Preface

Semi-Markov processes are a generalization of Markov and of renewal processes. They were independently introduced in 1954 by Lévy (1954), Smith (1955) and Takacs (1954), who essentially proposed the same type of process. The basic theory was given by Pyke in two articles (1961a,b). The theory was further developed by Pyke and Schaufele (1964), Çinlar (1969, 1975), Koroliuk and his collaborators, and many other researchers around the world.

Nowadays, semi-Markov processes have become increasingly important in probability and statistical modeling. Applications concern queuing theory, reliability and maintenance, survival analysis, performance evaluation, biology, DNA analysis, risk processes, insurance and finance, earthquake modeling, etc.

This theory is developed mainly in a continuous-time setting. Very few works address the discrete-time case (see, e.g., Anselone, 1960; Howard, 1971; Mode and Pickens, 1998; Vassiliou and Papadopoulou, 1992; Barbu et al., 2004; Girardin and Limnios, 2004; Janssen and Manca, 2006). The present book aims at developing further the semi-Markov theory in the discrete-time case, oriented toward applications.

This book presents the estimation of discrete-time finite state space semi-Markov chains under two aspects. The first one concerns an observable semi-Markov chain Z, and the second one an unobservable semi-Markov chain Z with a companion observable chain Y depending on Z. This last setting, described by a coupled chain (Z, Y), is called a hidden semi-Markov model (HSMM).

In the first case, we observe a single truncated sample path of Z and then we estimate the semi-Markov kernel $\mathbf{q}$, which governs the random evolution of the chain. Having an estimator of $\mathbf{q}$, we obtain plug-in-type estimators for other functions related to the chain. More exactly, we obtain estimators of reliability, availability, failure rates, and mean times to failure and we present their asymptotic properties (consistency and asymptotic normality) as the length of the sample path tends to infinity. Compared to the common use of

Markov processes in reliability studies, semi-Markov processes offer a much more general framework.

In the second case, starting from a truncated sample path of chain Y, we estimate the characteristics of the underlying semi-Markov chain as well as the conditional distribution of Y. This type of approach is particularly useful in various applications in biology, speech and text recognition, and image processing. A lot of work using hidden Markov models (HMMs) has been conducted thus far in these fields. Combining the flexibility of the semi-Markov chains with the advantages of HMMs, we obtain hidden semi-Markov models, which are suitable application tools and offer a rich statistical framework.

The aim of this book is threefold:

- To give the basic theory of finite state space semi-Markov processes in discrete time;
- To perform a reliability analysis of semi-Markov systems, modeling and estimating the reliability indicators;
- To obtain estimation results for hidden semi-Markov models.

The book is organized as follows.

In Chapter 1 we present an overview of the book.

Chapter 2 is an introduction to the standard renewal theory in discrete time. We establish the basic renewal results that will be needed subsequently.

In Chapter 3 we define the Markov renewal chain, the semi-Markov chain, and the associated processes and notions. We investigate the Markov renewal theory for a discrete-time model. This probabilistic chapter is an essential step in understanding the rest of the book. We also show on an example how to practically compute the characteristics of such a model.

In Chapter 4 we construct nonparametric estimators for the main characteristics of a discrete-time semi-Markov system (kernel, sojourn time distributions, transition probabilities, etc.). We also study the asymptotic properties of the estimators. We continue the example of the previous chapter in order to numerically illustrate the qualities of the obtained estimators.

Chapter 5 is devoted to the reliability theory of discrete-time semi-Markov systems. First, we obtain explicit expressions for the reliability function of such systems and for its associated measures, like availability, maintainability, failure rates, and mean hitting times. Second, we propose estimators for these indicators and study their asymptotic properties. We illustrate these theoretical results for the model described in the example of Chapters 3 and 4, by computing and estimating reliability indicators.

In Chapter 6 we first introduce the hidden semi-Markov models (HSMMs), which are extensions of the well-known HMMs. We take into account two types of HSMMs. The first one is called SM-M0 and consists in an observed sequence of conditionally independent random variables and of a hidden (unobserved) semi-Markov chain. The second one is called SM-Mk and differs from the previous model in that the observations form a conditional Markov chain of order k. For the first type of model we investigate the asymptotic properties of the nonparametric maximum-likelihood estimator (MLE), namely, the consistency and the asymptotic normality. The second part of the chapter proposes an EM algorithm that allows one to find practically the MLE of a HSMM. We propose two different types of algorithms, one each for the SM-M0 and the SM-M1 models. As the MLE taken into account is nonparametric, the corresponding algorithms are very general and can also be adapted to obtain particular cases of parametric MLEs. We also apply this EM algorithm to a classical problem in DNA analysis, the CpG islands detection, which generally is treated by means of hidden Markov models.

Several exercises are proposed to the reader at the end of each chapter. Some appendices are provided at the end of the book, in order to render it as self contained as possible. Appendix A presents some results on semi-Markov chains that are necessary for the asymptotic normality of the estimators proposed in Chapters 4 and 5. Appendix B includes results on the conditional independence of hidden semi-Markov chains that will be used for deriving an EM algorithm (Chapter 6). Two additional complete proofs are given in Appendix C. In Appendix D some basic definitions and results on finite-state Markov chains are presented, while Appendix E contains several classic probabilistic and statistical results used throughout the book (dominated convergence theorem in discrete time, asymptotic results of martingales, Delta method, etc.).

A few words about the title of the book. We chose the expression “toward applications” so as to make it clear from the outset that throughout the book we develop the theoretical material in order to offer tools and techniques useful for various fields of application. Nevertheless, because we speak only about reliability and DNA analysis, we wanted to specify these two areas in the subtitle. In other words, this book is not only theoretical, but it is also application-oriented.

The book is mainly intended for applied probabilists and statisticians interested in reliability and DNA analysis and for theoretically oriented reliability and bioinformatics engineers; it can also serve, however, as a support for a six-month Master or PhD research-oriented course on semi-Markov chains and their applications in reliability and biology.

The prerequisites are a background in probability theory and finite state space Markov chains. Only a few proofs throughout the book require elements of measure theory. Some alternative proofs of asymptotic properties of estimators require a basic knowledge of martingale theory, including the central limit theorem.

The authors express their gratitude to Mei-Ling Ting Lee (Ohio State University) for having drawn their attention to hidden semi-Markov models for DNA analysis. They are also grateful to their colleagues, in particular N. Balakrishnan, G. Celeux, V. Girardin, C. Huber, M. Iosifescu, V.S. Koroliuk, M. Nikulin, G. Oprisan, B. Ouhbi, A. Sadek, and N. Singpurwalla for numerous discussions and/or comments. Our thanks go to M. Boussemart, for his initial participation in this research project, and to the members of the statistical work group of the Laboratory of Mathematics Raphaël Salem (University of Rouen), for helpful discussions on various topics of the book. The authors also wish to thank J. Chiquet and M. Karaliopoulou who read the manuscript and made valuable comments. Particularly, we are indebted to S. Trevezas for having tracked mistakes in the manuscript and for our discussions on this subject.

The authors would equally like to thank Springer editor John Kimmel for his patience, availability, advice, and comments, as well as the anonymous referees who helped improve the presentation of this book by way of useful comments and suggestions.

It is worth mentioning that this book owes much, though indirectly, to the "European seminar" (`http://www.dma.utc.fr/~nlimnios/SEMINAIRE/`).

Rouen, France
Compiègne, France
March, 2008

Vlad Stefan BARBU
Nikolaos LIMNIOS