

Stochastic processes ⁽¹⁾ and statistical methods ⁽²⁾ in mathematical biology

Martina Favero, martina.favero@math.su.se

Mathematical biology is a broad research area in which mathematical developments are driven by the need to understand complex biological processes. In particular, mathematical population genetics and infectious disease dynamics aim to provide mathematical tools to analyse the spread and evolution of populations and diseases.

Within this context, the PhD project can be defined around two distinct but closely related research directions:

(1) the derivation of theoretical properties of stochastic processes (including random trees and graphs, and diffusion processes);

(2) the development and evaluation of modern statistical methods.

The PhD project will be developed in detail together with the successful applicant, considering their interests and skills. The project can focus on only one of the two directions, but developing a project at the interface of both is also encouraged. A brief description of each area is provided below.

(1) Stochastic processes in mathematical biology

While this field is motivated by real-world applications, it has given rise to mathematically rich objects and techniques that have gained their own place within probability theory.

In mathematical population genetics, a key role is played by coalescent processes and Wright-Fisher diffusions. Coalescent models are multitype random trees (or graphs) evolving backwards in time and representing the evolutionary history (genealogy) of a group of individuals. Wright-Fisher diffusions evolve forward in time and represent the dynamics of genetic frequencies. These two classes of processes are related through stochastic duality, which allows information from them to be combined, leading to further insight into the evolution of the underlying population. Both types of processes, as well as their duality relationships, have been generalised to include a wide range of genetic forces, leading to several complex models, including measure-valued diffusions, multiple-merger coalescents and models with spatial structure.

In infectious disease dynamics, the central role is played by an epidemic model describing the spread of a disease in a population, often accompanied by its branching process approximation (in the initial phase of an outbreak, when the population is fully susceptible, the epidemic grows like a branching process). In this setting as well, random trees naturally arise in the form of transmission trees, which encode who infected whom.

Due to the complexity of the stochastic processes involved, deriving explicit expressions for probabilities of interest (e.g. fixation probabilities, the probability of a major outbreak, sampling probabilities) remains an active area of research. When models become intractable (e.g. large-sample or large-parameter regimes), an asymptotic analysis becomes crucial, either to approximate the probabilities of interest or to characterise

scaling limits of the process itself and related fluctuations, often with infinitesimal generator techniques. Another promising research direction consists of expanding the scope of duality, for example by deriving new duality relationships for approximating processes and large-parameter regimes, or by applying duality methods, well established in population genetics, to epidemic models in order to provide new insight on transmission trees.

These brief examples are intended to give a flavour of the types of research problems that could form the focus of the PhD project, further directions and details can be discussed with interested applicants.

(2) Statistical methods for genetic and epidemic data analysis

Due to substantial advances in sequencing technology over recent decades, genetic data sets have become increasingly large, both in terms of sample size and length of sequenced segments. Infectious disease epidemiological data sets have also grown rapidly during and following the COVID-19 pandemic.

The analysis of these data sets requires modern, computationally intensive statistical methods, such as Monte Carlo methods. An active research area seeks to evaluate the efficiency of these methods, and to identify hidden biases, with the aim of improving existing methods and designing new ones. The theoretical toolbox provided by the study of the stochastic processes described in (1) plays a key role in bridging the gap between probability theory and data analysis.

Duality, for example, can be used to build importance sampling algorithms for the estimation of sampling probabilities and to construct exact simulation algorithms for Wright-Fisher diffusions. Scaling limits can be used to evaluate the efficiency of importance sampling algorithms and to approximate intractable likelihood functions. Moreover, theoretical distributions of key objects of interest can be used to benchmark tree- and graph-reconstruction methods (for genealogical trees or transmission trees) with the goal of identifying and correcting hidden biases.

The PhD project may focus on methodological developments along these lines, with applications to real data depending on the applicant's interests.