Preface

Over the past fifteen years or so, the domain of statistical and econometric methods for high-frequency financial data has been experiencing an exponential growth, due to the development of new mathematical methods to analyze these data, the increasing availability of such data, technological developments that made high-frequency trading strategies possible, and the correlative need of practitioners to analyze these data. So, the time seems ripe for a book devoted to this topic.

The purpose of this book is to introduce these recent methods and present some of the main new tools available to analyze high-frequency data, taking into account some of the distinguishing properties of financial data and the constraints they impose on the range of conceivable econometric methods. Indeed, from a statistical perspective, the analysis of high-frequency financial data presents a number of *specific characteristics.* As in many other time series settings, we are observing what is assumed to be an underlying continuous-time stochastic process, but on a grid of *discrete times.* Discrete observation of a path implies in particular that we need to make a distinction between the observed discrete increments and the complete path of the underlying process.

However, although observation times are discrete, the time interval Δ between successive observations is small, or very small: the *high-frequency asymptotics* we consider are all based on limiting results where the time interval Δ tends to zero, or equivalently the sampling frequency tends to infinity. By focusing on asymptotics, we make no real attempt toward an analysis of finite or small samples, although this question is mentioned occasionally. This is not to say that the analysis of small samples is unimportant; this is indeed crucial, since in real life the amount of data available is always finite (even if measured in gigabytes when ultra-high-frequency data is concerned). However, the properties of the various estimators or testing procedures in a small sample situation are always specific to each method, and in most cases can be ascertained only through simulation studies rather than through mathematical analysis. In a few instances, we discuss small sample refinements such as small sample bias corrections or Edgeworth expansions. But a useful treatment of small samples would require hundreds of pages, if at all feasible, hence our quasi exclusive focus on asymptotic theory.

Next, in this book we consider only inference problems on a *finite time horizon*, say $[0, T]$, unlike the usual time series asymptotics where T goes to infinity, or mixed asymptotics where both $\Delta \to 0$ and $T \to \infty$. Our rationale for keeping T fixed is twofold: first, high-frequency asymptotics make it possible to say much, although not everything, about the underlying process when observed within a finite time interval; second, in many cases the underlying models we consider fail to have the type of stationarity or ergodic properties that are crucial for long horizon asymptotics, which then requires different tools. One consequence of observing the price path on a finite horizon is the possibility of being subject to the *peso problem*: when jumps have finite activity, there is a positive probability that the path we observe has no jump on $[0, T]$, although the model itself may allow for jumps.

The class of problems we consider has another specific property: not only is the time horizon finite, but we also observe *a single path* of the process; for example a typical question is the estimation of the so-called integrated volatility, in a model with stochastic volatility: we want the integrated volatility over, say, a specific day, which is of course different from the integrated volatility over the next day, and averaging over many days does not make much sense in that case.

Two final distinguishing characteristics of high-frequency financial data are important, and call for the development of appropriate econometric methods. First, the time interval separating successive observations can be random, or at least time varying. Second, the observations are subject to *market microstructure noise*, especially as the sampling frequency increases. Market microstructure effects can be either information or non-information-related, such as the presence of a bid-ask spread and the corresponding bounces, the differences in trade sizes and the corresponding differences in representativeness of the prices, the different informational content of price changes due to informational asymmetries of traders, the gradual response of prices to a block trade, the strategic component of the order flow, inventory control effects, the discreteness of price changes, data errors, etc. The fact that this form of noise interacts with the sampling frequency raises many new questions, and distinguishes this from the classical measurement error problem in statistics.

Before describing in more detail the contents of this book, let us emphasize at the onset some of its general features:

- Our hope is that the book can be useful to econometricians, statisticians, mathematicians and high-frequency practitioners alike, starting at the graduate level. We have assumed basic knowledge of standard probabilistic and statistical principles, but have otherwise attempted to include the prerequisites that fall outside the standard graduate level econometric curriculum. This motivates the existence of the first two parts, which cover the required relevant elements about stochastic processes, convergence, and statistical experiments, plus a brief description of the specific qualitative features of financial data; a knowledgeable reader can skip these, although they do establish the notation we employ in the remainder of the book. Note that Chapter 5 also contains new material, to which the subsequent chapters occasionally refer.

- Because many methods developed in papers rely on different sets of assumptions, we have made a conscious effort to unify our treatment of the available methods by describing them, and their asymptotic properties, under a *common set of assumptions*. As a result, our proofs will often differ from those in the papers, and the results themselves are sometimes weaker, but more often stronger, than what appeared in the original papers.

- Many different problems are presented and, for most of them, many different methods have been proposed in the literature. However different these statistical methods may appear, they (almost) always hinge upon the same basic techniques and basic limit theorems concerning what we call *power variations*. The mathematical results about these power variations and some related functionals are gathered in Appendix A. They are all taken from Jacod and Protter (2011), and proofs are omitted.

- Apart from Appendix A, we have tried to make this book as self-contained as possible, as far as its mathematical content is concerned. This includes relying as little as possible on specific proofs contained in papers, and providing these proofs instead. On the other hand, writing its proof after each result slows down the exposition and tends to obscure the main ideas, at least for a non-mathematician reader. We have thus chosen to write the main part of the text without proofs, except when they are very simple and/or useful for the understanding of what immediately follows. Further, many results are not stated as formal theorems. Nevertheless, since this book is also intended to be a mathematical book, the proofs are given (some of them are new), and they are gathered in Appendix B.

- Since every rule has its exceptions, for a few results we do not give a proof and refer to the appropriate papers. This applies primarily to the results on microstructure noise and on non-equally spaced observations: these are topics in full development now, and the theory is not fully established yet. It also applies in places where we only sketch a description of several different methods which have been proposed for some specific problems.

- We have included an implementation on real data of some estimators or testing procedures described.

- We have tried to be as comprehensive as possible. This said, even under the restrictions imposed by the data (discrete sampling, a single path) and by our choice of methods (high-frequency asymptotics, finite horizon), we cannot pretend to cover all or almost all recent developments on the topic, let alone an exhaustive comparison between the different methods which have been proposed so far, while keeping the length of the book manageable. Inevitably, we did not describe *all* the available methods, and the – necessarily subjective – choice we made might be viewed as biased by our own interests. We apologize in advance to the authors who might feel that we have not done full justice to their work.

- We have left out some topics altogether, for instance forecasting (of volatility and other related quantities for example) and methods for processes driven by fractional Brownian motion or by fractal processes.

The book is divided into four parts, plus the two appendices A and B described above. The first two parts are devoted to preliminary material: the mathematical notions on stochastic processes and especially semimartingales which are necessary to proceed further, and a chapter that explains the specificities of financial data are gathered into Part I, whereas Part II introduces the asymptotic concepts that the analysis in this book relies upon.

Part III deals with estimation of the volatility part of the model, including methods that are robust to market microstructure noise. Part IV is devoted to estimation and testing questions involving the jump part of the model. The practical importance and relevance of jumps in financial data is universally recognized, but only recently have econometric methods become available to rigorously analyze jump processes. The objective of the methods we will describe here is to decide on the basis of statistical tests applied to high-frequency data which component(s) need to be included in the model (jumps, finite or infinite activity, continuous component, etc.) and determine their relative magnitude. We may then magnify specific components of the model if they are present, so that we can analyze their finer characteristics such as the degree of activity of jumps.

We would like to thank our co-authors of the various papers on which this book is partly based. Without them, most of the book would not exist: many thanks to Jianqing Fan, Arnaud Gloter, Tahaki Hayashi, Jia Li, Yingying Li, Loriano Mancini, Per Mykland, Mark Podolskij, Philip Protter, Markus Reiss, Mathieu Rosenbaum, Viktor Todorov, Matthias Vetter, Dacheng Xiu, Nakahiro Yoshida, Jialin Yu and Lan Zhang. We also thank Yaroslav Yevmenenko-Shul'ts for proofreading parts of the manuscript, and Peter Van Tassel and Michael Sockin for research assistance.

Last but not least, we warmly thank Sophie and Hadda for their love, patience and understanding while we wrote this book.

Yacine Aït-Sahalia, Princeton
Jean Jacod, Paris