Journal of Econometrics 194 (2016) 242–262

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Between data cleaning and inference: Pre-averaging and robust estimators of the efficient price☆

Per A. Mykland[a,*], Lan Zhang[b]

[a] The University of Chicago, United States

[b] University of Illinois at Chicago, United States

ARTICLE INFO

Article history:
Available online 1 June 2016

JEL classification:
C01
C02
C13
C14
C22
G11

Keywords:
Consistency
Cumulants
Contiguity
Continuity
Discrete observation
Efficiency
Equivalent martingale measure
High frequency data
Jumps
Leverage effect
M-estimation
Medianisation
Microstructure
Pre-averaging
Realised beta
Realised volatility
Robust estimation
Semi-martingale
Stable convergence

ABSTRACT

Pre-averaging is a popular strategy for mitigating microstructure in high frequency financial data. As the term suggests, transaction or quote data are averaged over short time periods ranging from 30 s to five min, and the resulting averages approximate the efficient price process much better than the raw data. Apart from reducing the size of the microstructure, the methodology also helps synchronise data from different securities. The procedure is robust to short term dependence in the noise.

Since averages can be subject to outliers, and since they can pulverise jumps, we have developed a broader theory which also applies to cases where M-estimation is used to pin down the efficient price in local neighbourhoods. M-estimation serves the same function as averaging, but we shall see that it is safer. Good choices of M-estimating function greatly enhance the identification of jumps. The methodology applies off-the-shelf to any high frequency econometric problem.

In this paper, we develop a general theory for pre-averaging and M-estimation based inference. We show that, up to a contiguity adjustment, the estimated process behaves as if one sampled from a semimartingale (with unchanged volatility) plus an independent error.

Estimating the efficient price is a form of pre-processing of the data, and hence the methods in this paper also serve the purpose of data cleaning.

1. "A tale full of sound and fury"

The recent literature on high frequency financial data has indeed been focused on sound (noise) and fury (jumps). While the tale is significant and important, one of the lessons from it is that both noise and jumps can severely impact *statistical* significance. Especially when they occur in combination.[1]

☆ We are grateful to for helpful comments and suggestions from Gib Bassett, Rong, Chen, Nour Meddahi, Neil Shephard, Kevin Sheppard, Dan Christina Wang, Jing Yu, and the Referees for *Journal of Econometrics*, as well as seminar participants in Chicago, Geneva, London, Oxford, Paris, and Zürich. We would also like to thank the Oxford-Man Institute, where the part of idea for this paper was hatched. Financial support from the National Science Foundation under grants SES 11-24526, DMS 14-07812, and DMS 14-07820 is gratefully acknowledged.

* Corresponding author.
E-mail address: per.mykland@gmail.com (P.A. Mykland).

[1] See, in particular, the discussions in Jacod and Protter (2012, Chapter 16.5, pp. 521-563) and Aït-Sahalia and Jacod (2014, Appendix A.4, p. 496-502).

http://dx.doi.org/10.1016/j.jeconom.2016.05.005

Unlike Shakespeare's Macbeth, we are fortunately not here faced with ultimate questions, but rather with the more prosaic one of finding a signal – something significant – in the middle of the sound and fury. The purpose of this paper is to introduce two (intertwined) approaches which we believe can be helpful: M-estimation, and contiguity.

The analysis of these data started with the work of Andersen and Bollerslev (1998a,b), Andersen et al. (2001, 2003), Barndorff-Nielsen and Shephard (2001, 2002), Barndorff-Nielsen (2004), Jacod and Protter (1998), Zhang (2001) and Mykland and Zhang (2006), and the group at Olsen and Associates (Dacorogna et al. (2001)), focusing on the concept of *realised volatility* (RV).[2] The work was based on the assumption that log prices follow a semimartingale of the form

$$dX_t = \mu_t dt + \sigma_t dW_t + dJ_t, \tag{1}$$

where J_t is a process of jumps.[3] W_t is Brownian motion; μ_t and σ_t are random processes that can be dependent with W. We also denote the continuous part of X_t by

$$dX_t^c = \mu_t dt + \sigma_t dW_t. \tag{2}$$

The semimartingale model for prices is required by the no-arbitrage principle in finance theory (Delbaen and Schachermayer, 1994, 1995, 1998).

Somewhat startlingly, the data had feedback to the theory: log prices are not semimartingales after all. The authors found that in actual data, the *RV* does not, in fact, converge as predicted by theory. This was clarified by the so-called *signature plot* (introduced by Andersen et al. (2000), see also the discussion in Mykland and Zhang (2005)). This led researchers to investigate a model where the efficient log price X_t is latent, and one actually observes a contaminated process Y_{t_j}:

$$Y_{t_j} = X_{t_j} + \epsilon_{t_j}. \tag{3}$$

The distortion ϵ_{t_j} is called either "microstructure noise" or "measurement error", depending on one's academic field (O'Hara, 1995; Hasbrouck, 1996). The t_j can be transaction times, or quote times.

The discovery of the impact of microstructure on inference led researchers to seek methods for high frequency data which allow for such noise. So far, five main approaches have come to light:

- Two- and Multi-scale estimation: weighted subsampled RVs (Zhang et al., 2005; Zhang, 2006, 2011)
- Realised Kernel: weighted autocovariances (Barndorff-Nielsen et al., 2008)[4]
- Pre-averaging: take weighted local averages before taking squares (Jacod et al., 2009a; Podolskij and Vetter, 2009b)
- Quasi-likelihood (Xiu, 2010)
- The local method of moments of Bibinger et al. (2014).

All methods can achieve up to $O_p(n^{-1/4})$ convergence rate for volatility, which is as good as for parametric inference (σ, μ constant), cf. Gloter (2000), Gloter and Jacod (2000, 2001).[5] The approaches mainly differ in treatment of edge effects. (See Mykland and Zhang, 2014 for a systematic discussion of edge effects.) Studies based on different microstructure models are also in development (Robert and Rosenbaum, 2009). A recent, more abstract, line of enquiry is based on equivalence of experiments (Hoffmann, 2008; Reiss, 2011; Jacod and Reiss, 2014; Bibinger et al., 2014). The latter path is related to our own; see Example 3 in Section 3.1.1.

However, existing literature has been confined to estimation of volatility and very closely related objects.[6] Also each estimator has been studied on a case by case basis. This is in contrast to the much greater generality which can be achieved when there is no microstructure, including high frequency regression, analysis of variance, powers of volatility (Mykland and Zhang, 2006, 2009; Kalnina, 2012; Jacod and Rosenbaum, 2013), empirically based trading strategies (Zhang, 2012), semivariances (Barndorff-Nielsen et al., 2009b), resampling (Kalnina and Linton, 2007; Gonçalves and Meddahi, 2009; Kalnina, 2011; Gonçalves et al., 2013), volatility risk premia (Bollerslev et al., 2011, 2009), the volatility of volatility (Vetter, 2011), robust approaches to volatility,[7] jump detection and estimation,[8] and so on. In other words, the research assuming no microstructure has flourished. To some extent, this is legitimate. As an old saying puts it, one has to learn to walk before one learns how to run. Also, there is the hope that either subsampling or pre-averaging can be used to eliminate the microstructure problem, and/or that data can be cleaned so hard that they do not have error any more. Even with this latter strategy, however, it is difficult to assess the impact of microstructure noise without including it in the model. Data processing, such as subsampling or pre-averaging, may also distort the jump characteristics of the data, and thus adversely affect subsequent inference.

This raises the question of whether we as a community will have to redo everything on an estimator-by-estimator basis for more realistic models that allow for microstructure noise and/or jumps.

The purpose of this paper is to find a way around this gargantuan task. We characterise the price process with sound and fury in presence. We develop a general theory that asymptotically separates the impact of the continuous evolution of a signal (i.e. latent efficient price), of the jumps, and of the microstructure. The theory covers both pre-averaging and M-estimation. On the one hand, our theory reduces the impact of microstructure, irrespective of the target of estimation. Our approach will not solve all problems for going between the noise and no-noise cases, but it is a step in the direction of typing these two together. On the other hand, our theory does not truncate jumps before analysis, and we show that we can tightly control the degree of modification of jumps when using a suitable M-estimator preprocessing before analysis. Thus the inference is transparent about how jump characteristics play a role in inference, again regardless of the "parameters".

We have two main clusters of results. One is Theorems 1–4 in Section 2.5, which show that by moving from pre-averaging to pre-M-estimation, one can to a great extent avoid the pulverisation of

[2] An instantaneous version of RV was earlier proposed by Foster and Nelson (1996) and Comte and Renault (1998). Antecedents can be found in Rosenberg (1972), French et al. (1987) and Merton (1980). For a number of other early papers, see the anthology (Shephard, 2005). For further references, see the review by Shephard and Andersen (2009).

[3] Some of the cited papers allow for jumps, others not.

[4] Realised kernel and Multi-scale estimation can be given adjustments to be asymptotically equivalent, see Bibinger and Mykland (2016).

[5] Other earlier methods based on parametric assumptions include, in particular, (Zhou, 1998; Curci and Corsi, 2005), which uses the famous parameter-free diagonalisation of the covariance matrix.

[6] Specifically Bi- and Multipower Variation (Podolskij and Vetter, 2009a; Jacod et al., 2009b) and integrated covariance under asynchronicity (Zhang, 2011; Barndorff-Nielsen et al., 2009a; Christensen et al., 2008a). The only other main classes of estimators that have been studied in the presence of noise are jump (see Footnote 8) and leverage effect (Wang and Mykland, 2014; Aït-Sahalia et al., 2013).

[7] In addition to the other papers cited, see, e.g., Andersen et al. (2012, 2014).

[8] References include Barndorff-Nielsen (2004), Aït-Sahalia (2004), Mancini (2004), Barndorff-Nielsen et al. (2006), Aït-Sahalia and Jacod (2007, 2008, 2009, 2012), Jacod and Todorov (2010), Jing et al. (2012), Lee (2005), Lee and Mykland (2008). (Huang and Tauchen, 2005; Fan and Wang, 2005; Jacod and Protter, 2012; Lee and Mykland, 2012; Aït-Sahalia and Jacod, 2014) do consider microstructure in connection with jumps.