Rencontres Mathématiques de Rouen

19-21 juin 2024

Modèles statistiques pour des données dépendantes et applications

Christophe Biernacki (Inria & U. Lille)

Model-based Co-Clustering: High Dimension and Estimation Challenges

Joint with Christine Keribin (U. Paris-Saclay) and Julien Jacques (U. Lyon 2)

Model-based co-clustering can be seen as a particularly important extension of model-based clustering. It allows for a significant reduction of both the number of rows (individuals) and columns (variables) of a data set in a parsimonious manner, and also allows interpretability of the resulting reduced data set since the meaning of the initial individuals and features is preserved. Moreover, it benefits from the rich statistical theory for both estimation and model selection. Many works have produced new advances on this topic in recent years, and we offer a general update of the related literature. It is the opportunity to advocate two main messages, supported by specific research material: (1) co-clustering requires further research to fix some well-identified estimation issues, and (2) co-clustering is one of the most promising approaches for clustering in the (very) high-dimensional setting, which corresponds to the global trend in modern data sets.

Michel Broniatowski (LPSM, Sorbonne Université Paris)

Stochastic minimization of directed distances

Joint with W. Stummer (FAU Erlangen-Nurnberg) et P. Bertrand (BNP)

On montre que des divergences de Csiszar entre deux mesures signées sur un ensemble fini de $\mathbb{R}^K$ peuvent être identifiées comme taux de grande déviation

pour des suites de vecteurs aléatoires qui peuvent être simulés de façon simple, dont la loi dépend du générateur de la divergence. Cette propriété permet l'optimisation de la divergence sur des ensembles de vecteurs d'intérieur non vide sous des conditions de régularité générales. La classe de divergences admettant cette représentation est celle dont le générateur s'écrit comme transformée de Legendre d'une fonction génératrice des moments. Cette propriété permet également de généraliser la construction ci-dessus à de grandes classes de divergences. Des versions spécifiques permettent l'optimisation de diverses entropies sous des contraintes variées.

Une conséquence en est aussi la minimisation d'une divergence de Csszar entre une loi de probabilité et un modèle, vu comme sous ensemble du simplex d'intérieur non vide. Des algorithmes explicites parallélisables utilisant des techniques d'Importance Sampling permettent d'estimer la divergence entre une mesure empirique et un modèle.

Ces méthodes permettent la minimisation de divergences de Bregman; couplées au Lemme de Varadhan elle permettent la minimisation de fonctions continues de K variables sur des sous ensembles de $\mathbb{R}^K$ sous des conditions de régularité souples. On présentera des algorithmes d'approximation des valeurs minimales de la fonction et de ses minimiseurs.

Une application au contexte de l'inférence pour des réseaux de neurones en classification supervisée d'images sera présentée, en grande dimension (ici 4×10^5) sous la forme d'un problème de minimisation de la norme $l1$ du vecteur des paramètres de nombre de composantes non nulles minimal sous une contrainte relative à la qualité de restitution des classes sur un échantillon témoin.

References

[M. Broniatowski and W. Stummer (2003)] Broniatowski, M. and Stummer, W. (2023) A precise bare simulation approach to the minimization of some distances. I. Foundations. *IEEE Trans. Inform. Theory*, 69:5, 3062–3120.

[M. Broniatowski and W. Stummer (2024)] Broniatowski, M. and Stummer, W. (2024) A precise bare simulation approach to the minimization of some distances. II. Further foundations. *arXiv:2402.08478*.

[P. Bertrand (2024)] Bertrand, P., Broniatowski, M, and Stummer, W. (2024) Optimizing Neural Networks through Bare Simulation search. *arXiv*.

Julien Chiquet (AgroParisTech)

Zero-inflation in the Multivariate Poisson Lognormal Family: variational inference and application to microbiome data

Analyzing high-dimensional count data is a challenge and statistical model-based approaches provide an adequate and efficient framework that preserves

explainability. The (multivariate) Poisson-Log-Normal (PLN) model is one such model: it assumes count data are driven by an underlying structured latent Gaussian variable, so that the dependencies between counts solely stems from the latent dependencies. However PLN doesn't account for zero-inflation, a feature frequently observed in real-world datasets. Here we introduce the Zero-Inflated PLN (ZIPLN) model, adding a multivariate zero-inflated component to the model, as an additional Bernoulli latent variable. The Zero-Inflation can be fixed, site-specific, feature-specific or depends on covariates. We estimate model parameters using variational inference that scales up to datasets with a few thousands variables and compare two approximations: (i) independent Gaussian and Bernoulli variational distributions or (ii) Gaussian variational distribution conditioned on the Bernoulli one. The method is assessed on synthetic data and the efficiency of ZIPLN is established even when zero-inflation concerns up to 90% of the observed counts. We then apply both ZIPLN and PLN to a cow microbiome dataset, containing 90.6% of zeroes. Accounting for zero-inflation significantly increases log-likelihood and reduces dispersion in the latent space, thus leading to improved group discrimination.

Emmanuelle Clément (Université Gustave Eiffel)

High-frequency estimation of stable CIR processes

Joint with Elise Bayraktar

We are interested in estimating the parameters of a stable Cox-Ingersoll-Ross (CIR) process from high-frequency observations on a fixed time period $(X_{i/n})_{0\leq i\leq n}$ where $(X_t)_{t\in[0,1]}$ solves the stochastic equation driven by a Brownian motion $(B_t)_{t\geq 0}$ and a spectrally positive stable Lévy process $(L_t^\alpha)_{t\geq 0}$ with jump activity $\alpha \in (1,2)$

$$X_t = x_0 + \int_0^t (a - bX_s)ds + \sigma \int_0^t \sqrt{X_s} dB_s + \delta \int_0^t X_{s-}^{1/\alpha} dL_s^\alpha.$$

In a first part, we consider a pure-jump stable CIR process ($\sigma = 0$ in the previous equation) and prove the existence of a consistent and asymptotically conditionally Gaussian estimator of all the parameters (drift coefficients a and b, scaling coefficient δ, jump activity α). Next we propose preliminary estimators easy to implement but not rate optimal and improve them by a one-step procedure.
In a second part, we assume that the process is driven both by $(B_t)_{t\geq 0}$ and $(L_t^\alpha)_{t\geq 0}$ ($\sigma > 0$ and $\delta > 0$) and we focus on the estimation of the volatility σ and the jump activity α. Since the stable Lévy process has infinite variation, it is well known that the estimation of these parameters becomes to be harder due to the presence of asymptotic bias in statistical quantities. To overcome this difficulty, we adapt to the stable CIR process the estimation method based on the real part of the characteristic function proposed by Jacod and Todorov (2014).

References

[1] E. Bayraktar and E. Clément. Estimation of a pure-jump Cox-Ingersoll-Ross process. *Bernoulli*, in press 2024.

[2] J. Jacod and V. Todorov. Efficient estimation of integrated volatility in presence of infinite variation jumps. *The Annals of Statistics,* 42(3):1029-1069, 2014.

Fabienne Comte (MAP5 & Université Paris Cité)

Nonparametric estimation for i.i.d. Stochastic Differential Equations with space-time dependent coefficients

Joint with Valentine Genon-Catalot, MAP5 & Université Paris Cité

We consider N *i.i.d.* one-dimensional inhomogeneous diffusion processes $(X_i(t), i = 1, \dots, N)$ with drift $\mu(t,x) = \sum_{j=1}^{K} \alpha_j(t) g_j(x)$ and diffusion coefficient $\sigma(t,x)$, where K, the functions $g_j(x)$ and $\sigma(t,x)$ are known. Our concern is the nonparametric estimation of the K-dimensional unknown function $(\alpha_j(t), j = 1, \dots, K)$ from the continuous observation of the sample paths $(X_i(t))$ throughout a fixed time interval $[0, \tau]$. A collection of projection estimators belonging to a product of finite-dimensional subspaces of $\mathbb{L}^2([0,\tau])$ is built. The $\mathbb{L}^2$-risk is defined by the expectation of either an empirical norm or a deterministic norm fitted to the problem. Rates of convergence for large N are discussed. A data-driven choice of the dimensions of the projection spaces is proposed. The theoretical results are illustrated by numerical experiments on simulated data.

Mitra Fouladirad (Centrale Méditerranée)

Degradation modelling for prognosis and maintennace

A system subject to degradation is considered. The latter is modelled by a stochastic process. The reliability at each moment is calculated. A condition-based maintenance policy is proposed and the properties of the maintained system are studied.

Ahmed Kebaier (Université d'Evry)

Multilevel Monte Carlo for pricing Barrier options under CIR, CEV and Heston models

Joint with Mouna Ben Derouich

In this work, we demonstrate how Multi-Level Monte Carlo (MLMC) techniques can be used to price barrier options for models possibly with non-globally Lipschitz diffusion coefficients. In the first part, we consider a general framework of one-dimensional models with diffusion coefficients that are not necessarily globally Lipschitz. We then introduce an interpolated implicit Euler scheme for which we prove a strong convergence result of order one. Following [1] works, we analyse the extreme trajectories of the diffusion process and its approximation, and prove that the MLMC method reaches its optimal regime $O(\varepsilon^{-2})$ for a given total precision ε. We apply these results to the pricing of barrier options in the CIR model and the CEV local volatility model, and develop semi-analytical formulas for the densities of the running minimum and maximum of these two processes. In the second part, we extend this approach to the more challenging problem of pricing barrier options under the two-dimensional log-Heston model. We develop an approximation scheme allowing us to analyse the variance of the MLMC method when combined with Brownian bridge techniques. We show that the aforementioned MLMC method has a lower order complexity than the standard Monte Carlo method, however it does not reach the optimal regime.

References

[1] Michael B. Giles, Kristian Debrabant, and Andreas RÃ¶ssler. "Analysis of multilevel Monte Carlo path simulation using the Milstein discretisation." Discrete Contin. Dyn. Syst. Ser. B 24, no. 8 (2019): 3881-3903.

Youri Koutoyants (Le Mans Université)

Hidden Markov Processes and Adaptive Filtration

We present several models of partially observed diffusion processes and describe the construction of adaptive Kalman-type filters in the situations where the systems depend on some unknown finite-dimensional parameters. The presented algorithms for the filters and estimators of the parameters have recurrent structure and the questions of their asymptotic optimality are discussed. The properties of the filters and estimators are studied in the asymptotics of small noise and large samples. For some nonlinear systems the construction and properties of the corresponding extended adaptive Kalman filters are discussed too.

Nikolaos Liminios (Université de Technologie de Compiègne)

Normal Deviation of Gamma Processes in Random Media

The aim of this presentation is to approximate, by a diffusion process, the deviation of a gamma processes from its average evolution, in a random environment. In fact, as the gamma process is an increasing one, the diffusion approximation requires an average approximation first. This averaged process will serve as an equilibrium to the initial gamma processes.

Catherine Matias, (CNRS; Sorbonne Université)

A guided tour on nodes clustering in hypergraphs

Joint with Luca Brusa and Veronica Poda

Over the past two decades, a wide range of models has been developed to capture pairwise interactions represented in graphs. However, modern applications in various fields have highlighted the necessity to consider high-order interactions, which involve groups of three or more nodes. To formalize these high-order interactions, hypergraphs provide the most general framework. Similar to a graph, a hypergraph consists of a set of nodes and a set of hyperedges, where each hyperedge is a subset of nodes involved in an interaction. In this talk, I will guide you to the main challenges posed by nodes clustering in hypergraphs and present the 3 main approaches for such clustering: spectral, modularity-based and model-based through stochastic blockmodels.

Thi Bao TrÃ¢m Ngô (Le Mans UniversitÃ©)

Optimal guaranteed estimation methods for the Cox-Ingersoll-Ross models

Joint with Mohamed BEN ALAYA and Serguei PERGAMENCHTCHIKOV, UniversitÃ© de Rouen Normandie

In this work, we study parameter estimation problems for the Cox-Ingersoll-Ross (CIR) processes. For the first time, for such models, the sequential estimation procedures are proposed. In the non-asymptotic setting, these proposed sequential procedures provide the estimation with non-asymptotic fixed mean square accuracy. For the scalar parameter estimation problems, the non-asymptotic normality properties for the proposed estimators are established even in the cases when the classical non sequential maximum likelihood estimators can not

be calculated. Moreover, the Laplace transformations for the mean observation durations are obtained. In the asymptotic setting, the limit forms for the mean observation durations are founded and it is shown, that the constructed sequential estimators uniformly converge in distribution to normal random variables. Then, using the Local Asymptotic Normality (LAN) property, it is obtained the asymptotic sharp lower bound for the minimax risks in the class of all sequential procedures with the same mean observation duration and as consequence, it is established, that the proposed sequential procedures are optimal in the minimax sens in this class.

Igor Nikiforov (LIST3N, Université de Technologie de Troyes)

Reliable detection of unknown transient change profile by the FMA test

Joint with F. E. Mana, B. K. Guépié, and L. Fillatre

The sequential reliable transient change detection by the Finite Moving Average (FMA) test is considered. Unlike the traditional quickest change detection, which assumes that the post-change period is infinitely long, sometimes it is necessary to detect a change with a delay upper-bounded by L. All detections that exceed the required time to alert L are assumed missed. A transient change occurs at an unknown (but non random) change-point ν. We wish to minimize the worst-case probability of missed detection

$$\overline{\mathbb{P}}\left(T\right)=\sup_{\nu\geq L}\mathbb{P}_{\nu}\left(T-\nu+1>L\mid T\geq\nu\right) \tag{1}$$

for the worst-case probability of false alarm during the reference period m_α :

$$\overline{\mathbb{P}}(T;m_{\alpha})=\sup_{\ell\geq L}\mathbb{P}_{\infty}(\ell\leq T<\ell+m_{\alpha}), \tag{2}$$

where T is a stopping time, $\mathbb{P}_\nu$ (resp. $\mathbb{P}_\infty$) stands for the probability under which the change occurs at ν (resp. the change never occurs).

Let us consider the following generative model of transient change :

$$y_{n}\sim\begin{cases}\mathcal{N}(0,\sigma^{2}) & \text{if}\quad 1\leq n<\nu\\ \mathcal{N}(\theta_{n-\nu+1},\sigma^{2}) & \text{if}\quad \nu\leq n\leq\nu+L-1\end{cases}, \tag{3}$$

where $\mathcal{N}(\mu,\sigma^2)$ stands for the Gaussian law with mean μ and variance σ^2 and the vector $\theta=(\theta_1,\ldots,\theta_L)^T$ stands for the transient change profile.

This presentation continues the line of research established in [1, 2, 3]. The original contributions are the following : the assumption of the known profile θ is relaxed; new versions of the FMA test are designed by using the generalized likelihood ratio; some extensions of model (3) are considered. These new quadratic FMA tests are compared against the linear FMA test based on the putative transient change profile θ^* in the case where the true profile θ differs from the putative profile θ^*.

References

[1] B. K. Guépié, L. Fillatre, and I. Nikiforov. 2017. "Detecting a suddenly arriving dynamic profile of finite duration." *IEEE Transactions on Information Theory*, 63(5):3039–3052.

[2] F. E. Mana, B. K. GuÃ©piÃ©, I. Nikiforov. 2023. "Sequential Detection of an Arbitrary Transient Change Profile by the FMA Test." *Sequential Analysis*, 42 (2), pp.91-111.

[3] G. Sokolov, V. S. Spivak, and A. G. Tartakovsky. 2023. "Detecting an intermittent change of unknown duration." *Sequential Analysis*, 42 (3), pp. 269-302

Gilles Pagès (LPSM, Sorbonne Université Paris)

Functional convex order for stochastic processes: a constructive (and simulable) approach

Joint with B. Jourdain, Y. Liu and C. Yeo

After a few reminders on the convex order $\preceq_{cv}$ between two integrable random vectors U and V defined by

$$U \preceq_{cv} V \text{ if } \mathbb{E}\, f(U) \le \mathbb{E}\, f(V) \text{ for every convex function} f : \mathbb{R}^d \mapsto \mathbb{R},$$

(with some variants like monotonic convex order) and their first applications in finance, we will explain how to extend this order in a functional way to various classes of stochastic processes, in particular to diffusions (Brownian, with jumps, or McKean Vlasov type), even to non-Markovian processes, such as the solutions of Volterra equations with singular kernels like those appearing in rough volatility modeling in Finance.

We systematically establish our comparison results by an approximation procedure of Euler scheme type, generally simulable. Thus, among other virtues, this approach makes it possible in finance to ensure that the prices of derivative products computed by simulation cannot give rise to arbitrages by lack of convexity.On our way, we will also establish the convexity of functionals $x \mapsto \mathbb{E}\, F(X^x)$ of such stochastic processes X^x when F is (l.s.c. and) convex and x is the starting value of X^x.

Other applications to American option pricing to stochastic control (with in mind swing option pricing) will be briefly mentioned. If time si not too short, we will conclude by a focus on $1D$-(scaled martingale) diffusions where an important convexity assumptions can be relaxed, in connection with former works by El Karoui-Jeanblanc-Shrece (tracking error) and Rüschendorf and co-authors.

Nathalie Peyrard (MIAT -INRAE)

Deux extensions du cadre HSMM : cas de plusieurs chaînes cachées et cas où la dynamique de la chaîne cachée dépend des observations

Joint with Hanna Bacave, Nikolas Limnios, Sam Nicol, Ronan Trépos, Régis Sabbadin

Les modèles semi markoviens cachés (Hidden Semi-Markov Models, HSMM) sont une extension des HMMs où la durée de séjour de la chaîne cachée dans un état n'est pas nécessairement de loi géométrique et peut être une loi quelconque. De part cette flexibilité, ils capturent mieux la dynamique des processus étudiés et ont été adoptés dans de nombreux domaines (traitement du signal, séismologie, écologie, . . .). Néanmoins, certaines applications soulèvent encore des verrous pour l'application des HSMM. Je présenterai deux exemples, motivés par deux applications en écologie. Le premier concerne l'inférence de chemins migratoires d'oiseaux lorsque les seules données disponibles sont des données de comptages aux sites de passage. Nous avons modélisé ce problème comme plusieurs chaînes de semi-Markov (une chaîne cachée par oiseaux) dont la dynamique est 'factorisée' par les données. La taille du problème rend l'estimation exacte par maximum de vraisemblance impossible et je présenterai différentes solutions. Le deuxième exemple concerne la dynamique des plantes avec dormance, qui ont la particularité de disposer d'un stock de graines en dormance dans le sol et qui peuvent survivre plusieurs années avant de germer lorsque les conditions sont favorables. La modélisation par un HSMM classique n'est pas possible ici car l'observation courante impacte l'état caché. En particulier, la notion de distribution de la durée de séjour ne peut plus être utilisée telle quelle. Nous proposons une solution de modélisation, consistant à se ramener dans un cadre markovien en augmentant le nombre de variables, et j'expliquerai comment l'utiliser pour l'estimation de paramètres clés de la dynamique des plantes avec dormance.

Stéphane Robin, LPSM, Sorbonne Université

Change-point detection in a Poisson process

Joint with Charlotte Dion-Blanc, Sorbonne Université; Émilie Lebarbier, Université Paris-Nanterre

Change-point detection aims at discovering behavior changes lying behind time sequences data. In this paper, we investigate the case where the data come from an inhomogenous Poisson process or a marked Poisson process. We will present an offline multiple change-point detection methodology based on minimum contrast estimator. In particular we will explain how to deal with the continuous nature of the process together with the discrete available observations. Besides, we will select the appropriate number of regimes through a cross-validation

procedure which is convenient here due to intrinsic properties of the Poisson process. Through experiments on synthetic and realworld datasets, we will assess their performances of the proposed method, which is implemented in the CptPointProcess R package. If time permits, we will describe a clustering version of the aforementioned procedure, which enables to classify segments into a limited number of categories corresponding to unobserved underlying behaviors.

Nicolas Verzelen (INRAE, Montpellier)

Optimal univariate change-point detection and Localization

Joint with Magalie Fromont, Matthieu Lerasle, Patricia Reynaud-Bouret

Given a times series Y in $\mathbb{R}^n$, with a piece-wise contant mean and independent components, the twin problems of change-point detection and change-point localization respectively amount to detecting the existence of times where the mean varies and estimating the positions of those change-points. In this work, we tightly characterize optimal rates for both problems and uncover the phase transition phenomenon from a global testing problem to a local estimation problem. Introducing a suitable definition of the energy of a change-point, we first establish in the single change-point setting that the optimal detection threshold is $\sqrt{2loglog(n)}$. When the energy is just above the detection threshold, then the problem of localizing the change-point becomes purely parametric: it only depends on the difference in means and not on the position of the change-point anymore. Interestingly, for most change-point positions, it is possible to detect and localize them at a much smaller energy level. In the multiple change-point setting, we establish the energy detection threshold and show similarly that the optimal localization error of a specific change-point becomes purely parametric. Along the way, tight optimal rates for Hausdorff and l_1 estimation losses of the vector of all change-points positions are also established. Two procedures achieving these optimal rates are introduced. The first one is a least-squares estimator with a new multiscale penalty that favours well spread change-points. The second one is a two-step multiscale post-processing procedure whose computational complexity can be as low as $O(nlog(n))$. Notably, these two procedures accommodate with the presence of possibly many low-energy and therefore undetectable change-points and are still able to detect and localize high-energy change-points even with the presence of those nuisance parameters.