Model-free Volatility Prediction

Dimitris N. Politis*
Department of Mathematics,
University of California at San Diego,
La Jolla, CA 92093-0112, USA
email: dpolitis@ucsd.edu

June 19, 2003

*Research partially supported by NSF grant DMS-01-04059.

Model-free Volatility Prediction

Abstract

The well-known ARCH/GARCH models with normal errors account only partly for the degree of heavy tails empirically found in the distribution of financial returns series. Instead of resorting to an arbitrary nonnormal distribution for the ARCH/GARCH residuals we propose a different viewpoint via a novel normalizing and variance–stabilizing transformation (NoVaS, for short) that can be seen as an alternative to parametric modelling. Some properties of this transformation are discussed, and algorithms for optimizing it are given. Special emphasis is given on the problem of volatility prediction and the issue of a proper measure for quality of prediction. A new prediction algorithm with favorable performance is given based on the NoVaS transformation. For motivation and illustration of this new general methodology, the NoVaS transformation is implemented in connection with three real data series: a foreign exchange series (Yen vs. Dollar), a stock index series (the S&P500 index), and a stock price series (IBM).

1 Introduction

Consider data $X_1, \ldots, X_n$ 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:

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

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

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

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 kyrtosis 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, non-parametric 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 for financial returns series is defined, and its properties are analyzed. Section 3 is devoted to the interesting (and quite challenging) problem of volatility prediction while Section 4 contains some conclusions.

For motivation and illustration throughout the paper we consider three datasets of daily returns taken from a foreign exchange rate, a stock price, and a stock index; a description of the 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.

2 Normalization and variance-stabilization

2.1 Definition of the NoVaS transformation

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

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

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 (causal) estimate of the standard deviation of X_t. Hence, we may define the new 'studentized' quantity

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

in the above, s_{t-1}^2 is an estimator of $\sigma_X^2 = Var(X_1)$ based on the data up to (but not including[1]) 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 transformation[2] (NoVaS, for short) under which the data series $\{X_t\}$ is

[1] The reason for not including time t in the variance estimator is just for notational clarity: we want to isolate and identify the effect of the coefficient a_0 associated with X_t^2 in the denominator of equation (4).

[2] 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.

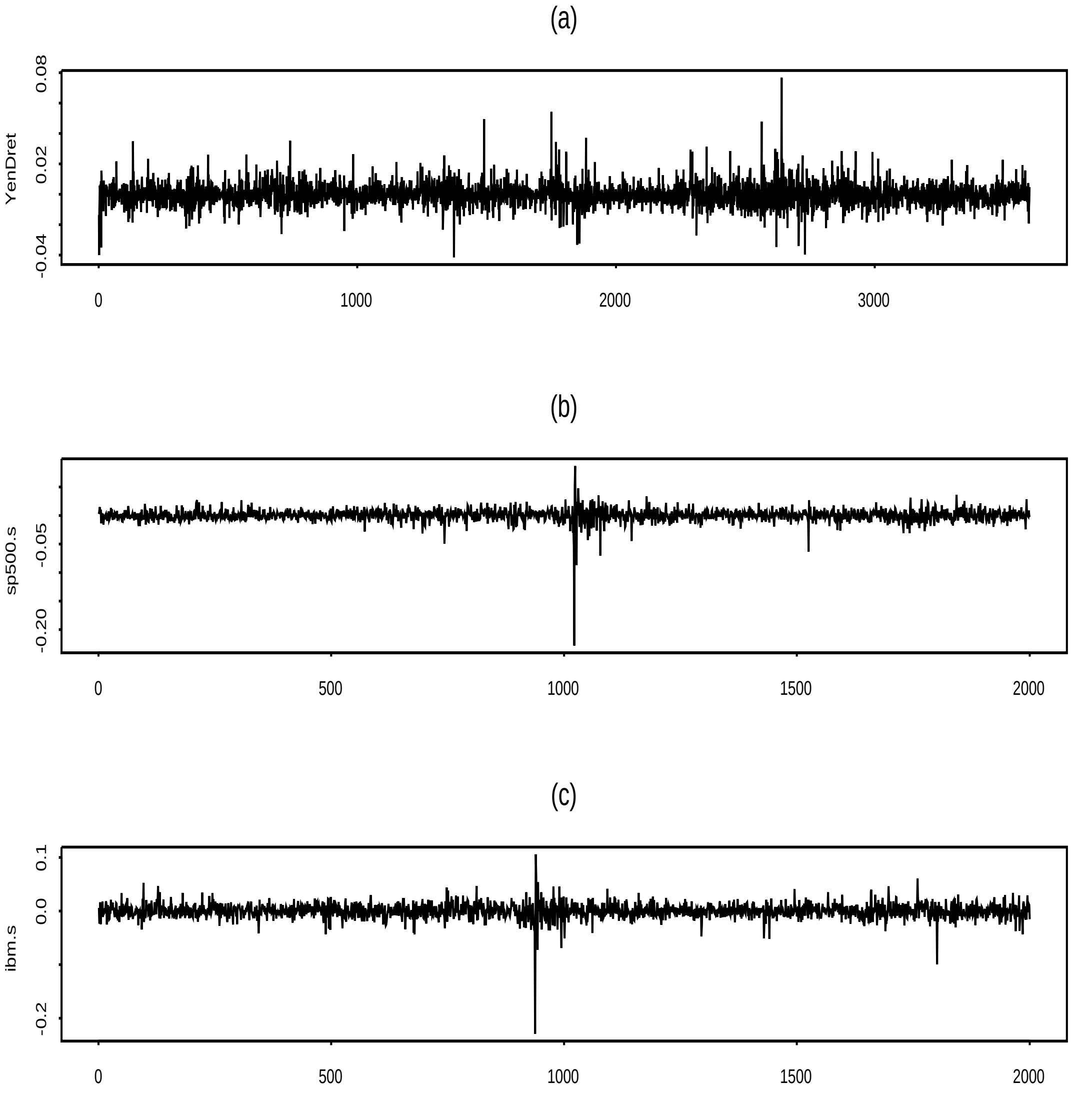

Figure 1: (a) Plot of the daily Yen/Dollar returns from December 31, 1987 up to August 1, 2002; (b) Plot of the daily S&P500 stock index returns from October 1, 1983 to August 30, 1991; (c) Plot of the daily returns of the IBM stock price from February 1, 1984 to December 31, 1991.

mapped to the new series $\{W_{t,a}\}$. The order $p(\geq 0)$ and the vector of nonnegative parameters $(\alpha, a_0, \ldots, a_p)$ 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:

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

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 $\{W_{t,a}, t = p+1, p+2, \cdots\}$ 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,

$$|W_{t,a}| \leq 1/\sqrt{a_0} \qquad (6)$$

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.

2.2 Choosing the parameters of NoVaS

In choosing the order p (≥ 0) and the parameters $\alpha, a_0, \ldots, a_p$ the twin goals of normalization and variance–stabilization of the transformed series $\{W_{t,a}\}$

are first taken into account. Secondarily, 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 3. 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

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

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 $\alpha, a_0, \ldots, a_p$ 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:

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

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 2.4. 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 $\alpha, a_0, a_1, \ldots, a_p$ 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 kyrtosis (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

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

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 $\sqrt{2\ln n}$.

2.3 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 kyrtosis 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 $KYRT_n(Y)$ denote the empirical kyrtosis of data $\{Y_t, t = 1, \ldots, n\}$, i.e.,

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

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 kyrtosis 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 kyrtosis for concreteness.

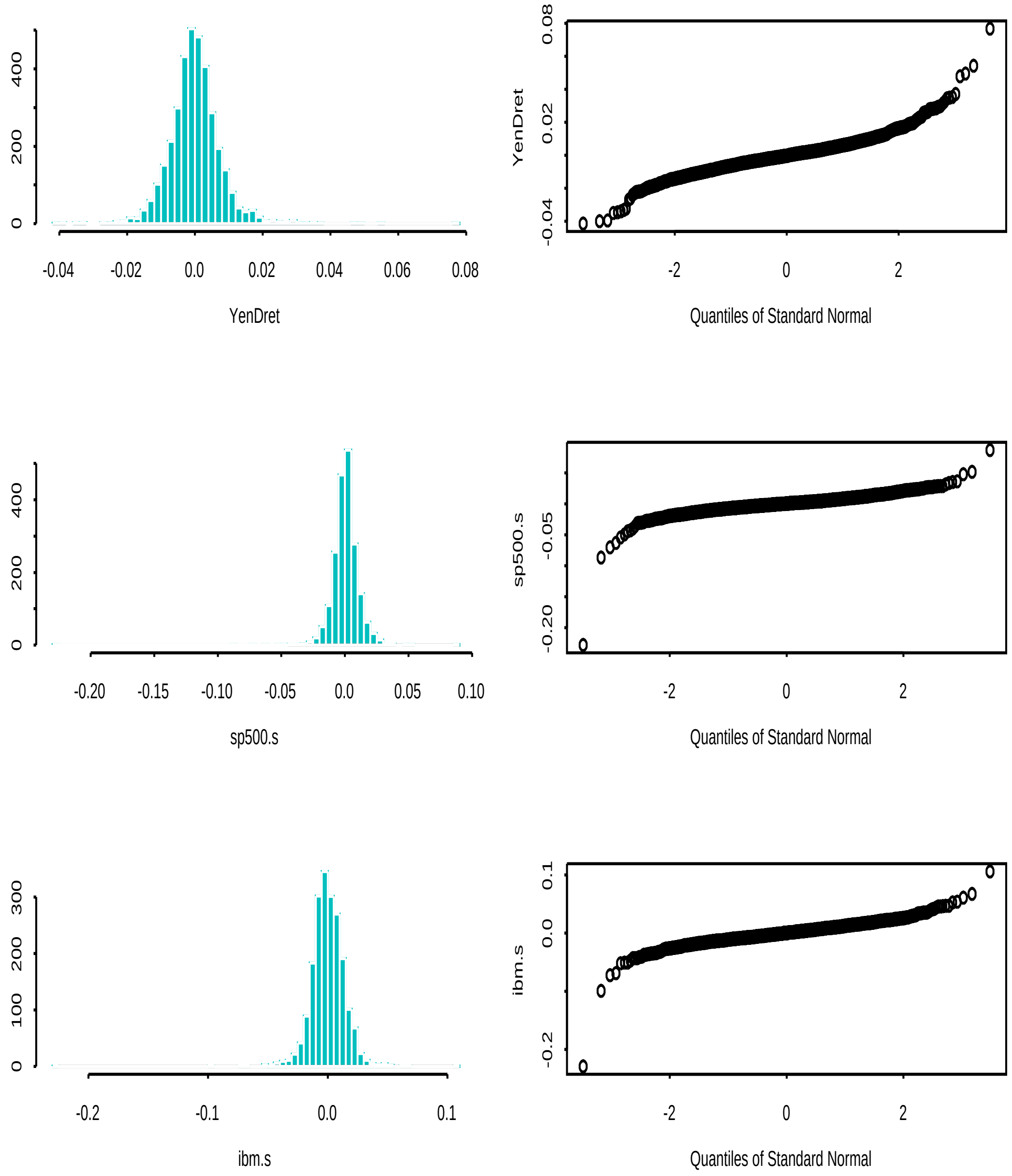

Figure 2: Histograms and Q-Q plots for the three returns series of Figure 1.

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 $|KYRT_n(W_{t,p}^S) - 3|$ is minimized.

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 $KYRT_n(W_{t,p}^S) \simeq 3$,

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

To see that the moment matching goal is a feasible one, note first that for $p = 0$ we have $a_0 = 1$, $W_{t,0}^S = sign(X_t)$, and $KYRT_n(W_{t,0}^S) = 1$. On the other hand, it is to be expected that for large p, $KYRT_n(W_{t,p}^S)$ will be bigger than 3. As a matter of fact, the law of large numbers implies that for increasing values of p, $KYRT_n(W_{t,p}^S)$ will tend to the 'true' kyrtosis of the random variable X_1 which is understood to be quite large (and may even be infinite—see the discussion in Section 3.1). Therefore, viewing $KYRT_n(W_{t,p}^S)$ as a (smooth) function of p, one would expect that for an intermediate value of p the level 3 would 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 $KYRT_n(W_{t,p}^S)$ for $p = 1, 2, \cdots$, stopping when $KYRT_n(W_{t,p}^S)$ first hits or just passes the value 3. Interestingly, $KYRT_n(W_{t,p}^S)$ 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-adjustement' 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

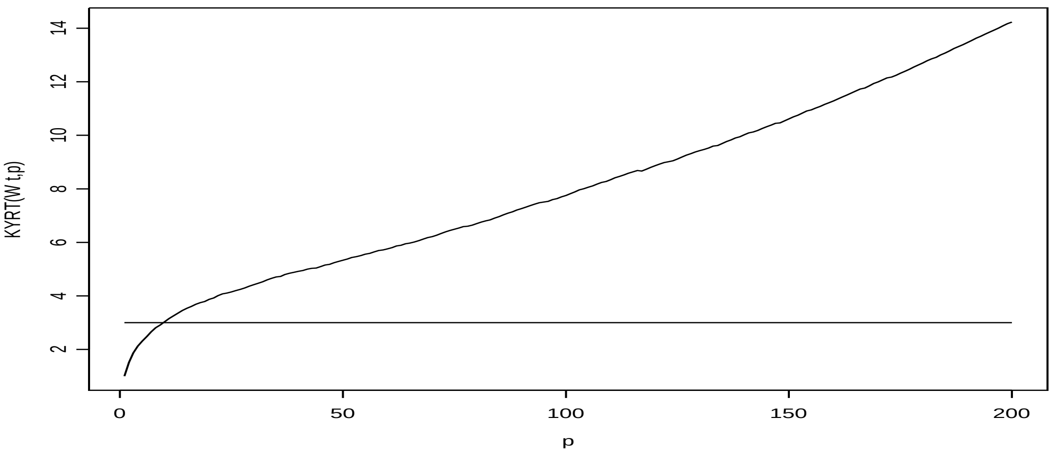

Figure 3: Illustration of the simple NoVaS algorithm for the Yen/Dollar dataset: plot of $KYRT_n(W_{t,p}^S)$ as a function of p; the solid line indicates the Gaussian kyrtosis of 3.

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-adjustement should be used with restraint, that is, the choice of C in (9) should be reasonably small, as it effectively over-rides the data-dependent character of choosing p; both concrete suggestions, i.e., $C = 3$ or $C \simeq \sqrt{2 \ln n}$ seem to work well in practice.

In Figure 3, an illustration of the simple NoVaS algorithm is given for the Yen/Dollar dataset. Figure 3 shows a plot of $KYRT_n(W_{t,p}^S)$ as a function of p; the monotonic increase of $KYRT_n(W_{t,p}^S)$ is apparent, rendering the NoVaS algorithm easy to implement. Notably, $KYRT_n(W_{t,p}^S)$ is closest to 3 for $p = 9$; actually, $KYRT_n(W_{t,9}^S) = 3.03$. 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 $n = 3600$.

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: 12 for the IBM dataset, and 10 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" kyrtosis to account for.

Focusing again on the Yen/Dollar dataset, it should be noted that with $p = 9$ the effective range of the Yen/Dollar NoVaS transform $\{W_{t,9}^S\}$ series is 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-adjustement step would be required leading to the choice $p = 15$; note that $KYRT_n(W_{t,15}^S) = 3.51$ which is still quite close to the target value of 3. As a matter of fact, a Q-Q plot (not shown) of the simple NoVaS Yen/Dollar $\{W_{t,15}^S\}$ series actually looks even closer to normal than the Q-Q plot of $\{W_{t,9}^S\}$ shown in Figure 5 in terms of showing less "clipping" near the upper right corner. The higher values of p in connection with the IBM and S&P500 datasets correspond to ranges of about 3.6 and 3.3 respectively, indicating even less of a need for possible range-adjustement, especially in view of their smaller sample size as well.

Remark 2.1 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 $KYRT_n(W_{t,p}^S) \simeq 3$; 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 kyrtosis 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

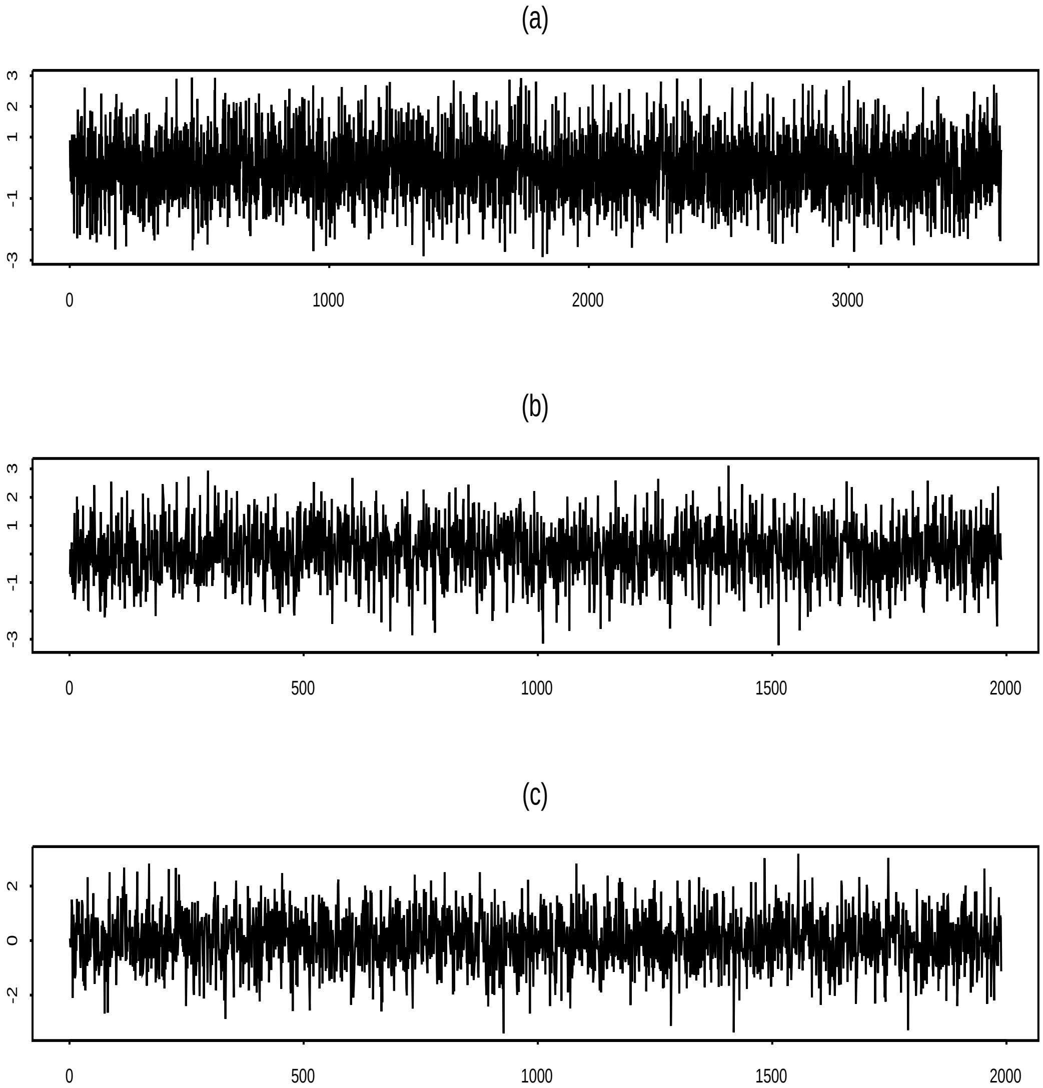

Figure 4: Plots of the Simple NoVaS transformed series corresponding to the three datasets of Figure 1.

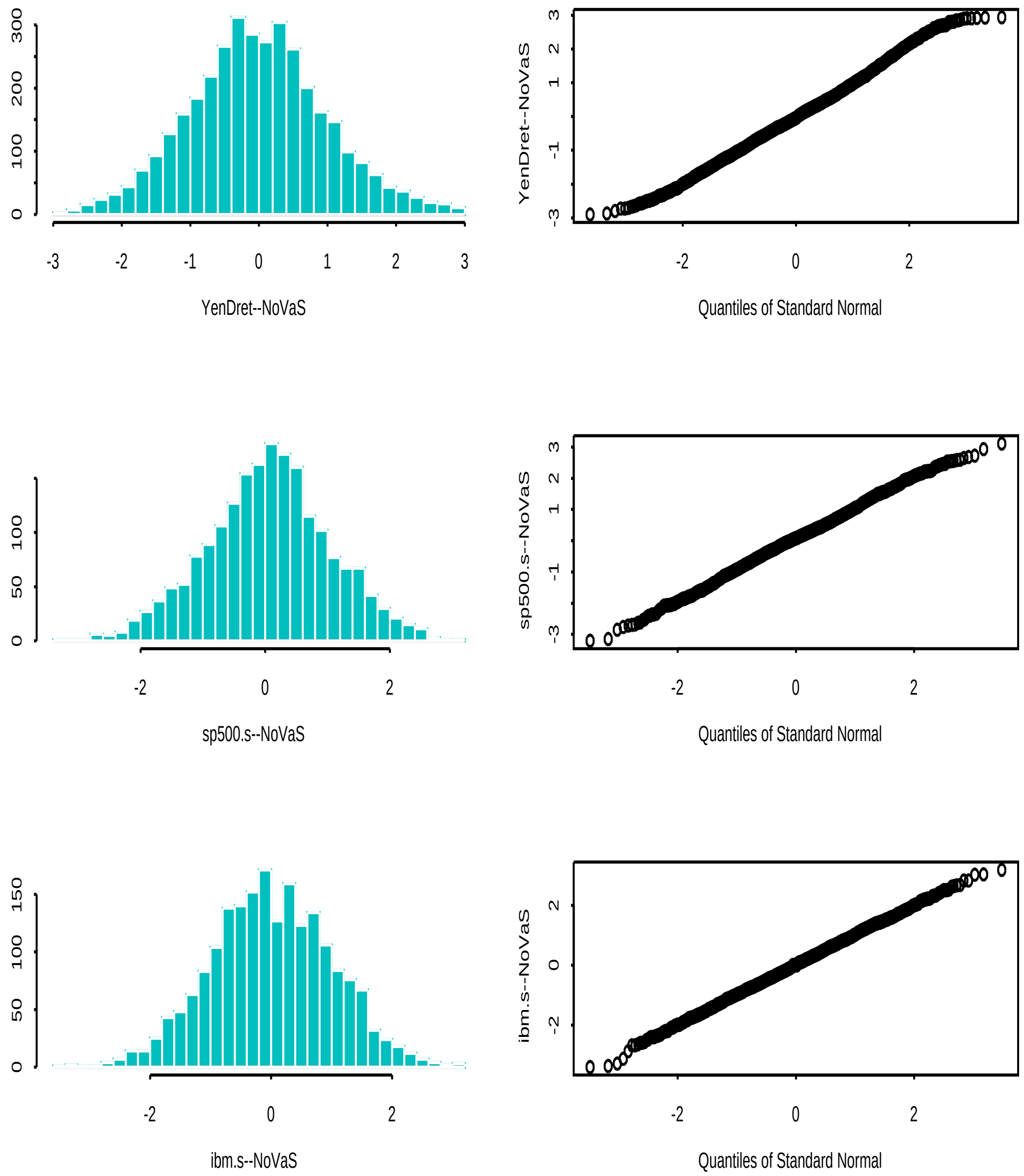

Figure 5: Histograms and Q-Q plots for the three NoVaS series of Figure 4.

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 $(W^S_{t,9}, W^S_{t-1,9})$.

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

Table 1: (Yen/Dollar example) Sample kyrtosis of $\tilde{W}^S_{t,9,i} = W^S_{t,9} + \lambda_i W^S_{t-1,9}$ 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 $W^S_{t,p}$ and $W^S_{u,p}$. The Cramér-Wold device suggests that we simultaneously consider some linear combinations of the type:

$$\tilde{W}^S_{t,p,i} = W^S_{t,p} + \lambda_i W^S_{u,p} \quad \text{for} \quad i = 1, \ldots, K,$$

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

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

2.4 Exponential NoVaS algorithm

In the Exponential NoVaS, to specify all the a_is, 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 $W^E_{t,c}$.

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 $|KYRT_n(W_{t,c}^E) - 3|$ 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-adjustement 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.

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_is so that their sum (for $i = 0, 1, \ldots, i_0$) equals one.

An illustration of the Exponential NoVaS algorithm is now given for the Yen/Dollar dataset. Figure 6 is a plot of $KYRT_n(W_{t,c}^E)$ as a function of c. Except for values of c very close to zero, $KYRT_n(W_{t,c}^E)$ seems to be monotonically decreasing hitting the value 3 for $c \simeq 0.0985$. Nevertheless, the behavior of $KYRT_n(W_{t,c}^E)$ 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 $KYRT_n(W_{t,c}^E)$ will be decreasing for all values of c.

To further elaborate, note that Figure 6 indicates $KYRT_n(W_{t,c}^E)$ 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 $i = 1, \ldots, p$ for the two values of c suggested by Figure 6; due to the truncation effect with ϵ= 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

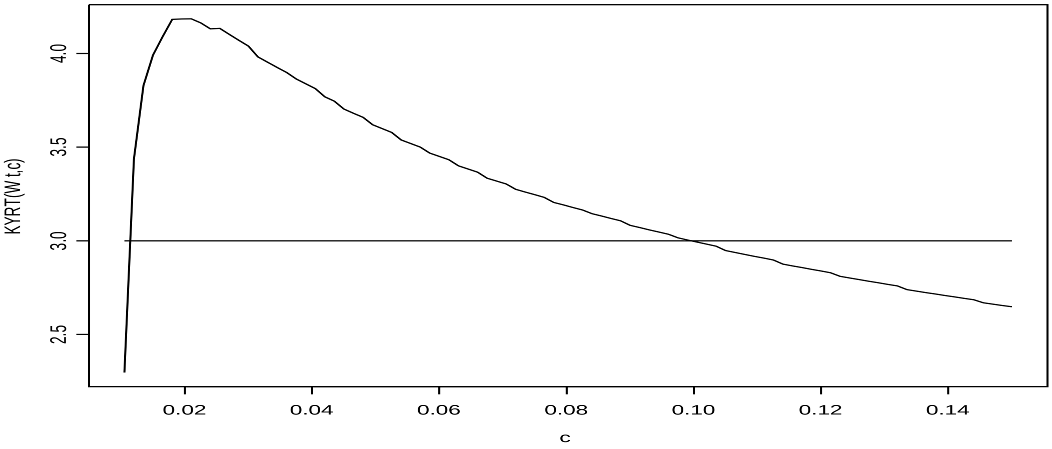

Figure 6: Illustration of the Exponential NoVaS algorithm for the Yen/Dollar dataset: plot of $KYRT_n(W^E_{t,c})$ as a function of c; the solid line indicates the Gaussian kyrtosis of 3.

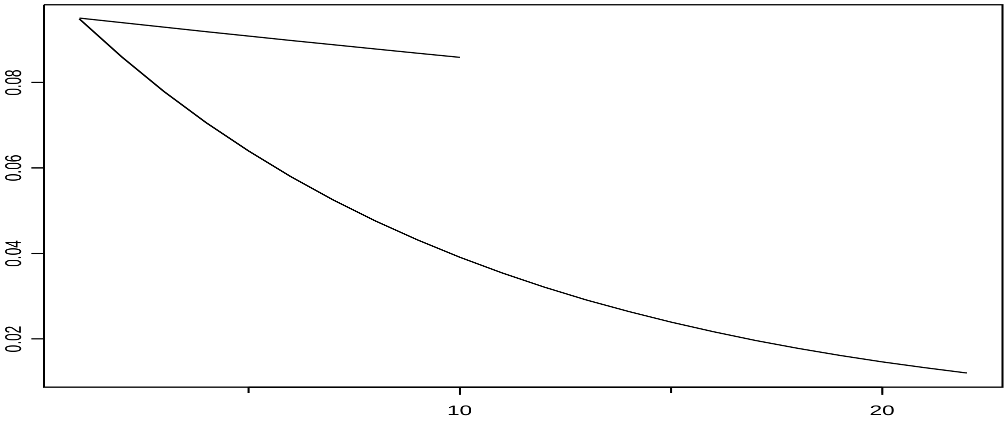

Figure 7: Plot of the exponential coefficients a_i versus the index $i = 1, \ldots, p$ for the two values of c suggested by Figure 6; note that $c \simeq 0.0113$ corresponds to $p = 10$, while $c \simeq 0.0985$ corresponds to $p = 22$.

$p = 10$; this is because the exponential coefficients decay so slowly that are close to being constant for $i = 1, \ldots, p$.

To sum up: a plot with shape such as Figure 6 is typical when a nonzero ϵ is used, suggesting that the function $|KYRT_n(W_{t,c}^E) - 3|$ 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.

Analogs 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 $a_i = c'e^{-ci} + d'e^{-di} \cdots$. The generalization may well include higher order moment matching and/or looking at linear combinations of higher order lags.

3 Volatility prediction

3.1 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.

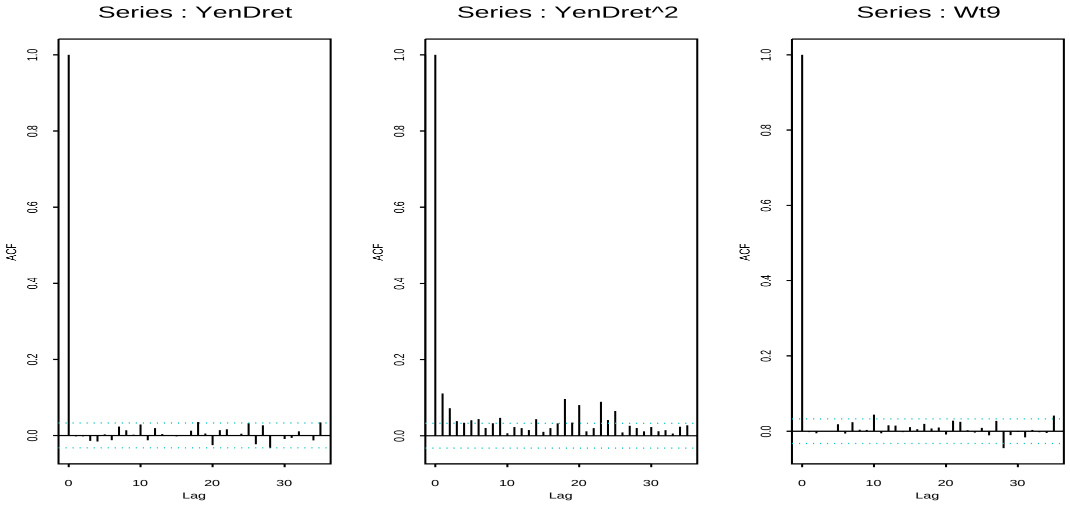

Figure 8: (Yen/Dollar example) (a) Correlogram of the returns series $\{X_t\}$; (b) Correlogram of the squared returns $\{X_t^2\}$; (c) Correlogram of the optimal Simple NoVaS series $\{W_{t,9}^S\}$.

However, the squared returns appear quite 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 predicting X_{n+1}^2 by a linear predictor of the type

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

here the b_i coefficients can be estimated by fitting an AR(r) model to the (de-meaned) 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 hardly 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 an AR(1) or maybe an ARMA(1,1) model in this situation.

Notably, 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).

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.

Predictor type	Yen/Dollar	S&P500	IBM
Eq. (10) with AIC	0.971	1.125	1.108
Eq. (2)—GARCH(1,1) with normal errors	1.005	1.164	1.140
Eq. (2)—GARCH(1,1) with t–errors	1.025	1.151	1.139

Table 2: Entries give the empirical Mean Squared Error (MSE) of prediction of squared returns relative to benchmark.

Table 2 focuses on the L_2 prediction performance of the three aforementioned predictors, namely 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 Splus. It is apparent that the performance of all three methods is rather abysmal as they seem to perform worse even than the (naive) benchmark. Due to 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".[3]

Nevertheless, the entries of Table 3a on the L_1 prediction performance tell a different story, namely that all three predictors outperform the benchmark when errors are measured in the L_1 norm. As intuitively expected, the GARCH with t–errors has the best performance among the three. Surprisingly, the GARCH with normal errors appears inferior to the linear model (10); the reason for this will be revealed shortly.

Predictor type	Yen/Dollar	S&P500	IBM
Eq. (10) with AIC	0.963	0.912	0.941
Eq. (2)—GARCH(1,1) with normal errors	0.971	0.982	0.980
Eq. (2)—GARCH(1,1) with t–errors	0.821	0.818	0.864

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

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 $KYRT_k(Y)$ denote the empirical (sample) variance and kyrtosis of dataset Y up to time k, i.e., $\{Y_1, \ldots, Y_k\}$. 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, $KYRT_k(Y)$ should tend to the kyrtosis of Y_1 be that infinite or not. Thus, plotting $VAR_k(Y)$ and $KYRT_k(Y)$ 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

[3] In turn, Andersen and Bollerslev (1998) define a 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.

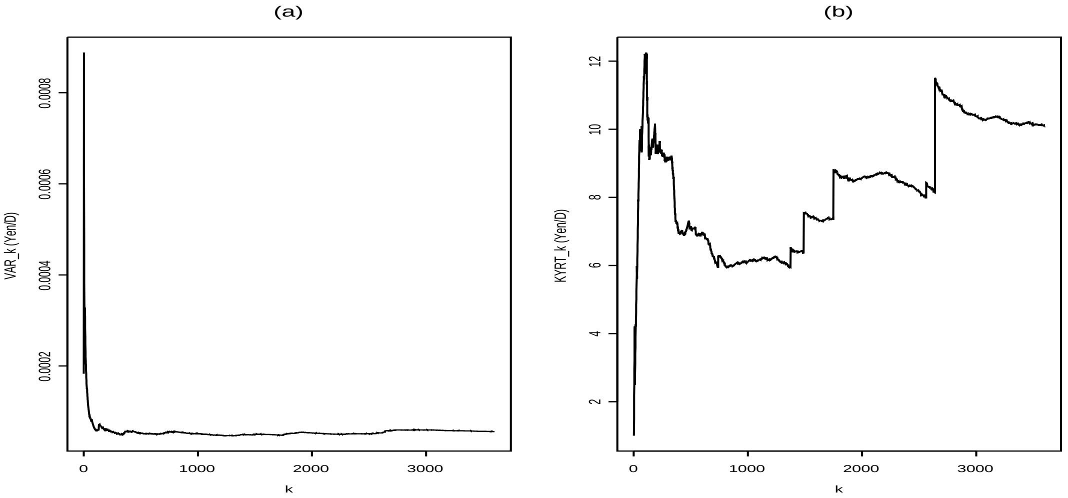

Figure 9: (Yen/Dollar example) (a) Plot of $VAR_k(X)$ as a function of k; (b) Plot of $KYRT_k(X)$ as a function of k.

plot in Figure 9 (a) seems to converge. Nevertheless, it seems that it has as 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.

Nevertheless, this is not the end of the story. To see why, note that the GARCH predictions for Tables 2 and 3a were performed—as customary—using the predictor (2). But eq. (2) gives the conditional 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. 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 predictor of X_{n+1}^2 in the L_1 sense is given by

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

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 (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	0.805	0.817	0.829
Eq. (11)—GARCH(1,1) with t–errors	0.793	0.799	0.831

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

Using the correct predictor leads, of course, 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. Furthermore, the GARCH(1,1) model with t–errors outperforms the benchmark by a nontrivial margin of the order of 20%.

Thus, by contrast to what is widely believed, ARCH/GARCH models *do* have predictive validity for the squared returns; this is particularly true for the GARCH with t–errors as expected. 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.

3.2 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 $\alpha, a_0, \ldots, a_p$ of NoVaS have already been chosen.

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

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

and

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

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

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

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.,

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

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

$$\widehat{X^2_{n+1}} = \mu_2 A_n^2 \tag{16}$$

where

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

Now observe that—up to the effect of initial conditions—the information set $\mathcal{F}_n = \{X_t, 1 \leq t \leq n\}$ is approximately[4] 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

$$\mu_2 \approx Median\left(\frac{W^2_{n+1,a}}{1 - a_0 W^2_{n+1,a}} | \tilde{\mathcal{F}}_n\right) \tag{17}$$

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 2, the joint distributions of the series $\{W_{t,a}, t = p+1, \ldots, n\}$ are also typically normalized by the NoVaS transformation. Thus, the series $\{W_{t,a}, t = p+1, \ldots, n\}$ 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)

[4] Note that the information set $\mathcal{F}_n$ is exactly equivalent to the information set $\{X_1, \ldots, X_p, W_{p+1,a}, W_{p+2,a}, \ldots, W_{n,a}\}$. Due to the stationarity and subject 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" $X_1, \ldots, X_p$ 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 $\{W_{t,a}, t > t_0 + p\}$.

approximately given by

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

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

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^2_{pred}—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 $W^2/(1 - a_0 W^2)$. 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 $\{W_{t,a}, t = p + 1, \ldots, n\}$ indicates no significant correlations—as is typically the case in practice[5]—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

$$\hat{\mu}_2 = median\{\frac{W^2_{t,a}}{1 - a_0 W^2_{t,a}};\ t = p+1, p+2, \ldots, n\} \tag{19}$$

and subsequently predict X^2_{n+1} by

$$\hat{\mu}_2 A^2_n. \tag{20}$$

[5] See, for example, the Yen/Dollar Simple NoVaS correlogram in Figure 8 (c).

Remark 3.1 Although in our examples the NoVaS series $\{W_{t,a}, t = p+1, \ldots, n\}$ 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 $t = r+1, \ldots, n$ 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 $\{e_t + \hat{W}_{n+1,a};\ t = r+1, \ldots, n\}$, 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 by[6]

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

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

Remark 3.2 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

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

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

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

[6] 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).

whereas under Case II the estimator becomes

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

.

Remark 3.3 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+h}^2 given $\mathcal{F}_n$, or in general predicting $g(X_{n+h})$ given $\mathcal{F}_n$, for some $h \geq 1$, can be handled in a similar vein; the details are omitted.

3.3 Optimizing NoVaS for volatility prediction

In section 3.2, 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 is apparent that the Simple NoVaS performs comparably to the optimal GARCH(1,1) with t–errors of Table 3b. In turn, the Exponential NoVaS performs uniformly better than the Simple NoVaS although they were both equally successful in normalizing/variance stabilizing the data.

The Exponential NoVaS is the best performing method among all previously considered candidates; for example, it outperforms the GARCH(1,1) with t–errors of Table 3b by a margin ranging from 1 to 5%.

Predictor type	Yen/Dollar	S&P500	IBM
Simple NoVaS	0.800	0.764	0.834
Exponential NoVaS	0.787	0.754	0.820

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 so well 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 Exponenential 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 kyrtosis matching ideas of Section 2. 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 kyrtosis 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 multi-parameter idea by allowing for the possibility of a nonzero value for the parameter α in (4) in connection with the best method so far, i.e., the Exponential NoVaS. We thus define the General Exponential NoVaS that has two free parameters, α and c, and will be denoted by $W^{GE}_{t;c,\alpha}$. The search is performed using a grid of possible values for α, say $\alpha_1, \alpha_2, \ldots, \alpha_K$; in picking the grid values, note that the kyrtosis matching goal may only be possible with small values of α.

Algorithm for General Exponential NoVaS:

- For $k = 1, \ldots, K$ 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 $|KYRT_n(W^{GE}_{t;c,\alpha_k}) - 3|$ is minimized, and denote by c_k the minimizing value.[7]
  - 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_is so that their sum (for $i = 0, 1, \ldots, p_k$) equals $1 - \alpha_k$ by eq. (7).
- Finally, compare the models $W^{GE}_{t;c_k,\alpha_k}$, for $k = 1, \ldots, K$, in terms of their volatility prediction performance, and pick the model with optimal such performance.

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

An illustration of the General Exponential NoVaS Algorithm in connection with our three datasets is presented in Table 5, where for each different value of α, the L_1 volatility prediction performance is given together with its corresponding optimal exponent c (in parentheses).

The results of 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 at all sensitive to the parameter α; the S&P500 index is more sensitive, while the single stock price (IBM) is most sensitive.

Note that, as α increases, c increases accordingly, and p decreases (the latter is not shown as it can be easily calculated). Table 5 was compiled using tolerance level $\epsilon = 0.01$; with that value, the optimal General Exponential NoVaS for the IBM dataset has $\alpha = 0.60,\ c = 0.580$ and $p = 4$ by contrast to the $p = 27$ that is associated with $\alpha = 0$. The extreme case where $\alpha = 0.7$ for the IBM dataset corresponds to $p = 0$, i.e., a model with no exponential term—just α and a_0 in the denominator of NoVaS. The interpretation is that, at least for the stocks datasets, it may be beneficial to use a very local (high c, low p) exponential, i.e., concentrating on just the last 3-4 days of data, paired with a relatively large value of α.

Finally, note that all (c, α) combinations in Table 5 are equally succesful in normalizing the NoVaS transformation in terms of achieving a kyrtosis 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 kyrtosis matching to be successful.

As a conclusion, recall that the Exponential NoVaS (with $\alpha = 0$) yielded a 1-5% improvement over the GARCH(1,1) with t–errors in terms of L_1 volatility prediction performance; see Tables 3b and 4. The introduction of a nonzero α in the General Exponential NoVaS yields only a small improvement over the Exponential NoVaS in the Yen/Dollar dataset but does yield appreciable improvements in the two stock price datasets, S&P500 and IBM. All in all, the General Exponential NoVaS is seen to outperform the GARCH(1,1) with t–errors by the margins 1.25%, 8.75%, and 5.25% for our three datasets Yen/Dollar, S&P500, and IBM respectively.

α	Yen/Dollar	S&P500	IBM
0.00	0.787	0.754	0.820
	(0.098)	(0.084)	(0.069)
0.05	0.786	0.750	0.815
	(0.109)	(0.095)	(0.075)
0.10	0.785	0.746	0.811
	(0.120)	(0.108)	(0.080)
0.20	0.785	0.739	0.806
	(0.140)	(0.135)	(0.098)
0.30	0.784	0.733	0.803
	(0.183)	(0.195)	(0.127)
0.40	**0.783**	**0.730**	0.797
	(0.260)	(0.300)	(0.180)
0.50	0.787	0.733	0.789
	(0.410)	(0.520)	(0.290)
0.60	0.787	N/A	**0.787**
	(0.813)	(—)	(0.580)
0.65	N/A	N/A	0.788
	(—)	(—)	(0.990)
0.70	N/A	N/A	0.796
	(—)	(—)	(2.740)
> 0.7	N/A	N/A	N/A
	(—)	(—)	(—)

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; below each entry in parentheses is the optimal exponent c from kyrtosis matching.

4 Conclusions

In this paper, a new methodology was introduced for dealing with financial returns data. By contrast to the customary viewpoint that is based on parametric and/or semi-parametris models (such as ARCH/GARCH) the new approach is totally nonparametric. This model-free methodology has at its core a novel normalizing and variance–stabilizing transformation (NoVaS, for short). For motivation and illustration of this new general methodology, the NoVaS transformation is implemented in connection with three real data series: 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 the returns data may not have a finite fourth moment in which case L_2 methods—such as conditional expectations—are inappropriate for the squared returns. Contrary to wide-spread beliefs, we show that ARCH/GARCH models actually do have predictive validity for the squared returns when applied properly, i.e., in an L_1 setting.

An algorithm for prediction of a general function $g(X_{n+1})$ given the data $X_1, \ldots, X_n$ was devised based on the NoVaS transformation. Finally, with some simple and intuitive choices for the NoVaS structure, e.g. Exponential or General Exponential NoVaS, it was shown that the prediction algorithm based on NoVaS empirically outperforms the popular 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] 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.

[3] Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity, *J. Econometrics*, 31, 307-327.

[4] Bollerslev, T., Chou, R. and Kroner, K. (1992). ARCH modelling in finance: a review of theory and empirical evidence, *J. Econometrics*, 52, 5-60.

[5] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods, 2nd ed.*, Springer, New York.

[6] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation, *Econometrica*, 50, 987-1008.

[7] Fama, E.F. (1965). The behaviour of stock market prices, *J. Business*, 38, 34-105.

[8] Gouriéroux, C. (1997). *ARCH Models and Financial Applications*, Springer Verlag, New York.

[9] Hamilton, J.D. (1994). *Time Series Analysis*, Princeton Univ. Press, Princeton, New Jersey.

[10] Mandelbrot, B. (1963). The variation of certain speculative prices, *J. Business*, 36, 394-419.

[11] Nelson, D. (1991). Conditional heteroscedasticity in asset returns: a new approach, *Econometrica*, 59, 347-370.

[12] 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.

[13] 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.