is strongly Markovian with generator

$$\mathcal{G}\phi(x,\theta,t) = \zeta(t) \int_{u=0}^{\infty} \int_{\theta'\in\Theta} \{\phi(x+u, \theta', 0) - \phi(x,\theta,t)\} F_U(du) H(d\theta') \\ - r(x,\theta)\partial_x \phi(x,\theta,t) + \partial_t \phi(x,\theta,t), \tag{79}$$

denoting by $\zeta(t) = g(t)/\int_{s=t}^{\infty} g(s)ds$ the hazard rate of the inter-intake times and provided that $\phi(.,\theta,.) : (x,t) \mapsto \phi(x,\theta,t)$ is a bounded function with bounded continuous first derivatives in x and t for all $\theta \in \Theta$.

In the above setting, the time origin $T_0 = 0$ does not necessarily correspond to an intake time. Given the time $A(0) = a$ since the last intake at time $t = 0$, we let ΔT_1 have the density

$g_a(t) = g(a+t)/\int_{s=a}^{\infty} g(s)ds$, making the renewal process $(\Delta T_n)_{n\in\mathbb{N}}$ possibly delayed, except naturally in the case when the inter-intake distribution G is exponential. However, the choice of such a memoryless distribution in the dietary context is clearly not pertinent, distributions with increasing hazard rate being more adequate. Here and throughout we denote by $\mathbb{P}_{x,a}$ the probability measure on the underlying space such that $(X(0), A(0)) = (x, a)$ and $\theta(0) \sim H$, and by $\mathbb{E}_{x,a}(.)$ the $\mathbb{P}_{x,a}$-expectation for all $x \geq 0$ and $a \in \text{supp}(G)$.

In the case when one neglects variability in the elimination process, this modeling boils down to a standard *storage model with a general release rate* (see [45] for instance). Basic communication and stochastic stability properties of the stochastic process $X = (X(t))_{t\geq 0}$ may be established in a fashion very similar to the ones of the latter processes, although the additional assumption that the renewal times are exponentially distributed is usually required in these studies, making the process X itself Markovian (which facilitates much the study but is not relevant to the present application as emphasized above).

Theorem 43 *(Bertail, Clémençon & Tressou, 2006a) Suppose that $G(dx) = g(x)dx$ has infinite tail. Assume further that either $g(x) > 0$ on $]0, \epsilon]$ for some $\epsilon > 0$ or else that $F_{\mathcal{U}}$ has infinite tail. Then X reaches any state $x > 0$ in finite time with positive probability whatever the starting point, i.e. for all $x_0 \geq 0$, $a \in \text{supp}(G)$, it holds*

$$\mathbb{P}_{x_0,a}(\tau_x < \infty) > 0, \tag{80}$$

with $\tau_x = \inf\{t \geq 0 : X_t = x\}$. Furthermore, if condition (C_1) is fulfilled, then (80) still holds for $x = 0$.
Besides, either X "heads to infinity" with probability one, i.e. is such that $\mathbb{P}_{x_0,a}(\{X(t) \to \infty$, as $t \to \infty\}) = 1$ for all $x_0 \geq 0$, or else X reaches any state $x > 0$ in finite time with probability one whatever the starting point, i.e. for all $x_0 \geq 0$, $a \in \text{supp}(G)$,

$$\mathbb{P}_{x_0,a}(\tau_x < \infty\}) = 1. \tag{81}$$

If (C_1) is satisfied, then (81) also holds for $x = 0$.

An important task is to find conditions ensuring that the limiting behavior of the exposure process X is represented by a stationary probability measure μ describing the equilibrium state to which the process settles as time goes to infinity. In particular, time averages over long periods, such as the mean time spent by the exposure process X over a threshold $u > 0$, $T^{-1}\int_0^T \mathbb{I}_{\{X_t \geq u\}}dt$, for instance, are then asymptotically described by the distribution μ. Beyond stochastic stability properties, evaluating the rate at which the process converges to the stationary state is also of critical importance. These questions are now tackled for linear rate models.

0.27 Probabilistic study in the linear rate case

We now focus on ergodicity properties of the exposure process $X(t)$ in the specific case when for a given metabolic state described by a real parameter θ, the elimination rate is proportional to the total body burden in contaminant, *i.e.*

$$r(x, \theta) = \theta x. \tag{82}$$

Here we suppose that Θ is a subset of $\mathbb{R}_+^*$, ensuring that (76) is satisfied. The linear case is of crucial importance in toxicology, insofar as it suitably models the pharmacokinetics behavior in man of numerous chemicals. In this case studying the long-term behavior of X can be reduced

to investigating the properties of the embedded chain $\tilde{X} = (X_n)_{n\geq 1}$ of which values correspond to the ones taken by the exposure process just after intake times : $X_n = X(T_n)$ for all $n \geq 1$. By construction, the chain $\tilde{X}$ satisfies the following autoregressive equation with random coefficients

$$X_{n+1} = e^{-\theta_n \Delta T_{n+1}} X_n + U_{n+1}, \text{ for all } n \geq 1, \tag{83}$$

and has transition probability $\Pi(x, dy) = \pi(x, y)dy$ with transition density

$$\pi(x, y) = \int_{\theta\in\Theta} \int_{t=\frac{1}{\theta}\log(1\vee\frac{x}{y})}^{\infty} f_U(y - xe^{-\theta t})G(dt)H(d\theta), \tag{84}$$

for all $(x, y) \in \mathbb{R}_+^{*2}$, where $a \vee b = \max(a, b)$. Ergodicity of such real-valued Markov chains Y, defined through stochastic recurrence equations of the form $Y_{n+1} = \alpha_n Y_n + \beta_n$, where $\{(\alpha_n, \beta_n)\}_{n\in\mathbb{N}}$ is a sequence of i.i.d. pairs of positive r.v.'s, has been extensively studied in the literature, such models being widely used in financial or insurance mathematics (see section 8.4 in [?] for instance). Specialized to our setting, well known results related to such processes enable to study the embedded chains $\tilde{X}$ is positive recurrent under the assumption that $\log(1 \vee U_1\})$ has finite expectation. Furthermore, the simple autoregressive form of Eq. (83) makes Foster-Lyapunov conditions easily verifiable for such chains, in order to refine their stability analysis. The following assumptions are required in the sequel.

- (H1) $\mathbb{E}[\log(1 \vee U_1)] < \infty$,
- (H2) there exists some $\gamma \geq 1$ such that $\mathbb{E}(U_1^\gamma) < \infty$,
- (H3) the r.v. U_1 is regularly varying with index $\kappa > 0$,
- (H4) there exists $\delta > 0$ such that $\mathbb{E}[exp(\delta\Delta T_2)] < \infty$.

Theorem 44 *(Bertail, Clémençon & Tressou, 2006a) Under the assumptions of Theorem 1 and supposing that* (H_1) *is fulfilled,* $X(t)$ *has an absolutely continuous limiting probability distribution* μ *given by*

$$\mu([u, \infty[) = m_G^{-1} \int_{x=u}^{\infty} \int_{t=0}^{\infty} \int_{\theta\in\Theta} t \wedge \frac{\log(x/u)}{\theta} \tilde{\mu}(dx)G(dt)H(d\theta), \tag{85}$$

in the sense that $T^{-1}\int_0^T \mathbb{I}_{\{X_t\leq u\}}dt \to \mu(]0, u])$, $\mathbb{P}_{x_0,a}$*-a.s., as* $t \to \infty$ *for all* $x_0 \geq 0$ *and* $a \in \text{supp}(G)$. *Furthermore,*

- *if* (H3) *holds and* Θ *is bounded, then* μ *is regularly varying with the same index as* F_U,
- *and if* (H2) *and* (H4) *hold and* G *has finite variance* σ_G^2, *then* μ *has finite moment of order* γ *and for all* $(x, a) \in \mathbb{R}_+^* \times \text{supp}(G)$ *there exist constants* $k \in]0, 1[$, $K_a < \infty$ *such that*

$$\sup_{\psi(z)\leq 1+z^\gamma} |\mathbb{E}_{x,a}[\psi(X_t)] - \mu(\psi)| \leq K_a(1 + x^\gamma)k^t. \tag{86}$$

Remark 5 When the U_n's are heavy-tailed, and under the assumption that the ΔT_n's are exponentially distributed (making $B(t)$ a time-homogeneous Lévy process), the fact that the stationary distribution μ inherits its tail behavior from F_U has been established in [9] for general deterministic release rates. Besides, when assuming G exponential and θ fixed, one may identify the limit distribution μ in some specific cases (see section 8 in [44] or section 2 in Chap. XIV of [8]) using basic level crossing arguments (X being itself markovian in this case). If F_U is also exponential for instance, μ is a Gamma distribution. And furthermore, due to the simple form of the generator in the latter case, one may establish an exponential rate of convergence to μ by standard drift criterion or coupling arguments (see section 5 in [169]).

In order to exhibit connections between the exposure process $X = (X(t))_{t\geq 0}$ and possible negative effects of the chemical on human health, it is appropriate to consider simple characteristics of the process X, easily interpretable from an epidemiology viewpoint. In this respect, the mean exposition over a long time period $T^{-1}\int_{t=0}^{T} X(t)dt$ is one of the most relevant features. Its asymptotic behavior is refined in the next result.

Proposition 45 *(Bertail, Clémençon & Tressou, 2006a) Under the assumptions of Theorem 1 and supposing that* (H2) *is fulfilled for* $\gamma = 1$*, we have for all* $(x_0, a) \in \mathbb{R}_+ \times \mathrm{supp}(G)$

$$\bar{X}_T = \frac{1}{T}\int_{t=0}^{T} X(t)dt \to m_\mu, \mathbb{P}_{x_0,a}\text{-}a.s. \ , \tag{87}$$

as $T \to \infty$ *with* $m_\mu = \int_{x=0}^{\infty} x\mu(dx)$*. Moreover, if* (H2) *is fulfilled with* $\gamma \geq 2$*, then there exists a constant* $0 < \sigma^2 < \infty$ *s.t. for all* $(x_0, a) \in \mathbb{R}_+ \times \mathrm{supp}(G)$ *we have the following convergence in* $\mathbb{P}_{x_0,a}$*-distribution*

$$\sqrt{T}(\bar{X}_T - m_\mu) \Rightarrow \mathcal{N}(0, \sigma^2) \ \ as \ T \to \infty. \tag{88}$$

Remark 6 • The asymptotic variance σ^2 in (88) may be related to the limiting behavior of a certain additive functional of the Markov chain $((X_n, \theta_n, \Delta T_{n+1}))_{n\geq 1}$. In [20] (see also [19]), an estimator of the asymptotic variance of such functionals based on pseudo-renewal properties of the underlying chain (namely, on renewal properties of a Nummelin extension of the chain) has been proposed and a detailed study of its asymptotic properties has been carried out.
• Beyond the asymptotic exposure mean or the asymptotic mean time spent by X above a certain threshold, other summary characteristics of the exposure process could be pertinently considered from an epidemiology viewpoint, among which the asymptotic tail conditional expectation $\mathbb{E}_\mu(X \mid X > u)$, denoting by $\mathbb{E}_\mu(.)$ the expectation w.r.t. μ, after the fashion of risk evaluation in mathematical finance or insurance.

0.28 Simulation-based statistical inference

We now consider the statistical issues one faces when attempting to estimate certain features of linear rate exposure models. The main difficulty lies in the fact that the exposure process X is generally unobservable. Food consumption data (quantities of consumed food and consumption times) related to a single individual over long time periods are scarcely available in practice. And performing measurements at all consumption times so as to record the food contamination levels appears as not easily realizable. Instead, practitioners have at their disposal some massive databases, in which information related to the dietary habits of large population samples over short periods of time is gathered. Besides, some contamination data concerning certain chemicals and types of food are stored in data warehouses and available for statistical purposes. Finally, experiments for assessing models accounting for the pharmacokinetics behavior in man of various chemicals have been carried out. And data permitting to fit values or probability distributions on the parameters of these models are consequently available. Estimation of steady-state or time-dependent features of the law $\mathcal{L}_X$ of the process X given the starting point $(X(0), A(0)) = (x_0, a) \in \mathbb{R}_+ \times \mathrm{supp}(G)$ could thus be based on preliminary computation of consistent estimates $\hat{G}$, $\hat{F}_U$ and $\hat{H}$ of the unknown df's G, F_U and H. Hence, when the quantity of interest $\mathcal{Q}(X)$ is not analytically available from (G, F_U, H), ruling out the possibility of computing plug-in estimates, a feasible method could consist in simulating sample paths starting from (x_0, a) of the approximate process $\hat{X}$ with law $\mathcal{L}_{\hat{X}}$ corresponding to the estimated df's $(\hat{G}, \hat{F}_U, \hat{H})$ and construct estimators based on the trajectories thus obtained. This leads up to investigate the *stability* of the stochastic exposure

model w.r.t. G, F_U and H, and consider the *continuity problem* consisting in evaluating a measure of closeness between $\mathcal{L}_X$ and $\mathcal{L}_{\hat{X}}$ making the mapping $\mathcal{L}_X \mapsto \mathcal{Q}(X)$ continuous for the functional of interest $\mathcal{Q}$ (refer to [160] for an account on this topic). Hence, convergence preservation results may be obtained via the *continuous-mapping approach* as described in [199], where it is applied to establish stochastic-process limits for queuing systems.

Let $0 < T < \infty$. Since the exposure process X has càd-làg sample paths, we use the $\mathcal{M}_2$ topology on the Skorohod's space $D([0,T],\mathbb{R})$ induced by the Hausdorff distance $m_{\mathcal{H}}^{(T)}$ on the space of completed graphs, allowing trajectories to be eventually close even if their jumps do not exactly match (the $\mathcal{J}_2$ topology would be actually sufficient for our purpose, refer to [199] for an account on topological concepts for sets of stochastic processes). In order to evaluate how close the approximating and true laws are, a specific coupling has been introduced in [26] for establishing an upper bound for the L_1-Wasserstein Kantorovich distance between the distributions $\mathcal{L}_{X^{(T)}}$ and $\mathcal{L}_{\hat{X}^{(T)}}$ of $X^{(T)} = (X(t))_{t\in[0,T]}$ and $\hat{X}^{(T)} = (\hat{X}(t))_{t\in[0,T]}$. This metric on the space of probability laws on $D([0,T],\mathbb{R})$ is defined as follows (refer to [160]):

$$W_1^{(T)}(\mathcal{L},\mathcal{L}') = \inf_{\substack{Z'\sim\mathcal{L}' \\ Z\sim\mathcal{L}}} \mathbb{E}[m_{\mathcal{M}_2}^{(T)}(Z',Z)], \tag{89}$$

where the infimum is taken over all pairs (Z',Z) with marginals $\mathcal{L}'$ and $\mathcal{L}$ and $m_{\mathcal{M}_2}^{(T)}(Z',Z) = m_{\mathcal{H}}^{(T)}(\Gamma_{Z'},\Gamma_Z)$, denoting by $\Gamma_{Z'}$ and Γ_Z the completed graphs of Z' and Z. It is well-known that this metric implies weak convergence. The law $\mathcal{L}_{\hat{X}^{(T)}}$ is shown to get closer and closer to $\mathcal{L}_{X^{(T)}}$ as the df's $\hat{G}$, $\hat{F}_U$ and $\hat{H}$ respectively tend to G, F_U and H in the Mallows sense. For $p \in [1,\infty)$, we denote by $M_p(F_1,F_2) = (\int_0^1 \left|F_1^{-1}(t) - F_2^{-1}(t)\right|^p dt)^{1/p}$ the L_p-Mallows distance between two df's F_1 and F_2 on the real line.

The next result now establishes the asymptotic validity of simulation estimators.

Theorem 46 *(Bertail, Clémençon & Tressou, 2006a) Let (G, F_U, H) (resp. $(\hat{G}^{(n)}, \hat{F}_U^{(n)}, \hat{H}^{(n)})$ for $n \in \mathbb{N}$) be a triplet of df's on $\mathbb{R}_+$ defining a linear exposure process X (resp. $\hat{X}_{(n)}$) starting from $x_0 \geq 0$ and fulfilling the assumptions of Theorem 5. Let $0 < T \leq \infty$.*

- *Let $\mathcal{Q}$ be a measurable function mapping $D((0,T),\mathbb{R})$ into some metric space $(\mathcal{S},D)$ with $\mathrm{Disc}(\mathcal{Q})$ as set of discontinuity points and such that $\mathbb{P}(X^{(T)} \in \mathrm{Disc}(\mathcal{Q})) = 0$. If $(\hat{G}^{(n)}, \hat{F}_U^{(n)}, \hat{H}^{(n)}) \to (G, F_U, H)$ in the L_1-Mallows distance, then we have the convergence in distribution*

$$Q(\hat{X}_{(n)}^{(T)}) \Rightarrow Q(X^{(T)}) \text{ in } (\mathcal{S},D). \tag{90}$$

- *Suppose that G (resp., $\hat{G}^{(n)}$) has finite variance σ_G^2 (resp. $\sigma_{\hat{G}^{(n)}}^2$) and H (resp., $\hat{H}^{(n)}$) has finite mean. If $\sigma_{\hat{G}^{(n)}}^2$ remains bounded and $(\hat{G}^{(n)}, \hat{F}_U^{(n)}, \hat{H}^{(n)}) \to (G, F_U, H)$ in the L_1-Mallows distance, then for any Lipschitz function $\phi : (D((0,T),\mathbb{R}), d_{\mathcal{M}}^{(T)}) \to \mathbb{R}$, we have*

$$\mathbb{E}[\phi(\hat{X}_{(n)}^{(T)})] \to \mathbb{E}[\phi(X^{(T)})]. \tag{91}$$

We conclude by giving several examples, illustrating how the results above apply to certain functionals of the exposure process in practice. Among the *time-dependent* and *steady-state* features of the exposure process, the following quantities are of considerable importance to practitioners in the field of risk assessment of chemicals in food and diet.

Mean exposure value. The mapping that assigns to any trajectory $X^{(T)} \in D((0,T),\mathbb{R})$ its mean value $T^{-1}\int_{t=0}^{T} X(t)dt$ is Lipschitz w.r.t the distance $m_{\mathcal{M}_2}^{(T)}$. Hence, given consistent estimates $\widehat{G}^{(n)}$, $\widehat{F}_U^{(n)}$ and $\widehat{H}^{(n)}$ of G, F_U and H, one may construct a consistent estimate of $\mathbb{E}[\int_{t=0}^{T} X(t)dt]$ by implementing a standard Monte-Carlo procedure for approximating the expectation $\mathbb{E}[\int_{t=0}^{T} \hat{X}_{(n)}(t)dt]$.

Maximum exposure value. In a similar fashion, the function $X^{(T)} \in D((0,T),\mathbb{R}) \mapsto \sup_{0\leq t\leq T} X(t)$ is Lipschitz w.r.t the distance $m_{\mathcal{M}_2}^{(T)}$ (see Theorem 13.4.1 in [199] for instance) and under the assumptions of Theorem 5, the expected supremum $\mathbb{E}[\sup_{0\leq t\leq T} X(t)]$ is finite and may be consistently estimated by Monte-Carlo simulations.

First passage times. Given the starting point x_0 of the exposure process X, the distribution of the first passage time beyond a certain (possibly critical) level $x \geq 0$, *i.e.* the hitting time $\tau_x^+ = \inf\{t \geq 0, X(t) \geq x\}$, is also a characteristic of crucial interest for toxicologists. The mapping $X \in D((0,\infty),\mathbb{R}) \mapsto \tau_x^+$ being continuous w.r.t. the $\mathcal{M}_2$-topology (refer to Theorem 13.6.4 in [199]), we have $\hat{\tau}_x^+ = \inf\{t \geq 0, \hat{X}(t) \geq x\} \Rightarrow \tau_x^+$ as soon as $\hat{X} \Rightarrow X$.

Steady State mean exposure. In practice, one is also concerned with *steady-state* characteristics, describing the long term behavior of the exposure process. The steady-state mean exposure m_μ can be pertinently used as a quantitative indicator for chronic risk characterisation. By virtue of Theorem 44 and Corollary 46, in an asymptotic framework stipulating that both $T \to \infty$ and $n \to \infty$, it can be consistently estimated by $\mathbb{E}[T^{-1}\int_{t=0}^{T} \hat{X}_{(n)}(t)dt$ since one may naturally write

$$\begin{aligned}\mathbb{E}[T^{-1}\int_{t=0}^{T} \hat{X}_{(n)}(t)dt] - m_\mu &= \mathbb{E}[T^{-1}\int_{t=0}^{T} \hat{X}_{(n)}(t)dt] - \mathbb{E}[T^{-1}\int_{t=0}^{T} X(t)dt] \\ &+ \mathbb{E}[T^{-1}\int_{t=0}^{T} X(t)dt] - m_\mu.\end{aligned}$$

Besides, with regard to statistical applications, Theorem 46 paves the way for studying the asymptotic validity of simulation estimators and in particular of bootstrap procedures in order to construct accurate confidence intervals (based on sample paths simulated from bootstrapped versions of the estimates $\widehat{G}^{(n)}, \widehat{F}_U^{(n)}, \widehat{H}^{(n)}$) as well. This is the subject of [27], in which these inference methods have been applied to the important case of dietary methyl mercury contamination.

Bibliography

[1] Abramovitz L., Singh K.(1985). Edgeworth Corrected Pivotal Statistics and the Bootstrap, *Ann. Stat.*, **13** ,116-132.

[2] Acerbi, C. & Tasche, D. (2001). On the coherence of expected shortfall, *Journal of Banking and Finance*, **26**, 1487–1503.

[3] Adamczak, R. (2005). Moment inequalities for U-statistics. Technical report, Institute of Mathematics of the Polish Academy of Sciences.

[4] Agarwal, S. Graepel, T., Herbrich, R., Har-Peled, S. & Roth, D (2005). Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425.

[5] Arcones, A. & Giné, E. (1993). Limit theorems for U-processes. *The Annals of Probability*, 21:1494–1542.

[6] Arcones, A. & Giné, E. (1994). U-processes indexed by Vapnik-Cervonenkis classes of functions with applications to asymptotics and bootstrap of u-statistics with estimated parameters. *Stochastic Processes and their Applications*, 52:17–38.

[7] Arzac, E.R. & Bawa, V.S. (1977). Portfolio choice and equilibrium in capital markets with safety-first investors, *Journal of Financial Economics*, **4**, 277–288.

[8] Asmussen, S. (2003). *Applied Probabilities and Queues.* Second edition. Springer.

[9] Asmussen, S. (1998). Extremal Value Theory for Queues Via Cycle Maxima. *Extremes*, **1**, No 2, 137-168.

[10] Athreya, K.B. & Atuncar, G.S. (1998). Kernel estimation for real-valued Markov chains. *Sankhya*, **60**, series A, No 1, 1-17.

[11] Athreya, K.B. & Fuh, C.D. (1989). Bootstrapping Markov chains: countable case. *Tech. Rep.* B-89-7, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC.

[12] Athreya, K.B. & Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.,* **245**, 493-501.

[13] Bach, F.R., Heckerman, D., & Horvitz, E. (2005). On the path to an ideal ROC Curve: considering cost asymmetry in learning classifiers. In Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS), 2005.

[14] Bachelier, L. (1900). *Théorie de la spéculation.* Annales Scientifiques de l'Ecole Normale Supérieure, 3ème série, **17**, 21-88. Translation: *The Random Character of Stock Market Prices.* Ed. Paul Cootner, Cambridge, MA: MIT Press.

[15] Back, A.D. & Weigend, A.S. (1997). A First Application of Independent Component Analysis to Extracting Structure from Stock Returns. *Int. J. Neur. Syst.*, **8**, 473–484.

[16] Bartlett, P.L., Jordan, M. & McAuliffe, J.D. (2005). Convexity, classification, and risk bounds. *Journal of the American Statistical Association.*

[17] Bartlett, P.L. & Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields*, 135.

[18] Bertail, P. (1997). Second order properties of an extrapolated bootstrap without replacement: the i.i.d. and the strong mixing cases,*Bernoulli*, **3**, 149-179.

[19] Bertail, P. & Clémençon, S. (2004a). Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. *Prob. Th. Rel. Fields*, **130**, 388–414.

[20] Bertail, P. & Clémençon, S. (2004b). Approximate Regenerative Block-Bootstrap for Markov Chains: second-order properties. In *Compstat 2004 Proc.* Physica Verlag.

[21] Bertail, P. & Clémençon, S. (2005a). Note on the regeneration-based bootstrap for atomic Markov chains. To appear in *Test.*

[22] Bertail, P. & Clémençon, S. (2005b). Regenerative Block Bootstrap for Markov Chains. To appear in *Bernoulli.*

[23] Bertail, P. & Clémençon, S. (2006a). Regeneration-based statistics for Harris recurrent Markov chains. In *Dependence in Probability and Statistics*, Eds P. Bertail, P. Doukhan & P. Soulier. Springer.

[24] Bertail, P. & Clémençon, S. (2006b). Regenerative Block Bootstrap for Markov Chains: some simulation studies. To appear in *Comp. Stat. Data Analysis.*

[25] Bertail, P. & Clémençon, S. & Rohmari, N. (2006a). Sharp probability inequalities for Harris Markov chains. Submitted.

[26] Bertail, P. & Clémençon, S. & Tressou, J. (2006b). A storage model with random release rate for modelling exposure to food contaminants. Submitted.

[27] Bertail, P., Clémençon, S. & Tressou, J. (2006b). Statistical Analysis of a Dynamic Model for Food Contamination Exposure with Applications to Dietary Methyl Mercury Contamination. Submitted.

[28] Bertail, P. & Politis, D. (2001). Extrapolation of subsampling distribution estimators in the i.i.d. and strong-mixing cases, *Can. J. Stat.*, **29**, 667-680.

[29] Bertail, P. & Tressou, J. (2006). Incomplete generalized U-statistics for food risk assessment. *Biometrics* **62** (1), 66-74.

[30] Bickel, P. & Freedman, D. (1981). SOme asymptotic theory for the bootstrap. *Ann. Stat.*, **9**, 1196-1217.

[31] Bickel, P.J. & Kwon, J. (2001). Inference for Semiparametric Models: Some Current Frontiers. *Stat. Sin.*, **11**, No. 4, 863-960.

[32] Bingham N.H., Goldie G.M. & Teugels J.L. (1989): *Regular Variation*, Cambridge University Press.

[33] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahr. verw. Gebiete,* **65**, 181-237.

[34] Blanchard, G., Lugosi, G. & Vayatis, N. (2003). On the rates of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894.

[35] Bolthausen, E. (1980). The Berry-Esseen Theorem for strongly mixing Harris recurrent Markov Chains. *Z. Wahr. Verw. Gebiete*, **54**, 59-73.

[36] Bolthausen, E. (1982). The Berry-Esseen Theorem for strongly mixing Harris recurrent Markov Chains. *Z. Wahr. Verw. Gebiete*, **60**, 283-289.

[37] Borovkova, S., Burton R. & Dehling H. (1999). Consistency of the Takens estimator for the correlation dimension. Ann. Appl. Prob., **9**, No. 2, 376-390.

[38] Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions, *Ann. Statist.*, 16, 1709-1722.

[39] Bouchaud, J.P., Sornette, D., Walter, C. & Aguilar, J.P. (1998). Taming Large Events: Optimal Portfolio Theory for Strongly Fluctuating Assets, *International Journal of Theoretical and Applied Finance*, **1**, 25–41.

[40] Boucheron, S., Bousquet, O. & Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM. Probability and Statistics*, 9:323–375.

[41] Boucheron, S., Bousquet, O., Lugosi, G. & Massart, P. (2005). Moment inequalities for functions of independent random variables. *The Annals Probability*, 33:514–560.

[42] Bradley, B.O. & Taqqu, M.S. (2004). An Extreme Value Theory Approach to the Allocation of Multiple Assets, *International Journal of Theoretical and Applied Finance*, **7**, 1031–1068.

[43] Breiman, L. (2004). Population theory for boosting ensembles. *Annals of Statistics*, 32:1–11.

[44] Brockwell, P.J., Resnick, S.J. & Tweedie, R.L. (1982). Storage processes with general release rules and additive inputs. *Adv. Appl. Probab.,* **14**, 392-433.

[45] Browne, S. & Sigman, K. (1992). Work-modulated queues with applications to storage processes. *J. Appl. Probab.,* **29**, 699-712.

[46] Bühlmann, P. (1997). Sieve Bootstrap for time series. *Bernoulli,* **3**, 123-148.

[47] Bühlmann, P. (2002). Bootstrap for time series. *Stat. Sci.*, **17**, 52-72.

[48] Callaert, H. & Veraverbeke, N. (1981). The order of the normal approximation for a Studentized statistic. *Ann. Stat.,* **9**, 194-200.

[49] Capobianco, E. (2002). Multiresolution approximation for volatility processes. *Quantitative Finance*, **2**, No2, 91-110.

[50] Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.*, **14**, 1171-1179.

[51] Chan, L.W. & Cha, S.M. (2001). Selection of independent factor model in finance, *Proceedings of the Third International Conference on Independent Component Analysis and Signal Separation*, (2001).

[52] Chen, X. (1999). *Limit Theorems for Functionals of Ergodic Markov Chains with General State Space.* Memoirs of the AMS, **139**, No 139.

[53] Clémençon, S. (2000). *Méthodes d'ondelettes pour la statistique non paramétrique des chaînes de Markov.* PhD thesis, Université Paris 7 Denis Diderot.

[54] Clémençon, S. (2000). Adaptive estimation of the transition density of a regular Markov chain. *Math. Meth. Stat.,* **9**, No. 4, 323-357.

[55] Clémençon, S. (2001). Moment and probability inequalities for sums of bounded additive functionals of regular Markov chains via the Nummelin splitting technique. *Stat. Prob. Letters,* **55**, 227-238.

[56] Clémençon, S. (2002). Nonparametric inference for some class of hidden Markov models. Rapport technique de l'université Paris X, No 03-9.

[57] Clémençon, S., Lugosi, G. & Vayatis, N. (2005a). From ranking problem: a statistical view. In *Studies in Classification, Data Analysis, and Knowledge Organization*, From Data and Information Analysis to Knowledge Engineering, Vol. 30, eds Myra Spiliopoulou, Rudolf Kruse, Andreas Nürnberger, C. Borgelt, Wolfgang Gaul (eds.): Proc. 29th Annual Conference of the GfKl, Otto-von-Guericke-University of Magdeburg, March 9-11, 2005, 214-221. Springer-Verlag, Heidelberg-Berlin, 2006.

[58] Clémençon, S., Lugosi, G. & Vayatis, N. (2005b). Ranking and Scoring Using Empirical Risk Minimization. In *Lecture Notes in Computer Science*, **3559**, Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005. Proceedings. Editors: Peter Auer, Ron Meir, 1-15. Springer-Verlag.

[59] Clémençon, S., Lugosi, G. & Vayatis, N. (2006a). Nonparametric scoring and U-processes. *Submitted.*

[60] Clémençon, S. & Slim, S. (2003). Statistical analysis of financial time series under the assumption of local stationarity. *Quantitative Finance.*

[61] Clémençon, S. & Slim, S. (2006). On Portfolio Selection under Extreme Risk Measure: the Heavy-Tailed ICA Model. To appear in *Int. J. Theoret. Appl. Finance.*

[62] Clémençon, S. & Vayatis, N. (2006b). On Ranking the Best Instances. *Submitted.*

[63] Cortes, C. & Mohri, M. (2004). AUC Optimization vs. Error Rate Minimization. In *Advances in Neural Information Processing Systems, eds Thrun, S. Saul, L. Schölkopf, B., MIT Press.*

[64] Costa, O.L.V. (1990). Stationary Distributions for Piecewise-Deterministic Markov Processes. *Journal of Applied Probability* **27** (1), 60-73.

[65] Cucker, F. & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49.

[66] Dahlaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Stat.,* **25**, 1-37.

[67] Datta, S. & McCormick W.P. (1993). Regeneration-based bootstrap for Markov chains. *Can. J. Statist.,* **21**, No.2, 181-193.

[68] Davis, M.H.A. (1991). *Applied Stochastic Analysis.* Taylor & Francis (Stochastics Monographs).

[69] Davis, M.H.A. (1984). Piecewise-Deterministic Markov Processes: A General Class of Non-Diffusion Stochastic Models. *Journal of the Royal Statistical Society. Series B (Methodological)* **46**, No. 3, 353-388.

[70] de la Peña, V.H. & Giné, E. (1999). *Decoupling: from Dependence to Independence.* Springer, New York.

[71] Deheuvels, P. Häusler, E. & Mason, D.M. (1988). Almost sure convergence of the Hill estimator. *Math. Proc. Camb. Philos. Soc.,* **104**, 371-381.

[72] Devroye, L., Györfi, L. & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York.

[73] Donoho, D.L., Mallat, S. & von Sachs, R. (1998). Estimating Covariances of Locally Stationary Processes: Rates of Convergence of Best Basis Methods. *Tech. Rep. 517, Statistics Department. Stanford University.*

[74] Donoho, D., Mallat, S., von Sachs, R. & Samuelides, Y. (2003). Locally Stationary Covariance and Signal Estimation with Macrotiles. *IEEE Transactions on Signal Processing*, Vol. 51, No. 3, 614- 627.

[75] Doukhan, P. (1994). *Mixing: Properties and Examples.* Lecture Notes in Statist., 85. Springer, New York.

[76] Doukhan, P. & Ghindès, M. (1983). Estimation de la transition de probabilité d'une chaîne de Markov Doeblin récurrente. *Stochastic Process. Appl.,* **15**, 271-293.

[77] Duffie, D. & Pan, J. (1997). An Overview of Value at Risk, The Journal of Derivatives, **4**, 7–49.

[78] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1-26.

[79] Embrechts, P. (2000). *Extremes and Integrated Risk Management*, London: Risk Books, Risk Waters Group.

[80] Embrechts, P., Klüppelberg, C. & Mikosch, T. (2001). *Modelling Extremal Events.* Springer-Verlag.

[81] Embrechts, P., Resnick, D. & Samorodnitsky, G. (1999). Extreme Value Theory as a Risk Management Tool, *North Amer. Act. Journ.* , **3**, 30–41.

[82] Engle, R. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica, Vol. 50, No. 4, 987-1008

[83] Feller, W. (1968). *An Introduction to Probability Theory and its Applications: vol. I.* John Wiley & Sons, NY, 2nd edition.

[84] Feller, W. (1971). *An Introduction to Probability Theory and its Applications: vol. II.* John Wiley & Sons, NY, 3rd edition

[85] Franke, J. , Kreiss, J. P. & Mammen, E. (2002). Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli*, **8**, 1–37.

[86] Freund, Y., Iyer,R., Schapire, R.E. & Singer, Y. (2004). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4(6):933–969.

[87] Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, 28:307–337.

[88] Fryzlewicz, S., van Bellegem, S. & von Sachs, R. (2002). Forecasting non-stationary time series by wavelet process modeling. *Tech. Rep.*, Department of Mathematics, University of Bristol.

[89] Fukasawa, M. (2005). Bootstrap and Edgeworth expansions for ergodic diffusions. *Submitted.*

[90] Giné, E., Latała, R. & Zinn, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II—Progress in Probability*, pages 13–38. Birkhauser.

[91] Giné, E. & Zinn, J. (1984). Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989.

[92] Glynn, W.P. & Zeevi, A. (2000). Estimating Tail Probabilities in Queues via Extremal Statistics. In *Analysis of Communication Networks: Call Centres, Traffic, and Performance* [D.R. McDonald and S.R. Turner, eds.] AMS, Providence, Rhode Island, 135-158.

[93] Glynn, W.P. & Whitt, W. (1995). Heavy-Traffic Extreme-Value Limits for Queues. Op. Res. Lett. **18**, 107-111.

[94] Goldie, C.M. (1991). Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Prob.,* **1**, 126-166.

[95] Götze & F., Hipp, C. (1983). Asymptotic expansions for sums of weakly dependent random vectors. *Zeit. Wahrschein. verw. Geb.*, **64**, 211-239.

[96] Götze, F. & Künsch, H.R. (1996). Second order correctness of the blockwise bootstrap for stationary observations. *Ann. Statist.*, **24**, 1914-1933.

[97] Greenblatt, S.A., (1996). Atomic Decomposition of Financial Data. *Computational Economics, 12,* No 3, 275-293. Kluwer Academic Publisher.

[98] Green, D.M. & Swets, J.A. (1966). *Signal detection theory and psychophysics.* Wiley, NY.

[99] de Haan, L. (1984). Slow variation and the characterization of domains of attraction. In *Statistical Extremes and Applications,* Ed. Tiago de Oliveira, Reidel, Dordrecht (1984) 31–48.

[100] Hall P. (1983). Inverting an Edgeworth Expansion. *Ann. Statist.*, **11**, 569-576.

[101] Hall, P. (1985). Resampling a coverage pattern. *Stoch. Process. Applic.*, **20**, 231-246.

[102] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer.

[103] Hall, P., Horowitz, J. & Jing, B.-Y. (1995), On blocking rules for the bootstrap with dependent data, *Biometrika*, 82, 561-574.

[104] Hanley, J.A. & McNeil, J. (1982). The meaning and use of the area under a ROC curve. Radiology, 143: 29-36.

[105] Harari-Kermadec, H. (2006). Regenerative Block Empirical Likelihood. *Submitted.*

[106] Härdle, W., Herwartz, H. & Spokoiny, V. (2001). Time Inhomogeneous Multiple Volatility Modelling. Discussion Paper 7, Sonderforschungsbereich 373, Humboldt University, Berlin.

[107] Harrison, J.M. & Resnick, S.J. (1976). The stationary distribution and first exit probabilities of a storage process with general release rule. *Math. Oper. Res.,* **1**, 347-358.

[108] Haussler, D. (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69:217–233.

[109] Helmers, R (1991). On the Edgeworth expansion and the bootstrap approximation for a studentized statistics. *Ann. Statist.* ,**19**, 470-484.

[110] Hill, B. (1975). A simple approach to inference about the tail of a distribution. *Ann. Stat.*, **3**, No. 5, 1163-1174.

[111] Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 10:293–325.

[112] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

[113] Horowitz, J. (2003). Bootstrap Methods for Markov Processes. *Econometrica*, **71**, No 4, 10-49.

[114] Houdré, C. & Reynaud-Bouret, P. (2003). Exponential Inequalities, with constants, for U-statistics of order two. Stochastic Inequalities and Applications - Progress in Probability, Birkhauser.

[115] Hsieh, F. & Turnbull, B.W. (1996). Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. *Ann. Stat.*, **24**, No. 1, 25-40.

[116] Hyung, N. & de Vries, C. (2004). Portfolio Selection with Heavy Tails, *Tinbergen Institute Discussion Papers* 05–009/2.

[117] Hyvärinen, A., Karhunen, J. & Oja, E. (2001) *Independent Component Analysis* , Wiley-Interscience.

[118] Jain, J. & Jamison, B. (1967). Contributions to Doeblin's theory of Markov processes. *Z. Wahrsch. Verw. Geb.,* **8**, 19-40.

[119] Jansen, D., Koedijk, K. & de Vries, C. (2000). Portfolio selection with limited downside risk, *Journal of Empirical Finance,* **7**, 247–269.

[120] Jiang, W. (2004). Process consistency for Adaboost (with discussion). *Annals of Statistics*, 32:13–29.

[121] Kalashnikov, V.V. (1978). *The Qualitative Analysis of the Behavior of Complex Systems by the Method of Test Functions.* Nauka, Moscow.

[122] Karlsen, H.A. & Tjøstheim, D. (2001). Nonparametric estimation in null recurrent time series. *Ann. Statist.,* **29** (2), 372-416.

[123] Kerkyacharian, G. & Picard, D. (2000). Thresholding algorithms, maxisets and well concentrated bases. *Test*, **9**, No 2, 283-344.

[124] Kiviluoto, K. & Oja, E. (1998). Independent component analysis for parallel financial time series. In *Proc. Proc. Int. Conf. on Neural Information Processing (ICONIP'98),* **2**, 895–989, Tokyo, Japan.

[125] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 36.

[126] Koltchinskii, V. & Panchenko, D. (2002). Empirical margin distribution and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:1–50.

[127] Kroisandt, G., Nason, G.P. & von Sachs, R. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectra. *J. R. Stat. Soc. Ser.* B, **62**, 271-292.

[128] Künsch, H.R. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.,* **12**, 843-863.

[129] Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.,* **17**, 1217-1241.

[130] Lahiri, S.N. (2003). *Resampling methods for dependent Data*, Springer.

[131] Leadbetter, M.R. & Rootzén, H. (1988). Extremal Theory for Stochastic Processes. *Ann. Prob.*, **16**, No. 2, 431-478.

[132] Ledoux, M. (1996). On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1997.

[133] Liu R. & Singh K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring The Limits of The Bootstrap.* Ed. Le Page R. and Billard L., John Wiley, NY.

[134] Lugosi, G. (2002). Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 5–62. Springer, Wien.

[135] Lugosi, G. & Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods (with discussion). *Annals of Statistics*, 32:30–55.

[136] Malevergne, Y. & Sornette, D. (2001). General framework for a portfolio theory with non-Gaussian risks and non-linear correlations. In *Proceedings of the 18th International Conference in Finance*, 26, 27 & 28 June 2001 Namur - Belgium.

[137] Malinovskii, V. K. (1985). On some asymptotic relations and identities for Harris recurrent Markov Chains. *In Statistics and Control of Stochastic Processes*, 317-336.

[138] Malinovskii, V. K. (1987). Limit theorems for Harris Markov chains I. *Theory Prob. Appl.*, **31**, 269-285.

[139] Malinovskii, V. K. (1989). Limit theorems for Harris Markov chains II. *Theory Prob. Appl.*, **34**, 252-265.

[140] Mallat, S., Papanicolaou, G. & Zhang, Z. (1998). Adaptive covariance estimation of locally stationary processes. *Ann. Stat.,* **26**, No 1, 1-47.

[141] Mallat, S. & Samuelides, Y. (2001). Non-stationary covariance estimation with macrotiles. *Tech. Rep., Ecole Polytechnique.*

[142] Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*, **7**, 77-91.

[143] Martin, R.D. & Yohai, V.J. (1986). Influence functionals for time series. *Ann. Stat.,* **14**, 781-818.

[144] Mason, D.M. (1982). Laws of large numbers for sums of extreme values, *Ann. Prob.,* **10**, 756–764.

[145] Massart, P. (2006). *Concentration inequalities and model selection.* Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer.

[146] Massart, P. & Nédélec, E. (2006). Risk bounds for statistical learning. *Annals of Statistics*, 34.

[147] McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge.

[148] Meyn, S.P. & Tweedie, R.L., (1996). *Markov chains and stochastic stability.* Springer.

[149] Müller, U.U., Schick, A. & Wefelmeyer, W., (2001). Improved estimators for constrained Markov chain models. *Stat. Prob. Lett.,* **54**, 427-435.

[150] Nummelin, E. (1978). A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete,* **43**, 309-318.

[151] Nummelin, E. (1984). *General irreducible Markov chains and non negative operators.* Cambridge University Press, Cambridge.

[152] Paparoditis, E. & Politis, D.N. (2002). The local bootstrap for Markov processes. *J. Statist. Plan. Infer.*, **108**, 301–328.

[153] Pepe, M. (2003). *Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

[154] Politis, D.N. & Romano, J.P. (1992). A General Resampling Scheme for Triangular Arrays of alpha-mixing Random Variables with Application to the Problem of Spectral Density Estimation, *Ann. Statist.*, **20**, 1985-2007.

[155] Politis, D.N. & Romano, J.P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.,* **22**, 2031-2050.

[156] Politis, D.N., Romano, J.P. & Wolf, T. (2000). *Subsampling.* Springer Series in Statistics, Springer, NY.

[157] Politis, D.N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science* , **18**, No. 2, 219-230.

[158] Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation.* Academic Press, NY.

[159] Priestley, M.B. (1965). Evolutionary spectra and non-stationary processes. *J. R. Statist. Soc. B,* **27**, 204-237.

[160] Rachev, S. T., Rüschendorf, L. (1998). *Mass Transportation Problems. Vol. I* and *II.* Springer.

[161] Ramsey, J.B. (1999). The Contribution of Wavelets to the Analysis of Economic and Financial Data. *Phil. Trans. R. Soc. Lond. A,* **357**, 2593-2606.

[162] Ramsey, J.B. & Zhang, Z (1996). The Application of Waveform Dictionaries to Stock Market Index Data. In *Predictability of Complex Dynamical Systems*, eds Y. Kravtsov & J. Kadtke, Springer.

[163] Ramsey, J.B. (2002). Wavelets in Economics and Finance: Past and Future. *Studies in Nonlinear Dynamics & Econometrics*, **6**, No 3, art. 1.

[164] Resnick, S. (1987). *Extreme Values, Regular Variation and Point Processes.* Springer, NY.

[165] Resnick, S. (1997). Heavy Tail Modeling And Teletraffic Data. *Ann. Stat., 25, 1805-1869.*

[166] Resnick, S. & Starica, C. (1995). Consistency of Hill estimator for dependent data. *J. Appl. Prob.*, **32**, 139-167.

[167] Revuz, D (1984). *Markov chains.* North-Holland, 2nd edition.

[168] Roberts, G.O. & Rosenthal, J.S. (1996). Quantitative bounds for convergence rates of continuous time Markov processes. *Electr. Journ. Prob.*, **9**, 1-21.

[169] Roberts, G.O. & Tweedie, R.L. (2000). Rates of Convergence of Stochastically Monotone and Continuous Time Markov Models. *J. Appl. Prob.*, **37**, (2), 359-373.

[170] Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference,* Ed. M. Puri, 199-210.

[171] Rootzén, H. (1988). Maxima and exceedances of stationary Markov chains. *Adv. Appl. Prob.*, **20**, 371-390.

[172] Roussas, G. (1969). Nonparametric Estimation in Markov Processes. *Ann. Inst. Stat. Math.*, 73-87.

[173] Roussas, G. (1991a). Estimation of transition distribution function and its quantiles in Markov Processes. In *Nonparametric Functional Estimation and Related Topics*, Ed. G. Roussas, 443-462.

[174] Roussas, G. (1991b). Recursive estimation of the transition distribution function of a Markov Process. *Stat. Probab. Letters*, **11**, 435-447.

[175] Roy, A.D. (1952). Safety first and the holding of assets, *Econometrica,* **20**, 431–449.

[176] Samuelides, Y. (2001). *Macrotile Estimation and Market Model with Jumps.* PhD thesis, Ecole Polytechnique.

[177] Schick, A. (2001). Sample splitting with Markov chains. *Bernoulli*, **7**, No 1, 33-61.

[178] Scott, C. & Nowak, R. (2006). Learning minimum volume sets. *Journal of Machine Learning Research*, **7**, 665–704.

[179] Scovel, S. & Steinwart, I. (2003). Fast Rates for Support Vector Machines. Technical Report LA-UR-03-9117, Los Alamos National Laboratory.

[180] Serfling, R.J. (1980). Approximation theorems of mathematical statistics. John Wiley & Sons.

[181] Serfling, R. & Zuo, Y. (2000). General Notions of Statistical Depth Function (in Data Depth). *Ann. Stat.*, **28**, No. 2., 461-482.

[182] Sigman, K. & Wolff, R. (1993). A Review of Regenerative Processes. *SIAM Review*, **35**, No 2, 269-288.

[183] Slim, S. (2005). *Nonparametric statistical methods for the analysis of stock returns and risk modelling.* PhD thesis, Université Paris X Nanterre.

[184] Smale, S. & Zhou, D.X. (2003). Estimating the approximation error in learning theory. Analysis and Applications, 1, pp. 17-41. Support Vector Machine Soft Margin Classifiers.

[185] Smith, W. L. (1955). Regenerative stochastic processes. *Proc. Royal Stat. Soc., A,* **232**, 6-31.

[186] Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93.

[187] Stute, W. (1991). Conditional U-statistics. *Annals of Probability*, 19:812–825.

[188] Stute, W. (1994). Universally consistent conditional U-statistics. *The Annals of Statistics*, 22:460–473.

[189] Szegö, G. (2004). *Risk Measures for the 21-th Century*, Edited by G. Szegö Wiley (2004).

[190] Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563.

[191] Tjøstheim, D. (1990). Non Linear Time series, *Adv. Appl. Prob.*, **22**, 587-611.

[192] Tsybakov, A.B. (2002). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166.

[193] Thorisson, H. (2000). *Coupling, Stationarity and Regeneration.* Springer.

[194] Tressou, J. (2005). *Méthodes statistiques pour l'évaluation du risque alimentaire.* Thèse de l'Université Paris X.

[195] Tressou, J. (2006). Nonparametric Modelling of the Left Censorship of Analytical Data in Food Risk Exposure Assessment. To appear in *J.A.S.A.*

[196] van de Geer, S. (2000). *Empirical Processes in M-Estimation* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[197] van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press.

[198] Vapnik, V.N. & Chervonenkis, A.Y. (1974). *Theory of Pattern Recognition.* Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

[199] Whitt, W. (2002). *Stochastic-Process Limits. An Introduction to Stochastic-Process Limits and their Application to Queues.* Springer.

[200] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization (with discussion). Annals of Statistics, 32:56–85.