

Model-free vs. Model-based Volatility Prediction*

Dimitris N. Politis
Depts. of Mathematics and Economics,
University of California at San Diego,
La Jolla, CA 92093-0112, USA
email: dpolit@ucsd.edu

*This paper was presented as a keynote address at the “International Workshop on Recent Advances in Time Series Analysis”, Cyprus, June 2004, as well as the 2nd International Symposium “Advances in Financial Forecasting”, October 2005, Loutraki, Greece. Many thanks are due to the participants for helpful feedback, and in particular to Professors E. Paparoditis and D. Thomakos. The support of the National Science Foundation through grants DMS-01-04059 and SES-04-18136 is also gratefully acknowledged.

Abstract

The well-known ARCH/GARCH models for financial time series have been criticized of late for their poor performance in volatility prediction, i.e., prediction of squared returns.¹ Focusing on three representative data series, namely a foreign exchange series (Yen vs. Dollar), a stock index series (the S&P500 index), and a stock price series (IBM), the case is made that financial returns may not possess a finite fourth moment. Taking this into account, we show how and why ARCH/GARCH models—when properly applied and evaluated—actually *do* have nontrivial predictive validity for volatility. Furthermore, we show how a simple model-free variation on the ARCH theme can perform even better in that respect. The model-free approach is based on a novel normalizing and variance-stabilizing transformation (NoVaS, for short) that can be seen as an alternative to parametric modelling. Properties of this transformation are discussed, and practical algorithms for optimizing it are given.

Keywords: ARCH/GARCH models, forecasting, L_1 methods, volatility.

JEL classification codes: C3; C5.

Introduction

Consider data X_1, \dots, X_n arising as an observed stretch from a financial *returns* time series $\{X_t, t \in \mathbf{Z}\}$ such as the percentage returns (or, equivalently, the differences of the logarithms) of a stock price, stock index or foreign exchange rate; the returns may be daily, weekly, or calculated at different (discrete) intervals. The returns series $\{X_t\}$ will be assumed (strictly) stationary with mean zero which—from a practical point of view—implies that trends and other nonstationarities have been successfully removed.

Bachelier’s (1900) pioneering work suggested the Gaussian random walk model for (the logarithm of) stock market prices. Because of the aforementioned equivalence of percentage returns to differences in the logarithm of the price series, the implication of Bachelier’s thesis was that the returns series $\{X_t\}$ can be modelled as independent, identically distributed (i.i.d.) random variables with Gaussian $N(0, \sigma^2)$ distribution.

The assumption of Gaussianity was challenged in the 1960s when it was noticed that the distribution of returns seemed to have fatter tails than the

normal; see e.g. Fama (1965). The adoption of some non-normal, heavy-tailed distribution for the returns seemed—at the time—to be the solution. However, in the early paper of Mandelbrot (1963) the phenomenon of ‘volatility clustering’ was pointed out, i.e., the fact that days with high volatility are clustered together and the same is true for days with low volatility; this is effectively negating the assumption of independence of the returns in the implication that the absolute values (or squares) of the returns are positively correlated.

The popular ARCH (Auto-Regressive Conditional Heteroscedasticity) models of Engle (1982) were designed in order to capture the phenomenon of volatility clustering by postulating a particular structure of dependence for the time series of squared returns $\{X_t^2\}$. A typical ARCH(p) model is thus described by an equation of the type:

$$X_t = Z_t \sqrt{a + \sum_{i=1}^p a_i X_{t-i}^2} \quad (1)$$

where the series $\{Z_t\}$ is assumed to be i.i.d. $N(0, 1)$ and p is an integer indicating the order of the model. Note that under this ARCH(p) model, the best (in a Mean Squared Error sense) prediction of X_{n+1}^2 based—i.e., conditional—on the observed past $\mathcal{F}_n = \{X_t, 1 \leq t \leq n\}$ is given by

$$E(X_{n+1}^2 | \mathcal{F}_n) = a + \sum_{i=1}^p a_i X_{n+1-i}^2; \quad (2)$$

the quantity on the RHS of (2) is commonly referred to as the ‘volatility’ (although the same term is sometimes also used for its square root).

Volatility clustering as captured by model (1) does indeed imply a marginal distribution for the $\{X_t\}$ returns that has heavier tails than the normal. However, model (1) can account only partly for the degree of heavy tails empirically found in the distribution of returns, and the same is true for the Generalized ARCH (GARCH) models of Bollerslev (1986); see Bollerslev et al. (1992) or Shephard (1996) for a review. For example, the market crash of October 1987 is still an outlier 6-7 standard deviations away even after the best ARCH/GARCH model is employed; see Nelson (1991).

Consequently, researchers and practitioners have been resorting to ARCH models with heavy-tailed errors. A popular assumption for the distribution

of the $\{Z_t\}$ is the t -distribution with degrees of freedom empirically chosen to match the apparent degree of heavy tails in the residuals; see Shephard (1996) and the references therein.

Nevertheless, this situation is not very satisfactory since the choice of a t -distribution seems quite arbitrary. In a certain sense, it seems that we have come full-circle back to the 60s in trying to model the excess kurtosis by an arbitrarily chosen heavy-tailed distribution. Perhaps the real issue is that a simple and neat parametric model such as (1) could not be expected to perfectly capture the behavior of a complicated real-world phenomenon such as the evolution of financial returns that—almost by definition of market ‘efficiency’—ranks at the top in terms of difficulty of modelling/prediction.

As a more realistic alternative, one may resort to an exploratory, model-free approach in trying to understand this type of data; such an approach is outlined in the paper at hand. In the next section, a normalizing and variance-stabilizing transformation (NoVaS, for short) for financial returns series is defined, and its properties are analyzed; a preliminary announcement of the NoVaS transformation was given in Politis (2003a,b). Section 2 is devoted to the interesting (and quite challenging) problem of volatility prediction, while Section 3 contains some conclusions.

Throughout the paper we focus on three representative datasets of daily returns taken from a foreign exchange rate, a stock price, and a stock index; a description of our main datasets is as follows.

- **Example 1: Foreign exchange rate.** Daily returns from the Yen vs. Dollar exchange rate from January 1, 1988 to August 1, 2002; the data were downloaded from Datastream. A plot of the returns is shown in Figure 1a; the sample size is 3600 (weekends and holidays are excluded).
- **Example 2: Stock index.** Daily returns of the S&P500 stock index from October 1, 1983 to August 30, 1991; the data are available as part of the `garch` module in `Splus`. A plot of the returns is shown in Figure 1b; the sample size is 2000.
- **Example 3: Stock price.** Daily returns of the IBM stock price from February 1, 1984 to December 31, 1991; the data are again available as part of the `garch` module in `Splus`. A plot of the returns is shown in Figure 1c; the sample size is 2000.

The phenomenon of volatility clustering is quite apparent in the three returns series of Figure 1. Note, in particular, the extreme volatility and outlying values around the mid-point of Figure 1(b) and slightly before the mid-point of Figure 1(c); those points of time correspond to the aforementioned market crash of October 1987.

Returning to the ARCH model (1), it should be stressed that this is not just a model for the conditional variance; the ARCH model is a model for the *whole* data generating process (DGP) of the data series X_t . In this connection, observe that the one-step-ahead ARCH-based predictor for X_{n+1} given $\mathcal{F}_n = \{X_t, t \leq n\}$ is trivial, i.e., zero, essentially due to the random sign of Z_t in eq. (1).

Nevertheless, the litmus test for a model is its predictive ability. Since the ARCH can not predict the signed return X_{n+1} , it should at least have some predictive ability for the squared returns. Recall also that if $\{X_t\}$ follows the ARCH(p) model (1), then $\{X_t^2\}$ follows an AR(p) model—see e.g. Gouriéroux (1997); this is further corroboration of the fact that an ARCH model *should* have some predictive power for the squared returns.

It is the widely reported failure of ARCH (*and* GARCH) models to predict squared returns that served as the motivation for this paper. Reassuringly, one of the findings of the paper is that indeed ARCH/GARCH models do have predictive ability for the squared returns when an appropriate performance measure (such as L_1 loss) is used, and furthermore when the optimal predictor for the chosen loss function is employed. More details are given in Section 2.

The literature on volatility prediction is already quite large and appears to be continuously expanding. The recent articles by Poon and Granger (2003), and Andersen, Bollerslev, Christoffersen and Diebold (2006) provide comprehensive reviews of the subject. We further mention here some papers that are related to the problem at hand: Barndorff-Nielsen, Nielsen, Shephard and Ysusi (1996) for an early treatment of forecasting volatility; Meddahi (2001) for an eigenfunction volatility modeling approach; Hansen, Lunde and Nason (2003) on selecting volatility models; Andersen, Bollerslev and Meddahi (2004) on analytic evaluation of volatility forecasts; Hansen and Lunde (2005, 2006) for comparing forecasts of volatility models against the standard GARCH(1,1) model and for consistent ranking of volatility models; Koopman, Jungbacker and Hol (2005) and Patton (2005) for volatility forecast evaluation; and Ghysels, Santa Clara and Valkanov (2006) for predicting volatility using data sampled at different frequencies.

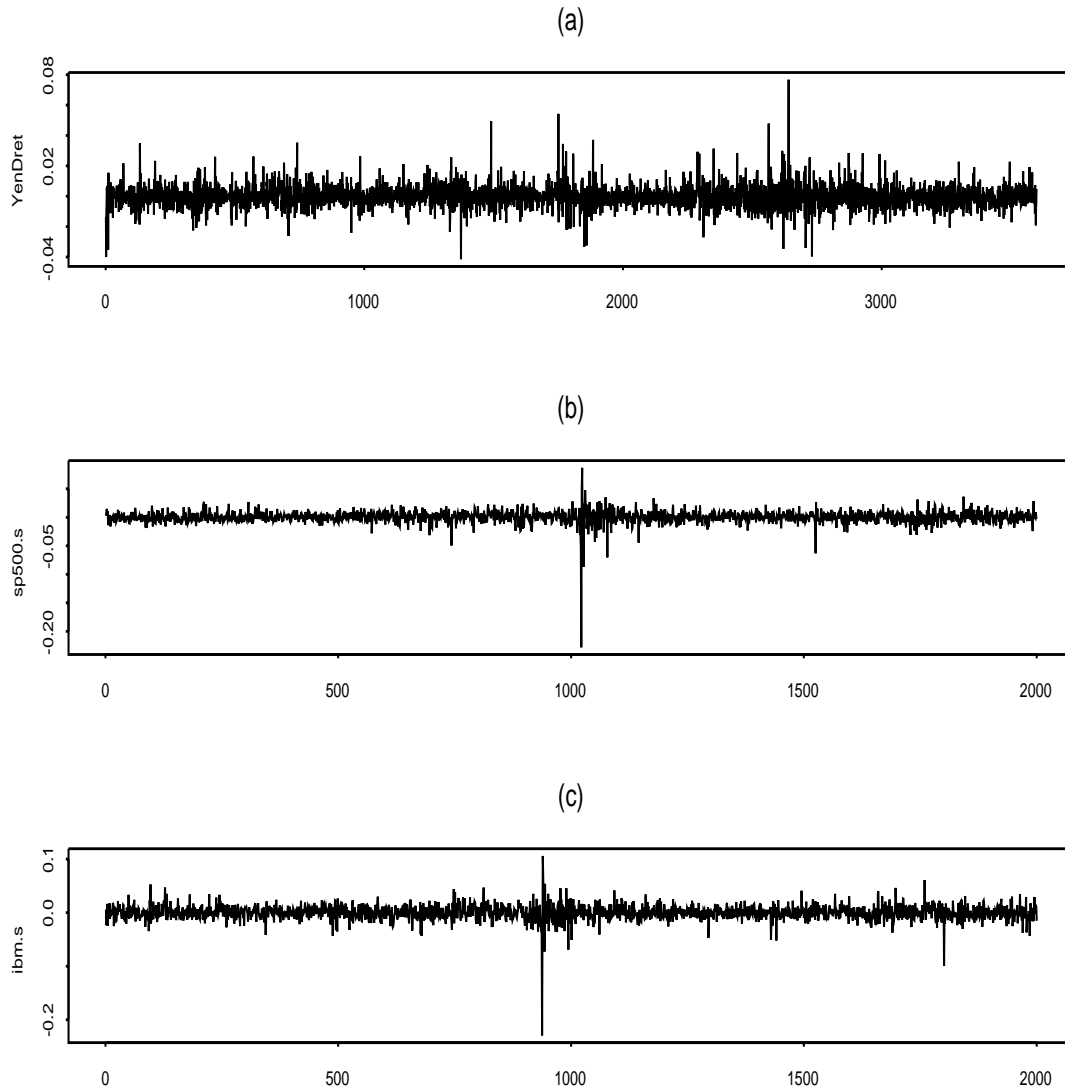


Figure 1: (a) Plot of the daily Yen/Dollar returns from December 31, 1987 up to August 1, 2002; (b) Plot of the daily S&P500 stock index returns from October 1, 1983 to August 30, 1991; (c) Plot of the daily returns of the IBM stock price from February 1, 1984 to December 31, 1991.

1 Normalization and variance-stabilization

1.1 Definition of the NoVaS transformation

Observe that, under the ARCH model (1), the quantity

$$\frac{X_t}{\sqrt{a + \sum_{i=1}^p a_i X_{t-i}^2}} \quad (3)$$

is thought of as perfectly normalized and variance-stabilized as it is assumed to be i.i.d. $N(0, 1)$. From an applied statistics point of view, the above ratio can be interpreted as an attempt to ‘studentize’ the return X_t by dividing with a (time-localized) measure of the standard deviation of X_t .

Nevertheless, there seems to be no reason—other than coming up with a neat model—to exclude the value of X_t from an empirical, causal² estimate of the standard deviation of X_t . Hence, we may define the new ‘studentized’ quantity

$$W_{t,a} := \frac{X_t}{\sqrt{\alpha s_{t-1}^2 + a_0 X_t^2 + \sum_{i=1}^p a_i X_{t-i}^2}} \quad \text{for } t = p+1, p+2, \dots, n; \quad (4)$$

in the above, s_{t-1}^2 is an estimator of $\sigma_X^2 = \text{Var}(X_1)$ based on the data up to (but not including³) time t ; under the zero mean assumption for X_1 , the natural estimator is $s_{t-1}^2 = (t-1)^{-1} \sum_{k=1}^{t-1} X_k^2$.

Equation (4) describes our proposed normalizing and variance-stabilizing transformation⁴ (NoVaS, for short) under which the data series $\{X_t\}$ is mapped to the new series $\{W_{t,a}\}$. The order $p(\geq 0)$ and the vector of non-negative parameters $(\alpha, a_0, \dots, a_p)$ are chosen by the practitioner with the twin goals of normalization/variance-stabilization in mind that will be made more precise shortly.

The NoVaS equation (4) can be re-arranged to yield:

$$X_t = W_{t,a} \sqrt{\alpha s_{t-1}^2 + a_0 X_t^2 + \sum_{i=1}^p a_i X_{t-i}^2}. \quad (5)$$

Note that the only real difference between the NoVaS eq. (5) and the ARCH eq. (1) is the presence of the term X_t^2 paired with the coefficient a_0 . Replacing the term a in eq. (1) by the term αs_{t-1}^2 in (5) is only natural since the former

has—by necessity—units of variance; in other words, the term a in eq. (1) is not scale invariant, whereas the term α in (5) is.

Equation (5) is very useful but should not be interpreted as a “model” for the $\{X_t\}$ series; rather, the focus should remain on equation (4) and the effort to render the transformed series $\{W_{t,a}, t = p + 1, p + 2, \dots\}$ close—in some sense to be described shortly—to behaving like the standard normal ideal.

A further note of caution on viewing eq. (5) as a “model” comes from the observation that *exact* normality is not feasible for the series $\{W_{t,a}\}$ as the latter comprises of bounded random variables; to see this, note that

$$\frac{1}{W_{t,a}^2} = \frac{\alpha s_{t-1}^2 + a_0 X_t^2 + \sum_{i=1}^p a_i X_{t-i}^2}{X_t^2} \geq a_0$$

if all the parameters are nonnegative. Therefore,

$$|W_{t,a}| \leq 1/\sqrt{a_0} \tag{6}$$

almost surely, assuming of course that $a_0 \neq 0$. However, with a_0 chosen small enough, the boundedness of the $\{W_{t,a}\}$ series is effectively (and practically) not noticeable.

1.2 Choosing the parameters of NoVaS

In choosing the order p (≥ 0) and the parameters α, a_0, \dots, a_p the twin goals of normalization and variance–stabilization of the transformed series $\{W_{t,a}\}$ are first taken into account. Secondly, the NoVaS parameters may be further optimized with a specific criterion in mind, e.g., optimal volatility prediction; this approach is expanded upon in Section 2. We now focus on the primary goals of normalization and variance–stabilization.

The target of variance–stabilization is easier and—given the assumed structure of the return series—amounts to constructing a local estimator of scale for studentization purposes; for this reason we require

$$\alpha \geq 0, \quad a_i \geq 0 \quad \text{for all } i \geq 0, \quad \text{and} \quad \alpha + \sum_{i=0}^p a_i = 1. \tag{7}$$

Equation (7) has the interesting implication that the $\{W_{t,a}\}$ series can be assumed to have an (unconditional) variance that is (approximately) unity.

Nevertheless, note that p and α, a_0, \dots, a_p must be carefully chosen to achieve a degree of conditional homoscedasticity as well; to do this, one must necessarily take p small enough—as well as α small enough or even equal to zero—so that a local (as opposed to global) estimator of scale is obtained. An additional intuitive—but not obligatory—constraint may involve monotonicity:

$$a_i \geq a_j \quad \text{if} \quad 1 \leq i < j \leq p. \quad (8)$$

It is practically advisable that a simple structure for the a_i coefficients is employed satisfying (7) and (8). The simplest such example is to let $\alpha = 0$ and $a_i = 1/(p + 1)$ for all $0 \leq i \leq p$; this specification will be called the ‘*simple*’ NoVaS transformation, and involves only one parameter, namely the order p , to be chosen by the practitioner. Another example is given by the *exponential* decay NoVaS where $\alpha = 0$ and $a_i = c'e^{-ci}$ for all $0 \leq i \leq p$. The exponential scheme involves choosing two parameters: p and $c > 0$ since c' is determined by (7); nevertheless, the parameter p is now of secondary importance—see section 1.4. The simple and exponential NoVaS schemes are most intuitive as they correspond to the two popular time series methods of obtaining a ‘local’ average, namely moving average (of the last $p + 1$ values) and ‘exponential smoothing’; see e.g. Hamilton (1994).

Subject to the variance stabilization condition (7)—together with (8) if desirable—one then proceeds to choose (the parameters needed to identify) p and $\alpha, a_0, a_1, \dots, a_p$ with the optimization goal of making the $\{W_{t,a}\}$ transformed series as close to normal as possible. To quantify this target it is suggested that one matches the empirical kurtosis (and/or possibly some higher order even moments) of $W_{t,a}$ to those of a standard normal random variable. In order to render joint distributions of the $\{W_{t,a}\}$ series more normal, one may also apply the previous moment matching idea to a few specific linear combinations of $W_{t,a}$ random variables; more details are given in the next subsection.

However, in view of the bound (6), one must be careful to ensure that the $\{W_{t,a}\}$ random variables have a large enough range such that the boundedness is not seen as spoiling the normality. Thus, we also require

$$1/\sqrt{a_0} \geq C \quad \text{i.e.,} \quad a_0 \leq 1/C^2 \quad (9)$$

for some appropriate C of the practitioner’s choice. Recalling that 99.7% of the mass of the $N(0, 1)$ distribution is found in the range ± 3 , the simple

choice $C = 3$ can be suggested; this choice seems to work reasonably well—at least for the usual samples sizes. Alternatively, one may let C depend on the sample size n . Taking into account that the maximum of n i.i.d. $N(0, 1)$ random variables is of the order of $\sqrt{2 \ln n}$, one may let C be equal (or proportional) to $\sqrt{2 \ln n}$. Such a choice, however, may lead to an impractically large value of C especially in connection with large datasets; see Remark 1.2 for an example.

1.3 Simple NoVaS algorithm

We now give specific algorithms for optimizing the NoVaS transformation in the two previously mentioned examples, simple and exponential NoVaS. First note that it is a matter of common practice to assume that the distribution of financial returns is symmetric (at least to a first approximation); therefore, the skewness of financial returns is often ignored. In contrast, the kurtosis is typically very large, indicating a heavy tailed distribution. The above claims, i.e., approximate symmetry and heavy tails, are confirmed by Figure 2 where histograms and Q-Q plots for our three returns series are presented.

Let $KURT_n(Y)$ denote the empirical kurtosis of data $\{Y_t, t = 1, \dots, n\}$, i.e.,

$$KURT_n(Y) = \frac{n^{-1} \sum_{t=1}^n (Y_t - \bar{Y})^4}{(n^{-1} \sum_{t=1}^n (Y_t - \bar{Y})^2)^2}$$

where $\bar{Y} = n^{-1} \sum_{t=1}^n Y_t$ is the sample mean. For our three datasets, Yen/Dollar, S&P500 and IBM, the empirical kurtosis was 10.1, 94.0 and 38.3 respectively. Although even moments of order higher than four may also be used to measure deviation from normality, in what follows we focus on the kurtosis for concreteness.

Note that the only free parameter in simple NoVaS is the order p ; therefore, the simple NoVaS transformation will be denoted by $W_{t,p}^S$.

ALGORITHM FOR SIMPLE NOVAS:

- Let $\alpha = 0$ and $a_i = 1/(p + 1)$ for all $0 \leq i \leq p$.
- Pick p such that $|KURT_n(W_{t,p}^S) - 3|$ is minimized.

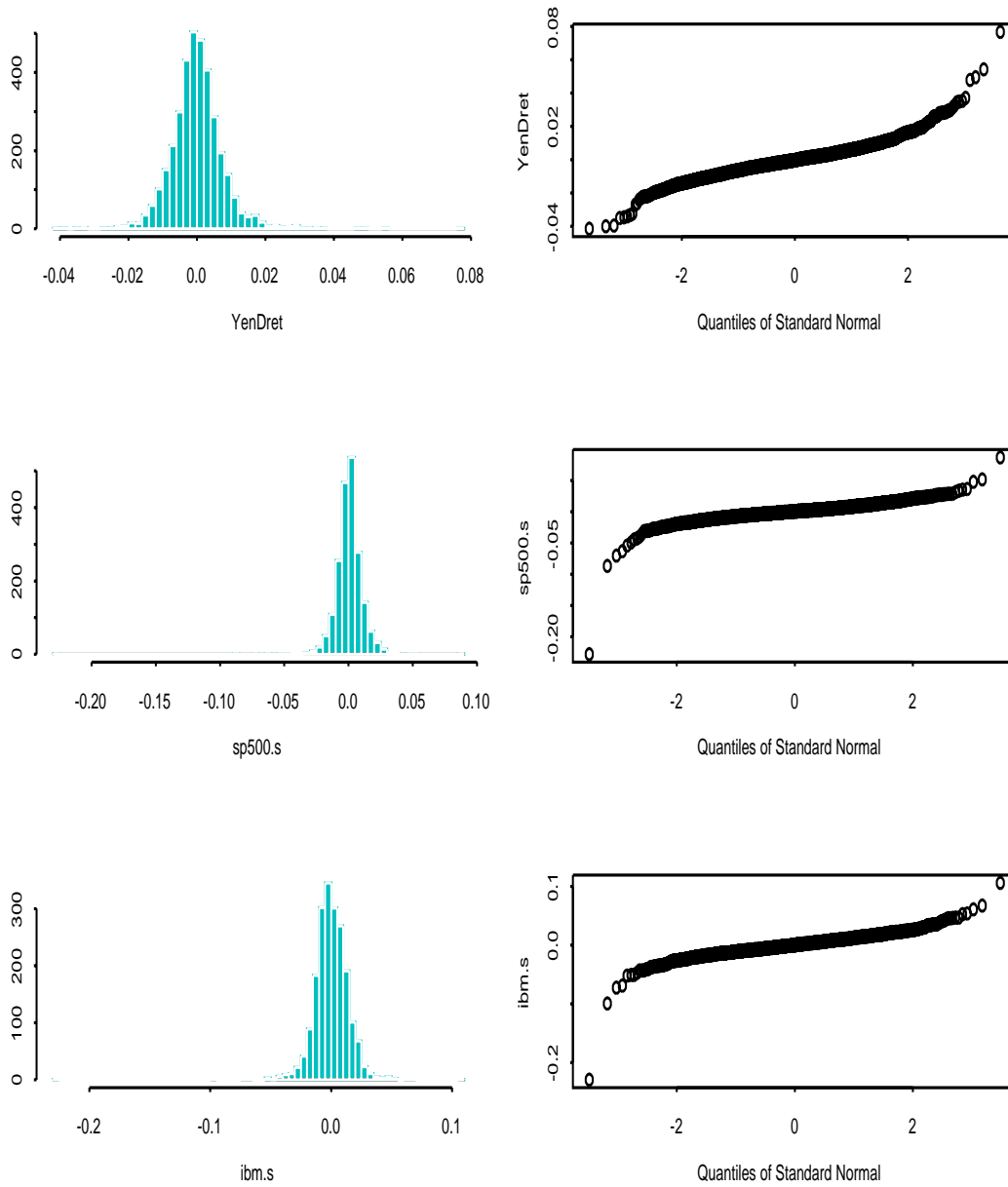


Figure 2: Histograms and Q-Q plots for the three returns series of Figure 1.

The last step of the above algorithm was described as an optimization problem for mathematical concreteness. Nevertheless, it could be better understood as a moment matching, i.e.,

- Pick p such that $KURT_n(W_{t,p}^S) \simeq 3$,

where of course the value 3 for kurtosis corresponds to the Gaussian distribution.

Remark 1.1 [Feasibility of normalization by moment matching.] To see that the moment matching goal is a feasible one, note first that for $p = 0$ we have $a_0 = 1$, $W_{t,0}^S = \text{sign}(X_t)$, and $KURT_n(W_{t,0}^S) = 1$. On the other hand, it is to be expected that for large p , $KURT_n(W_{t,p}^S)$ will be bigger than 3. As a matter of fact, the law of large numbers implies that for increasing values of p , $KURT_n(W_{t,p}^S)$ will tend to the ‘true’ kurtosis of the random variable X_1 which is understood to be quite large (and may even be infinite—see the discussion in Section 2.1). Therefore, viewing $KURT_n(W_{t,p}^S)$ as a (smooth) function of p , the intermediate value theorem would suggest that, for an intermediate value of p , the level 3 can always be (approximately) attained; this is actually what happens in practice.

Thus, to actually carry out the search for the optimal p in the Simple NoVaS Algorithm, one sequentially computes $KURT_n(W_{t,p}^S)$ for $p = 1, 2, \dots$, stopping when $KURT_n(W_{t,p}^S)$ first hits or just passes the value 3. Interestingly, $KURT_n(W_{t,p}^S)$ is typically an increasing function of p which makes this scheme very intuitive; see Figure 3(a).

The above simple algorithm seems to work remarkably well. A caveat, however, is that the range condition (9) might not be satisfied. If this is the case, the following ‘range-adjustment’ step can be added to algorithm.

- If p (and a_0) as found above are such that (9) is not satisfied, then increase p accordingly; in other words, redefine p to be the smallest integer such that $1/(p+1) \leq 1/C^2$, and let $a_i = 1/(p+1)$ for all $0 \leq i \leq p$.

It goes without saying that this range-adjustment should be used with restraint, that is, the choice of C in (9) should be reasonably small, as it

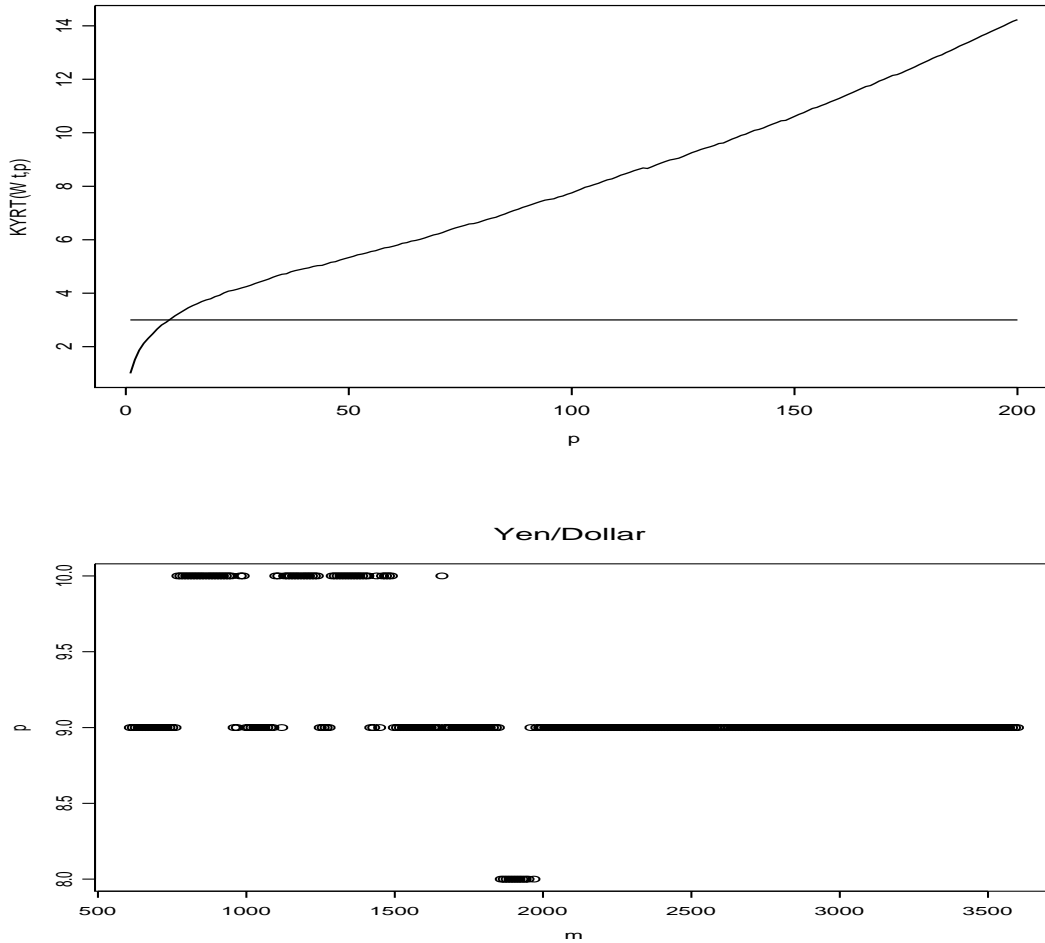


Figure 3: Illustration of the simple NoVaS algorithm for the Yen/Dollar dataset. (a) Plot of $KURT_n(W_{t,p}^S)$ as a function of p with the solid line indicating the Gaussian kurtosis of 3; here the full sample size $n = 3600$ is used. (b) Plot of the optimal values of p in Simple NoVaS corresponding to a subseries of the type $x[1 : m]$; here x is the Yen/Dollar returns dataset.

effectively over-rides the data-dependent character of choosing p . The conservative choice of letting $C = 3$ seem to work well in practice; see Remark 1.2 for an example.

In Figure 3 (a), an illustration of the simple NoVaS algorithm is given for the Yen/Dollar dataset. Figure 3 (a) shows a plot of $KURT_n(W_{t,p}^S)$ as a function of p ; the monotonic increase of $KURT_n(W_{t,p}^S)$ is apparent, rendering the NoVaS algorithm easy to implement. Notably, $KURT_n(W_{t,p}^S)$ is closest to 3 for $p = 9$; actually, $KURT_n(W_{t,9}^S) = 3.03$. Interestingly, the data-dependent choice $p = 9$ seems very stable; estimating p over different subsamples of the Yen/Dollar dataset typically yielded the value 9 ± 1 even for subsamples with length one tenth of $n = 3600$; see Figure 3 (b).

The optimal simple NoVaS transformed series $\{W_{t,9}^S\}$ for the Yen/Dollar dataset is plotted in Figure 4(a). Although $\{W_{t,9}^S\}$ is related in a simple way to the original data of Figure 1(a), the regions of “volatility clustering” corresponding to the $\{X_t\}$ series are hardly (if at all) discernible in the plot of the NoVaS series $\{W_{t,9}^S\}$.

Similar calculations were performed for our other two datasets; the optimal p values were: 13 for the IBM dataset, and 11 for the S&P500 dataset. Figure 4 depicts plots of the Simple NoVaS transformed series for the three datasets of Figure 1. The variance stabilization effect is quite apparent; in particular, note that the market crash of October 1987 is hardly (if at all) noticeable in Figure 4 (b) and (c). A comparison with Figure 1 is quite striking.

Figure 5 shows histograms and Q-Q plots for the three NoVaS series of Figure 4. Comparing Figure 5 to Figure 2, it is visually apparent that the goal of normalization has been largely achieved. The histograms look quite normal and the Q-Q plots look quite straight; there is no indication of heavy-tails and/or outlying values in Figure 5, i.e., no “left-over” kurtosis to account for. A formal Kolmogorov-Smirnov (K-S) test of the hypothesis that the transformed series are normal confirms the conclusions of the visual inspection of the three Q-Q plots; the P -values were found to be 0.1654, 0.3638 and 0.4646 for the Yen/Dollar, S&P500 and IBM datasets respectively, so the hypothesis of normality is not rejected.

Remark 1.2 [On range-adjustment.] Focusing again on the Yen/Dollar data, note that—perhaps not surprisingly—the lowest P -value in K-S testing is associated with the smallest value of p for Simple NoVaS; recall that $p = 9$

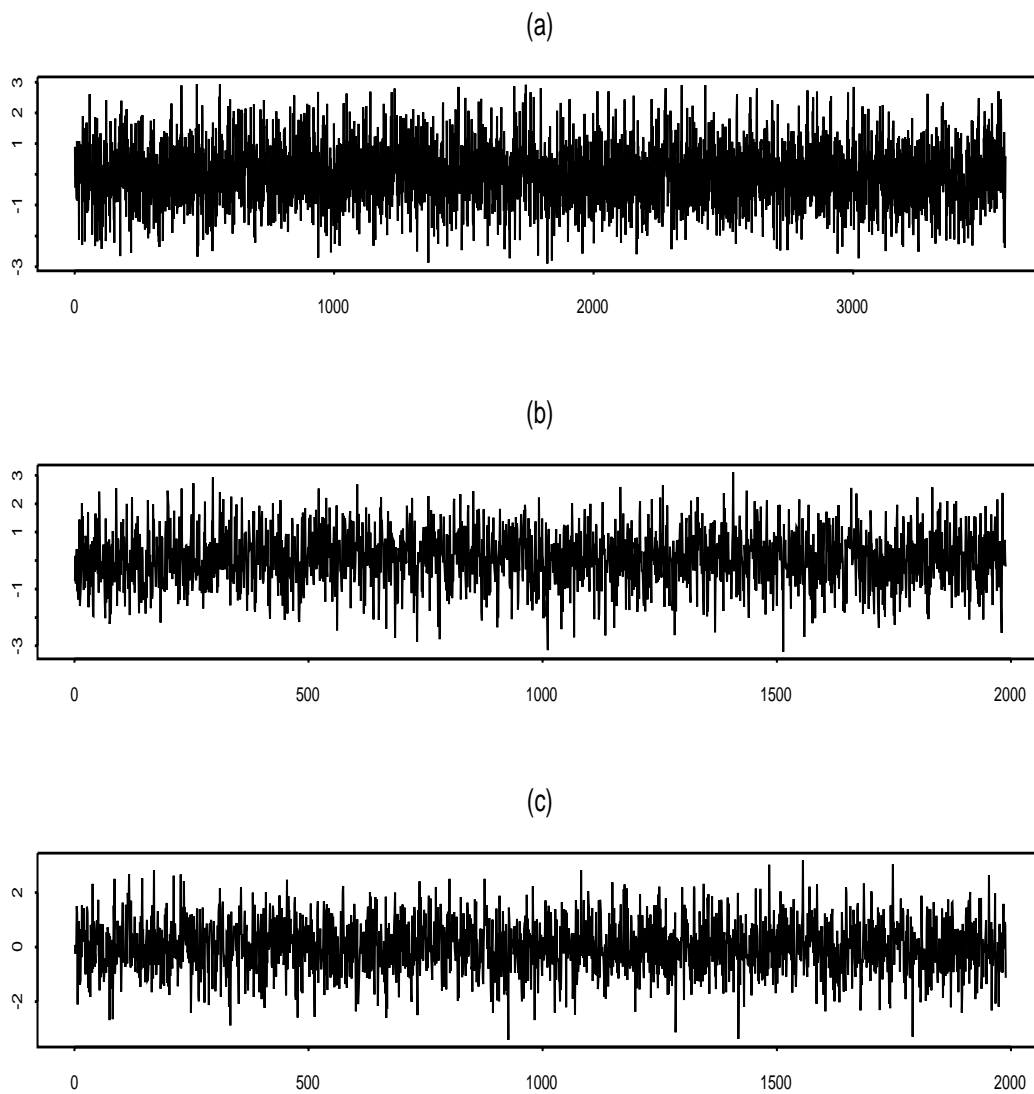


Figure 4: Plots of the Simple NoVaS transformed series corresponding to the three datasets of Figure 1.

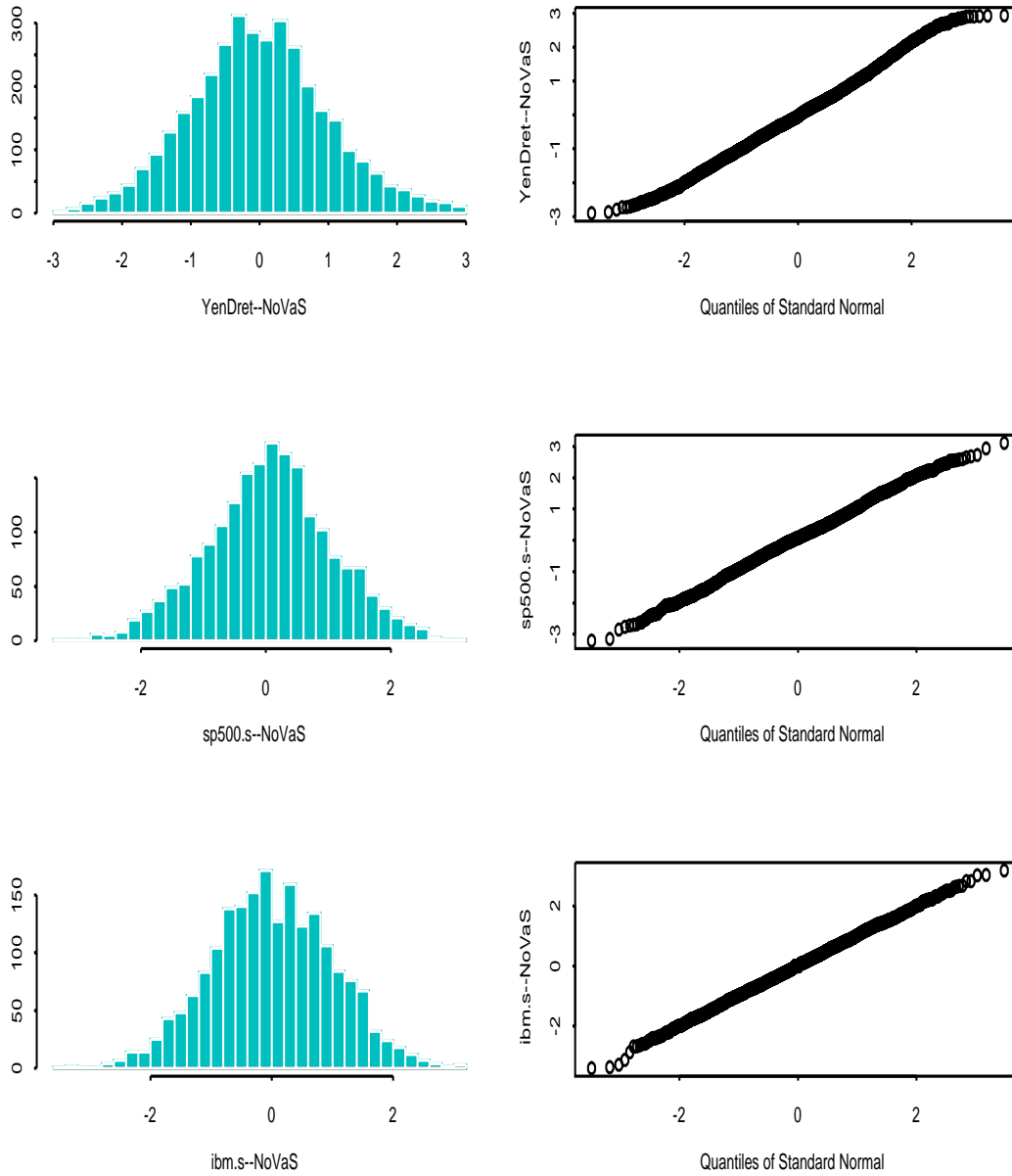


Figure 5: Histograms and Q-Q plots for the three NoVaS series of Figure 4.

for Yen/Dollar, whereas p was found to be 11 for S&P500, and 13 for IBM. A low value of p arising from our kurtosis matching algorithm may indicate an adverse effect of truncation to our normalization goal, and a subsequent need for range-adjustment.

Still the value $p = 9$ is high enough to yield an effective range of the Yen/Dollar NoVaS transform $\{W_{t,9}^S\}$ series of about 3.16 which is acceptable in terms of (9) being satisfied with $C = 3$. However, if one opted for the choice $C = \sqrt{2 \ln n}$, then in this case C would be about 4 and a range-adjustment step would be required leading to the choice $p = 15$; note that $KURT_n(W_{t,15}^S) = 3.51$ which is still relatively close to the target value of 3. Nevertheless, the K-S test of $W_{t,15}^S$ has a P -value of 0.0008, leading us to reject the hypothesis that $p = 15$ succeeds in transforming the Yen/Dollar series to normality. Therefore, the simple choice of $C = 3$ seems to be preferable to the alternative $C = \sqrt{2 \ln n}$; with the simple choice $C = 3$ there is no need for range-adjustment in connection with the Yen/Dollar data. The higher values of p in connection with the S&P500 and IBM datasets correspond to ranges of about 3.3 and 3.6 respectively, indicating even less of a need for possible range-adjustment.

Remark 1.3 [Normalization of joint distributions.] In the simple NoVaS algorithm, the target was 4th moment matching of $W_{t,p}^S$ to the corresponding Gaussian moment, i.e., to obtain $KURT_n(W_{t,p}^S) \simeq 3$; this procedure has the goal of (approximately) normalizing the marginal distribution of $W_{t,p}^S$. Interestingly, this simple procedure seems to be somehow effective in normalizing joint distributions as well, e.g. the joint distribution of $W_{t,p}^S$ and its lagged version $W_{t-1,p}^S$, which is a highly desirable objective. Table 1 gives the sample kurtosis of the series $\tilde{W}_{t,9,i}^S = W_{t,9}^S + \lambda_i W_{t-1,9}^S$ (in the case of the Yen/Dollar dataset) for different values of λ_i . Notably, all the entries of Table 1 are close to the nominal value of 3 supporting the claim of approximate normalization of the *joint* distribution of the pair $(W_{t,9}^S, W_{t-1,9}^S)$.

λ_i	-4	-1	-0.5	0	0.5	1	4
$KURT_n(\tilde{W}_{t,9,i}^S)$	2.92	2.89	2.98	3.03	3.03	3.10	3.12

Table 1: (Yen/Dollar example) Sample kurtosis of $\tilde{W}_{t,9,i}^S = W_{t,9}^S + \lambda_i W_{t-1,9}^S$ for different values of λ_i .

However, if one wanted to *ensure* that some joint distributions are also normalized—at least as far as 4th moments are concerned—then the moment matching criterion of the algorithm can be modified. To fix ideas, consider the target of normalizing the joint distribution of $W_{t,p}^S$ and $W_{u,p}^S$. The Cramér-Wold device suggests that we simultaneously consider some linear combinations of the type:

$$\tilde{W}_{t,p,i}^S = W_{t,p}^S + \lambda_i W_{u,p}^S \quad \text{for } i = 1, \dots, K,$$

where the λ_i 's are some chosen constants as in Table 1. The simple NoVaS algorithm is then altered to focus on the kurtosis of $\tilde{W}_{t,p,i}^S$ instead of that of $W_{t,p}^S$; to elaborate, the last step of the simple NoVaS algorithm would read:

- Pick p such that $\max_i |KURT_n(\tilde{W}_{t,p,i}^S) - 3|$ is minimized.

1.4 Exponential NoVaS algorithm

In the Exponential NoVaS, to specify all the a_i s, one just needs to specify the two parameters p and $c > 0$, in view of (7). However, because of the exponential decay, the parameter p is now of secondary significance as the following algorithm suggests; thus, we may concisely denote the exponential NoVaS transformation by $W_{t,c}^E$.

ALGORITHM FOR EXPONENTIAL NOVAS:

- Let p take a very high starting value, e.g., let $p \simeq n/4$ or $n/5$.
- Let $\alpha = 0$ and $a_i = c'e^{-ci}$ for all $0 \leq i \leq p$, where $c' = 1/\sum_{i=0}^p e^{-ci}$ by eq. (7).
- Pick c in such a way that $|KURT_n(W_{t,c}^E) - 3|$ is minimized.

It is apparent that the above search will be practically conducted over a discrete grid of c -values; let c_0 denote the resulting minimizer. Consequently, the following range-adjustment safeguard may be added.

- If c_0 as found above is such that (9) is not satisfied, then decrease c stepwise (starting from c_0) over the discrete grid until (9) is satisfied.

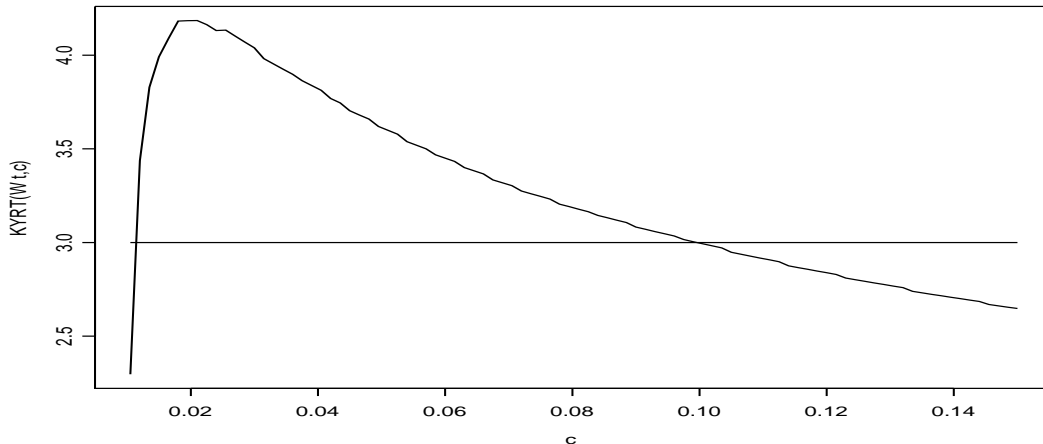


Figure 6: Illustration of the Exponential NoVaS algorithm for the Yen/Dollar dataset: plot of $KURT_n(W_{t,c}^E)$ as a function of c ; the solid line indicates the Gaussian kurtosis of 3.

Finally, the value of p must be trimmed for efficiency of usage of the available sample; to do this we can simply discard the a_i coefficients that are close to zero, i.e., those that fall below a certain threshold ϵ which is the practitioner’s choice. A threshold value of 0.01 is reasonable in connection with the a_i which—it should be stressed—are normalized to sum to one.

- Trim the value of p by a criterion of the type: if $a_i < \epsilon$, then let $a_i = 0$. Thus, if $a_i < \epsilon$, for all $i \geq i_0$, then let $p = i_0$, and renormalize the a_i s so that their sum (for $i = 0, 1, \dots, i_0$) equals one.

An illustration of the Exponential NoVaS algorithm is now given for the Yen/Dollar dataset. Figure 6 is a plot of $KURT_n(W_{t,c}^E)$ as a function of c . Except for values of c very close to zero, $KURT_n(W_{t,c}^E)$ seems to be monotonically decreasing hitting the value 3 for $c \simeq 0.0985$. Nevertheless, the behavior of $KURT_n(W_{t,c}^E)$ for c close to zero is not a fluke; rather it is a predictable outcome of our truncation/clipping of all coefficients that are less than ϵ (which was equal to 0.01 for the purposes of Figure 6). If a very low value for ϵ is used—say even that ϵ is set to zero—then the plot of $KURT_n(W_{t,c}^E)$ will be decreasing for all values of c .

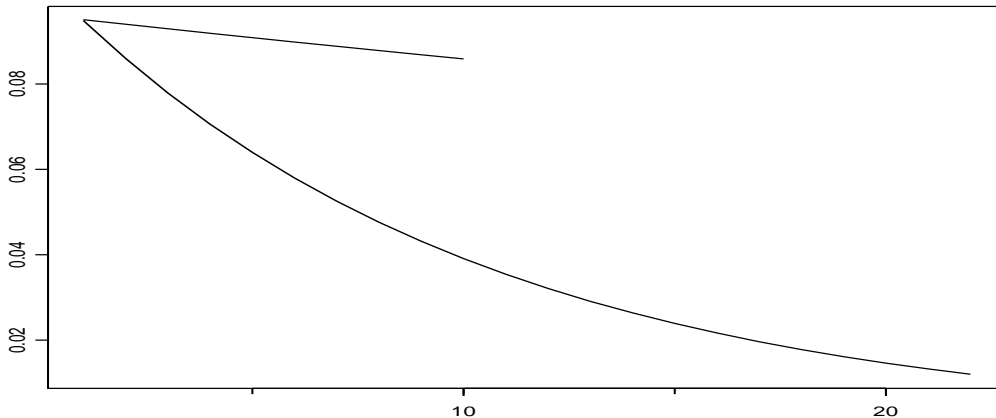


Figure 7: Plot of the exponential coefficients a_i versus the index $i = 1, \dots, p$ for the two values of c suggested by Figure 6; note that $c \simeq 0.0113$ corresponds to $p = 10$, while $c \simeq 0.0985$ corresponds to $p = 22$.

To further elaborate, note that Figure 6 indicates $KURT_n(W_{t,c}^E)$ hitting the value 3 for another value of c as well, namely for $c \simeq 0.0113$. Figure 7 shows a plot of the exponential coefficients a_i versus the index $i = 1, \dots, p$ for the two values of c suggested by Figure 6; due to the truncation effect with $\epsilon = 0.01$, we have $c \simeq 0.0113$ corresponding to $p = 10$, while $c \simeq 0.0985$ corresponds to $p = 22$. Note that the ultra-slow decay of the a_i coefficients in the case $c \simeq 0.0113$, combined with the truncation effect at $p = 10$, makes the Exponential NoVaS with $c \simeq 0.0113$ very similar to a Simple NoVaS with $p = 10$; this is because the exponential coefficients decay so slowly that are close to being constant for $i = 1, \dots, p$.

To sum up: a plot with shape such as Figure 6 is typical when a nonzero ϵ is used, suggesting that the function $|KURT_n(W_{t,c}^E) - 3|$ may have two values of c minimizing it. The higher of those two c values is the *bona fide* exponential decay constant; the lower of the two c values is typically not useful—but the p corresponding to that lower c value is a good indicator of the optimal p in Simple NoVaS.

Analogues of Figures 4 and 5 can be constructed using the Exponential NoVaS algorithm on our three datasets; they are not given here to save space

as they are visually very similar to the Simple NoVaS results of Figures 4 and 5. The optimal c values were: 0.070 (with $p = 27$) for the IBM dataset, and 0.084 (with $p = 24$) for the S&P500 dataset.

Note that as in the simple NoVaS algorithm, for the Exponential NoVaS as well we could focus on moment matching for the linear combinations of $W_{t,c}^E$ of $W_{u,c}^E$ (say) instead of $W_{t,c}^E$. In addition, the Exponential NoVaS algorithm could be extended to include a sum of two or more exponentials, i.e., a situation where $a_i = c'e^{-ci} + d'e^{-di} \dots$. The generalization may well include higher order moment matching and/or looking at linear combinations of higher order lags.

2 Volatility prediction

2.1 Some basic notions: L_1 vs L_2

In this section, we consider the problem of prediction of X_{n+1}^2 based on the observed past $\mathcal{F}_n = \{X_t, 1 \leq t \leq n\}$. Under the zero mean assumption, a first predictor is given by a simple empirical estimator of the (unconditional) variance σ_X^2 of the series $\{X_t, 1 \leq t \leq n\}$, for example, $s_n^2 = n^{-1} \sum_{k=1}^n X_k^2$; this will serve as our ‘benchmark’ for comparisons.

The above predictor is quite crude as it implicitly assumes that the squared returns $\{X_t^2, 1 \leq t \leq n\}$ are independent which is typically not true. As a matter of fact, the basic premise regarding financial returns is that they are dependent although uncorrelated—hence the typical assumption of nonlinear/non-normal models in that respect. For example, Figure 8(a) confirms that for the Yen/Dollar dataset the returns indeed appear uncorrelated. However, the squared returns appear to be correlated even for lags as high as 25 days; see Figure 8(b).

An immediate improvement over the above benchmark should thus be obtainable by a simple forecasting method such as ‘exponential smoothing’; see e.g. Hamilton (1994). In our context, the exponential smoothing predictor of X_{n+1}^2 is of the form $\sum_{k=1}^q \delta^k X_{n-k}^2 / \sum_{j=1}^q \delta^j$ where δ is a number in $(0, 1)$ and q an appropriate practical truncation limit.⁵ The ‘discount’ factor δ is typically chosen by a cross validation (CV) step with the forecasting goal in mind; that is, δ is chosen to minimize the Mean Squared Error (MSE) of prediction within the available dataset.

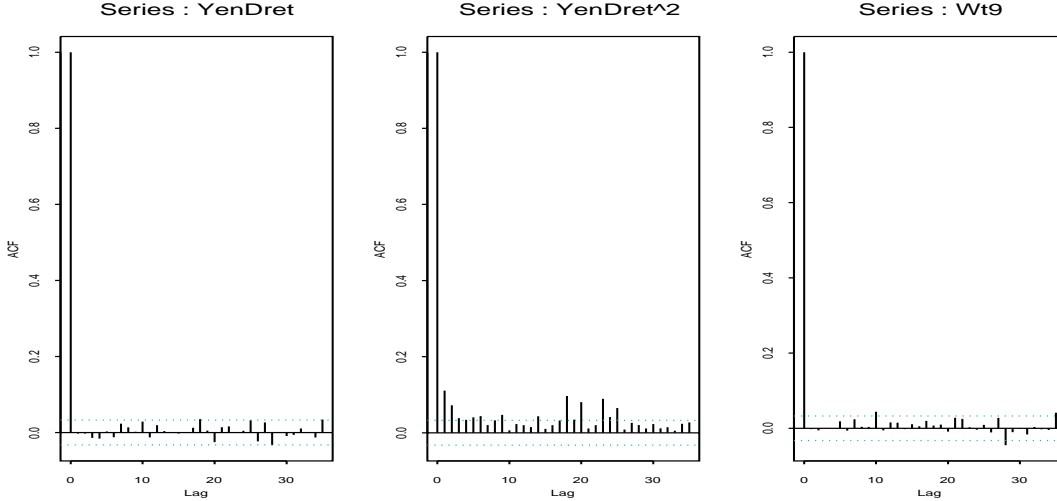


Figure 8: (Yen/Dollar example) (a) Correlogram of the returns series $\{X_t\}$; (b) Correlogram of the squared returns $\{X_t^2\}$; (c) Correlogram of the optimal Simple NoVaS series $\{W_{t,9}^S\}$.

The CV method is intuitive and ubiquitous; for example, it is built-in in many statistical software packages such as the ITSM specialized time series software accompanying the highly influential book by Brockwell and Davis (1991). Despite its appeal, however, rigorous analysis of the performance of the CV method is difficult and has been lacking in the literature. A most notable exception is the recent paper by Gijbels et al. (1999) where the performance of the CV method is successfully analyzed under a model of the type: “deterministic trend function plus error”. Unfortunately, such a model can not be reasonably assumed in connection with our (squared) returns series. Consequently, the results in the first row of Table 2 should not come as a surprise. The exponential smoothing predictor is seen to perform very poorly; in fact, it is performing quite worse than our naive benchmark predictor which is our estimate of the (unconditional) variance σ_X^2 .

Note that the ‘exponential smoothing’ predictor is linear in the variables $\{X_t^2, 1 \leq t \leq n\}$ but the coefficients in the linear combination are not chosen according to an optimality criterion. As a matter of fact, exponential smoothing is analogous to fitting an AR(1) model to the squared returns.

Thus, a further step in constructing a good predictor of X_{n+1}^2 may be to consider the best linear predictor which is of the type

$$(1 - \sum_{i=1}^r b_i) s_n^2 + \sum_{i=1}^r b_i X_{n+1-i}^2. \quad (10)$$

In the above, the b_i coefficients can be estimated by fitting an AR(r) model to the (de-means) squared returns $\{X_t^2, 1 \leq t \leq n\}$ with the order r typically determined by minimizing Akaike's AIC criterion—see e.g. Brockwell and Davis (1991).

It should be noted though that this linear predictor is typically suboptimal since the series $\{X_t^2\}$ is generally non-normal and nonlinear. However, the main reason that eq. (10) may give a poor predictor in practice is the following: the correlogram of the squared returns $\{X_t^2, 1 \leq t \leq n\}$ does *not* give an accurate estimation of the true correlation structure mainly due to the underlying heavy tails (and non-linearities); see e.g. Resnick et al. (1999). For example, using the AIC criterion to pick the order r in connection with the squared Yen/Dollar returns yields $r = 26$; this is not surprising in view of the correlogram of Figure 8 (b), but it is hard to seriously entertain a model of such high order for this type of data. An experienced researcher might instead fit a low order AR or ARMA model in this situation.

As mentioned before, fitting an AR(1) model for prediction is closely related to the exponential smoothing forecaster. Interestingly, fitting an ARMA(1,1) to the squared returns is in the spirit of a GARCH(1,1) model since the GARCH(1,1) predictor of X_{n+1}^2 has the same form as predictor (10) with the b_i coefficients decaying exponentially as in an ARMA(1,1) model.

The GARCH(1,1) model is the most popular among the GARCH(p, q) models of Bollerslev (1986) as it is believed to achieve the most parsimonious fit with returns data. Recall that the ARCH family is a subset of the GARCH family since an ARCH(p) model is equivalent to a GARCH($p, 0$); in addition, a GARCH(p, q) model is equivalent to an ARCH(∞) with a special structure (typically exponential) for its a_i coefficients—see Hamilton (1994) or Gouriéroux (1997). The GARCH(1,1) model is described by the equation:

$$X_t = h_t Z_t \quad \text{with} \quad h_t^2 = C + AX_{t-1}^2 + Bh_{t-1}^2;$$

as before, the $\{Z_t\}$ s are i.i.d. errors, and h_t^2 is the 'volatility' as in eq. (2).

In order to compare the different predictors of squared returns, we will use two popular performance measures: Mean Squared Error (MSE) of prediction and Mean Absolute Deviation (MAD) of prediction both relative to the benchmark; these are of course nothing other than the L_2 and L_1 norms of the prediction error respectively, divided by the corresponding L_2 or L_1 norm of the benchmark’s prediction error. Hansen et al. (2003) have also recently compared volatility predictions using both L_2 and L_1 loss functions.

Predictor type	Yen/Dollar	S&P500	IBM
Exponential Smoothing with CV	1.085	1.207	1.299
Eq. (10)—AR fit with AIC	1.008	0.988	1.010
Eq. (2)—GARCH(1,1) with normal errors	1.020	1.187	1.260
Eq. (2)—GARCH(1,1) with t -errors	1.039	1.190	1.277

Table 2: Entries give the empirical Mean Squared Error (MSE) of prediction of squared returns relative to benchmark; note that the MSE of prediction achieved by the benchmark was 2.98e-008, 1.72e-006, 1.74e-006 in the three cases Yen/Dollar, S&P500, IBM respectively.

Table 2 reports the L_2 prediction performance of the aforementioned predictors, namely exponential smoothing with CV, the linear model (10) with order chosen by minimizing the AIC, and the GARCH(1,1) with normal and t -errors (the latter having degrees of freedom estimated from the data); all computations were done in S+. It is apparent that the performance of all methods is rather poor as they seem to perform worse even than the (naive) benchmark. In particular, the performance of the GARCH(1,1) predictor is abysmal, be it with normal or t errors.

Due to empirical results such as those in Table 2, it has been widely believed that ARCH/GARCH models are characterized by “poor out-of-sample forecasting performance vis-a-vis daily squared returns”; see Andersen and Bollerslev (1998) and the references therein. To further quote Andersen and Bollerslev (1998): “numerous studies have suggested that ARCH and stochastic volatility models provide poor volatility forecasts”.⁶

Nevertheless, the entries of Table 3a on the L_1 prediction performance tell a different story, namely that all aforementioned predictors—with the

exception of exponential smoothing—outperform the benchmark when errors are measured in the L_1 norm.

Predictor type	Yen/Dollar	S&P500	IBM
Exponential Smoothing with CV	1.013	1.187	1.132
Eq. (10)—AR fit with AIC	0.989	0.943	0.983
Eq. (2)—GARCH(1,1) with normal errors	0.986	0.995	0.994
Eq. (2)—GARCH(1,1) with t -errors	0.830	0.875	0.918

Table 3a: Entries give the empirical Mean Absolute Deviation (MAD) of prediction of squared returns relative to benchmark; note that the MAD of prediction achieved by the benchmark was $6.28\text{e-}005$, $1.57\text{e-}004$, $2.29\text{e-}004$ in the three cases Yen/Dollar, S&P500, IBM respectively.

To see why such a big discrepancy exists between the two performance measures, L_1 and L_2 , we return to our data. Let $VAR_k(Y)$ and $KURT_k(Y)$ denote the empirical (sample) variance and kurtosis of dataset Y up to time k , i.e., $\{Y_1, \dots, Y_k\}$. By the (strong) law of large numbers, as k increases, $VAR_k(Y)$ should tend to the variance of the random variable Y_1 be that infinite or not. Similarly, $KURT_k(Y)$ should tend to the kurtosis of Y_1 be that infinite or not. Thus, plotting $VAR_k(Y)$ and $KURT_k(Y)$ as functions of k one may be able to visually gauge whether Y_1 has finite second and/or fourth moments; this is done in Figure 9 for the Yen/Dollar dataset.

It appears that the Yen/Dollar dataset may have finite variance as the plot in Figure 9 (a) seems to converge. Nevertheless, it seems that it has an infinite fourth moment as the plot in Figure 9 (b) seems to diverge with each extreme value ‘jolt’. The same conclusions, namely finite variance but infinite fourth moment, seem to also apply to our other two datasets.

Therefore, it is hardly surprising that the L_2 measure of prediction performance yields unintuitive results: the MSE of predicting X_{n+1}^2 is essentially a fourth moment, and the data suggest that fourth moments may be infinite! It is unreasonable to use an L_2 measure of performance in a set-up where L_2 norms may not exist.

Note that the GARCH predictions for Tables 2 and 3a were performed—as customary—using the predictor (2). But eq. (2) gives the conditional

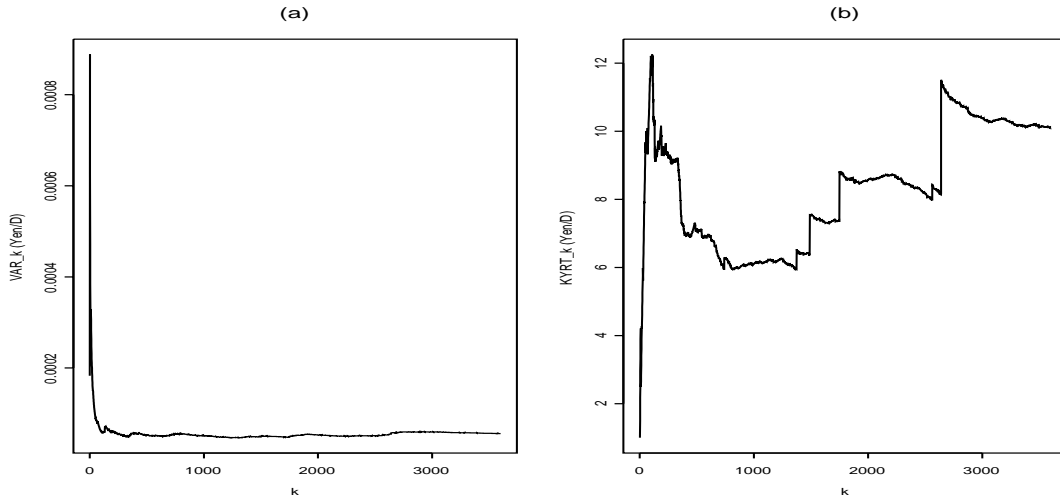


Figure 9: (Yen/Dollar example) (a) Plot of $VAR_k(X)$ as a function of k ; (b) Plot of $KURT_k(X)$ as a function of k .

expectation of X_{n+1}^2 given \mathcal{F}_n under the ARCH(p) model (1) with standard normal errors $\{Z_t\}$. If the errors are *not* standard normal, then eq. (2) is not the conditional expectation any longer.⁷ For example, $E(t_5^2) \simeq 1.67$ which is far from the value of one which holds under normality; here t_5 denotes a random variable distributed according to Student's t distribution with 5 degrees of freedom—a typical value for the degrees of freedom associated with our data.

Furthermore, under our objective of L_1 prediction, the optimal predictor is the conditional median—*not* the conditional expectation. Hence, the optimal GARCH(1,1) predictor of X_{n+1}^2 in the L_1 sense is given by

$$Median(X_{n+1}^2 | \mathcal{F}_n) = (a + \sum_{i=1}^p a_i X_{n+1-i}^2) Median(Z_{n+1}^2); \quad (11)$$

note that $Median(Z_{n+1}^2) \simeq 0.45$ if $Z_t \sim N(0, 1)$, whereas $Median(Z_{n+1}^2) \simeq 0.53$ if $Z_t \sim t_5$.

Table 3b shows the L_1 prediction performance of our two GARCH(1,1) models using the optimal L_1 predictor (11); again note that eq. (11) in the GARCH(1,1) setting should be interpreted as having $p = \infty$ (or very

large), and a_i coefficients decaying exponentially according to the GARCH structure—see e.g. Gouriéroux (1997, Ch. 4.1.5).

Predictor type	Yen/Dollar	S&P500	IBM
Eq. (11)—GARCH(1,1) with normal errors	0.812	0.867	0.892
Eq. (11)—GARCH(1,1) with t -errors	0.800	0.856	0.893

Table 3b: Entries give the empirical Mean Absolute Deviation (MAD) of prediction of squared returns relative to benchmark.

As expected, using the correct predictor leads to ameliorated performance as a comparison of Table 3b to Table 3a shows. In particular, as was expected, *both* GARCH(1,1) models outperform the lineal predictor (10) in the L_1 sense. Interestingly, assuming the t distribution for the errors does not seem to give any appreciable advantage—if at all—over the customary normal errors assumption.

To conclude, by contrast to what is widely believed, ARCH/GARCH models *do* have predictive validity for the squared returns. However, to appreciate and take advantage of this one must: (a) use a more meaningful measure of prediction such as L_1 , and (b) use the proper predictor, i.e., the conditional median in the L_1 prediction case.

In the sequel we will focus exclusively on the L_1 measure of performance and the Mean Absolute Deviation (MAD) of prediction. Although we have seen that GARCH models do have reasonable predictive validity, in what follows we show how we can obtain even better volatility predictions using the NoVaS transformation.

Remark 2.1 [On ‘honest’ predictions] All predictions reported in this paper are ‘honest’ in the sense that to predict X_{t+1}^2 , only information set $\{X_1, \dots, X_t\}$ was used in estimating the particulars of the predictor, be it GARCH parameters, AIC for AR-fitting, the exponential smoothing constant, or NoVaS parameters in what follows. Furthermore, we did not follow the usual practice of splitting the dataset in half, estimating parameters from the first half and predicting the second half; this is unrealistic as parameters would/should be updated constantly—even daily—in practice. For the purposes of the paper’s numerical work, however, it was deemed unnecessary—

and computationally too expensive—to update the parameters daily as the parameters are quite stable for that to be of any influence. The updating was typically performed every $n/10$ days⁸ for a dataset of size n .

2.2 Volatility prediction using NoVaS

To give an alternative procedure for prediction of X_{n+1}^2 based on the observed past \mathcal{F}_n we now focus on volatility forecasting based on NoVaS. To describe this, suppose that the order $p(\geq 0)$ and the parameters α, a_0, \dots, a_p of NoVaS have already been chosen.

First note that we can further re-arrange the NoVaS equation (4) to yield:

$$X_t^2 = \frac{W_{t,a}^2}{1 - a_0 W_{t,a}^2} \left(\alpha s_{t-1}^2 + \sum_{i=1}^p a_i X_{t-i}^2 \right) \quad (12)$$

and

$$X_t = \frac{W_{t,a}}{\sqrt{1 - a_0 W_{t,a}^2}} \sqrt{\alpha s_{t-1}^2 + \sum_{i=1}^p a_i X_{t-i}^2}. \quad (13)$$

The one-step ahead prediction problem can be generally defined as follows. Let g be some (measurable) function of interest; examples include $g_0(x) = x$, $g_1(x) = |x|$, and $g_2(x) = x^2$, the latter being the function of interest for volatility prediction. From eq. (13) it follows that the predictive (given \mathcal{F}_n) distribution of $g(X_{n+1})$ is identical to the distribution of the random variable

$$g \left(A_n \frac{W}{\sqrt{1 - a_0 W^2}} \right) \quad (14)$$

where $A_n = \sqrt{\alpha s_n^2 + \sum_{i=1}^p a_i X_{n+1-i}^2}$ is treated as a constant given the past \mathcal{F}_n , and the random variable W has the same distribution as the conditional (on \mathcal{F}_n) distribution of the random variable $W_{n+1,a}$.

Therefore, our best (in an L_1 sense) prediction of $g(X_{n+1})$ given \mathcal{F}_n is given by the median of the conditional (given \mathcal{F}_n) distribution of $g(X_{n+1})$, i.e.,

$$\widehat{g(X_{n+1})} := \text{Median} \left(g \left(A_n \frac{W_{n+1,a}}{\sqrt{1 - a_0 W_{n+1,a}^2}} \right) \mid \mathcal{F}_n \right) \quad (15)$$

Specializing to the case of interest, i.e., volatility prediction and the function $g_2(x) = x^2$, we then have

$$\widehat{X_{n+1}^2} = \mu_2 A_n^2 \quad (16)$$

where

$$\mu_2 = \text{Median} \left(\frac{W_{n+1,a}^2}{1 - a_0 W_{n+1,a}^2} \middle| \mathcal{F}_n \right).$$

Now observe that—up to the effect of initial conditions—the information set $\mathcal{F}_n = \{X_t, 1 \leq t \leq n\}$ is approximately⁹ equivalent to the information set $\tilde{\mathcal{F}}_n = \{W_{t,a}, p < t \leq n\}$. To see this, note that eq. (13)—when iterated—gives an expression for X_t in terms of $\tilde{\mathcal{F}}_n$; conversely, eq. (4) defines $W_{t,a}$ in terms of \mathcal{F}_n . Thus, we can use the expression

$$\mu_2 \approx \text{Median} \left(\frac{W_{n+1,a}^2}{1 - a_0 W_{n+1,a}^2} \middle| \tilde{\mathcal{F}}_n \right) \quad (17)$$

in connection with the predictor given in eq. (16).

Our task now is significantly simplified: find the predictive distribution of the random variable $W_{n+1,a}$ based on its own recent past $\tilde{\mathcal{F}}_n$. But—by construction— $W_{t,a}$ should be approximately equal to a normal random variable. In addition, as mentioned in Section 1, the joint distributions of the series $\{W_{t,a}, t = p+1, \dots, n\}$ are also typically normalized by the NoVaS transformation. Thus, the series $\{W_{t,a}, t = p+1, \dots, n\}$ may be thought of as an (approximate) *Gaussian series* in which case optimal prediction is effectively *linear* prediction since all dependencies should be captured in the correlogram; see e.g. Brockwell and Davis (1991).

Under this Gaussian/linear dependence structure, the conditional (on $\tilde{\mathcal{F}}_n$) distribution of $W_{n+1,a}$ should be close to a normal with mean (and median) approximately given by

$$\hat{W}_{n+1,a} = \sum_{i=1}^q c_i W_{n-i+1,a}, \quad (18)$$

and *constant* variance σ_{pred}^2 , i.e., σ_{pred}^2 not depending on $\tilde{\mathcal{F}}_n$. Here again the order q is usually chosen in practice by minimizing Akaike's AIC criterion, and the coefficients c_i can easily be found by fitting an AR(q) model to the

$\{W_{t,a}, t = p + 1, \dots, n\}$ series. Fitting an AR(q) model, e.g. by the Durbin-Levinson algorithm, also gives an estimate of the prediction error variance σ_{pred}^2 .

Note that the simplified expression (17) still represents an unknown quantity but it could conceivably be approximated by Monte Carlo, for example using the normal predictive density that has mean given by (18) and variance σ_{pred}^2 —recall though that this normal density should be truncated to an effective range of $\pm 1/\sqrt{a_0}$. However, a very large number of replications would be required due to the heavy tails of the distribution of $W^2/(1 - a_0W^2)$. In addition, it should be stressed that the normal (conditional or unconditional) density for $W_{n+1,a}$ is only an approximation; thus, it is more appropriate to estimate μ_2 empirically from the data without resort to the normal distribution.

To fix ideas, note that if the correlogram of the series $\{W_{t,a}, t = p + 1, \dots, n\}$ indicates no significant correlations—as is typically the case in practice¹⁰—then we can infer that the series $\{W_{t,a}\}$ is not only uncorrelated but also independent (by the approximate joint normality of its marginal distributions). Therefore, the conditional (on $\tilde{\mathcal{F}}_n$) distribution of $W_{n+1,a}$ would equal the unconditional distribution of $W_{n+1,a}$. Hence, we may estimate μ_2 by a sample median, i.e., let

$$\hat{\mu}_2 = median\left\{\frac{W_{t,a}^2}{1 - a_0W_{t,a}^2}; t = p + 1, p + 2, \dots, n\right\} \quad (19)$$

and subsequently predict X_{n+1}^2 by

$$\hat{\mu}_2 A_n^2. \quad (20)$$

Remark 2.2 Although in our examples the NoVaS series $\{W_{t,a}, t = p + 1, \dots, n\}$ turned out to be effectively uncorrelated, one can not preclude the possibility that for other datasets the series $\{W_{t,a}\}$ may exhibit some correlations; in that case, the c_i coefficients in eq. (18) are not all zero, and a slightly more complicated procedure is suggested in order to estimate μ_2 . First, the predictive residuals must be collected from the data; to do this, let $e_t = W_{t,a} - \hat{W}_{t,a}$ for $t = r + 1, \dots, n$ where $r = \max(p, q)$. Then the conditional (on $\tilde{\mathcal{F}}_n$) distribution of $W_{n+1,a}$ may be approximated by the empirical distribution of the points $\{e_t + \hat{W}_{n+1,a}; t = r + 1, \dots, n\}$, i.e., by

the empirical distribution of the predictive residuals shifted to give it mean $\hat{W}_{n+1,a}$. In that case we would estimate μ_2 by¹¹

$$\hat{\mu}_2 = \text{median}\left\{\frac{(e_t + \hat{W}_{n+1,a})^2}{1 - a_0(e_t + \hat{W}_{n+1,a})^2}; t = r + 1, r + 2, \dots, n\right\} \quad (21)$$

and again predict X_{n+1}^2 by eq. (20).

Remark 2.3 We can generalize the previous discussion to an interesting class of prediction functions g as in eq. (14), namely the power family where $g(x) = x^k$ for some fixed k , and the power–absolute value family where $g(x) = |x|^k$. Let $g_k(x)$ denote either the function x^k or the function $|x|^k$; then eq. (15) suggests that our best predictor of $g_k(X_{n+1})$ given \mathcal{F}_n is $\widehat{g_k(X_{n+1})} = \mu_k A_n^k$, where

$$\mu_k = \text{Median}\left(g_k\left(\frac{W_{n+1,a}}{\sqrt{1 - a_0 W_{n+1,a}^2}}\right) \mid \mathcal{F}_n\right).$$

As before, μ_k can be estimated by an appropriate sample median. Let us consider the two cases separately, Case I (where the NoVaS series $\{W_{t,a}\}$ can be assumed uncorrelated), and Case II (where the NoVaS series appears correlated). Under Case I, we estimate μ_k by

$$\hat{\mu}_k = \text{median}\left\{g_k\left(\frac{W_{t,a}}{\sqrt{1 - a_0 W_{t,a}^2}}\right); t = p + 1, p + 2, \dots, n\right\}$$

whereas under Case II the estimator becomes

$$\hat{\mu}_k = \text{median}\left\{g_k\left(\frac{e_t + \hat{W}_{n+1,a}}{\sqrt{1 - a_0(e_t + \hat{W}_{n+1,a})^2}}\right); t = r + 1, r + 2, \dots, n\right\}$$

Remark 2.4 In this section, the procedure of one-step ahead volatility prediction using NoVaS was outlined. The multi-step ahead prediction problem, i.e., predicting X_{n+m}^2 given \mathcal{F}_n , or in general predicting $g(X_{n+m})$ given \mathcal{F}_n , for some $m \geq 1$, can be handled in a similar vein; the details are omitted.

2.3 Optimizing NoVaS for volatility prediction

In section 2.2, the methodology for volatility prediction based on NoVaS was put forth. Using this methodology the L_1 prediction performance of the Simple and Exponential NoVaS was quantified and tabulated in Table 4. It appears that, in the foreign exchange data, Exponential NoVaS offers some improvement over Simple NoVaS but not enough to outperform the optimal GARCH(1,1) predictors of Table 3b. The situation is different in stock-based returns where Simple and Exponential NoVaS perform comparably, and they are both slightly better than the two optimal GARCH(1,1) predictors.

Predictor type	Yen/Dollar	S&P500	IBM
Simple NoVaS	0.837	0.853	0.880
Exponential NoVaS	0.815	0.854	0.884

Table 4: Entries give the empirical Mean Absolute Deviation (MAD) of prediction of squared returns relative to benchmark.

It is interesting to note that the NoVaS methodology performs competitively in volatility prediction despite its extreme parsimony: both Simple and Exponential NoVaS have just one free parameter (p and c respectively—since the p in Exponential NoVaS is determined by the tolerance level ϵ). By contrast, the GARCH(1,1) with t -errors has *four* free parameters—the fourth being the degrees of freedom for the t -distribution.

The single free parameter in Simple and Exponential NoVaS is identified using the kurtosis matching ideas of Section 1. Nevertheless, one can entertain more general NoVaS schemes with two (or more) free parameters. In such set-ups, one (or more) of the parameters can be identified by kurtosis matching (of the data or lagged linear combinations thereof); the remaining free parameters can then be identified by specific optimality criteria of interest, e.g. optimal volatility prediction.

Although many different multi-parameter NoVaS schemes can be devised, we now elaborate on the possibility of a nonzero value for the parameter α in (4) in connection with the Exponential NoVaS. We thus define the General Exponential NoVaS that has two free parameters, α and c , and will be denoted by $W_{t;c,\alpha}^{GE}$. The search is performed using a grid of possible

values for α , say $\alpha_1, \alpha_2, \dots, \alpha_K$. In picking the grid values, note that the kurtosis matching goal may only be possible with small values of α ; else, the intermediate value argument of Remark 1.1 may fail.

ALGORITHM FOR GENERAL EXPONENTIAL NOVAS:

- For $k = 1, \dots, K$ perform the following steps.
 - Let p take a very high starting value, e.g., let $p \simeq n/4$ or $n/5$.
 - Let $\alpha = \alpha_k$ and $a_i = c'e^{-ci}$ for all $0 \leq i \leq p$, where $c' = (1 - \alpha_k) / \sum_{i=0}^p e^{-ci}$ by eq. (7).
 - Pick c in such a way that $|KURT_n(W_{t;c,\alpha_k}^{GE}) - 3|$ is minimized, and denote by c_k the minimizing value.¹²
 - Trim the value of p to some value p_k as before: if $a_i < \epsilon$, then set $a_i = 0$. Thus, if $a_i < \epsilon$, for all $i \geq i_k$, then let $p_k = i_k$, and renormalize the a_i s so that their sum (for $i = 0, 1, \dots, p_k$) equals $1 - \alpha_k$ by eq. (7).
- Finally, compare the models $\{W_{t;c_k,\alpha_k}^{GE}, k = 1, \dots, K\}$ in terms of their volatility prediction performance, and pick the model with optimal such performance.

An illustration of the General Exponential NoVaS Algorithm in connection with our three main datasets is presented in Table 5, where for each different value of α , the L_1 volatility prediction performance is given.

The results in Table 5 are very interesting. Firstly, the L_1 measure appears convex in α making the minimization very intuitive; a unique value of the optimal α (given in bold-face font) is easily found in each of the three datasets. Secondly, although all three datasets seems to benefit from a nonzero value of α , it is apparent that the significance of α differs according to the type of data involved: the Yen/Dollar series is not so sensitive on the value of the parameter α ; the S&P500 index is more sensitive, while the single stock price (IBM) is most sensitive. Still the improvement offered by a nonzero α in the Yen/Dollar data is big enough to make the General Exponential NoVaS competitive with the best GARCH(1,1) predictor of Table 3b.

α	Yen/Dollar	S&P500	IBM
0.0	0.815	0.854	0.884
0.1	0.807	0.839	0.862
0.2	0.805	0.835	0.849
0.3	0.803	0.825	0.843
0.4	0.813	0.826	0.837
0.5	N/A	N/A	0.823

Table 5: Entries give the empirical Mean Absolute Deviation (MAD) of prediction of squared returns relative to benchmark using the General Exponential NoVaS with parameter α ; the minimum MAD is given with boldface.¹³

α	Yen/Dollar	S&P500	IBM
0.0	0.0975	0.0825	0.0675
0.1	0.1150	0.1025	0.0800
0.2	0.1400	0.1350	0.0950
0.3	0.1825	0.1900	0.1250
0.4	0.2600	0.2925	0.1800
0.5	0.4100	0.5300	0.2900

Table 6: Entries give the optimal exponent c from kurtosis matching in the General Exponential NoVaS with parameter α ; note that the values in this table were computed using the whole available sample sizes.

Note that, as α increases, c increases accordingly, and p decreases (the latter is not shown as it can be easily calculated). This phenomenon is captured in Table 6 that was compiled using tolerance level $\epsilon = 0.01$. Finally, note that all (c, α) combinations in Table 6 are equally successful in normalizing the NoVaS transformation in terms of achieving a kurtosis of about 3. However, as previously alluded to, the N/A entries in Table 5 indicate values of α that are too big for the kurtosis matching to be successful.

2.4 Summary of data-analytic findings

- Because of the lack of finite fourth moments, the MSE is not a good measure of performance of prediction of squared returns; see Table 2 and Table 3a where the *same* predictors are compared with respect to MSE and MAD respectively.
- Using the L_1/MAD measure of performance, it is apparent that GARCH models *do* have predictive validity for the squared returns.
- As expected, once the L_1 setting is assumed, using the optimal predictor of eq. (11) gives an appreciable difference; compare Table 3a to Table 3b.
- Once the optimal GARCH predictor is used, it seems to make little difference whether the normal or t distribution is assumed for the errors; see Table 3b.
- Simple NoVaS seems comparable to Exponential NoVaS in volatility prediction, and they both outperform the best GARCH(1,1) predictors in stock-based returns—see Table 4.
- General Exponential NoVaS gives appreciable improvements over either Simple and/or Exponential NoVaS; these are of the order of a 2% reduction in MAD in the Yen/Dollar dataset, and about 5% reduction in the stock-based datasets—see Table 5.
- In stock-based datasets, the General Exponential NoVaS outperforms the optimal GARCH(1,1) predictors of Table 3b by a margin of about 5–10% in MAD.
- Although the model-free NoVaS predictor (in any of its forms) invariably outperforms GARCH models for volatility prediction in stock-based returns, the fine-tuned GARCH predictor, i.e., eq. (11) with t errors, is competitive with the best NoVaS predictor in our foreign exchange data; this may well indicate an important difference in the underlying stochastic structures between stocks and foreign exchange.

3 Conclusions

In this paper, a new methodology was introduced for dealing with financial returns time series data. By contrast to the customary viewpoint that is based on parametric/semi-parametric models (such as ARCH/GARCH), the new approach is totally model-free, and has at its core a novel normalizing and variance-stabilizing transformation (NoVaS). The NoVaS transformation was implemented in connection with three main datasets: a foreign exchange series (Yen vs. Dollar), a stock index series (the S&P500 index), and a stock price series (IBM).

Properties of the NoVaS transformation were discussed, and intuitive algorithms for optimizing it were presented in detail. Special emphasis was given on the problem of volatility prediction and the issue of a proper measure for quality of prediction. In particular, the case was made that financial returns data may not have a finite fourth moment in which case L_2 methods—such as conditional expectations—are inappropriate for predicting the squared returns. Thus, contrary to wide-spread beliefs, we show that the popular ARCH/GARCH models actually do have predictive validity for the squared returns when applied properly, i.e., in an L_1 setting.

A general algorithm for prediction of a function $g(X_{n+1})$ given the data X_1, \dots, X_n was also given based on the NoVaS transformation. With some very simple and intuitive choices for the NoVaS structure, it was shown that the NoVaS prediction methodology generally outperforms the aforementioned ARCH/GARCH models in the case at hand where $g(x) = x^2$.

References

- [1] Andersen, T.G. and Bollerslev, T. (1998). Answering the sceptics: yes, standard volatility models do provide accurate forecasts, *International Economic Review*, vol. 39, no.4, 885-905.
- [2] Andersen, T.G., Bollerslev, T., Christoffersen, P.F., and F. X. Diebold, (2006). “Volatility and Correlation Forecasting” in G. Elliott, C.W.J. Granger, and Allan Timmermann (eds.), *Handbook of Economic Forecasting*, Amsterdam: North-Holland, 778-878.

- [3] Andersen, T.G., Bollerslev, T. and Meddahi, N. (2004). Analytic evaluation of volatility forecasts, *International Econ. Rev.*, 45, 1079-1110.
- [4] Bachelier, L. (1900). *Theory of Speculation*. Reprinted in *The Random Character of Stock Market Prices*, P.H. Cootner (Ed.), Cambridge, Mass.: MIT Press, pp. 17-78, 1964.
- [5] Barndorff-Nielsen, O.E., Nielsen, B., Shephard, N. and Ysusi, C. (1996). Measuring and forecasting financial variability using realized variance with and without a model, *State Space and Unobserved Components Models: Theory and Applications*, A.C. Harvey, S.J. Koopman and N. Shephard (Eds.), Cambridge Univ. Press, pp. 205-235.
- [6] Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity, *J. Econometrics*, 31, 307-327.
- [7] Bollerslev, T., Chou, R. and Kroner, K. (1992). ARCH modelling in finance: a review of theory and empirical evidence, *J. Econometrics*, 52, 5-60.
- [8] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods, 2nd ed.*, Springer, New York.
- [9] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation, *Econometrica*, 50, 987-1008.
- [10] Fama, E.F. (1965). The behaviour of stock market prices, *J. Business*, 38, 34-105.
- [11] Gijbels, I., Pope, A. and Wand, M.P. (1999), Understanding exponential smoothing via kernel regression, *J.R.S.S. Ser. B*, vol. 61, pp. 39-50.
- [12] Ghysels, E., Santa-Clara, P. and Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies, *J. Econometrics*, vol. 131, Issues 1-2, pp. 59-95.
- [13] Gouriéroux, C. (1997). *ARCH Models and Financial Applications*, Springer Verlag, New York.
- [14] Hamilton, J.D. (1994). *Time Series Analysis*, Princeton Univ. Press, Princeton, New Jersey.

- [15] Hansen, P.R. and Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH (1,1)?, *J. Appl. Econometrics*, vol. 20, pp. 873-889.
- [16] Hansen, P.R. and Lunde, A. (2006). Consistent ranking of volatility models, *J. Econometrics*, Vol. 131, pp. 97-121.
- [17] Hansen, P.R., Lunde, A. and Nason, J.M. (2003). Choosing the best volatility models: the model confidence set approach, *Oxford Bulletin of Economics and Statistics*, 65, pp. 839-861.
- [18] Koopman, S.J., Jungbacker, B. and Hol, E. (2005). Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements, *J. Empirical Finance*, 12, 445-475.
- [19] Mandelbrot, B. (1963). The variation of certain speculative prices, *J. Business*, 36, 394-419.
- [20] Meddahi, N. (2001). An eigenfunction approach for volatility modeling, Technical report, CIRANO Working paper 2001s-70, Univ. of Montreal.
- [21] Nelson, D. (1991). Conditional heteroscedasticity in asset returns: a new approach, *Econometrica*, 59, 347-370.
- [22] Patton, A. (2005). Volatility forecast evaluation and comparison using imperfect volatility proxies, preprint available from the website: <http://fmg.lse.ac.uk/~patton/>
- [23] Politis, D.N. (2003a), ‘Model-free volatility prediction’. UCSD Dept. of Economics Discussion Paper 2003-16.
- [24] Politis, D.N. (2003b), ‘A normalizing and variance-stabilizing transformation for financial time series’, in *Recent Advances and Trends in Non-parametric Statistics*, (M.G. Akritas and D.N. Politis, Eds.), Elsevier (North Holland), 2003, pp. 335-347.
- [25] Poon, S. and C. Granger, (2003). “Forecasting Volatility in Financial Markets: A Review”, *Journal of Economic Literature*, 41, 478539.

- [26] Resnick, S., Samorodnitsky, G. and Xue, F. (1999). How misleading can sample ACF's of stable MA's be? (Very!). *Ann. Appl. Probab.*, vol. 9, no. 3, 797-817.
- [27] Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields*, D.R. Cox, David V. Hinkley and Ole E. Barndorff-Nielsen (eds.), London: Chapman & Hall, pp. 1-67.

Notes

1. Here, and throughout the paper, the term ‘volatility prediction’ will be taken to mean ‘prediction of squared returns’.
2. A causal estimate is one involving present and past data only, i.e., data $\{X_s, s \leq t\}$.
3. The reason for not including time t in the variance estimator is for purposes of notational clarity as well as the easy identifiability of the effect of the coefficient a_0 associated with X_t^2 in the denominator of equation (4).
4. Note that—unlike the usual i.i.d. framework—the normalizing and variance-stabilizing transformation in this time series setting is not an instantaneous function of each data point; rather, it is a function of a whole stretch of past data points.
5. The specification of q here resembles closely the choice of p in our Exponential NoVaS of Section 1.4. This is not a coincidence but due to the following analogy: in Exponential NoVaS, the technique of exponential smoothing is used with the purpose of constructing a (local) estimate of the variance of X_t for the subsequent NoVaS studentization. By contrast, in the present section exponential smoothing is considered in its usual function of forecasting/predicting the (future) value of the random variable X_{n+1}^2 .
6. In turn, Andersen and Bollerslev (1998) define the notion of ‘latent’ volatility based on an assumed underlying continuous-time diffusion structure, and show that ARCH/GARCH models are successful in predicting future ‘latent’ volatility instead.
7. For this reason, however, many practitioners use a re-scaled t distribution family, the re-scaling being such that $E(t_k^2) = 1$; to avoid any confusion, we will not use this type of re-scaling in the paper.
8. For the S&P500 dataset the updating was done every $n/7$ days; the choice of the latter was dictated by convergence issues in the numerical MLEs involved in GARCH modeling.
9. Note that the information set \mathcal{F}_n is exactly equivalent to the information set $\{X_1, \dots, X_p, W_{p+1,a}, W_{p+2,a}, \dots, W_{n,a}\}$. Due to the stationarity and sub-

ject to a usual weak dependence condition—such as mixing—on the series $\{X_t\}$, the random variables $\{X_t, t > t_0\}$ will be approximately independent of the “initial conditions” X_1, \dots, X_p for some t_0 that is typically only moderately large with respect to p . In other words, the initial conditions are quickly “forgotten” in the subsequent evolution of the $\{X_t\}$ series; the effect of the initial conditions is minimal on the $\{X_t, t > t_0\}$ random variables, and the same is true for the random variables $\{W_{t,a}, t > t_0 + p\}$.

10. See e.g. the Yen/Dollar Simple NoVaS correlogram in Figure 8 (c).

11. Note that the ratio in eq. (19) is always positive and finite since its denominator is bigger than zero by eq. (6). Because of the approximate nature of obtaining the predictive residuals, the same is not necessarily true for the denominator of eq. (21). However, the sample median is robust against such anomalies and would trim away negative values and/or infinities of the ratio in eq. (21).

12. As before, if c_k is such that (9) is not satisfied, then decrease it stepwise over its discrete grid until (9) is satisfied.

13. A value of N/A indicates that kurtosis matching was not possible for the given value of α . Note, however, that the values in Table 6 were computed using the whole available sample sizes, whereas the results of Table 5 were computed using subsets of the data for the purpose of “honest” predictions. That is how and why the two entries that are reported as N/A in Table 5, are successfully computed for the purposes of Table 6.

APPENDIX

In order to confirm the findings of the paper, the analysis was repeated over all remaining financial datasets available in the `garch` module of S+, namely the New York Stock Exchange Data corresponding to the companies: HP, AMOCO, FORD, and MERCK. As before, those were daily data from February 2, 1984 to December 31, 1991, with a sample size of 2000.

Analogues of Tables 2, 3, 4, 5, and 6 are given below for the additional four datasets. The analysis of these new datasets gives qualitatively similar results to those reported concerning the Yen/Dollar, S&P 500 and IBM data, and thus leads to similar conclusions.

Predictor type	HP	AMOCO	FORD	MERCK
Exponential Smoothing with CV	1.234	1.283	1.085	1.174
Eq. (10)—AR fit with AIC	1.029	1.017	0.983	1.020
Eq. (2)—GARCH(1,1) with normal errors	1.224	1.275	1.078	1.097
Eq. (2)—GARCH(1,1) with t -errors	1.256	1.276	1.091	1.120

Table 2*: Entries give the empirical Mean Squared Error (MSE) of prediction of squared returns relative to benchmark; note that the MSE of prediction achieved by the benchmark was $2.07\text{e-}006$, $1.44\text{e-}006$, $1.07\text{e-}006$, $3.54\text{e-}007$ in the four cases HP, AMOCO, FORD, MERCK respectively.

Predictor type	HP	AMOCO	FORD	MERCK
Exponential Smoothing with CV	1.077	1.054	0.998	1.085
Eq. (10)—AR fit with AIC	1.020	0.980	0.972	1.041
Eq. (2)—GARCH(1,1) with normal errors	1.037	0.999	0.950	1.043
Eq. (2)—GARCH(1,1) with t -errors	1.010	0.892	0.881	0.932

Table 3*a: Entries give the empirical Mean Absolute Deviation (MAD) of prediction of squared returns relative to benchmark; note that the MAD of prediction achieved by the benchmark was $5.51\text{e-}004$, $2.55\text{e-}004$, $3.76\text{e-}004$, $2.29\text{e-}004$ in the four cases HP, AMOCO, FORD, MERCK respectively.

Predictor type	HP	AMOCO	FORD	MERCK
Eq. (11)—GARCH(1,1) with normal errors	0.895	0.855	0.834	0.901
Eq. (11)—GARCH(1,1) with t -errors	0.927	0.846	0.837	0.898

Table 3*b: Entries give the empirical Mean Absolute Deviation (MAD) of prediction of squared returns relative to benchmark.

Predictor type	HP	AMOCO	FORD	MERCK
Simple NoVaS	0.863	0.860	0.812	0.862
Exponential NoVaS	0.872	0.868	0.890	0.871

Table 4*: Entries give the empirical Mean Absolute Deviation (MAD) of prediction of squared returns relative to benchmark.

α	HP	AMOCO	FORD	MERCK
0.0	0.872	0.868	0.890	0.871
0.1	0.860	0.837	0.799	0.860
0.2	0.851	0.842	0.796	0.857
0.3	0.845	0.836	0.792	0.853
0.4	0.842	0.835	0.798	0.852
0.5	0.836	0.836	0.804	0.850

Table 5*: Entries give the empirical Mean Absolute Deviation (MAD) of prediction of squared returns relative to benchmark using the General Exponential NoVaS with parameter α ; the minimum MAD is given with boldface.

α	HP	AMOCO	FORD	MERCK
0.0	0.0625	0.0675	0.0500	0.0650
0.1	0.0700	0.0800	0.0625	0.0775
0.2	0.0825	0.0925	0.0725	0.0925
0.3	0.1050	0.1175	0.0975	0.1200
0.4	0.1475	0.1650	0.1350	0.1700
0.5	0.2275	0.2525	0.2175	0.2750

Table 6*: Entries give the optimal exponent c from kurtosis matching in the General Exponential NoVaS with parameter α ; the values in this table were computed using the whole available sample sizes.