A Long History of Realized Volatility

Blake LeBaron *

International Business School

Brandeis University

March, 2018

Abstract

Several unique data sets are brought together to build approximate daily realized volatility estimates back to the early 1930's. Estimators are tested extensively on modern data to see how well they line up with common estimators using high frequency pricing information. Estimators are also shown to pass several diagnostic tests from the early samples as well. Finally, well known qualitative features are tested for their stability across subsamples. This includes empirical volatility term structures which give reasonable estimates for perceived persistence to volatility shocks. Recommendations are made as to best practice for estimating long horizon realized volatility.

Keywords: Realized volatility, volatility dynamics, kernel ridge regression, long memory

*415 South Street, Mailstop 32, Waltham, MA 02453 - 2728, blebaron@brandeis.edu, www.brandeis.edu/~blebaron.

1 Introduction

Modeling volatility has been one of the more important areas of research in financial econometrics. Information about volatility is critical for risk management, and option pricing. Also, the dynamics of volatility has proved to be a rich and interesting space for developing useful time series models. A revolution in volatility modeling occurred with the introduction of high frequency data sets. Intraday data was used to build estimates of the conditional variance on a given day, known as realized volatility (RV).[1] This has led to a new set of features, and more detailed dynamic models of the volatility process. Unfortunately, very high frequency intraday data is only available for relatively recent periods where intraday data is available.[2] This paper utilizes several special data sets to build realized volatility estimators back to the early 1930's, expanding the length of previous RV time series by nearly a factor of 4.

Several data sets are critical in this analysis. First, is a long intraday series of hourly observations of the Dow Jones Industrials going back to 1933. This is provided by Global Financial Data (GFD) on the U.S. stock database. A second series, also from GFD, provides special high/low range information using individual high's and low's for components of the index. This is how most index high/low ranges were estimated in earlier series, but GFD has maintained this cruder methodology into the present. Finally, the high quality intraday RV series provided on the Oxford-Man website is important for model building, estimation and testing.

The goals of this paper are to first demonstrate the methodology for generating long range RV series. Second, to test the reliability of these series both in the recent series where the actual high frequency series are available, and in the past, by demonstrating that some stylized features are relatively constant through time. Finally, the long series will be used to turn to questions of stability of the volatility process itself, leading to some incites into how traders view risk and its persistence.

The econometric methodology in this paper is simple but unusual in what it is trying to accomplish. In the modern data it is assumed that the feature of interest, realized volatility, is measured with high precision. Predictive models are then built projecting the high quality series on the coarser series. In this recent data, this exercise should only be of limited interest. It provides some information on the structure of volatility, but the need for a lesser quality RV estimator is probably not great. The key is that these fitted models can then be used "out of sample" into the past to generate very long RV time series. There is no

[1]This area is now very large, and is a standard for volatility modeling. See Barndorff-Nielsen & Shephard (2010) and Andersen, Bollerslev, Christoffersen & Diebold (2013) for two of the many surveys available.

[2]Most series begin in the early 1990's through the 2000's.

obvious way to model this, and several estimators will be raced against each other. The paper will give some results on best possible RV estimators in terms of performance.

This paper does not add to the literature on estimating realized volatility. A standard 5 minute RV estimator is taken as the benchmark. Also, most of the features that are fit are well known. These include the normality of RV estimates and standardized returns. Also, the extreme persistence, or long memory, of RV time series will also be examined in many cases. [3] Finally, the autocorrelation patterns of absolute standardized residuals are used as a new diagnostic test. This paper purposefully avoids detailed questions about volatility forecasting. Having a longer series of high quality realized volatility should lead to better predictive models. The evidence has been strong that building volatility models incorporating RV measures yields improvements in forecasting close to close volatility.[4]

Most of the RV model fits use linear models. However, an attempt is made to use nonlinear modeling at this stage. There is little if any justification for why linear models would dominate. A kernel ridge regression is implemented and displays promising improvements. However, in the diagnostic testing stage it did not show dramatic improvements versus a linear approach.[5]

The paper begins in section 2 with a description of the data sources, a description of the empirical methodology, and a motivation for the value added of using and combining lower quality series to estimate daily conditional variances. Section 3 details the fitting of various realized volatility estimators to the benchmark 5 minute RV series. In the next section, section 4, well known features of the high quality RV time series are compared with the coarser series in the modern sample. These features are then explored in the early data part of the data set (starting in 1933) in section 5. Finally, long range stability of several well known features for RV are explored across subsamples in section 6. The final section, 7, concludes, gives guidance on the use of long range RV estimators, and suggests future research.

[3]See Andersen, Bollerslev, Diebold & Ebens (2001), Andersen, Bollerslev, Diebold & Labys (2003), and Barndorff-Nielsen & Shephard (2002), for early evidence.

[4]A few early examples are Shephard & Sheppard (2010) Hansen, Huang & Shek (2011). In terms of modeling RV per se, Bollerslev, Patton & Quaedvlieg (2016) is an example incorporating high frequency fourth moments to help in controlling dynamic weights in predictive models.

[5]Other work using modern machine learning approaches in volatility are Audrino & Knaus (2016) and Chen, Hardle & Jeong (2010). Both of these studies operate on the problem of forecasting.

2 Empirical setup

2.1 Data sources

The results in this paper are driven by combining several unique time series. The most important of these are the hourly observations of the Dow Jones Industrials which are on the equities part of Global Financial Data (GFD). Hourly data may not seem very impressive, but these series extend back to 1933 giving a much longer span than we get with real time modern series. The series have been used before. Gerety & Mulherin (1994) use the price series to document volatility patterns across the day, and Gerety & Mulherin (1992) use the intraday volume information to measure the demand for trading as a function of expected volatility at the open and the close. The series moves to a 30 minute interval starting in January, 1987. They form the core of this study by giving a crude proxy for a daily realized volatility measure built from the intraday price levels.

A second important series from GFD is the theoretical Dow Industrials series. Many data sets contain time series which track both open and close, and also high/low information on each day. For individual securities, there is no confusion about what this information represents. However, for the case of an index the recording methodology has changed over time. In older data sets the high (low) of an index is measured as the high (low) for all the individual components regardless of when they occurred. This was done, because high frequency tracking of the index itself through the day was not available. Markets only kept track of high/low information for the individual components. This biases high/low range estimation because the recorded high will be larger than the true index high, and the recorded low smaller than the true low. Using most historical high/low range information therefore yields an inconsistency in recording when this methodology was changed. One cannot go back and fill in the index high, but one can estimate the modern data with the cruder old technology of individual high's and low's. This is done in the GFD *theoretical* index series. It provides a consistent, albeit biased, picture of high's and low's on the Dow index.

The final series used will be the modern realized volatility series from the Oxford-Man library (Heber, Lunde, Shephard & Sheppard 2009). This continuously updated website contains many volatility series on a range of financial instruments, including the Dow. This paper uses the 5 minute realized volatility measure for the Dow Jones Industrials which is available from January 2000 and continuously updated. This RV measure is often believed to be a good compromise between extremely high frequency volatility estimation, and the problems of microstructure noise. Many more sophisticated methods are available, but many studies have continued to stay with the 5 minute benchmark methodology.[6]

It should be noted that for all series used here, cleaning is necessary. The Oxford-Man website details a very precise cleaning procedure. The raw hourly GFD data contains clear errors where prices were entered incorrectly, efforts have been taken to eliminate most of the easy mistakes. However, it is probably the case that errors persist. The combination of the two long range time series should accomplish some smoothing of errors that have not been captured by hand.

2.2 Empirical methodology and motivation

The econometric tools used in this paper are simple, but the empirical strategy is unusual. This section gives a background justification and some contributing monte-carlo evidence for what is being done. As mentioned at the start of the paper, the goal is to build and test some very long run realized volatility time series. This is done by taking the high quality RV series from Oxford-Man and fitting it to series designed to replicate the lower quality series that are available back to 1933. The fitted models are then used to build an estimated, or proxy, realized volatility series over the long sample.

This forces one to make some unusual choices about throwing out data. First, in the modern (post 2000) era, the GFD series contains observations every 30 minutes. Estimating using this data would not be consistent with the older hourly observations, so this sampling frequency is taken down to the hourly frequency by removing observations. [7] Second, data before May 1952 contains half day trading on Saturdays. These days are dropped from the sample. Price range information has long been known to be a powerful tool for estimating volatility.[8] The data problem here is that the high/low information is not estimated consistently in standard series. Again, the choice is to use the cruder, but consistent value, the theoretical index, which estimates high's and lows's based on individual components.[9]

The hourly data for each day is used to build a robust standard deviation estimate. Let $r_{t,h}$ represent the hourly log returns on day t. A robust estimate of the daily standard deviation is given by the adjusted

[6]The Oxford-Man data set provides many RV estimators itself. For simplicity, and since the evidence suggests the 5 minute RV values are generally difficult to beat, (Liu, Patton & Sheppard 2013), this paper will use only the 5 min RV. Another set of estimators are those that are try to separate jumps from the diffusion variance as in the bipower estimator Barndorff-Nielsen & Shephard (2004).

[7]Sampling takes place at even hours starting at 10AM, and ending at 4pm. For much of the modern data this drops an observation at 9:30AM.

[8]Garman & Klass (1980) and Parkinson (1980) are early examples. More modern research using high frequency data is in Alizadeh, Brandt & Diebold (2002).

[9]For the GFD data set the actual and theoretical ranges diverge starting in July 2010.

mean absolute return, assuming a zero mean.

$$\hat{\sigma}_t^H = \sqrt{N}\sqrt{\frac{\pi}{2}}(\frac{1}{N})\sum_{h=1}^{N}|r_{t,h}| \tag{1}$$

The adjustment factor is the standard relationship between mean absolute deviation and standard deviation for a normal distribution. The $\sqrt{N}$ adjustment is to scale it up to daily frequency, for an open to close standard deviation measure. This will give the hourly RV estimate for the series.

For high/low range estimation, the basic Parkinson estimator is used,

$$\hat{\sigma}_t^P = \sqrt{\frac{\log(H_t) - \log(L_t)}{4\log 2}}. \tag{2}$$

This estimator was derived in Parkinson (1980). Estimators adding open and close information were also derived in Garman & Klass (1980). They offer some improvement, but early experiments suggest that behavior of opening prices change across the 90 years of the sample, making the simpler high/low estimator more reliable. The high/low values used here are the "theoretical" values derived from individual prices, and not the true index levels. All these series are available back to 1933, and can therefore be used for long range realized volatility estimation.

To test the usefulness of these estimators for finding daily volatility numbers, a simple monte-carlo experiment is performed. Trading days with different volatility levels are generated, with underlying prices sampled at the 1 second frequency. The Oxford-Man realized volatility levels are assumed log/normal and simulated variances are drawn from this normal distribution. The high and low price for each trading day is recorded, but has noise added which is calibrated to the deviations between the actual and theoretical high and low prices from the Dow data. It is important to stress that this monte-carlo is unrealistically optimistic in that it simulates a true Brownian motion with no microstructure noise, and Gaussian increments where all higher order moments exist.

Several volatility estimators are tested by sweeping through the time duration for high frequency price observations from 1 minute through 2 hours. A regression of the true log standard deviation on various estimators is performed,

$$\log\sigma_t = \beta_0 + \beta_1 \log\sigma_t^M + \beta_2 \log\sigma_t^P + \epsilon_t. \tag{3}$$

Figure 1 shows estimated goodness of fit, measured with the simple R^2, for the volatility estimates. The black line, labeled RV, sets $\beta_2 = 0$, and uses just the lower frequency RV estimators. The red line allows for

β_2 nonzero, and a contribution from the range estimator. The flat green line is the range estimator alone for comparison. Finally, the blue line is based on a regression on the absolute open/close daily return.

The black line shows a steadily falling level of fit for the prices as sampling frequency decreases. This represents the simple linear impact of sample size on the squared error. The red line shows that combining the two volatility measures improves the fit dramatically for relatively low frequency observations in the range of 50 to 120 minutes. An interesting benchmark is the case of a 5 minute sampling level. At this frequency the fit is near perfect, and there would be no need to add the range estimator. Also, the range estimator on its own is not very impressive with only an R^2 of 0.65. However, its near 5 fold improvement over a simple open/close estimator demonstrates the weakness of daily returns by themselves for volatility measurement. With hourly data, the simulation demonstrates that it likely will be useful to combine the hourly RV value with a range estimator. In this case a relatively good fit with an R^2 of 0.8 can be achieved, suggesting that accurate long range daily volatility time series are possible.[10]

3 Fitting RV estimators

This section estimates and compares several RV estimators. Remember, the strategy is to take the modern 5 minute RV estimate as the target. Given this value, the objective is to get the most precise estimator in the recent data, using inputs which are available back to 1933. These can then be used to build long time series of daily volatility estimates.

Given our series, the target function $g()$ is what will be estimated.

$$\log RV_t = g(\log(\sigma_t^H), \log(\sigma_t^P)) + \epsilon_t \tag{4}$$

RV_t refers to the 5 minute realized volatility from the Oxford-Man data. The hourly robust volatility estimate is given by σ_t^H, and the Parkinson range estimate is given by σ_t^P. These are all in units of open/close standard deviations. Estimation is all in logged volatility space. This respects the fact that most realized volatility estimates have been shown to be near log normal. This is the case for the 5 minute RV value used here.

[10]It is interesting to note that several values in figure 1 line up with estimates from Garman & Klass (1980). First the relative efficiency of the range estimator versus the open close estimator is about a factor of 5 which is the point of these early papers. Their range estimators using open/close information show an efficiency relative to the open/close of 7-8 which could make a difference here, but as mentioned, in the actual data, market open values appear to change through time. Parkinson (1980) compares relative efficiency of subsampling methods to range estimators and finds equivalent efficiency when sample size for the intra day is about 5 prices per day. This is not far off the crossing point at 75 minutes in figure 1.

Most of the practical estimates of $g()$ will assume a linear model,

$$\log(RV_t) = \beta_0 + \beta_1 \log(\sigma_t^H) + \beta_2 \log(\sigma_t^P) + \beta_3 |r_t|. \tag{5}$$

The last term, $|r_t|$ represents the absolute open/close return and is added primarily for comparison purposes. It will not be used in the later RV functions. Simple OLS model estimates are given in table 1. All estimation uses the late era sample which extends from January 2000 through October 31, 2017. The table presents parameter estimates, standard errors are given for each. It also shows the R^2 along with the sample mean squared error (MSE), and mean absolute error (MAD) on the RV target. Obviously, all the coefficients are highly significant, as should be expected. These are all measures of the same daily volatility level. The first two rows of the table show that the hourly information gives a more accurate RV fit then the range estimator alone, with a MSE of 0.107 versus 0.142. As is well known both these estimators offer dramatic improvements over the baseline absolute return which gives a MSE of 0.195, corresponding to an R^2 of only 0.364. Combining hourly and range information improves the fit with a MSE of 0.088. As should be expected, adding the open/close return does not provide much improvement to the hourly/range model.

The final line of the table provides a check with much simpler estimates. In this case the two volatility estimates themselves are taken as the estimates for $\log(RV_t)$. No regression is performed. Since these values are all estimators of the same thing this is a useful comparison. In both cases the fit is now much worse. This demonstrates clearly that the primitive estimators need some form of bias adjustment. This is completely consistent with the first two rows of the table, where for unbiased estimators we would have $(0,1)$ for the constants and coefficients.

There is no theoretical justification for why $g()$ should be linear, and it would be important if a nonlinear fit might improve the RV estimator. The key question is just what sort of nonlinear econometric technology should be used for $g()$. Modern machine learning provides many approaches to approximating $g()$.[11] It is beyond the scope of this paper to test all of them, but some rudimentary decisions and comparisons have been made. The model comparison and search is greatly assisted by the Python Scikit Learn package. Not only does it contain many of the tools mentioned here, but its consistent interface across models allows for easy comparison and cross validation.

For a modern data analytic example this one is not all that standard. Modern "big data" problems usually consist of large cross sections and many potential right hand side variables. In this problem the

[11] This field, which has a long history, is best summarized by Hastie, Tibshirani & Friedman (2009).

data set is only moderately sized (about 4000 observations), and the number of right side variables is small (2). The sample size makes it unlikely that highly parameterized, deep learning, neural networks will succeed.[12] There is still a set of nonlinear models which may be useful. Particularly important are kernel based approaches. Support vector regressions have been used in several finance applications, but they are designed more for classification, and usually involve an objective function which is nonstandard for time series analysis.[13] A relatively new tool is kernel ridge regression. This is both more directly applicable to standard least square objectives. Also, it has recently seen some early success in macroeconomic time series forecasting.[14] Kernel ridge combines a large sequence of nonlinear kernels on the right-hand side along with the standard ridge regression (L^2) penalty function to avoid over fitting. This model was the most reliable of the nonlinear models used.

For the small data set, small explanatory variable problem, two other nonlinear methods also should be considered. Nearest neighbor regression, and kernel regression.[15] Experiments were tried with both methods. The appropriately cross validated nearest neighbor model yielded results similar to the kernel ridge regression. The kernel regression results were unstable, and bandwidth calibration was tricky. Neither method gave any indication for improvements beyond kernel ridge.

The kernel ridge model contains two crucial meta parameters that control its fit. The ridge penalty weight is given by α, and the bandwidth for the kernel is γ. These correspond exactly to the parameters in the SciKit learn system. The parameters are determined using a grid search procedure, and the kernel function is the radial basis function (RBF) kernel. The objective is out of sample MSE, using 5-fold randomized cross validation on the sample.[16] The grid search range was guided by ranges suggested in Exterkate (2013). The grid search found values of $\alpha = 0.075$, and $\gamma = 0.4$, although the results are relatively stable across a general range of parameters.

Model testing and comparison is done for a set of models in the late sample from Jan 2000 - Oct 2017. MSE and MAE are estimated on training and test data sets generated by 5 fold randomized cross validation. Selection into training (4/5) and validation/testing (1/5) is repeated 500 times recording model accuracy measures in each run. Table 2 reports the results for various candidate models for mapping $\log(RV_t)$, the

[12] For some early attempts with limited data see LeBaron & Meng (2012).

[13] See Chen et al. (2010) for an application of support vector machines for volatility prediction.

[14] See Exterkate, Groenen, Heij & van Dijk (2016). These papers also include much of the mathematical structure and details of the model that are skipped here.

[15] Cleveland (1979) is an early primer on the basic technology, also Hardle (1990) is an early textbook. Diebold & Nason (1990), LeBaron (1992),Meese & Rose (1990), Mizrach (1992) are all early examples of applications in economics and finance.

[16] Randomized cross validation chooses a noncontiguous set of training and testing periods. This is important for time series such as this, which may experience different regimes moving through the sample period.

5 minute realized volatility measure, onto the two coarser measures, $\log(\sigma_t^H)$ and $\log(\sigma_t^P)$. The first four columns report the mean and standard deviation of the MSE and MAE across the different models. The final two columns report the fraction of runs for which a given model beats one of two benchmark targets, the kernel ridge, and the hourly only case which refers to the linear projection of log(RV) on the hourly volatility measure. These will be important comparisons for which models to continue with.

The first line shows a mean MSE and MAE for the kernel ridge of 0.074, and 0.216 respectively. These represent the smallest MSE for the model set, and all simulations beat the simple hourly only case. The second line uses the same right hand side predictors, but replaces the nonlinear ridge specification with a basic linear specification as estimated in table 1. Estimation consistently follows the 5-fold cross validation in all cases. Mean MSE drops to 0.089, and the second to last column indicates that none of the linear model estimates beat the kernel ridge model in terms of MSE.[17] The last column shows that in the space of linear models the addition of range information to the linear specification is important in terms of model fit. All of the estimated hour/range models beat the hourly only specification.

The row labeled 50/50 is a naive benchmark using an equal weight of both hourly and range estimates. It does surprisingly well, but not as well as the estimated linear model. The final two rows look at the hourly alone, and the range alone. The hourly alone performs relatively well with a MSE of 0.107, but the linear and kernel models have already been demonstrated to be better fits. Following figure 1 the range only regression model seems to practically and statistically do much worse than the others with a MSE of 0.142.

General features and comparisons of the different fitted RV models are given in table 3 and figure 2. Both show a large amount of correlation between the fitted values and their target, which is as to be expected. The range alone model may exhibit the least amount of correlations, but this is difficult to take a strong stand on. Visually it is the one scatter plot which stands out as being less sharp than the other 3. As a final comparison a picture of the time series of the 5 Min RV and the kernel ridge is shown in figure 3. It is difficult to see much of a difference.

4 Testing RV estimators against RV 5 min

Since they were first estimated and created, realized volatility measures have yielded a set of reliable empirical features. The purpose of this section is to use the latter sample to compare the candidate RV estimators

[17] Note that both the MSE and MAE are estimated with tight precision. These are standard deviations on the MSE, not on the mean MSE.

to the 5 min value to see how they compare. It has already been shown that the fit is not perfect for any of the estimators. However, some may still capture the latent integrated volatility process as well as the 5 min RV values. These tests will help decide which, if any, of the estimators are fit to be used in the earlier data.

Figure 4 displays estimated densities for logged volatility measures. These are logs of the volatility measured as a daily standard deviation. Similar to the previous results there is a general similarity with normal distributions. However, for two measures, the kernel and range, there are strong visual deviations from normality. The kernel case appears to be somewhat skewed, while the range exhibits leptokurtosis.

The next figures examine the standardized returns. These are daily returns (close/close) divided by appropriate volatility measures in units of standard deviation. For comparison, figure 5 presents distributional properties for the logged daily returns in the sample both in a histogram, and a qqplot. They are strongly nonnormal and leptokurtotic as is well known. Normalizing returns by an estimate of the conditional standard deviation should help to reduce the nonormality in the series. In a pure mixtures of normal distribution world returns would follow,

$$r_t = \sigma_t z_t \qquad z_t \sim N(0,1), \tag{6}$$

and therefore dividing by the appropriate conditional standard deviation should yield a normal.[18] Figure 6 reports this for returns standardized by 4 of the RV estimators in the 2000-2017 sample, and figure 7 repeats this display using qqplots. Visually, the returns standardized by the 5 minute RV appears to be the best as should be expected. However, the linear model is not that far off normality. Both the kernel, and especially the range show pretty strong deviations from normality. The range normalized returns are strongly platykurtic with tails which are drawn in from a normal distribution. Figure 7 also adds the restricted hourly based RV model which drops the range out of the regression. This does not appear as close to normality as the model with both range and hourly information (labeled linear in the plot).

These graphical features are backed up by various tests in table 4. This table presents empirical measures of normality to get a sense of how close the standardized returns are to a normal distribution. Three different measures are presented. The kurtosis is shown along with the Jarque-Bera test for normality. Finally, the table displays a quantile range ratio,

$$Q = \frac{q_{0.995} - q_{0.005}}{q_{0.95} - q_{0.05}}. \tag{7}$$

[18] See Clark (1973) for the early work on this process for returns.

This ratio measures the extreme quantile range (0.01) relative to a smaller range (0.10). It is a more robust measure of a deviation from normality, and the table presents this value divided by the ratio from a normal distribution. The table clearly rejects normality for all the series. However, as others have noted the reduction in kurtosis and the changes in the quantile ratios are dramatic. Again, the most normal looking series is the 5 minute RV normalized returns, but the linear model is very close. The kernel and the range both appear relatively far off normality with extreme tails which are too thin relative to a normal as was shown in the figures.

If the pure mixtures model were completely true, then σ_t is capturing the rate of information flow, or economic activity in the market. One of the best documented facts in all of finance, is the persistence of σ_t through time.[19] Correctly normalizing should leave $|z_t|$, or z_t^2 as white noise. Figure 8 displays autocorrelations for the absolute values of the standardized returns, and raw returns for comparison. Any correlation displayed here reveals remaining volatility structure not captured in the RV estimator. The figure displays a general lack of correlation for 3 of the 4 volatility estimators. The key exception is the range only estimator which reveals a large amount of residual autocorrelation.[20] All four standardized return series display dramatic reductions in the persistence of absolute returns relative to the raw, unadjusted, returns series, which is large and positive as is well known. As a secondary diagnostic to better understand what is happening with the standardized returns, they are all plotted in figure 9. visually one should expect no patterns in volatility. Only the range series shows some indication of what appears to be a dramatic downward shift in volatility beyond year 2010. It appears likely that something has changed in this series across the sample.

The final key feature that has been documented for realized volatility is their extreme persistence. Autocorrelations of the RV series display a hyperbolic decay pattern which goes out well beyond a year. It is a classic signature of a long memory, or fractionally integrated series. Figure 10 shows the autocorrelation for the RV estimates in the modern data set. The 5 minute RV series shows the most persistence, but all 4 series display a similar pattern. The ACF decay is very slow and positive correlations continue out until at least 250 days. As a performance test, it appears that all the RV estimators are close to the 5 minute RV in terms of persistence.

[19] It has been well known since Mandelbrot (1963) and Black (1976), and the more modern development of ARCH/GARCH machinery in Engle (1982) and Bollerslev (1986).

[20] All three of the good RV estimators display a small amount of negative correlation at lag 1. This could be created by a miss timing around the open. The Oxford-Man RV drops off the first part of each trading day, as does the hourly price information. This might cause volatility to get logged a day after it hits returns. If that happens it might generate a small amount of negative correlation as seen in the figures.

Another signature of long memory is the rate at which the variance of partial sums expands,

$$\operatorname{var}\sum_{i=1}^{n}(y_i) \sim n^{2H}, \tag{8}$$

where H is the self-similarity parameter. It is related to the rate of decay of the autocorrelations through,

$$\rho_k \sim k^{-\gamma} \tag{9}$$

$$\gamma = 2H - 2 \tag{10}$$

It also connects to the more commonly used level of fractional differencing, d, through

$$d = \frac{1}{2}(2H - 1). \tag{11}$$

The slope of the variance of partial sums in 8 can then be used to estimate H and d.[21]

Figure 11 displays the log/log plot of the variance expansions from equation 8. The estimated slopes yield levels of fractional differencing given in table 6. For example, for the 5 minute RV series, the estimated value of $d = 0.44$, and for log it is $d = 0.47$. It is clear from the table that the estimates for all 4 RV estimators yield similar results. The values estimated here are slightly larger than those previously reported. Also, it is curious that the log volatility levels are higher, but this is consistent with other estimates. Finally, the log volatility estimate for d is close to the stationarity boundary of $d = 0.5$.[22]

It has been known for a long time that the long memory in volatility could be represented as the sum of a small number of short memory processes.[23] Recently, Corsi (2009) has developed a simple 3 factor model which has proved useful in fitting the dynamics of persistent financial volatility processes. This model takes the form

$$rv_t = \beta_0 + \beta_1 rv_{t-1} + \beta_2 rv_{t-1|t-5} + \beta_3 rv_{t-1|t-22}. \tag{12}$$

The model can be estimated in levels or logs. For this paper the log framework will be used throughout with $rv_t = \log(RV_t)$. This is done both acknowledging the near log normality of the various RV estimators,

[21] A good exposition of long memory technology is Beran (1994). Also, LeBaron (2001) estimates this for simple daily returns. LeBaron also shows that tail behavior of raw returns is well replicated by the dynamic volatility process. This feature is also reported in Warusawitharana (2016).

[22] See Andersen et al. (2003) for comparison estimates.

[23] See Granger (1980) for the general theory, and Granger & Ding (1996) for examples generating long memory. LeBaron (2001) shows that a 3 factor model with different past horizons works very well in terms of displaying long memory features, though the actual model is a short memory process.

and also it allows for easy monte-carlo simulation that does not need to impose positivity on the dynamics of the model. Following the standard notation of the model, the lags represent means at different horizons,

$$rv_{t-h|t-k} = \frac{1}{k-h+1} \sum_{j=t-k}^{j=t-h} rv_j. \tag{13}$$

Therefore, the model tries to represent both short (daily) and longer (weekly, monthly) horizons for volatility. The framework is a restricted AR model with restrictions that correspond to natural calendar dates.[24]

Parameter estimates are given in table 6. Parameter estimates across the different RV estimators are similar, but the fit of the model as measured by the R^2 does drop off when moving to the lower quality RV estimators. This indicates that dynamically they must be adding some noise to the predictable features of the underlying latent volatility process captured by the RV 5 min series. The table provides some basic information on how the goodness of fit is affected by the different volatility measures. The column labeled, "R^2 alternative" restricts the parameters to the model estimated on alternative RV estimators. In the case of Kernel, Linear, and Range, it imposes the dynamic model estimated from the RV 5 min series. For the first row, the RV 5min data is predicted with the model estimated using the Linear RV specification. In all cases there is little qualitative change in the goodness of fit, indicating similar dynamics must be at work.[25]

The alternative RV estimators have now been extensively tested in the later data set, where detailed high frequency price information is available. The overall results show that for the most part they display features almost exactly the same as the higher quality volatility estimator. The two "best" estimators, the Kernel, and Linear estimators both perform very well. The linear model does slightly better in terms of producing standardized returns which look similar to those standardized by the 5 min RV. It is also the case that the RV estimator using range information alone was inferior to the other models. The results suggest that these new estimators may be effective in the early parts of the sample where no high frequency data is available. The next section turns to testing on the early data.

[24]See work by Dacorogna, Gencay, Muller, Olsen & Pictet (2001) and Lynch & Zumbach (2003) for evidence supporting the HAR model, and some of its alignments with calendar inspired breaks.

[25]A more formal goodness of fit was implemented by imposing the parameter restriction of fixing the coefficients to the RV 5 min estimates and then applying to the other samples. In each case for this test the F-statistic strongly rejected the specification. These tests results are not reported because this test is not formally a correct use of an F-test. The fact that the alternative RV estimators are actually generated values that use the RV 5min series itself in their construction makes the correct specification test nontrivial, and the traditional test may be inappropriately sized.

5 Testing out of sample (early periods)

Having a set of slightly inferior RV estimators is not the mission of this paper. The key is that data is available to take these estimators back to the start of the hourly data in 1933. This section repeats the previous set of diagnostics, but on the early data. This data can be viewed as taking the models "out of sample" since they are being used on data far from where their parameters were estimated. However, this is not a standard forecasting experiment. The target series, the RV 5 min estimate, is not available in the past, so no direct test of model fit is possible. Only the indirect tests can be repeated here.

Figure 12 displays the linear RV estimator fit to both hourly and range information across the early sample years, 1933-1999. The RV is displayed in units of open/close daily standard deviations. Generally, the figure appears typical for many RV estimated series. There is a large, obvious spike in 1987. Also, the early 1930's display the well known amount of large volatility as the market dropped at the beginning of the Great Depression.

Figure 13 repeats the display of the distribution of the log of the various RV estimators. The figure adds the 1 HR RV estimator as well. This repeats the same linear based model used in the other specifications, but limits the input to the hourly volatility level. No range information is used. This tests the value added of range information in early RV model generation. Other models are those used before, with all parameters coming from the 2000-2017 period. In all cases, the general normality result holds. There is a slight skew for the kernel, and some potential leptokurtosis for the range, but not all that strong.

The more important diagnostic looks at the distributions for the standardized returns. These are the returns divided by the estimated RV on each day. These densities are displayed in figure 14. Normality looks like a fair approximation as we saw in the late data. However, there is a little more evidence for platykurtic behavior (thin tails). This can be seen in more detail for the qq plots displayed in figure 15. Again, approximations to normality appear good, with some slight deviations in the tails.

Table 5 repeats the quantitative tests performed in the later sample. All series statistically reject normality, but levels of kurtosis and the quantile ratios are not too far off normality. The basic features of the qq plots are supported in that the distributions are close to normal, but with tails that are a little thin.[26]

The previous section looked at the autocorrelation for the absolute standardized returns to see if they left any dynamic structure. These autocorrelations are shown in figure 16, which should be compared to figure

[26]It is possible that this feature comes from the fact that the RV estimators are estimated in log space. Standardized returns require exponentiating the estimator back. Formally there should be a bias adjustment at this stage. It is not done both for simplicity, and the fact that this would involve estimating another parameter, the variance of the log volatility. Future work, may implement this correction.

8 that displays the same autocorrelation on the later data. All the volatility standardized return series show a dramatic reduction in comparison to the raw returns. Again, the range only based estimator performs poorly with a large amount of remaining absolute return autocorrelation. The other three RV estimators are similar. Their autocorrelation is small, but positive, going out many lags. A pattern such as this could be indicative of a regime shift in the data. To check this the standardized returns themselves are plotted in figure 17 for the linear and range estimators. There is no obvious regime shift in the data, but the graphs do show some amount of residual conditional heteroskedasticity. This is very pronounced for the range only estimator. Historically, the early 1930's show the well known pattern of high volatility even in the standardized residuals. It is interesting that the estimated RV values cannot remove this, but it is also an empirical challenge that is not part of the subsample where the models were fit.

The autocorrelation patterns for the RV estimators are shown in figure 18. The strong persistent autocorrelations from figure 10 are repeated here. The estimators are close to each other, except for the range only estimator which consistently shows more persistence across the entire sample. The behavior again seems consistent with a possible long memory process even on this data set which ends before the other sample starts.

6 Dynamic features and stability

It has been demonstrated that several of the coarser RV estimators can be taken back into the earlier data, and they give general performance similar to more modern, high frequency RV estimators. This section takes the longer range RV estimate as given, and applies it to some early tests of stability of RV and return dynamics across the much longer time series that are now available. This analysis would get complicated by carrying the set of 5 different estimators around, so one champion is determined. The linear model using both hourly and range information has generally performed well. In most tests it is comparable to the kernel based estimator, which did generate smallest MSE on the late sample. It will be used for two basic reasons. First, it is a much simpler model than the kernel ridge. Turning on the extra nonlinear kernel technology appears to not be useful or necessary. Second, in terms of standardized residuals it has performed slightly better than the kernel ridge model. In summary, out of a set of close competitors, the linear model with hourly and range information combined will be declared the winner and used for subsample estimation in this section.

The dynamics of volatility have always been the core of empirical volatility modeling. Figure 19 turns to the basic results on the realized volatility properties across the full sample and several subsamples. The subsamples are designed to be roughly equal, and to cut on even decade dates. The first subsample is slightly unusual since it technically contains the period where Saturday morning trading is active. For the others, this is not an issue. The patterns in figure 19 replicate the general extreme persistence observed before for all subsamples. However, quantitatively there is a moderate amount of dispersion between the different subsamples.

This graphical result is repeated in table 7. The variance scaling estimates for the level of fractional integration are repeated. For the raw data, there is now much more dispersion in the estimated values for d_v, for the log data the values are very similar, but it is interesting that the subsamples are now lower than the full sample.[27] The HAR model of Corsi (2009) is again estimated for the full sample and the subsamples. Predictability across the various subsamples varies, but parameter estimates are close. A formal test of stability is reported in the column labeled "Chow". This is a simple Chow test for stability between the subsample given in the row, and the 2000-2017 sample, where the linear RV model was fit. Stability is rejected for 2/3 other subsamples. Interestingly, it is not rejected for the earliest period, 1933-1949. The R^2 column again reports the R^2 for the fitted model along with a pseudo R^2 imposing the restriction of using the parameters from the 2000-2017 subsample. There is essentially no change which is interesting given other indications of model instability.

Use of long time series enables a more detailed analysis of the behavior of long range volatility forecasts. The question of how quickly investors should expect volatility to return to normal when a big increase is observed is important for the performance of markets, and better understanding of models with learning and dynamics. The long range data is used for a simple analysis of the dynamics of volatility forecasting many periods out into the future. This is an empirical volatility term structure. To the extent that it represents possible investor beliefs it will have a big impact on their behavior.[28]

Empirically, a very simple approach will be used here. Volatility forecasts are built by projecting future volatility onto current using standard deviation units,

$$RV_{t+j} = \beta_{0,j} + \beta_{1,j} RV_t, \tag{14}$$

a direct long range forecast. The term structure is obtained by rolling the forecast out across the estimated forecast parameters, indexed by j and estimating this over overlapping periods. Rather than presenting

[27] This may be an indication that they are running out of data. In future work monte-carlo standard errors will be necessary here.

[28] For recent results on empirical persistence of volatility shocks see Wang & Yang (2017).

the actual parameters a graphical representation of the term structure is given in figure 20. The figure represents the conditional expectation of future standard deviations given a 2 standard deviation shock in RV_t. The term structure presented here was estimated on the post 2000 sample, and is used to compare the 5 min RV with the fitted linear RV model. The figure shows that the results are similar for the two volatility estimators which again gives support for using the linear model on the long range data series. They are particularly close over the first 100 day range, showing a shock half life of about 25 days, and a quarter life of about 100 days.

Figure 21 repeats the term structure experiment from figure 20 for the various subsamples using the linear RV estimator. The figure shows a relatively diverse set of results across subsamples. For example, in the 1980-1999 period, most of a 2 standard deviation volatility shock is eliminated after 100 days. However, for the 1950-1979 period, the forecast stays close to 0.5 (one quarter of the shock) out nearly 200 days. There appears to be little overall agreement, suggesting that in a long range sense the older data are not helpful in understanding how long a shock to volatility will last in current periods.

Figure 22 uses a subsampling experiment to emphasize this uncertainty.[29] The full sample from 1933-2017, and the linear RV series are used to give a large population from which to draw. The structure of the experiment is to simulate a possible draw of a decade long sample that an investor might use to estimate the half life of a volatility shock. The simulated forecast term structure is performed on blocks of 2500 days of contiguous data drawn randomly from the entire sample. In each case a forecast curve as in our earlier figures is estimated. This is repeated for 500 redraws of the subsamples, with the median and 95 percent confidence bands presented in figure 22. This figure confirms the earlier results that there is a lot of uncertainty in the long sample about the persistence of volatility shocks. It suggests that investors may be very cautious in their forecasting of future volatility, or else one would possibly observe dramatic changes in their beliefs about volatility dynamics as we move through time. For example, the figure loosely shows that the range for 3/4 of a volatility shock to disappear could be as short as a few days, and as long as 100 days. The series suggests a very imprecise forecast.[30]

Another feature of volatility which can be analyzed for stability of the long series, is the impact of past returns on future volatility. This is often referred to as the "leverage effect".[31] There is a well documented asymmetric impact from past returns to future volatility. Volatility increases are stronger in falling markets.

[29] See Lahiri (2003) for some of the basic information on subsampling and resampling methods.

[30] It should be noted that this forecasting model is very crude. Bringing more lags and more information might reduce some of this uncertainty.

[31] Originally documented in Christie (1982) as related to the amount of leverage at various firms, it has recently been reexamined in Hasanhodzic & Lo (2011) who show it being present even with firms carrying no debt, so its causes are still not well understood.

A graphical approach to examine it is the news impact curve developed in Engle & Ng (1993).[32] Figure 23 estimates the curve over several different subsamples. The estimated lines using a nearest neighbor forecast with the number of neighbors (250) estimated with a 5 fold cross validation using the most recent subsample. The news impact schedules appear surprisingly similar over the inner range of the return data. In a range from about -0.03 through 0.03 the volatility impacts appear similar, and the asymmetry is quite pronounced.

There are two exceptions to this regularity. First, the extreme returns show dispersion across the subperiods, but this may be driven by a few extreme data points. Some robustness testing in the tails would seem appropriate here. Second, there is a deviation in the 1980-1999 period where the schedule diverges even in the range of returns near 0.00. This is strange, and may be driven by the crash of 1987 which is an extreme outlier for the volatility series.

7 Conclusions

The main point of this paper is simple. Extending daily volatility measures well into the early 20th century is feasible and reliable. Longer time series either allow for more precise parameter estimates, or explorations of stability, and the potential impact of changing institutional structure on market behavior. All of this is done for financial price volatility where, after decades of detailed research, we still have limited theoretical models for what generates many of its interesting empirical features.

This paper has shown that it is possible to build reliable long range daily realized volatility (RV) estimators using hourly intraday data along with high/low range information. The results show that using both of these empirical measures together improves the RV estimators. The best format appears to be a simple linear model. One conjecture for why this performs well is that both the hourly and range estimators are based on true latent volatility plus noise. As long as the noise in the two pieces of information is not perfectly correlated then some combination will improve on using them individually. It is possible that adding them is the best way to eliminate this noise. The linear combination model replicates many stylized features of realized volatility in the recent sample, and in the early data as well. From the standpoint of basic diagnostics it appears useable as a good proxy for having high frequency data in periods where it is unavailable.

[32]This has been incorporated into many models of conditional variance dynamics including Glosten, Jagannathan & Runkle (1993) and Nelson (1991). For some richer dynamics see Curci & Corsi (2012).

The paper also presents some interesting results using a nonlinear approach with a kernel ridge regression. These results show a reliable nonlinear relationship on the recent data. However, taking this model into the past did not improve on the linear model, and along some dimensions might have done worse. The puzzle of just what nonlinear structure is in recent data, or how the model is working with volatility remains an interesting question for the future.

The paper ends with a final demonstration of what is possible with such long range series. Subsamples are used to show a few things. First, the dynamics and persistence of volatility does not appear to be stable. Trader beliefs about volatility should at least be viewed with a large amount of uncertainty, or the data may support a wide range of different beliefs about the empirical volatility term structure. Second, the asymmetric impact of returns on near term volatility appears to be relatively stable across much of the data. It is interesting that this empirical feature, which has little theoretical support, is so robust.

This study has to be viewed as a preliminary view of an under explored time series. It demonstrates its reliability and uses, but remains quiet on some big issues for the future. One major area that is ignored is volatility prediction. It does not look at volatility forecasting in a detailed fashion. The dynamics of the RV process are examined, but the further question of extending this to close-to-close volatility estimates, or the incorporation into RV-GARCH models and dynamic volatility control strategies, is left for future research. A long range RV estimator may be extremely useful in all these practical applications of volatility prediction. It remains to be seen whether daily volatility is necessary for longer range volatility modeling. For example if one is interested in the dynamics of monthly volatility, then a monthly volatility estimate built from daily returns might be all that is needed.[33]

Theoretical models that generate persistent volatility may be rare, but models where agents possess heterogeneous beliefs generate persistent volatility dynamics almost generically.[34] Some of these models can become quite detailed in which features they are able to replicate.[35] The detailed volatility series provide a much needed resource for detailed model calibration and validation in this area.

Extending realized volatility time series to long time horizons is possible. It is facilitated by the availability of very long range intraday data. Hopefully, longer series such as this one will lead to better understanding of what the causes are for changing volatility in financial prices, a puzzle that still remains after many years of study.

[33]See French, Schwert & Stambaugh (1987) and LeBaron (2015) for examples of this simpler approach.
[34]See Hommes (2006) and LeBaron (2006) for many early examples.
[35]LeBaron (2013) qualitatively goes after many aspects of volatility dynamics.

References

Alizadeh, S., Brandt, M. W. & Diebold, F. X. (2002), 'Range-based estimation of stochastic volatility models', *Journal of Finance* **57**, 1047–1091.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F. & Diebold, F. X. (2013), Financial risk measurement for financial risk management, *in* G. Constantinedes, M. Harris & R. Stulz, eds, 'Handbook of the Economics of Finance', Vol. 2, Elsevier, pp. 1127–1220.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Ebens, H. (2001), 'The distribution of realized stock return volatilty', *Journal of Financial Economics* .

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **96**, 579–625.

Audrino, F. & Knaus, S. D. (2016), 'Lassoing the HAR model: A model selection perspective on realized volatility dynamics', *Econometric Reviews* **35**(8-10), 1485–1521.

Barndorff-Nielsen, O. E. & Shephard, N. (2002), 'Econometric analysis of realized volatility and its use in estimating stochastic volatility models', *Journal of the Royal Statistical Society B* **64**(2), 253–280.

Barndorff-Nielsen, O. E. & Shephard, N. (2004), 'Power and bipower variation with stochastic volaitlity and jumps', *Journal of Financial Econometrics* **2**, 1–37.

Barndorff-Nielsen, O. E. & Shephard, N. (2010), Measuring and modeling volatility, *in* R. Cont, ed., 'Encyclopedia of Quantitative Finance', John Wiley, pp. 1898–1901.

Beran, J. (1994), *Statistics for Long-Memory Processes*, Chapman and Hall, New York, NY.

Black, F. (1976), 'Studies of stock price volatility changes', *Proceedings of the American Statistical Association, Business and Economics Statistics Section* pp. 177–181.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **21**, 307–328.

Bollerslev, T., Patton, A. J. & Quaedvlieg, R. (2016), 'Exploiting the errors: A simple approach for improved volatility forecasting', *Journal of Econometrics* **192**, 1–18.

Chen, S., Hardle, W. K. & Jeong, K. (2010), 'Forecasting volatility with support vector machine-based garch model', *International Journal of Forecasting* **29**(4), 406–433.

Christie, A. A. (1982), 'The stochastic behavior of common stock variances: Value, leverage and interest rate effects', *Journal of Financial Economics* **10**, 407–432.

Clark, P. K. (1973), 'A subordinated stochastic process model with finite variance for speculative prices', *Econometrica* **41**, 135–55.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Corsi, F. (2009), 'A simple approximate long-memory model of realized volatility', *Journal of Financial Econometrics* **7**(2), 174–196.

Curci, G. & Corsi, F. (2012), 'Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling', *Journal of Business and Economic Statistics* **30**(3), 368–380.

Dacorogna, M. M., Gencay, R., Muller, U. A., Olsen, R. B. & Pictet, O. V. (2001), *An Introduction to High-Frequency Finance*, Academic Press, San Diego, CA.

Diebold, F. X. & Nason, J. M. (1990), 'Nonparametric exchange rate prediction?', *Journal of International Economics* **28**, 315–332.

Engle, R. F. (1982), 'Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation', *Econometrica* **50**, 987–1007.

Engle, R. F. & Ng, V. K. (1993), 'Measuring and testing the impact of news on volatility', *Journal of Finance* **48**, 1749–1778.

Exterkate, P. (2013), 'Model selection in kernel ridge regression', *Computational Statistics and Data Analysis* **64**, 1–16.

Exterkate, P., Groenen, P. J. F., Heij, C. & van Dijk, D. (2016), 'Nonlinear forecasting with many predictors using kernel ridge regression', *International Journal of Forecasting* **32**, 736–753.

French, K. R., Schwert, G. W. & Stambaugh, R. F. (1987), 'Expected stock returns and volatility', *Journal of Financial Economics* **19**, 3–29.

Garman, M. B. & Klass, M. J. (1980), 'On the estimation of security price volatilities from historical data', *Journal of Business* **53**, 67–78.

Gerety, M. S. & Mulherin, J. H. (1992), 'Trading halts and market activity: An analysis of volume at the open and the close', *The Journal of Finance* **47**(5), 1765–1784.

Gerety, M. S. & Mulherin, J. H. (1994), 'Price formation on stock exchanges: The evolution of trading within the day', *Review of Financial Studies* **7**(3), 609–629.

Glosten, L. R., Jagannathan, R. & Runkle, D. E. (1993), 'On the relation between the expected value and the volatility of the nominal excess return on stocks', *Journal of Finance* **48**, 1779–1801.

Granger, C. W. (1980), 'Long memory relationships and the aggregation of dynamic models', *Journal of Econometrics* **14**, 227–238.

Granger, C. W. J. & Ding, Z. (1996), 'Varieties of long memory models', *Journal of Econometrics* **73**, 61–77.

Hansen, P. R., Huang, Z. & Shek, H. H. (2011), 'Realized GARCH: A joint model for returns and realized measures of volatility', *Journal of Applied Econometrics* .

Hardle, W. (1990), *Nonparametric Regression*, Cambridge University Press.

Hasanhodzic, J. & Lo, A. W. (2011), Black's leverage effect is not due to leverage, Technical report, Boston University.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Heber, G., Lunde, A., Shephard, N. & Sheppard, K. K. (2009), Oxord-Man Institute's Realized Library, Technical Report 0.2, Oxford-Man Institute, University of Oxford, Oxford, UK.

Hommes, C. H. (2006), Heterogeneous agent models in economics and finance, *in* L. Tesfatsion & K. L. Judd, eds, 'Handbook of Computational Economics', Elsevier, pp. 1109–1186.

Lahiri, S. N. (2003), *Resampling methods for dependent data*, Springer-Verlag, Heidelberg.

LeBaron, B. (1992), 'Forecast improvements using a volatility index', *Journal of Applied Econometrics* **7**, S137–S150.

LeBaron, B. (2001), 'Stochastic volatility as a simple generator of apparent financial power laws and long memory', *Quantitative Finance* **1**, 621–631.

LeBaron, B. (2006), Agent-based computational finance, *in* L. Tesfatsion & K. L. Judd, eds, 'Handbook of Computational Economics', Elsevier, pp. 1187–1233.

LeBaron, B. (2013), Heterogeneous agents and long horizon features of asset prices, Technical report, Brandeis International Business School.

LeBaron, B. (2015), Dynamic risk avoidance and the attraction of short memory volatility forecasts, Technical report, International Business School, Brandeis University.

LeBaron, B. & Meng, X. (2012), Nonlinear connections between realized volatility and high/low range information, Technical report, International Business School, Brandeis University.

Liu, L. Y., Patton, A. J. & Sheppard, K. (2013), Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes, Technical report, Duke University.

Lynch, P. & Zumbach, G. (2003), 'Market heterogeneities and the causal structure of volatility', *Quantitative Finance* **3**, 320–331.

Mandelbrot, B. B. (1963), 'The variation of certain speculative prices', *Journal of Business* **36**, 394–419.

Meese, R. A. & Rose, A. K. (1990), 'Nonlinear, nonparametric, nonessential exchange rate estimation', *American Economic Review* **80**, 192–196. %K %X %F.

Mizrach, B. (1992), 'Multivariate nearest-neighbor forecasts of EMS exchange rates', *Journal of Applied Econometrics* **7**, S151–63.

Nelson, D. B. (1991), 'Conditional heteroskedasticity in asset returns: A new approach', *Econometrica* **59**, 347–370.

Parkinson, M. (1980), 'The extreme value method for estimating the variance of the rate of return', *Journal of Business* **53**, 61–65.

Shephard, N. & Sheppard, K. (2010), 'Realising the future: Forecasting with high-frequency–based volatility (heavy) models', *Journal of Applied Econometrics* **25**, 197–231.

Wang, J. & Yang, M. (2017), Conditional volatility persistence, Technical report, University of Technology Sydney.

Warusawitharana, M. (2016), Time-varying volatility and the power law distribution of stock returns, Technical report, Federal Reserve Board.

$$\log(RV_t) = \beta_0 + \beta_1 \log(\sigma_t^H) + \beta_2 \log(\sigma_t^P) + \beta_3 |r_t|$$

Table 1: Realized Volatility function estimates

Base	β_0	β_1	β_2	β_3	R^2	MSE	MAE
Hourly	-1.415 (0.039)	0.680 (0.007)			0.651	0.107	0.255
Range	-0.975 (0.055)		0.815 (0.011)		0.534	0.142	0.304
Abs(OC)	-5.214 (0.009)			40.6 (0.806)	0.364	0.195	0.346
Hourly+Range	-0.579 (0.044)	0.485 (0.009)	0.380 (0.012)		0.714	0.088	0.235
All	-1.106 (0.065)	0.475 (0.009)	0.294 (0.014)	8.207 (0.749)	0.721	0.085	0.233
Hourly alone					0.327	0.206	0.341
Range alone					0.485	0.158	0.314

Model estimation: Fitted parameters, late sample 2000-Oct 2017. Estimation is by OLS. MSE and MAE are mean squared error, and mean absolute error respectively. The last two rows do not estimate regressions, but simply use the volatility estimates from hourly data, and range information.

Table 2: Model Comparisons

Base	Mean(MSE)	σ(MSE)	Mean(MAE)	σ(MAE))	MSE<MSE(Kernel)	MSE<MSE(Hourly)
Kernel ridge	0.074	0.004	0.216	0.005		1.00
Linear	0.089	0.004	0.236	0.005	0.00	1.00
50/50	0.100	0.006	0.250	0.006	0.00	0.95
Hourly	0.107	0.006	0.256	0.006	0.00	
Range	0.142	0.006	0.303	0.007	0.00	0.00

Model comparisons: MSE and MAE are estimated using 5 fold randomized cross validation on the late sample (2000-2017). Estimates correspond to means and standard deviations from 500 simulated training/testing sample splits. The last two columns correspond to the fraction of model simulations that beat the kernel ridge model, and the hourly only linear specification respectively.

Table 3: RV estimator correlations

	RV 5min	Kernel ridge	Linear	50/50	Hourly	Range
RV 5min	1.00	0.81	0.77	0.84	0.81	0.73
Kernel ridge		1.00	0.95	0.94	0.99	0.69
Linear			1.00	0.89	0.95	0.66
50/50				1.00	0.94	0.89
Hourly					1.00	0.69
Range						1.00

Estimated contemporaneous correlations across realized volatility (RV) estimators. All estimates from the post 2000 sample.

Table 4: Normality Tests: returns and standardized returns

	Skewness	Kurtosis	JB Test	Quantile ratio
Returns	-0.26	10.4	0.00	1.50
RV 5min	0.03	2.7	0.01	0.93
Kernel	0.02	2.3	0.00	0.80
RV Hour	0.07	3.5	0.00	1.08
Linear	0.02	2.6	0.00	0.91
Range	-0.09	2.2	0.00	0.80

Distributional diagnostics, 2000-2017 subsample. JB test is the p-value from a Jarque-Bera normality test. Quantile ratio reports the ratio of the 0.01 to 0.10 quantile ranges. This number is normalized by dividing by the value for a standard normal distribution. Therefore, 1 corresponds to normal, >1 would represent leptokurtosis.

Table 5: Normality Tests: returns and standardized returns

	Skewness	Kurtosis	JB Test	Quantile ratio
Returns	-1.18	43	0.00	1.45
RV Hour	0.00	2.9	0.00	0.99
Kernel	-0.01	2.6	0.00	0.92
Linear	-0.01	2.6	0.00	0.95
Range	-0.06	3.1	0.00	0.98

Distributional diagnostics, 1933-1999 subsample.

Table 6: Modern sample: HAR model estimates and fractional integration (d)

	β_0	β_1	β_2	β_3	R^2	R^2 alternative	d_v	$d_{\log(v)}$
RV 5min	-0.29 (0.05)	0.31 (0.02)	0.43 (0.03)	0.20 (0.03)	0.69	0.68	0.44	0.47
Kernel	-0.43 (0.06)	0.16 (0.02)	0.50 (0.03)	0.25 (0.03)	0.57	0.56	0.43	0.48
Linear	-0.36 (0.06)	0.16 (0.02)	0.48 (0.03)	0.29 (0.03)	0.55	0.54	0.43	0.48
Range	-0.39 (0.07)	0.07 (0.02)	0.57 (0.04)	0.28 (0.03)	0.51	0.49	0.42	0.49

This table displays some of the quantitative features for the various $\log(RV)$ measures estimated in the 2000-2017 period. The first columns are parameter estimates for the HAR model. R^2 measures the R-squared of the HAR model regression. The column labeled "R^2 alternative" reports the R-squared forcing the parameters from the RV 5 min estimation on the alternative RV estimators. For the case of the RV 5min value, the parameters from the linear hourly/range model are used. The final two columns report estimates of fractional difference levels, d, for both the volatility and log volatility series.

Table 7: Subsamples: HAR model estimates and fractional integration (d)

	β_0	β_1	β_2	β_3	R^2	Chow	d_v	$d_{\log(v)}$
Full sample	-0.36 (0.03)	0.19 (0.01)	0.40 (0.01)	0.33 (0.01)	0.51		0.41	0.43
2000-2017	-0.36 (0.06)	0.16 (0.02)	0.48 (0.03)	0.29 (0.03)	0.55		0.43	0.48
1933-1949	-0.32 (0.07)	0.18 (0.02)	0.43 (0.04)	0.32 (0.03)	0.56/0.56	0.67	0.42	0.47
1950-1979	-0.52 (0.07)	0.23 (0.01)	0.31 (0.03)	0.34 (0.03)	0.41/0.41	0.00	0.39	0.46
1980-1999	-0.67 (0.09)	0.14 (0.02)	0.42 (0.03)	0.30 (0.03)	0.32/0.32	0.00	0.36	0.48

This table displays some of the quantitative features for the various $\log(RV)$ measures estimated in the 2000-2017 period. The first columns are parameter estimates for the HAR model. R^2 measures the R-squared of the HAR model regression. The column labeled R^2 reports both the fitted R-squared on the subsample, but also the R^2 using the model parameters from the 2000-2017 period. "Chow" reports results of a Chow test (p-values on the F-test) for parameter stability between the given subsample, and the 2000-2017 period. The final two columns report estimates of fractional difference levels, d, for both the volatility and log volatility series.

Figure 1: Sampling frequency experiments

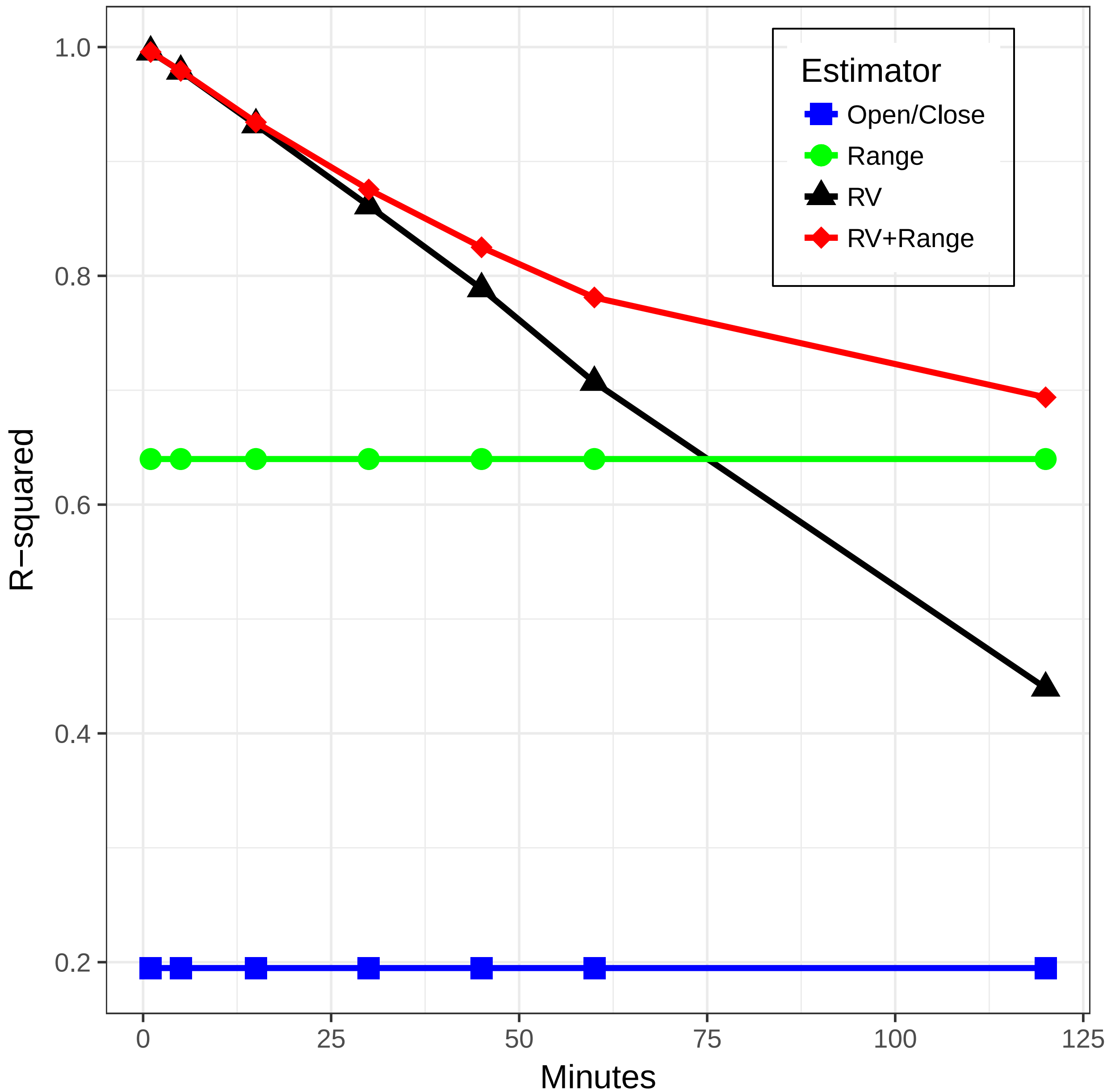

Figure 2: RV 5 min versus estimators

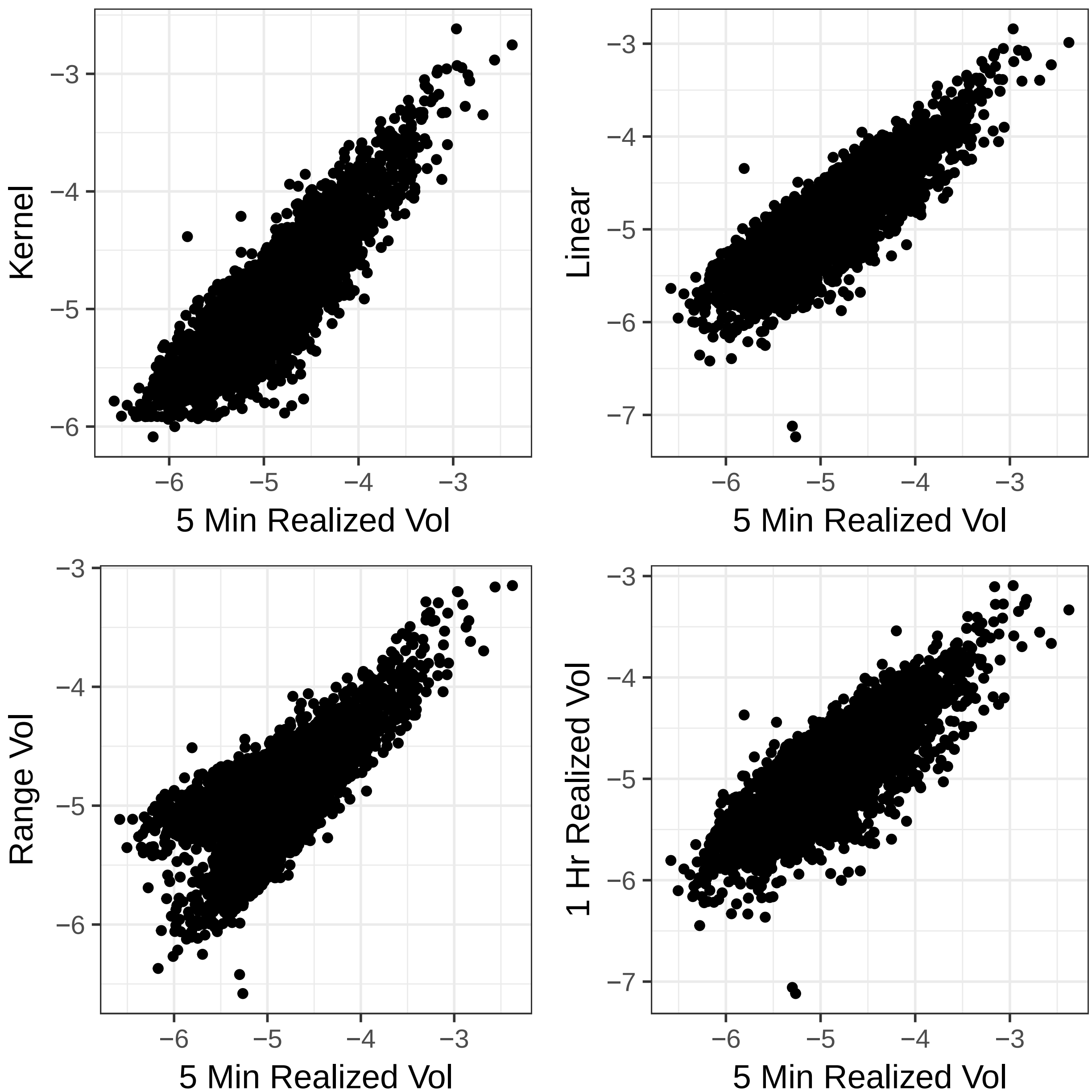

Figure 3: Volatility series: Late sample

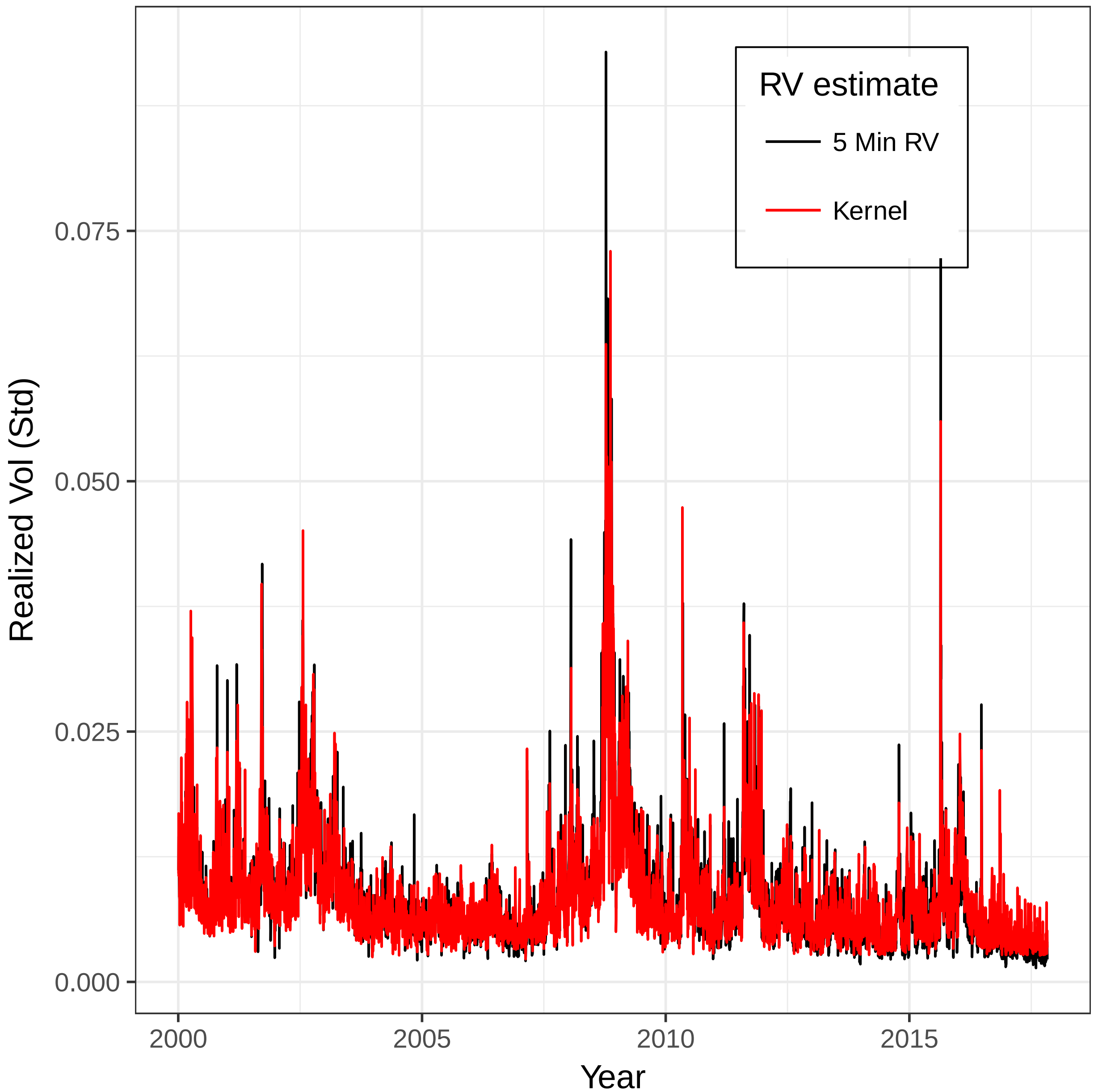

Figure 4: Log volatility densities

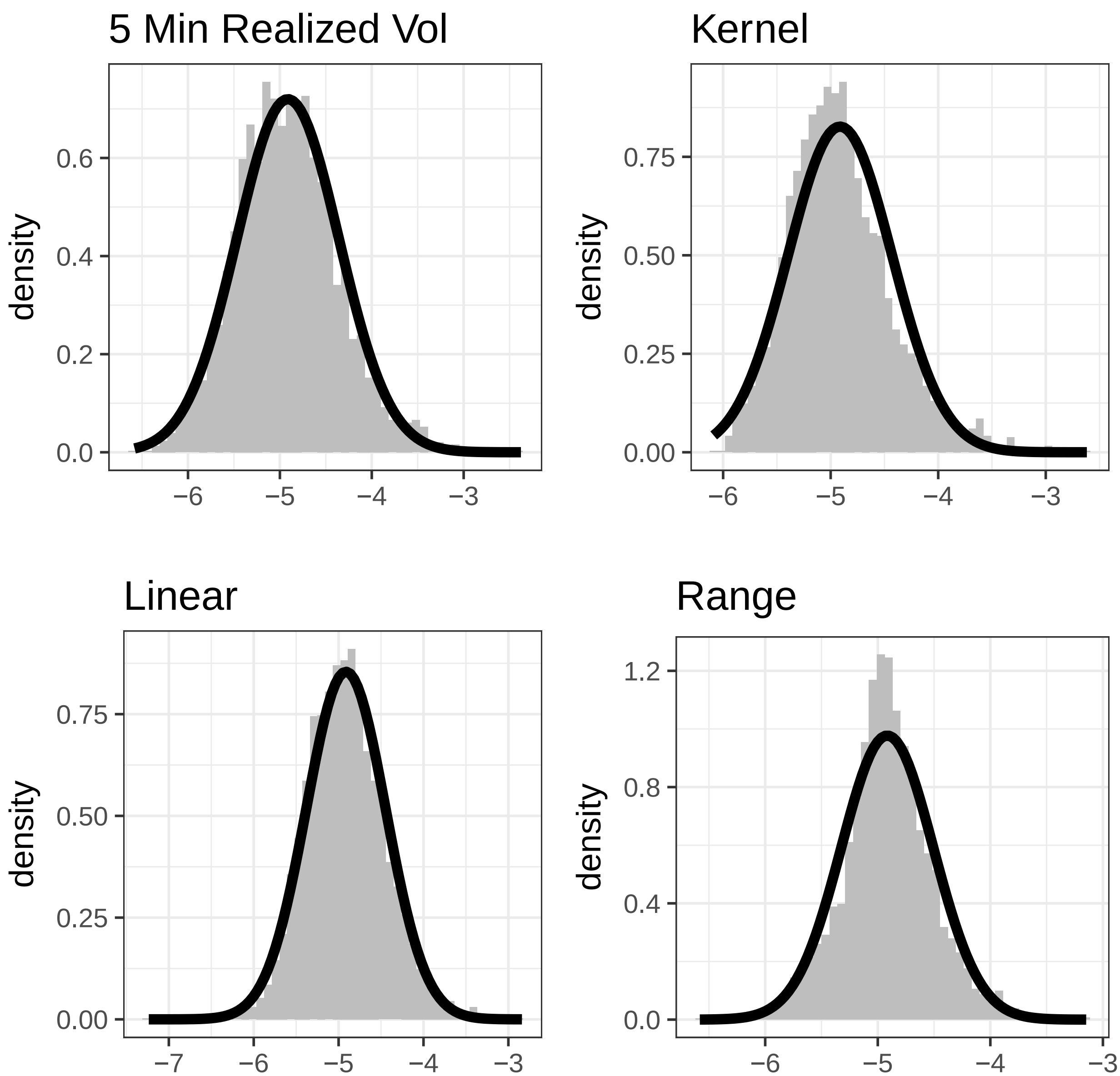

Figure 5: Raw return densities

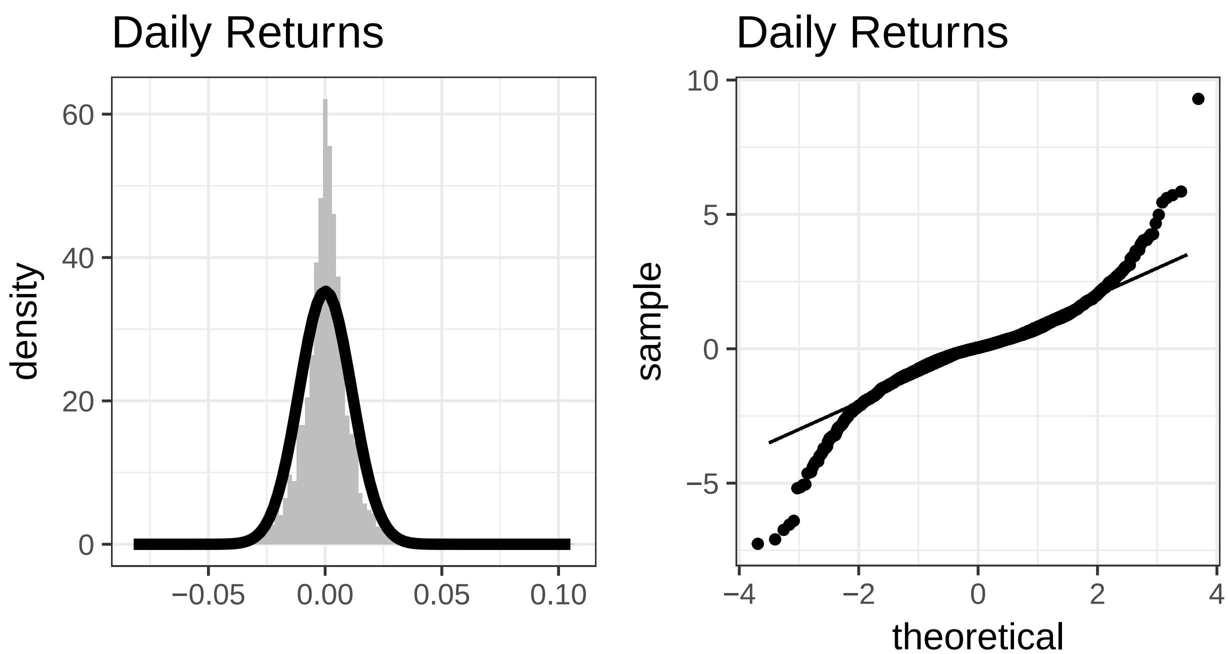

Figure 6: Standardized returns: Densities

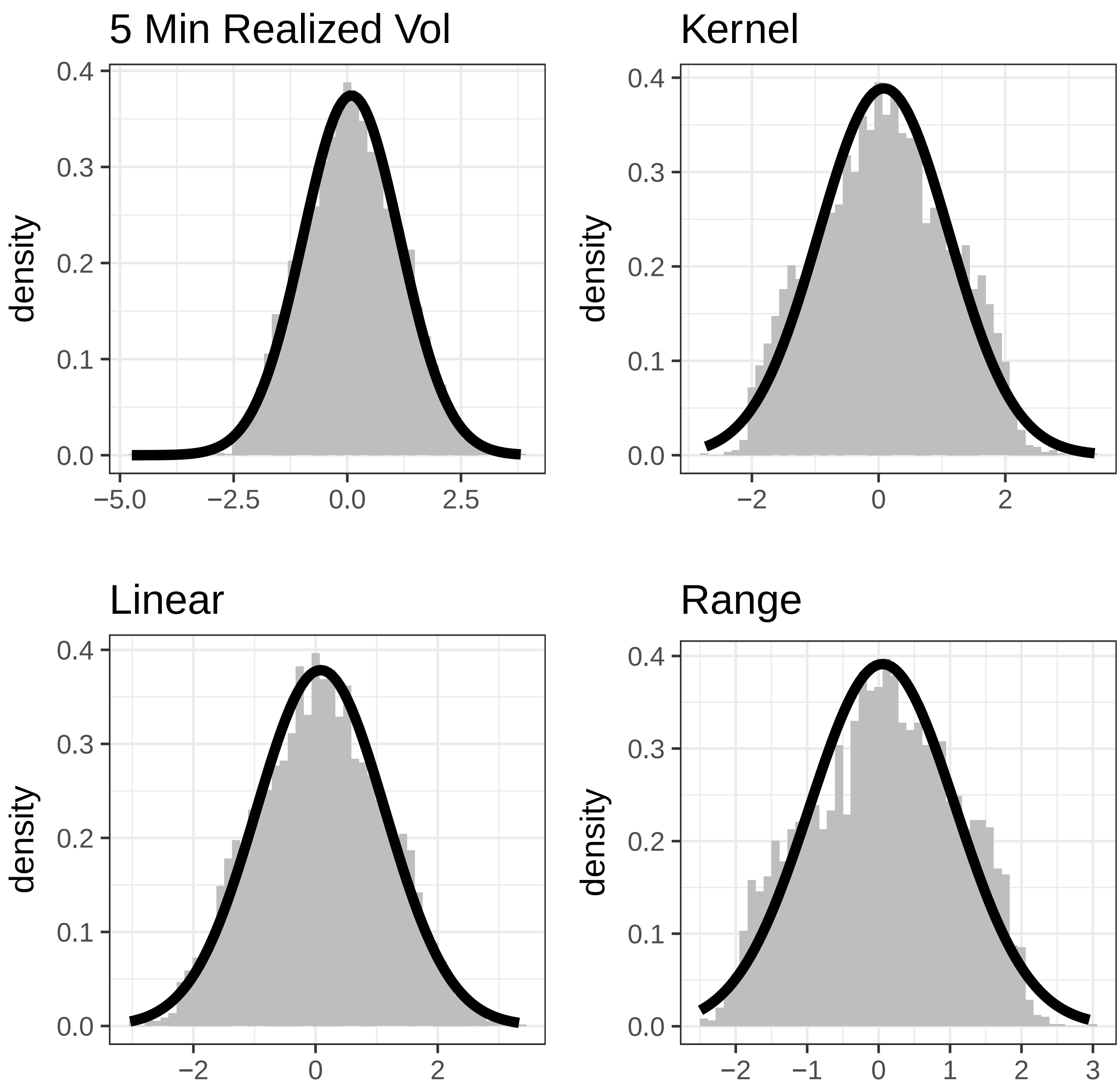

Figure 7: Standardized returns: qqplots against normal

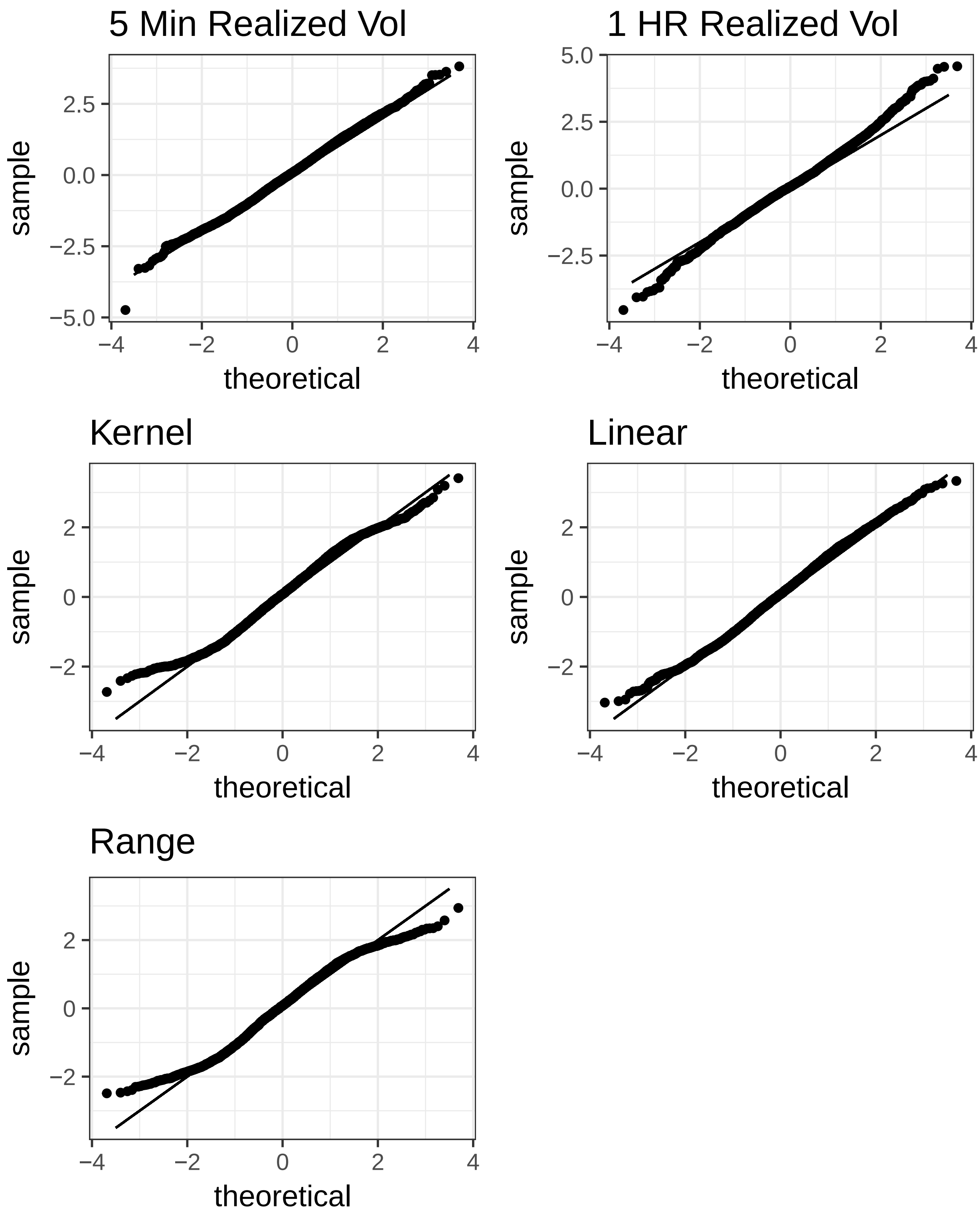

Figure 8: Absolute standardized return autocorrelations

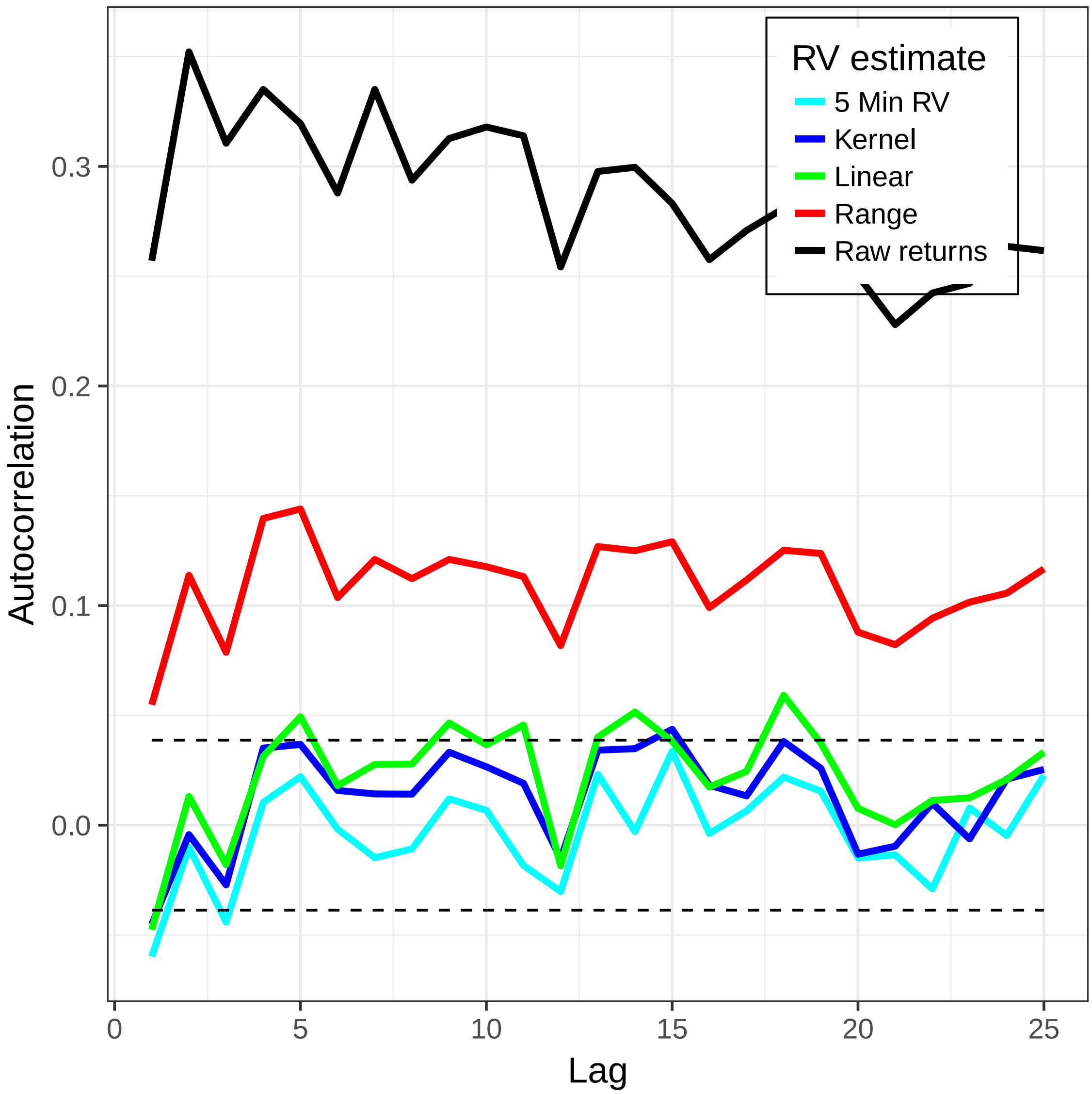

Figure 9: Standardized returns

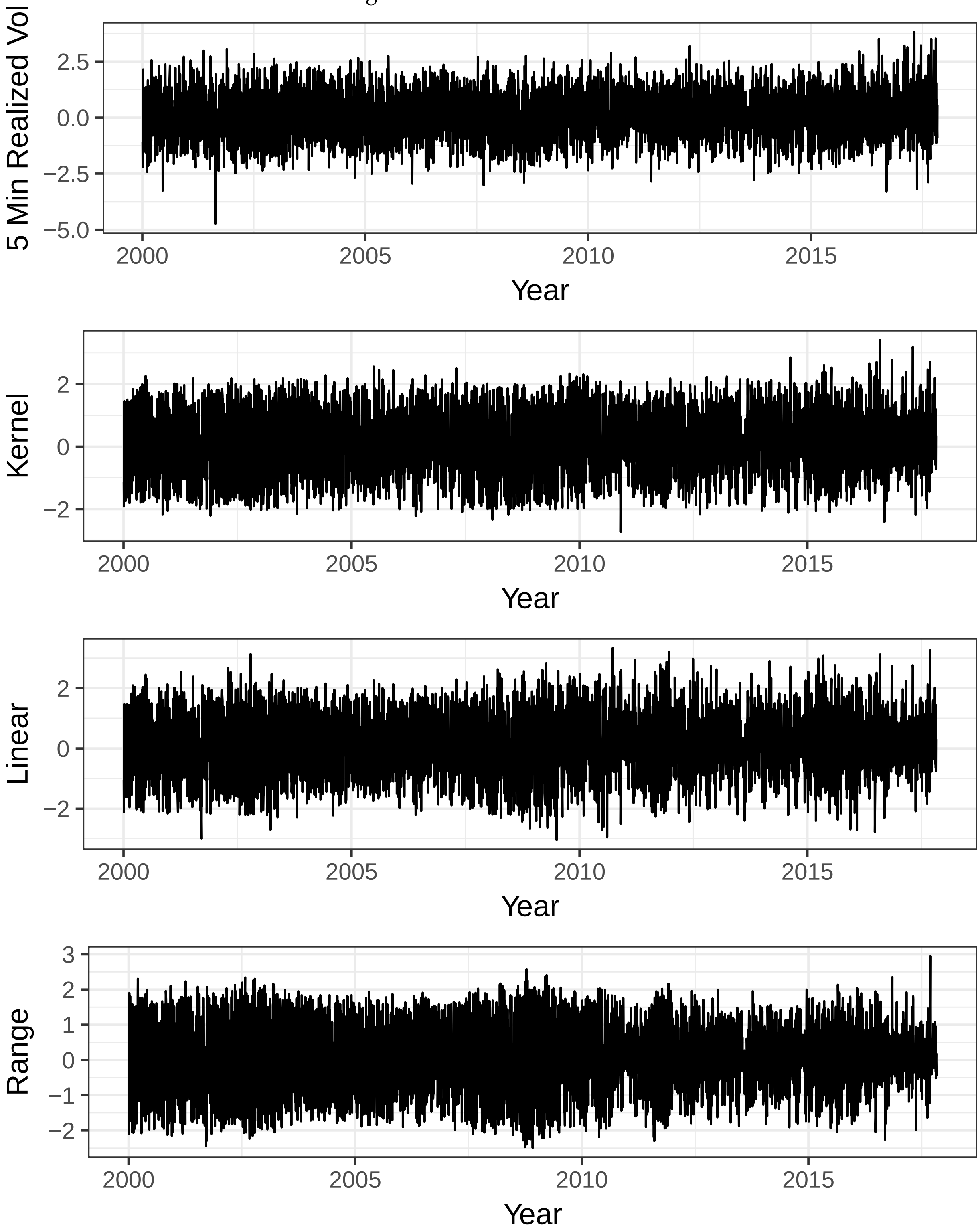

Figure 10: Volatility autocorrelations

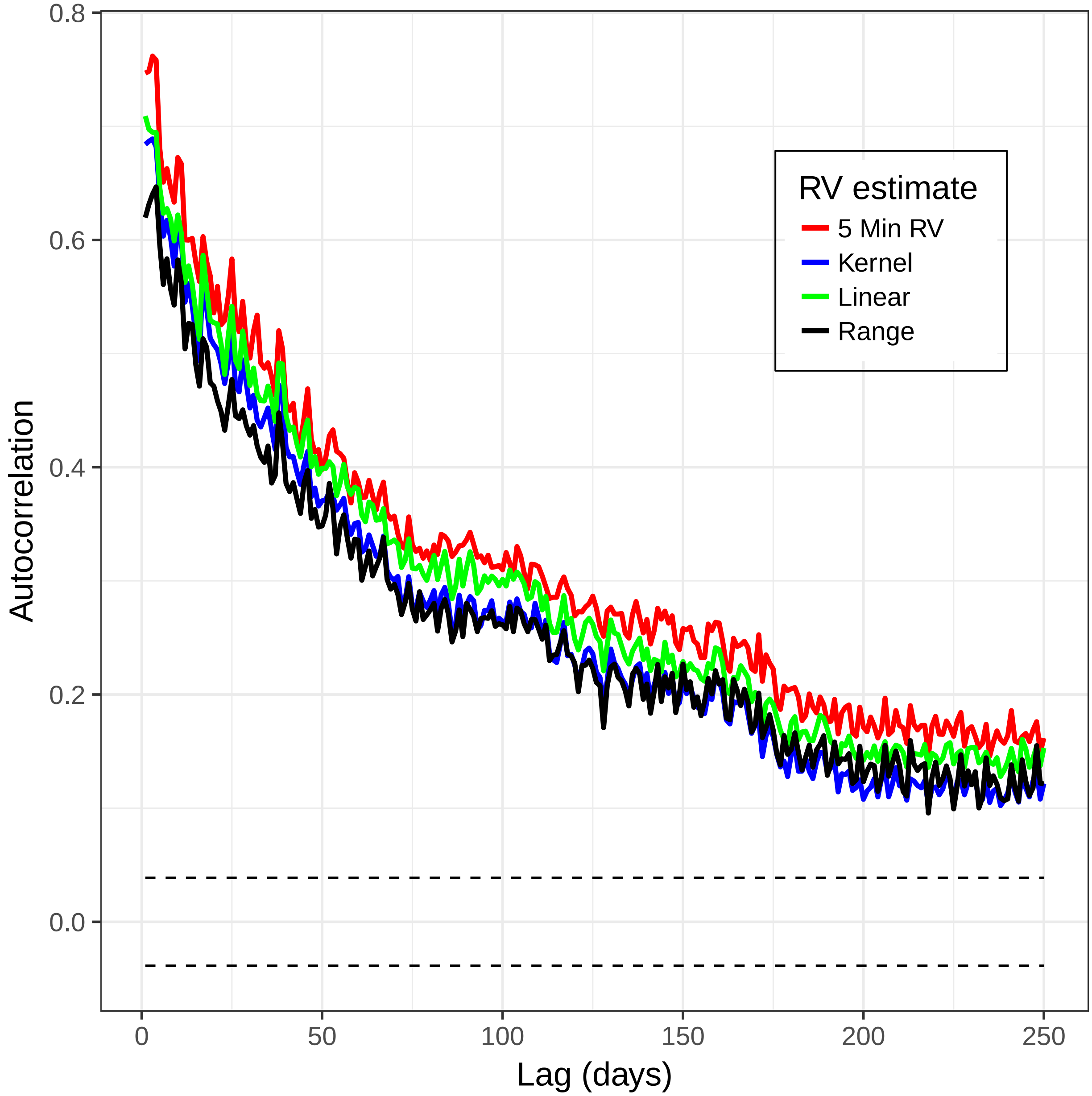

Figure 11: Variance of partial sums

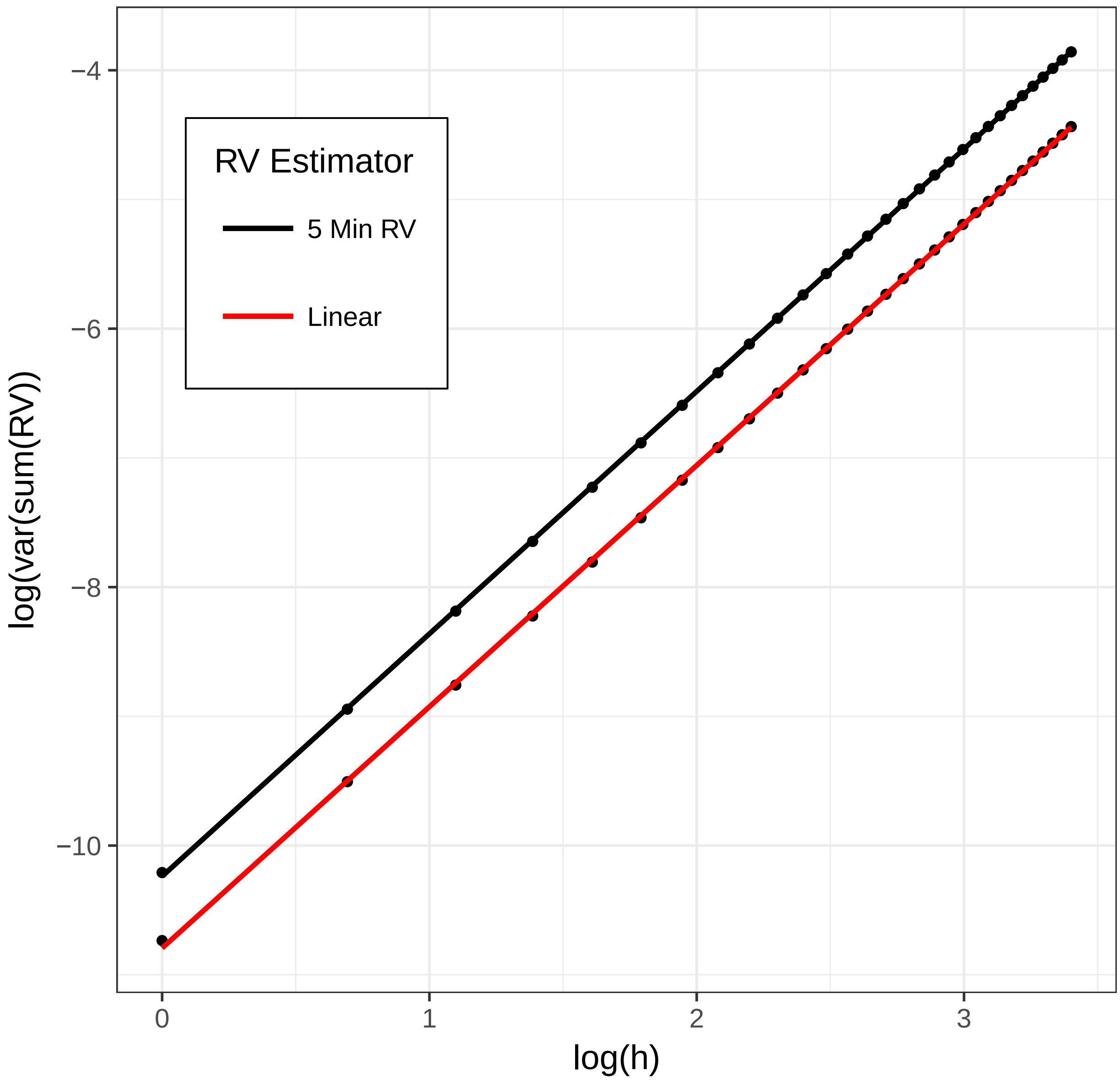

Figure 12: Linear RV fit to early series

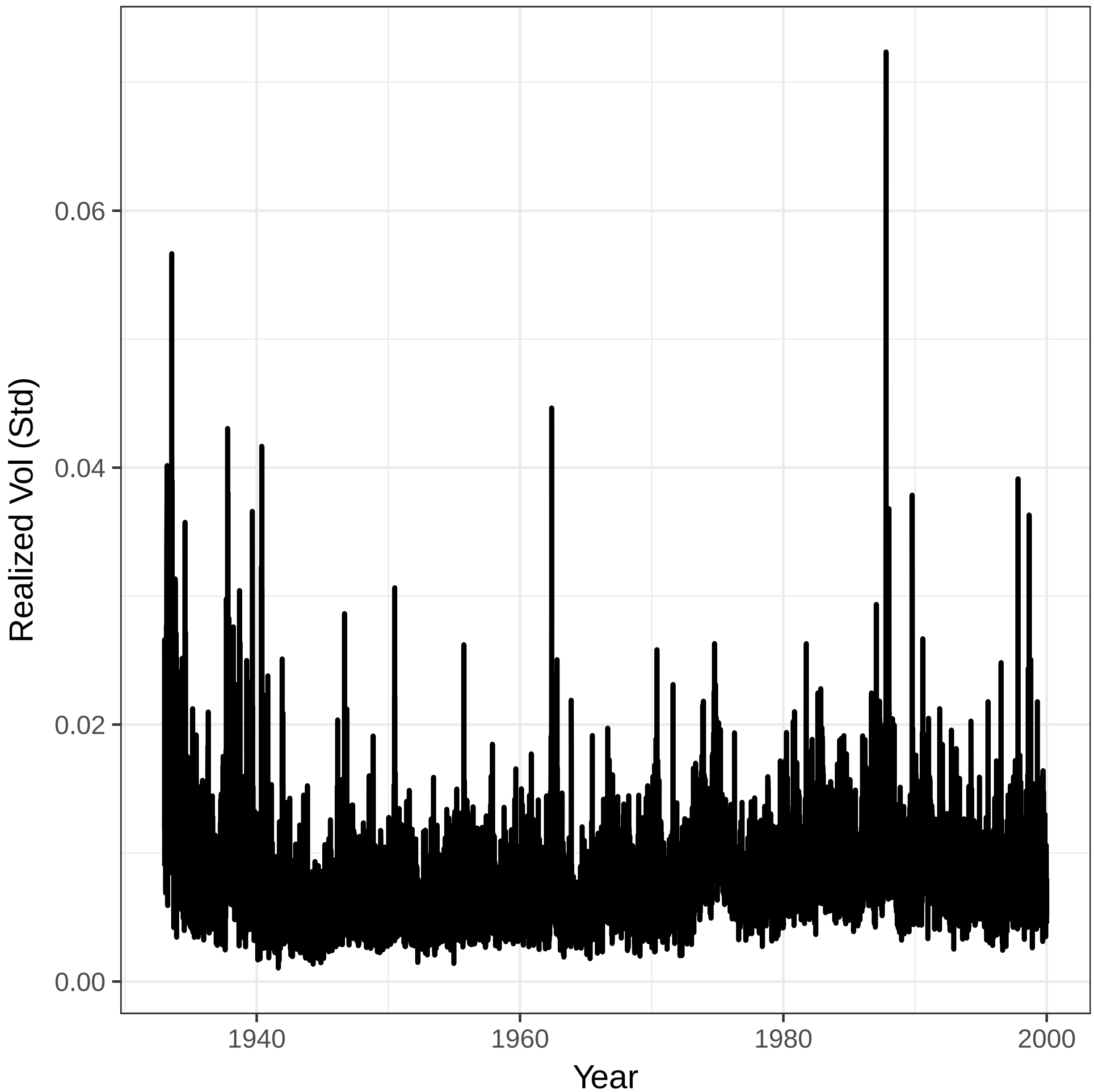

Figure 13: Log Volatility Densities: 1933-1999

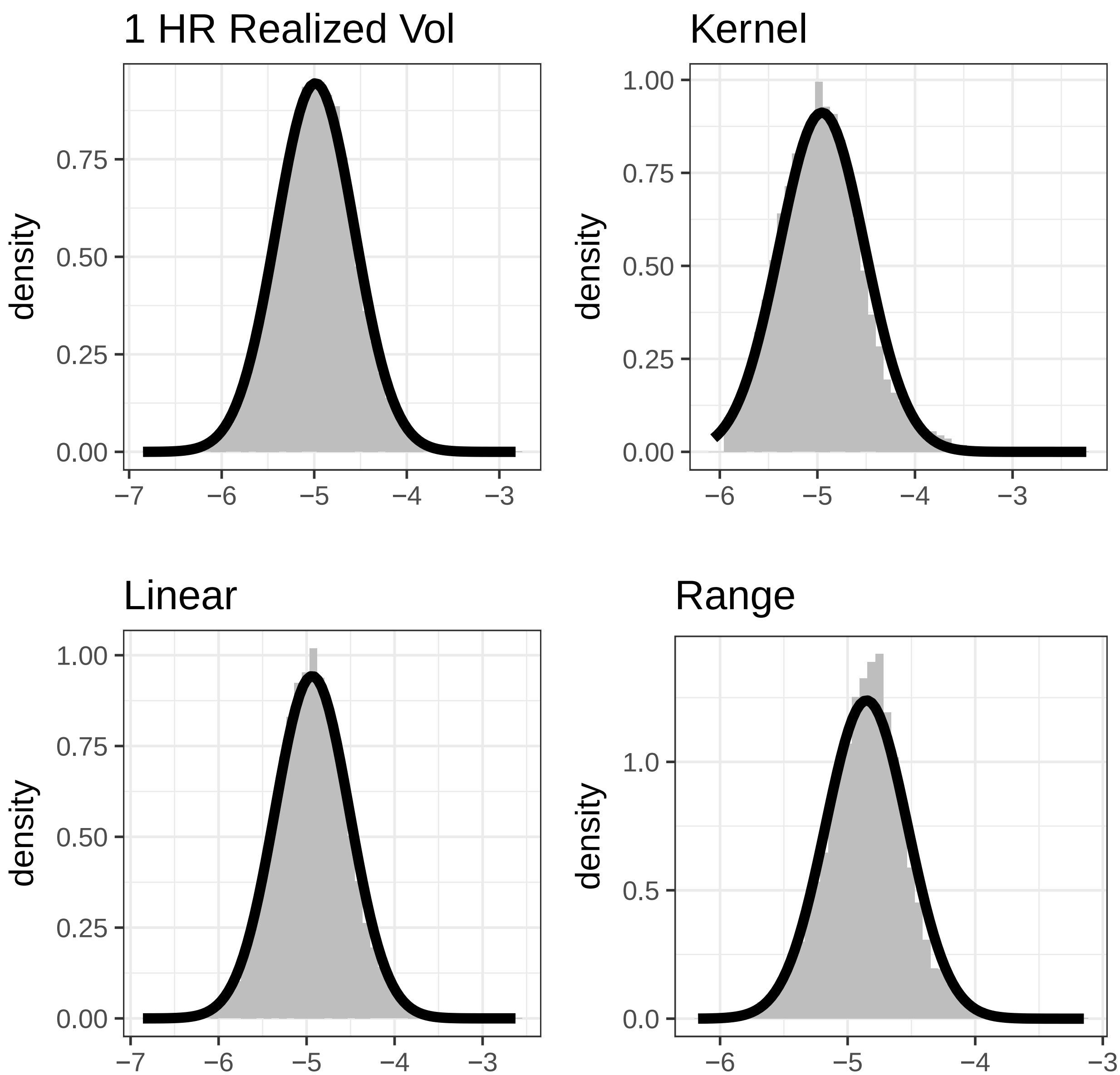

Figure 14: Standardized returns: Histograms (1933-1999)

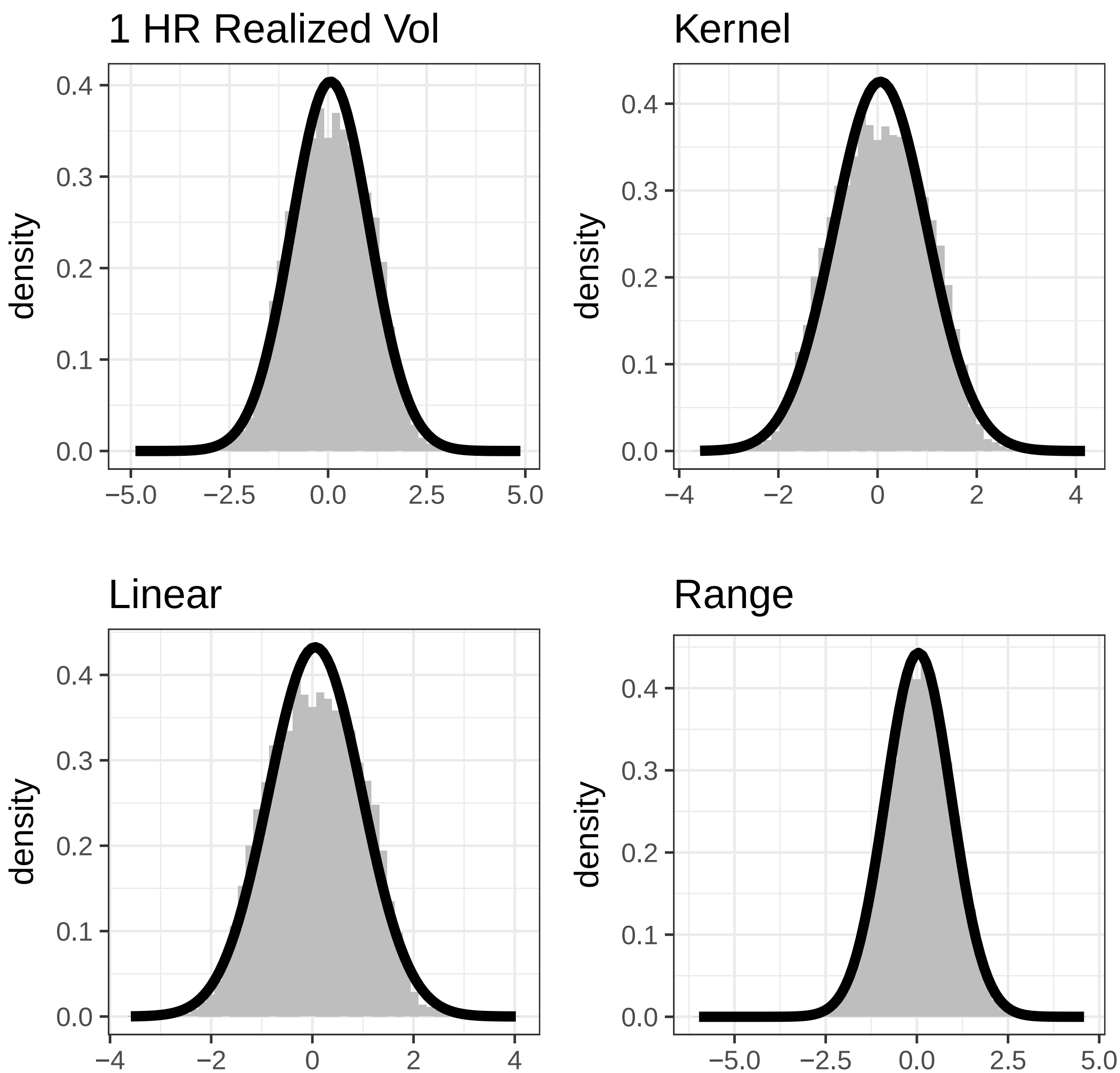

Figure 15: Standardized returns: QQ plots (1933-1999)

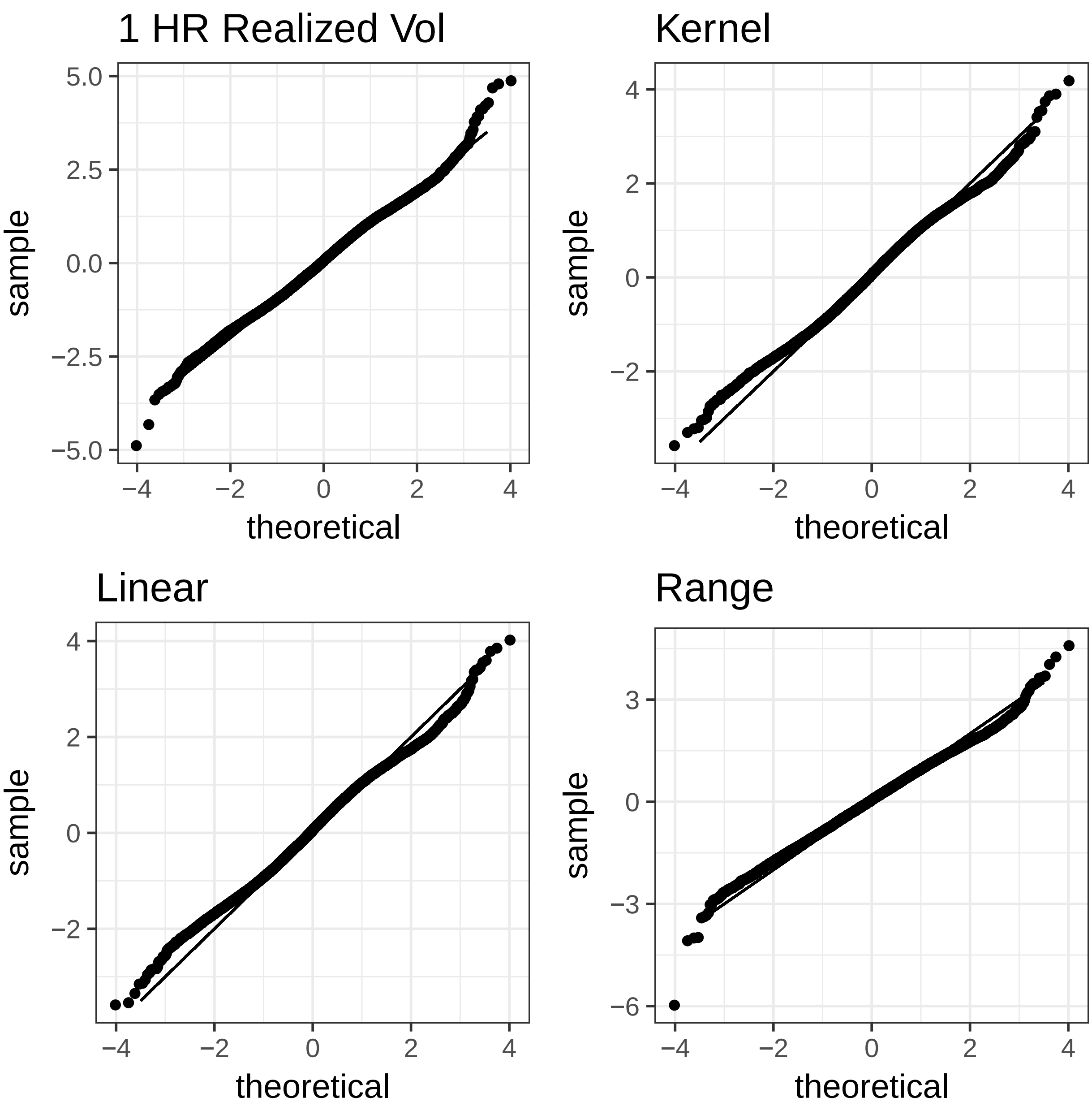

Figure 16: Absolute standardized returns autocorrelations (1933-1999)

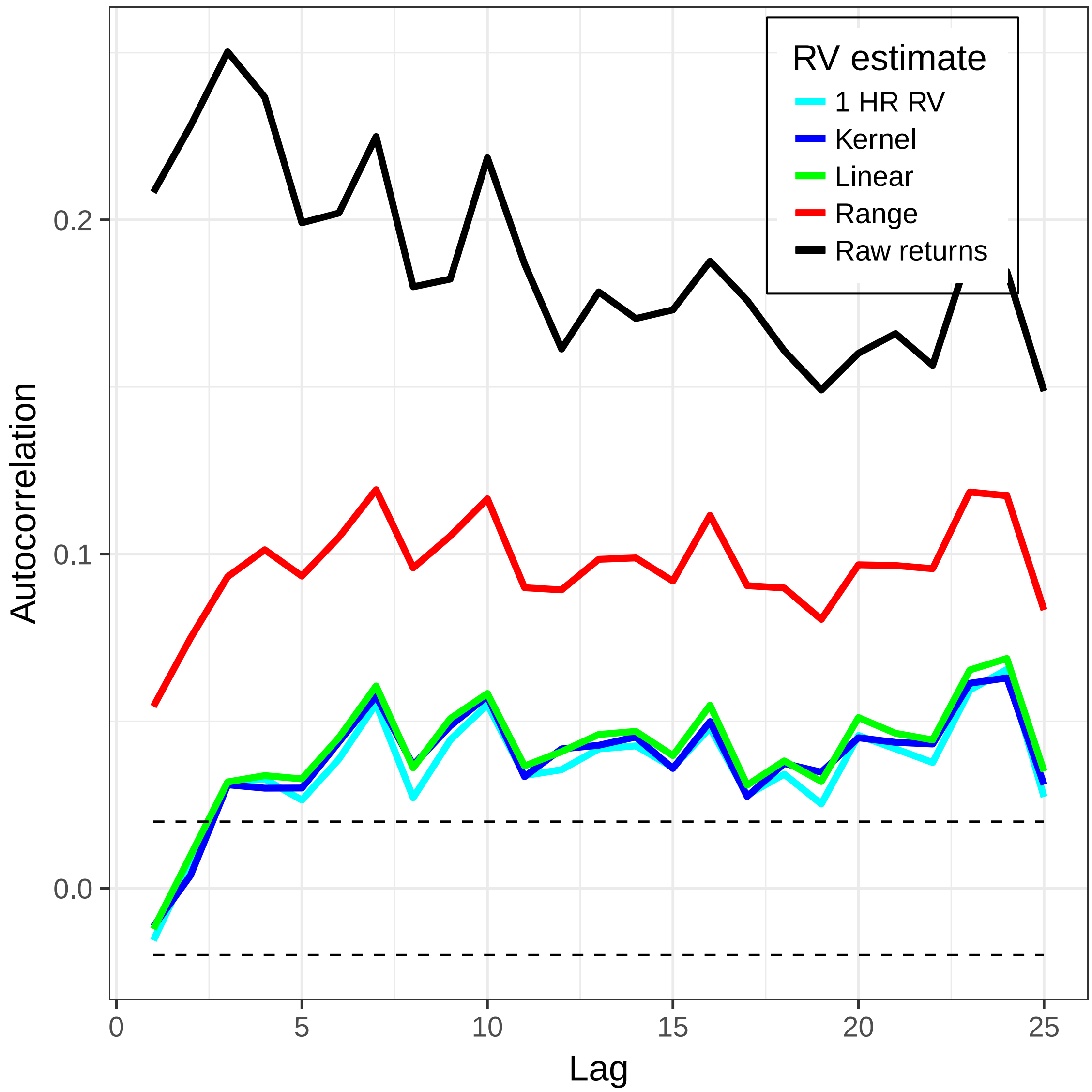

Figure 17: Standardized returns (1933-1999)

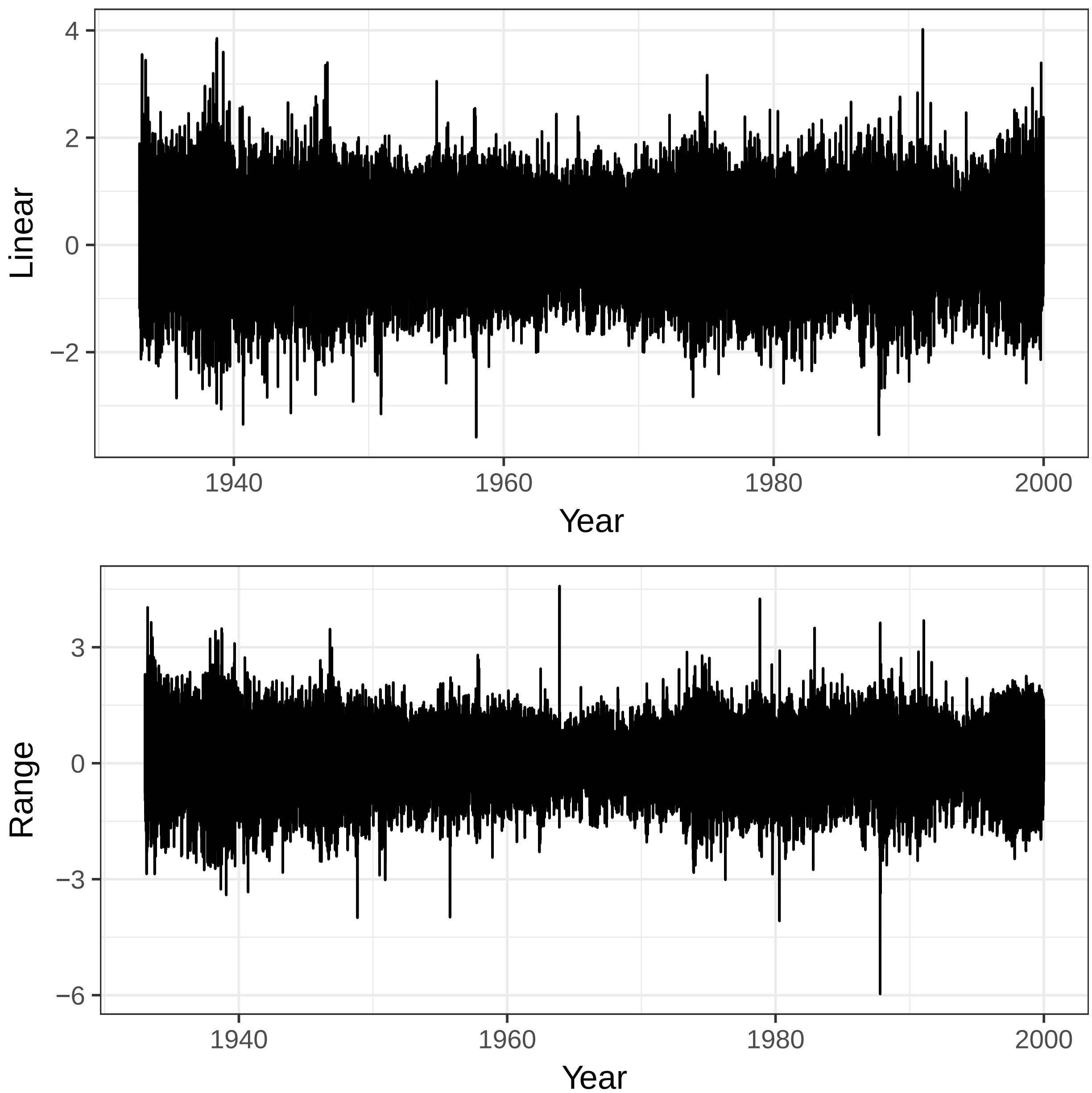

Figure 18: RV autocorrelation: 1933-1999

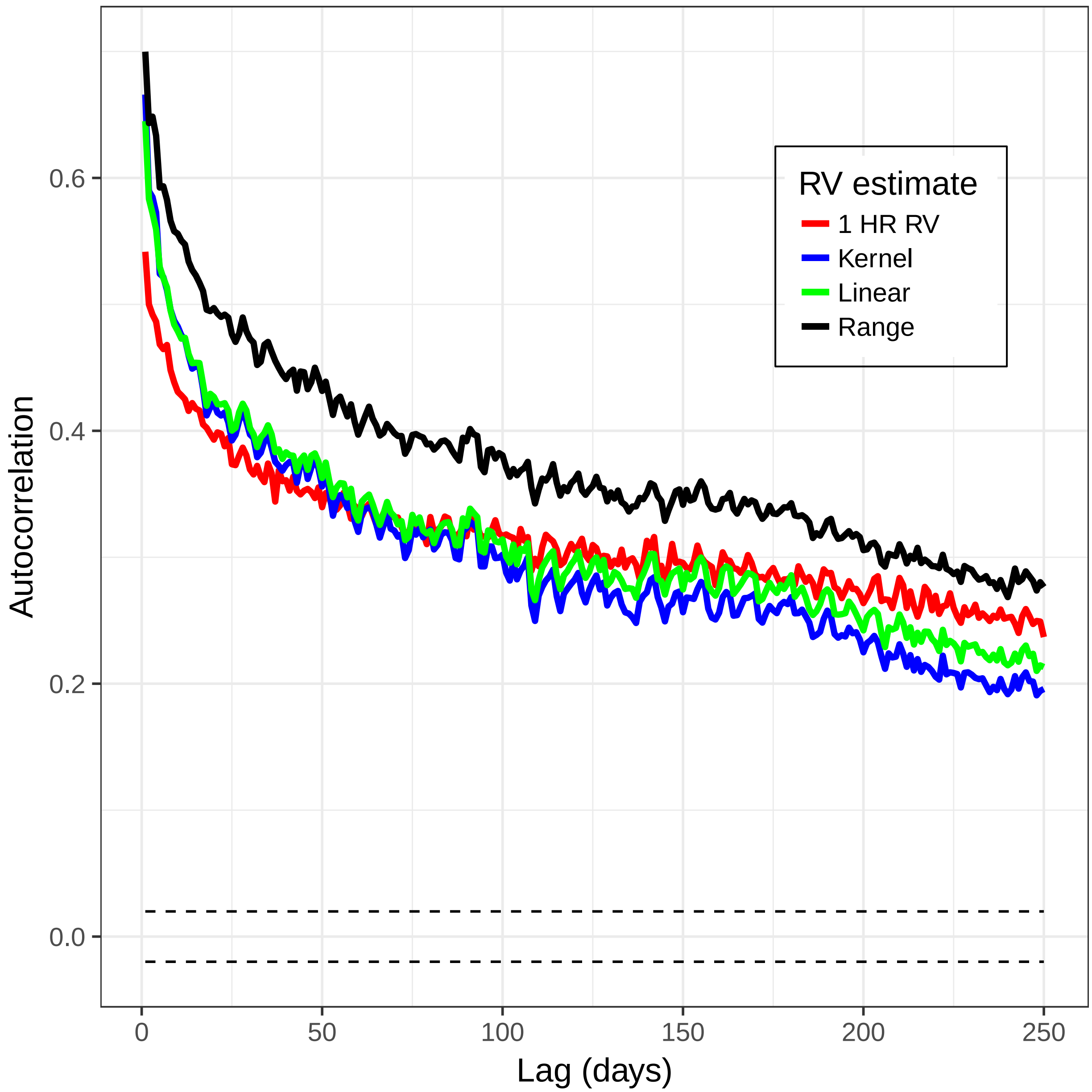

Figure 19: Linear RV autocorrelation: subsamples

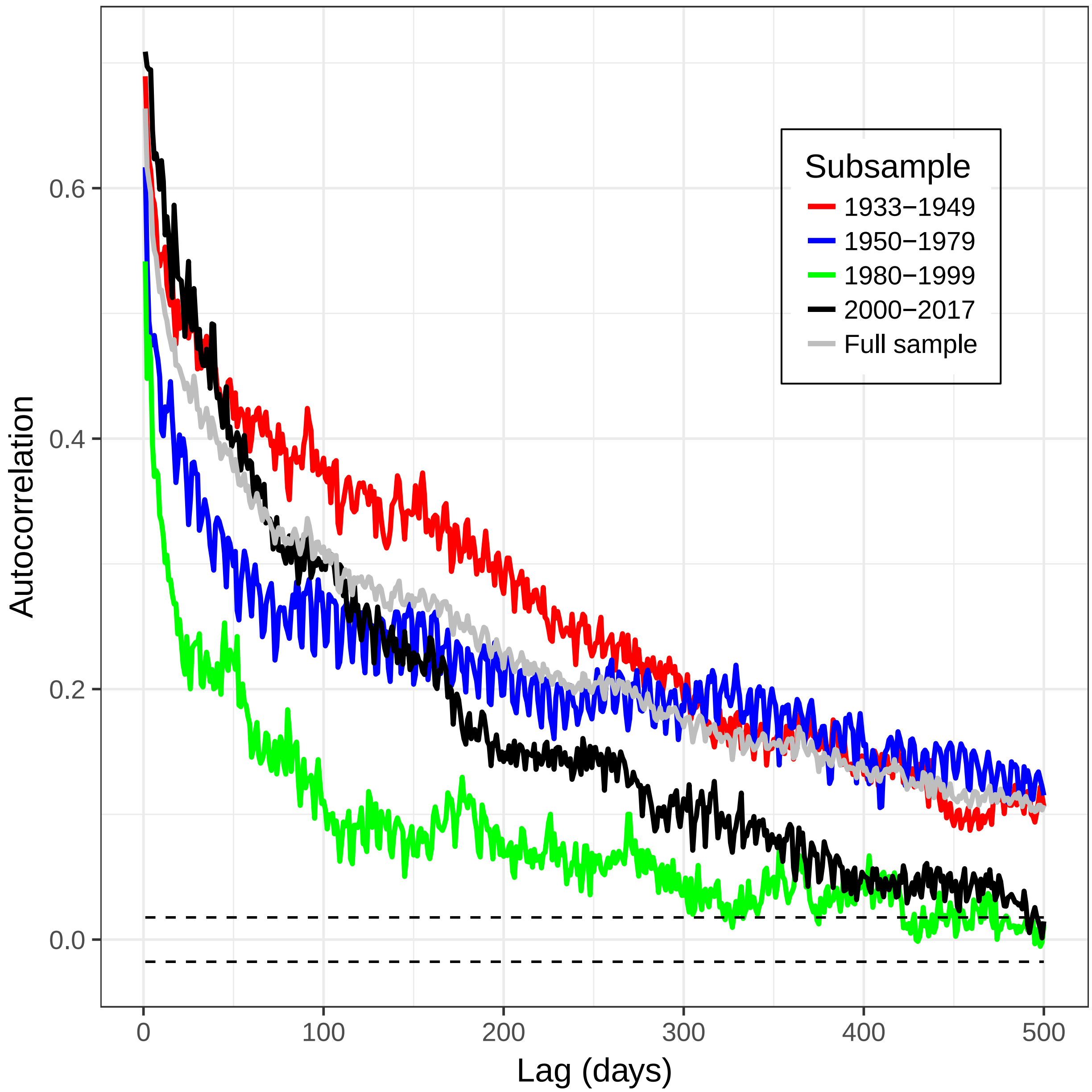

Figure 20: Volatility forecasts: 2000-2017

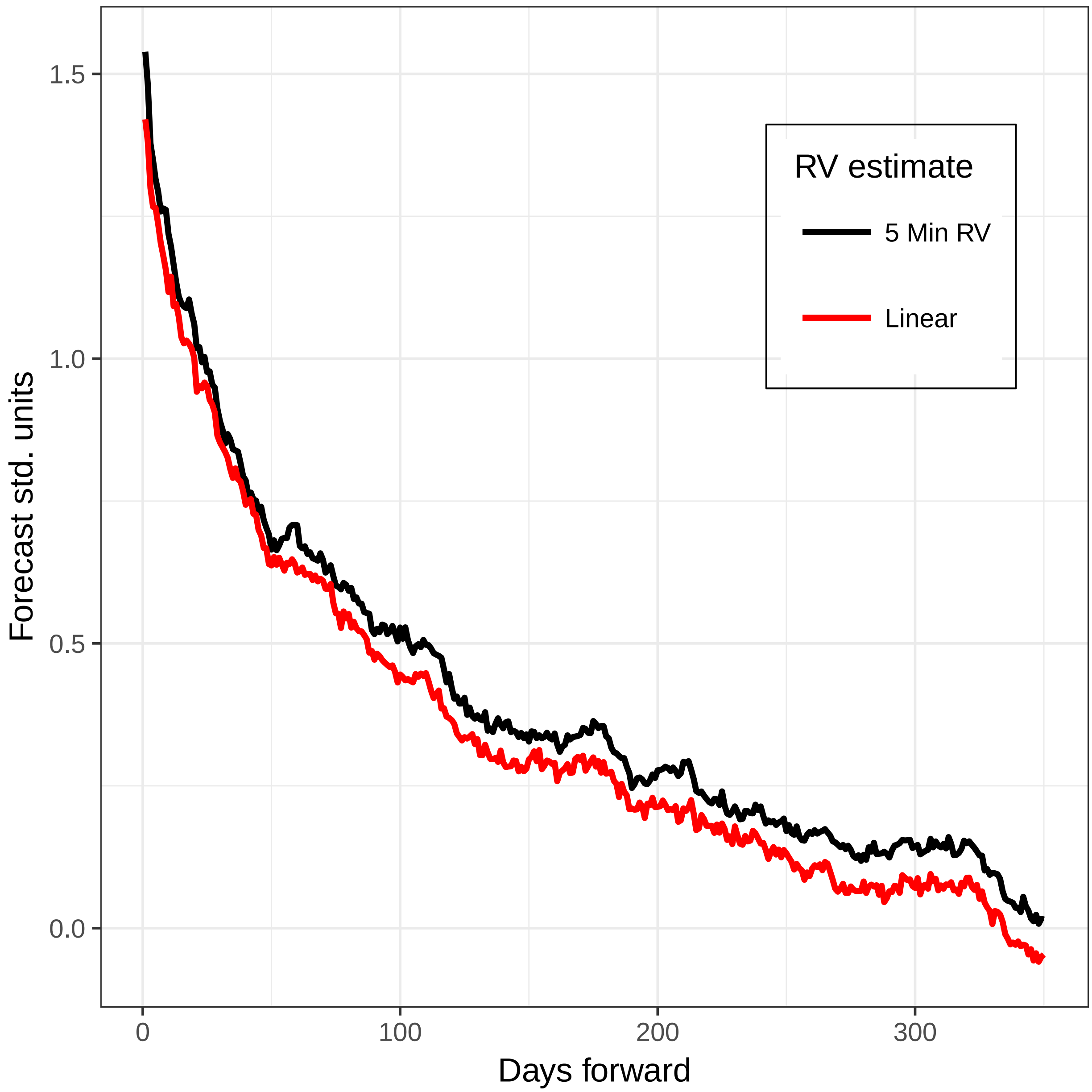

Figure 21: Volatility forecasts: Subsamples

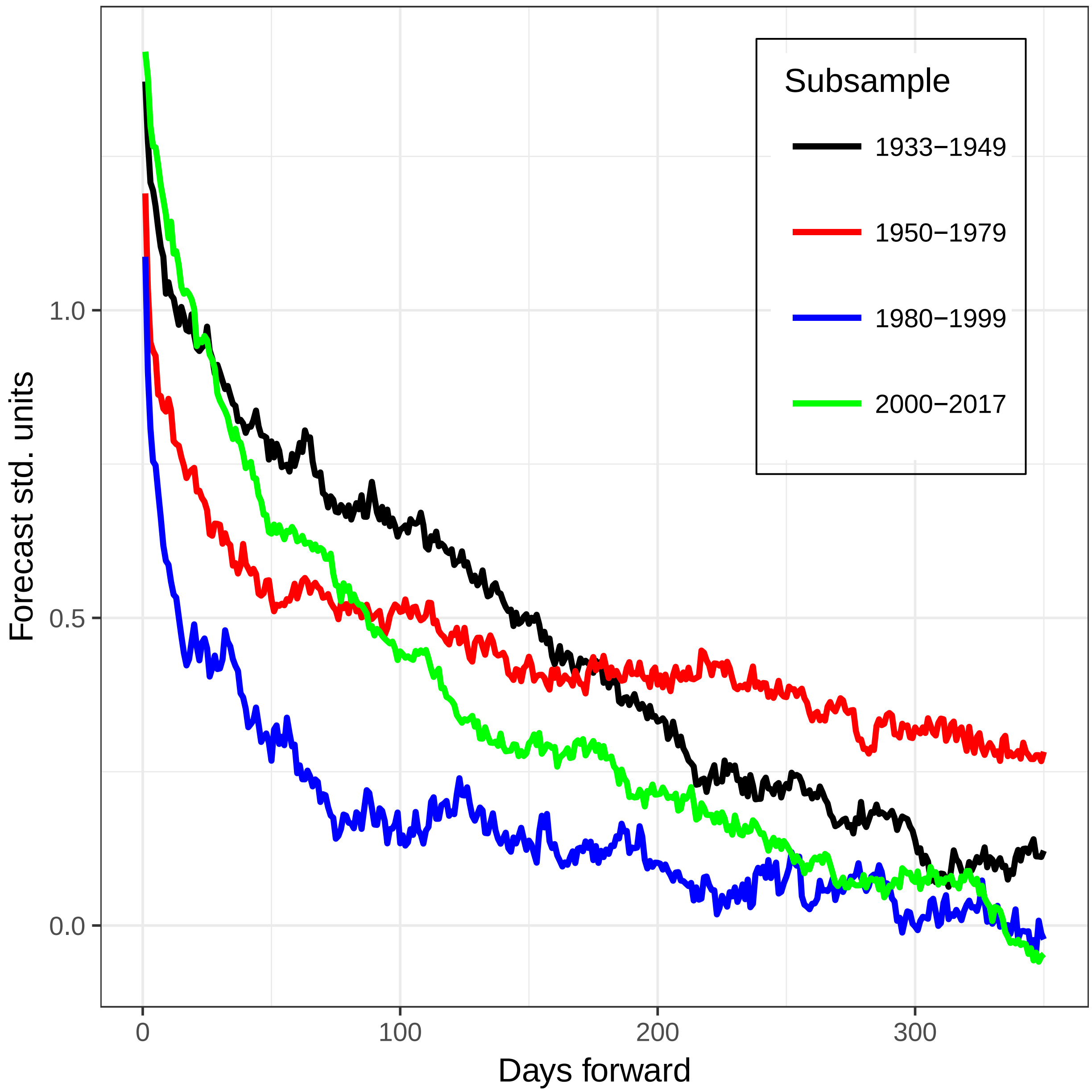

Figure 22: Volatility forecasts: Subsampling confidence bands

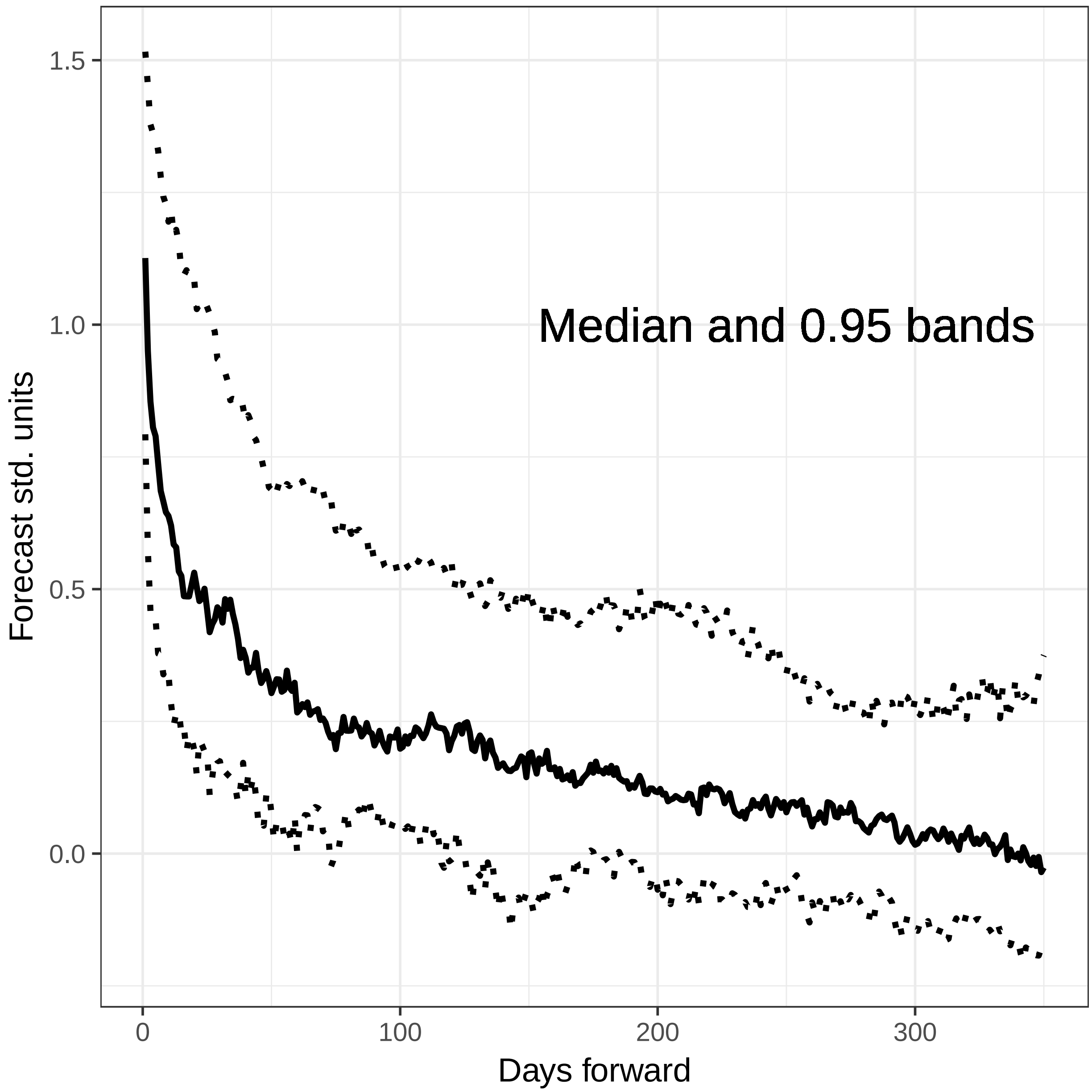

Figure 23: News impact: Subsamples