Article

GARMA, HAR and Rules of Thumb for Modelling Realized Volatility

David Edmund Allen [1,2,3,*] and Shelton Peiris [1]

1 School of Mathematics and Statistics, University of Sydney, Camperdown, NSW 2006, Australia; shelton.peiris@sydney.edu.au
2 Department of Finance, Asia University, Taichung 41354, Taiwan
3 School of Business and Law, Edith Cowan University, Joondalup, WA 6027, Australia
* Correspondence: profallen2007@gmail.com

Abstract: This paper features an analysis of the relative effectiveness, in terms of the Adjusted R-Square, of a variety of methods of modelling realized volatility (RV), namely the use of Gegenbauer processes in Auto-Regressive Moving Average format, GARMA, as opposed to Heterogenous Auto-Regressive HAR models and simple rules of thumb. The analysis is applied to two data sets that feature the RV of the S&P500 index, as sampled at 5 min intervals, provided by the OxfordMan RV database. The GARMA model does perform slightly better than the HAR model, but both models are matched by a simple rule of thumb regression model based on the application of lags of squared, cubed and quartic, demeaned daily returns.

Keywords: GARMA; Gegenbauer processes; HAR models; realized volatility; rules of thumb

1. Introduction

This paper features the application of time series modelling techniques to the modelling of RV. In this particular case, RV sampled at 5 min, as supplied by the OxfordMan Institute (see: https://realized.oxford-man.ox.ac.uk/data, accessed on 22 March 2022), was used. We compare the effectiveness of Gegenbauer processes in Auto-Regressive Moving Average format, GARMA models, as opposed to Heterogenous Auto-Regressive HAR models and other rules of thumb, based on demeaned squared, cubed and quartic daily returns, as methods for modelling and forecasting daily 5 min RV. The aim of the paper is to explore whether sophisticated models out-perform simple rules of thumb.

Over the past 100 years, considerable advances have been made in time series modelling. Yule (1926) and Slutsky (1927) developed the stochastic analysis of time series and developed the concepts of autoregressive (AR) and moving average (MA) models. Box and Jenkins (1970) suggested methods for applying autoregressive moving average (ARMA) or autoregressive integrated moving average (ARIMA) models to find the best fit of a time-series model to past values of a time series.

A contrary view of the approriateness of this approach has been promoted by Commandeur and Koopman (2007), who suggested that the Box–Jenkins approach is fundamentally problematic. They have championed the adoption of alternative state-space methods to counter the contention that many real economic series are not truly stationary, despite differencing.

In the 1980s, attention switched to the consideration of the issues related to stationarity and non-stationary time series, fractional integration and cointegration. Granger and Joyeux (1980) and Hosking (1981) focussed attention on fractionally integrated autoregressive moving average (ARFIMA or FARIMA) processes. Unit root testing to assess the stationarity of a time series became established via the application of the Dickey–Fuller test, following the work of Dickey and Fuller (1979).

Engle and Granger (1987) developed the concept of cointegration, whereby two time series might be individually integrated and non-stationary I(1), but some linear combination of them might possess a lower order of integration and be stationary, in which case the series are said to be cointegrated. Many of these conceptual developments have important applications to economic and financial time series, and to economic theory in these discipline areas.

One of the common features of many time series of financial data sets is that the variance of the series is not homoscedastic, and that these features concerned are autocorrelated. The Autoregressive Conditional Heteroskedasticity (ARCH) model that incorporates all past error terms was developed by Engle (1982). Bollerslev (1986) further generalised it to GARCH by including lagged term conditional volatility. Therefore, GARCH predicts that the best indicator of future variance is a weighted average of long-run variance, the predicted variance for the current period, and any new information in this period, as captured by the squared residuals. GARCH models provide an estimate of the conditional variance of a financial price time series.

The concept of RV and associated metrics were developed by Andersen et al. (2001), Andersen et al. (2003), and Barndorff-Nielsen and Shephard (2002). This alternative approach measures the variance directly from the observed values of the price series. These RV measures are theoretically sound, high-frequency, nonparametric-based estimators of the variation of the price path of an asset during the times at which the asset trades frequently on an exchange.

The modelling of the variance of financial time series and the use of RV is the focus of this paper. In the empirical analysis, we use RV 5 min estimates from Oxford-Man for the S&P500 Index as the RV benchmark. Their database contains daily (close to close) financial returns and a corresponding sequence of daily realized measures $rm_1, rm_2, \ldots \ldots, rm_T$.

Corsi (2009, p. 174) suggests "an additive cascade model of volatility components defined over different time periods. The volatility cascade leads to a simple AR-type model in the realized volatility with the feature of considering different volatility components realized over different time horizons, and which he termed as being a "Heterogeneous Autoregressive model of Realized Volatility". We make use of the Corsi (2009) HAR model to model RV in some of the empirical tests included in this paper.

However, the main focus of this paper is the application of Gegenbauer processes to the modelling of RV. Gegenbauer processes were introduced by Hosking (1981) and further developed by Anděl (1986) and Gray et al. (1989). The latter proposed the class of time series models known as Gegenbauer ARMA, or, as abbreviated, GARMA processes, which are the central focus of this paper.

In the current paper, we compare the effectiveness of GARMA models, as opposed to HAR models and other rules of thumb, based on demeaned squared daily returns, as methods for modelling and forecasting daily 5 min RV.

The paper is a further companion piece to two previous studies in the topic's general area, namely, Allen and McAleer (2020) and Allen (2020), that compared the effectiveness of stochastic volatility (SV), vanilla GARCH and HAR models, as opposed to simple rules of thumb, in their effectiveness as tools for capturing the RV of major stock market indices.

The current paper concentrates on the S&P500 index and examines whether GARMA, HAR or simple rules of thumb better capture the RV sampled at 5 min intervals, as provided by Oxford-Man, of the S&P500 Index. Thus, the central concern is what is the best method of capturing the long memory properties of a historical time series of RV5 for the S&P500 index. This is in contrast with the two previously mentioned studies, which contrasted the effectiveness of the volatility models per se.

The benchmark is provided by the estimates of RV5, in a sample of daily estimates of RV5, running from 8 May 1997 to 30 August 2013 with 4096 observations, plus a longer-period sample of RV5, also based on the S&P500 Index, running from 4 January 2000 to 30 April 2020, comprising 5099 observations. This is the same data set as used in Allen (2020).

The motivation for this paper is provided by Poon and Granger (2005, p. 507), who observed that: "as a rule of thumb, historical volatility methods work equally well compared with more sophisticated ARCH class and SV models." This paper similarly seeks to explore whether simple rules of thumb, in this case based on the use of a regression model featuring squared demeaned daily returns, with the subsequent addition of cubed and quartic powers of them, perform as well as more sophisticated time series models.

The paper is divided into four sections: Section 2 reviews the literature and econometric methods employed. Section 3 presents the results, and Section 4 presents the conclusions.

2. Previous Work and Econometric Models

Recent reviews of the literature on the nature and applications of Gegenbauer processes are provided by Hunt et al. (2021) and Dissanayake et al. (2018). Peiris et al. (2005) and Peiris and Thavaneswaran (2007) considered long memory models driven by heteroskedastic GARCH errors. Peiris and Asai (2016) returned to this topic, and Guegan (2000) combined Gegenbauer processes with integrated GARCH (GIGARCH) to include the attributes of long memory, seasonality and heteroskedasticity at the same time, in the modelling of volatility.

2.1. The Basic GEGENBAUER Model

Let $\{X_t, t = 1, 2, \ldots,\}$ be a stationary random process with the autocovariance $\gamma(k) = Cov(X_t, X_{t+|k|})$, and the autocorrelation function $\rho(k) = \gamma(k)/\gamma(0)$, where $k = 1, 2 \ldots$. The spectral density function (sdf) is denoted by:

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \rho(k)e^{-\omega k}, \quad -\pi \leq \omega \leq \pi,$$

where ω is the Fourier frequency.

There are various ways in which the long memory component of the Gegenbauer model can be specified, as discussed in Dissanayake et al. (2018). In the analysis that follows, we utilise the R package GARMA, as developed by Hunt (2022a).

A Gegenbauer process is a long memory process generated by the dynamic equation:

$$(1 - 2uB + B^2)^\delta X_t = \epsilon_t, \tag{1}$$

where $| u | < 1$, $\delta \in (0, 0.5)$ and ϵ_t is a short memory process characterised by a positive and bounded spectral density, $f_\epsilon(\omega)$. If $\epsilon_t \sim WN(0, \sigma^2)$, (1) is a Gegenbauer process of order δ or a GARMA $(0, \delta, 0, u)$ process. Dissanayake et al. (2018) mention that (1) complies with the definition of a long memory process at the frequency $\omega_0 arccos(u)$. According to (1), X_t arises from filtering the process ϵ_t by the infinite impulse response

$$(1 - 2uB + B^2)^{-\delta} = \sum_{j=0}^{\infty} C_j^\delta(u)B^j.$$

It can be shown that a stationary Gegenbauer process contains an unbounded spectrum at ω_0 and is long memory when $0 < \delta < \frac{1}{2}$. This special frequency ω_0 is called the Gegenbauer or G-frequency, as outlined by Dissanayake et al. (2018, p. 416).

The k-factor GARMA model as fit by the GARMA package, as in Hunt (2022a), is specified as:

$$\phi(B) \prod_{i=1}^{k} (1 - 2u_i B + B^2)^{d_i} (1 - B)^\delta (X_t - u) = \theta(B)\epsilon_t \tag{2}$$

- where $\phi(B)$ represents the short-memory autoregressive component of order p,
- $\theta(B)$ represents the short memory moving average component of order q,
- $(1 - 2u_i B + B^2)^{d_i}$ represents the long-memory Gegenbauer component (there may in general be k of these),

- d represents integer differencing (currently only = 0 or 1 is supported),
- X_t represents the observed process,
- ϵ_t represents the random component of the model—these are assumed to be uncorrelated but identically distributed variates,
- B represents the Backshift operator, defined by $BX_t = X_{t-1}$.

When $k = 0$, then this is just a short memory model, as would be represented by an ARIMA model.

Dissanayake (2015) considered a class of Gegenbauer processes generated by Gaussian white noise and GARCH errors and suggested that many other models could be nested within this framework. He explored a number of related issues and used state space modelling to explore seasonal Gegenbauer processes.

Phillip (2018) applied a Gegenbauer long memory stochastic volatility model with leverage and a bivariate Student's t-error distribution to describe the innovations of the observation and latent volatility jointly and demonstrated its effectiveness in applications to Cryptocurrency time series. He demonstrated by means of MCMC runs for various values of $[u; d]$, that when d is low, and as $u \to 0$. the convergence of u to its true value gets slower. This is an expected result, since as $d \to 0$, the process has less information and becomes "less long-memory".

2.2. Heterogenous Autoregressive Model (HAR)

Corsi (2009, p. 174) suggests "an additive cascade model of volatility components defined over different time periods. The volatility cascade leads to a simple AR-type model in the realized volatility with the feature of considering different volatility components realized over different time horizons and which he termed as a Heterogeneous Autoregressive model of Realized Volatility". Corsi (2009) suggests that his model can reproduce the main empirical features of financial returns (long memory, fat tails, and self-similarity) in a parsimonious way. He writes his model as:

$$\sigma_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \widetilde{\omega}_{t+1d}^{(d)}, \tag{3}$$

where $\sigma^{(d)}$ is the daily integrated volatility, and $RV_t^{(d)}$, $RV_t^{(w)}$ and $RV_t^{(m)}$ are, respectively, the daily, weekly, and monthly (ex post) observed realized volatilities.

2.3. Historical Volatility Model (HISVOL)

Poon and Granger (2005) discuss various practical issues involved in forecasting volatility. They suggest that the HISVOL model has the following form:

$$\hat{\sigma}_t = \phi_1 \sigma_{t-1} + \phi_2 \sigma_{t-2} + \ldots + \phi_\tau \sigma_{t-\tau}, \tag{4}$$

where $\hat{\sigma}_t$ is the expected standard deviation at time t, ϕ is the weight parameter, and σ is the historical standard deviation for periods indicated by the subscripts. Poon and Granger (2005) suggest that this group of models include the random walk, historical averages, autoregressive (fractionally integrated) moving average, and various forms of exponential smoothing that depend on the weight parameter ϕ.

We use a simple form of this model in which the estimate of σ is the previous day's demeaned squared return. Poon and Granger (2005) review 66 previous studies and suggest that implied standard deviations appear to perform best, followed by historical volatility and GARCH, which have roughly equal performance.

Barndorff-Nielsen and Shephard (2003) point out that taking the sums of squares of increments of log prices has a long tradition in the financial economics literature. See, for example, Poterba and Summers (1986), Schwert (1989), Taylor and Xu (1997), Christensen and Prabhala (1998), Dacorogna et al. (1998), and Andersen et al. (2001). Shephard and Sheppard (2010) p. 200, footnote 4) note that: "Of course, the most basic realized measure is the squared daily return". We utilise this approach as the basis of our historical volatility

model. Furthermore, Perron and Shi (2020) show how the squared low-frequency returns can be expressed in terms of the temporal aggregation of a high-frequency series in relation to volatility measures.

3. Results

3.1. The Data Sets

To expedite a direct comparison with previous work on the HAR model, we use the R library package 'HARModel' by Sjoerup (2019). This contains data featuring RV measures from the SP500 index from April 1997 to August 2013, and we use the RV5 measures on the S&P500 Index sampled at 5 min intervals.

Table 1 provides a statistical description of this RV5 data set, together with that of another slightly longer data set, taken from 2000 to 2020, also featuring S&P500 index RV5 data taken from Allen and McAleer (2020). Both data sets feature RV5 estimates taken from the Oxford-Man Realized library.

Table 1. Descriptive statistics RV5 data sets.

Descriptor	S&P500 1997–2013	S&P500 2000–2020
Number of Observations	4096	5099
Minimum	0.04329	0.00000122
Maximum	60.56	0.0074
median	0.6294	0.0000471
mean	1.1752	0.000112
Standard Deviation	2.3151	0.000269

NB: The data taken from the R library package HARModel on RV5 was scaled up in the package.

One of the features of estimates of RV is that the data time series displays long-memory characteristics. Long memory refers to the association between observations in a time series that are ever larger sample intervalling and is also referred to as long-range dependence. It basically refers to the level of statistical dependence between two points in the time series sampled at increasing intervals.

Figure 1 displays the long memory characteristics of the two RV5 time series that we analyse. The first panel in the two plots displays the basic series of RV5, and the two large spikes in RV5 correspond to the effects on volatility of the Global Financial Crisis (GFC) that occurred in 2008. The two panels marked ACF and PACF refer to the autocorrelation and partial autocorrelation statistics.

The R GARMA package, Hunt (2022a), was used to generate the two graphs. The program was instructed to use 100 lags of daily observations. The blue lines in the bottom two sets of panels display the standard error bands. The long memory properties of RV5 are apparent in both sets of diagrams in that the ACF statistics remain well outside the error bands for 100 lags, and the PACF is outside the error bands for up to 30 lags.

These long memory characteristics are used in Corsi's (2009) HAR model and will be a feature of the Gegenbauer models that we fit to the data sets.
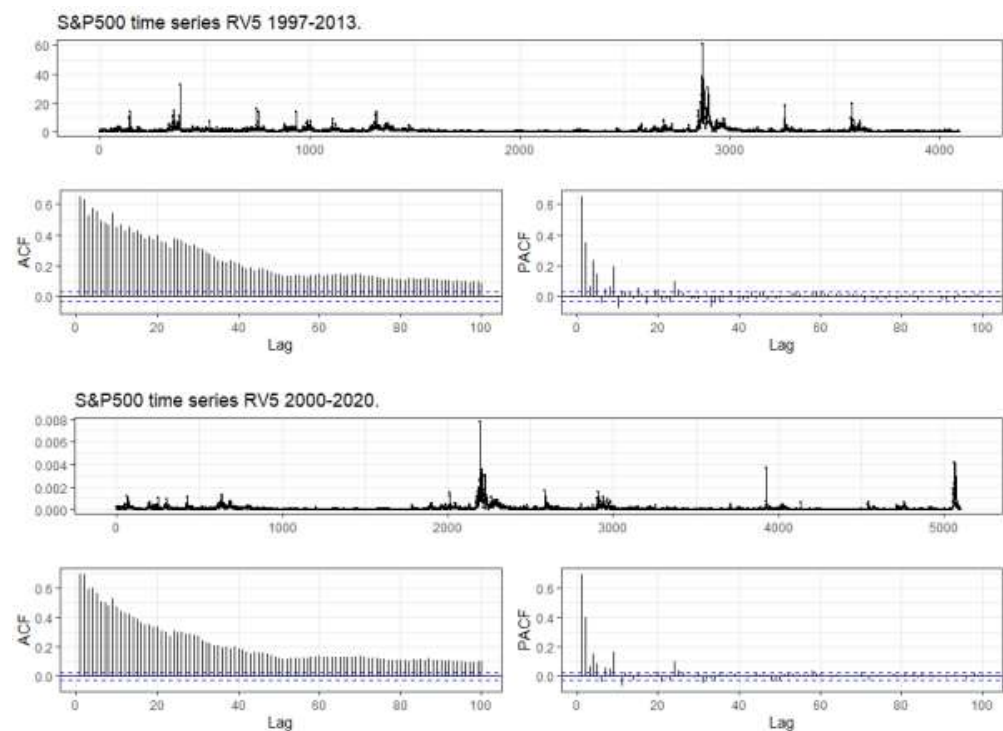
Figure 1. Plots of the RV5 samples for the S&P500 index and their long-range dependence. (Dashed blue line represents error bands).

3.2. The Basic HAR Model

Table 2 provides summary descriptions of the HAR models fitted to the two RV5 data sets, and Figure 2 provides plots of the fits. The results presented in Table 2 show that the basic HAR model does an excellent job of capturing the time series properties of RV5. All the estimates are significant at a one percent level, as are the F statistics, and the Adjusted R-squares are 52 percent for the period 1997–2013 and 56 percent for the period 2000–2020, respectively.

Table 2. Summary of the HAR models fitted to the RV5 data sets.

Coefficient	Estimate	Standard Error	*t*. Value
	S&P500 1997–2013 RV5		
beta0	0.11231	0.03065	3.664 ***
beta1	0.22734	0.01870	12.157 ***
beta5	0.49035	0.03144	15.595 ***
beta22	0.18638	0.02813	6.624 ***
Adjusted R-squared	0.5221		
F-Statistic	1484 ***		
	S&P500 2000–2020 RV5		
beta0	1.218×10^{-5}	2.877×10^{-6}	4.235 ***
beta1	2.703×10^{-1}	1.704×10^{-2}	15.858 ***
beta5	5.295×10^{-1}	2.633×10^{-2}	20.108 ***
beta22	9.134×10^{-2}	2.225×10^{-1}	4.105***
Adjusted R-squared	0.5608		
F-Statistic	2162 ***		

Note: *** Indicates significant at the 1% level.

The plots in Figure 2 confirm this but do suggest that the large periodic peaks in RV5 are not captured so effectively by the HAR model. The question remains as to whether the Gegenbauer model will perform more effectively.
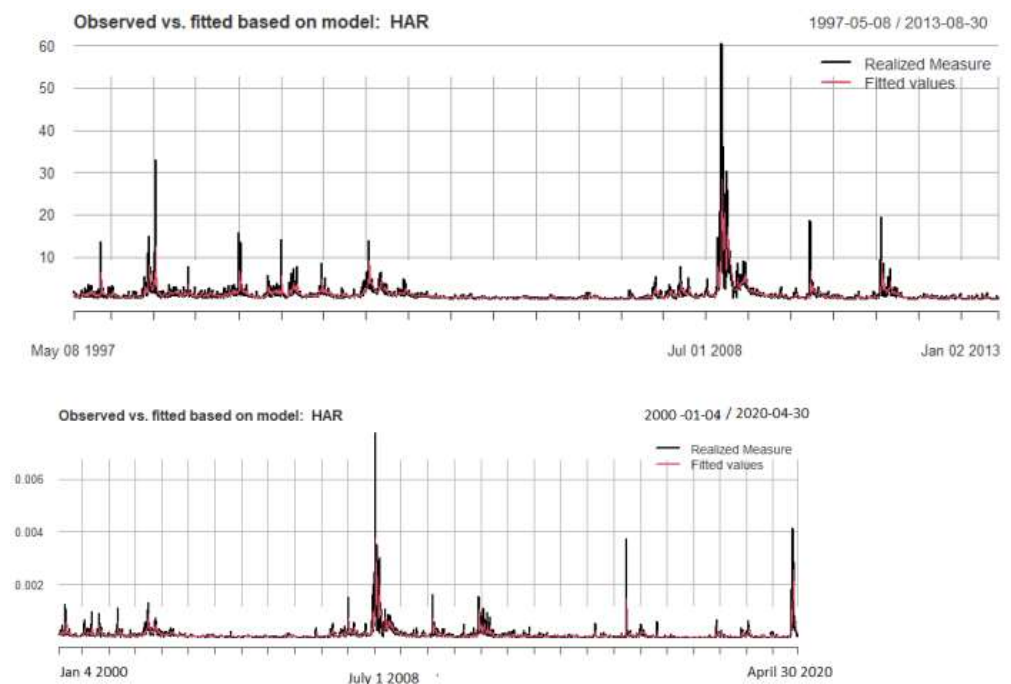
Figure 2. Plots of fitted HAR models.

3.3. Gegenbauer Results

The R library package (GARMA), Hunt (2022a), was used to fit GARMA models to the RV5 series for the S&P500 index as sourced from the Oxford-Man library. This shorter RV series, for the S&P500 from 1997 to 2013, was taken from the HARmodel R library package, Sjoerup (2019).

A detailed summary of the methods used in the GARMA package is available in Hunt (2022b). The GARMA package provides the ability to fit stationary, univariate GARMA models to a time series and to forecast from those models. The garma() function in the GARMA package is the main function for estimating the parameters of a GARMA model. It provides three methods of parameter estimation: the Whittle method, (Whittle 1953), the conditional sum-of-squares (CSS) method, (for a discussion, See Hunt (2022b) chp. 2, eq. 2.3.2) and the WLLS method. The latter, the Whittle Log Least-Squares method, was proposed by Hunt (2022b, chp. 3). The Whittle method was used in the estimations reported in the paper.

A summary of the Gegenbauer model estimated for this data is shown in Table 3. A potential advantage of the Gegenbauer model is that it is non-linear and more flexible than the HAR model.

The results of a regression of the fit from the Gegenbauer model for this data on the actual RV5 estimates for the S&P500 index for this period are shown in Table 4. The HAR model regression for this period, shown in the first half of Table 2, had an Adjusted R-squared of 0.52. The result for the Gegenbauer model estimation is an Adjusted R-squared of 0.567, and so the non-linear model does show an increased explanatory power.

We also fitted the Gegenbauer model to the longer time-period of RV estimates running from 2000 to 2020 and present the results in Table 5.

We regressed the actual daily RV5 series for the longer period from 2000 to 2020, and the results are shown in Table 6. The slope coefficient is significant at the 1 per cent level and is very close to 1, whilst the Adjusted R-square is 0.59, and the F statistics for the regression, with a value of 7326, is also significant at the 1 percent level. The Adjusted R-square for the HAR model for the same period was 0.56, so the Gegenbauer model provides a marginally better fit than the HAR model.

Table 3. Gegenbauer estimation for RV5 with GARMA for the S&P500 from 1997–2013, constant included with no trend.

Series	Intercept	U_1	δ_1	ar1	ar2	ar3	ar4	ar5
coefficient	1.139×10^{-4}	0.9794239	0.33368	−0.27435	2.215×10^{-11}	$−5.991 \times 10^{-11}$	0.09145	0.08230
S.E.	7.430×10^{-8}	0.0001457	0.03893	0.07677	5.745×10^{-2}	3.259×10^{-2}	0.02167	0.01137
Series	**ar6**	**ar7**	**ar8**	**ar9**	**ar10**	**ar11**	**ar12**	**ar13**
coefficient	−0.02743	9.408×10^{-11}	0.02743	0.2469	0.16461	0.08230	0.1097	0.10974
S.E.	0.01075	1.034×10^{-2}	0.01050	0.0118	0.02678	0.02949	0.0260	0.02663
Series	**ar14**	**ar15**	**ar16**	**ar17**	**ar18**	**ar19**	**ar20**	**ar21**
coefficient	0.02743	0.0823	0.0823	0.02743	1.320×10^{-10}	1.125×10^{-10}	$−6.332 \times 10^{-12}$	−0.03658
S.E.	0.02655	0.0205	0.0208	0.02007	1.578×10^{-2}	1.237×10^{-2}	1.103×10^{-2}	0.01056
Series	**ar22**	**ar23**	**ar24**	**ar25**	**ar26**	**ar27**	**ar28**	**ar29**
coefficient	−0.02743	−0.10974	0.04572	−0.02743	−0.02743	0.02743	4.256×10^{-11}	$−4.386 \times 10^{-11}$
S.E.	0.01127	0.01226	0.01763	0.01204	0.01327	0.01359	1.129×10^{-2}	1.117×10^{-2}
Series	**ar30**	**Gegenbauer frequency**	**Gegenbauer period**	**Gegenbauer Exponent**				
coefficient	0.02743	0.0323	30.9197	0.3337				
S.E.	0.01050							

Table 4. Regression of RV5 on Gegenbauer model estimates, 1997–2013.

	Coefficient	S.E.
Constant	0.0003596	0.0287058
RV5	0.9994219 ***	0.0136477
Adjusted RSQ	0.567	
F. Statistic	5363	

Note: *** Indicates significant at the 1% level.

Table 5. Gegenbauer estimation for RV5 with GARMA for the S&P500 from 2000 to2020, constant included with no trend.

Series	Intercept	U_1	δ_1	ar1	ar2	ar3	ar4	ar5
coefficient	1.175	0.9776774	0.12974	0.09841	0.18564	−0.06645	0.11861	0.1606
S.E.	8.315	0.0004903	0.03286	0.06504	0.02608	0.01162	0.01376	0.0119
Series	**ar6**	**ar7**	**ar8**	**ar9**	**ar10**	**ar11**	**ar12**	**ar13**
coefficient	−0.01016	−0.02717	0.02901	0.23708	−0.02952	0.01219	0.05422	0.05570
S.E.	0.01662	0.01295	0.01187	0.01259	0.02260	0.01466	0.01351	0.01408
Series	**ar14**	**ar15**	**ar16**	**ar17**	**ar18**	**ar19**	**ar20**	
coefficient	−0.02716	0.05100	0.05079	−0.04578	−0.03084	0.01445	0.04350	
S.E.	0.01417	0.01183	0.01241	0.01240	0.01136	0.01165	0.01133	
Series	**Gegenbauer Frequency**	**Gegenbauer period**	**Gegenbauer Exponent**					
coefficient	0.0337	29.6812	0.1297					

Table 6. Regression of RV5 on Gegenbauer model estimates, 2000–2020.

	Coefficient	S.E.
Constant	6.72543×10^{-7}	2.77936×10^{-6}
RV5	0.993054 ***	0.0116015
Adjusted RSQ	0.589660	
F. Statistic	7326.850 ***	

Note: *** Indicates significant at the 1% level.

3.4. How Do Rule of Thumb Approaches Perform?

The next issue is how squared demeaned end-of-day returns perform as a simple rule-of-thumb method to explain RV5. Table 7 presents the results of the regression of RV5 for the longer period of 2000 to 2020 on 20 lags of squared demeaned daily returns.

Table 7. Regression of RV5 on squared demeaned daily returns for 2000 to 2020; OLS, using observations 21–5099 ($T = 5079$); Dependent variable: rv5.

Variable	Coefficient	Std. Error	t-Ratio	p-Value
const	0.0000302614	0.00000284505	10.64	0.000 ***
SQSPRET_1	0.174294	0.00555060	31.4	0.000 ***
SQSPRET_2	0.0935967	0.00558548	16.76	0.000 ***
SQSPRET_3	0.0527928	0.00587581	8.99	0.000 ***
SQSPRET_4	0.0280810	0.00587693	4.78	0.000 ***
SQSPRET_5	0.0591188	0.00587284	10.07	0.000 ***
SQSPRET_6	0.0386591	0.00589632	6.56	0.000 ***
SQSPRET_7	0.0127360	0.00592375	2.15	0.031 **
SQSPRET_8	0.0125674	0.00590739	2.127	0.033 **
SQSPRET_9	0.0460141	0.00590538	7.79	0.000 ***
SQSPRET_10	0.0109619	0.00588044	1.86	0.062 *
SQSPRET_11	−0.0125986	0.00588112	−2.14	0.032 **
SQSPRET_12	0.00679795	0.00590660	0.15	0.249
SQSPRET_13	0.000835242	0.00590814	0.14	0.888
SQSPRET_14	−0.00784497	0.00592586	−1.32	0.186
SQSPRET_15	−0.00255057	0.00589736	−0.433	0.665
SQSPRET_16	−0.0109373	0.00587609	−1.861	0.063
SQSPRET_17	0.00499043	0.00587928	0.848	0.396
SQSPRET_18	0.0124227	0.00592245	2.098	0.036 **
SQSPRET_19	0.0124595	0.00563006	2.213	0.027 **
SQSPRET_20	−0.00716919	0.00559011	−1.282	0.12

Mean dependent var	0.000112	S.D. dependent var		0.000269
Sum squared resid	0.000168	S.E. of regression		0.000182
R^2	0.543145	Adjusted R^2		0.541339
$F(20, 5058)$	300.6678	p-value(F)		0.000000
Log-likelihood	36531.92	Akaike criterion		−73021.83
Schwarz criterion	−72884.64	Hannan–Quinn		−72973.79
$\hat{\rho}$	0.281412	Durbin–Watson		1.437144

Note: ***, ** and * denotes significance at 1%, 5% and 10%.

It can be seen in Table 7 that 20 lags of squared demeaned returns do not perform quite as well as the Gegenbauer or HAR models but still have an Adjusted R-Square of 54 percent, which is a marginal 2 percent less than the HAR model and 5 percent less than the Gegenbauer model. Only 4 of the 20 lags used in this rule-of-thumb approach are insignificant. The Durbin–Watson statistic of 1.43 suggests that a considerable amount of autocorrelation remains in the residuals.

The application of Ramsey Reset tests (Ramsey 1969), suggests that squares and cubes of the explanatory variable could add to the power of the regression. Table 8 reports the results of adding 10 lags of cubed demeaned SPRET and 10 lags of demeaned SPRET to the power 4.

This is essentially another non-linear model, but admittedly, we now have 40 explanatory variables in the model in the form of lags of three explanatory variables. The Adjusted R-Square now increases to over 59 percent, which matches the power of the Gegenbauer model. Admittedly, the Durbin–Watson statistic is still a relatively low 1.54. This suggests that there is still autocorrelation in the residuals, which could be exploited further in enhanced modifications of the model.

The above regression is quite clumsy and contains some redundant terms. An anonymous reviewer suggested that we apply a *lasso* technique to reduce the size of the regression.

This is not entirely consistent with using a 'rule of thumb' approach, and it would be a simple matter to drop the redundant terms in the regression above.

Table 8. Regression of RV5 on squared, cubed and quartic daily demeaned returns for 2000 to 2020; OLS, using observations 21–5099 (T = 5079); variable: rv5.

	Coefficient	Std. Error	t-Ratio	p-Value
const	0.000082971	0.0000303009	1.594	0.1110
SQSPRET_1	0.231351	0.00958455	24.14	0.000 ***
SQSPRET_2	0.110377	0.00959229	11.51	0.000 ***
SQSPRET_3	0.118282	0.00966599	12.24	0.000 ***
SQSPRET_4	0.0666821	0.00969419	6.879	0.000 ***
SQSPRET_5	0.0642053	0.00980063	6.551	0.000 ***
SQSPRET_6	0.0669488	0.00989797	6.764	0.000 ***
SQSPRET_7	0.0345805	0.00995553	3.473	0.0005 ***
SQSPRET_8	0.0169334	0.00994948	1.702	0.0888 *
SQSPRET_9	0.0254747	0.00993460	2.564	0.0104 **
SQSPRET_10	0.0227310	0.00975198	2.331	0.0198 **
SQSPRET_11	−0.00478774	0.00583425	−0.8206	0.4119
SQSPRET_12	0.0213587	0.00588183	3.631	0.0003 ***
SQSPRET_13	0.00934631	0.00582412	1.605	0.1086
SQSPRET_14	−0.00178946	0.00578095	−0.3095	0.7569
SQSPRET_15	−0.000564490	0.00574051	−0.09833	0.9217
SQSPRET_16	−0.00229510	0.00575739	−0.3986	0.6902
SQSPRET_17	0.00247291	0.00572209	0.4322	0.6656
SQSPRET_18	0.00617666	0.00576039	1.072	0.2837
SQSPRET_19	0.00692568	0.00554060	1.250	0.2114
SQSPRET_20	−0.0188372	0.00539186	−3.494	0.0005 ***
CUSPRET_1	−0.534907	0.0592631	−9.026	0.0000 ***
CUSPRET_2	−0.574605	0.0614957	−9.344	0.0000 ***
CUSPRET_3	−0.542658	0.0625408	−8.677	0.0000 ***
CUSPRET_4	−0.421881	0.0625029	−6.750	0.0000 ***
CUSPRET_5	−0.581211	0.0623480	−9.322	0.0000 ***
CUSPRET_6	−0.406321	0.0625585	−6.495	0.0000 ***
CUSPRET_7	−0.306419	0.0622486	−4.923	0.0000 ***
CUSPRET_8	−0.0728304	0.0624320	−1.167	0.2434
CUSPRET_9	0.0440999	0.0614311	0.7179	0.4729
CUSPRET_10	−0.143643	0.0603871	−2.379	0.0174
sq_SQSPRET_1	−11.1016	0.991204	−11.20	0.0000 ***
sq_SQSPRET_2	−5.00005	0.992337	−5.039	0.0000 ***
sq_SQSPRET_3	−10.1474	1.01023	−10.04	0.0000 ***
sq_SQSPRET_4	−7.20797	1.01302	−7.115	0.0000 ***
sq_SQSPRET_5	−2.06748	1.01765	−2.032	0.0422 **
sq_SQSPRET_6	−5.51711	1.01948	−5.412	0.0000 ***
sq_SQSPRET_7	−4.90765	02938	−4.768	0.0000 ***
sq_SQSPRET_8	−0.981766	1.03193	−0.9514	0.3415
sq_SQSPRET_9	2.00887	1,02378	1.962	0.0498 **
sq_SQSPRET_10	−2.40620	1.00723	−2.389	0.0169 **

Mean dependent var	0.000112	S.D. dependent var		0.000269
Sum squared resid	0.000148	S.E. of regression		0.000172
R^2	0.596908	Adjusted R^2		0.593707
$F(40, 5038)$	186.5095	p-value(F)		0.000000
Log-likelihood	36849.86	Akaike criterion		−73617.72
Schwarz criterion	−73349.87	Hannan–Quinn		−73523.92
$\hat{\rho}$	0.228344	Durbin–Watson		1.543216

Note: ***, ** and * denotes significance at 1%, 5% and 10%.

The adaptive lasso regression uses different penalties (weights) for different regressors when running a lasso regression. Under certain conditions, as applied to those weights, the

results will have the so-called oracle property, see Zou (2006) as opposed to the standard lasso approach.

Zou (2006) derives a necessary condition for the lasso variable selection to be consistent. His version of the lasso, the adaptive lasso, employs adaptive weights that are used for penalizing different coefficients in the λ_1 penalty. He demonstrates that the adaptive lasso enjoys oracle properties, namely, that it performs as well as if the true underlying model were given in advance. Zou (2006) suggests that $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response vector and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T, j = 1, \ldots, p$, are the linearly independent predictors. He lets $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ be the predictor matrix. He assumes that $E[y \,|\, \mathbf{x}] = \beta^*_1 x_1 + \cdots + \beta^*_p x_p$. The data are assumed to be centred so the intercept is not included in the regression equation. He lets $A = \{j: \beta_j^* \neq 0\}$ and then assumes that $|A| = p_0 < p$. This implies that the true model depends only on a subset of the predictors.

If we denote by $\widehat{\beta}(\delta)$, the coefficient estimator is produced by a fitting procedure δ. Then, the arguments of Fan and Li (2001) can be adopted and δ can be termed as being an oracle procedure if $\widehat{\beta}(\delta)$ (asymptotically) has the following oracle properties:

- Identifies the right subset model, $\{j: \widehat{\beta}_j \neq 0\} = A$

- Has the optimal estimation rate, $\sqrt{n}\,\widehat{\beta}\,((\delta)A - \boldsymbol{\beta^*}A) \to d\,\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma^*})$, where $\boldsymbol{\Sigma^*}$ is the covariance matrix knowing the true subset model.

The lasso is a regularization technique for simultaneous estimation and variable selection Tibshirani (1996). The lasso estimates are defined as:

$$\widehat{\beta}(lasso) = \underset{\beta}{\arg\min} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is a nonnegative regularization parameter. The second term in the expression above is the so-called "ℓ_1 penalty," which is crucial for the success of the lasso. The lasso continuously shrinks the coefficients toward 0 as λ increases, and some coefficients are shrunk to exactly 0 if λ is sufficiently large.

Zou proposes the adaptive lasso, in which adaptive weights are used for penalizing different coefficients in the ℓ_1 penalty and he demonstrates that the adaptive lasso enjoys the oracle properties. He employs $\widehat{\beta}(ols)$ to construct the adaptive weights in the adaptive lasso; and suggests that the computation procedure involves the employment of the LARS algorithm (Efron et al. 2004).

We implemented this version of the adaptive lasso technique using a GRETL function package containing code submitted by Schreiber (2023). Table 9 shows the regression screening results chosen by the technique as applied to the variables in Table 8.

Table 9. Regressors chosen by the adaptive lasso technique applied to the variables in Table 8.

Variable	Weight
sq_DMSPRET_1	0.23908
sq_DMSPRET_2	0.12086
sq_DMSPRET_3	0.12002
sq_DMSPRET_4	0.063485
sq_DMSPRET_5	0.059401
sq_DMSPRET_6	0.067899
sq_DMSPRET_7	0.026704
sq_DMSPRET_9	0.012712
sq_DMSPRET_12	0.012814
CUSPRET_1	−0.50290
CUSPRET_2	−0.54575
CUSPRET_3	−0.48000

Table 9. *Cont.*

Variable	Weight
CUSPRET_4	−0.36132
CUSPRET_5	−0.55055
CUSPRET_6	−0.32513
CUSPRET_7	−0.26524
CUSPRET_9	0.026202
CUSPRET_10	−0.074585
sq_sq_DMSPRET_1	−11.184
sq_sq_DMSPRET_2	−5.4840
sq_sq_DMSPRET_3	−9.7038
sq_sq_DMSPRET_4	−6.4668
sq_sq_DMSPRET_5	−1.3524
sq_sq_DMSPRET_6	−4.7695
sq_sq_DMSPRET_7	−3.7897
sq_sq_DMSPRET_8	1.0471
sq_sq_DMSPRET_9	2.6983

These variables chosen were then applied in a new variant of the rule of thumb OLS regression of RV5 on the various lags of squared, cubed and quartic demeaned returns chosen by the adaptive lasso technique. The results are shown in Table 10.

The results of the application of the adaptive lasso technique shown in Table 10 confirm the previous findings reported in Table 8, namely, that a rule-of-thumb method using squared, cubed, and quartic end-of-day returns in the explanation of RV5 matches those obtained by application of the Gegenbauer technique.

Table 10. Regression of RV5 on squared, cubed and quartic daily demeaned returns for 2000 to 2020 as chosen by the adaptive lasso technique.

	Coefficient	Std. Error	*t*-Ratio	*p*-Value
Const	0.00000637199	0.00000298585	2.134	0.0329 **
sq_DMSPRET_1	0.232117	0.00945099	24.56	6.33×10^{-126} ***
sq_DMSPRET_2	0.117260	0.00937219	12.51	2.15×10^{-35} ***
sq_DMSPRET_3	0.118641	0.00943350	12.58	9.71×10^{-36} ***
sq_DMSPRET_4	0.0668396	0.00939540	7.114	1.28×10^{-12} ***
sq_DMSPRET_5	0.0652956	0.00954500	6.841	8.81×10^{-12} ***
sq_DMSPRET_6	0.0725886	0.00967025	7.506	7.14×10^{-14} ***
sq_DMSPRET_7	0.0420426	0.00958814	4.385	1.18×10^{-5} ***
sq_DMSPRET_9	0.0243582	0.00964983	2.524	0.0116 **
sq_DMSPRET_12	0.0203625	0.00534193	3.812	0.0001 ***
CUSPRET_1	−0.523256	0.0581730	−8.995	3.29×10^{-19} ***
CUSPRET_2	−0.577212	0.0587275	−9.829	1.35×10^{-22} ***
CUSPRET_3	−0.525699	0.0607949	−8.647	7.00×10^{-18} ***
CUSPRET_4	−0.399551	0.0586792	−6.809	1.10×10^{-11} ***
CUSPRET_5	−0.600792	0.0592911	−10.13	6.67×10^{-24} ***
CUSPRET_6	−0.360412	0.0582064	−6.192	6.41×10^{-10} ***
CUSPRET_7	−0.307560	0.0577694	−5.324	1.06×10^{-7} ***
CUSPRET_9	0.0320326	0.0565113	0.5668	0.5709
CUSPRET_10	−0.0800320	0.0544286	−1.470	0.1415

Table 10. *Cont.*

	Coefficient	Std. Error	*t*-Ratio	*p*-Value
sq_sq_DMSPRET_1	−10.9763	0.978161	−11.22	7.02×10^{-29} ***
sq_sq_DMSPRET_2	−5.62131	0.964744	−5.827	6.00×10^{-9} ***
sq_sq_DMSPRET_3	−9.91128	0.993007	−9.981	3.02×10^{-23} ***
sq_sq_DMSPRET_4	−7.04655	0.991848	−7.104	1.38×10^{-12} ***
sq_sq_DMSPRET_5	−2.34791	1.00553	−2.335	0.0196 **
sq_sq_DMSPRET_6	−5.40256	1.00500	−5.376	7.97×10^{-8} ***
sq_sq_DMSPRET_7	−5.33317	1.00213	−5.322	1.07×10^{-7} ***
sq_sq_DMSPRET_8	0.835833	0.550288	1.519	0.1289
sq_sq_DMSPRET_9	1.69941	0.996423	1.706	0.0882 *
Mean dependent var	0.000112	S.D. dependent var	0.000269	
Sum squared resid	0.000149	S.E. of regression	0.000172	
R-squared	0.594745	Adjusted R-squared	0.592578	
F(27, 5051)	274.5461	*p*-value(F)	0.000000	
Log-likelihood	36,836.27	Akaike criterion	−73,616.54	
Schwarz criterion	−73,433.62	Hannan-Quinn	−73,552.48	
rho	0.223409	Durbin-Watson	1.553007	

Note: ***, ** and * denotes significance at 1%, 5% and 10%.

4. Conclusions

In this paper, we have explored the use of the Gegenbauer process or GARMA model to capture the behaviour of RV5 volatility of the S&P500 index, as reported by the OxfordMan database. The results suggest that the non-linear Gegenbauer model performs slightly better than the HAR model in capturing RV5. However, a simplified rule of thumb model based on the use of lagged, squared, cubed, and quartic, demeaned daily returns performed equally well. These results, for the S&P500 index, suggest that non-linear models perform better than linear ones in the capture of long memory properties of RV5 and that sophisticated models do not necessarily dominate rules of thumb.

Author Contributions: Conceptualization, D.E.A. and S.P.; methodology, D.E.A. and S.P.; software, D.E.A. and S.P.; validation, D.E.A. and S.P.; formal analysis, D.E.A.; investigation, D.E.A.; resources, S.P.; data curation, D.E.A.; writing—original draft preparation, D.E.A.; writing—review and editing, D.E.A. and S.P.; visualization, D.E.A.; project administration, S.P.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

Allen, David E., and Michael McAleer. 2020. Do we need stochastic volatility and Generalised Autoregressive Conditional Heteroscedasticity? Comparing squared end-of-day returns on FTSE. *Risks* 8: 12. [CrossRef]

Allen, David Edmund. 2020. Stochastic volatility and GARCH: Do squared end-of-day returns provide similar information? *Journal of Risk and Financial Management* 13: 202. [CrossRef]

Anděl, Jiří. 1986. Long memory time series models. *Kybernetika* 22: 105–23.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. 2001. The distribution of realized stock return volatility. *Journal of Financial Economics* 61: 43–76. [CrossRef]

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys. 2003. Modeling and forecasting realized volatility. *Econometrica* 71: 529–626. [CrossRef]

Barndorff-Nielsen, Ole E., and Neil Shephard. 2002. Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* 63: 253–80. [CrossRef]

Barndorff-Nielsen, Ole E., and Neil Shephard. 2003. Realized power variation and stochastic volatility models. *Bernoulli* 9: 243–65. [CrossRef]

Bollerslev, Tim. 1986. Generalized Autoregressive Conditional Heteroscedasticity. *Journal of Econometrics* 31: 307–327. [CrossRef]

Box, George E. P., and Gwilym M. Jenkins. 1970. *Times Series Analysis. Forecasting and Control*. San Francisco: Holden-Day.

Christensen, Bent J., and Nagpurnanand R. Prabhala. 1998. The relation between implied and realized volatility. *Journal of Financial Economics* 37: 125–50. [CrossRef]

Commandeur, Jacques J. F., and Siem Jan Koopman. 2007. *Introduction to State Space Time Series Analysis*. Oxford: Oxford University Press.

Corsi, Fulvio. 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7: 174–96. [CrossRef]

Dacorogna, Michel M., Ulrich A. Muller, Richard B. Olsen, and Olivier V. Pictet. 1998. Modelling short term volatility with GARCH and HARCH. In *Nonlinear Modelling of High Frequency Financial Time Series*. Edited by C. Dunis and B. Zhou. Chichester: Wiley.

Dickey, David A., and Wayne A. Fuller. 1979. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association* 74: 427–31.

Dissanayake, G. S., M. Shelton Peiris, and Tommaso Proietti. 2018. Fractionally Differenced Gegenbauer Processes with Long Memory: A Review. *Statistical Science* 33: 413–26. [CrossRef]

Dissanayake, Gnanadarsha Sanjaya. 2015. Advancement of Fractionally Differenced Gegenbauer Processes with Long Memory. Ph.D. thesis, School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. Least Angle Regression. *The Annals of Statistics* 32: 407–99. [CrossRef]

Engle, Robert F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987–1007. [CrossRef]

Engle, Robert F., and Clive W. J. Granger. 1987. Co-integration and error correction: Representation, estimation and testing. *Econometrica* 55: 251–76. [CrossRef]

Fan, Jianqing, and Runze Li. 2001. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* 96: 1348–60. [CrossRef]

Granger, Clive W. J., and Roselyne Joyeux. 1980. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1: 15–29. [CrossRef]

Gray, Henry L., Nien-Fan Zhang, and Wayne A. Woodward. 1989. On generalized fractional processes. *Journal of Time Series Analysis* 10: 233–57, Erratum in *Journal of Time Series Analysis* 15: 561–62. [CrossRef]

Guegan, Dominique. 2000. A new model: The k-factor GIGARCH process. *Journal of Signal Processing* 4: 265–71.

Hosking, J. R. M. 1981. Fractional differencing. *Biometrika* 68: 165–76. [CrossRef]

Hunt, Richard. 2022a. Garma: Fitting and Forecasting Gegenbauer ARMA Time Series Models, R Package Version 0.9.11. Available online: https://CRAN.R-project.org/package=garma (accessed on 3 October 2023).

Hunt, Richard. 2022b. Investigations into Seasonal ARMA Processes. Ph.D. thesis, School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia.

Hunt, Richard, Shelton Peiris, and Neville Weber. 2021. Estimation methods for stationary Gegenbauer processes. *Statistical Papers* 63: 1707–41. [CrossRef]

Peiris, M. Shelton, and Manabu Asai. 2016. Generalized fractional processes with long memory and time dependent volatility revisited. *Econometrics* 4: 37. [CrossRef]

Peiris, Shelton, David Allen, and Udara Peiris. 2005. Generalised autoregressive models with conditional heteroscedasticity: An application to financial time series modelling. In *The 2004 Workshop on Research Methods: Statistics and Finance, Proceedings of the 2004 Workshop on Research Methods: Statistics and Finance*. Edited by Eric J. Beh, Robert G. Clark and J. C. W Rayner, Wollongong: University of Wollongong, pp. 75–83. ISBN 1 74128 107 5.

Peiris, M. Shelton, and Aera Thavaneswaran. 2007. An introduction to volatility models with indices. *Applied Mathematics Letters* 20: 177–82. [CrossRef]

Perron, Pierre, and Wendong Shi. 2020. Temporal aggregation and Long Memory for asset price volatility. *Journal of Risk and Financial Management* 13: 181. [CrossRef]

Phillip, Andrew. 2018. On Gegenbauer long memory stochastic volatility models: A Bayesian Markov chain Monte Carlo approach with applications. Ph.D. thesis, School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia.

Poon, Ser-Huang, and Clive Granger. 2005. Practical issues in forecasting volatility. *Financial Analysts Journal* 61: 45–56. [CrossRef]

Poterba, James M., and Lawrence H. Summers. 1986. The persistence of volatility and stock market fluctuations. *American Economic Review* 76: 1124–41.

Ramsey, James Bernard. 1969. Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society, Series B* 31: 350–71. [CrossRef]

Schreiber, Sven. 2023. adalasso.gfn GRETL Code Function. Available online: https://gretl.sourceforge.net/cgi-bin/gretldata.cgi?opt=SHOW_FUNCS (accessed on 3 October 2023).

Schwert, G. William. 1989. Why does stock market volatility change over time? *Journal of Finance* 44: 1115–53. [CrossRef]

Shephard, Neil, and Kevin Sheppard. 2010. Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics* 25: 197–231. [CrossRef]

Sjoerup, Emil. 2019. HARModel: Heterogeneous Autoregressive Models. R Package Version 1.0. Available online: https://CRAN.R-project.org/package=HARModel (accessed on 3 October 2023).

Slutsky, Eugen. 1927. The summation of random causes as the source of cyclic processes. *Econometrica* 5: 105–46. [CrossRef]

Taylor, Stephen J., and Xinzhong Xu. 1997. The incremental volatility information in one million foreign exchange quotations. *Journal of Empirical Finance* 4: 317–40. [CrossRef]

Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B* 58: 267–88. [CrossRef]

Whittle, Peter. 1953. The analysis of multiple stationary time series. *Journal of the Royal Statistical Society. Series B (Methodological)* 15: 125. [CrossRef]

Yule, G. Udny. 1926. Why do we sometimes get nonsense correlations between timeseries? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society* 89: 1–63. [CrossRef]

Zou, Hui. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101: 1418–29. [CrossRef]