Economics Letters 120 (2013) 224–228

Contents lists available at SciVerse ScienceDirect

Economics Letters

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

A factor approach to realized volatility forecasting in the presence of finite jumps and cross-sectional correlation in pricing errors

Alev Atak [a,*], George Kapetanios [b]

[a] City University, Department of Economics, Social Sciences Building, Whiskin Street, London EC1R 0JD, United Kingdom

[b] School of Economics and Finance, Queen Mary, University of London, Mile End Road, London E1 4NS, United Kingdom

HIGHLIGHTS

- This paper examines the role of approximate factors in forecasting future realized volatility.
- We identify the discontinuous components using the jump tests before applying factors.
- We extend the factor models to the derivation of our realized measure.
- We relate the common component to unobservable financial characteristics.
- Our model outperforms the currently available approaches.

ARTICLE INFO

Article history:
Received 10 March 2013
Accepted 30 March 2013
Available online 16 April 2013

JEL classification:
C32
C50
C52
C53
C58

Keywords:
Realized volatility
Bipower variation
Jump tests
Factor models
Volatility forecasting
Model selection

ABSTRACT

There is a growing literature on the realized volatility (RV) forecasting of asset returns using high-frequency data. We explore the possibility of forecasting RV with factor analysis; once considering the significant jumps. A real high-frequency financial data application suggests that the factor based approach is of significant potential interest and novelty.

1. Introduction

Recently, there has been increasing interest in forecasting methods that utilize large, high frequency data sets. Andersen and Bollerslev (1998), Andersen et al. (2003), Barndorff-Nielsen and Shephard (2002) (termed BNS henceforth), among others, advocated the use of nonparametric realized volatility (RV). The consistency of the RV as an estimator is violated by the presence of the market microstructure noise (henceforth 'noise') which emerges due to market frictions. Another backdrop is that the nonparametric RV literature has concentrated less on distinguishing jump from non jump movements. Corsi et al. (2010) reveal that dividing volatility into jumps and continuous variation yields a substantial improvement in volatility forecasting.

There is an alternative way of looking at these problems. The limitations of the traditional procedures motivate our diverse approach for measuring and forecasting the realized equity return volatility. We apply the methodology of approximate factor modeling on the nonparametric RV and also on the realized bipower variation (BV) (Barndorff-Nielsen and Shephard, 2004) when it is required after separately measuring the continuous sample path variation and the discontinuous jump part of the quadratic variation (QV) process. Factor methods are very appealing and extensively used for forecasting; providing a theoretical device for summarizing large data sets without running

* Corresponding author. Tel.: +44 0 20 7040 8500.
E-mail addresses: aatak@qmul.ac.uk (A. Atak), g.kapetanios@qmul.ac.uk (G. Kapetanios).

0165-1765/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.econlet.2013.03.051

into degrees of freedom problem, while taking into account the marginal benefits that increasing information brings to forecasting. As argued by Ludvigson and Ng (2009), the fluctuations and comovements of a large number of economic and financial variables are produced by a handful of observable or unobservable factors, which in this case represent the omitted unobservable factors in the noise. Our new factor-based realized volatility model (FB-RV-J) fits well for large dimensional panels.

2. Theory

The dynamics of the logarithmic price process, p_t, is usually assumed to be a jump-diffusion process of the form:

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

where μ_t denotes the drift term with a continuous and locally bounded variation, σ_t is the diffusion parameter and W_t is a standard Brownian motion. J_t is the jump process at time t, defined as $J_t = \sum_{j=1}^{N_t} k_{t_j}$ where k_{t_j} represents the size of the jump at time t_j and N_t is a counting process, representing the number of jumps up to time t. The QV of the price process up to a certain point in time t is:

$$\mathrm{QV}_t = \int_0^1 \sigma_s^2 ds + \sum_{j=1}^{N_t} k_{t_j}^2 \tag{2}$$

where $\int_0^1 \sigma_s^2 ds = \mathrm{IV}_t$ is the integrated variance or volatility. Thus, QV has two parts; the diffusion component and the jump component. The two components have a different nature and should be separately analyzed and modeled. The IV is characterized by persistence, whereas jumps have an unpredictable nature.

Let the interval $[0, t]$ be split into n equal subintervals of length m. The jth intra-day return r_j on day t is defined as $r_j = p_{t-1+jm} - p_{t-1+(j-1)m}$. QV_t can be estimated by the realized volatility, or variation, (RV_t), defined as (Andersen and Bollerslev, 1998):

$$\mathrm{RV}_t = \sum_{j=1}^{n} r_j^2 \xrightarrow{p} \mathrm{QV}_t, \quad \text{for } m \longrightarrow 0 \tag{3}$$

where $\xrightarrow{p}$ stands for convergence in probability. Hence, in the absence of discontinuities and noise the RV_t is consistent for the IV_t. Most of the jump detection procedures are based on the comparison between RV_t and a robust to jump estimator. We need to highlight that none of these procedures can test for the absence or presence of jumps in the model or the data generating process. Hence, it is difficult to judge whether the realization of the process is continuous or not, within a certain time interval or at a certain moment without a jump test. We turn now to the jump detection methods.

2.1. Jump tests

We use two tests, the adjusted ratio statistic of Barndorff-Nielsen and Shephard (2006) and the Lee and Mykland test (Lee and Mykland, 2008, termed LM henceforth), in order to check whether the two tests give consistent results. The BNS test tells whether a jump occurred during a particular day and how much the jump-squared contributes to the total realized variance, i.e. $\int_{t-1}^{t} J_s^2 dq_s / \mathrm{RV}_t$. The significant jump component of RV_t is:

$$\widehat{J_t} \equiv \mathrm{sign}(r_t) \times \sqrt{(\mathrm{RV}_t - \mathrm{BV}_t) \times I_{t,(\mathrm{ZJ_{(bv)}} \geq \Phi_\alpha^{-1})}} \tag{4}$$

where $\mathrm{BV}_t = 1.57 \sum_{j=2}^{n} |r_j| \, |r_{j-1}|$. The BNS test can only identify days that contain jumps. Hence, we use the "intra-day" LM test which has the additional capability of identifying specific returns that can be classified as jumps. We compute the LM test statistics for every moment t_j within a trading day and then pick up the maximum statistic as the final test for that day to determine whether both tests are consistently detecting the presence of jumps. We effectively observe the consistency in both methods.

2.2. Model

We now put the idea of separately measuring the jump component and continuous variation. The contribution to the QV_t process due to the discontinuities in the underlying price process can be estimated by:

$$\mathrm{RV}_t - \mathrm{BV}_t \to \sum_{j=1}^{N_t} k_{t_j}^2, \quad \text{for } m \longrightarrow 0. \tag{5}$$

Under this central insight and based on the above mentioned test statistics and threshold requirements, we use BV_t in our analysis if we detect jumps in the data, otherwise RV_t. So, $C_t = I_{t,(\mathrm{ZJ_{(bv)}} < \Phi_\alpha^{-1})}\mathrm{RV}_t + I_{t,(\mathrm{ZJ_{(bv)}} \geq \Phi_\alpha^{-1})}\mathrm{BV}_t$. This recognition motivates our model. We propose that our nonparametric jump-free 'realized' measure can be decomposed into the common and idiosyncratic components. We relate the common component to unobservable financial characteristics, in particular, to cross sectional correlation in pricing errors. For simplicity, we abbreviate our model FB-RV-J:

$$h_{it} = \alpha_i' f_t + u_{it}, \quad t = 1, \ldots, T \text{ and } i = 1, \ldots, N \tag{6}$$

where h_{it}, is the realized measure, which is the element in the tth row and ith column of the data matrix, $T \times N$. f_t is a r-dimensional vector of common factors with $t = 1, \ldots, T$ and α_i refers to the ith row of the corresponding matrix of factor loadings. $\alpha_i' f_t = W_{it}$ is the set of common components. In addition, u_{it} is the idiosyncratic component of h_{it}. We assume that in general the idiosyncratic terms are also weakly dependent processes with mild cross-sectional dependence. α_i and f_t are clearly not jointly identified since the factors can be pre-multiplied by an invertible $r \times r$ matrix without having to make changes in the model. The most crucial point here is that $r \ll N$, so that substantial dimension reduction can be achieved.

Factor identification and estimation of (6) is based on the set of assumptions that are used in Bai and Ng (2002, 2006). Estimation is divided into steps; we start with determining the number of factors, which is followed by estimating them along with the loadings. We estimate common factors in large panels by the method of asymptotic principal components. This approach fits well for the large panel of realized volatilities because it does not suffer from the curse of dimensionality problem.

2.2.1. The number of factors

We now focus on checking robustness with respect to the number of factors and consider two approaches; Bai and Ng (2002) information criteria forming a nonparametric method to determine the statistically important factors and the Onatski (2010) estimator described by an algorithm named edge distribution (ED). Kapetanios (2010) suggests a method of the determination of the number of factors using a bootstrap method, which is robust to considerable cross-sectional and temporal dependence, but we prefer to follow a simpler approach by Onatski (2010). As it is shown in the empirical application, the two methods indicate that there exist three common factors.

3. Empirical application

The data used in this paper are extracted and compiled from the Trade and Quote (TAQ). We use the 50 largest capitalization stocks included in the S&P500 index. The data consists of full record transaction prices from January 2007 to December 2010. As in Müller et al. (1993) linear interpolation of logarithmic five-minute