Statistical inference for stochastic dynamical models

Michael Sørensen

Ever since I was a student, I have been fascinated by stochastic dynamical systems and the problems of modelling and obtaining information about them. In more technical terms, I have been interested in stochastic processes and their statistical inference. My main interest has been stochastic differential equations, which are differential equations influenced by one or more random terms. A solution is a stochastic process, and the random terms typically depend on the present value of this process and increments of Lévy processes (often a Wiener process). Models of this type can be analysed by the powerful tools of stochastic calculus, not least martingale theory. Depending on the nature of the random terms, the solution process can be a continuous process or it can have jumps. We refer to a solution as a diffusion process or a diffusion with jumps, respectively, in these two cases. Suppose a diffusion process has been observed at discrete time points. Then a main challenge for statistical inference is that usually there is no closed form expression for the likelihood function. In the following I'll briefly sketch three problems that I have worked on in recent years.

Diffusion bridges. In the hypothetical situation, where a diffusion process has been observed at all real time points in a finite interval, an expression for the likelihood function often exists (by Girsanov's theorem). One way forward is therefore to think of the more realistic situation with observations at discrete time points as a missing data problem, where we have not observed the values of the process between the discrete observation times. If we can simulate the missing data conditionally on the actual observations, we can apply classical statistical methods for missing data. Because the solution is a Markov process, the missing data are independent diffusion bridges. A diffusion bridge is a solution to a stochastic differential equation in a finite time interval conditional on the values at the two endpoints. Motivated by this idea and other inference problems for diffusions, Mogens Bladt and I have developed methods for simulating diffusion bridges. In the one-dimensional case our approach is easy to explain. We simply simulate two independent diffusions, one moving forward in time from the given value at the left endpoint of the interval, the other moving backwards in time from the given value at the right endpoint. If the two processes meet (which can be shown to usually happen with a reasonable probability), we have a process which connects the two given values at the endpoints like a diffusion bridge. If the two diffusions do not meet, we simply start over and repeat until they meet. This is called a rejection sampler. It can be shown that the resulting bridge process is not a diffusion bridge. However, it is, in a certain sense, not far from being a diffusion bridge (as one would expect from the strong Markov property and the time-reversibility of ergodic one-dimensional diffusions), and the approximate diffusion bridge can be used as input to a (very high dimensional) Markov chain Monte Carlo algorithm that produces diffusion bridges. In the case of multivariate diffusions, the situation is more complicated, because with probability

one two independent diffusions will not meet. Therefore the two diffusions must be suitably dependent, which can be obtained by applying so-called coupling methods developed in the 1980s as a tool to solve problems in probability theory. Presently we work on the problem of non-synchronous sampling of multivariate diffusions, where the coordinates have not been observed at the same time points. This is a considerably more complicated missing data problem.

Estimating functions. I have always been interested in the mathematical structure of statistical problems. An interesting problem is to understand why some estimators for diffusion processes work particularly well. Since the maximum likelihood estimator is difficult to calculate, a considerable number of alternative estimators have been proposed in the literature. Many estimators can be associated with so-called estimating functions, which contain important information about the estimators (and can be used to calculate them). Suppose a diffusion process has been observed at the time points Δi, $i = 1, \ldots, N$, and consider the asymptotic scenario, where $\Delta N \to \infty$ and $\Delta \to 0$ (high frequency asymptotics with infinite time horizon). In this situation, I have found easily verifiable conditions on an estimating function which ensure that the corresponding estimator converges to the true parameter value at the fastest possible rate, and that the asymptotic variance of the estimator is as small as possible. Estimating functions satisfying the conditions can easily be constructed. In joint work with Nina Munkholt Jakobsen this work has been generalized in two directions. In the case where $\Delta N = 1$ and $N \to \infty$ (high frequency asymptotics with bounded time horizon), similar conditions have been obtained by quite different mathematical methods. Here the situation is more complicated because the limit distribution is not a normal distribution. In a recently accepted paper, we have shown that for diffusions with jumps the conditions for rate-optimality and minimal asymptotic variance are considerably more restrictive. For instance, parameters related to the jumps can only be estimated with minimal variance by estimators that are closely related to the maximum likelihood estimator. More generalizations are in the pipeline.

Diffusion processes on the torus. In a third strand of research, diffusion processes have been used as part of a model for the evolution of protein structure. This is joint work with my former postdoc Eduardo García Portugués (who is now in Madrid) and colleagues at the SCIENCE Bioinformatics Centre and University of Oxford. At the Bioinformatics Centre a hidden Markov model of protein structure had been developed, and in order to turn this into a dynamical model that can describe evolution, it was combined with a Markov model for the evolution of the sequence of amino acids, provided from Oxford. However, a model for the evolution of angles was needed too. For each amino acid there are two angles, so Eduardo and I developed suitable models for diffusion processes on the torus. Because our model is only a part of a large and computationally demanding model, a parsimonious model allowing an easily calculated approximation to the likelihood function was needed. In order to fit data on the evolution of proteins, it turned out that the model had to include jumps. In a separate paper we have investigated mathematical and computational problems for diffusion processes on the torus.