Journal of the Korean Statistical Society 43 (2014) 251–262

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Optimal sampling frequency for high frequency data using a finite mixture model

Hyunhee Choi [a,b], Moonsu Kang [c,*]

[a] Department of Statistics, Korea University, Seoul, 136-720, Republic of Korea

[b] Laboratory, IBM Korea, 135-750, Republic of Korea

[c] Department of Information Statistics and Research Institute of Natural Science of Gangneung-Wonju National University, Gangneung-Wonju National University, Republic of Korea

ARTICLE INFO

Article history:
Received 28 December 2012
Accepted 5 September 2013
Available online 5 October 2013

AMS 2000 subject classifications:
primary 62P20
62P05
secondary 91B82
62F03

Keywords:
Optimal sampling frequency
Mixture model
Time series
Modified likelihood test
Volatility
AICc

ABSTRACT

Advances in IT technology have led to the feasibility of high frequency sampling for stock price data in financial time series. A standard approach to obtaining a sampling cycle is to calculate n by minimizing the mean squared error (MSE) which is not appropriate for a nonlinear time series mixture and does not account for the number of parameters included in a model and targeted statistical power. The objective of this article is to show two methods for the calculation of optimal sampling frequency under the framework of a finite mixture model. First we investigate how sampling frequency can be obtained through a modified likelihood ratio test to obtain a designated statistical power. Second, we propose a new method based upon the minimization of AICc with a penalty for the number of parameters. Numerical studies show that it performs better than other methods.

1. Introduction

With the rapid growth in financial markets, price data have been sampled at extremely high frequencies over the past few years. Since most researchers in financial markets generally use thousands of financial data for model estimation, sample size has not typically been a great concern. However, microstructure noise effects have caused a severe bias in daily statistical calculations (McAleer & Medeiros, 2008). The fundamental causes of microstructure noise include price discreteness, price changes caused by price limits, non-synchronous trading, reporting errors, quote delays and bid–ask spread (de Pooter, Martens, & van Dijk, 2008). The effect becomes huge when the sampling frequency (sample size for a stock in a given day) is large, which motivates us to fit a more complex model rather than a simple one, requiring the utility of a finite mixture model in the data. There is a vast literature on computing sampling frequency for high frequency sampling. In general, a conventional sampling frequency calculation method is a high frequency sampling method with the objective of minimum squared error (MSE) minimization (Bandi & Russell, 2006), but it has two major drawbacks. First, it cannot be applied for

* Corresponding author. Tel.: +82 10 6331 0853.
E-mail addresses: choihh00@gmail.com (H. Choi), moonsukang0223@gmail.com (M. Kang).

1226-3192/$ – see front matter
http://dx.doi.org/10.1016/j.jkss.2013.09.003

(nonlinear) financial time series which have often been based on a finite mixture model. A mixture distribution for a time series is designed from the point in which stock return volatility is high at the open and close of trading and low in the middle of the day. Moreover, an auto-correlated series should be sampled sparsely considering the correlation between samples (Craig, 1984). Many researchers have studied a finite mixture model for time series auto-correlated data (Albert, 1991; Huerta, Jiang, & Tanner, 2003; Jin & Li, 2006; McQuarrie & Tsai, 1998). Second, it does not fully consider the number of parameters involved in a model (Akaike, 1974). A few researchers have studied the alternative methods to MSE in the literature. The aim of this paper is to present two methods for the calculation of optimal sampling frequency under the framework of a finite mixture model.

In this sense, we show how we apply the modified likelihood ratio test based on a finite mixture model to obtain sampling frequency first (Chen, Chen, & Kalbfleisch, 2001; Chen & Kalbfleisch, 2005), so that the model should obtain a targeted statistical power unlike MSE. In fact, it tests the homogeneity in the finite mixture model (McLachlan & Peel, 2000) with a great parametric kernel distribution family. Pros and cons of this algorithm should be considered in the light of target data in determining an appropriate algorithm. It explains the reason for choosing such a model (Chen et al., 2001; Chen, Chen, & Kalbfleish, 2004; Chen & Kalbfleisch, 2005; Qin & Smith, 2004) and the reason why the modification of a likelihood ratio test for the mixture model is needed (McLachlan & Peel, 2000).

Second, we propose a new method based on the minimization of AICc (Akaike information criterion (AIC) with a correction) involving a penalty for the number of parameters under the framework of a finite mixture model. AICc was first proposed by Hurvich and Tsai (1989) who ground their high opinion of AICc on extensive simulation work with time series. To penalize for the number of parameters is very crucial in model assessment. We avoid the fact that as the number of parameters increases, model assessment criterion should be inflated.

We introduce a motivating example for our study in Korean stock market data in Section 2. In Section 3, we describe our model with efficient prices and microstructure noises in detail. In Section 4, a modified likelihood ratio test (MLRT) for our data is conducted and we discuss how to determine sample size for a designated statistical power. Section 5 shows how the optimal sampling frequency for a mixture of generalized autoregressive conditional heteroskedasticity models is obtained by taking AICc calculations into account (Bollerslev, 1986). In Section 6, we illustrate simulation studies in various settings of parameters. We demonstrate Korean stock market data for illustration in Section 7.

2. Motivation of this study

Many researchers have developed the computation of sampling frequency for high frequency sampling in stock price data in the literature. They usually advocated using MSE for this but ignored the following aspects arising from a financial stock market. In Korea, the stock market shows the relationship between volatility and herd behavior which creates high volatility and also increases trading volume. It leads to the transformation of the shape of the volatility pattern and the market is also highly affected by new information. The effect on volatility is very strong in the morning.

Those effects on the stock market make a totally U-shaped volatility pattern. It has been widely documented that volatility varies systematically over the trading day and that this pattern is highly correlated with the intraday variation of trading volume and bid–ask spreads as properties of the financial market. Volatility exhibits a U-shaped pattern, in particular, high at the open and close of trading and low in the middle of the day, which implies that data consist of a few of distinct models with different variance as well as different means. In addition, there is a vast literature of time series data based on a finite mixture model (Albert, 1991; Huerta et al., 2003; Jin & Li, 2006). These perspectives motivate us to apply a finite mixture model for the time series to our data. Unfortunately, MSE does not account for this. For instance, Fig. 1 illustrates that the stock price return of the Korean stock price index (KOSPI) was distributed over 5 min intervals in a day between December 1, 2011 and February 29, 2012. In this data, morning volatility is bigger than afternoon volatility so that an appropriate mixture of distributions for the stock market index should be applied.

Two or more volatility clusters can be mixed in Fig. 1. For this reason, two time series models, the ARMA and GARCH models discussed in Tsay (2005) will be considered. On the other hand, when computing an appropriate sample size in stock price data, the number of parameters involved in the model is not considered in most cases. So, we want to determine sample size to account for this aspect by utilizing AICc criterion.

3. Model description

The basic setup for the stock price model is as follows.

$$p(t) = \mu(t) + \sigma(t)dW(t), \quad t = 1, 2, \ldots, n, \tag{1}$$

where $p(t)$ is a stock price at time t, $\mu(t)$ is a drift component, $\sigma(t)$ is an instantaneous volatility (or standard deviation) and $W(t)$ is a standard Brownian motion. In addition, suppose also that $\sigma(t)$ is orthogonal to $W(t)$ such that there is no leverage effect.

Stock price is assumed to be a sum of an efficient price and microstructure noises which are independent of each other. Using similar notation in Zhang (2006), $\log(p(t))$, $t = 1, 2, \ldots, n$ can be modeled below.

$\log(p(t)) = (\log(p(t)))^* + \varepsilon_t$, where $(\log(p(t)))^*$ is an efficient price and ε_t is a microstructure noise.