ARTICLE IN PRESS

Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

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

Inference from high-frequency data: A subsampling approach[☆]

K. Christensen [a,*], M. Podolskij [b,a], N. Thamrongrat [c], B. Veliyev [a]

[a] CREATES, Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark

[b] Department of Mathematics, Aarhus University, Ny Munkegade 118, 8000 Aarhus C, Denmark

[c] Institute of Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form
15 April 2016
Accepted 18 July 2016
Available online xxxx

JEL classification:
C10
C80

Keywords:
Bipower variation
High-frequency data
Microstructure noise
Positive semi-definite estimation
Pre-averaging
Stochastic volatility
Subsampling

ABSTRACT

In this paper, we show how to estimate the asymptotic (conditional) covariance matrix, which appears in central limit theorems in high-frequency estimation of asset return volatility. We provide a recipe for the estimation of this matrix by subsampling; an approach that computes rescaled copies of the original statistic based on local stretches of high-frequency data, and then it studies the sampling variation of these. We show that our estimator is consistent both in frictionless markets and models with additive microstructure noise. We derive a rate of convergence for it and are also able to determine an optimal rate for its tuning parameters (e.g., the number of subsamples). Subsampling does not require an extra set of estimators to do inference, which renders it trivial to implement. As a variance–covariance matrix estimator, it has the attractive feature that it is positive semi-definite by construction. Moreover, the subsampler is to some extent automatic, as it does not exploit explicit knowledge about the structure of the asymptotic covariance. It therefore tends to adapt to the problem at hand and be robust against misspecification of the noise process. As such, this paper facilitates assessment of the sampling errors inherent in high-frequency estimation of volatility. We highlight the finite sample properties of the subsampler in a Monte Carlo study, while some initial empirical work demonstrates its use to draw feasible inference about volatility in financial markets.

1. Introduction

Volatility is a key ingredient in the assessment and prediction of financial risk, be it in asset- and derivatives pricing (e.g., Black and Scholes, 1973; Sharpe, 1964), portfolio selection (e.g., Markowitz, 1952), or risk management and hedging (e.g., Jorion, 2006).

Around the turn of the millennium, the advent of financial high-frequency data led to a surge in the nonparametric measurement of volatility (see, e.g., Andersen et al., 2010; Barndorff-Nielsen and Shephard, 2007). High-frequency data are recorded at the tick-by-tick level and store information about the time, price (i.e., a bid–ask quote or transaction price), and size of individual orders and executions. In theory, the harnessing of high-frequency information leads to a perfect, error-free measure of ex-post volatility via the realized variance; a sum of squared intraday log-returns (e.g., Andersen and Bollerslev, 1998; Barndorff-Nielsen and Shephard, 2002).

After the initial – pioneering – work, the literature turned toward addressing two inherent shortcomings of realized variance. Firstly, realized variance can only estimate quadratic variation, and it does not separate continuous, diffusive volatility from discontinuous jump risk. This motivated the development and application of estimators that can robustly measure very general functionals of volatility, also in the presence of jumps (e.g., Aït-Sahalia and Jacod, 2012; Andersen et al., 2012; Barndorff-Nielsen and Shephard, 2004; Barndorff-Nielsen et al., 2006;

☆ We thank Yacine Aït-Sahalia (co-managing editor), Ilze Kalnina, Per Mykland, two anonymous referees, participants at the "Market Microstructure and High-Frequency Data" meeting in Chicago, USA; the "High-Frequency Financial Econometrics" conference in Barcelona, Spain; the AHOI workshop at Imperial College, London, UK; the Dynstoch workshop in Lund, Sweden; the "Econometrics of High-Dimensional Risk Networks" conference at the Stevanovich Center in Chicago, USA; the "9th International Conference on Computational and Financial Econometrics" in London, UK; the "9th Annual SoFiE Conference" in Hong Kong, China and in seminars at U. of Padova, Venice and Verona for helpful comments and suggestions. Christensen and Podolskij appreciate research funding from the Danish Council for Independent Research (DFF – 4182-00050). In addition, Podolskij thanks the Villum Foundation for supporting his research on "Ambit fields: Probabilistic properties and statistical inference". Thamrongrat gratefully acknowledges money from Deutsche Forschungsgemeinschaft through the Research Training Group (RTG – 1953). This work was also supported by CREATES, which is funded by the Danish National Research Foundation (DNRF78).

* Corresponding author.
E-mail address: kim@econ.au.dk (K. Christensen).

http://dx.doi.org/10.1016/j.jeconom.2016.07.010

Please cite this article in press as: Christensen, K., et al., Inference from high-frequency data: A subsampling approach. Journal of Econometrics (2017), http://dx.doi.org/10.1016/j.jeconom.2016.07.010

Christensen et al., 2010; Corsi et al., 2010; Mancini, 2009; Mykland et al., 2012). Secondly, when applied to tick-by-tick data realized variance is severely biased by common sources of noise, which form an integral part of any realistic model for securities' prices (e.g., Hansen and Lunde, 2006b; Zhou, 1996). This paved the way for the next cohort of estimators that were designed to be more resistant to noise, e.g., Aït-Sahalia et al. (2005), Barndorff-Nielsen et al. (2008), Jacod et al. (2009) and Podolskij and Vetter (2009a); Zhang (2006). As such, much progress has been made and today there is no shortage of estimators that can provide consistent estimates of volatility functionals in various contexts (either plain vanilla, or robustly to jumps or noise—or both).

The large battery of estimators at our disposal also brings with it an increasing demand for assessing estimation errors and drawing inference about volatility—e.g., in the form of confidence intervals or hypothesis tests. This is because whether the sample is small or large, as long as it is finite, there is necessarily some sampling error left in the estimate, and when confidence intervals are computed in practice, high-frequency estimators of volatility are often found to contain sizable errors (e.g., Barndorff-Nielsen et al., 2008). The distinction between a realized measure of volatility and its population target is critical, because failing to properly take sampling uncertainty into account can severely distort parameter estimation of stochastic volatility models and be detrimental to the construction and evaluation of forecasts of volatility (e.g., Andersen and Bollerslev, 1998; Andersen et al., 2005, 2011; Hansen and Lunde, 2006a, 2014; Patton, 2011).

There are several problems associated with drawing inference about volatility functionals in high-frequency data. The first and foremost is of course to figure out the relevant distribution theory. The next hurdle is then to find a good proxy for the asymptotic (conditional) variance of the estimator. This is a formidable challenge in practice, because the asymptotic variance often depends on parameters that are substantially more difficult to back out from the available sample of high-frequency data. The expression for the asymptotic variance typically also rests heavily on the properties of the data and it is bound to change depending on these. This is an unpleasant concern with real high-frequency data, which are contaminated by market microstructure frictions. While the noise is often assumed to be i.i.d. and independent of the efficient price, there is some empirical and theoretical support for a serially correlated, heteroscedastic, and, potentially, endogenous noise process at the tick level (e.g., Aït-Sahalia et al., 2011; Diebold and Strasser, 2013; Hansen and Lunde, 2006b; Kalnina and Linton, 2008). An estimator of the asymptotic variance designed for i.i.d. and exogenous noise cannot be expected to give valid inference, if the underlying conditions are violated. In practice, it is not trivial to verify the conditions imposed on the noise (e.g., Hautsch and Podolskij, 2013), which makes it more pressing to find estimators that are robust against modeling criteria. Finally, in multivariate analysis, inference would at some stage require an estimate of the asymptotic covariance matrix. Here, the proposed estimator should ideally be positive semi-definite, while, in contrast, some existing estimators of the asymptotic covariance matrix in the high-frequency setting are not assured to be that. As we show in this paper, this runs smack into problems in practically relevant and realistic settings (see Table 1 in Section 4).

In this paper, we propose to use subsampling for assessing the uncertainty embedded in high-frequency estimation of functionals of financial volatility. Subsampling is based on creating several – properly rescaled – estimates of the parameter(s) of interest using local stretches of sample data and then studying the sampling variation of these. It was originally developed in the context of stationary time series in the long-span domain (e.g., Politis and Romano, 1994; Politis et al., 1999). The term appeared in the high-frequency literature in Zhang et al. (2005), who proposed a two-scale realized variance based on price subsampling. This is different from traditional subsampling and actually does not work for asymptotic variance estimation, because it leads to an overlapping samples problem in the subsampled returns, causing the subsample estimates to be too strongly correlated in large samples. This was pointed out by Kalnina and Linton (2007) and Kalnina (2011), who propose an inference strategy based on various alternative subsampling schemes, which lead to better asymptotic properties. Kalnina (2015) extends these ideas to inference about a multivariate parameter, while Ikeda (2016) and Varneskov (2016) consider subsample estimation of the asymptotic variance of the realized kernel.

As an inferential tool, subsampling has several attractive features from a practical point of view. First, subsampling is intuitive and relatively easy to compute, because it does not require an extra set of estimators; it uses copies of the original statistic. Second, in the multivariate context, it leads to variance–covariance matrix estimates that are positive semi-definite by construction. And, third, subsampling does not explicitly take the structure of the asymptotic variance into account. It is to a large degree automatic and has an innate ability to adapt to the problem at hand, which makes it highly robust against design criteria, as shown by Kalnina (2011). This type of analysis, where inference is effectively carried out by bypassing the asymptotic variance, is also emphasized by Mykland and Zhang (forthcoming), who propose a so-called Observed Asymptotic Variance, which, as our approach, is based on the comparison of adjacent estimators.

This paper builds on these ideas. It contributes to extent literature in several directions. First, we propose to subsample bipower variation as a means to estimate the asymptotic variance–covariance matrix of this statistic. We devise an estimator, which involves fewer tuning parameters compared to Kalnina (2011). Second, we derive an asymptotic theory within this framework in both frictionless and noisy markets. We show our estimator is consistent under weak assumptions on the data-generating process, accommodating jumps in the price and volatility, while allowing the noise to be either heteroscedastic or autocorrelated. Third, with stronger conditions, we provide a decomposition of the leading errors of the subsampler, from which we get insights about how to configure it by optimally choosing its tuning parameters (e.g., the number of subsamples). This yields a rate of convergence for our statistic; a result that has – to the best of our knowledge – not been derived in earlier work. It reveals that the robustness of subsampling is not free of charge, but leads to a loss of efficiency compared to existing estimators in the form of a slower rate of convergence. It implies a trade-off in that if, for example, one is prepared to use an estimator, which is not positive semi-definite, a better rate can potentially be achieved. Or, if prior knowledge about the asymptotic variance matrix is available or parametric assumptions can be verified from the data, it is typically better to construct estimators which attempt to exploit that information relative to doing subsampling.[1] Still, in finite samples we show in a realistic setting with microstructure noise the subsampler produces convincing results compared to some available alternatives.

The rest of this paper goes as follows. Section 2 introduces the setting. In Section 3, we derive the theory first without and then with noise. In Section 4, we do numerical simulations in order to inspect the finite sample performance of our estimator. In Section 5, we confront our framework with some real high-frequency data, while the Appendix contains the proofs of our results.

[1] To paraphrase Politis et al. (1999), subsampling is "a robust starting point toward even more refined procedures".

Please cite this article in press as: Christensen, K., et al., Inference from high-frequency data: A subsampling approach. Journal of Econometrics (2017), http://dx.doi.org/10.1016/j.jeconom.2016.07.010