Special feature: statistics for high-frequency data

Hiroki Masuda¹

Published online: 1 April 2021

© Japanese Federation of Statistical Science Associations 2021

Statistics for high-frequency data is about half-century old, initially launched as a fusional research area based on stochastic analysis and asymptotic statistics. Nowadays, its implication has become much broader compared with that in the early days (70s–80s); thanks to the ever advancing information-processing power in the modern period, people are now able to use a broad spectrum of large-scale and/or high-frequency data having various dependence structures. To statistically handle such data in a practical and interpretable manner, simple adaptations of currently available statistics are not enough.

This special feature consists of nine papers, designed for exhibiting recent works and reviews relevant to statistical modeling of large-scale and/or high-frequency data. Readers will see therein a wide variety of theories, methodologies, and empirical analyses.

Historically, statistics for high-frequency data was developed through a stochastic differential equation (SDE), and more generally, a semimartingale model; rather roughly, a semimartingale is defined to be a sum of diffusion and point processes. Statistical inference for such models can be regarded as that for approximate dynamics in a high-frequency limit; for example, one can think of the diffusion approximation with or without jumps of a Markov step process. The first golden era of development in fundamental theory seems to be from the 80s to 90s. Nevertheless, we still have many related statistical and computational issues, in conjunction with building statistical softwares.

- Jørgensen and Sørensen introduced a class of prediction-based estimating functions for a weak solution to a SDE of the ergodic-diffusion type. Under the high-frequency sampling design, the authors formulated a Monte Carlo method for evaluating asymptotic variance of the proposed estimator. Also provided is a vision for further developments toward optimal estimation in multiple-predictor settings.
- Recently, estimation of a stochastic partial differential equation (SPDE) based on high-frequency data has attracted much attention. Kaino and Uchida considered

✉ Hiroki Masuda
hiroki@math.kyushu-u.ac.jp

¹ Faculty of Mathematics, Kyushu University, 744 Motooka Nishi-ku, Fukuoka 819-0395, Japan

estimation of parametric linear parabolic SPDE with small space-time noise, which has a complicated dependence structure. The authors formulated an adaptive (stepwise) method based on explicit contract functions with theoretical properties.

- Mixed-effects modeling has a long history in statistics, especially in biomedical science. Because it is designed for modeling longitudinal data and because dynamic structure is common in a wide range of applications, it is natural to think of SDE modeling with mixed-effects. However, it is quite recent that systematic studies in this direction began. Delattre wrote a very useful review of the state-of-the-art in relevant theoretical aspects, mentioning an existing software in R.

In the 90s, the terminology “high-frequency data” attracted much attention in econometrics, where one of the important fundamental purposes is estimation of the so-called integrated volatility. The quantity equals the quadratic variation, for which the so-called realized volatility (RV) is directly compatible. The statistic RV is rather easy-to-use and, most importantly, theoretically valid in a great variety of model setups, thereby, triggering and driving the trend in those days. Today, we can see a great deal of extended versions of RV, and this special feature contains some of them.

- Shigemoto and Morimoto presented a framework for predicting an integrated volatility matrix, with a view toward constructing an optimal portfolio. The method utilizes a realized kernel with a graphical lasso, and then a conditional autoregressive Wishart model. Using Nikkei NEEDS-TICK data, the authors conducted thorough empirical analyses. The proposed method is practical. High-dimensional and long-memory extensions are left as.
- Kunitomo and Kurisu considered a detection problem of latent factors hidden in a class of noisy semimartingales with jumps. Based on the SIML (separating information maximum likelihood) method developed earlier, which involves some fine-tuning parameters, the authors proposed how to estimate the characteristic roots and vectors of the latent quadratic variation, followed by numerical experiments and empirical analysis.
- Koike studied estimation of a time-varying lead-lag relationship, together with testing its absence. It is concerned with a method of modeling time delay between a leader and a follower in more than one stochastic process. Empirical studies based on the S&P 500 index and its two derivative products illustrate the theoretical results. Here, and also in some of the above-mentioned papers, the so-called stable convergence in law plays an indispensable role in the construction of an approximate confidence interval.
- Concerned with evaluating market quality by liquidity and volatility, Hayashi and Takahashi proposed to adopt collaborative filtering combined with a regression framework, termed as the regression-based latent factor model (RLFM); here again, RV plays an important role in both estimation and prediction. The proposed methods are demonstrated by extensive empirical analyses of high-frequency limit-order book data from the Tokyo Stock Exchange.

The terminology “high frequency” typically refers to a very fine timescale (possibly finer than a (milli)second), but it is not the only the case. Any space-time massive data may be regarded as high-frequency one, just by changing the time and/or space-time scale; for example, quite informally speaking, daily or weekly time-series data from a very long time ago to the present could be regarded as a high-frequency data. We have two articles in relevant directions.

- Bartoszek *et al.* considered an SIR (susceptible–infected–recovered) type dynamical system, which is piecewise smooth and randomly perturbed by a finite-state Markov chain. In addition to a review of the theory, the authors also discussed how one can simulate high-frequency SIR-model characteristics, where we let the intensity matrix of the driving Markov chain increase.
- Mobile positioning data are large-scale, high-frequency, and dynamic. Iacus *et al.* proposed a simple and robust quantitative method for detecting mobility anomalies in such data sets. It has three user-input tuning parameters, for which the meanings are intuitively clear. An extensive empirical analysis in the case of Italy is presented. I would like to take this opportunity to thank the authors for their effort to make their empirical analyses public.

This fiscal year has suffered under the COVID-19 pandemic. I would like to extend my sincere gratitude to all the authors for their great efforts in this unusual period. I do hope that this special feature will encourage readers’ greater interest in statistics for high-frequency data.

Last but certainly not least, I would like to thank the Editor-in-Chief, Professor Makoto Aoshima, for providing me with this great opportunity of managing this special feature.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.