Forecasting Volatility in Financial Markets with High Frequency Data

Marta Campi
(Lyudmyla Hvozdyk)

University of Essex / University College London

ucabmc2@ucl.ac.uk

February 25, 2016

Overview

1. Introduction
2. Literature Review
3. Methodology
4. Empirical Analysis
5. Conclusions
6. Future Research

Introduction

What is high frequency?

- Annual
- Quarterly
- Monthly
- Weekly
- Daily
- Hourly
- 30 - minutes
- 10 - minutes
- 5 - minutes
- 1 - minutes
- Seconds
- Transaction time

Why is it relevant to financial markets?

Volatility

Central Role

- Asset Pricing
- Asset Allocation
- Risk Management

Description

"The extent to which the price of a security or commodity, or the level of a market, interest rate or currency, changes over time". - *Financial Times Lexicon*

The 1990s

Main changes:

- Database Availability
- Use of new tools
- Improved computing power

Estimation

- Parametric approach
- Semi- parametric methods
- **Non-parametric Measures**

Merton (1980) - Nelson (1992) Volatility may be estimated through sufficiently finely sampled high-frequency returns over a certain time interval.

Supporting this idea, Andersen, Bollerslev, Diebold and Labys (ABDL)(2001), Barndorff-Nielsen and Shephard (2002a,b), Meddhai (2002) propose the use of **non-parametric simple measures** know as *Realized Variance or Realized Variation*.

- Avoiding misspecification (GARCH versus stochastic volatility models) ABDL (2001)
- Retaining most of the relevant information at intradaily level Barndorff-Nielsen and Shephard (2002a)
- Simple model of Realized Volatility outperform popular GARCH and SV models ABDL (2003)

Main Features

Financial data show a series of **stylized facts** that have to be taken into within the considered model (i.e autocorrelation of square and absolute returns, fat-tails of return distribution, persistence, long-memory, multi-scaling behavior), ABDL (1997), Ding, Granger and Engle (1993)

A nonparametric approach allows to build measures that are more robust to discontinuities of the price-process. Hence, the considered model has to employ robust measure in terms of the **jump component** ABDL(2006), Barndorff-Nielsen and Shephard (2002), Ghysels, Santa-Clara and Valkonov (2003)

Modelling Volatility

- Corsi (2003) offers a simple model able to reproduce the statistical behavior of financial data: *The Heterogeneous Autoregressive Model*
- At the same time, Ghysels, Santa-Clara and Valkonov (2003) propose *Mixed Data Sampling Regression Models* to combine data sampled at different frequencies

Both models:

- reproduce the **memory persistence** observed in financial data
- reduce the **number of parameters** to estimate
- allow to employ predictors robust to **jumps**
- consider **multiple horizons forecast**

Why Exchange Rates Market?

Jumps of interest rates are strictly related to **macroeconomic news events** Barndorff-Nielsen and Shephard (2006), ABDL(2006), Andersen, Bollerslev, Diebold and Vega(2003, 2005), Johannes (2004)

"Exchange rates are among the most active and **liquid market** in the world, permitting high-frequency sampling without contamination of microstructure effects" Andersen, Bollerslev, Diebold and Vega (2003)

"The first motivation is obviously the possibility of refining our understanding of the fundamental determinants of exchange rates, the central and still largely-unresolved question of **exchange rate economics**" Andersen, Bollerslev, Diebold and Vega (2003)

Why does sampling frequency matter

- Asymptotic theory suggests that the higher the sampling frequency, the closer it is to continuously observed prices, the better.
- Market frictions bring to high market microstructure noise.
- Use 5 minutes frequency represents **the balanced trade-off between bias and variance of the estimators** ABDL(2001, 2003), Barndorff-Nielsen and Shephard (2002a,b)

Motivations and Contributions

- High Frequency Data provide better forecast performances
- HAR-RV models are employed given their easy computation
- MIDAS-RV models are used because of their flexibility
- They both are able to include robust regressors to jumps
- **HAR models can be considered as restricted MIDAS models**

The main contribution is a comparison in terms of performances IN SAMPLE and OUT of SAMPLE through different regressors over three horizons of a relative recent timespan of data.

Theoretical Background

Classic continuous time stochastic volatility model for the logarithmic price p(t) at time t:

$$dp\left(t\right) = \mu\left(t\right)dt + \sigma\left(t\right)dW\left(t\right)$$

- $\mu\left(t\right)$ is a continuous and locally bounded variation process (drift)
- $\sigma\left(t\right)$ strictly positive stochastic volatility process
- W(t) is a standard Brownian Motion

The increment of the quadratic variation over some horizon H:

$$\sigma^{|2|}_{t,t+H} = \int_{t}^{t+H} \sigma^{2}\left(s\right)ds$$

However, the Quadratic Variation is not observable. Therefore the empirical stochastic specification of the continuous-time jump diffusion process is considered:

$$dp(t) = \mu(t)\,dt + \sigma(t)\,dW(t) + k(t)\,dq(t)$$

- $q(t)$ is a counting process that identifies the presence of jumps
- $k(t)$ represents the jump size

The quadratic variation for the return process is then:

$$QV_{t,t+H} = \int_t^{t+H} \sigma^2(s)\,ds + \sum_{\{s\in[t,t+H]:dq(s)=1\}} k^2(s)$$

Realized Variance and Bipower Variation

Barndorff-Nielsen and Shephard (2002) proposed the following estimator for the quadratic variation:

$$RV_{t,t+H}^{M} = \sum_{j=1}^{MH} \left(r_{(t+H)-(j-1)/M,(t+H)-(j-2)/M} \right)^2$$

$$\lim_{M \to \infty} RV_{t,t+H}^{M} \to QV_{t,t+H}$$

Barndorff-Nielsen and Shephard (2004b) present the *Bipower Variation* as:

$$BPV_{t,t+H}^{M}(k) = \mu_1^2 \sum_{j=k+1}^{HM} |r_{(t+H)-(j-1)/M,(t+H)-(j-2)/M}||r_{(t+H)-(j-1-k)/M,(t+H)-(j-2-k)/M}|$$

Jumps'detection

$$RV^{M}_{t,t+H} - BPV^{M}_{t,t+H} \rightarrow \sum_{\{s\in[t,t+H]:dq(s)=1\}} k^{2}\left(s\right)$$

By exploiting this quantity, Huang and Tauchen (2005) define a statistical test to identify the jump component:

$$Z_{t,t+H} = \frac{\left[RV_{t,t+H} - BV_{t,t+H}\right] RV^{-1}_{t,t+H}}{\left[\left(\mu_1^{-4} + 2\mu_1^{-2} - 5\right)/M max\{TQ_{t,t+H} BV t, t+H^{-2}\}\right]^{1/2}}$$

Then, the jump component and the continuous sample path are:

$$J^{\alpha}_{t,t+H} = I\left(Z_{t,t+H} > \Phi_{\alpha}\right)\left(RV_{t,t+H} - BV_{t,t+H}\right)$$

$$C^{\alpha}_{t,t+H} = I\left(Z_{t,t+H} \leq \Phi_{\alpha}\right) RV_{t,t+H} + I\left(Z_{t,t+H} > \Phi_{\alpha}\right) BV_{t,t+H}$$

HAR-RV Model

Corsi (2003) proposed the following model in order to predict the variable RV, known as the *Heterogeneous Autoregressive model* over some horizon H:

$$RV_{t,t+H} = \beta_0 + \beta_D RV_t + \beta_W RV_{t,t-5} + \beta_M RV_{t,t-20} + \epsilon_{t+H}$$

where the normalised realised variance over the horizon H is:

$$RV_{t,t+H} = H^{-1}\left[RV_{t+1} + RV_{t+2} + ... + RV_{t+H}\right]$$

HAR-RV-CJ Model

ABDL (2003) provided an extension by splitting the continuous sample path and the jump component of RV and define the HAR-RV-CJ model through these quantities:

$$C_{t,t+H} = H^{-1}\left[C_{t+1} + C_{t+2} + \ldots + C_{t+H}\right]$$

$$J_{t,t+H} = H^{-1}\left[J_{t+1} + J_{t+2} + \ldots + J_{t+H}\right]$$

$$RV_{t,t+H} = \beta_0 + \beta_{CD} C_t + \beta_{CW} C_{t,t-5} + \beta_{CM} C_{t,t-20} + \beta_{JD} J_t + \beta_{JW} J_{t,t-5} + \beta_{JM} J_{t,t-20} + \epsilon_{t+H}$$

MIDAS Regression Model

Ghysels, Santa-Clara and Valkanov (2002, 2004) introduced *Mixed Data Sampling Regression Models* allowing regression at different frequencies. A general MIDAS Model is defined as:

$$Y_t = \beta_0 + B\left(L^{1/m}\right) X_t^{(m)} + \epsilon_t^{(m)}$$

- Y_t is a variable sampled at certain frequency.
- X_t^m is a variable sampled m times faster.
- $B\left(L^{1/m}\right) = \sum_{j=0}^{jmax} L^{j/m}$ denotes a polynomial of maximum length equal to j^{max} in the $L^{1/m}$ operator.

Several specifications of the weighting scheme:

- Exponential Almon Lag function
- **Beta lag function**
- Polynomial with step functions
- Unrestricted coefficients

MIDAS-RV Models

Forsberg and Ghysels (2004), MIDAS-RV Models are exploited by using several regressors as in the HAR-RV models. A MIDAS-RV model is a standard MIDAS model which has RV as dependent variable:

$$RV_{t,t+H} = \mu_H + \phi_1 \sum_{k=0}^{k_{max}-1} w_k X_{t-k,t-k-1} + \epsilon_{H_t}$$

- H is the prediction horizon
- w_k is a chosen weighting scheme
- $X_{t-k,t-k-1}$ represents the regressor from t-k to t-k-1.

The Beta lag function is employed and specified as:

$$b_H(k;\theta) = \frac{f\left(\frac{k}{k_{max}}\theta_1, \theta_2\right)}{\sum_{j=1}^{k_{max}} f\left(\frac{k}{k_{max}}\theta_1, \theta_2\right)}, \text{ with } \theta_1 = 1, \theta_2 > 1$$

HAR-RV Models vs MIDAS-RV Models

HAR-RV Models

$$HAR-RV-RV : RV_{t,t+H} = \beta_0 + \beta_D RV_t + \beta_W RV_{t,t-5} + \beta_M RV_{t,t-20} + \epsilon_{t+H}$$

$$HAR-RV-C : RV_{t,t+H} = \beta_0 + \beta_D C_t + \beta_W C_{t,t-5} + \beta_M C_{t,t-20} + \epsilon_{t+H}$$

$$HAR-RV-CJ : RV_{t,t+H} = \beta_0 + \beta_{CD} C_t + \beta_{CW} C_{t,t-5} + \beta_{CM} C_{t,t-20} + \beta_{JD} J_t + \beta_{JW} J_{t,t-5} + \beta_{JM} J_{t,t-20} + \epsilon_{t+H}$$

MIDAS-RV Models

$$MIDAS-RV-RV : RV_{t,t+H} = \mu_H + \phi_1 \sum_{k=0}^{k_{max}} b(k,1,\theta_2) RV_{t-k,t-k-1} + \epsilon_{t+H}$$

$$MIDAS-RV-C : RV_{t,t+H} = \mu_H + \phi_1 \sum_{k=0}^{k_{max}} b(k,1,\theta_2) C_{t-k,t-k-1} + \epsilon_{t+H}$$

$$MIDAS-RV-CJ : RV_{t,t+H} = \mu_H + \phi_1 \sum_{k=0}^{k_{max}} b\left(k,1,\theta_2^C\right) C_{t-k,t-k-1} + \phi_2 \sum_{k=0}^{k_{max}} b\left(k,1,\theta_2^J\right) J_{t-k,t-k-1} \epsilon_{t+H}$$

Performance Measurements

IN SAMPLE and OUT-of-SAMPLE analysis are conducted in order to evaluate performances of the models and identify the best one at every selected horizon.

- Mean Squared Error (and related transformated versions)
- Median of the Heteroscedastic adjusted squared error (and related transformated versions)

$$MSE = N^{-1} \sum_{i=1}^{N} \left(RV_{i,i+H} - \hat{RV}_{i,i+H} \right)^2$$

$$MedHSE = Median \left(\frac{RV_{i,i+H}}{exp\left(ln\hat{RV}_{i,i+H} \right)} - 1 \right)^2$$

Forecast Evaluation

Diebold and Mariano (1995) define a test for the forecast accuracy. This test exploits *the difference in forecast errors* between two models analyzed over the same forecast window.

$H_0 : E\left[g\left(e_{i,t}\right)\right] = E\left[g\left(e_{j,t}\right)\right]$ or $E\left[d_t\right] = 0$ where $d_t = \left[g\left(e_{i,t}\right) - g\left(e_{j,t}\right)\right]$

Diebold-Mariano test

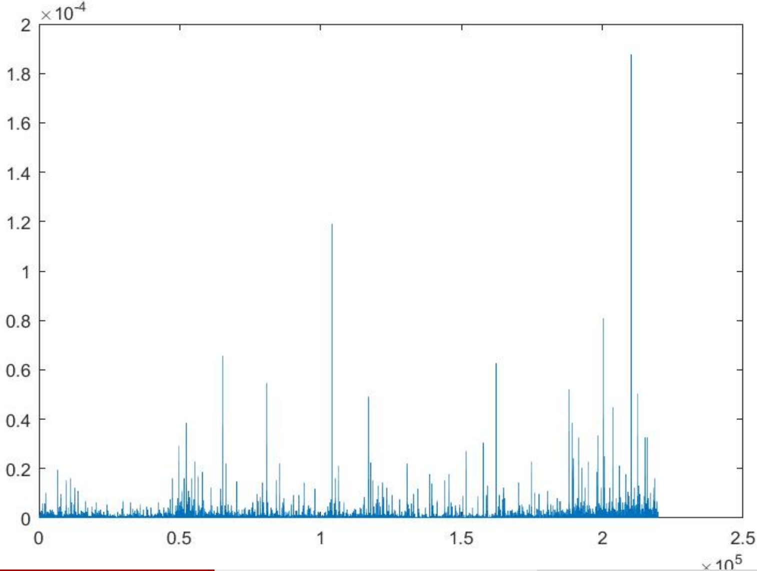
$$S = \frac{\bar{d}}{\left(\frac{L\hat{R}V}{T_0}\right)^{1/2}}$$

- $\bar{d} = \frac{1}{T_0} \sum_{t=t_0}^{T} d_t$
- $L\hat{R}V = \gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j$ where $\gamma_j = cov\left(d_t, d_{t-j}\right)$
- The asymptotic distribution of S is a standard normal distribution

Data Description

- 5-minutes choice and market microstructure noise
- FX spot prices of most liquid markets:
  - GBP/EUR
  - USD/GBP
  - USD/EUR
  - USD/CHF
- Three years time span: 28 May 2012 8:00am to 25 June 2015 7:55am
- Holidays and weekends are removed from 21:05 GMT of a night to 21:00 GMT of the next one
- 219744 intradaily 5-mins-returns corresponding to 763 daily observations

Squared Return

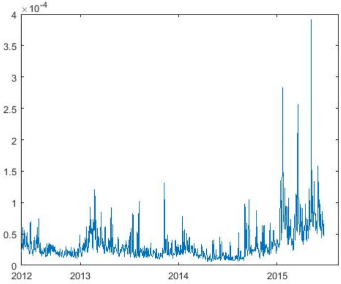


Statistics

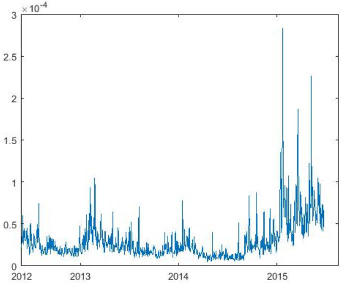
Table: GBB/EUR - Summary Statistics Realized Volatilities, Continuous paths and Jumps.

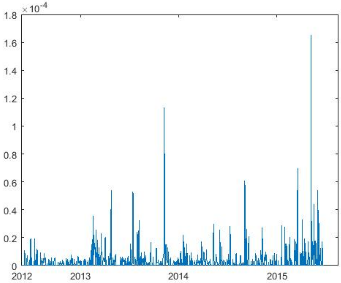
	RV_t	C_t	J_t	$\mathrm{RV}_t^{1/2}$	$\mathrm{C}_t^{1/2}$	$\mathrm{J}_t^{1/2}$
Mean	3.3363e-05	2.9597e-05	3.7663e-06	0.0054	0.0051	0.0011
St. Dev	2.9683e-05	2.4898e-05	1.0096e-05	0.0020	0.0018	0.0016
Skew	4.6053	3.5892	8.6144	1.8413	1.6230	1.9728
Kurt	41.5111	25.5343	111.6708	9.3622	7.3644	9.8301
Min	5.9000e-06	4.8086e-06	0.0000	0.0024	0.0022	0.0000
Max	3.9102e-04	2.8327e-04	1.6522e-04	0.0198	0.0168	0.0129
LB_{10}	1520	2239.3	25.79	2400.2	3072.5	25.7721

	$\ln(\mathrm{RV}_t$	$\ln(\mathrm{C}_t)$	$\ln(J_t + 1)$
Mean	-10.6542	-10.6542	3.7663e-06
St. Dev	0.6410	0.6399	1.0096e-05
Skew	0.5384	0.4771	8.6139
Kurt	3.3860	3.2531	111.6579
Min	-12.0406	-12.2451	0.0000
Max	-7.8468	-8.1691	1.6521e-04
LB_{10}	2749.7	3405.4	25.80



1. The left panel shows the RV_t.
2. The left-below represents the Continuous path C_t.
3. The right-below shows the Jump componentJ_t.





IN SAMPLE ANALYSIS

Table: The table shows the RV Models for the GBP/EUR FX

	HAR - RV			MIDAS - RV		
Horizon	RV	C	CJ	RV	C	CJ
MSE						
1 day	1.27E-10	1.39E-10	1.27E-10	8.366E-10	8.362E-10	8.362E-10
1 week	**3.241E-11**	**3.850E-11**	**3.220E-11**	**5.844E-10**	**5.885E-10**	**5.885E-10**
1 month	4.884E-11	4.251E-11	4.010E-11	6.079E-10	6.086E-10	6.086E-10
MEDhse						
1 day	0.065	0.062	0.067	0.078	0.073	0.073
1 week	**0.010**	**0.010**	**0.013**	**0.012**	**0.011**	**0.011**
1 month	0.021	0.018	0.022	0.025	0.022	0.022

OUT of SAMPLE ANALYSIS

Table: The table shows the RV Models for the GBP/EUR FX

	HAR - RV			MIDAS - RV		
Horizon	RV	C	CJ	RV	C	CJ
MSE						
1 day	9.11E-10	8.93E-10	9.11E-10	1.700E-09	1.681E-09	1.681E-09
1 week	**2.681E-10**	**2.705E-10**	**2.570E-10**	6.698E-10	6.511E-10	6.511E-10
1 month	3.915E-10	3.781E-10	3.756E-10	**2.269E-10**	**2.169E-10**	**2.169E-10**
MEDhse						
1 day	0.073	0.067	0.074	0.089	0.077	0.077
1 week	**0.030**	**0.035**	**0.032**	0.067	0.062	0.062
1 month	0.041	0.045	0.043	**0.029**	**0.030**	**0.030**

DM TEST

PERFORMANCE WITHIN HAR CLASS MODELS

	HAR-RV-RV vs HAR-RV-C		HAR-RV-RV vs HAR-RV-CJ		HAR-RV-C vs HAR-RV-CJ	
H	**S**	**P-Value**	**S**	**P-Value**	**S**	**P-Value**
1	0.375	0.7076	0.541	0.5886	0.422	0.673
5	**-4.961**	**7.014E-07**	0.689	0.4908	**4.991**	**6.01E-07**
20	**-2.583**	**0.0098**	-0.781	0.4348	**2.603**	**0.0092**

PERFORMANCE WITHIN MIDAS MODELS

	MIDAS-RV-RV vs MIDAS-RV-C		MIDAS-RV-RV vs MIDAS-RV-CJ		MIDAS-RV-C vs MIDAS-RV-CJ	
H	**S**	**P-Value**	**S**	**P-Value**	**S**	**P-Value**
1	**-7.847**	**4.22E-11**	**-7.847**	**4.22E-11**	0	1.000
5	**-3.489**	**4.85E-04**	**-3.489**	**4.85E-04**	0	1.000
20	-1.807	0.0708	-1.807	0.0708	0	1.000

Final Results

	BEST MODEL
	General
h = 1	HAR-RV-RV
h = 5	HAR-RV-RV or CJ
h = 20	MIDAS-RV-RV or CJ
h = 1	HAR-ln(RV)-ln(CJ)
h = 5	MIDAS-ln(RV)-ln(CJ)
h = 20	HAR-ln(RV)
h = 1	HAR-(RV)1̂/2-(RV)1̂/2
h = 5	HAR-(RV)1̂/2-(CJ)1̂/2
h = 20	MIDAS-(RV)1̂/2-(CJ)1̂/2

Conclusions

- IN SAMPLE analysis shows a non-linear behavior for both models with error measures smaller at a weekly horizon.
- The same results are shown in the OUT of SAMPLE analysis for the HAR Models. While, MIDAS Models provide a better performance at a monthly horizon.
- To pick the best model at every horizon, DM test are carried out. Overall results show that:
    - $\text{H} = 1$ HAR Models provide a better forecast accuracy.
    - $\text{H} = 5$ it seems again that HAR Models outperform MIDAS.
    - $\text{H} = 20$, MIDAS provide a better forecast.

Future Research

- Forecast encompassing test for forecast combination of the two models.
- Macro-economic announcement to better predict the jump component.

THANK YOU!