International Journal of Forecasting 31 (2015) 349–363

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Forecasting financial market activity using a semiparametric fractionally integrated Log-ACD

Yuanhua Feng *, Chen Zhou

University of Paderborn, Germany

ARTICLE INFO

Keywords:
Approximately best linear predictor
Realized volatility
Financial forecasting
Long memory
Nonparametric scale function
Semi-FI-Log-ACD

ABSTRACT

This paper considers the modeling and forecasting of long memory and a smooth scale function in different nonnegative financial time series aggregated from high-frequency data based on a fractionally integrated Log-ACD (FI-Log-ACD) and its semiparametric extension (Semi-FI-Log-ACD). Necessary and sufficient conditions for the existence of a stationary solution of the FI-Log-ACD and its properties under the log-normal assumption are studied in detail. An approximately best linear predictor based on the truncated AR(∞) form of the logarithmic process is proposed, and approximate variances of the prediction errors for an individual observation and for the conditional mean are obtained. Forecasting intervals for these quantities in the log-transformed data and in the original process are calculated under the log-normal assumption. Finally, applications to realized volatility, trading volumes and other data sets show that the proposal works very well in practice.

1. Introduction

This paper considers the forecasting of long memory and a smooth scale function in financial time series aggregated from high-frequency data, such as (daily) trading volumes, trading numbers, average transaction durations, and realized volatilities. Here, long memory is probably caused by aggregation. In the current context, well-known short-memory models include the ARCH (autoregressive conditional heteroskedasticity; Engle, 1982) and GARCH (generalized ARCH; Bollerslev, 1986) for returns, and the ACD (autoregressive conditional duration, Engle & Russell, 1998) for transaction durations. The ACD can also be applied to trading volumes (Manganelli, 2005) and other quantities. Furthermore, the (first type) Log-ACD (Allen, Chan, McAleer, & Peiris, 2008; Bauwens, Galli, & Giot, 2008; Bauwens & Giot, 2000; Karanasos, 2008) is also proposed, where the log-data are modeled by an ARMA. The idea ensures that the estimates of the original data are always nonnegative. The modeling of a smooth scale function in volatility caused by a changing macroeconomic environment was investigated by Engle and Rangel (2008) and Feng (2004).

Well-known long memory volatility models include the FIGARCH (fractionally integrated GARCH; Baillie, Bollerslev, & Mikkelsen, 1996), the LM-GARCH (long memory GARCH, Karanasos, Psaradakis, & Sola, 2004), the FIACD (Jasiak, 1998), and the LM-ACD (Karanasos, 2004). Baillie and Morana (2009) proposed an adaptive FIGARCH for modeling long memory and structural breaks in volatility. As far as we know, the estimation of smooth scale functions in volatility models with long memory has not yet been studied extensively. Most recently, Beran, Feng, and Ghosh (2014) proposed that short memory, long memory and a nonparametric scale function be modeled in nonnegative financial time series based on the log-transformation. They assumed that the process under consideration has a log-normal marginal distribution and proposed that the stochastic component of the log-data be modeled using

* Corresponding author. Tel.: +49 5251 60 3379.
E-mail address: yuanhua.feng@wiwi.upb.de (Y. Feng).

http://dx.doi.org/10.1016/j.ijforecast.2014.09.001
0169-2070/

a Gaussian FARIMA (fractional autoregressive integrated moving average; Beran, Feng, Ghosh, & Kulik, 2013, and Hosking, 1981). The log-data themselves are analyzed using a SEMIFAR (semiparametric fractional autoregressive; Beran & Feng, 2002a). Hence, their proposals are referred to as EFARIMA (exponential FARIMA) and ESEMIFAR, respectively, which can be estimated easily using existing software packages.

In this paper, a possible origin of long memory in the data under consideration is discussed. The EFARIMA and ESEMIFAR models are extended to the case of general marginal distributions and represented as FI-Log-ACD and Semi-FI-Log-ACD models, respectively. Necessary and sufficient conditions for the existence of a stationary solution of the FI-Log-ACD are obtained, and the properties of the model under the log-normal assumption are investigated in detail. In particular, the long memory parameter is no longer affected by the log-transformation (see also Dittman & Granger, 2002). Our focus is on forecasting using the Semi-FI-Log-ACD, based on an improved data-driven SEMIFAR algorithm. The forecasting is carried out using a truncated linear predictor based on the AR(∞) form, which is simpler and runs faster than the best linear SEMIFAR predictor of Beran and Ocker (1999). The properties of the proposal are investigated in detail. For an ARMA model, the proposed predictor is an approximately best linear predictor (Brockwell & Davis, 2006, p. 184). We show that this is still true in the presence of long memory. The asymptotic variances of the prediction errors for an individual observation and for the conditional mean are obtained, and the calculation of approximate forecasting intervals under the log-normal assumption is discussed. The effects of the errors in the estimated trend on the asymptotic properties of the proposed predictor are investigated. Applications to (daily) trading volumes, trading numbers, average durations and realized volatilities show that the proposals work very well in practice and the log-normal distribution is quite reasonable. Finally, our empirical results reveal that, in addition to long memory and smooth scale change, realized volatility may also exhibit structural breaks.

The motivations for this study and the definitions of the models are given in Section 2. Section 3 describes the properties and estimation of the proposed models. The linear predictor is proposed and studied in Section 4. Section 5 reports the empirical results. Final remarks in Section 6 conclude the paper. The proofs of our results are given in the Appendix.

2. A semiparametric multiplicative long memory model

2.1. Origins of long memory in aggregated financial data

Non-negative financial time series often exhibit long memory. Long memory in realized volatilities and trading volumes has been studied by Andersen, Bollerslev, Diebold, and Ebens (2001) and Fleming and Kirby (2011), among others. Deo, Hsieh, and Hurvich (2010) revealed that transaction durations, trading numbers, squared returns and realized volatilities may exhibit long memory at potentially the same level. Beran et al. (2014) found evidence of long memory in average durations.

A well-known theoretical origin of long memory in economic time series is the cross-sectional aggregation of microeconomic data. For instance, Granger (1980) and Leipus, Philippe, Puplinskaite, and Surgailis (2014) showed that the aggregation of random-coefficient ARMA processes under certain conditions will result in long-memory processes. Zaffaroni (2007) discussed the aggregation of GARCH models and indicated that this does not lead to long memory in volatility. See also Beran et al. (2013) and references therein. The time series considered in this paper can be thought of as special cross-sectional aggregates of high-frequency data, i.e., aggregations of micro-financial data in some sense. For instance, define the realized volatility based on 1-min (log-) returns as the sum of squared returns (Andersen et al., 2001). If the returns at a given time point on a trading day follow a GARCH model, the realized volatility will be an aggregate of squared GARCH processes. Although a squared GARCH process corresponds to a nonlinear ARMA model, the stationary conditions for such (squared) processes are quite different from those for linear ARMA models. Hence, the results of Granger (1980) and Leipus et al. (2014) do not apply to realized volatility. Zaffaroni's (2007) results on the aggregation of GARCH models also do not apply to the aggregation of squared returns. The realized volatility used in the application in Section 5 is indeed defined as the sum of the squared ultra-high-frequency returns (without considering the effect of the microstructure noise). Now, discussing the origin of long memory in such time series would be even more difficult. To the best of our knowledge, theoretical models explaining the origin of long memory in financial time series aggregated from high-frequency data are still unknown; however, we believe that it is caused mainly by aggregation.

2.2. Simultaneously modeling long memory and scale changes

A well known model for a stationary nonnegative financial time series $X_t,\ t = 1, \dots, n$, is the MEM (multiplicative error model; Engle, 2002), defined as

$$X_t = \nu\lambda_t\eta_t, \tag{1}$$

where $\nu > 0$ is a scale parameter, $\lambda_t > 0$ is the conditional mean of $X_t^* = X_t/\nu$, and $\eta_t \geq 0$ are i.i.d. random variables. In order to model long memory and a slowly changing scale function simultaneously, we propose the use of a semiparametric MEM model by replacing ν in Eq. (1) with a nonparametric scale function $\nu(\tau) > 0$:

$$X_t = \nu(\tau_t)X_t^* = \nu(\tau_t)\lambda_t\eta_t, \tag{2}$$

where $\tau_t = t/n$ denotes the rescaled time. Let $Y_t = \ln(X_t)$, $\mu(\tau_t) = \ln[\nu(\tau_t)]$, $\zeta_t = \ln(\lambda_t)$, $\varepsilon_t = \ln(\eta_t)$ and $Z_t = \zeta_t + \varepsilon_t$, where $\mu(\tau_t)$ and ζ_t are the local and conditional means of Y_t, respectively. Following Beran et al. (2014), we assume that $E(\varepsilon_t) = 0$, var $(\varepsilon_t) = \sigma_\varepsilon^2$, and the stochastic component Z_t follows a FARIMA

$$\pi(B)Z_t = \psi(B)\varepsilon_t, \tag{3}$$

where $\pi(B) = (1 - B)^d\phi(B) = 1 - \sum_{i=1}^{\infty} \pi_i B^i$ with $\pi_i \approx c_\pi i^{d-1}$ for large i, $d \in (-0.5, 0.5)$ is the fractional differencing parameter, and $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$