Processus markoviens, semi-markoviens et leurs applications

Date et lieu

5-7 juin 2023 - IMAG Montpellier, France

Programme

--- Lundi 05/06 ---

13h30 Accueil café

Mot d'ouverture

- 14h - 15h : **Irene VOTSI**. *Reliability indicators for hidden Markov and (hidden) semi-Markov models.*
- 15h - 15h20 : **Sophie Lèbre, Charles Lecellier, Laurent Bréhélin.** Apprentissage supervisé de HMM pour la reconnaissance de sites de fixation de facteurs de transcription.
- 15h20 - 15h40 : **Jingqi Zhang, Mitra Foualdirad, Nikolaos Limnios.** A semi-Markov model with geometric renewal processes.

15h40 - 16h10 Pause café

- 16h10 - 16h30 : **Ibrahim Bouzalmat**. Modélisation et estimation des paramètres d'un modèle multi-chaîne cachée de la fièvre typhoïde à Mayotte.
- 16h30 - 16h50 : **Bartholomé Vieille.** Markovian approximation of a hidden Hawkes process for epidemic propagation modelling.
- 16h50-17h10 : **José G. Gómez-García.** Sur la stabilité des modèles CHARME.

--- Mardi 06/06 ---

- 9h - 10h : **Gersende FORT.** Quand les Chaînes de Markov contrôlent la simulation Monte Carlo.

- 10h - 10h20 : **Jean-Baptiste. Durand, Hanna Bacave, Alain Franc, Sandra Plancade, Nathalie Peyrard, Régis Sabbadin.** Modèles de Markov et de semi-Markov cachés multichaînes.

10h20 - 10h50 Pause café

- 10h50 - 11h10: **Nicolas Vergne, Vlad Barbu, Corentin Lorthordé, Caroline Berard.** Modèles dérivants semi-markoviens : estimation, simulation et utilisation à travers le package dsmmR.
- 11h10 – 11h30 : **Nicolas Chopin, Hai-Dang Dau.** On backward smoothing algorithms.
- 11h30 - 11h50 : **Hanna Bacave, Pierre-Olivier Cheptou, Nikolaos Limnios, Nathalie Peyrard.** HMM non paramétriques pilotés par les observations.

12h - 13h30 pause déjeuner (prise en charge par le colloque)

- 13h30 - 14h30 : **Marie-Pierre ETIENNE.** *Les chaînes de Markov cachées : un outil flexible pour analyser le déplacement des animaux et la manière dont ils appréhendent l'espace.*
- 14h30 - 14h50 : **Olivier Gimenez.** Écologie statistique, modèles de Markov cachés et gestion des grands carnivores en Europe.
- 14h50-15h10 : **De Larochelambert Quentin, Hamri Imad, Difernand Audrey, Barlier Kilian, Antero Juliana, Sedeaud Adrien, Toussaint Jean-François, Coulmy Nicolas, Louis Pierre-Yves.** Young performance determines the adulthood performance? A parametric and non-parametric Markovian multi-state development model in French alpine skiing.

15h10 - 15h40 Pause café

- 15h40 - 16h: **David Métivier.** Interpretable hidden Markov model for stochastic weather generation and climate change analysis.
- 16h - 16h20: **Jean-Baptiste Durand, Sandra Plancade.** Impact d'une erreur de spécification dans les H(S)MM multi-réponse - application aux arbres fruitiers.
- 16h20 - 16h40: **Sandra Plancade, Nathalie Peyrard, Nathan Ranc, Nicolas Morellet.** Analyse des activités des cervidés par HSMM à partir de données d'accélérometrie.
- 16h40 - 17h00 : **Ibrahim Kaddouri.** Clustering Hidden Markov Models (HMM) data.

19h30 Dîner au Trinque Fougasse (prise en charge par le colloque)

--- Mercredi 07/06 ---

- 9h -10h : **Odalric-Ambrym MAILLARD**. *Optimal regret minimization strategies in Markov Decision Processes.*
- 10h00 - 10h20 : **Emmanuel Hyon, Alain Jean-Marie**. Les bibliothèques marmote et marmoteMDP.
- 10h20 - 10h40 : **Orlane Le Quellennec**. Un exemple d'optimisation de traitement médical avec modèle incertain.

10h40 - 11h10 Pause café

- 11h10 - 11h30 : **Johannes Wirtz.** Time Reversal of a Markov Chain on trees in a population genetical context.
- 11h30 - 11h50 : **Jean Peyhardi**. Integer autoregressive models based on quasi Po'lya thinning operator.

PRÉSENTATIONS INVITÉES

Reliability indicators for hidden Markov and (hidden) semi-Markov models

Irene VOTSI

(LMM, Le Mans Université)

Abstract. This work concerns different reliability indicators for random repairable and non-repairable systems based on models that are partially observed. Both observation and hidden processes are defined in discrete time and have finite state spaces. Different statistical estimators are proposed and their asymptotic properties are studied. As a particular case, in a semi-Markov framework, an application to real data is presented that describes sustainable vibration levels.

Quand les Chaînes de Markov contrôlent la simulation Monte Carlo

Gersende FORT

(CNRS, Institut de Mathématiques de Toulouse)

Résumé. Les méthodes de Monte Carlo sont des outils numériques consistant à simuler des points approchant une loi "cible" donnée, ceci dans l'optique par exemple de mieux comprendre les zones de forte probabilité de cette loi cible, ou proposer un estimateur de moments sous cette loi qui n'ont pas d'expression explicite. Certains échantillonneurs Monte Carlo sont par nature markoviens : ils consistent en la définition d'une chaîne de Markov ergodique ayant la loi cible comme unique loi stationnaire. D'autres le sont car ils incluent dans leur mécanisme, l'apprentissage en ligne d'un paramètre d'implémentation, apprentissage mené à l'aide des points passés produits par l'algorithme. Par suite, la théorie des Chaînes de Markov est au coeur des études de bien-fondé de beaucoup de méthodes Monte Carlo. La dimension adaptative de certains échantillonneurs Monte Carlo complexifie l'étude théorique, en faisant appel à des résultats sur les chaînes de Markov "contrôlées". Dans cet exposé nous illustrerons la façon dont la théorie des Chaînes de Markov (contrôlées) justifie certains algorithmes de Monte Carlo. Pour ce faire, nous considérerons des exemples d'échantillonneurs de type algorithme de Monte Carlo par Chaînes de Markov adaptatifs, et de type Echantillonnage d'Importance adaptatifs; et considérerons un contrôle dont la dynamique suit une procédure d'Approximation Stochastique.

Les chaînes de Markov cachées : un outil flexible pour analyser le déplacement des animaux et la manière dont ils appréhendent l'espace

Marie-Pierre ETIENNE

(IRMAR, AgroCampus Ouest, Rennes)

Résumé. L'écologie du déplacement consiste à étudier le lien entre des population animales et leur environnement au travers de leur déplacement. On équipe un ou plusieurs individus de capteurs type GPS et on essaie de comprendre et d'analyser leur déplacement. Cette branche du biologging connait actuellement un intérêt grandissant et un développement important dans les modèles de déplacement. Les chaînes de Markov cachées jouent un rôle clé dans ces modèles pour lesquels le temsp est considré comme discret ou conitnu, la chaîne de Markov sous jacente est homogène ou non. L'exposé présentera les résultats récents sur le sujet et les questions soulevées par l'utilisation demodèle en temps discret ou continu.

Optimal regret minimization strategies in Markov Decision Processes

Odalric-Ambrym MAILLARD

(Inria Lille - Nord Europe, CRIStaL)

Abstract. We consider reinforcement learning in a discrete, undiscounted, infinite-horizon Markov Decision Problem (MDP) under the average reward criterion, and focus on the minimization of the regret with respect to an optimal policy, when the learner does not know the rewards nor the transitions of the MDP. In light of their success at regret minimization in multi-armed bandits, popular bandit strategies, such as the optimistic UCB, KL-UCB or the Bayesian Thompson sampling strategy, have been extended to the MDP setup. Despite some key successes, existing strategies for solving this problem either fail to be provably asymptotically optimal, or suffer from prohibitive burn-in phase and computational complexity when implemented in practice. In this work, we shed a novel light on regret minimization strategies, by extending to reinforcement learning the computationally appealing Indexed Minimum Empirical Divergence (IMED) bandit algorithm. Traditional asymptotic problem-dependent lower bounds on the regret are known under the assumption that the MDP is ergodic. Under this assumption, we introduce IMED-RL and prove that its regret upper bound asymptotically matches the regret lower bound. Rewards are assumed light-tailed, semi-bounded from above. Last, we provide numerical illustrations on classical tabular MDPs, ergodic and communicating only, showing the competitiveness of IMED-RL in finite-time against state-of-the-art algorithms. IMED-RL also benefits from a light complexity. We then discuss extensions of these promising strategy to communicating, but not necessarily ergodic MDPs

RESUMES

Apprentissage supervisé de HMM pour la reconnaissance de sites de fixation de facteurs de transcription

Sophie Lèbre (IMAG), Charles Lecellier (IGMM, LIRMM), Laurent Bréhélin (LIRMM)

Résumé. Les HMMs sont largement utilisés en bioinformatique pour la modélisation de séquences protéiques et génomiques [2]. Les problématiques adressées couvrent la segmentation de séquence et la classification supervisée. Ils sont notamment utilisés en routine pour discriminer les familles de domaines protéiques au travers de bases de données telles que Pfam [4] et SuperFamily [6]. Dans ce cas précis, les HMMs utilisés (dénommés HMM profils) sont sans cycles, et sont entraînés sur la base d'un alignement de séquences issues de la famille en question. Dans le cas plus général où la structure du HMM comporte des cycles, l'algorithme classiquement utilisé pour l'entraînement est l'algorithme de maximisation de la vraisemblance de type Expectation-Maximisation dénommé algorithme de Baum-Welch [7]. Quelques approches ont été proposés dans la littérature pour entraîner un HMM de manière supervisé, c-à-d en minimisant l'erreur de classification plutôt qu'en maximisant la vraisemblance [1, 5, 3]. Malgré des performances en théorie meilleures sur les problèmes de classification, ils restent à notre connaissance peu, voire pas, utilisés en bioinformatique. Dans ce travail, on s'intéresse à un algorithme supervisé dédié à l'apprentissage d'un HMM pouvant prédire si une séquence génomique est, ou pas, fixée par un facteur de transcription (TF) particulier. Ces molécules clefs de la régulation de l'expression génique se fixent sur le génome à des positions spécifiques, qu'ils reconnaissent sur la base de la présence de motifs génomiques particuliers. Il existe de l'ordre de 2000 TFs différents dans les cellules humaines, chacun d'eux reconnaissant des éléments génomiques différents. Identifier les éléments qui contrôlent la fixation d'un TF donné est l'un des grands enjeux de la génomique moderne. Dans notre travail, on s'intéresse au rôle des séquences de faible complexité (c-à-d enrichies pour 1 nucléotide ou un k-mer specifique) dans cette reconnaissance. On propose pour cela un HMM cyclique permettant de modéliser la présence d'une région de faible complexité au sein de la séquence génomique : l'objectif est de prédire si la séquence est fixée par le TF d'intérêt en fonction de la longueur de la zone de faible complexité identifiée. Ce HMM est entraîné sur la base d'un ensemble de séquences d'apprentissage labellées (fixée vs. non fixée) grâce à un nouvel algorithme d'apprentissage supervisé.

References

[1] L. Bahl, P. Brown, P. de Souza, and R. Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 11, pages 49–52, April 1986.

[2] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, 1998.

[3] A. Krogh. Hidden Markov models for labeled sequences. In Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5), volume 2, pages 140–144 vol.2, October 1994.

Clustering Hidden Markov Models (HMM) data

Ibrahim Kaddouri
(Université Paris-Saclay)

Abstract. Clustering is a statistical method which aims at grouping data into homogeneous subgroups. It is used as an exploratory tool to detect structure and discover latent variables explaining observed heterogeneity. Clustering is used in all sorts of applications such as marketing, psychology, sociology, bio-informatics and epidemic surveillance [5, 3]. It can be performed on data with dependence structure or not. For independent observations, the problem have been studied extensively. For example, Verzelen et al. have shown in [4] that the miss-classification error of the celebrated K-means clustering algorihm decays exponentially with respect to a signal to noise ratio in the case of observations derived from a mixture of subgaussians. They have also shown that this SNR is optimal in the sense that the misclassification error cannot decay at a faster rate since the Bayes error is lower bounded by the same term up to constants. However, similar results regarding the case of dependent observations have not been proved yet. This is mainly the case for HMMs. When studying clustering, mixture models arise naturally in modelling observations since observations are supposed to belong to different communities. Given a convex combination of unknown probability distributions, it is impossible to retrieve the components of the mixture without prior assumptions on the mixture components or the structure of the observations because the model is not identifiable. A particular situation in which this becomes possible is when observations derive from a hidden Markov model. In these models, the random variables indicating the true labels form a Markov chain. Theoretical guarantees for parametric and nonparametric learning of hidden Markov models have been established recently [1, 2, 6]. The purpose of the work is to study the Bayes risk in this framework by understanding its dependence with respect to the model parameters while keeping a nonparametric modelling of the emission densities. First, after defining the clustering risk appropriately, we identify the limit of the Bayes risk when the number of observations grows to infinity in the offline and online clustering frameworks and identify the convergence rate. Second, we exhibit some bounds on this limiting Bayes risk which allow identifying the appropriate signal-to-noise ratio with respect to which the Bayes risk decays. Third, we control the excess risk of the plug-in classifier using estimates of the model parameters by the errors of estimating the model parameters (the initial distribution ν, the transition matrix Q and the emission densities f0, f1). Finally, using appropriate
estimators, we exhibit a rate on the excess risk. The SNR identified here coincides with the one obtained for clustering iid observations. However, the added value under the HMM modelling is that the model is identifiable in a nonparametric context and one can construct consistent estimators allowing to reach the Bayes risk at least asymptotically. This is not possible in the nonparametric iid framework.

References

[1] Yohann de Castro, Elisabeth Gassiat, and claire Lacour. “Minimax adaptive estimation of non-parametric hidden Markov models”. In: Journal of Machine Learning Research, 17, 11, (43p) (2016).

[2] Yohann de Castro, Elisabeth Gassiat, and Sylvain Le Corff. “Consistent estimation of the filtering and marginal smoothing distributions in non-parametric hidden Markov models”. In: IEEE Trans. Info. th, 63 (8), 4758-4777 (2017).
[3] Mark Gales and Steve Young. “The Application of Hidden Markov Models in Speech Recognition”. In: Foundations and Trends in Signal Processing (Feb 2008).
[4] Christophe Giraud and Nicolas Verzelen. “Partial recovery bounds for clustering with the relaxed Kmeans.”In: Mathematical Statistics and Learning (April 2019).
[5] Diksha Khiatani and Udayan Ghose. “Weather Forecasting Using Hidden Markov Model”. In: 2017International Conference on Computing and Communication Technologies for Smart Nation (Oct 2017). [6] Luc Leh ricy. “Estimation adaptative pour les mod les de Markov cach s non-param triques”. PhD thesis. 2018.

Modélisation et estimation des paramètres d'un modèle multi-chaîne cachée de la fièvre typhoïde à Mayotte

Ibrahim Bouzalmat

Résumé : L'objectif principal de ce travail est de proposer et d'étudier un nouveau modèle épidémiologique qui capture efficacement la dynamique de transmission de la fièvre typhoïde à Mayotte à partir des caractéristiques médicales de cette maladie et d'un jeu de données d'hospitalisations fourni par l'Agence Régionale de Santé. Nous présentons une approche paramétrique qui consiste à utiliser un processus Markovien à deux compartiments pour comptabiliser les personnes exposées et infectées par la maladie, et à estimer les paramètres critiques du modèle, tels que les taux de contamination de personne à personne, contamination par l'environnement, incubation et guérison. Notre méthode d'estimation est innovante, car elle permet de relever deux défis majeurs liés aux données disponibles. Premièrement, les observations ne sont disponibles en temps continu, mais seulement à des dates fixes (hospitalisations journalières), et deuxièmement, à ces dates, le nombre total de personnes infectées n'est pas observé, seuls les cumuls des nouveaux cas déclarés sont comptabilisés, ce qui rend l'estimation des paramètres plus complexe. Pour surmonter les spécificités des données disponibles, des expressions explicites des estimateurs des paramètres sont obtenues en utilisant les moments du processus exposé-infecté. Ensuite, l'algorithme de Baum-Welch est adapté au cas de multi-chaîne de Markov cachée pour estimer ces paramètres. Le modèle et la méthodologie d'estimation proposés ont des implications significatives pour la compréhension et le contrôle de la transmission de la fièvre typhoïde à Mayotte, et potentiellement dans d'autres régions présentant des caractéristiques épidémiologiques similaires. Les résultats de cette étude pourraient contribuer à orienter les politiques de santé publique visant à atténuer la propagation de la fièvre typhoïde à Mayotte.

Markovian approximation of a hidden hawkes process for epidemic propagation modelling

Bartholomé Vieille

Abstract. In recent years, numerous studies grounded on Hawkes processes have been carried out in many fields including finance, biology and social network. Hawkes processes form a class of self-exciting simple point processes. In this communication, I will introduce a markovian approximation of a specific hidden multivariate Hawkes process considered in a spatial setting and used to model the spatio-temporal spread of an epidemic. The spatial domain is composed of multiple disjoint regions. The baseline intensity is time-dependent, and the jump size is constant and equal to 1. Furthermore, the exciting function is a general one. The closed-form expression of the multivariate characteristic function of such a markovian process will be presented. This allows us to obtain a closed-form formula for the temporal structure of the first moments. I will also discuss other points such as parameter estimation of the state-space epidemic model by Sequential Monte-Carlo.

Sur la stabilité des modèles CHARME

José G. Gómez-García

Résumé. Le modèle CHARME (Conditional Heteroscedastic Autoregressive Mixture of Experts) est un modèle de mélange généralisé de séries temporelles ARARCH nonlinéaires et (non)paramétriques. Ce modèle peut modéliser les séries temporelles à changement abrupts de régime comme les électroencéphalogrammes (EEG) et les changements structurels dans les données hydrologiques. Dans cet exposé, nous aborderons la stabilité (ergocicité, stationnarité et dépendance) de ce modèle dans un cadre plus général que ceux des articles [1] et [2] en considérant au moins le processus de changement de régime faiblement dépendant et le nombre de régime inconnu. Nous montrerons quelques simulations et nous discuterons des applications pour les données climatiques.

References

[1] Gómez-García, J.G., Fadili, J. and Chesneau, C. (2021) Learning CHARME models with neural networks. À apparaitre dans Statistical Papers. Arxiv 2002.03237
[2] Stockis, J-P., Franke, J., Tadjuidje Kamgaing, J. (2010) On geometric ergodicity of Charme models. J. Time Series Analysis, 31:141-152

Modèles de Markov et de semi-Markov cachés multichaînes

J.-B. Durand, H. Bacave, A. Franc, S. Plancade, N.Peyrard, R. Sabbadin

Résumé : Les Chaînes de Markov Cachées (CMC) sont très populaires pour modéliser et analyser des processus dynamiques lorsque la variable d'intérêt catégorielle n'est pas observable. Dans de nombreuses applications, il y a plus d'une variable d'intérêt et les chaînes cachées sont en interaction, directe ou indirecte. C'est par exemple le cas pour l'étude de métapopulations ou métacommunautés en écologie, ou pour l'étude de la propagation d'une épidémie dans un réseau. Quelques modèles ont été proposés pour étendre les CMC au cas où plusieurs chaînes sont utilisées pour modéliser la dynamique des variables cachées. Mais l'éventail des structures d'interaction possibles entre les chaînes est plus large que celles formalisées dans ces modèles. Nous proposons

ici une définition du cadre des Modèles de Markov Cachés Multichaines (MMCM) qui apporte un regard unifié sur les modèles de la littérature et permet de formaliser de nouvelles structures d'interaction. Il en résulte une typologie des MMCM définie à partir des relations d'indépendance conditionnelles existant ou non dans le modèle entre variables cachées et observées à l'instant présent, et variables cachées et observées à l'instant précédent. Cette présentation unifiée des MMCM inclut en particulier les CMC factorielles de Ghahramani et Jordan (1997) de même que les modèles de Le Coz et al. (2019) et Toupoulou et al. (2020) et décrit aussi de nouvelles structures. Nous montrons ensuite des exemples de domaines applicatifs pour lesquels ces nouvelles structures sont pertinentes. L'inférence et l'estimation dans les MMCM sont en général plus complexes que pour les CMC. Nous illustrons, sur l'algorithme EM, comment cette complexité dépend de la structure d'interaction et peut rester raisonnable ou devenir rédhibitoire lorsque le nombre de chaînes augmente. Nous discutons des options possibles pour une estimation approchée. Enfin, nous discutons d'extensions du cadre MMCM au cas semi-Markovien, notamment le modèle de Toupoulou et al. (2020), et des questions de modélisation non triviales que cela soulève.

Références

Z. Ghahramani et M. Jordan. Factorial hidden Markov models. In: Advances in Neural Information Processing Systems, vol. 8, 1995.

S. Le Coz, P.-O. Cheptou et N. Peyrard. A spatial Markovian framework for estimating regional and local dynamics of annual plants with dormancy. Theoretical Population Biology, 127, 120-132, 2019.

P. Touloupou, B. Finkenstädt et S.E.F. Spencer. Scalable Bayesian inference for coupled hidden Markov and semi-Markov models. Journal of Computational and Graphical Statistics, 29(2), 238–249, 2020

Modèles dérivants semi-markoviens : estimation, simulation et utilisation à travers le package dsmmR

Nicolas Vergne, V. Barbu, C. Lorthordé, C. Berard

Résumé. Les modèles dérivants semi-markoviens permettent de combiner les particularités des modèles semi-markoviens et des modèles de Markov dérivants. Le modèle inclut donc la loi du temps de séjour dans un état et permet au noyau semi-markovien de varier le long de la séquence. Trois types de modèles dérivants semi-markoviens seront introduits et la construction des estimateurs associés sera présentée. Enfin, le package dsmmR sera utilisé pour l'estimation des modèles dérivants semi-markoviens ainsi que la simulation de données issues de ces modèles.

On backward smoothing algorithm

Nicolas Chopin (et Hai-Dang Dau, Université d'Oxford)

Abstract. In the context of state-space (hidden Markov) models, backward smoothing algorithms rely on a backward sampling step, which by default has a O(N^2) complexity (where N is the

number of particles). An alternative implementation relying on rejection sampling has been proposed in the literature, with stated O(N) complexity. We show that the running time of such algorithms may have an infinite expectation. We develop a general framework to establish the convergence and stability of a large class of backward smoothing algorithms that may be used as more reliable alternatives. We propose three novel algorithms within this class. The first one mixes rejection with multinomial sampling; its running time has finite expectation, and close-to-linear complexity (in a certain class of models). The second one relies on MCMC, and has deterministic O(N) complexity. The third one may be used even when the transition of the model is intractable. We perform numerical experiments to confirm the good properties of these novel algorithms.

HMM non paramétrique pilotés par les observations

Hanna Bacave, Pierre-Olivier Cheptou, Nikolaos Limnios, Nathalie Peyrard

Résumé. Les modèles de Markov cachés (HMM) sont largement utilisés dans de nombreux domaines pour étudier la dynamique d'un processus qui ne peut être observé directement. Cependant, dans certains cas, la structure des dépendances d'un HMM est trop simple pour décrire la dynamique du processus caché. En particulier dans certaines applications en finance et en écologie, les probabilités de transition de la chaîne de Markov cachée peuvent également dépendre de l'observation courante. Dans ce travail, nous nous intéressons à l'extension du HMM classique à cette situation. Nous définissons un nouveau modèle, le modèle de Markov caché piloté par l'observation (Observation-Driven HMM, OD-HMM). Nous présentons ensuite une étude complète du modèle dans le cadre non paramétrique avec des espaces d'états discrets et finis pour les variables cachées et observées. Nous étudions tout d'abord l'identifiabilité du modèle. Puis nous étudions la consistance de l'estimateur du maximum de vraisemblance. Nous établissons ensuite les équations forward-backward associées à l'étape E de l'algorithme EM. La qualité des estimations obtenues est testée sur des jeux de données simulées. Enfin, nous illustrons l'utilisation du modèle sur une application à l'étude de la dynamique des plantes annuelles. Ces travaux posent les bases théoriques et pratiques d'un nouveau cadre qui pourra ensuite être étendu au cas paramétrique, pour simplifier l'estimation, ou au cas semi-markovien pour aller vers plus de réalisme.

Écologie statistique, modèles de Markov cachés et gestion des grands carnivores en Europe

Olivier Gimenez

Résumé. Les grands carnivores (ours, loups, lynx) ont fait leur retour récemment en Europe et en France en particulier. Ces espèces sont à nouveau présentes sur des territoires fortement anthropisés avec souvent des interactions négatives avec les activités humaines. Pour aider à la gestion de ces conflits, il est nécessaire de fournir des informations écologiques aux politiques telles que les zones où se distribue l'espèce et combien d'animaux sont présents sur le territoire. Toutefois, les grands carnivores sont des espèces animales difficiles à suivre sur le terrain : ils sont discrets et vivent à de faibles densités sur de vastes zones. Dans cette présentation, j'illustrerai

comment notre équipe et nos collaborateurs de l'Office Français de la Biodiversité contribuent avec l'écologie statistique à la gestion des grands carnivores en Europe. Je présenterai l'inférence sur l'abondance, la démographie et la distribution des grands carnivores grâce au développement et à l'application des modèles de Markov cachés qui permettent un usage pertinent des données disponibles et fortement bruitées

---

Young performance determines the adulthood performance ? A parametric and non-parametric Markovian multi-state development model in French alpine skiing

De Larochelambert Quentin, Hamri Imad, Difernand Audrey, Barlier Kilian, Antero Juliana, Sedeaud Adrien, Toussaint Jean-François, Coulmy Nicolas, Louis Pierre-Yves

Résumé. Estimating the potential of alpine skiers is an unresolved question, especially because of the complexity of sports performance. Yet, an unsolved question is: the best performers at young age will be the best ones when they get older? Our study aims to objectify this problem and to determine the influence the rate of progression on skier's future performance. We categorized the states of the model by five Performance Lanes (PL), defined as a categorisation of performances by age group .We modelled and estimated the probabilities of performance progression among French alpine skiers in slalom using parametric and non-parametric multi-state Markovian models and a time-inhomogeneous Markov chain to calculate future PL probability. , We calculated the rate of individual performance progression based on a Bayesian nonlinear mixed model and we estimated the impact of this rate of progression on the skiers future PL. The probability to be the first PL (PL1, i.e highest performance by age) at 20 years old is rather similar among the skiers who were at the first or the last PL at 10 years old. These probabilities are more heterogeneous with age. The probability to reach the PL1 at 20 years old according the PL 1,2,3,4 and 5 at 15 years old are 0.13, 0.07, 0.04, 0.02, 0.01 respectively. At 12 years old, the rate of progression is a better estimator of future performance at 20 years old than the PL the skier is. We show that at the youngest age, the probability difference for reaching the highest PL according to the starting PL are rather similar, meaning that performance at the youngest age are not a good predictor of future performance. Yet, such difference increases with age. We also showed that the rate of progression among the youngest is a better predictor of future performance than the skiers PL.

Interpretable hidden Markov model for stochastic weather generation and climate change analysis

David Métivier (CR MISTEA, INRAE Montpellier)

Résumé : The challenges of climate change force industrials to carefully analyze the resilience of their assets to anticipate future weather conditions. In particular, the estimation of future extreme hydrometeorological events, like the frequency of long-lasting dry spells, is critical for hydropower or nuclear generation. Stochastic Weather Generators (SWG) are essential tools to determine these future risks, as they can quickly sample climate statistics from models. They can

be either trained on historical data or from simulated data like expert climate change scenarios. In our work, the SWG described and validated with France historical data is based on a spatial Hidden Markov Model (HMM). It generates correlated multisite rain occurrences and amounts, with special attention to the correct reproduction of the distribution of dry and wet spells. The hidden states are viewed as global climate states, e.g., dry all over France, rainy in the north, etc. The resulting model is fully interpretable. We describe how the model can approximately recover large-scale structure such as North Atlantic Oscillation even if it was trained only on French weather stations. The model achieves very good performances, specifically in terms of extremes, where for example, statistics of drought at the scale of France are well replicated. The model architecture allows easy integration of other weather variables like temperature. In a last part, we show how the model parameters evolve with RCP climate scenarios and the impact of these change on extreme climatic events. This is a joint work with S. Parey (EDF) and E. Gobet (École polytechnique - CMAP).

Impact d'une erreur de spécification dans les H(S)MM multi-réponse - application aux arbres fruitiers.

Jean-Baptiste Durand, Sandra Plancade

Résumé. Dans le cadre de la croissance des arbres, Guédon (2001) a développé un modèle de HSMM multi-réponse, appliqué notamment aux arbres fruitiers par Mezaros (2020), dans lequel les états latents correspondent à des phases de croissance. Ce modèle suppose que les différentes variables observées sur un même arbre sont régies par une dynamique commune, pilotée par les états latents. Dans le cadre d'une étude sur la floraison des pommiers juvéniles, l'analyse exploratoire des données soulèvent des doutes sur la validité de cette hypothèse.
Nous définissons un modèle alternatif potentiel où les dynamiques des variables observées peuvent différer. En simulant sous ce modèle et en estimant selon le modèle classique, nous mettons en évidence les biais d'estimation et les erreurs d'interprétation biologique qui en découlent. Nous proposons ensuite des pistes pour sélectionner entre ces deux modèles.
Cette présentation utilise les travaux du stage de master de Sarah Assaf.

Analyse des activités des cervidés par HSMM à partir de données d'accélérometrie

Sandra Plancade Nathalie Peyrard, Nathan Ranc, Nicolas Morellet

Résumé. Les accéléromètres sont des dispositifs permettant l'enregistrement des mouvements d'un animal (ou humain), dans le but d'inférer ses phases d'activités. On distingue le contexte non-supervisé, dans lequel les phases d'activités sont des variables latentes, et le contexte supervisé dans lequel on enregistre à la fois les mesures d'accélérométrie et les séquences d'activité pour une partie des données (jeu d'apprentissage), avec pour objectif de prédire les séquences d'activités inconnues à partir des mesures d'accélérométrie pour un jeu d'entrainement. Dans le cadre supervisé, la restauration des activités est le plus souvent basée sur une prédiction temps-par-temps par des méthodes de machine learning, parfois suivie d'un lissage par des modèles de markov

cachés (HMM). Néanmoins, l'analyse des résultats se limite à la proportion du temps consacré à chaque activité au cours d'une période, sans inclure l'aspect séquentiel. Notre objectif est d'évaluer le potentiel des modèles de semi-markov cachés (HSMM) pour restaurer les séquences d'activités dans le cadre d'une étude sur le chevreuil. En effet, l'étude des dynamiques d'activités permettrait de répondre à de nouvelles questions biologiques et affinerait la compréhension du comportement de l'animal. Notre étude se situe dans un contexte supervisé où les activités de l'animal (toilettage, marche, course...) pour les séquences du jeu d'entrainement sont extraites de vidéos. Nous considérons un HSMM dans lequel les états cachés sont les activités de l'animal et les observations sont les prédictions par random forest à partir de mesures d'accélérométrie. Le classifieur random forest, les durées de séjour et les distributions d'émissions sont estimées directement sur le jeu d'apprentissage, et les séquences d'activités du jeu de test sont restaurées par l'algorithme de Viterbi. Dans cette présentation, nous évoquerons différents enjeux de l'implémentation de cette approche (qui constitue un travail en cours). En particulier, nous avons mis en évidence l'impact de la fréquence d'échantillonnage sur la restauration par l'algorithme de Viterbi, qui calcule la chaine d'états sous-jacente qui maximise la vraisemblance des observations, pour une valeur donnée des paramètres du modèle. Par des développements approchés de la vraisemblance, nous mettons en évidence l'impact d'une variation de la fréquence d'échantillonnage sur le rapport entre l'adéquation au modèle et l'adéquation aux données, confirmée sur nos données expérimentales. Ainsi, de manière contre-intuitive, dans certaines situations, un sous-échantillonnage des données pourrait conduire à une meilleure restauration. Nous évoquerons également d'autres pistes pour jouer sur cet équilibre entre adéquation au modèle et aux données. Cette question s'étend au delà de l'accélérométrie et concerne toute application des H(S)MM où les observations mesurées en temps discret sont issues d'une variable en temps continu.

A semi-Markov model with geometric renewal processes

Jingqi Zhang, Mitra Foualdirad, Nikolaos Limnios

Résumé. Nous considérons un système dont l'évolution de son état est modélisé par un processus Semi-Markov où les temps de séjour sont issus d'un processus géométrique. Nous supposons qu'il existe deux causes de défaillance, interne ou externe. Les défaillances externes arrivent suivant un processus de Poisson. La défaillance externe est réparable avec une probabilité p. Les défaillances internes sont dues au vieillissement et elles ne sont pas réparables. Après une défaillance non-réparable, le système est remplacé par un système neuf. Après chaque réparation le système se dégrade. La durée d'une réparation augmente avec le nombre de défaillance. Initialement, Pérez-Ocon et Torres-Castro ont étudié ce modèle. Dans ce travail, en utilisant les propriétés standards d'un processus Semi-Markov, quelque soit la loi initiale, nous calculons toutes les mesures de fiabilité telles que la disponibilité, le temps moyen avant la défaillance et l'intensité de défaillance. Pour ce faire, un espace d'état étendu est défini en considérant d'une part les états de fonctionnement et d'autre part les états de défaillance ou réparation. Par la suite le noyau de transition est calculé et les différentes mesures sont déduites.

Les bibliothèques marmote et marmoteMDP

Emmanuel Hyon, Alain Jean-Marie

Résumé. Nous présentons les bibliothèques logicielles marmote et marmoteMDP, respectivement d´edi´ees à la manipulation de chaînes de Markov, et de processus de décision markoviens. Leur intention est de fournir aux scientifiques de toutes disciplines les outils pour construire, analyser formellement, résoudre numériquement ou simuler des modèles markoviens, `a espace d'états discrets, en temps discret ou continu. La bibliothèque marmote, écrite en C++, fournit une interface de programmation pour les chaînes de Markov en g´en´eral. Elle est multiplateforme : Windows, Mac, Linux. La bibliothèque que marmoteMDP, aussi écrite en C++, est une brique logicielle construite au-dessus comme un plugin. Elle est monoplateforme mais dispose d'une interface python. Cette bibliothèque fournit les méthodes spécifiques pour les MDP. C'est un exemple de bibliothèque applicative comme il pourrait y en avoir dédiées aux processus de Markov cach´es, aux mod`eles de particules, de bioinformatique, files d'attente, etc. marmote s'articule autour de quatre abstractions principales. En premier viennent les espaces d'états, discrets et multi-dimensionnels. Puis les structures de transitions, pouvant prendre la forme de matrices mais pas nécessairement. Enfin les concepts de distribution de probabilités et de chaîne de Markov proprement dit. La bibliothèque fournit les outils pour construire et parcourir des espaces d'´etat complexes, ce qui rend la fabrication des matrices de transition/ g´en´erateurs infinitésimaux plus simple et réduit le risque de bugs. Le fait d'abstraire la structure de transition permet également de construire des modèles à espaces d'états infinis. Côté résolution numérique, marmote fournit des méthodes exactes pour les modèles qui ont une solution analytique, et des méthodes numériques, principalement it´eratives, pour le cas g´en´eral. marmoteMDP permet la spécification de processus de décision markoviens avec espaces d'états et d'action discrets en utilisant les abstractions de marmote. Les diff´erents mod`eles proposés couvrent l'éventail des critères d'optimisation : horizon fini ou infini, avec ou sans actualisation et critère moyen. Les méthodes de résolution sont celles de programmation dynamique. Des expérimentations numériques montrent que marmoteMDP résout les modèles plus vite que MDPToolBox, une bibliothèque python populaire actuellement. D'autres expérimentations ont montré la possibilité de calculer une politique optimale pour des grands espaces (voir partie 6 de https://arxiv.org/abs/2104.14879). D'un point de vue utilisation, des exemples sont fournis pour aider à la prise en main. Par ailleurs, notre objectif à court terme est de construire des interfaces R et python qui allieraient rapidité de calcul du C++ et facilité d'utilisation.

Un exemple d'optimisation de traitement médical avec modèle incertain

Orlane Le Quellennec

Résumé. Les maladies humaines telles que le cancer impliquent un suivi à long terme. Un·e patient·e alterne des phases de rémission et de rechutes. Un biomarqueur est monitoré tout au long du suivi. Sa dynamique est modélisée par un processus de Markov déterministe par morceaux (PDMP) contrôlé. Le PDMP évolue en temps et en espace
continus, le processus est partiellement observé à travers un bruit et certains de ses paramètres sont inconnus, ce qui rend le problème du contrôle particulièrement difficile. A notre connaissance, il n'existe pas de méthode pour contrôler un tel PDMP, c'est-à-dire pour maximiser la vie du·de la patient·e tout en minimisant le coût du traitement et les effets secondaires. Dans un premier temps,

nous considérons des dates discrètes uniquement pou les décisions, transformant ainsi le PDMP contrôlé en un processus de décision markovien (à espaces d'états continus) partiellement observé (POMDP). Ensuite, par le biais de simulations, nous comparons les méthodes d'apprentissage par renforcement bayésiennes et non-bayésiennes pour résoudre ce POMDP.

Time Reversal of a Markov Chain on trees in a population-genetical context

Johannes Wirtz

Abstract. In the neutral Moran Model of finite size $n \in \mathbb{N}$, a classic model of population genetics, the stochastic transformations affecting the population can be realized directly on its genealogy and give rise to a Markov Chain on a state space consisting of n-sized binary trees. This process admits time reversal (in the sense of e.g. [Kolmogorov1936] or [Kelly1979]), which makes it possible to simplify the mechanisms determining state changes, and allows for a thorough investigation of the Most Recent Common Ancestor process. Additionally, the construction admits an extension to some non-recurrent versions of the process, such as the path to fixation of a mutation, in the context of both neutrality and selection. Time reversal can again be achieved on the genealogical state space, this time relying on the theory of [Hunt1960]. We will discuss possible implications and end by pointing out some other applications of markovian time reversal of processes in theoretical biology.

Integer autoregressive models based on quasi Po'lya thinning operator

Jean Peyhardi

Abstract. A thinning operator is a stochastic operator that shrinks a random count variable into another one. For instance, consider a population size at a certain time t, then the thinning operator gives the number of survivors at time t + 1. It allows to define integer value autoregressive (INAR) models for count time series, preserving some properties of usual gaussian autoregressive models. The binomial thinning operator is the most popular, it assumes that individuals independently die with the same probability. It has been shown that the Poisson distribution emerges as the natural choice for the residuals, in this case, since the Poisson distribution is preserved at stationarity. The present work proposes an unified approach by introducing the class of quasi P´olya thinning operators. It includes the binomial thinning operator as a special case, but also the hypergeometric, beta binomial, quasi binomial and the new case of quasi beta binomial thinning operator. It is shown that the natural choice for the residual distribution of this INAR(1) model is the class of (additive) modified power series distribution. Therefore, the proposed class of INAR(1) models includes the usual INAR(1) model with Poisson margins as special case, but also those with binomial, negative binomial, generalized Poisson margins or generalized negative binomial margins. It allows to cover a high range of dispersion that are strictly ordered from the binomial case to the generalized negative binomial case. Moreover, usual characteristics of INAR(1) models, such as the auto-correlation function, are described. Asymptotic normality of the maximum likelihood estimator is obtained. The class of INMA(q) based on quasi P´olya thinning

operator is also introduced. Finally, the proposed INAR(1) models are applied on simulated and real datasets.
Key words: Binomial thinning operator; Quasi P´olya distribution; INAR model