Forecasting Realized Volatility using a Long Memory Stochastic Volatility Model: Estimation, Prediction and Seasonal Adjustment

Rohit Deo*, Clifford Hurvich †and Yi Lu‡

December 6, 2004

Abstract

We study the modeling of large data sets of high frequency returns using a long memory stochastic volatility (LMSV) model. Issues pertaining to estimation and forecasting of large datasets using the LMSV model are studied in detail. Furthermore, a new method of de-seasonalizing the volatility in high frequency data is proposed, that allows for slowly varying seasonality. Using both simulated as well as real data, we compare the forecasting performance of the LMSV model for forecasting realized volatility to that of a linear long memory model fit to the log realized volatility. The performance of the new seasonal adjustment is also compared to a recently proposed procedure using real data.

KEY WORDS: Realized Volatility, Long Memory Stochastic Volatility Model, High Frequency Data, Seasonal Adjustment

JEL Classification: C13, C32, C53

*Corresponding author: rdeo@stern.nyu.edu, Stern School of Business, New York University: 44 W. 4'th Street, New York NY 10012, USA.

†churvich@stern.nyu.edu, Stern School of Business, New York University: 44 W. 4'th Street, New York NY 10012, USA.

‡ylu@stern.nyu.edu, Stern School of Business, New York University: 44 W. 4'th Street, New York NY 10012, USA.

1 Introduction

The availability of high frequency (intra day) data on returns of financial assets has sparked a great amount of research in modeling both these high frequency returns as well as the realized volatility (RV) computed from them. It is now well accepted in the literature (see, for example, Andersen, Bollerslev, Diebold and Labys, 2003, henceforth referred to as ABDL) that realized volatility is an important quantity in itself. An important question which arises is whether there is an advantage to be gained from predicting RV by modeling the high frequency returns or whether RV is a sufficient statistic by itself, in the sense that it contains almost all the relevant information useful for predicting its future values. Most of the literature (see for e.g., ABDL, Martens (2001), Martens, Chang and Taylor (2002)) related to this issue has modeled the high frequency returns using observation driven conditionally heteroscedastic models such as the GARCH model and its variants fit to seasonally adjusted high frequency returns. Furthermore, the seasonal adjustments considered in the literature have not allowed for time varying seasonality. However, to the best of our knowledge there seems to be no work in the literature that attempts to model the high frequency returns with latent variable conditionally heteroscedastic volatility models such as the stochastic volatility (SV) models, probably due to the fact that SV type models are not easily manipulable for estimation and forecasting purposes. We attempt to bridge this gap in the literature by studying in depth issues related to model estimation and prediction of high frequency returns using the Long Memory SV (LMSV) model (Breidt, Crato, and De Lima, 1998 and Harvey, 1998). Some newly proposed algorithms for solving large Toeplitz systems are exploited to provide efficient methods of forecasting from LMSV models with large data sets. Furthermore, we also propose a new way of estimating the seasonality in volatility that is present in high frequency returns, allowing for slowly varying seasonality. We then apply our methods to compare the RV forecasting performance of the LMSV model fit to high frequency data to that of linear long memory models fit to the log RV itself. The approach of modeling the log RV by a linear long memory process has been proposed by ABDL. The performance of the newly proposed seasonal adjustment procedure is also compared to a recently proposed seasonal adjustment for volatility in high frequency data.

The LMSV model for the returns r_t that we consider in this paper is given by

$$r_t = \sigma \exp\left(h_t/2\right) \varepsilon_t, \tag{1}$$

where $\varepsilon_t \sim$ IID(0,1), $\sigma > 0$ and $\{h_t\}$ is a stationary zero mean Gaussian long-memory process assumed to be independent of $\{\varepsilon_t\}$. Two popular choices for $\{\varepsilon_t\}$ are the standard normal distribution and a normalized t-distribution with ν degrees of freedom, with the normalization such that the variance of ε_t is unity. For the sake of simplicity, we will assume in this paper that $\{h_t\}$ follows an Autoregressive Fractionally Integrated

Moving Average ARFIMA(p, d, q) given by

$$\Phi(B)(1-B)^d h_t = \Theta(B)\eta_t,$$

where B denotes the backshift operator, $\eta_t \sim$ IID N$(0,\sigma_\eta^2)$, $0 < d < 0.5$, $\Phi(B)$ and $\Theta(B)$ are polynomials of order p and q respectively with all roots outside the unit circle. It should be emphasized that all of our procedures can be extended in simple ways to accommodate other long memory model specifications for h_t, such as a Fractional Exponential model. See Hurvich (2002) for details on the FEXP model.

The fact that the LMSV model is expressed in terms of latent variables makes it extremely amenable to the establishment of some of its theoretical properties, such as the behavior of the covariance function of powers of absolute returns. However, since the conditional variance of r_t is not available in analytic form, unlike in the GARCH family of models, it is not possible to write down the likelihood of the data on returns as a product of conditional likelihoods. This raises problems in the estimation of the model parameters as well as in computing forecasts of future squared returns. We thus turn our attention to addressing these twin problems in the following sections.

2 LMSV Model Estimation

There are several procedures available in econometrics for the estimation of parameters of SV models. Prime amongst these are the Generalized Method of Moments (GMM), Efficient Method of Moments (EMM) and Frequency Domain Quasi Maximum Likelihood (FDQML) estimation, which we discuss next.

2.1 GMM Estimation

For GMM estimation, one specifies a set of sample moments based on n observations denoted by $M_n = (M_{1n}, ..., M_{qn})$, where $M_{in} = (n-j)^{-1} \sum_{t=j+1}^{n} g_i(r_t, r_{t-j})$, j is the maximum lag being used, g_i is some smooth function and q, the number of selected moments, is at least as large as the dimension of the parameter vector θ to be estimated. The GMM estimator, $\hat{\theta}$, minimizes the distance $(M_n - M(\theta))' \Lambda^{-1} (M_n - M(\theta))$, where $M(\theta) = E(M_n)$ and Λ is some suitably chosen weight matrix. Under suitable regularity conditions, $\hat{\theta}$ is $\sqrt{n}$ consistent and asymptotically normal (Hansen 1982). These suitable conditions include the requirement that the vector of moments M_n be a $\sqrt{n}$ consistent estimator of $M(\theta)$. In the literature (see, for example, Andersen and Sorensen 1996), the moment conditions that have been used for SV models are obtained by using functions g_i of the form

$$g_i(r_t, r_{t-j}) = |r_t|^{a_i} |r_{t-j}|^{b_i} \tag{2}$$

for some integer valued non-negative a_i, b_i. Using the form of the LMSV model, one sees for such a choice of g_i that

$$M_{in} = (n-j)^{-1} \sum_{t=j+1}^{n} \sigma^{a_i+b_i} \exp\left(\left(a_i h_t + b_i h_{t-j}\right)/2\right) |\varepsilon_t|^{a_i} \left|\varepsilon_{t-j}\right|^{b_i}.$$

Now appealing to the fact that $a_i h_t + b_i h_{t-j}$ is a stationary Gaussian long memory series with memory parameter d which is independent of $\{\varepsilon_t\}$ and using the moment generating function of a Gaussian random variable, it is easy to show that $Var\left(M_{in}\right) \sim Cn^{2d-1}$ for some positive constant C, implying that the rate of convergence of M_{in} is of order $n^{1/2-d}$. Thus, GMM estimation in the LMSV case will result in estimators which are slower than $\sqrt{n}$ consistent since $0 < d < 0.5$. As a matter of fact, since most empirical estimates of d for the volatility of high frequency returns tend to be between 0.3 and 0.35 (Andersen and Bollerslev, 1997a), the GMM estimators will have a rate of convergence of the order of approximately $n^{0.2}$ or $n^{0.15}$.

2.2 EMM Estimation

In EMM estimation, the score of an approximating likelihood function with a tractable form is used to generate the moment conditions. In principle, one estimates the parameters of a mis-specified auxiliary model and then maps the estimates of the parameters of the auxiliary model to the parameters of the assumed data generating process. The better the auxiliary process is in approximating the assumed true data generating process, the more efficient the EMM estimators will be. However, in the context of linear long memory models, Chen and Deo (2003) have shown that the estimators of the parameters of a mis-specified model fit to a linear long memory process can have a rate of convergence that is slower than $\sqrt{n}$. Chen and Deo (2003) derive the condition under which $\sqrt{n}$ consistency is retained for the estimators of the parameters of the auxiliary model, but the condition is unverifiable in practice since it depends on the model parameter values which are unknown. Hence, their work implies that more work needs to be done on the problem of choosing an appropriate auxiliary model to ensure that EMM estimation provides $\sqrt{n}$ consistent estimators of the LMSV model.

2.3 FDQML Estimation

The QML estimation procedure exploits the fact that for the LMSV model, the transformed series $Z_t = \log(r_t^2)$ can be expressed as a sum of a Gaussian long memory signal plus a zero mean noise series, given by $Z_t = \log(r_t^2) = \mu + h_t + \xi_t$, where $\xi_t = \log(\epsilon_t^2) - E(\log(\epsilon_t^2))$ and $\mu = \log(\sigma^2) + E(\log(\epsilon_t^2))$. The QML estimators are found by treating the series Z_t as a Gaussian time series and maximizing the Gaussian log likelihood of Z_t. Even though Z_t is not a Gaussian process, Deo (1995) has shown that the resultant

estimators are $\sqrt{n}$ consistent and asymptotically normal. In practice, it is easier to minimize the frequency domain approximation to the time domain Gaussian negative log likelihood (Brockwell and Davis, 1991), called the Whittle approximation, and given by

$$\mathfrak{L}_n(\theta) = \sum_{j=1}^{[\frac{n-1}{2}]} \{\log(f_\theta(\omega_j)) + \frac{I_j}{f_\theta(\omega_j)}\} \tag{3}$$

where θ is a candidate parameter vector, $I_j = (2\pi n)^{-1} \left|\sum_{t=1}^{n} Z_t \exp(-i\omega_j t)\right|^2$ is the periodogram of Z_t at the j-th Fourier frequency $\omega_j = \frac{2\pi j}{n}$ and f_θ is the theoretical spectral density function of Z_t, given by

$$f_\theta(\omega_j) = \frac{\sigma_\eta^2}{2\pi} \frac{|\Theta(-i\omega_j)|^2}{|\Phi(-i\omega_j)|^2 |1-\exp(-i\omega_j)|^{2d}} + \frac{\sigma_\xi^2}{2\pi} \tag{4}$$

where σ_ξ^2 is the variance of ξ_t. We call the resultant estimator $\hat{\theta}$ the FDQML estimator. The periodogram I_j of Z_t can be evaluated at all the Fourier frequencies ω_j very quickly using the Fast Fourier Transform, even for large data sets, and hence the FDQML method is extremely easy from a computational viewpoint. When ε_t is assumed to be standard normal, σ_ξ^2 is known to be $\pi^2/2$ and hence need not be estimated. When ε_t is assumed to have the normalized $t-$distribution with unknown ν degrees of freedom, with the normalization such that the variance of ε_t is unity, then σ_ξ^2 needs to be estimated. There is a one-to-one relationship between σ_ξ^2 and ν given by $\sigma_\xi^2 = \psi'(\nu/2) + \pi^2/2$ where ψ' is the derivative of the digamma function. Thus, this relationship in conjunction with an estimate of σ_ξ^2 yields an estimate of the degrees of freedom, ν. Note that as $\nu \to \infty$, the t-distribution gets closer to the standard normal and $\psi'(\nu/2) \to 0$ and hence $\sigma_\xi^2 \to \pi^2/2$, as is to be expected.

Thus, the FDQML estimator is the only estimator out of the three popular estimators described above that is guaranteed to be $\sqrt{n}$ consistent when estimating the parameters of the LMSV model. Though the FDQML estimator is based on the Gaussian likelihood, one would expect it to be inefficient since the series Z_t is typically non-Gaussian. This prompts us to consider a modified version of the FDQML estimator which we describe next.

2.4 Enhanced FDQML

Since Z_t is non-Gaussian, better estimates might be obtained by applying the Whittle method to some transformation of high frequency absolute returns which is closer to Gaussianity than log squared returns. One such potential transformation is $|r_t|^c$ for a judicious choice of c which, for example, sets the theoretical skewness of $|r_t|^c$ to zero, thus making its distribution closer to the Gaussian than that of $\log r_t^2$. To make FDQML estimation feasible for the transformed series $|r_t|^c$, one has to first overcome two issues: (i) The choice of the power c and (ii) the calculation of the model spectral density of $|r_t|^c$. To solve the first issue,

we suggest a two stage procedure. In the first stage, FDQML estimation is employed on $\log r_t^2$ to get initial parameter estimates of the LMSV model. Using these parameter estimates, one can compute the theoretical skewness of $|r_t|^c$ for every value of c as a nonlinear function of the model parameters as follows. First we note that the skewness $E[|r_t|^c - \mu_c]^3$, where $\mu_c = E[|r_t|^c]$, can be expressed up to a constant multiple as

$$E[|r_t|^c - \mu_c]^3 = E|r_t|^{3c} - {\mu_c}^3 - 3\mu_c E|r_t|^{2c} + 3\mu_c^2 E|r_t|^c. \tag{5}$$

By using the moment generating function of a Gaussian random variable and setting $\sigma_h^2 = Var(h_t)$, we get for any positive integer k,

$$E|r_t|^{kc} = E|\varepsilon_t|^{kc} \exp({\sigma_h}^2 k^2 c^2/8)\sigma^{kc}. \tag{6}$$

From Harvey (1998), when ε_t follows a standardized $t-$distribution with unit variance, we have for $c < \upsilon$,

$$E|\epsilon_t|^c = (\upsilon - 2)^{\frac{c}{2}} \frac{\Gamma(\frac{c+1}{2})\Gamma(\frac{\upsilon - c}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\upsilon}{2})},$$

whereas when ε_t follows a standard normal distribution we have

$$E|\epsilon_t|^c = \frac{2^{c/2}\Gamma\left((c+1)/2\right)}{\Gamma\left(1/2\right)}.$$

These expressions in conjunction with Equation (6) and Equation (5) give us an explicit expression for the skewness of $|r_t|^c$. A simple bisection method can be used to find the root of the skewness as a function of c. We also considered finding the c that minimized an equally weighted sum of squares of skewness and excess kurtosis. The resulting value of c was extremely close to that obtained by minimizing skewness alone. This is exemplified in Fig 1, where we plot the sample skewness and sample excess kurtosis respectively as a a function of c for seasonally adjusted r_t^c.

Once we find an appropriate c value, the Whittle estimation of the model with transformed data $|r_t|^c$ next requires the model spectral density of $|r_t|^c$. Unfortunately, there does not exist a simple formula for the model spectral density of $|r_t|^c$, unlike that given in (4) for the density of $\log r_t^2$. However, it is known that for any stationary time series with covariances $c_j(\theta)$ at lag j, the spectral density $f_\theta(\omega)$ at frequency ω is also given by

$$f_\theta(\omega) = \frac{1}{2\pi} \sum_{|j|<\infty} c_j(\theta) \exp(ij\omega). \tag{7}$$

The model covariance function $c_j(\theta)$ of $|r_t|^c$ can be obtained by noting that for $a, b > 0$,

$$E|r_t|^a = \sigma^a E|\epsilon_t|^a \exp(a^2 \frac{\sigma_h^2}{8})$$

and

$$E(|r_t|^a |r_{t-j}|^b) = \sigma^{a+b} E|\epsilon_t|^a E|\epsilon_t|^b \exp((a^2+b^2)\frac{\sigma_h^2}{8} + ab\gamma_h(j)/4),$$

where $\sigma_h^2 = Var(h_t)$ and $\gamma_h(j) = Cov(h_t, h_{t-j})$. Though there is no explicit formula for the covariance function $\gamma_h(j)$ of an ARFIMA process, Bertelli and Caporin (2002) provide a simple algorithm for computing it, which we use. Combining this algorithm with the expressions for $E|\epsilon_t|^c$ given above, we can easily compute the $c_j(\theta)$. Finally, it should be noted that in computing the Whittle likelihood, the infinite sum in expression (7) needs to be truncated at some value to make the computation of the spectral density feasible. However, a naive truncation at some point can result in a spectral density which is negative, which is naturally undesirable. Hence, we choose to weight the covariances $c_j(\theta)$ with a properly chosen smooth function w_j before truncating, which ensures a positive spectral density given by

$$f_{\theta,M}(\omega) = \frac{1}{2\pi} \sum_{|j|<M} w_j c_j(\theta) \exp(ir\omega),$$

where M is the truncation point. In this paper we used a popular weight sequence called the Bartlett window, which is guaranteed to yield positive densities and is given by $w_j = 1 - |j|/M$. In our study we chose the truncation to be $M = n$, where n is the number of observations. It can be shown that as $M \to \infty$, $f_{\theta,M}(\omega) \to f_\theta(\omega)$ for every $\omega > 0$.

Though there are no theoretical results on the asymptotic distribution of the Enhanced FDQML estimator that we have proposed, we have chosen to include it in our study. In Section 6 below, we report on a simulation study in which we compare the FDQML and the Enhanced FDQML estimators based on $\log r_t^2$ and $|r_t|^c$ respectively.

3 Forecasting with the LMSV model

We are interested in predicting r_{n+L}^2 from the available returns $r_n, r_{n-1}, \ldots, r_1$ for $L \geq 1$. The best such prediction of r_{n+L}^2 would be $E[r_{n+L}^2 | r_n, r_{n-1}, \ldots, r_1]$. Unfortunately, this conditional expectation is not available in an analytical form for the SV model, unlike in the observation driven models. One solution is to use the best linear prediction of r_{n+L}^2 based on $|r_n|^c, |r_{n-1}|^c, \ldots, |r_1|^c$ for some value of $c \leq 2$. As noted by Ding, Granger & Engle (1993), the magnitude of the correlations of $|r_t|^c$ tends to be stronger for $c < 2$ than for $c = 2$. Hence, it might be worthwhile to predict squared returns from $|r_t|^c$ with $c < 2$. This idea has similar features as the PARCH model of Ding, Granger & Engle (1993). In order to compute this best linear predictor, we need coefficients $A_{j,L}$ for $j = 0, 1, \ldots, n-1$ which minimize

$$E\{r_{n+L}^2 - \mu_{r,2} - \sum_{j=0}^{n-1} A_{j,L}[|r_{n-j}|^c - \mu_{r,c}]\}^2. \tag{8}$$

where $\mu_{r,2} = E(r_t^2)$ and $\mu_{r,c} = E(|r_t|^c)$.

Once the coefficients $A_{j,L}$ have been obtained, the best L-step ahead linear predictor of r_{n+L}^2 is given by

$$\hat{r}_{n,L,c}^2 = \mu_{r,2} + \sum_{j=0}^{n-1} A_{j,L}[|r_{n-j}|^c - \mu_{r,c}] \tag{9}$$

It is well known that the coefficients $A_{j,L}$ which minimize (8) are the solution of the set of linear equations

$$\mathbf{\Sigma_c A}_L = \gamma_{2,c,L}, \tag{10}$$

where $\mathbf{A}_L = (A_{0,L}, \ldots, A_{n-1,L})'$, $\mathbf{\Sigma_c} = Cov(|r_n|^c, \ldots, |r_1|^c)$ and

$$\gamma_{2,c,L} = [Cov(r_{n+L}^2, |r_n|^c), \ldots, Cov(r_{n+L}^2, |r_1|^c)]'.$$

Once the parameters of the LMSV model are known, the entries of $\mathbf{\Sigma_c}$ and $\gamma_{2,c,L}$ can be found as described in Section 2. It thus remains to solve the set of linear equations in (10), which involve a Toeplitz matrix $\mathbf{\Sigma_c}$. A numerically efficient algorithm to solve such a system was provided by Levinson (1946). However, Levinson's algorithm needs $O(n^2)$ operations and with a large high-frequency data set such as ours, is still not efficient enough. Chen, Hurvich and Lu (2004) have proposed an algorithm for solving the system of equations (10). Their results suggest that the number of operations used by this algorithm is $O(n \log^3(n))$. For $n = 30000$, the factor of improvement in efficiency afforded by this new algorithm compared to Levinson's algorithm is close to 30 at least.

Though we forecast the squared returns using the best linear forecast, this forecast will not be optimal since squared returns are not Gaussian. However, as discussed in Section 2, we can find the power c such that $|r_t|^c$ have zero skewness and are closer to normal. This suggests that it might be preferable to get the best linear forecast of $|r_{n+L}|^c$ based on $|r_n|^c, ..., |r_1|^c$ and then convert this forecast to one of r_{n+L}^2. The best linear forecast, say $|\hat{r}_{n+L}|^c$, of $|r_{n+L}|^c$ based on $|r_n|^c, ..., |r_1|^c$ is found in a manner similar to that described above, by solving a system of linear equations. Then, conditional on $r_n, ..., r_1$, the distribution of $|r_{n+L}|^c$ is treated as Gaussian with mean $|\hat{r}_{n+L}|^c$ and variance $\sigma_{L,c}^2 = E\left(|r_{n+L}|^c - |\hat{r}_{n+L}|^c\right)^2$. Now,

$$E[r_{n+L}^2 | r_n, r_{n-1}, \ldots, r_1] = E[(|r_{n+L}|^c)^{2/c} | r_n, r_{n-1}, \ldots, r_1].$$

Thus, the forecast of r_{n+L}^2 based on the power transformation, denoted by $\hat{r}_{n+L,P}^2$, is now computed as

$$\hat{r}_{n+L,P}^2 = E\left(|Y|^{2/c}\right), \tag{11}$$

where Y is a normal random variable with mean $|\hat{r}_{n+L}|^c$ and variance $\sigma_{L,c}^2$. Though there is no explicit formula for this expectation, it can be computed very easily using numerical integration.

In our simulation study as well as the empirical study in later sections, we study the performance of both the prediction procedures described above.

4 Seasonal Adjustment of Volatility

It is an empirically observed fact that high frequency returns exhibit seasonality in volatility and any attempt to model high frequency returns with an eye towards prediction must deal with estimating this seasonal component. There are several ways in which the seasonal component in volatility has been estimated in the literature (See, for e.g., Martens et al, 2002). However, a common theme in all of the proposed methods of seasonal adjustment is that the seasonal pattern is assumed to remain constant with some periodicity, be it a day or a week or a month. Based on our high frequency data, we argue below that the seasonality in volatility may be actually slowly varying over time and we propose a new way of estimating it which accounts for this slow variation. The basic model that we will assume to motivate our seasonal adjustment procedure is

$$R_t = \exp\left(S_t/2\right) r_t, \tag{12}$$

where R_t is the high frequency return which has been demeaned using the sample mean, S_t is the seasonal component and r_t is the de-seasonalized high frequency return. We find it convenient for two reasons to work with the transformed series $X_t = \log R_t^2$ when estimating the seasonal component. The first reason is that log squared returns are less prone to outliers and hence, from a data analytic point of view, it seems a sensible transformation to consider. The second reason is that the log squared de-seasonalized return $\log r_t^2$ has certain desirable properties which can be exploited to construct statistical tests of significance to test for the presence of seasonality. It should also be noted here that Martens et al (2002) concluded that it is preferable to estimate seasonal adjustments based on log squared returns. The fact that we use the demeaned high frequency data generally assures us that there are no zero values after the demeaning, thus avoiding any problems with taking logarithms.

In addition to working with the log squared returns, we also find it convenient to study the seasonal pattern in volatility in the frequency domain. Towards this end, we construct the periodogram of the log squared returns at Fourier frequencies ω_j,

$$I_X\left(\omega_j\right) = \left(2\pi n\right)^{-1} \left| \sum_{t=1}^{n} X_t \exp\left(-it\omega_j\right) \right|^2 .$$

In Fig 2(a), we plot the log periodogram of a data set of log squared high frequency returns versus the Fourier frequencies. The data set which we have considered here, explained in detail in Section 5 below, consists of 51000 observations of half hourly returns on the S&P500 index, with 12 returns per day.

If the seasonality were exactly periodic with period 12, we would expect a peak in the periodogram at every Fourier frequency with an index j that is an integer multiple of $n/12$, n being the sample size of the data set. However, closer examination of such frequencies in Fig 2(a) shows that their neighboring Fourier frequencies

also have large periodogram values. This behavior can be seen in more detail in Fig 2(b), which shows the log periodogram in the neighborhood of the Fourier frequency $\omega_{n/12} = \frac{2\pi}{n}\frac{n}{12} = \pi/6$. We argue next that this phenomenon, where the log periodogram shows peaks not just at Fourier frequencies with indices that are integer multiples of $n/12$ but also at their neighboring frequencies, is typical of a seasonal component that is of the form $S_t = P_t F_t$, where P_t is an exactly periodic function with period 12 and F_t is a smooth function which is slowly varying around 1. For any series u_t, $t = 1, 2, ..., n$, let $G_{u,j}$ denote its Discrete Fourier Transform (DFT) at Fourier frequency ω_j, given by

$$G_{u,j} = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} u_t \exp\left(-it\omega_j\right).$$

If the seasonal component S_t does indeed have the form $S_t = P_t F_t$, then it is known (Bloomfield, 1976, pg 86) that

$$G_{S,j} = \sum_{k=0}^{n-1} A_{P,k} A_{F,(j-k) \bmod n}. \tag{13}$$

Now the DFT of P_t would have peaks at exactly the Fourier frequencies with indices that are integer multiples of $n/12$, since P_t is exactly periodic with period 12. On the other hand, since F_t is a smooth slowly varying function, its DFT would be peaked at zero frequency and then taper down gradually at other Fourier frequencies (Bloomfield, 1976, pg 57-60). Thus, from the convolution (13) we see that a peak in the DFT of P_t at a Fourier frequency which is an integer multiple of $n/12$ will *leak* into neighboring Fourier frequencies in the DFT of S_t, due to the widening of the DFT of F_t around zero frequency, resulting in the observed behavior of the periodogram.

Having argued that the pattern in seasonality is consistent with a slowly varying function, we next provide a procedure to remove the seasonality. We model the seasonality S_t as a linear combination of sines and cosines evaluated at the Fourier frequencies that show seasonal peaks. More specifically, we write

$$S_t = \sum_{p=1}^{k} a_p \cos \omega_{j_p} t + \sum_{p=1}^{k} b_p \sin \omega_{j_p} t, \tag{14}$$

where $B = \left\{\omega_{j_p}\right\}_{p=1}^{k}$ is the collection of all Fourier frequencies with indices that are integer multiples of $n/12$ and their neighboring Fourier frequencies that exhibit large values. A statistical test of significance which can be used to decide which Fourier frequencies to use in this set B is provided later in this section. The coefficients a_p and b_p are estimated by

$$\hat{a}_p = 2n^{-1} \sum_{t=1}^{n} X_t \cos \omega_{j_p} t$$

and

$$\hat{b}_p = 2n^{-1} \sum_{t=1}^{n} X_t \sin \omega_{j_p} t\ .$$

The computation of these coefficients can be done in a numerically efficient manner, using the Fast Fourier Transform, since $\hat{a}_p$ and $\hat{b}_p$ are the just multiples of real and imaginary parts respectively of the DFT of X_t computed at the Fourier frequency ω_{j_p}. The seasonal component estimated in this fashion would be identical to that obtained by running the regression

$$X_t = \sum_{p=1}^{k} a_p \cos \omega_{j_p} t + \sum_{p=1}^{k} b_p \sin \omega_{j_p} t + e_t, \tag{15}$$

since the regressors in (15) are all orthogonal to each other. The fitted values in this regression yield the estimated seasonal component S_t, while the residuals in this regression yield the de-seasonalized log squared returns. The seasonal specification in (14) may be easily extended to include dummy variables to account for any day-specific effect if need be. If dummy variables are included, the seasonal adjustment would have to be estimated using a regression similar to the one in (15), with dummy variables included, since there is no simple method of estimating the seasonal effect coefficients in the frequency domain in that case. It should however be noted that the regression (15) is computationally very efficient, even in the case when dummy variables are included, since the regressors $\cos \omega_{j_p} t$ and $\sin \omega_{j_p} t$ are orthogonal to each other. Due to this fact, the normal equations that one would have to solve would involve a matrix which is almost diagonal, except for the first several rows and columns, which would correspond to the dummy variables. Solving such a system of normal equations is a computationally simple task.

In Fig 3, we plot the log periodogram of the de-seasonalized log squared returns. On comparing this plot with Fig 2(a), we note that the de-seasonalized series now no longer shows any seasonal peaks in its periodgram as is to be expected. One important benefit of our seasonal adjustment worth pointing out at this stage is that the peridogram of the de-seasonalized series remains untouched at all other frequencies, as is evidenced by the two plots. In particular, the frequencies closest to the origin are left undisturbed, implying that the long memory dynamics of the series have not been disturbed in the de-seasonalizing procedure.

To get an idea of how the estimated seasonal component looks like for our data, we plotted in Fig 4 the estimated seasonal component for a two week period, at two different points in time in the sample. It is observed that though the overall pattern of the estimated seasonality remains the same in the two plots, the exact details vary, as is to be expected.

When one views our proposed seasonal adjustment in the light of the regression (15), it becomes easier to interpret it as a generalization of the Flexible Fourier Form (FFF) method of seasonal adjustment, which has been generally used in the literature (see, for e.g., Andersen and Bollerslev, 1997b, Martens et al (2002)). The simplest form of the FFF essentially runs a regression of the form (15) but only at the Fourier frequencies whose indices are integer multiples of $n/12$, viz. $\{\omega_j\}$, $j = n/12,\ n/6, ..., sn/12$, where s is some small arbitrary integer, necessarily less than $L/2$, where L, in general, is the number of observations per day (In

our current empirical exercise, $L = 12$). In addition, the regression also includes a low order polynomial, such as a quadratic, which is periodic with period L. When $s = L/2$, this seasonal adjustment is identical to subtracting the sample means of the data for each of the L periods. In such a case, the low order polynomial in the regression becomes redundant since the L cosine and sine terms form a basis for the L dimensional space. However, in the literature, FFF has generally been applied with s quite small relative to the value of L and the polynomial chosen generally being quadratic. The rationale behind this approach is to estimate the seasonal component in a parsimonious and smooth way, rather than by estimating one mean level for each period within the day. However, FFF does not easily provide a method of identifying the value of s and the degree of the polynomial, a shortcoming which is overcome by our approach. In the next paragraph, we provide an objective method of deciding which Fourier frequencies should be included.

The de-seasonalization procedure we describe above requires us to specify the collection of Fourier frequencies B which are to be included in the regression (14). Though some of the frequencies to be included could possibly be easily detected by visual inspection of the plot of the log periodogram against frequency, there might be some frequencies which leave room for doubt. Hence, a statistical test of significance is essential to help make a decision on the inclusion of a frequency for deseasonalizing and the following Theorem provides just such a test. We omit the proof as it is quite simple, exploiting the fact that h_t is a Gaussian series independent of ξ_t and that the DFT's of each of the two series, at distinct Fourier frequencies bounded away from the origin, are asymptotically Gaussian and independent, with variances that are proportional to their respective spectral densities evaluated at the corresponding Fourier frequencies. See Moulines and Soulier, 1999.

Theorem 1 *For any fixed integer k, let $\{\omega_{j_p}\}_{p=1}^{k}$ be a set of Fourier frequencies such that $\liminf_{n\to\infty} \min_p \omega_{j_p} > 0$. Let $f_X\left(\omega_{j_p}, \hat{\theta}\right)$ be the estimated spectral density of $\log r_t^2$, where $\hat{\theta}$ is a vector of estimated parameters which is $\sqrt{n}$ consistent. Assuming that $a_s = b_s = 0$ for $s = j_1, ..., j_k$, we have*

$$\sum_{p=1}^{k} \frac{I_X\left(\omega_{j_p}\right)}{f_X\left(\omega_{j_p}, \hat{\theta}\right)} \xrightarrow{D} \frac{1}{2} V,$$

where V is a χ^2_{2k} random variable.

Theorem 1 allows us to test the joint null hypothesis of the lack of seasonal peaks at any specified set of Fourier frequencies. The test needs some $\sqrt{n}$ consistent estimate $\hat{\theta}$ of the parameters of the LMSV model. Our suggestion is that in order to obtain this estimate $\hat{\theta}$ for the seasonal adjustment stage, the following modified version of the Whittle likelihood in (3) be used,

$$\mathfrak{L}_{n,M}(\theta) = \sum_{j \in M} \{\log(f_\theta(\omega_j)) + \frac{I_j}{f_\theta(\omega_j)}\}. \tag{16}$$

where $M = \{1, 2, ..., [(n-1)/2]\} \cap D^c$ and D is some initial collection of frequencies which one suspects of having seasonal peaks. For example, D could consist of the each of the Fourier frequencies with indices j that are integer multiples of $n/12$ and 15 Fourier frequencies to the left and right of each of them. The exclusion of a fixed number of frequencies in D will not affect the asymptotic properties of the initial estimate $\hat{\theta}$ and yet at the same time will ensure that it is not affected by any seasonality in the volatility.

In closing this Section, we note that all of our procedures and suggestions generalize to high frequency data sets with an arbitrary number of observations per day, say L, by merely replacing the ratio $n/12$ by n/L.

Having described the estimation, forecasting and seasonal adjustment procedures for LMSV models in detail, we now turn to applying them to real data in the next Section.

5 Empirical Analysis

In this section, we use real data to compare various procedures with respect to their ability to forecast RV based on high frequency returns. The data set that we consider was obtained from Tick Data, Inc. and consists of half hourly returns on the S&P500 index. There were 12 returns per day, computed from 10:00am to 3:30pm, with the 12th return being the overnight return computed using the closing price at 3:30 pm on the given day and the 10:00 am closing price on the following day. The data spanned a period from 2/1/1983 to 6/30/2000. From the high frequency data, we also constructed a corresponding series of RV, where the RV for day t was defined as the sum of squares of the 12 half hourly returns for that day. Andersen, Bollerslev, Diebold & Labys (2002) proposed RV as an error-free measure of volatility and theoretically speaking, the finer we sample during the day, the more accurate the measure becomes. However, so-called market microstructure problems such as bid-ask bounce and asynchronous trading in the equity market create bias in the estimation of parameters such as autocorrelation if excessively high frequency financial return data is used. Thus, there is a trade-off between the crudeness of the realized volatility measure and the microstructure bias. We choose 30-minute returns as the balance point for these issues.

In Fig 5(a), we plot the autocorrelations for the log squared sample mean adjusted high frequency returns up to lag 3120, which corresponds to about 1 year. The extremely slow decay of the sample autocorrelations points to the existence of long memory, while the periodic peaks at lags which are integer multiples of 12, which is the number of observations per day, points towards the existence of seasonality. The existence of long memory is also observed in the strong linear relation between the log periodogram for $\log(|r_t|^2)$ versus the log frequency(for $j = 1$ to 1000) in Fig 5(b).

We compare the forecasting performance of our seasonal adjustment in combination with an LMSV model fit to high frequency data to that of a linear long memory model fit to log realized volatility itself. The LMSV model that we chose was such that the log volatility process h_t followed an ARFIMA$(1, d, 0)$ while the errors ε_t were assumed to follow a normalized $t-$distribution with ν degrees of freedom, where the normaliztion was such that ε_t had unit variance. In addition, we also include the forecasting performance of a GARCH(1,1) and a component GARCH (Engle and Lee, 1999) model fit to seasonally adjusted high frequency data. The component GARCH has two components in the volatility function, one transient and the other persistent, and is essentially a GARCH(2,2) model with certain restrictions on the parameters. We include the GARCH(1,1) model and the component GARCH model as competitors, to evaluate how well they can account for the long memory in the volatility in spite of being, theoretically, short memory models for volatility. The linear long memory model that we fit to the log realized volatility was an ARFIMA$(1, d, 0)$, as was done in ABDL.

We compare the out-of-sample forecasting performance of the competing models using two different estimation window sizes: 15000 (corresponding to about 5 years of data) and 30000 (about 10 years of data). We also compare 3 different forecasting horizons: 1 day, 1 week (5 days), and 4 weeks.

We now describe in detail the design of the empirical experiments for the various models. Consider an estimation window of 15000 and forecast horizon of 1 week as an example. Note that we have a total of 51000 S&P 30-minute returns in our dataset.

5.1 Modeling the high frequency returns

i) First we demeaned the data using the sample mean of the returns, denoted by $\hat{\mu}$. Henceforth in this section, when we refer to the returns, we imply the sample mean adjusted returns. One advantage of demeaning the data in this fashion is that it ensures there are no zero values, thus avoiding problems when making the logarithmic transformation. Then, we estimated the seasonal component S_t using the log squared returns. In order to do this, we first estimated the parameters of the LMSV model using the modified Whittle log likelihood $\mathfrak{L}_{n,M}$ in (16) and where D consisted of each of the Fourier frequencies with indices j that are integer multiples of $n/12$ and 30 Fourier frequencies to the left and right of each of them. Once the initial estimates were obtained, we tested the periodogram at each Fourier frequency in the set D, using the test given in Theorem 1 to detect which peaks were statistically significant. The level of significance used at each frequency was set at 5%. Thus, the frequencies to be included in the set B for the seasonal adjustment using the regression (14) were determined and the seasonal component S_t estimated.

ii) Using the model definition in (12), the high frequency de-seasonalized returns were obtained by the calculation $r_t = \exp(-S_t/2)\, R_t$, where R_t was the original high frequency return. Then the first 15000 seasonally adjusted 30-minute returns $\{r_t\}_{t=1}^{15000}$ were used to estimate an LMSV, GARCH(1,1) or component GARCH model.

iii) Next, we constructed a one-step (30-minute) ahead forecast $\hat{r}_{15001}^2$ of r_{15001}^2 using the observations $\{r_t\}_{t=1}^{15000}$. The exact way in which this forecast was computed depended on the model. For the GARCH(1,1) and component GARCH, the calculation of the forecast is built into the volatility equation in the model. For the LMSV model, we used two different methods of forecasting, The first method used the best linear predictor given in (9). The second method used the predictor based on the power transformation given in (11). The power c in this procedure was obtained by setting the theoretical skewness of $|r_t|^c$ to zero, as described in Section 2 above.

iv) Then we excluded the 1st observation r_1 and included the square root of the forecast we just generated, to get a "new" data set $\{r_t\}_{t=2}^{15000} \cup \hat{r}_{15001}^{1/2}$ of 15000 observations. Then step (iii) was repeated to get another one-step ahead forecast $\hat{r}_{15002}^2$ of r_{15002}^2. We repeated this process 60 times, obtaining a set of forecasts $\{\hat{r}_t^2\}_{t=15001}^{15060}$ of seasonally adjusted squared returns $\{r_t^2\}_{t=15001}^{15060}$ for each 30-minute period in the future week.

v) We then re-seasonalized the forecasts $\{\hat{r}_t^2\}_{t=15001}^{15060}$ by extrapolating the most recent in-sample estimated seasonal component to generate forecasts of future squared high frequency returns with seasonality in them, denoted by $\hat{R}_t^2$, as

$$\hat{R}_t^2 = \exp(S_{t-60})\hat{r}_t^2 \qquad t = 15001, ..., 15060. \tag{17}$$

Since we had been working with mean adjusted returns all along, it was essential to undo this mean adjustment to get a forecast of the "unprocessed" high frequency returns, denoted by $\hat{R}_{t,\mu}^2$. This was achieved by computing

$$\hat{R}_{t,\mu}^2 = \hat{R}_t^2 + \hat{\mu}^2 \qquad t = 15001, ..., 15060.$$

Finally, we aggregated the 60 forecasts $\{\hat{R}_{t,\mu}^2\}_{t=15001}^{15060}$ to get one forecast

$$\hat{V}_1 = \sum_{t=15001}^{15060} \hat{R}_{t,\mu}^2$$

of the realized volatility, RV_1, of the future week.

vi) Next, we excluded the data from the very first week in our first estimation window of 30-minute returns, and included the data for the week we had just forecast, yielding a new estimation window $\{r_t\}_{t=61}^{15060}$. Then steps (i)-(v) above were repeated. This cycle was repeated 150 times and we got 150 out-of-sample forecasts $\{\hat{V}_i\}_{i=1}^{150}$ for the 150 realized volatility values $\{RV_i\}_{i=1}^{150}$ computed from the S&P data directly.

It should be noted here that the maximum number of weekly forecasts we could have computed in this manner for an estimation window size of 15000 is actually 600. However, the maximum number of monthly forecasts we can compute is only 150. Thus, we chose to compute only 150 forecasts for all three horizons, daily, weekly and monthly, when the estimation window size was 15000. This allows us to see how the various forecasting measures change across the three forecasting horizons based on the same number of forecasts. When the estimation window size was 30000, we however computed the maximum number of forecasts permissible across each horizon, viz. 90 monthly (4-week) forecasts, 360 weekly forecasts and 1800 daily forecasts.

In the next subsection, we describe how we modeled the realized volatility directly. Once again, for the sake of exposition, we describe the procedure for the case where we were predicting the realized volatility one week ahead.

5.2 Modeling the realized volatility

For modeling the realized volatility, we first calculated all daily realized volatility values, each such value being the sum of squares of the 30 minute returns on that day. We then fit an ARFIMA(1,d,0) model to the first 1250 values of daily log RV, in keeping with the approach of ABDL. Using the fitted ARFIMA model, forecasts were generated of the log RV for the next 5 days, corresponding to one week. These forecasts were exponentiated to obtain five forecasts of the daily RV and then these five forecasts were aggregated to obtain a forecast of the future weekly RV. We then rotated the data forward by one week, excluding the daily realized volatility for the first week and including the daily RV for the week we had just forecast, and then used this new set of 1250 daily realized volatility values to estimate the ARFIMA(1,d,0) model and forecast 1-week ahead again. We repeated this process 600 times and obtained 600 forecasts, each forecast being for non-overlapping weeks.

In the next sub-section, we compare the performance of the different forecasting procedures.

5.3 Comparison of the forecasts based on different models

There is no universal agreement about the best measure for evaluating the forecasting performance for volatility models. Hence, we compared the different forecasting procedures based on five measures that have commonly been used in the literature. These measures were: (i) The R^2 from the regression of log realized volatility on the forecast of log realized volatility (ii) The R^2 from the regression of square root realized volatility on the forecast of square root realized volatility (iii) The mean squared error (MSE) (iv) The mean

absolute deviation (MAD) (v) Mean absolute percentage deviation (MAPD).

The log and square root transformations were used as variance stabilizing transformations in the regressions since the RV is very heteroscedastic. We did not compute any $t-$statistics on the slope and intercept coefficients in these regressions since there is no reason to expect the forecasts to be unbiased after a square root or log transformation.

The LMSV model was always estimated using the Enhanced FDQML method of Sub-section 2.4 above. For the LMSV model, the forecasts based on the c power transformation given in (11) are denoted by LMSV-C. The c was chosen to make the theoretical skewness of the distribution of $|r_t|^c$ equal to 0. In our estimations, the mean c value was 0.2562311. We denote the method of using ${r_t}^2$ to forecast ${r_t}^2$ as given in (9) by LMSV-Square. Forecasts based on fitting an ARFIMA(1,d,0) model to log realized volatility are denoted by ABDL.

The forecasting results with 15000 (5 years of data) as estimation window size are summarized in Table 1. The results are based on 150 forecasts for each combination of forecasting model and horizon. The number in bold corresponds to the best performance for a particular measure for a particular forecasting horizon.

Main observations for Table 1 are:

- Overall, LMSV models forecast better if R^2 or MSE measures are used. The ABDL procedure generally forecasts better if MAD or MAPD measures are used, with the ABDL procedure having consistently the lowest MAPD across all forecasting horizons.
- Better modeling of long memory does improve forecasts significantly as we can see, with the short memory GARCH(1,1) consistently doing worst, and the long memory ABDL and LMSV procedures consistently doing best.
- The ABDL model delivers a very impressive performance given that it just fits a simple ARFIMA(1,d,0) model and does not model the high-frequency seasonality directly.
- The component GARCH model also does well, particularly since it is not a true long memory model. Its performance is extremely close to and even better than that of the ABDL model especially for weekly and daily horizons. This seems to suggest that for a horizon up to a week, the component GARCH model is able to capture the long memory dynamics fairly well.
- R^2 as an increasing function of forecast horizon for all procedures.
- LMSV C-Fore performed uniformly better than LMSV-Square across all horizons with respect to the

MSE, MAD and the MAPD, but this uniform superiority over LMSV-Square was not retained for the R^2 measures.

- Though we do not report the results here, we also attempted to fit component GARCH and GARCH(1,1) models directly to the non-seasonally adjusted high frequency data. This approach performed uniformly worse than the other procedures considered here.

The forecasting results with 30000 (10 years of data) as estimation window size are summarized [1] in Table 2. The results are based on 90 monthly (4-week) forecasts, 360 weekly forecasts and 1800 daily forecasts.

Major observations for Table 2 are:

- Overall, LMSV models forecast better if R^2 or MSE measures are used. The ABDL procedure forecasts better if MAD or MAPD measures are used, with the ABDL procedure having consistently the lowest MAPD across all forecasting horizons. This is consistent with the results from Table 1 when the window size was 15000.
- In view of the results from simulation study reported below, it seems the increased forecastiblity from $n = 15000$ to $n = 30000$ is not because of the increase of n, but probably due to the fact that when n is larger, the seasonality is stronger and seasonal adjustment has much larger effect on forecasting performance.

It is also of interest to see whether the seasonal adjustment that we propose does better than a naive adjustment which estimates the seasonal component by using the sample means for each period. As a comparison, we therefore also generated forecasts of RV using this naive seasonal adjustment in conjunction with our LMSV model to forecast RV for the window size of 15000. We present the forecasting results using this approach Table 3.

On comparing the results in Table 3 with those in Table 1, we see that in almost all the cases, forecasts using the new seasonal adjustment perform better than the seasonal means adjustment method. Forecasts using LMSV C-Fore always do better using the new adjustment than when using the seasonal means. However, there are some measures and horizons where forecasts using LMSV-Square perform better when in conjunction with the seasonal means adjustment than with the new adjustment. In all, our procedure does better than the seasonal means adjustment in 25 out of the 30 cases.

[1] Component GARCH and GARCH results are not reported due to the memory constraints to run the program in Eviews. This happens since we need to use Splus to do the seasonal adjustment and stored the seasonal adjusted data in Eviews for further analysis

We also compare the performance of our seasonal adjustment to that of the FFF method mentioned earlier. We followed Martens, Chang and Taylor(2002) specification of FFF with second order polynomial and $s = 4$. The forecasting results obtained by FFF are presented in Table 4. On comparing these results with Table 1, we see that the forecasts using the new seasonal adjustment do better than those based on the FFF method in 21 out of 30 situations. Forecasts using the LMSV-Square method in conjunction with the new seasonal adjustment outperform the FFF method uniformly based on the R-squared measures.

In closing, we also present superimposed time series plots of actual monthly realized volatility and its forecasts based on the different models in Fig 8 to Fig 12 at the end of paper.

Since seasonality is an inherent part of high frequency data, it is not possible to separate the forecastiblity of RV due to the modeling of seasonality from that due to the modeling of long memory. To overcome this problem, we carried out a simulation study in which no seasonality is present, so that we could directly compare the forecasting performance of two long memory models, namely, the LMSV model and the ABDL method without any confounding factors. The simulation study also permitted us to compare the performance of FDQML and enhanced FDQML estimation procedure. We report our results from the simulation study in the next section.

6 Simulation Study with the LMSV model

The model we used to simulate high frequency returns was given by

$$r_t = \sigma \exp(h_t/2)\epsilon_t \tag{18}$$

where $\sigma > 0$ and $\{h_t\}$ is a stationary zero mean Gaussian process assumed to be independent of $\{\epsilon_t\}$. The $\{\epsilon_t\}$ series was assumed to be i.i.d. and distributed as $\sqrt{\frac{v-2}{v}}t_v$, where t_v has a t-distribution with v degrees of freedom. The log volatility series $\{h_t\}$ was assumed to follow an ARFIMA $(1, d, 0)$ given by

$$(1 - \alpha B)(1 - B)^d h_t = \eta_t,$$

where η_t is i.i.d. Gaussian with mean zero and with variance σ_η^2. The model parameters, chosen to correspond to typical values estimated from seasonally adjusted S&P 500 high frequency return data, were given by $\alpha = 0.35$, $d = 0.370549$, $\sigma_\eta^2 = 0.27$, $\sigma_\epsilon^2 = 4.94847$, $\nu = 147.3237$ and $\sigma = 0.00144019$.

To be consistent with the way we dealt with the real data, we only considered two choices of 15000 and 30000 as the estimation window size. A window of 15000 corresponds to approximately 5 years of 30-minute returns data while a window of 30000 corresponds to 10 years of 30-minute returns data. We considered

three choices of lead times: a lead-time of 12 that corresponds to 1-day realized volatility, a lead-time of 60 that corresponds to 1-week realized volatility, and a lead-time of 240 that corresponds to 1-month(4-week) realized volatility.

6.1 Comparison of Estimation Procedures

We simulated samples of size 30000 from the LMSV model in (18) and estimated the parameters of the model using the FDQML and Enhanced FDQML procedures described in Section 2 above. This exercise was replicated 300 times. Boxplots of the estimates of the four parameters across the 300 replications are shown in Fig 6 and Fig 7. It is immediately apparent from the plots that Enhanced FDQML provides a vast improvement in the estimation of the AR(1) parameter α as well as the noise variance σ_η^2 in the log volatility process h_t.

The mean squared error (MSE) and the absolute bias for the various estimates are shown in Table 5 and Table 6 respectively. Both the MSE and the absolute bias for the estimates of α is obviously lower using the Enhanced FDQML method as also observed in Fig 6(a). The MSE for the estimates of d is lower for enhanced FDQML even though the absolute bias is higher. That indicates the much lower variance of the estimate for d using Enhanced FDQML as also observed in Fig 6(b). On the other hand, the MSE for the estimates of σ_η^2 is a little higher for Enhanced FDQML even though the absolute bias is much lower as observed in Fig 7(a). For estimates of σ_ϵ^2, the MSE of Enhanced FDQML is higher due to larger absolute bias as can be seen from Table 6 and Fig 7(b).

6.2 Forecasting Results with Simulated Data

In our simulation study, the three forecasting procedures that we employed were identical to those described in Subsection 5.3 above, viz. LMSV-Square, LMSV-C and ABDL. The first two procedures, as detailed in Subsection 5.3, model the high frequency returns to generate predictions of future high frequency squared returns, which are then aggregated to obtain predictions of future RV, whereas the ABDL procedure fits an ARFIMA$(1, d, 0)$ to the log RV to generate predictions of future RV.

In the comparison of the different forecasting procedures, we chose to use both the estimation methods, the FDQML as well as the Enhanced FDQML. Furthermore, we also computed forecasts using the true parameter values, which allows us to quantify the extent to which parameter estimation degrades the quality of the forecasts across the various procedures.

The estimation window was was allowed to be 15000 and 30000 and for each value of the estimation window, the forecasts were compared across three different forecasting horizons: daily realized volatility (12 data periods), weekly (5-day) realized volatility (60 data periods) and monthly(4-week) realized volatility (240 data periods). The measures of comparison of the forecasts were identical to the five used in Subsection 5.3 above. The results, all based on 300 forecasts and 30000 as estimation window size, are given in Table 7. The numbers in bold correspond to the best performance for a given measure and a given horizon.

For monthly volatility forecasting, our observations based on Table 7 are:

- As is to be expected, knowing the true parameter values helps the forecast, but not by very much for the LMSV-Square method.
- The Enhanced FDQML estimation method does not help much in terms of forecasting. Specifically, it does not make much difference for the LMSV Square-Fore method, and it actually makes the forecast worse for the LMSV-C forecasting method.

For weekly volatility forecasting, our observations based on Table 7 are:

- Obviously, knowing the true parameter gives the best forecast performance. Somewhat surprisingly, the LMSV-Square method performs better than the LMSV-C method.
- Again, we see the insensitivity of the LMSV-Square method to parameter estimation. That is both good and bad: good because it implies that we should not worry too much about parameter uncertainty, bad since it could be the case that the forecasting method is not powerful enough to make good use of more accurate parameter estimates
- Now the Enhanced FDQML estimation method does help the LMSV-C forecasting method. The reason being that the performance of this method is very sensitive to the estimation of c.
- As the horizon becomes shorter, the forecastibility decreases.

For daily volatility forecasting, our observations based on Table 7 are very similar to those at the longer horizons.

Our observations for the ABDL method based on Table 7 are:

- The forecasting performance of the ABDL method is consistently best at all horizons based on the MAD measure. It is best at both the weekly and daily horizon based on all measures except the

MSE. However, at the monthly horizon, ABDL gets worse except when using the MAD measure. One possible explanation for this is that accurate modeling of the long memory is not that important when forecasting at the "short" daily and weekly horizons but becomes more crucial at the longer monthly horizons.

The forecasting results with 15000 as estimation window size are summarized in Table 8. The bold numbers correspond to the best performance for a given measure at a given horizon.

For monthly volatility forecasting, the use of the true parameter values generally yields the best forecasting performance. The advantage of using the Enhanced FDQML estimation method is at best only marginal. Specifically, it does not make much difference for the LMSV Square-Fore method, and it actually makes the forecast worse for the LMSV C-Fore forecasting method. Similar observations hold for the weekly and daily horizon.

Noticeably, larger estimation window size (30000 versus 150000) tends to have better forecasting performance only for short horizons (1 day). There is no significant difference in the forecasting performance for 4-week or 1-week horizon.

The ABDL method does fairly well at the daily and weekly horizon, having the best performance for almost all five measures. At the daily horizon, the ABDL method is beaten only by the infeasible LMSV forecasts which use the true parameter values, but beats all the feasible LMSV forecasts. At the monthly horizon, the ABDL method is outperformed by the LMSV forecasts using the R^2 measures and the MAD. Once again, the superior behavior of ABDL at the daily/weekly horizon is probably attributable to its ability to capture the long memory fairly well when forecasting over a short horizon.

7 Conclusion and Future Directions for Research

Predicting RV based on a simple ARFIMA(1,d,0) model applied to the log realized volatility series is a very good competitor to the method which predicts RV using a long memory stochastic volatility model applied to high frequency return data while accounting for the strong slowly varying intra-day seasonality in volatility. It is worth noting that the Component GARCH model, as an approximation of a long memory model, gives a surprisingly good performance even though still inferior in most cases to the LMSV model. The simple GARCH(1,1) gets outperformed by all the other models, which is not surprising, considering that it is not capable of mimicking long memory in a reasonable way. We also feel that our study raises some questions that are worth pursuing for future research. For example, in our study, we compared the different

forecasting procedures using five different measures. However, it is currently unclear which of these measures are better at comparing competitive forecasting procedures in the context of RV. Research which attempts to resolve this issue would be helpful. Also, the data set which we used had 12 observations per day, which were used in computing daily RV. In studies of exchange rates, (see, for example, ABDL), the RV is based on 48 observations per day. It would be interesting to compare the ABDL approach to the LMSV model fit to seasonally adjusted data for such data sets, where the level of aggregation is increased. Furthermore, it would also be interesting to see what effect seasonal adjustments involving dummy variables to account for announcements may have on the forecasts.

8 Bibliography

Andersen, T. and T. Bollerslev (1997a), Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long Run in High Frequency Returns. Journal of Finance, 52, 975-1005

Andersen, T. and T. Bollerslev (1997b), Intraday Periodicity and Volatility Persistence in Financial Markets, Journal of Empirical Finance, 4, 115-58

Andersen, T. and T. Bollerslev (1998a), Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. International Economic Review, 39(4), 885-905

Andersen, T. and T. Bollerslev (1998b), DM-dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements and Longer Run Dependencies, Journal of Finance, 53, 219-65

Andersen, T., Bollerslev, T., Diebold, F.X. and H. Ebens (2001), The Distribution of Realized Stock Return Volatility, Journal of Financial Economics, 61, 43-76

Andersen, T., Bollerslev, T., Diebold, F.X. and P. Labys (2003), Modeling and Forecasting Realized Volatility, Econometrica, forthcoming

Areal N.M.P and Taylor S. J. (2002), The Realized Volatility of FTSE-100 Futures Prices, Journal of Futures Markets, 22, 627-648

Bertelli and Caporin (2002), A Note on Calculating Autocovariances of Long-Memory Processes, Journal of Time Series Analysis, 23(5), 503-508

Bloomfield, P. (1976), Fourier Analysis of Time Series: An Introduction, New York: Wiley

Bollerslev, T. (1986), Generalized Autoregressive Conditional Heteroskedasticity, Journal of Econometrics, 31, 307-327

Breidt, F.J., Crato, N. and P. de Lima (1998), On the Detection and Estimation of Long Memory in Stochastic Volatility. Journal of Econometrics, 83, 325-348

Brockwell, P. and R. Davis (1996), Time Series: Theory and Methods, Second Edition, Springer

Chen, W., Hurvich, C.M. and Y. Lu (2004), On the Correlation Matrix of the Discrete Fourier Transform and the Fast Solution of Large Toeplitz Systems For Long-Memory Time Series

Chen, W and R. Deo (2003) Estimation of Mis-specified Long Memory Models with Potential Implications, Preprint.

Deo, R. (1995), Tests for Unit Roots in Multivariate Autoregressive Processes, Unpublished Ph.D. Thesis, Iowa State University

Deo, R. and C. Hurvich (2003), Estimation of Long Memory in Volatility. Theory and Applications of Long-Range Dependence, edited by Paul Doukhan, George Oppenheim, Murad S. Taqqu. Part A, pp 313-324, Birkhauser

Ding, Z., Granger, C.W.J., and R.F. Engle (1993), A Long Memory Property of Stock Market Returns and A New Model. Journal of Empirical Finance, 1, 83-106

Engle, R.F. (1982), Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, Econometrica, 50(4), 987-1007

Engle, R.F. and Gary G.J. Lee. (1999), A Permanent and Transitory Component Model of Stock Return Volatility, in Cointegration, Causality, and Forecasting: A Fetschrift in Honour of Clive W. J. Granger. Robert F. Engle and Halbert White, eds. Oxford: Oxford University Press, 475-97

Engle, R.F. and A.J. Patton (2001), What Good is a Volatility Model? Quantitative Finance, 1(2), 237-245

Gray, H. L., Zhang, N.F. and W. Woodward (1989), On Generalized Fractional Processes, Journal of Time Series Analysis, 10(3), 1-26

Hansen, P.R. and A. Lunde(2002), A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? Working paper, Department of Economics, Brown University

Harvey, Andrew (1998), Long Memory in Stochastic Volatility, in Forecasting Volatility in Financial Markets. J Knight and S Satchell (eds), 307-320. Oxford: Butterworth-Heineman

Hosking, J.R.M. (1981), Fractional Differencing, Biometrika, 68, 165-176

Hurvich, C.M. (2002), Multi-step Forecasting of Long Memory Series using Fractional Exponential Models, International Journal of Forecasting, 18, 167-179

Levinson, N. (1946), The Wiener RMS (root mean square) Error Criterion in Filter Design and Prediction, J.Math.Phys., 25, 261-178

Martens, M., Chang, Y.C. and S. Taylor (2002), A Comparison of Seasonal Adjustment Methods when Forecasting Intraday Volatility, Journal of Financial Research, 25, 283-299

Moulines, E. and P. Soulier (1999), Broadband Log-periodogram Regression of Time Series with Long-range Dependence, Annals of Statistics 27, 1415-1439

Pagan, A.R. and G.W. Schwert (1990), Alternative Models for Conditional Volatility, Journal of Econometrics, 45, 267-290

Sowell, F.B. (1992), Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models, Journal of Econometrics, 53, 165-188

Taylor, S. (1986), Modeling Financial Time Series, New York: John Wiley Sons

Whittle, P. (1962), Gaussian Estimation in Stationary Time Series, Bull. Int. Statist. Inst., 39, 105-129

Table 1: Forecasting Results for S&P 500 Realized Volatility With n=15000

Model	Horizon	R^2 logVol	R^2 SqrtVol	MSE	MAD	MAPD
LMSV-Square	4 Weeks	**0.479**	0.430	1.042e-06	6.137e-04	0.570
LMSV-C		0.457	**0.497**	**6.947e-07**	**5.145e-04**	0.498
ABDL		0.433	0.390	1.236e-06	5.917e-04	**0.387**
Component GARCH		0.430	0.408	9.969e-07	5.835e-04	0.563
GARCH(1,1)		0.165	0.155	1.291e-06	7.136e-04	0.758
LMSV-Square	1 Week	**0.287**	0.260	7.926e-08	1.935e-04	0.793
LMSV-C		0.182	0.207	6.596e-08	1.684e-04	0.698
ABDL		0.249	0.242	7.565e-08	**1.504e-04**	**0.398**
Component GARCH		0.274	**0.274**	**6.567e-08**	1.673e-04	0.674
GARCH(1,1)		0.171	0.155	8.850e-08	1.903e-04	0.786
LMSV-Square	1 Day	**0.082**	**0.081**	6.901e-09	6.747e-05	1.596
LMSV-C		0.038	0.047	**5.618e-09**	4.763e-05	0.896
ABDL		0.063	0.067	5.641e-09	**4.667e-05**	**0.787**
Component GARCH		0.074	0.079	5.636e-09	5.388e-05	1.118
GARCH(1,1)		0.025	0.020	6.966e-09	5.779e-05	1.151

Table 2: Forecasting Results for S&P 500 Realized Volatility with n=30000

Model	Horizon	R^2 logVol	R^2 SqrtVol	MSE	MAD	MAPD
LMSV-Square	4 Weeks	0.694	**0.673**	**7.218e-07**	**4.864e-04**	0.415
LMSV-C		0.611	0.575	9.278e-07	6.301e-04	0.635
ABDL		**0.710**	0.658	1.368e-06	6.166e-04	**0.336**
LMSV-Square	1 Week	**0.631**	**0.604**	**7.787e-08**	1.489e-04	0.558
LMSV-C		0.527	0.497	1.024e-07	1.910e-04	0.863
ABDL		**0.631**	0.593	1.195e-07	**1.620e-04**	**0.377**
LMSV-Square	1 Day	**0.422**	**0.412**	**8.158e-09**	4.656e-05	1.334
LMSV-C		0.353	0.338	9.219e-09	5.306e-05	1.820
ABDL		0.417	0.397	9.680e-09	**4.283e-05**	**0.811**

Table 3: Forecasting Results with Seasonal Means Adjustment for n=15000

Model	Horizon	R^2 logVol	R^2 SqrtVol	MSE	MAD	MAPD
LMSV-Square	4-Week	0.479	0.425	9.912e-07	5.903e-04	0.539
LMSV-C		0.281	0.223	1.357e-06	6.787e-04	0.620
LMSV-Square	1-Week	0.249	0.219	8.173e-08	1.955e-04	0.782
LMSV-C		0.152	0.176	7.221e-08	1.892e-04	0.860
LMSV-Square	1-Day	0.078	0.070	7.665e-09	6.953e-05	1.591
LMSV-C		0.010	0.013	5.708e-09	5.485e-05	1.260

Table 4: Forecasting Results with Flexible Fourier Form (FFF) Seasonal Adjustment for n=15000

Model	Horizon	R^2 logVol	R^2 SqrtVol	MSE	MAD	MAPD
LMSV-Square	4-Week	0.479	0.424	9.945e-07	5.912e-04	0.540
LMSV-C		0.411	0.370	1.043e-06	6.051e-04	0.595
LMSV-Square	1-Week	0.247	0.217	8.204e-08	1.956e-04	0.783
LMSV-C		0.229	0.224	7.178e-08	1.835e-04	0.790
LMSV-Square	1-Day	0.078	0.071	7.584e-09	6.933e-05	1.589
LMSV-C		0.052	0.057	5.695e-09	5.628e-05	1.283

Table 5: Mean Squared Error of Estimation Results with 300 Simulations

MSE	$\hat{\alpha}$	$\hat{d}$	$\hat{\sigma_{\eta}^2}$	$\hat{\sigma_{\epsilon}^2}$
Enhanced FDQML	0.0140	0.000416	0.0077	0.0057
FDQML	0.0174	0.000557	0.0076	0.0053

Table 6: Absolute Bias of Estimation Results with 300 Simulations

Absolute Bias	$\hat{\alpha}$	$\hat{d}$	$\hat{\sigma_\eta^2}$	$\hat{\sigma_\epsilon^2}$
Enhanced FDQML	0.001885	0.003672	0.01356	0.0400
FDQML	0.050566	0.003150	0.03127	0.0318

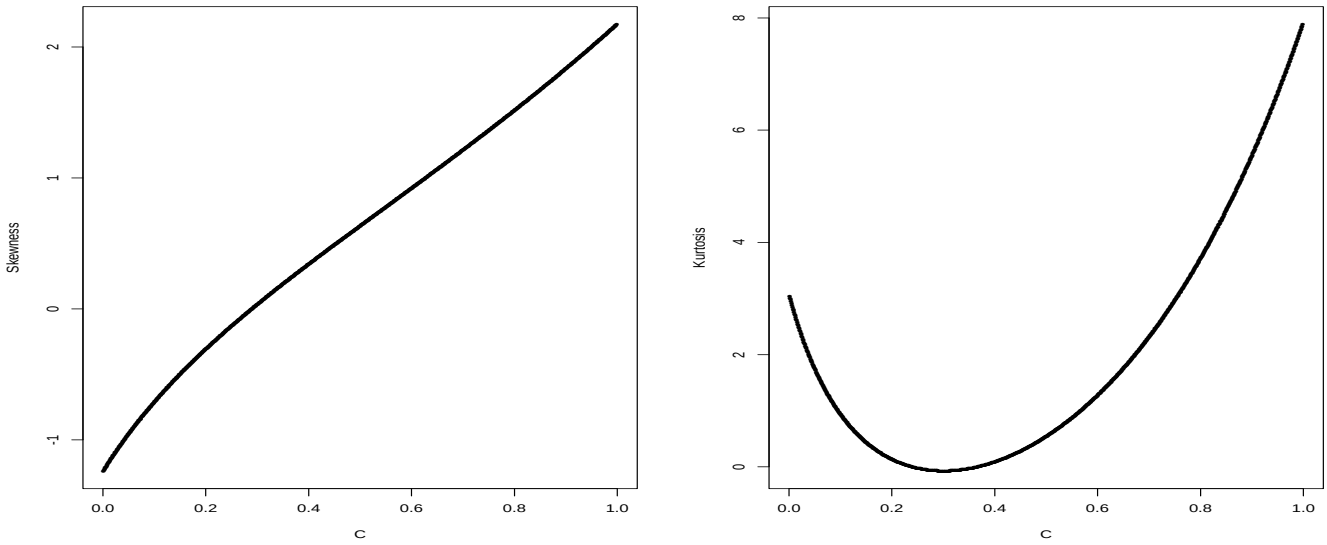
Table 7: Forecasting results for Data Simulated from LMSV with n=30000

Estimate(Forecast) Methods	Horizon	R^2 logVol	R^2 SqrtVol	MSE	MAD	MAPD
True (LMSV-Square)	4 Weeks	0.494	0.493	**9.797e-08**	2.286e-04	0.361
True (LMSV-C)		**0.524**	**0.549**	1.941e-07	2.714e-04	0.303
FDQML (LMSV-Square)		0.494	0.493	9.799e-08	**2.283e-04**	0.360
FDQML (LMSV-C)		0.451	0.443	1.763e-07	2.623e-04	0.317
Enhanced FDQML (LMSV-Square)		0.494	0.493	**9.797e-08**	2.286e-04	0.361
Enhanced FDQML (LMSV-C)		0.366	0.373	1.908e-07	2.744e-04	0.353
ABDL		0.508	0.493	1.480e-07	2.561e-04	**0.298**
True (LMSV-Square)	1 Week	0.344	0.330	**1.036e-08**	7.107e-05	0.519
True (LMSV-C)		0.337	0.329	1.167e-08	6.884e-05	0.396
FDQML (LMSV-Square)		0.345	0.330	1.037e-08	7.106e-05	0.518
FDQML (LMSV-C)		0.278	0.272	1.153e-08	7.183e-05	0.454
Enhanced FDQML (LMSV-Square)		0.344	0.329	1.039e-08	7.119e-05	0.519
Enhanced FDQML (LMSV-C)		0.291	0.289	1.138e-08	7.182e-05	0.475
ABDL		**0.373**	**0.361**	1.129e-08	**6.822e-05**	**0.382**
True (LMSV-Square)	1 Day	0.317	0.286	9.365e-10	1.950e-05	0.898
True (LMSV-C)		0.285	0.265	1.036e-09	1.847e-05	0.682
FDQML (LMSV-Square)		0.319	0.288	**9.288e-10**	1.942e-05	0.898
FDQML (LMSV-C)		0.210	0.185	1.095e-09	1.956e-05	0.786
Enhanced FDQML (LMSV-Square)		0.317	0.286	9.315e-10	1.943e-05	0.901
Enhanced FDQML (LMSV-C)		0.239	0.213	1.051e-09	1.958e-05	0.794
ABDL		**0.330**	**0.295**	9.793e-10	**1.802e-05**	**0.642**

Table 8: Forecasting results for Data Simulated from LMSV with n=15000

Estimate(Forecast) Methods	Horizon	R^2 logVol	R^2 SqrtVol	MSE	MAD	MAPD
True (LMSV-Square)	4 Weeks	**0.503**	**0.478**	1.447e-07	**2.454e-04**	0.358
True (LMSV-C)		0.418	0.397	1.515e-07	2.621e-04	0.301
FDQML (LMSV-Square)		0.496	0.471	1.459e-07	2.469e-04	0.360
FDQML (LMSV-C)		0.376	0.363	1.444e-07	2.683e-04	0.347
Enhanced FDQML (LMSV-Square)		0.497	0.471	1.459e-07	2.468e-04	0.361
Enhanced FDQML (LMSV-C)		0.383	0.358	1.802e-07	2.786e-04	0.373
ABDL		0.474	0.450	**1.293e-07**	2.514e-04	**0.297**
True (LMSV-Square)	1 Week	0.358	0.329	1.346e-08	7.803e-05	0.499
True (LMSV-C)		0.330	0.298	1.654e-08	7.973e-05	0.391
FDQML (LMSV-Square)		0.360	0.331	**1.342e-08**	7.778e-05	0.498
FDQML (LMSV-C)		0.257	0.221	1.660e-08	8.266e-05	0.473
Enhanced FDQML (LMSV-Square)		0.358	0.330	**1.342e-08**	7.791e-05	0.501
Enhanced FDQML (LMSV-C)		0.239	0.213	1.693e-08	8.718e-05	0.520
ABDL		**0.377**	**0.346**	1.567e-08	**7.695e-05**	**0.374**
True (LMSV-Square)	1 Day	**0.261**	0.245	**1.912e-09**	2.077e-05	0.896
True (LMSV-C)		0.238	0.208	2.094e-09	1.977e-05	**0.660**
FDQML (LMSV-Square)		0.258	0.242	1.917e-09	2.080e-05	0.899
FDQML (LMSV-C)		0.205	0.191	1.996e-09	2.070e-05	0.796
Enhanced FDQML (LMSV-Square)		0.259	0.244	1.913e-09	2.079e-05	0.902
Enhanced FDQML (LMSV-C)		0.195	0.174	2.030e-09	2.171e-05	0.857
ABDL		0.259	**0.255**	1.960e-09	**1.949e-05**	0.661

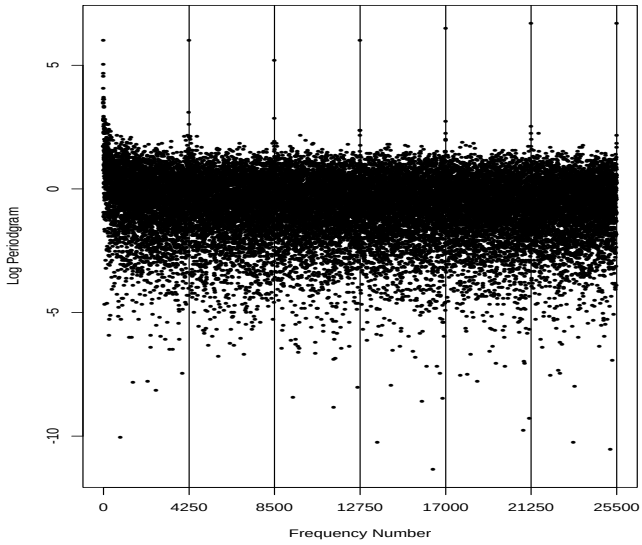
Figure 1: Sample Moments of De-seasonalized r_t^c versus c



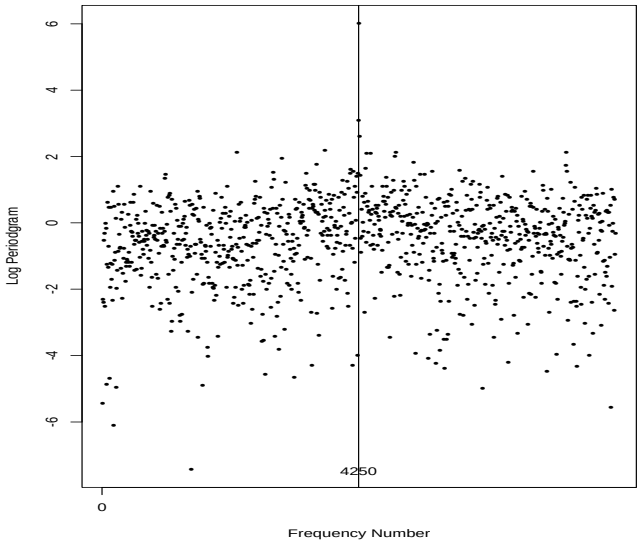
(a) Sample Skewness

(b) Sample Excess Kurtosis

Figure 2: Log Periodogram for High Frequency Log Squared Returns



(a) Log Periodogram for $\log |r_t|^2$



(b) Log Periodogram around the first Peak

Figure 3: Logged Periodogram for De-seasonalized $\log |r_t|^2$

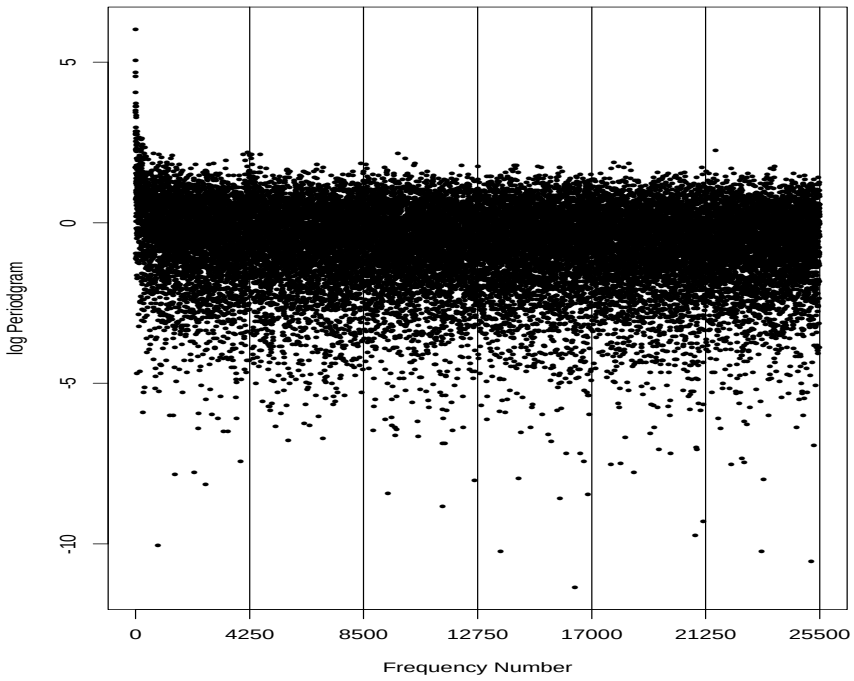


Figure 4: Seasonal Component in Volatility during Different Time Periods

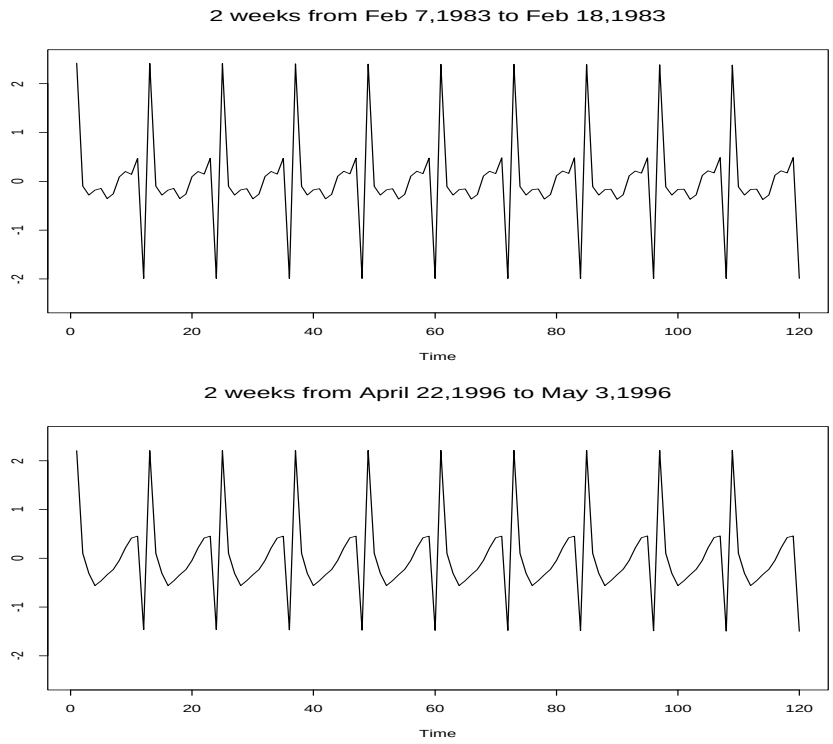
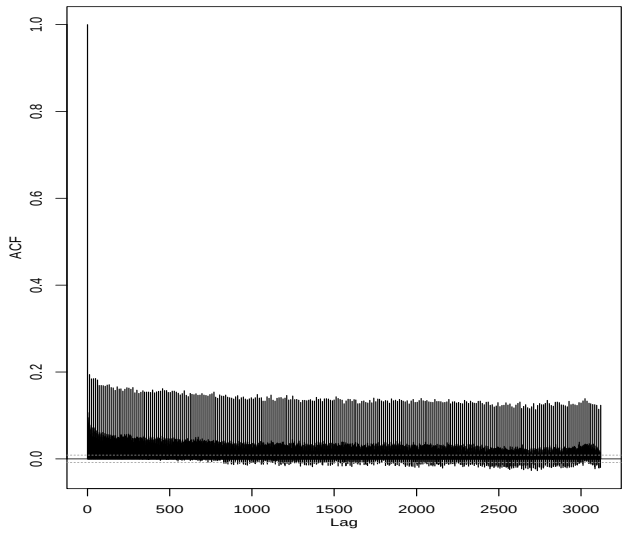
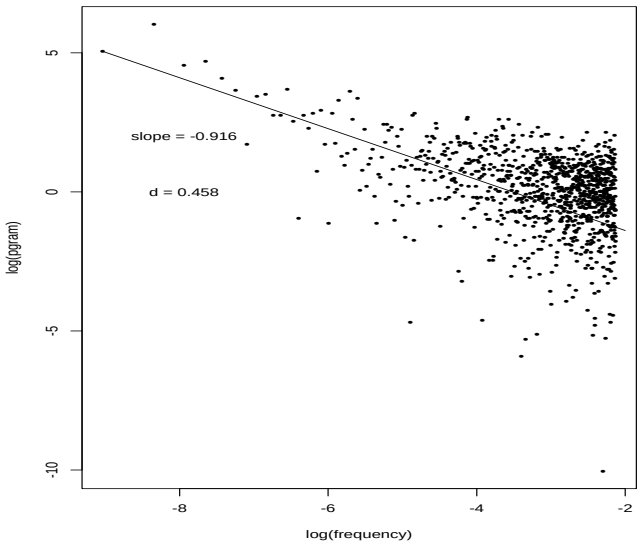


Figure 5: Autocorrelations and log(Periodogram) of $\log(|r_t|^2)$

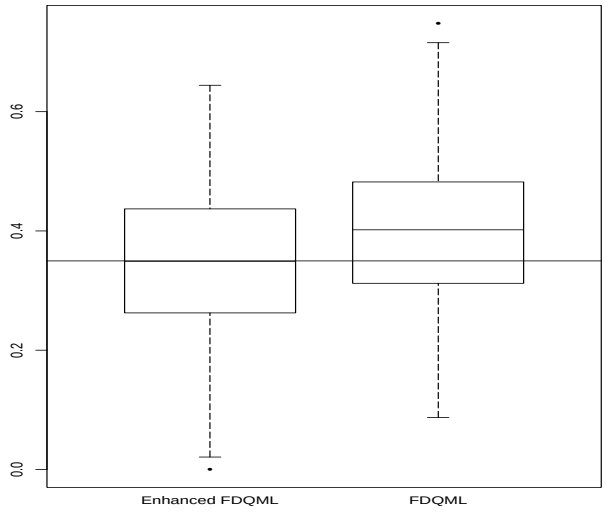


(a) Autocorrelation plot of $\log(|r_t|^2)$

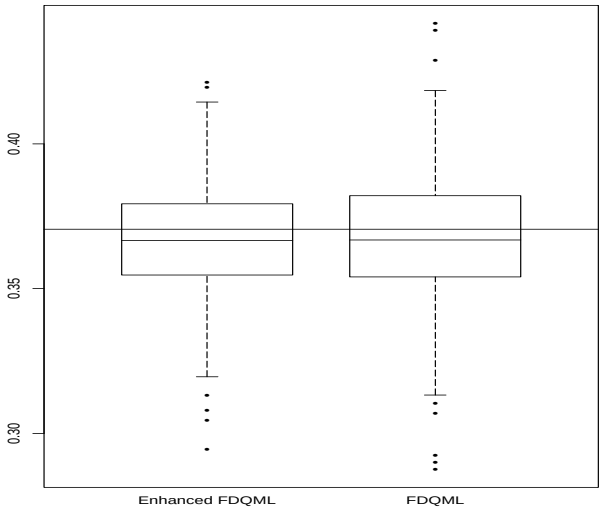


(b) Log Periodogram versus Log Frequency

Figure 6: Boxplots of Estimated α and d from 300 Simulations

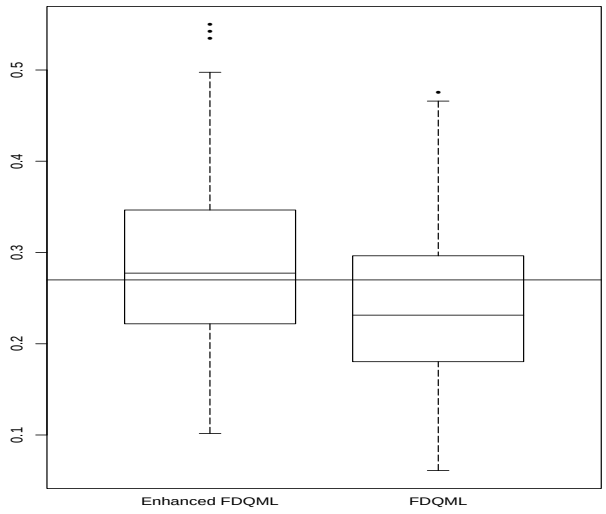


(a) $\hat{\alpha}$ ($\alpha = 0.35$)

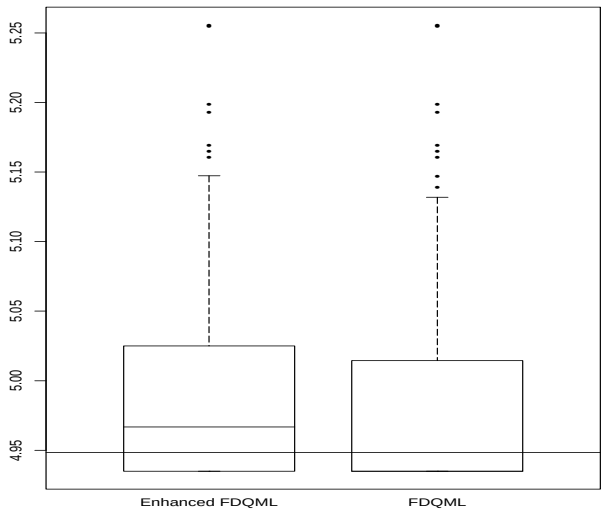


(b) $\hat{d}$ ($d = 0.370549$)

Figure 7: Boxplots of Estimated σ_η^2 and σ_ϵ^2 from 300 Simulations

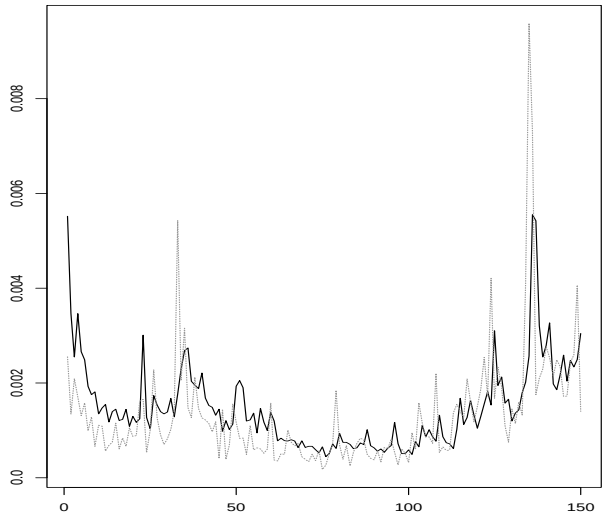


(c) $\hat{\sigma_\eta^2}$ ($\sigma_\eta^2 = 0.27$)

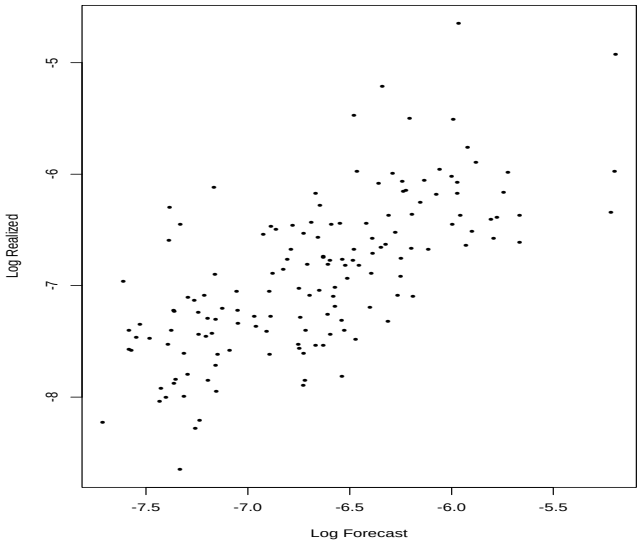


(d) $\hat{\sigma_\epsilon^2}$ ($\sigma_\epsilon^2 = 4.94847$)

Figure 8: LMSV-Square Method Forecasting Results

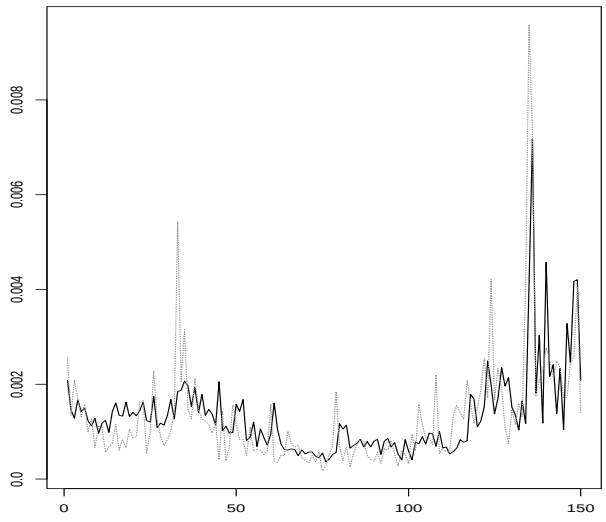


(a) Realized Vol (.....) and Forecasts (——)

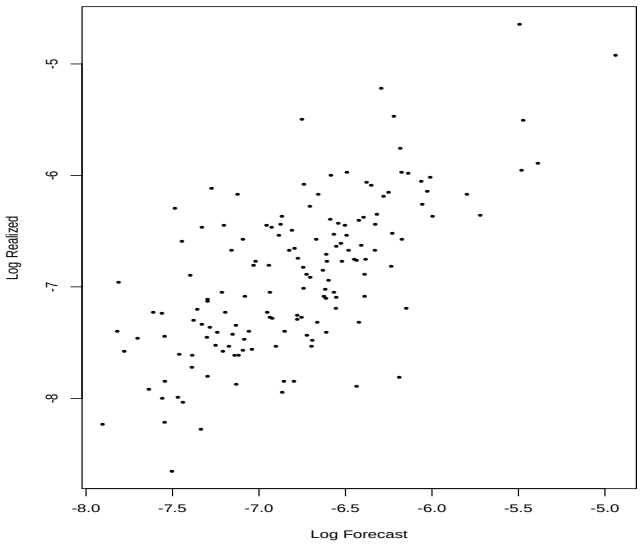


(b) Log Realized Volatility versus Log Forecasts

Figure 9: LMSV-C Method Forecasting Results (C decided by Data)

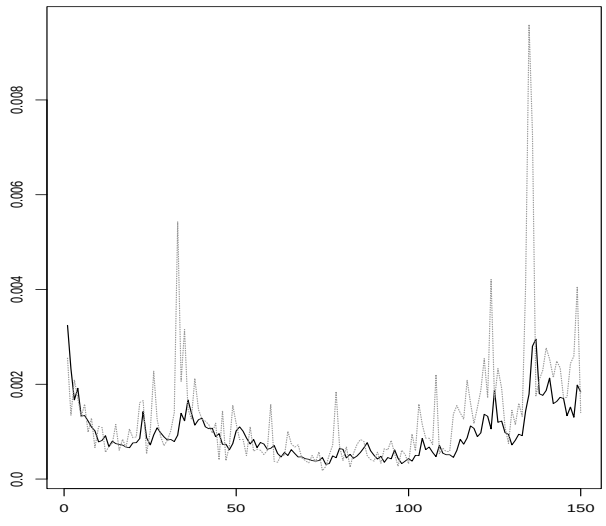


(a) Realized Vol (.....) and Forecasts (——)

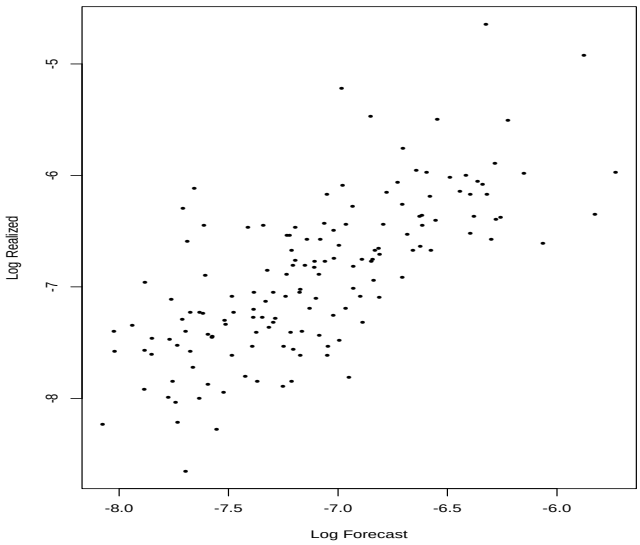


(b) Log Realized Volatility versus Log Forecasts

Figure 10: LMSV-C Method Forecasting Results ($C = 1.4$)

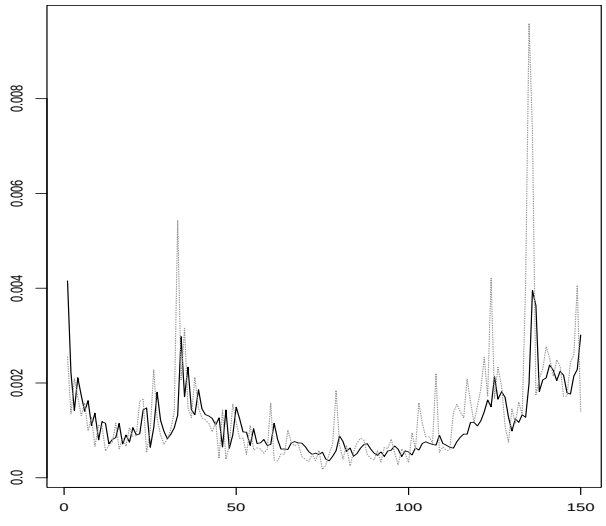


(a) Realized Vol (.....) and Forecasts (——)

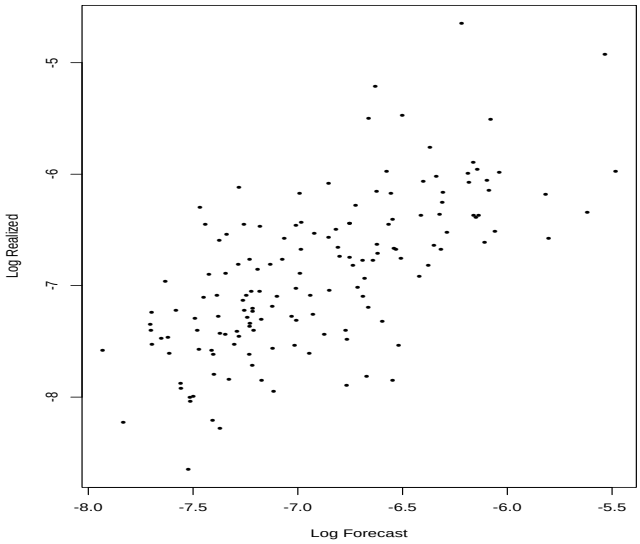


(b) Log Realized Volatility versus Log Forecasts

Figure 11: ABDL Forecasting Results

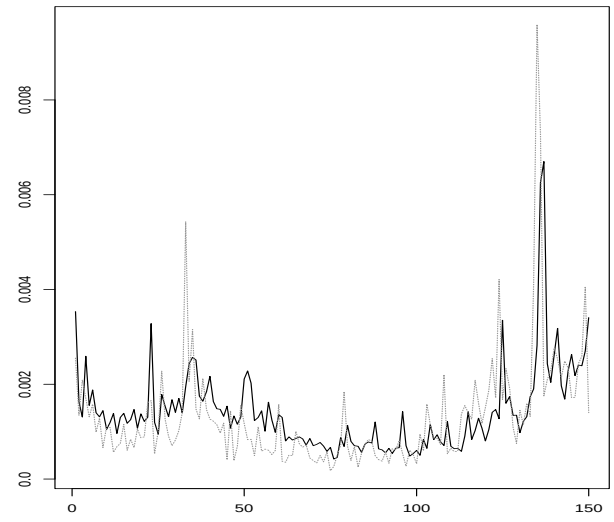


(a) Realized Vol (.....) and Forecasts (——)

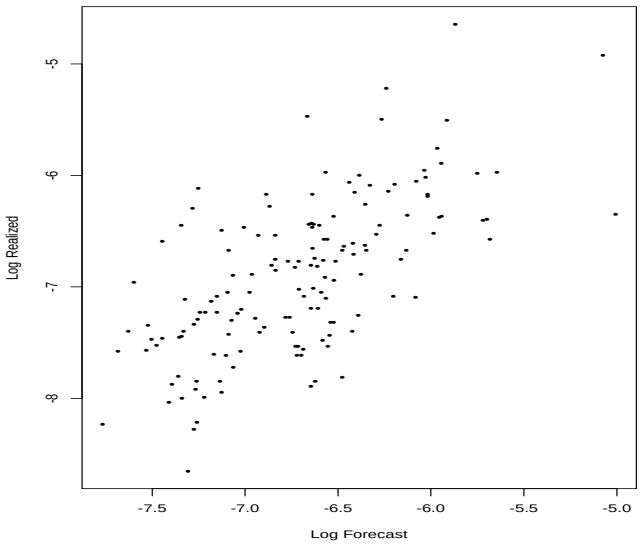


(b) Log Realized Volatility versus Log Forecasts

Figure 12: Component GARCH(1) Forecasting Results

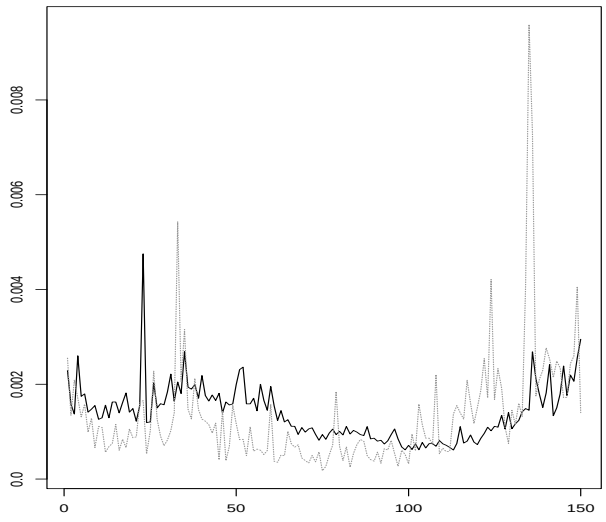


(a) Time series plot of Realized Vol (.....) and Forecasts (——)

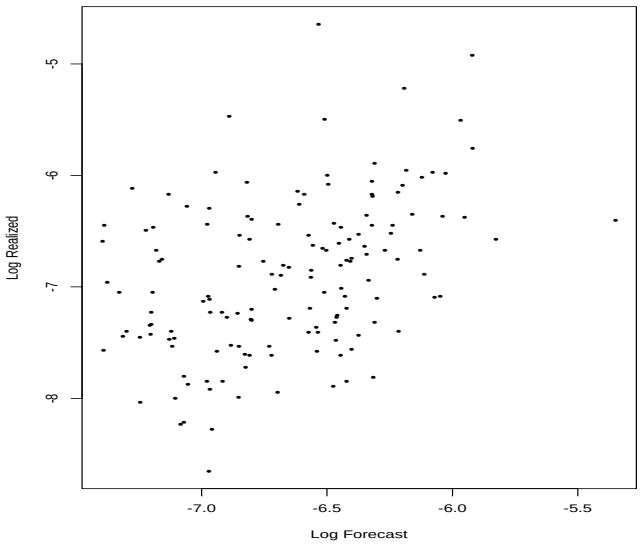


(b) Log Realized Volatility versus Log Forecasts

Figure 13: GARCH(1,1) Forecasting Results



(a) Realized Vol (.....) and Forecasts (——)



(b) Log Realized Volatility versus Log Forecasts