Point processes and ultra high frequency data: limit order book modeling

Nakahiro Yoshida

Graduate School of Mathematical Sciences, University of Tokyo *

Japan Science and Technology Agency, CREST

1 Ultra high frequency data: sampling and microstructure

It is recognized that the Epps effect is caused by non-synchronicity in sampling and minute structure of trading mechanism. The former was settled by non-synchronous estimation techniques and the latter was approached in statistics by introducing measurement errors and denoising techniques. Lead-lag phenomena[1] are also requesting an explanation by a natural model.

The latest trend of financial statistics is toward analysis of ultra high frequency phenomena by modeling more precise mechanisms in a more precise time-scale. There is no Brownian motion as a driving process of the system since the central limit theorem is not effective at this level of description, differently from the standard framework.

Developments in measurement and storage technologies are enhancing feasibility of continuous-observation paradigm, and statistical inference for stochastic processes is advancing back to there. Point process modeling gives a promising approach to a description of microstructure. In this talk, we shall discuss some topics in statistical modeling with point processes.

2 Point process regression model and quasi likelihood analysis

In order to construct price models and limit order book models, we proposed a point process regression model (PPRM). The point process regression model can express asynchronicity of observations, lead-lag effects and microstructure. This model can incorporate nonstationarity

*Graduate School of Mathematical Sciences, University of Tokyo: 3-8-1 Komaba, Meguro-ku, Tokyo 153-8914, Japan. e-mail: nakahiro@ms.u-tokyo.ac.jp

[1] cf. [2], [1].

under finite time horizon and self-exciting/self-correcting effects of the point processes as well as exogenous effects.

Suppose that a d-dimensional point process $N^n = (N^{n,\alpha})_{\alpha\in\mathcal{I}}$, $\mathcal{I} = \{1, ..., \mathsf{d}\}$, on the time interval $I = [T_0, T_1]$ has a d-dimensional intensity process $n\lambda^n(t,\theta)$ admitting a representation

$$\lambda^n(t,\theta) \quad = \quad g^n(t,\theta) + \int_{\hat{T}_0}^{t-} K^n(t,s,\theta)dX^n_s, \tag{1}$$

where θ is a statistical parameter. The intensity process of N^n refers to the functional covariate processes g^n and K^n, and a covariate process X^n.[2] More precisely, on a stochastic basis $\mathcal{B} = (\Omega, \mathcal{F}, \mathbf{F}, P)$, $\mathbf{F} = (\mathcal{F}_t)_{t\in\hat{I}}$, $\hat{I} = [\hat{T}_0, T_1] \supset I$, for each $n \in \mathbb{N}$ and $\theta \in \Theta$, $(g^n(t,\theta))_{t\in I}$ is a d-dimensional predictable process, $(K^n(t,s,\theta))_{s\in[\hat{T}_0,t)}$ is a $\mathsf{d}\times\mathsf{d}_0$-matrix valued process on I, $\mathcal{I}_0 = \{1, ..., \mathsf{d}_0\}$, and $(X^n_t)_{t\in\hat{I}}$ is a d_0-dimensioanl $\mathbf{F}$-adapted càd increasing process. The multivariate point process N^n is compensated by $(\int_{T_0}^t n\lambda^n(s,\theta)ds)_{t\in I}$ when θ is the true value.

The quasi likelihood analysis (QLA) is a systematic analysis of the quasi likelihood random field and the associated estimators, with a large deviation method that derives precise tail probability estimates for the random field and estimators ([4]). For the PPRM (1), in high activity as $n \to \infty$, the quasi likelihood analysis (QLA) was constructed in [3]. In summary,

(i) Statistics is non-ergodic in general

(ii) The quasi likelihood is locally asymptotically mixed normal.

(iii) A polynomial type large deviation estimate for the quasi likelihood random field is obtained.

(iv) The quasi maximum likelihood estimator and the quasi Bayesian estimator are asymptotically mixed normal.

(v) Convergence of moments of these estimators holds.

These results apply to the Hawkes type processes, i.e., $X^n = n^{-1}N^n$. Generalization to marked point processes is straightforward.

3 Modeling limit order book

The non-ergodic model (1) can be interpreted as a collection of local models on time intervals $[t, t + n^{-1/2}]$, and as the first step, it is important to model the local model. The authors are analyzing limit order book data by modeling intensities of point processes of limit order, market order and cancellation. [3] We found dependency of the intensity processes on various covariates. Among them, a limit order intensity model will be presented in the talk.

[2]The integral part can be included in $g^n(t,\theta)$. However the present expression is more convenient for models of Hawkes type.

[3]Joint work with Ioane Muni Toke.

4 Other topics

(i) A related topic is a nonparametric method for estimating covariation between latent intensity processes.

(ii) A similar formulation of QLA is possible when the time horizon $T_1 \to \infty$, instead of increasing the intensities. Then establishing ergodicity of point processes becomes an issue.

(iii) Lead-lag modeling by point processes is in progress.

References

[1] Abergel, F., Huth, N.: High frequency lead/lag relationships. Empirical Facts (2012)

[2] Hoffmann, M., Rosenbaum, M., Yoshida, N.: Estimation of the lead-lag parameter from non-synchronous data. Bernoulli **19**(2), 426–461 (2013). DOI 10.3150/11-BEJ407. URL http://dx.doi.org/10.3150/11-BEJ407

[3] Ogihara, T., Yoshida, N.: Quasi likelihood analysis for point process regression models, preprint (2015)

[4] Yoshida, N.: Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. Ann. Inst. Statist. Math. **63**(3), 431–479 (2011). DOI 10.1007/s10463-009-0263-z. URL http://dx.doi.org/10.1007/s10463-009-0263-z