# Realistic Processes for Stocks from one Day to one Year

**Gilles Zumbach, Luis Fernández, Caroline Weber**

Swissquote Bank
Ch. de la Crétaux 33
Case Postale 319
CH-1196 Gland, Switzerland
www.swissquote.ch

December 21, 2010

**Abstract**

A realistic ARCH process is set so as to duplicate for all practical purposes the properties of stock time series from 1 day to 1 year. The process includes heteroskedasticity with long memory, leverage, fat-tail innovations, relative return, price granularity, and holidays. Its adequacy to describe empirical data is controlled over a broad panel of statistics, including (robust L-statistics) skew, (robust) kurtosis, shape factor for the volatility distribution, and lagged correlations between combinations of return and volatility. These statistics are computed for returns and volatilities with characteristic time intervals ranging from 1 day to 1 year. This wide cross-check between stock time series and simulations ensures that the most important features of the data are correctly captured by the process up to 1 year. The by-products of the statistical analyses and estimations are 1) a positive skew, 2) a cross-sectional relation between kurtosis and heteroskedasticity, 3) the very similar cross-sectional distribution for the statistics evaluated over the empirical data set or for the process with one set of parameters and 4) the heteroskedasticity is very close to an integrated volatility process.

# 1 Introduction

Finance is relying increasingly on Monte Carlo simulations. The reasons are many, including for example the greater complexity of some contracts, American optionality, the large size of portfolios, the inclusion of fat-tails and heteroskedasticity, and the availability of cheap computational power. Even options that have analytical solutions are often computed using Monte Carlo simulations in order to ensure a consistent pricing for all positions across a portfolio. As the goal of analytical tractability is abandoned, processes more accurate than a constant volatility Gaussian random walk can be used. Our goal in this paper is to set a process that can reproduce the major empirical stylized facts for stocks, for time intervals ranging from 1 day to 1 year. In particular, we want the process to capture the fat-tail distributions, the persistence of the kurtosis, the heteroskedasticity with its long memory, the leverage effect, the time reversal asymmetry, and the influence of holidays and price granularity. The process is a parsimonious extension of a long memory ARCH process, with one set of parameters that can reproduce the major statistics over the desired time intervals for a large cross-section of stocks. This realistic process will be used in a subsequent paper [] to price European options and to compute implied volatilities.

The description of financial time series by GARCH data generating processes has a long history, going back to the seminal works of [Engle, 1982, Engle and Bollerslev, 1986] and [Bollerslev, 1986]. There exist numerous extensions of the basic ideas in order to include more stylized facts. Interesting extensions relevant for the present work are the inclusion of fat-tailed innovations [Engle and Bollerslev, 1986], of the leverage [Glosten et al., 1993] and of the long memory for the heteroskedasticity [Granger and Ding, 1996, Zumbach, 2004]. Our approach differs from most of the literature, where the parameters are estimated using a log-likelihood maximization. The log-likelihood approach focuses on the daily scale and is weakly sensitive to long time properties of the data. Instead, the equations and parameters are tuned in this work so as to reproduce best the empirical statistics over a broad range of time horizons and stocks. This multi-scales approach allows us to obtain a data generating process that can duplicate financial time series for many practical purposes. Another difference is the use of relative returns instead of logarithmic returns. This choice follows an analysis of [O'Neil and Zumbach, 2009] showing that processes with fat-tailed innovations have well defined properties in this set-up (i.e. expectations are finite) whereas logarithmic random walks lead to infinite integrals.

In a first step, we analyse the empirical statistics over the desired time intervals. A broad panel of estimators is used which cover distributional properties (mean, variance, skew, kurtosis and other shape statistics) and dynamical properties (lagged correlation between combinations of return and volatility). Some care is required due to the fat-tail distributions of the returns, and robust estimators are used. Among the results, the skew of the relative return distributions is close to zero or positive. Another significant result is the relation between shape factors for the return distribution and the lagged correlation for the volatility. This relation shows that the shape of the return distribution is directly related to the dynamical properties of the volatility.

In a second step, an ARCH process is set so as to reproduce the empirical statistics from one day to one year. Beside the basic heteroskedasticity, several effects should be included in the equations in order to reach the desired level of accuracy. These effects include fat-tail innovations to reproduce the skewness, long memory aggregated ARCH process to reproduce the volatility dynamics, and the leverage effect to reproduce the correlation between return

and volatility. Moreover, the short term kurtosis and the rate of zero returns require to introduce a finite price granularity together with holidays.

Interestingly, a single set of parameters is able to reproduce the observed statistics. More precisely, the empirical cross-section for the statistics over a large set of stocks is reproduced by Monte Carlo simulations using one set of parameter values. This shows that the possibly large dispersions of empirical statistics could be due only to the random realizations, but the underlying process is fairly constrained in its analytical form and parameter values.

The plan of the subsequent sections follows roughly the line of the above introduction. The next section presents the wide set of statistics used to quantify the properties of the time series. These statistics are applied to the empirical stock time series in Sec. 3. The theoretical processes are presented in Sec. 4 and the most important statistics shown in Sec. 5. The conclusions summarize the findings.

## 2 Statistics for the return and volatility

The price time series is denoted with $p(t)$. The annualized *relative* return over $\Delta T = n$ days at time $t$ is defined by

$$r[\Delta T](t) = \sqrt{\frac{1 \text{ year}}{\Delta T}} \, \frac{p(t) - p(t - \Delta T)}{p(t - \Delta T)} \tag{1}$$

where 1 year = 260 days, $\Delta T = n\,\delta t$ and $\delta t = 1$ day is the time increment of the data. The pre-factor annualizes the returns, so that the leading random walk scaling is already absorbed into the definition. In this way, the variance of the returns is essentially independent of the time span $\Delta T$. In order to compare both definitions, a plot is presented for logarithmic returns, with $r_{\log}[\Delta T](t) = \sqrt{1 \text{ year}/\Delta T} \cdot \log(p(t)/p(t - \Delta T))$. The (historical) volatility at time $t$ and evaluated over $\Delta T = n$ days is defined by

$$\sigma^2[\Delta T](t) = \frac{1}{n} \sum_{t - \Delta T < t' \leq t} r^2 \, [1 \text{ day}] \, (t') \tag{2}$$

where $n$ is the number of terms in the sum. The sample size for the return is denoted by $N$ and the sample expectation of $x$ by $\langle x \rangle$. A few cross-sectional expectations are computed, and are denoted by $\overline{x}$.

The summary statistics and correlations are computed using a de-trended price time series. With relative returns, the appropriate definition is $p'(t) = p(t) \cdot (1 + \langle r \rangle)^{-j}$ where $(1 + \langle r \rangle)^{-j}$ is the discount factor with $\langle r \rangle$ the daily mean return and $j$ the number of days since the start of the sample $t_0$, namely $t = t_0 + j\,\delta t$. The annualized return computed from the de-trended prices $p'$ is denoted by $r'$ and is $r'[j](t) = \sqrt{\frac{1 \text{ year}}{j}} \, \frac{p(t)(1+\langle r \rangle)^{-j} - p(t-j)}{p(t-j)}$.

Distributional properties are usually summarized by the first reduced moments, in particular the skew and kurtosis. Yet, the distributions of returns are fat-tailed, with tail exponents of the order of 3 to 5. While the mean and variance are well defined, it is clearly dangerous to compute empirical moments up to order 4. Therefore, we want to define similar statistics in order to measure asymmetry and kurtosis, while having well defined theoretical values for such low tail exponents. For the empirical statistics, the gain of the modified definitions is to be less sensitive to extreme events.

Beside the usual reduced moment-estimators, we have investigated three families of more robust definitions for the location, size, skew and kurtosis. First, the abs-estimators use definitions with lower exponents and absolute value for the returns. For example, the abs-skew is defined as $\langle r' \, |r'| \rangle \, / \, \langle |r'| \rangle^2$. Second, the quantiles estimators are based on combinations of a few quantiles, typically quartiles and the median. For example, the quantile-skew is defined as $(r_{[0.75\,N]} - 2r_{[0.50\,N]} + r_{[0.25\,N]})/(r_{[0.75\,N]} - r_{[0.25\,N]})$ where $r_{[i]}$ is the $i$ ordered statistics of the total sample of size $N$. Third, L-statistics [Hosking, 1990] are defined as linear sample expectations of the ordered statistics. The L-moments of order $k$ are given by $l_k = \sum_i g_k(N,i)\, r_{[i]}$ and the weights $g_k(N,i)$ are defined with combinatorial coefficients. Because these moments are linear in the returns, the L-skew is defined as $l_3/l_2$.

An extensive cross-sectional comparison of the estimators over the time series in the data set shows the following. For all the moments, the L-estimators and abs-estimators give consistent values, the quantile-estimators can be inconsistent, while the usual moment-estimators can be very different. For the second moment (the size), no large difference between estimators is seen and it is therefore not very important to take one or another definitions. For the skew, the quantile-skew is very dependent on the chosen quantile, and the above 1/4-3/4 quantile above is negatively (!) related to the other skew measures. The L-skew and abs-skew are closely related, while the moment-skew is roughly related but with a very wide distribution. For the kurtosis, the L-kurtosis and abs-kurtosis are similar, the quantile-kurtosis is somewhat related, but the moment-kurtosis is weakly related with the others and with a huge dispersion. Clearly, the message is consistent with the known fat-tail distribution of the returns, namely it is dangerous to use the usual reduced moment estimators, and they are becoming worst for increasing exponents. The quantile estimators are better, but the dependence on the selected quantile for measuring the skew makes them less appealing. Therefore, two good families are the L-estimators and the abs-estimators. We have decided to present the analysis below using L-estimators as they have been studied in the statistical literature. Our definitions include some pre-factors so as to be consistent with the usual definitions for a normal distribution.

The sample mean and annualized volatility are defined by the moment estimators

$$\mu = \langle r \rangle \tag{3a}$$
$$s_M^2 = \langle r'^2 \rangle \, . \tag{3b}$$

The L-estimators are linear, with the following definition for the first four moments of a sample of size $N$:

$$l_1 = \binom{N}{1}^{-1} \sum_j r_{[j]} \tag{4a}$$
$$l_2 = \frac{1}{2} \binom{N}{2}^{-1} \sum_j \left\{ \binom{j-1}{1} - \binom{N-j}{1} \right\} r_{[j]} \tag{4b}$$
$$l_3 = \frac{1}{3} \binom{N}{3}^{-1} \sum_j \left\{ \binom{j-1}{2} - 2\binom{j-1}{1}\binom{N-j}{1} + \binom{N-j}{2} \right\} r_{[j]} \tag{4c}$$
$$l_4 = \frac{1}{4} \binom{N}{4}^{-1} \sum_j \left\{ \binom{j-1}{3} - 3\binom{j-1}{2}\binom{N-j}{1} + 3\binom{j-1}{1}\binom{N-j}{2} - \binom{N-j}{3} \right\} r_{[j]}. \tag{4d}$$

The first moment is equal to the sample mean $l_1 = \mu$. We define the L-size, L-skew and

L-kurtosis by

$$s_L = \sqrt{\pi}\, l_2 \tag{5a}$$
$$\tau_L = \frac{l_3}{l_2} \tag{5b}$$
$$\kappa_L = \frac{1}{0.1226}\,\frac{l_4}{l_2} \tag{5c}$$

where the pre-factors are the inverse of the corresponding values for a normal distribution. For a Gaussian distribution with variance $\sigma^2$, the L-estimators are $s_L = s_M = \sigma$, $\tau_L = 0$ and $\kappa_L = 1$.

The statistics for the volatility are much simpler as only the first two moments are used. The mean, variance and shape factor for the volatility are defined respectively by

$$\mu_\sigma = \langle \sigma \rangle \tag{6a}$$
$$s_\sigma^2 = \left\langle (\sigma - \mu_\sigma)^2 \right\rangle \tag{6b}$$
$$\gamma_\sigma = s_\sigma / \mu_\sigma. \tag{6c}$$

The standard deviation for the volatility is sometimes called the *vol-of-vol*. The shape factor measures the width of the distribution in unit of the mean. As the return $r[\Delta T]$ and volatility $\sigma[\Delta T]$ depend on a time horizon, all the statistical estimators above depend on the scale $\Delta T$.

In order to quantify the dynamical properties of the empirical time series and processes, lag $\Delta T$ correlations are computed. The return and volatility definitions (1) and (2) are *historical* quantities, namely they depend on the prices between $t - \Delta T$ and $t$. The *realized* return and volatility are defined similarly but with a shift in the time index so as to depend on the prices between $t$ and $t + \Delta T$. Equivalently, the realized quantities are the historical values moved forward by $\Delta T$, which we denote simply with a forward operator $F$ by $Fr$ and $F\sigma$ for the realized return and volatility respectively.

The (time series) correlations between return and volatility, one historical and one realized, characterize the dynamical properties of the time series. Essentially, these correlations measure the dependence between the immediate past and immediate future, probed both with a depth $\Delta T$ around $t$. The lag $\Delta T$ correlations are $\rho(r, Fr)$, $\rho(r, F\sigma)$, $\rho(\sigma, Fr)$, and $\rho(\sigma, F\sigma)$. For the volatility, the correlation between the historical volatility over $\Delta T$ and the realized volatility over $\Delta T'$ is also investigated, namely $\rho(\sigma[\Delta T], F\sigma[\Delta T'])$. The statistical studies have been done both for the usual Pearson measure of correlation, and for a robust measure of correlation based on the mean absolute deviation [Zumbach, 2006]. As there is no significant difference between both estimators, the results are reported for the Pearson estimator.

# 3 Empirical statistics

The data set used for the empirical analysis contains daily close prices for the stocks of three indexes: SPI, DJ EuroStoxx and S&P 500. The stocks are filtered for too short histories or too low liquidity (remove stocks less than 2000 days long, more than 20% of zero daily transaction values and more than 10 consecutive zero returns). The final set contains 777 stocks, with a length between 2000 and 7000 days up to February 2010.

The (time series) statistics defined on the previous section have been computed for the 777 stock time series and for the time horizons $n = 1, 2, 5, 10, 21, 42, 65, 130, 260$ days. The cross-sections of these statistics have been studied with three types of graphs. First, scatter plots for pair of statistics at given time horizons are used to reveal dependence, or independence. This study is mainly used to select the effects and their magnitudes which should be incorporated into the processes. Second, scaling plots for each statistics are used to study the temporal aggregation properties. This study allows us to select the time structure that should be included in a process. Third, the correlations between historical and realized volatility are computed for all pairs of time horizons, and then averaged over the time series. The corresponding graph measures the mean heteroskedasticity across time horizons, and in particular its asymmetry is one measure of the time reversal asymmetry of the stock time series [Zumbach, 2009b]. Taken together, all these graphs give a fairly exhaustive view of the properties of the data. A small subset is presented in this section, covering the most salient properties of the stocks.

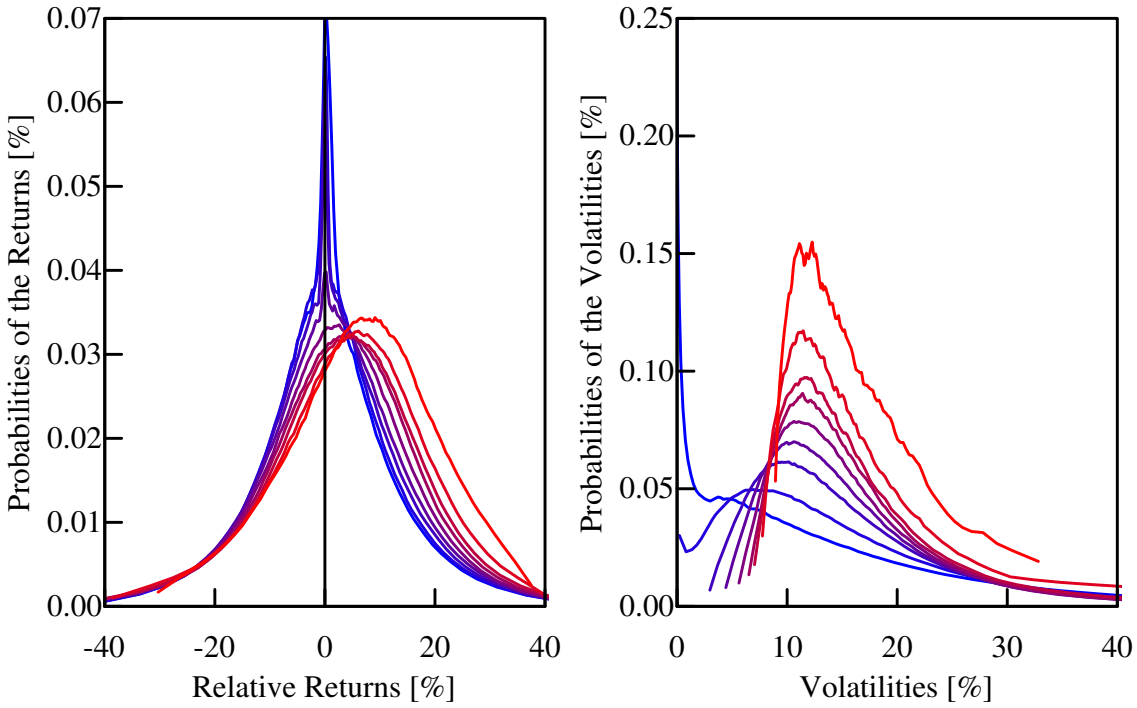


Figure 1: *The mean empirical probability density shape for the return and volatility (annualized and in %). The time horizons range from 1 day (blue) to 1 year (red).*

The mean empirical distributions for the return and volatility are displayed on Fig. 1. For each time series, the returns are first normalized by the standard deviation and 100 quantiles are computed. The volatilities are normalized by the mean volatility, before computing the quantiles. Then, the quantiles and probability densities are averaged over the time series in order to obtain the mean cdf and pdf. As the input time series are normalized for the variance, this procedure extracts the mean shape of the distributions. Many features can be observed on Fig. 1, such as the fat-tail distributions and the near stationarity of the return density. An important feature is the peak at zero return for short $\Delta T$. This peak is due to holidays and to the granularity of the prices imposed by the exchange rules. A more accurate measure of the height of the peaks is obtained by computing the probability of zero returns. A small threshold given by $10^{-6}$ of the interquartile range is used to discriminate zero returns. The mass included in the zero-return peak is of the order of 10% at one day, 4% at two days

and 2.4% at one week. Such large values are incompatible with an explanation in terms of observed holidays only, and a finite price granularity should be introduced. Both of these effects should be included in a process in order to duplicate the scaling behaviour of the zero return probability. The significant masses in zero return peaks influence several statistics, like for example the short term kurtosis. Another interesting feature of the mean probability density is the wide distribution for the volatility at long time horizons, a characteristic due to the heteroskedasticity.

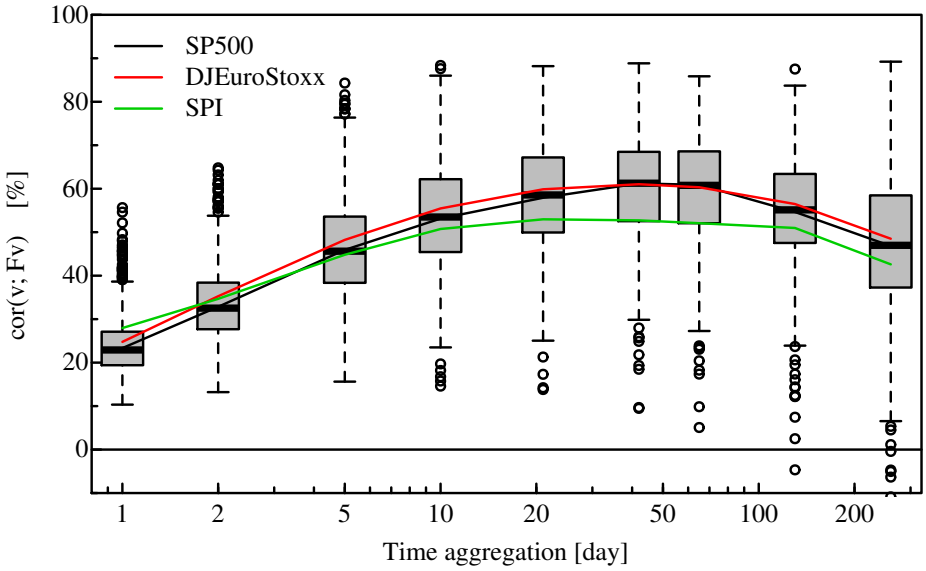


Figure 2: *Scaling box plot for the heteroskedasticity, measured by the historical–realized volatility correlation* $\rho(\sigma[\Delta T], F\sigma[\Delta T])$.

One of the most important characteristics of financial time series is the heteroskedasticity, namely the volatility is changing with time and is clustered. This description is quantified by the correlation between the historical and realized volatilities $\rho(\sigma[\Delta T], F\sigma[\Delta T])$. Figure 2 presents a scaling plot of the cross-section of this correlation. More precisely, the (time series) correlations $\rho(\sigma[\Delta T], F\sigma[\Delta T])$ for all the stocks are displayed as one box plot for each $\Delta T$. The centre of the box is given by the (cross-sectional) median, the box extends from the first to the third quartile, and the whiskers extend to 1.5 times the interquartile range or to the most extreme data point (whichever is closer to the median). The coloured lines give the mean values for the three data sets (all time series are gathered together for the boxes). This representation gives us a good view of the cross-section properties for characteristic times $\Delta T$ ranging from 1 day to 1 year. The historical–realized volatility correlation is the largest correlation in financial time series, with typical values ranging from 20 to 60%. Because of these large values, the heteroskedasticity is certainly the most important effect to capture in a process. The maxima of these correlations occur for time ranges between 1 to 3 months. For shorter time intervals, the volatility is not well estimated as the number of daily returns entering into the formula for computing $\sigma(t)$ is small. As the volatility for small $\Delta T$ contains a large statistical amount of noise, subsequent correlations are lower. For time ranges longer than a few months, the historical/realized volatility correlation decreases due to the decreasing memory on the volatility. Both effects lead to the observed hump shape.

More generally, the correlation $\rho(\Delta T, \Delta T') = \rho(\sigma[\Delta T], F\sigma[\Delta T'])$ between the historical volatilities $\sigma[\Delta T]$ and the realized volatilities $F\sigma[\Delta T']$ can be studied. This correlation gives more detailed information about the dependencies between historical and realized volatilities

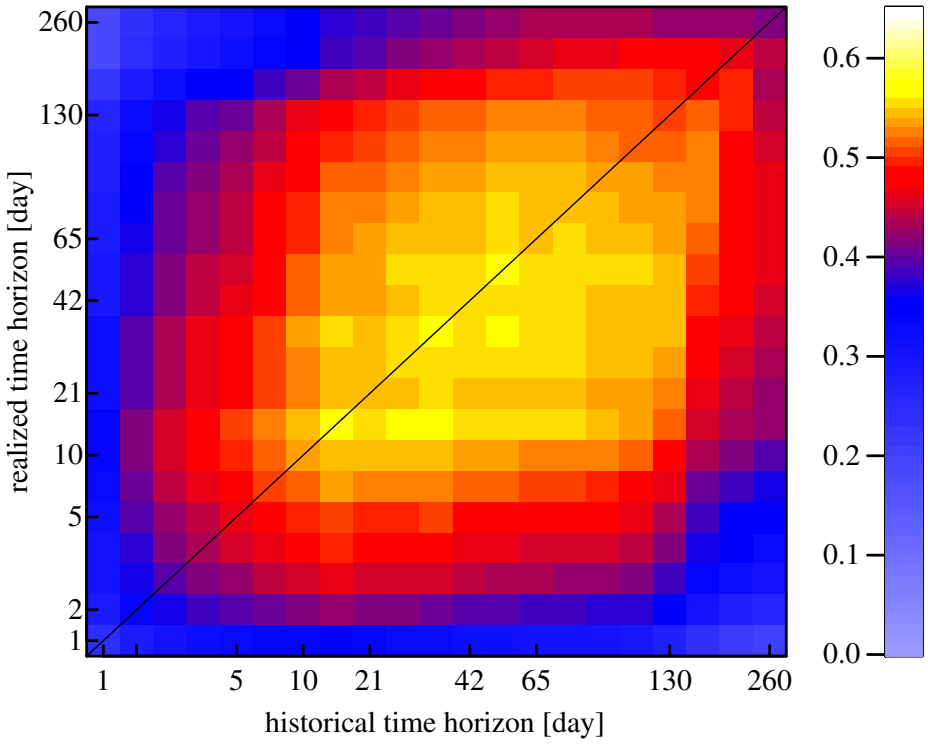


Figure 3: *The mean historical–realized volatility correlation* $\rho(\sigma[\Delta T], F\sigma[\Delta T'])$. *The correlation is computed for each stock time series, and the average is displayed. The historical time interval* $\Delta T$ *is given on the horizontal axis,* $\Delta T'$ *on the vertical axis.*

at various time horizons. In particular, the exchange of the arguments $\Delta T$ and $\Delta T'$ corresponds to exchanging the historical and realized volatilities, or to swap past and future. The difference between the value of $\rho(\Delta T, \Delta T')$ and $\rho(\Delta T', \Delta T)$ is therefore a measure of time reversal asymmetry of the financial time series. In [Zumbach, 2009b], this measure was used along with two other statistics to show that foreign exchange time series are not time reversal invariant. The definition of the volatility used in the computation is slightly different from (2) as the returns are aggregated for increasing $\Delta T$, namely

$$\sigma^2[\Delta T](t) = \frac{1}{n} \sum_{t-\Delta T+\Delta T_r \le t' \le t} r^2\left[\Delta T_r\right](t') \tag{7}$$

with $\Delta T_r = 1 + \lfloor \Delta T/15 \rfloor$ and $\lfloor \cdot \rfloor$ returns the integer part of its argument. Results for stocks are shown on Fig. 3. The observed asymmetry (by exchanging both axes) for stocks is very similar to the asymmetry for foreign exchange rates. As shown in [Zumbach, 2009b], this asymmetry gives a strong selection criterion for processes, which is well reproduced by the realistic ARCH process introduced in the next section.

Figures 4 and 5 show scaling box plots for the skews computed in two different ways. The first one is the usual computation, namely the third reduced moment of the logarithmic returns. The second figure displays the relative return L-skew. The important differences between both figures deserve a thorough discussion. The usual computation applied on our data set gives mostly negative values, but with a very large dispersion and a large number of outliers beyond the box whiskers. This finding is in agreement with the current consensus that the skews are negative for stocks [Cont, 2001]. The same computation but for relative returns gives similar dispersions, but with a slightly positive mean and median, and more positive

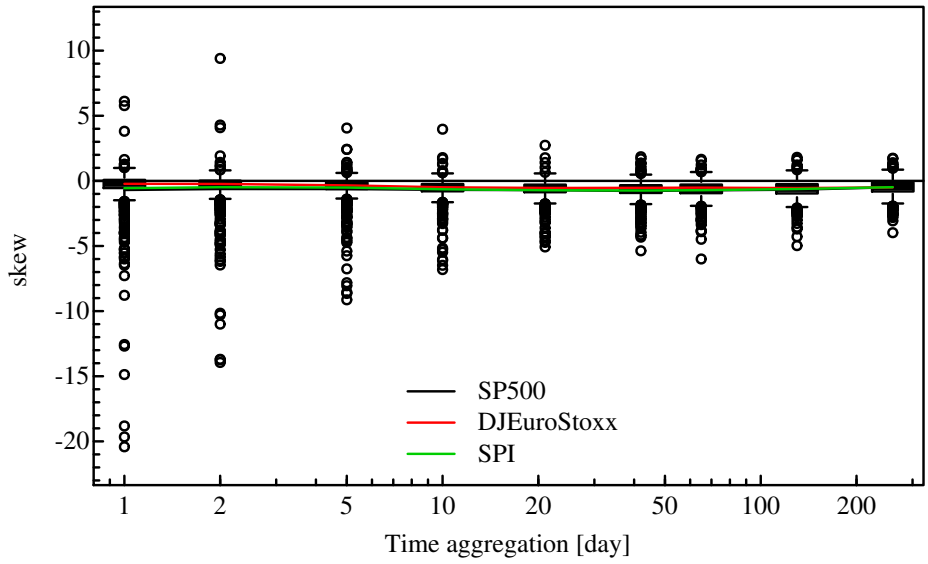


Figure 4: *Scaling box plot for the skew (evaluated with the third reduced moment) of the logarithmic returns.*

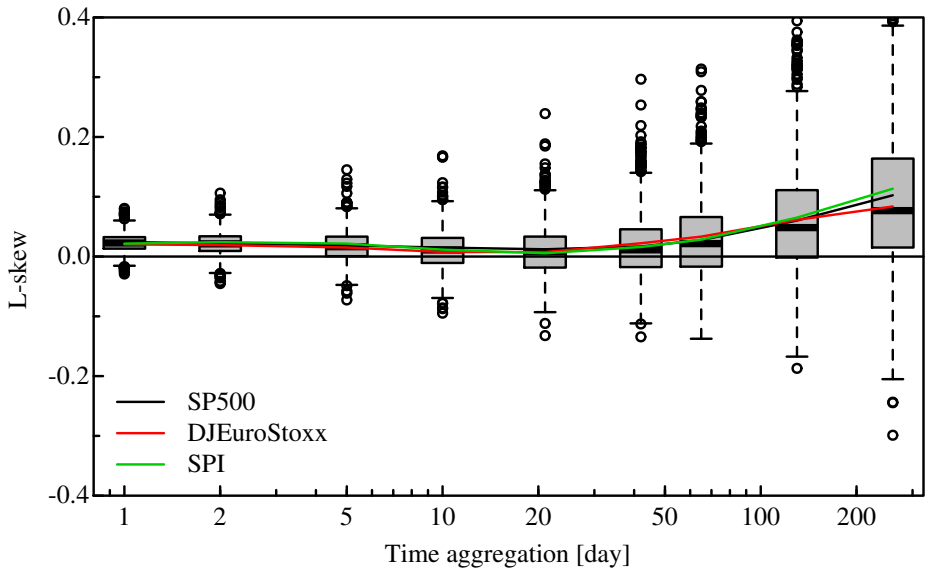


Figure 5: *Scaling box plot for the L-skew of the relative returns.*

outliers. The difference in the sign is explained by the convex mapping between relative and logarithmic returns given by $r_{\text{rel}} = \exp(r_{\log}) - 1$, which increases (decreases) large positive (negative) returns. Yet, the large dispersions and the number of outliers show that it is dangerous to compute moments for fat-tailed data, and that the conclusion about negative skews is fragile. The computation of the L-skew with logarithmic return gives a zero median skew at $\Delta T = 1$ day, and then an increasing negative skew for increasing $\Delta T$. Figure 5 plots the L-skew for relative returns. Surprisingly, the L-skew is *slightly positive*, and therefore differs from today's "negative skew" consensus. The increasing skews and dispersions at large $\Delta T$ are reproduced by a simple constant volatility random walk (see Fig. 12). This shows that the right side of the plot is simply due to the aggregation of the returns. The

small positive L-skew for $\Delta T$ between 1 day and 1 week is the remaining stylized fact about the skew. Because of the small numerical values, we have decided to neglect the skew in the processes hereafter. If needed, the skew at short $\Delta T$ can be introduced easily by using a skew distribution for the innovations, like a skewed Student.

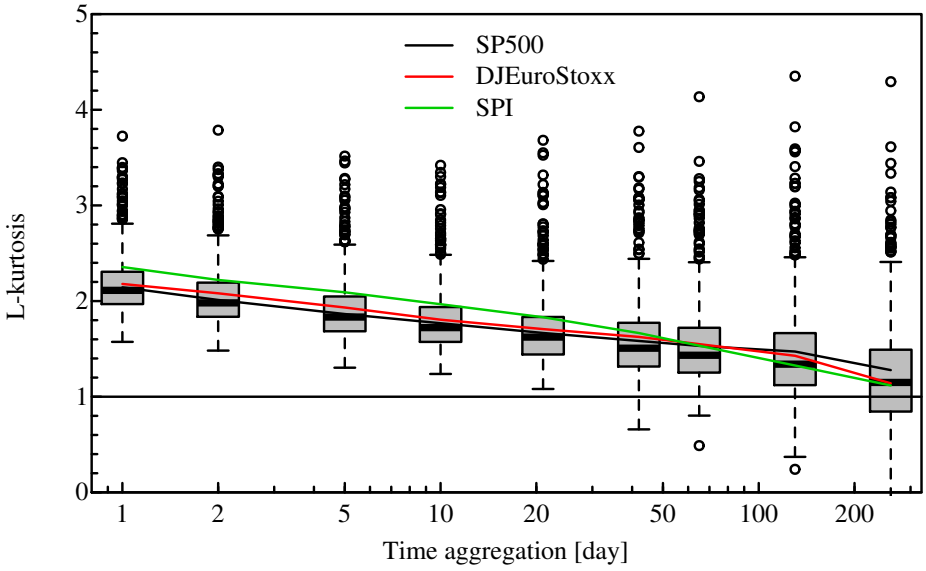


Figure 6: *Scaling box plot for the L-kurtosis.*

Figure 6 shows the scaling box plot for the L-kurtosis. Together with Fig. 5, these graphs give the main shape factors for the relative return distributions. Because the normalisation for the L-kurtosis $\kappa_L$ has been chosen so that a normal distribution corresponds to $\kappa_L = 1$, the empirical data are always leptokurtic. The interesting feature is the slow decay of the kurtosis. Monte Carlo simulations for a constant volatility random walk with Student innovations show a decrease of the kurtosis roughly similar to a $1/\Delta T$ behaviour. Therefore, the persistence of the kurtosis for large $\Delta T$ is a feature of the empirical time series that we would like to reproduce in a process.

The leverage effect is another important stylized fact for stock time series. This corresponds to the negative correlations between historical returns and realized volatilities, namely a large decrease in the stock value is likely to be followed by a large volatility. The scaling box plot shown in Fig. 7 is compatible with this description. Interestingly, the effect is increasing with the time interval $\Delta T$, with correlations from -10 to -30%. Therefore, this is also a quantitatively large effect that should be captured in a process, including its peculiar scaling structure.

An interesting question is the dependence between the mean volatility and the shape of the volatility distribution. Figure 8 shows a scatter plot for these statistics for $\Delta T = 10$ days. Interestingly, there is essentially no cross-sectional dependence between these statistics, showing that the shape of the volatility distribution is essentially independent of the mean volatility. The absence of dependence can be re-phased as the *vol-of-vol* is proportional to the volatility, but is not an independent parameter.

Figure 9 displays a cross-section for the heteroskedasticity intensity $\rho(\sigma, F\sigma)$ versus the shape of the return distribution as measured by the L-kurtosis, both at 1 day. A similar figure is obtained between the heteroskedasticity and the shape of the volatility distribution as

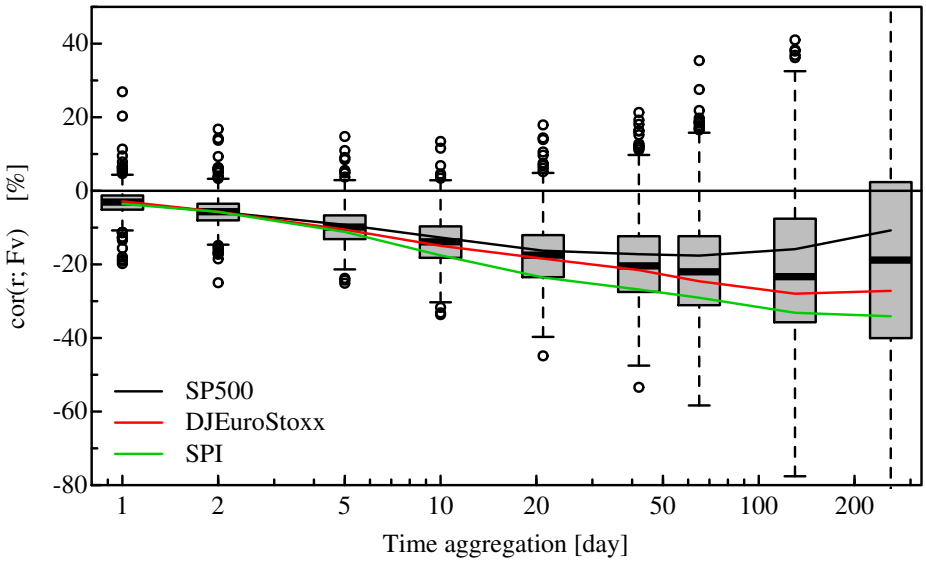


Figure 7: *Scaling box plot for the the leverage effect, namely correlation between the return and the forward volatility.*

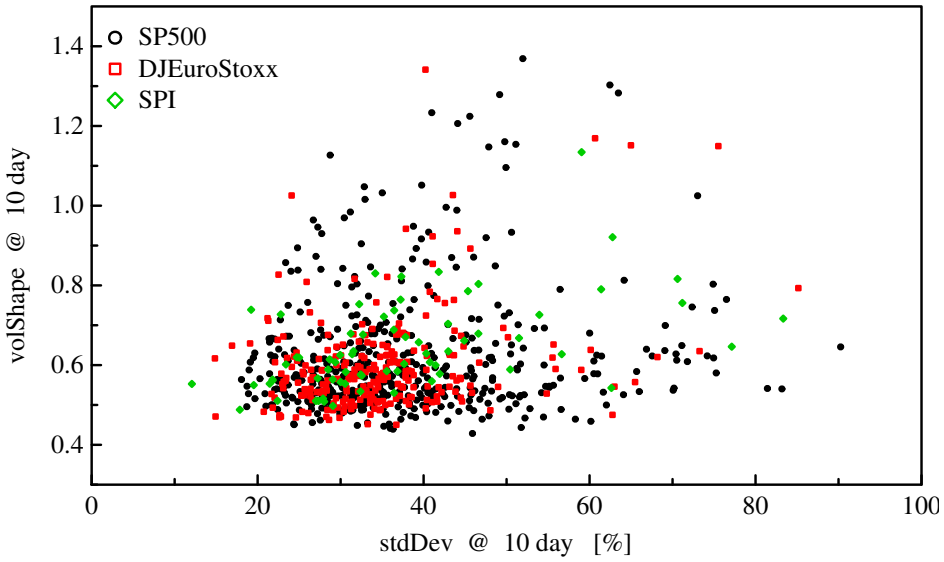


Figure 8: *Scatter plot for the volatility shape versus the standard deviation, both at 10 days.*

measured by $\gamma_\sigma$. This dependence decreases for an increasing time span $\Delta T$. The strong dependence is very interesting as it relates a (static) distributional property ($\gamma_\sigma$) in terms of a dynamical quantity ($\rho(\sigma, F\sigma)$). Essentially, this shows that the shape of the return and of the volatility distribution is induced largely by the heteroskedasticity.

## 4 The processes

The Monte Carlo simulations use a process built using a long memory aggregated ARCH for the heteroskedasticity, with leverage and fat-tailed innovations. The core idea is to have a

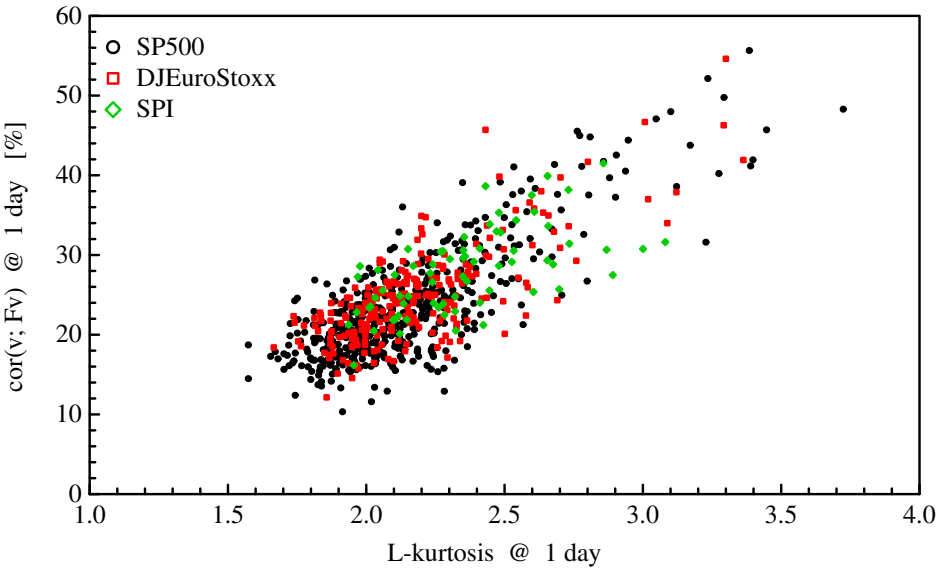


Figure 9: *Scatter plot for the heteroskedasticity versus the L-kurtosis of the return distribution, both at 1 day.*

set of exponential moving averages that capture the volatility at increasing time horizons, and with an asymmetry for positive versus negative returns (the leverage effect). A convex combination of these squared volatilities produces the volatility for the forthcoming returns, with innovations distributed according to a Student distribution. Then, holidays and price granularity are introduced.

The process is based on a set of increasing time intervals built with a geometric progression:

$$\Delta T_k = \Delta T_1 \, \rho^{k-1} \quad \text{and} \quad k = 1, ..., n \tag{8a}$$

$$\tau_k = \frac{\tau_1}{\Delta T_1} \, \Delta T_k \tag{8b}$$

$$\mu_k = \exp\left(-\,\delta t / \tau_k\right). \tag{8c}$$

The returns are computed over the time interval $\Delta T_k$ while the volatilities are computed with a characteristic time $\tau_k$ proportional to $\Delta T_k$. For a process with a daily increment, we take $\Delta T_1 = 1$ day and the progression factor of the series is $\rho = 2$. The coefficients $\mu_k$ are the exponential moving average coefficients corresponding to $\tau_k$. The variance at the scale $\tau_k$ is essentially computed by an exponential moving average (EMA) of the squared return $r^2[\Delta T_k]$:

$$r_k(t) = r[\Delta T_k](t) - \frac{\Delta T}{1 \text{ year}} \, \mu_{\text{eff}} \tag{9a}$$

$$\sigma_k^2(t) = \mu_k \, \sigma_k^2 \left(t - \delta t\right) + \left(1 - \mu_k\right) \left\{1 - \lambda_{\text{lev}} \tanh\left(\frac{r_k}{\lambda_{\text{range}} \, \sigma_\infty}\right)\right\} \left(r_k(t)\right)^2. \tag{9b}$$

The de-meaned return $r_k$ at the scale $\Delta T_k$ is obtained by subtracting the mean return at the scale $\Delta T$. The coefficient $\mu_{\text{eff}}$ is the annualised drift of the process, namely $E\left[r[\delta t]\right] = \mu_{\text{eff}}$ and where the returns $r[\delta t]$ are annualized. The historical variance $\sigma_k^2$ is an EMA of the past squared returns, and the term in braces introduces the leverage effect. The intensity for the

leverage is given by the parameter $\lambda_{\text{lev}}$ while the tanh function provides for a smooth leverage effect for small returns. The range for the smoothing is given by the parameter $\lambda_{\text{range}}$, in units of the mean volatility. Removing $\mu_{\text{eff}}$ in Eq. 9a has a small impact on the resulting process.

The return $r[\Delta T_k]$ entering the measure for the historical volatility $\sigma_k$ is an *aggregated* return over a similar time interval. Another definition can use the *microscopic* returns, namely $r[\delta t]$. Only the aggregated definition leads to the correct asymmetry for the historical–realized volatility correlation presented in Fig. 3, in agreement with [Zumbach, 2009b]. This difference is also important for the leverage effect as a microscopic definition does not allow us to obtain the correct magnitude as a function of $\Delta T$.

The effective variance $\sigma^2_{\text{eff}}$ is obtained by a convex combination of the variance $\sigma^2_k$ and the parameter $\sigma^2_\infty$

$$\sigma^2_{\text{eff}}(t) = \sum_{k=1}^{n} w_k \, \sigma^2_k(t) + w_\infty \, \sigma^2_\infty. \tag{10}$$

The weights $w_k$ are given by

$$w_k = \frac{1}{C}\left(1 - \frac{\log(\tau_k)}{\log(\tau_0)}\right) \tag{11}$$

and the normalization $C$ is fixed by the condition $\sum_k w_k + w_\infty = 1$. The parameter $\tau_0$ fixes the decay of the weights for the EMA volatilities, while the parameter $w_\infty$ gives the weight for the mean volatility parameter $\sigma_\infty$. The parameter $\sigma_\infty$ fixes the mean volatility (for a logarithmic process without leverage, $\sigma^2_\infty$ is equal to the mean variance). The weight $w_\infty$ for the mean variance parameter takes value in the range $[0, 1]$. A constant volatility process is obtained for $w_\infty = 1$, while the limit $w_\infty = 0$ is singular and corresponds to an integrated process (see [Zumbach, 2004] for a discussion of the affine versus integrated processes).

The process with relative returns is given by

$$r_{\text{rel}}[\delta t](t+\delta t) = \mu_{\text{eff}} + \sigma_{\text{eff}}(t)\,\epsilon(t+\delta t) \tag{12}$$

$$p(t+\delta t) = p(t)\left(1 + \sqrt{\frac{\delta t}{1 \text{ year}}}\, r_{\text{rel}}[\delta t](t+\delta t)\right). \tag{13}$$

The parameters $\mu_{\text{eff}}$ and $\sigma_\infty$ are annualized, the returns $r_{\text{rel}}[\Delta T]$ are annualized, and therefore the effective volatility $\sigma_{\text{eff}}$ and the random returns $r_{\text{rel}}[\delta t](t+\delta t)$ are also annualized. The ratio $\sqrt{\delta t / 1 \text{ year}}$ scales back the returns at the scale $\delta t$. This convenient convention allows us to have intuitive values for the parameters. The innovations $\epsilon$ are drawn from a fixed distribution (Student or Gaussian).

The empirical analysis shows that it is necessary to introduce holidays and a price granularity. The holidays are drawn at random, with a probability $p_{1 \text{ day}}$ and $p_{2 \text{ days}}$ for the one and two days holidays. During a Holiday, the price history used to compute $r_k$ is not updated, the variances $\sigma_k$ are not updated, and the price is left unchanged. After a Holiday, a period of 5 days without holidays is enforced. The price granularity $g$ is introduced using a simple rule that makes the minimum price difference roughly proportional to the price. From the raw price $p$ generated by the random walk algorithm, a granular price $p'$ is obtained by

$$p' = \frac{\lfloor p\, g + 0.5 \rfloor}{g} \tag{14}$$

where $\lfloor \cdot \rfloor$ rounds the argument to the nearest integer closer to zero (e.g. $\lfloor 0.8 \rfloor = \lfloor -0.8 \rfloor = 0$, $\lfloor -1.8 \rfloor = -1$, $\lfloor 1.8 \rfloor = 1$). The granularity depends on a base granularity $g_{\text{base}}$ and the price range according to the following table. Empiricaly, a value of $g_{\text{base}} = 4$ reproduces well the observed rate of zero returns.

| Price range | Tick size $g$ | Tick size for $g_{\text{base}} = 4$ |
|---|---|---|
| $0 \leq p < 5$ | $1/(g_{\text{base}} \cdot 100)$ | 0.0025 |
| $5 \leq p < 50$ | $1/(g_{\text{base}} \cdot 10)$ | 0.025 |
| $50 \leq p \leq 500$ | $1/g_{\text{base}}$ | 0.25 |
| $500 < p \leq 5000$ | $10/g_{\text{base}}$ | 2.5 |
| $5000 < p$ | $100/g_{\text{base}}$ | 25 |

This completes the description of the most general process used. With $n = 1$ component and $\lambda_{\text{lev}} = 0$, this process reduces to the GARCH(1,1) process. The simulations are started from an initial price of $p(t_{\text{initial}}) = 10$, followed by 1000 MC steps that are discarded. This initialization randomizes the volatilities and prices at the start of the sample. Then, the simulations used to compute the statistics are done so as to duplicate the empirical data set, namely the number of simulations and their respective lengths are identical to the empirical set. The process parameters are optimized so as to reproduce at best a broad panel of statistics from 1 day to 1 year. A small subset of those statistics are shown in Sec. 3, and a few interesting points of the processes are presented below. Let us emphasize that the parameter estimation is not done by log-likelihood maximization. A log-likelihood procedure focuses on a one day scale, and gives only a poor diagnostic about the adequacy of a process with respect to empirical data. By contrast, our heuristic procedure leads to a good data generating process over a broad range of time scales. The adequacy is measured by many statistics that test systematically important features of the data. This constitutes a much more stringent test than a single number computed at the scale $\delta t$.

The parameters have been optimized by trial and error, with the following values for the best model. The time structure of the volatility influences mostly the heteroskedasticity measured by $\rho(\sigma, F\sigma)$. The following values lead to the observed broad maximum: $n = 8$, $\Delta T_1 = 1$ day, $\Delta T_8 = 64$ days, $\rho = 2$, $\tau_1 = 5$ days and the decay for the weights $w_k$ given by $\tau_0 = 4160$ days $= 8$ years. The value for $\tau_1$ is such that the EMAs include enough independent returns, so that the $\sigma_k$ are not too noisy. With this choice, the EMA characteristic time ranges roughly from 1 week to 1 year. The $\tau_1$ parameter is the most important in this group as it fixes the location for the maximum of the $\rho(\sigma, F\sigma)$ correlation. The leverage effect is well reproduced by $\lambda_{\text{lev}} = 0.55$, while the parameter $\lambda_{\text{range}} = 0.5$ is less important. The one and two days holidays probability are $p_{1\text{ day}} = 0.04$ and $p_{2\text{ day}} = 0.005$. With a base granularity of $g_{\text{base}} = 4$, these parameters reproduce well the rate of zero returns, and the short-term kurtosis. The mean volatility parameter $\sigma_\infty = 0.7$ leads to the observed standard deviation for the return $\overline{\langle r^2 \rangle^{1/2}} = 0.38$. The historical volatilities $\sigma_k$ are initialized with this last value. The larger parameter $\sigma_\infty$ compared to the mean volatility is needed to compensate from various effects due to the relative return, the leverage and the price granularity. The coupling constant for the mean volatility fixes mainly the level of the heteroskedasticity, and $w_\infty = 0.02$ leads to the observed maximum with a 60% correlation. The drift parameter $\mu_{\text{eff}} = 0.17$ reproduces the mean observed drift $\overline{\langle r \rangle} = 0.17$. This fairly large value for the drift likely originates in the selection bias induced by taking only the stocks constituant of indexes. The distribution for the innovations is a Student with 5 degrees of freedom.

For comparison purposes, the results for a few simpler processes are included below. These processes are a constant volatility with Student innovations, a GARCH(1,1) process with

Student innovations (and without leverage), and a long memory ARCH process with leverage and Student innovations but without granularity nor holidays. The parameters are set according to the same global screening for the statistics. For the GARCH(1,1) process, the parameters are $\tau = 32$ days and $w_\infty = 0.115$.

## 5 Simulation results

As mentioned in the previous section, the simulations are done so as to replicate the empirical data set in terms of the number of time series and their respective lengths. One set of parameters is used, and the fluctuations in the statistics computed from the simulations are due only to the different realizations of the random price paths. The set-up for the scaling graphs are similar to the empirical ones, with the boxes computed from the LM-ARCH process with granularity, and the curves corresponding to the mean for selected processes.

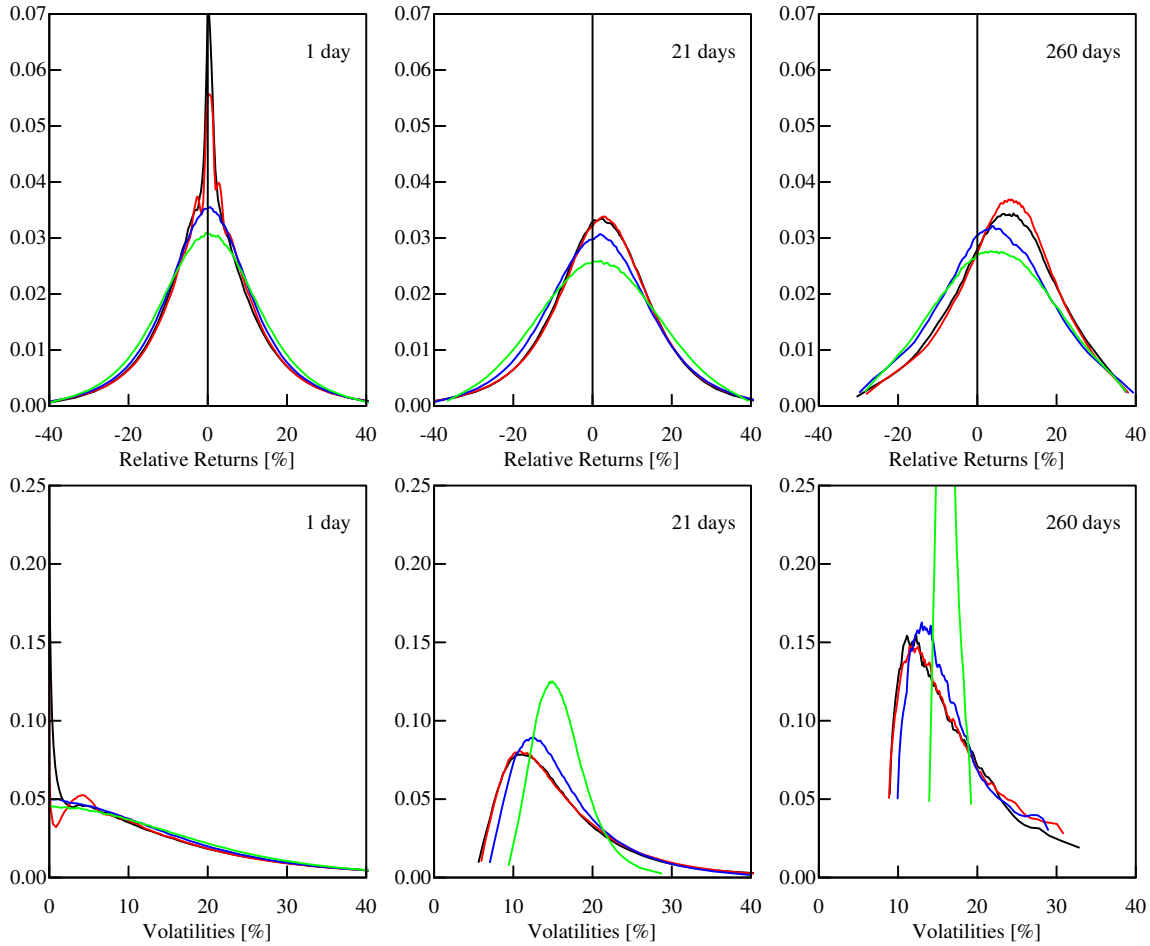


Figure 10: *The probability density for the returns (top line) and volatility (bottom line), for the time horizons of 1 day (left), 1 month (center), 1 year (right). The return and volatility are annualized and in %. The colours correspond to the empirical data (black), LM-ARCH with granularity (red), GARCH(1,1) (blue) and constant volatility + Student innovations (green).*

A first overall diagnostics of the adequacy between processes and empirical time series is obtained by comparing the probability density for the return and volatility on various time

horizons. Figure 10 displays these distributions for the time horizons of 1, 21 and 260 days. The leading $\sqrt{\Delta T}$ scaling is removed by annualizing returns and volatilities. In this way, the same ranges are used regardless of $\Delta T$, and are identical to Fig. 1. The long tails at 1 day are cropped so that the bulks of the distributions are clearly visible. Obviously, the constant volatility process (green) is not very good, however a GARCH(1,1) process (blue) is much better. The agreement between the LM-ARCH with granularity (red) and the empirical time series (black) is excellent, and indeed the red curves are often covering the black ones.

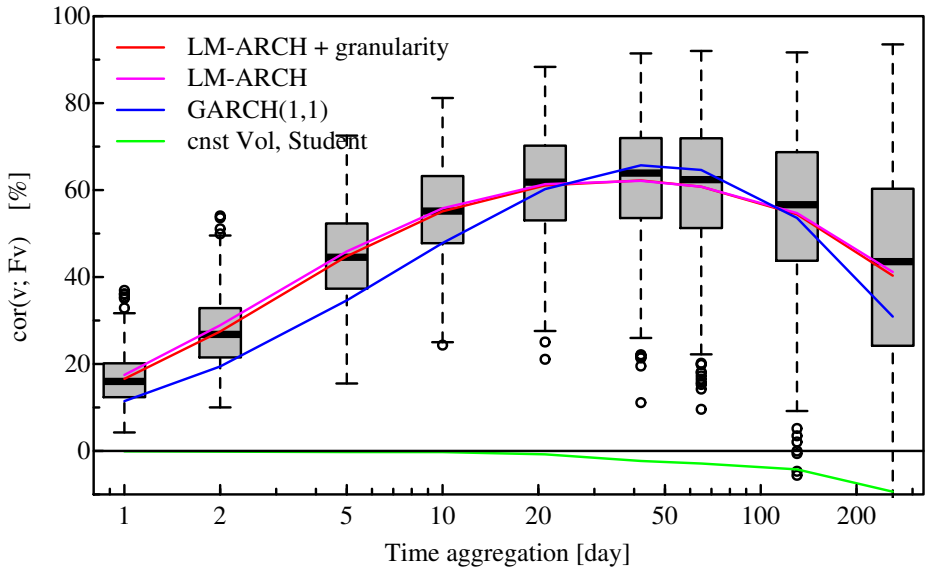


Figure 11: *The heteroskedasticity for the LM-ARCH process with granularity (box plot) and the means for a few processes (colored curves).*

The heteroskedasticity for a few processes is observed in Fig. 11. The too narrow range provided by a GARCH(1,1) process is clearly revealed. Let us emphasize that location and height of the maximum can be adjusted by changing parameters, but the shape cannot be made broader. The heteroskedasticity with the long memory process is essentially accurate, however too small in the small $\Delta T$ region. This is due to the simulations based on a daily increment: the historical volatilities $\sigma_k$ are fairly noisy and this noise limits the induced correlation. In order to reach the empirical level for the heteroskedasticity, the parameter $w_\infty$ should be very small.

The scaling plot for the L-skew is reported in Fig. 12. Interestingly, the skew is increasing with the aggregation, due to the definition of the process based on relative return. This shows that the right hand side of the scaling plot for the empirical L-skew is due only to the multiplicative random walk. The small skew in the empirical data at small $\Delta T$ has been neglected in the process.

The scaling behaviour of the L-kurtosis is plotted on Fig. 13. The process with constant volatility and Student innovations is converging fast towards the kurtosis of the normal distribution. Yet, the convergence is slower than $1/\Delta T$, as would be obtained by an Edgeworth expansion for a logarithmic random walk. The kurtosis for the GARCH(1,1) process is too small, while the kurtosis for the long memory ARCH process is essentially accurate. The scaling for the leverage effect is essentially correct for the LM-ARCH process with granularity, as well as the scaling of all other statistics.

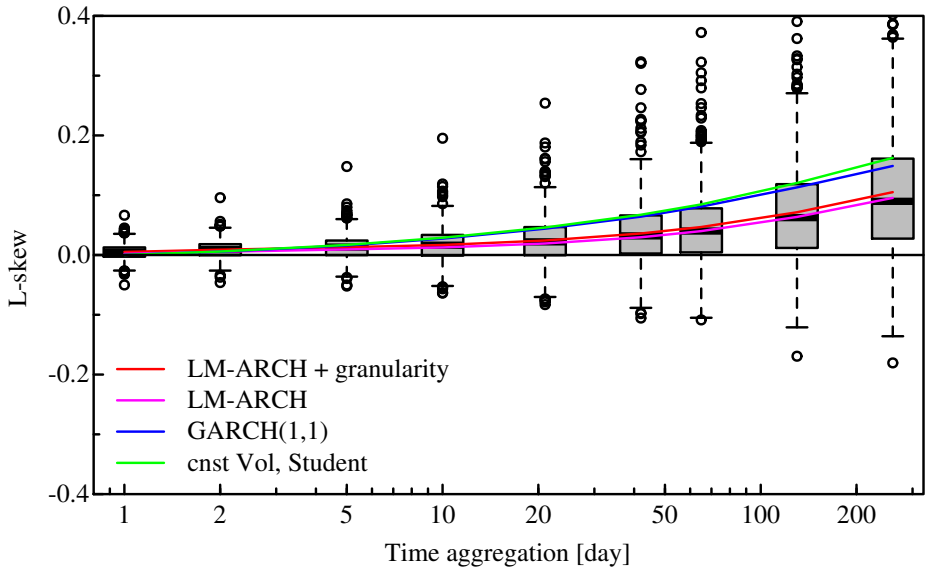


Figure 12: *The L-skew for the LM-ARCH process with granularity (box plot) and the means for a few processes (coloured curves).*

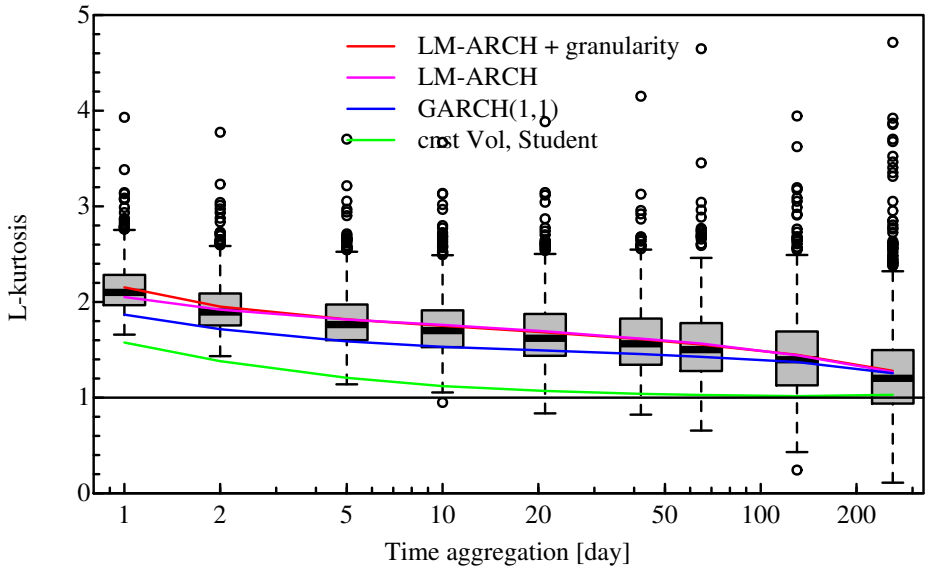


Figure 13: *The L-kurtosis for the LM-ARCH process with granularity (box plot) and the means for a few processes (coloured curves).*

A surprising fact of the scaling graphs is that boxes and outliers are very similar between empirical and simulation plots. These are rough measures of dispersion for the corresponding statistics. Indeed, all the scatter plots are very alike, with a dispersion in the empirical sample and in the simulations that are comparable. This is the case for the LM-ARCH process with granularity, while GARCH(1,1) simulations exhibit smaller dispersions (and constant volatility processes have a much smaller dispersion). Figure 14 compares the cross-sectional distributions of the mean sample volatility, for the data and two processes. Let us emphasize that simulations are done with a unique set of parameters. The natural conclusion is that empirical time series and their cross-sectional properties can be well described by a single

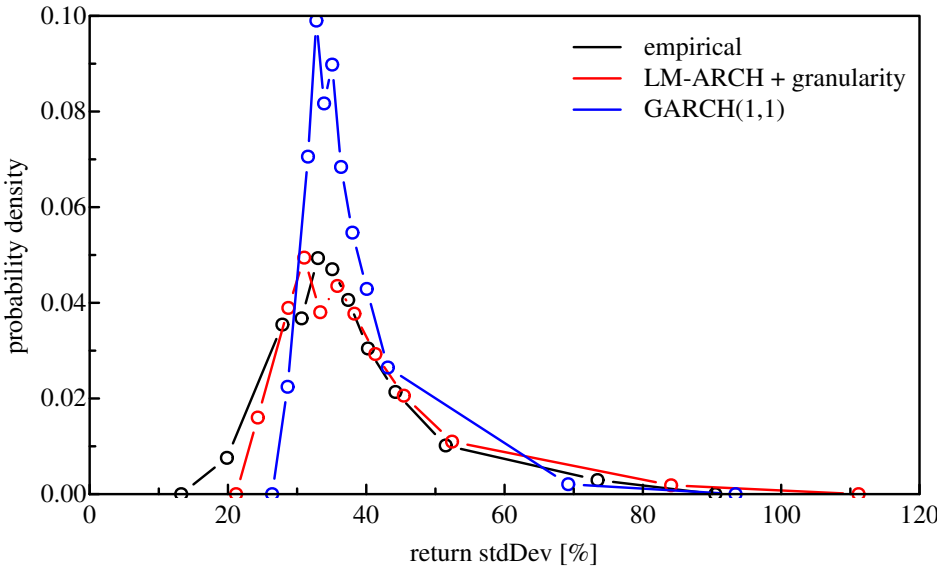


Figure 14: *The cross sectional probability density for the mean standard deviation for the 1 day return (i.e. the mean sample volatility), for the empirical data set and for 2 processes.*

process, with a unique set of parameters. For risk applications, the empirical distribution of innovations was studied in [Zumbach, 2006] and was found to be well described by a Student with 5 degrees of freedom for all asset classes. In [Zumbach, 2009a], a study devoted to the heteroskedasticity reached a similar conclusion, namely the shape of the lagged correlation for measures of the volatility is similar for a broad set of financial time series, and therefore can be characterized by the same function with the same parameters. In our case, the implication is that all stocks in our universe can be described by a unique set of parameters, including mean volatility, heteroskedasticity, leverage, drift and distribution for the innovations.

## 6 Conclusions

Our analysis shows that one process can reproduce a broad panel of statistics (location, size, skew, kurtosis, correlation and lagged correlation between return and volatility) computed for time intervals ranging from 1 day to 1 year for stock time series. The process needs to include heteroskedasticity with long memory, leverage, fat-tail innovations, relative returns, price granularity, and holidays, while possible other effects are small and can be neglected. A systematic screening of the return statistics, either in the form of scaling plots or two-dimensional scatter plots, shows only minor differences between empirical stock data or generated data. Therefore, large scale Monte Carlo simulations can be done with this process, for example to optimize portfolios, to evaluate market risks, or to price derivative contracts. Beyond the availability of a convenient tool for simulations, the method and some results deserve a few comments.

The value for the parameter $w_\infty$ is found to be small, in the range of 0.02 to 0.04 for the long memory ARCH process. Essentially, the process is very close to the singular limit $w_\infty = 0$ where it becomes integrated, namely the mean volatility is no longer fixed by the parameters but by the initial conditions. In this limit, the process is "unstable" and does not have well defined asymptotic distributions [Nelson, 1990, Zumbach, 2004]. The opposite

limit is $w_\infty = 1$ where the volatility is constant and the process is trivially stable. The structure of the process gives us a hint about the explanation. Intuitively, the coupling between a particular time series and the physical world is contained in the innovations, which are described by a simple iid process. The innovations should have a fat-tail distribution, but no time dependence needs to be introduced here. On the other hand, the historical volatilities $\sigma_k$ measure the responses of the market participants induced by the past. The optimal parameter values indicate that the heteroskedasticity is mostly endogenous to the market. It remains to be explained why the collective response of the market participants is described by a process very close to an instability threshold.

Some care needs to be taken with respect to the skew and kurtosis statistics related to the shape of the return distribution. The daily innovations are described by a Student with 5 degrees of freedom. This gives a lower bound on the tail exponent of the return distribution, and the dynamics induced by the heteroskedasticity can only decrease the return tail exponent. Clearly, computing the skew and kurtosis with the usual reduced moments is dangerous. A cross-sectional study of the usual estimators against robust estimators confirms the extreme sensitivity of the moment estimators, and the corresponding L-estimators give simple (robust) alternatives to the skew and kurtosis. The L-skew estimator applied to relative returns results in zero skew (at short $\Delta T$) or positive skew (at long $\Delta T$), in contrast with the present consensus about negative skews for stocks. A cross-sectional study for the L-kurtosis shows the strong dependence with measures of heteroskedasticity, hence relating distributional properties with dynamical ones.

The process equations have been written in such a way as to separate the effects and parameters included in the process. Consequently, it is fairly easy to find a good set of parameter values that describe well the data over a broad range of time horizons. This is different from a log-likelihood procedure that focuses only on the shortest time increment and maximises a single number for each time series. Although the log-likelihood maximisation offers a mathematical framework to decide about the relevance of parameters for nested processes, it is blind to missing effects inside a family, and is weakly sensitive to long term effects. On the other hand, our heuristic procedure focuses on a broad set of statistical properties. These properties are indeed what we care about to describe financial time series.

Finally, a surprising outcome is the similar dispersion between empirical stock time series and simulations for a single set of parameter values. This similar dispersion is observed for all statistical estimators. This finding points to a fairly universal description of financial data, with most parameters fixed at their default values. Parameters in this class can be the volatility time structure, the shape for the leverage effect and the distribution for the innovations. A small set of parameters can be taken as time series dependent and estimated by simple moment estimators on empirical and simulated time series, like for example the drift, mean volatility and magnitude of the leverage effect. Even for these parameters, the Figure 14 suggests that a large part of the cross-sectional variability is due to different statistical realizations of a process with the same parameter values. Regardless of a split between universal parameters versus time series dependent parameters, the large statistical dispersions for the statistics suggest that it is irrelevant to quote parameter values with more that 2 significant figures.

## References


[Bollerslev, 1986] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327.

[Cont, 2001] Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236.

[Engle, 1982] Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U. K. inflation. *Econometrica*, 50:987–1008.

[Engle and Bollerslev, 1986] Engle, R. F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, 5:1–50.

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

[Granger and Ding, 1996] Granger, C. and Ding, Z. (1996). Varieties of long memory models. *J. Econometrics*, 73:61–77.

[Hosking, 1990] Hosking, J. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society B*, 52(1):105–124.

[Nelson, 1990] Nelson, D. (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, 6:318–34.

[O'Neil and Zumbach, 2009] O'Neil, C. and Zumbach, G. (2009). Using relative returns to accommodate fat tailed innovations in processes and option pricing. Technical report, RiskMetrics Group. Submitted; Available at www.ssrn.com.

[Zumbach, 2004] Zumbach, G. (2004). Volatility processes and volatility forecast with long memory. *Quantitative Finance*, 4:70–86.

[Zumbach, 2006] Zumbach, G. (2006). The riskmetrics 2006 methodology. Technical report, RiskMetrics Group. available at: www.riskmetrics.com and www.ssrn.com.

[Zumbach, 2009a] Zumbach, G. (2009a). Characterizing heteroskedasticity. Technical report, RiskMetrics Group. preprint, available at www.ssrn.com.

[Zumbach, 2009b] Zumbach, G. (2009b). Time reversal invariance in finance. *Quantitative Finance*.