Tartu Workshop on Markov Modelling, 28-29 November 2024

Concentration inequalities for some infinite memory models

Paul Doukhan, CY Cergy Paris University
Xiequan Fan, Northeastern University at Qinhuangdao

Markov processes are well known to be represented as iterative random models $X_t = F_t(X_{t-1})$, for some independent sequence of random functions F_t. More general $m-$Markov, represent as solutions of $X_t = F_t(X_{t-1}, ..., X_{t-m})$. One possible extension is that of contractive infinite memory models introduced in Doukhan and Wintenberger (2008), $X_t = F_t(X_{t-1}, X_{t-2}, ...)$. In case F_t are identically distributed a stationary solution is easily obtained and we proved weak dependence conditions in Doukhan and Louhichi (1999).

More recently with Xiequan Fan we derived concentration inequalities for such Markov models through martingale type arguments in Dedecker, Doukhan, Fan (2019) and Alquier, Doukhan, Fan (2022).

The talk dedicates to derive analogue exponential type and concentration inequalities for the case of infinite memory models. Many applications are still to be derived and we propose the study of stochastic algorithms in this setting.

Hidden Markov processes and fields

Evgeny Verbitskiy, Leiden University/University of Groningen

Thermodynamic properties of hidden Markov processes such as Gibbsianity and computation of entropy will be discussed in the talk.

Evolutionary models for sets and networks, that satisfy detailed balance

Chris Watkins; Royal Holloway, University of London

The talk presents extensions of the evolutionary model of [1] to the evolution of sets and networks. This enables modelling of more complex and perhaps more realistic evolutionary scenarios. In addition, an efficient evolutionary algorithm for approximating the infinite-population limit using only small finite populations will be presented. Empirical investigations will be given of the scaling of these algorithms on the evolution of complex networks.

[1] Lember and Watkins, An evolutionary model that satisfies detailed balance, Methodology and Computing in Applied Probability, 2019.

On the robustness of pairwise Kalman predictors

Marc Escudier, Telecom Sudparis

Let us consider two discrete-time processes X and Y, the pair (X, Y) being Gaussian, homogeneous and Markovian. Such models, called "Gaussian homogeneous pairwise Markov models" (GH-PMMs), extend the classical Gaussian homogeneous hidden Markov models (GH-HMMs), also called Gaussian homogeneous state space models. Being significantly more general, GH-PMMs allow still different processing, like Bayesian filtering or smoothing, similar to those used in GH-HMMs, the latter being commonly used for time series processing in many applications such as tracking, epidemiology, weather forecasting and much more. In this presentation, we will compare the robustness of forecasting with both models, by first conducting a theoretical study of the mean squared error of a GH-HMM predictor for data following a GH-PMM. We will then present some experiments on real data, forecasting meteorological variables, to compare the two predictors when the data distribution follows none of the two models. The theoretical study shows that in some cases, forecasting with GH-HMMs instead of GH-PMMs can lead to a high error increase, while the experiments confirm that GH-PMMs based forecasting can be significantly more robust than the GH-HMMs based one.

This is a joint work with Clément Fernandes and Wojciech Pieczynski.

No Viterbi for triplet Markov models — a method to find the most likely sequence of states of a marginal process

Oskar Soop, University of Tartu

A generalization of a hidden Markov model, a triplet Markov model is a triplet $(U, X, Y)(t)$ with the Markov property, where at each time point t, we have Y_t - *class of observations*, X_t - *class of hidden states of interest*, U_t - *class of hidden auxiliary variables*. Maximum A Posteriori estimation problem is to find the most likely sequence of $X_1, \ldots, X_n$, given observations $y_1, \ldots, y_n$. As posterior is Markov and - in case of finite state space - easily obtainable with forward-backward algorithm, the problem really is about finding the most likely sequence of $\hat{X}_1, \ldots, \hat{X}_n$ from Markov chain $(\hat{U}, \hat{X})$. This problem is known to be NP-hard [1,2]. In this talk I will provide an efficient algorithm for solving the problem exactly. If an exact solution is not found within a time limit, the algorithm provides an approximate solution with error bounds.

[1] Joshua Goodman, Parsing inside-out, 1998.
[2] Rune B. Lyngsø, Christian N.S. Pedersen. The consensus string problem and the complexity of comparing hidden Markov models, 2002.

Markov properties of Gaussian processes on metric graphs

David Bolin, King Abdullah University of Science and Technology

There has recently been much interest in Gaussian fields on linear networks and, more generally, on compact metric graphs. We derive an explicit link between Gaussian Markov random fields on metric graphs and graphical models, and in particular show that a Markov random field restricted to the vertices of the graph is, under mild regularity conditions, a Gaussian graphical model with a distribution which is faithful to its pairwise independence graph, which coincides with the neighbor structure of the metric graph. This is used to show that there are no Gaussian random fields on general metric graphs which are both Markov and isotropic in some suitably regular metric on the graph. We then focus on the generalized Whittle-Matérn fields, which is a class of Gaussian processes obtained as solutions to a fractional-order stochastic differential

equation on the metric graph. We show that these fields are Markov only for certain values of the fractional exponent and discuss some implications of these Markov properties.

Can Markov modelling help advance the novel Autofluorescence-Raman technology to diagnose and treat skin and breast cancers?

Alexey Koloydenko; Royal Holloway, University of London

An automated method to detect Basal Cell Carcinoma (BCC), currently trialled by several medical centres, relies on Autofluorescence (AF) imaging guiding Raman microscopy to obtain biochemical information for tissue classification. The guidance is provided via an image segmentation technique aiming to reduce the risk of overlooking cancer. Ongoing efforts to improve performance of the technology include statistical analysis of shape of the AF segments and combining shape data with Raman spectroscopy features. After presenting these ideas and current results, I will share some thoughts on how Markov modelling fits in this context. For example, besides the obvious Eucledian distance measuring proximity of sites and segments in a tissue sample, a suitable shape distance may provide an alternative network structure on which several Markov models may be defined with a view to improving accuracy of tissue classification. This work has been a long term collaboration with biophysics and medical colleagues at the University of Nottingham, which has more recently also involved Professor Jüri Lember and Dr Kristi Kuljus of Tartu University, Estonia.

Posterior analyses and decoding of hidden Markov models with applications to identification and characterization of archaic fragments in human genomes

Asger Hobolth, Aarhus University

In hidden Markov model (HMM) applications the main interest is often to compute summary statistics such as the time spent in a hidden state or the number of visits to a hidden state. Summary statistics can be obtained from global decoding of the HMM (the Viterbi sequence, i.e. the hidden state sequence with the largest posterior probability) or from local decoding (the posterior sequence, i.e. the hidden state with the largest posterior probability in each position). However, these two sequences are often poor representatives for the posterior distribution of the hidden state sequence. It is therefore more useful to compute the distribution of the summary statistics from the full posterior distribution. In the first part of this talk we review the Aston-Martin framework for obtaining pattern distributions from the posterior of the hidden state sequence. We apply the method on a large data set of human genomes from the present. In particular we identify and characterize archaic fragments from past introgression events in modern human genomes. In our application we resolve several computational issues and use simulations to create an efficient framework for the large data set. In the second part of the talk we apply the Lember-Koloydenko-Kuljus framework for improved decoding of an HMM. We demonstrate that hybrid decoding shows increased performance compared to global or local decoding, and we introduce a novel procedure for choosing the tuning parameter in the hybrid algorithm.

This is joint work with Moises Coll Macia, Laurits Skov and Zenia Elise Damgaard Bæk.

A new stochastic order applied to branching random walks

Daniela Bertacchi, University of Milano-Bicocca

We consider general discrete-time branching random walks on a countable set X. According to these processes, a particle at $x \in X$ generates a random number of children and places them at (some of) the sites of X, not necessarily independently nor with the same law at different starting vertices x.

We introduce a new type of stochastic ordering of branching random walks, generalizing the germ order introduced by Hutchcroft in 2022, which relies on the generating function of the process. We prove that given two branching random walks with law μ and ν respectively, with $\mu \geq \nu$, then in every set where there is survival according to ν, there is survival also according to μ. Moreover, in every set where there is strong local survival according to ν, there is strong local survival also according to μ, provided that the supremum of the global extinction probabilities, for the ν-process, taken over all starting points x, is strictly smaller than 1.

New conditions for survival and strong survival for inhomogeneous branching random walks are provided. We also extend a result of Moyal which claims that, under some conditions, the global extinction probability for a branching random walk is the only fixed point of its generating function, whose supremum over all starting coordinates may be equal to 1.

Based on joint work with Fabio Zucca.

Non-homogeneous and hidden Markov multistate models for intermittently observed processes: application to partially observed treatment outcomes among patients with nAMD

Sangita Kulathinal and Dario Gasbarra, University of Helsinki

Markov processes have a wide range of applications. When applying them in a scenario where only intermittent observations regarding state occupancy are available, the underlying dynamics can be captured by hidden Markov models (HMMs). Moreover, in many situations the underlying process is a non-homogeneous Markov process and estimation of the model parameters as well as the hidden states are of interest. Here the transition probabilities and the initial distribution comprise the parameters. The observations themselves may provide complete or partial information about the hidden states. In this paper, we consider a non-homogeneous Markov (nhm) process $\{X(t), t \geq 0\}$ with state space $S = \{1, 2, 3, 4\}$ and eight possible transitions $\{1 \to 2, 1 \to 4, 2 \to 1, 2 \to 3, 3 \to 2, 3 \to 4, 4 \to 1, 4 \to 1\}$. The states are observed only during (some and not all) pre-fixed clinical visits. Additionally, the clinical visits involve another intermittently observed but related process(es) $\{Y(t), t \geq 0\}$ and a binary process which is always observed. Our aims are to estimate the parameters of the nhm model, to predict the hidden state at a future time given the process history and to predict a change in the hidden state in a pre-defined interval. This work is motivated by the partial observations of treatment responses in patients of neovascular Age-related Macular Degeneration (nAMD). The characteristic feature of nAMD is the fluid accumulated under the macula due to leakage from abnormal blood vessels and the target of the treatment is to reduce the fluid. The fluid state is determined from the Optical Coherence Tomography (OCT) image and the vision is assessed using so-called visual acuity (VA, integer-valued ranges between 20 to 85 letters). In clinical practice, VA is measured more frequently than taking OCT images and the treatment needs to administered frequently to maintaining the fluid status low and thereby, the vision good. The interest at each clinical visit is in determining the time when the fluid will appear in the near future. This then helps in adjusting the frequency of the treatment to be administered.

This is an ongoing work with several PhD students and ophthalmologists from the Helsinki University Hospital.

Continuous-time Hidden Markov Model for colorectal cancer – approaches to parameter estimation and common problems

Aapeli Nevala, University of Helsinki

Commonly used approaches rely on the hidden Markov models (HMM) to incorporate true underlying states. Each approach needs to account for the underlying data generating process and related external information, and requires assumptions for estimation.

In this talk, we present Bayesian HMM developed for natural history of colorectal cancer (CRC), combining data on latent disease states from randomised screening study and on observed clinical cancers from the population-based cancer registry. We focus on parameter estimation in Bayesian setting. With our modelling approach and study design provide estimates for latent state occupancy probabilities not only for screening-attenders, but also for the control group and those who never attended screening — despite data on latent states only existing for the attenders.

We particularly focus on the sampling process for this type of model: we present how to use simulation-based calibration to ensure that posterior distributions can be reliably estimated despite the challenges brought in by the sampling scheme. We discuss why one should apply Bayesian methods to obtain parameter estimates. Two algorithms, Hamiltonian Monte Carlo (HMC) and Automatic Differentiation Variational Inference (ADVI), are applied and compared, first by using simulated data and then with a real dataset. Additionally, we show how different sampling mechanisms can yield peculiar correlations in the model posterior despite the marginal parameter estimates looking seemingly similar.

Based on the work by Nevala et al. "Bayesian hidden Markov model for natural history of colorectal cancer: handling misclassified observations, varying observation schemes and unobserved data".

Hidden Markov models for longitudinal data in multistate models: comparing algorithms for parameter and hidden state estimation

Etienne Sebag, University of Helsinki

Hidden Markov Models (HMMs) are powerful tools which can assist in the understanding and analysis of longitudinal data within a multistate modelling framework. Their structure consists of latent (hidden) states that follow the Markov property and observed states that are independent conditioned on the hidden states. These conditional distributions are given via emission probabilities. Specification of the initial distribution completes the parametrization of the HMM, which we denote by a parameter vector $\theta \in R^d$. In statistics, simultaneous estimation of these parameters given the data, as well as prediction of the hidden path (latent states) are central themes. An important area of ongoing research focuses on the investigation of optimal algorithms that can provide accurate parameter estimates and/or path prediction for HMMs in survival (time-to-event) settings while ensuring computational efficiency.

Overall, this work aims to tackle and understand HMM parameter estimation in a bottom-up approach, starting from the most basic kind of model and working upwards to more complex ones as we gain progressive insights. The study is performed on an ecological model setup of capture-recapture data, whereby the latent states represent whether an animal is alive or dead. Fundamentally, this is the simplest case of a survival model, allowing just one transition from the alive state to the dead state. Due to imperfect detection, we do not directly observe whether an animal is alive or not, and rather, only have access to data about the detection status. Hence, we can establish an HMM and perform inference on the transition probability (survival probability) and the detection probability. By comprehending how various algorithms behave and function in the estimation of parameters within this foundational case, we can later expand the model and allow reversible transitions between the latent states, and finally progress to a three-state survival model with an absorbing state.

The goal of the study is therefore to compare various estimation algorithms, both from a frequentist and a Bayesian perspective in order to generate increased knowledge regarding which algorithm is the most appropriate for a specific HMM context. This can pave the way for the development of more sophisticated algorithms (which are perhaps a combination of several methods) that can ultimately handle estimation adequately in more complex HMMs. Comparison of these algorithms will consider the rate of convergence of the algorithms, the bias of the resultant estimates, the computational efficiency and execution time, their performance in handling missing data, and finally their ease of implementation by studying their assumptions. Due to the inherent hidden and observed structure, one natural candidate that will serve as a benchmark for comparison is the Baum-Welch algorithm (which is a special case of the Expectation-Maximization (EM) algorithm). Additionally, this study will also examine the MCEM algorithm (Markov Chain EM), Variational Bayes Inference, more classical MCMC (Markov Chain Monte Carlo) methods, and direct numerical maximization of the (log)-likelihood using quasi-Newton methods like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, among others.

Finally, the project also raises many pertinent questions, including whether more reliable results can be obtained when working with the marginalized likelihood, instead of the joint likelihood consisting of the hidden and observed estimated together. Additionally, it is of significant interest to explore the performance of these algorithms in the prediction of the hidden path, and to indeed assess whether the same algorithms can also be utilized within this prediction purpose. The results of the analysis are then intended to be expanded into broader areas of application. This work is being conducted jointly with Sangita Kulathinal and is in a very early stage.

Behaviour of extremal optimal alignments for Markov chains

Joonas Sova, University of Tartu

The length of the longest common subsequence is a common measure of similarity between two finite sequences. In general there are numerous different alignments which produce a longest common subsequence. All of those alignments are said to be optimal. The distance between the lowest and highest optimal alignments has been proposed as an alternative measure of similarity between two sequences. We show that when two random sequences comprise a Markov chain and the dependence between them is strong in some sense, then this distance grows at most logarithmically. We also show that the asymptotic similarity score (also known as Chvatal-Sankoff constant) can be bounded from above through the entropy rate of the Markov chain.