Univariate Volatility Models: ARCH and GARCH

Massimo Guidolin
Dept. of Finance, Bocconi University

1. **Introduction**

Because volatility is commonly perceived as a measure of risk, financial economists have been traditionally concerned with modeling the time variation in the volatility of (individual) asset and portfolio returns. This is clearly crucial, as volatility, considered a proxy of risk exposure, leads investors to demand a premium for investing in volatile assets. The time variation in the variance of asset returns is also usually referred to as the presence of *conditional heteroskedasticity* in returns: therefore the risk premia on conditionally heteroskedastic assets or portfolios may follow a dynamics that depends on their time-varying volatility. The concept of conditional heteroskedasticity extends in general to all patterns of time-variation in conditional second moments, i.e., not only to conditional variances but also to conditional covariances and hence correlations. In fact, you will recall that under the standard (conditional) CAPM, the risk of an asset or portfolio is measured by its conditional beta vs. the returns on some notion of the market portfolio. Because a conditional CAPM beta is defined as a ratio of conditional covariance with market portfolio returns and the conditional variance of returns on the market itself, patterns of time-variation in covariances and correlations also represent ways in which time-varying second moments affects investors' perceptions of risk exposure. Moreover, as already commented in chapter 1, banks and other financial institutions apply risk management (e.g., value-at-risk, VaR) models to high frequency data to assess the risks of their portfolios. In this case, modelling and *forecasting* volatilities and correlations becomes a crucial task for risk managers.

The presence of conditional heteroskedastic patterns in financial returns is also intimately related to the fact that there is overwhelming evidence that the (unconditional) distribution of realized returns on most assets (not only stocks and bonds, but also currencies, real estate, commodities, etc.) tends to display considerable departures from the classical normality assumption. We shall document that conditional heteroskedasticity implies that the unconditional, long-run distribution of asset returns is non-normal.[1] This is well-known to be potentially responsible for strong departures

[1] We shall define the technical terms later on, but for the time being, the unconditional distribution of a time series process is the overall, long-run distribution of the data generated by the process. Drawing on one familiar example, if $X_{t+1} = \phi X_t + \epsilon_{t+1}$ with $\epsilon_{t+1} \sim \mathcal{N}(0,1)$, it is clear that the conditional distribution of X_{t+1} at time t (i.e., given information observed at time t) is $\mathcal{N}(\phi X_t, 1)$; however, in the long-run, when one averages over infinite draws from

of observed derivative prices from simple but still widely employed pricing frameworks that are built on the classical results by Black and Scholes (1973) that rely on normality of financial returns.

Given these motivations, in this chapter we develop and compare alternative variance forecasting models for each asset (or portfolio) individually and introduce methods for evaluating the performance of these forecasts. In the following chapters, we extend these methods to a framework that may capture any residual deviations of the distribution of asset returns from normality, after any models of conditional heteroskedasticity have been applied. Additionally, we show how it is possible to connect individual variance forecasts to covariance predictions within a correlation model. The variance and correlation models together will yield a time-varying covariance model, which can be used to calculate the variance of an aggregate portfolio of assets

This chapter has two crucial lessons that go over and above the technical details of each individual volatility model or its specific performance. First, one should not be mislead by the naive notion that because second moments change over time, this implies that the time series process characterized by such moments becomes "wild", in the sense of being non-stationary. On the contrary, under appropriate technical conditions, one can prove that even though the conditional variance may change in heteroskedastic fashion, the underlying time series process may still be stationary.[2] In practice, this means that even though the variance of a series may go through high and low periods, the unconditional (long-run, average) variance may still exist and be actually constant.[3] Second, one can read this chapter as a detailed survey of a variety of alternative models used to forecast variances. However, there is no logical contradiction in the fact that many different models have been developed and compared in the literature: in the end we only care for their forecasting performance, and it is possible that in alternative applications and sample periods, different models may turn out to outperform the remaining ones.

Section 2 starts by offering a motivating example that connects conditional heteroskedasticity to a few, easily checked and commonly observed empirical properties of financial returns. Section 3 introduces a few simple, in fact as simple as to be naive, variance models that have proven rather resilient in the practice of volatility forecasting, in spite of their sub-optimality in a statistical perspective. Section 4 represents the core of this chapter and contains material on forecasting volatility that is tantamount to you having ever attended a financial econometrics course: we introduce and develop several strands of the GARCH family. Section 5 presents a particularly

the process, because (under stationarity, i.e., $|\phi| < 1$) $E[X_{t+1}] = 0$ and $Var[X_{t+1}] = 1/(1-\phi^2)$, you know already that $X_{t+1} \sim \mathcal{N}(0, 1/(1-\phi^2))$ so that conditional and unconditional distributions will differ unless $\phi = 0$.

[2] Heuristically, stationarity of a stochastic process $\{X_t\}$ means that for every $k \geq 0$, $\{X_t\}_{t=k}^{\infty}$ has the same distribution as $\{X_t\}_{t=1}^{\infty}$. In words, this means that whatever is the point from which one starts sampling a time series process, the resulting overall (unconditional) distribution is unaffected by the choice: under stationarity, the implied distribution of returns over the last 20 years is the same as the distribution over 20 years of data to be sampled 10 years from now, say. Intuitively, this is related to the concept that a stationary time series will display "stable" long-run statistical properties—as summarized by its *unconditional distribution*—over time. Here the opposition between the unconditional natural of a distribution and time-varying *conditional variance* is important.

[3] However, if the unconditional variance of a time series is not constant, then the series is non-stationary.

useful and well-performing family of GARCH models that capture the evidence that past negative (shocks to) returns tend to increase the subsequent predicted variance more than positive (shocks to) returns do. Section 6 explains how models of conditional heteroskedasticity can be estimated in practice and leads to review some basic notions concerning maximum likelihood estimation and related inferential concepts and techniques. Section 7 explains how alternative conditional variance models may be evaluated and, in some ways, compared to each other. This seems to be particularly crucial because this chapter presents a range of different models, so that deciding whether a model is "good" plays a crucial role. Section 8 closes by introducing a more advanced GARCH model based on the intuition that the dynamics of variance in the short- vs. the long-run may be different. The Appendix presents a fully worked set of examples in Matlab.

2. One Motivating Example: Easy Diagnostic of Conditional Heteroskedasticity

As a motivating example, consider the (dividend-adjusted) realized *monthly* returns on a value-weighted index (hence, this is a portfolio) of all NYSE, AMEX, and NASDAQ stocks over the sample period is January 1972 - December 2009.[4] Even though this is not among the practice time series to be used in this class, the series is similar to the typical ones that appear in most textbooks.[5] Figure 1 plots the time series of the data.

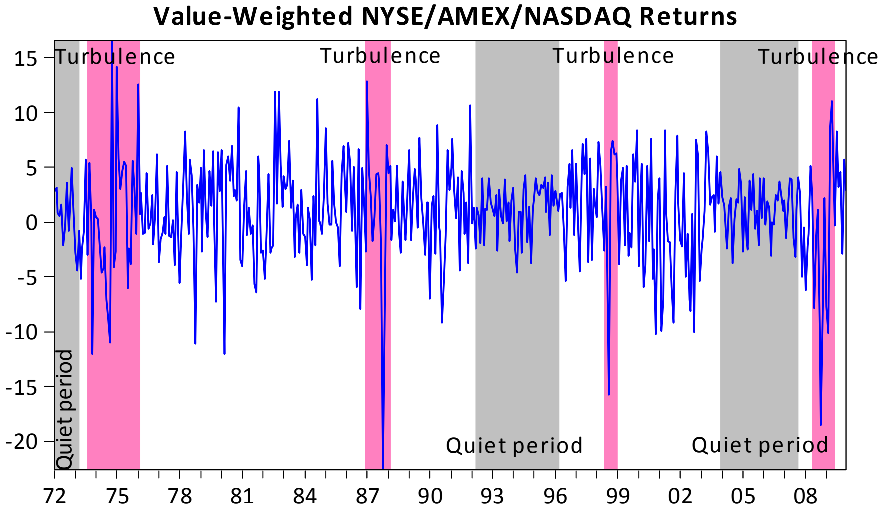


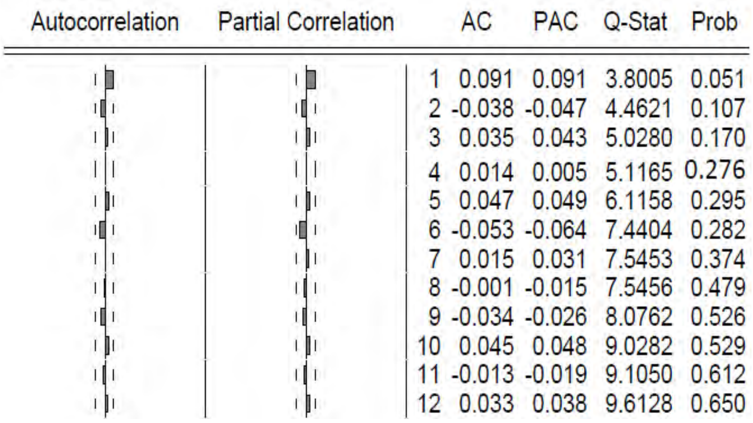
Figure 1: Value-weighted U.S. CRSP monthly stock returns

Visibly, volatility "clusters" in time: high (low) volatility tends to be followed by high (low) volatility. Casual inspection does have its perils, and formal testing is necessary to substantiate any first impressions. In fact, our objective in this chapter is to develop models that can fit this typical sequence of calm and turbulent periods. And especially forecast them.

[4] The data are compiled by the Center for Research in Security Prices (CRSP) and are available to the general public from the data repository by Kenneth French, at http://mba.tuck.dartmouth.edu/pages/faculty /ken.french/data_library.html.

[5] Do not worry: we shall take care of examining your typical class data during your MATLAB sessions as well as at the end of this chapter.

Let's now take this data a bit more seriously and apply the very methods of analysis that you have learned over the initial 5 weeks of Financial Econometrics II. As you know, a good starting point consists of examining the autocorrelogram of the series. Table 1 shows the autocorrelogram function (ACF), the partial autocorrelogram function (PACF), as well as new statistics introduced below, for the same monthly series in Figure 1.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.091	0.091	3.8005	0.051
		2	-0.038	-0.047	4.4621	0.107
		3	0.035	0.043	5.0280	0.170
		4	0.014	0.005	5.1165	0.276
		5	0.047	0.049	6.1158	0.295
		6	-0.053	-0.064	7.4404	0.282
		7	0.015	0.031	7.5453	0.374
		8	-0.001	-0.015	7.5456	0.479
		9	-0.034	-0.026	8.0762	0.526
		10	0.045	0.048	9.0282	0.529
		11	-0.013	-0.019	9.1050	0.612
		12	0.033	0.038	9.6128	0.650

Table 1: Serial correlation properties of value-weighted U.S. CRSP monthly stock returns

As you would expect of a series sampled at a relatively high frequency (such as monthly), there is weak serial correlation in U.S. stock returns. This lack of correlation means that, given past returns, the forecast of today's expected return is unaffected by knowledge of the past. However, more generally, the autocorrelation estimates from a standard ACF can be used to test the hypothesis that the process generating observed returns is a series of independent and identically distributed (IID) variables. The asymptotic (also called Bartlett's) standard error of the autocorrelation estimator is approximately $1/\sqrt{T}$, where T is the sample size. In table 1, such a constant $\pm 2/\sqrt{T}$ 95% confidence interval boundary is represented as the short vertical lines that surround the bars that represent the sample autocorrelation estimates also reported in the AC column of the table (these bars are to the left of the vertical line representing the 0 in the case of negative autocorrelations and to the right of the vertical zero-line in the case of positive autocorrelations).[6] Visibly, there is only one "bin"—in correspondence to the first lag, $\tau = 1$ (an AC of 0.091)—that touches the vertical line corresponding to the $2/\sqrt{T}$ upper limit of the 95% confidence interval; also in this case, because the upper limit is 0.094 and 0.091 fails to exceed it, the null hypothesis of $\rho_1 = 0$ can in principle not be rejected, although it is clear that we are close to the boundaries of the no-rejection area. However, for all other values of τ between 2 and 12, the table emphasizes that all sample autocorrelations fall inside the 95% confidence interval centered around a zero serial correlation, which is consistent with the absence of any serial correlation and hence independence of the series of monthly stock returns.

[6] To be precise, the 2 in the confidence interval statement $\pm 2/\sqrt{T}$ should be replaced by 1.96:

$$\Pr\{-1.96/\sqrt{T} \leq \rho_\tau \leq 1.96/\sqrt{T}\} = 0.95.$$

Notice that this confidence interval only obtains as an approximation, as $T \to \infty$.

However, as we shall see, the absence of serial correlation is insufficient to establish independence.[7]

2.1. *Testing the independence hypothesis and conditional heteroskedasticity*

The independence hypothesis can also be tested using the Portmanteau Q-statistic of Box and Pierce (1970), $\hat{Q}_k$, calculated from the first k autocorrelations of returns as:[8]

$$\hat{Q}_k \equiv T\sum_{\tau=1}^{k} \hat{\rho}_\tau^2 \overset{a}{\sim} \chi_k^2 \quad \text{where} \quad \hat{\rho}_\tau \equiv \frac{\sum_{t=1}^{T-\tau}(R_t - \bar{R})(R_{t+\tau} - \bar{R})}{\sum_{t=1}^{T-\tau}(R_t - \bar{R})^2}$$

(where $\tau > 0$). Here the notation $\overset{a}{\sim}$ means that asymptotically, as $T \rightarrow \infty$, the distribution of the $\hat{Q}_k$ statistic, under the null of an IID process (i.e., assuming that the null hypothesis holds), is chi-square, with k degrees of freedom.[9] In fact, the last two columns of table 1 report both $\hat{Q}_k$ for k between 1 and 12 and the corresponding p-value under a χ_k^2. In this case, the availability of 456 monthly observations lends credibility to the claim that, at least as an approximation, $\hat{Q}_k \sim \chi_k^2$. It is typically suggested to use values for the upper limit k up to $T/4$, although here we have simply set a maximum of $k = 12$ to save space. Consistently with earlier evidence for $\hat{\rho}_1 = 0.091$, the table shows that none of the levels of k experimented up to this point leads to rejecting the null hypothesis of IIDness of U.S. stock returns.

Does this evidence allows us to conclude that stock returns are (approximately) IID? Unfortunately not: it turns out that the *squares* and *absolute* values of stock and bond returns display high and significant autocorrelations. Here the conceptual point is that while

$$R_t \text{ is IID} \implies \hat{Q}_k \simeq 0 \text{ for all } k \geq 1$$

the opposite does not hold:

$$\hat{Q}_k \simeq 0 \text{ for all } k \geq 1 \not\Longrightarrow R_t \text{ is IID.}$$

The reason is that the definition of independence of a time series process has the following characterization:[10]

$$\begin{aligned} R_t \text{ is IID} \quad &\Longleftrightarrow \quad \hat{Q}_k^g \simeq 0 \text{ for all } k \geq 1 \\ \hat{Q}_k^g \quad &\equiv \quad T\sum_{\tau=1}^{k}(\hat{\rho}_\tau^g)^2 \overset{a}{\sim} \chi_k^2 \quad \text{where} \quad \hat{\rho}_\tau^g \equiv \frac{\sum_{t=1}^{T-\tau}(g(R_t) - \overline{g(R_t)})(g(R_{t+\tau}) - \overline{g(R_t)})}{\sum_{t=1}^{T-\tau}(g(R_t) - \overline{g(R_t)})} \end{aligned}$$

[7] Note that the fact that $\{X_t\}$ is independently distributed (over time) implies that the all autocorrelation coefficients $\rho_\tau = 0$, $\forall\tau \geq 1$, does not imply the opposite: even though $\rho_\tau = 0$, $\forall\tau \geq 1$, independence does not follow. We shall expand on this point below.

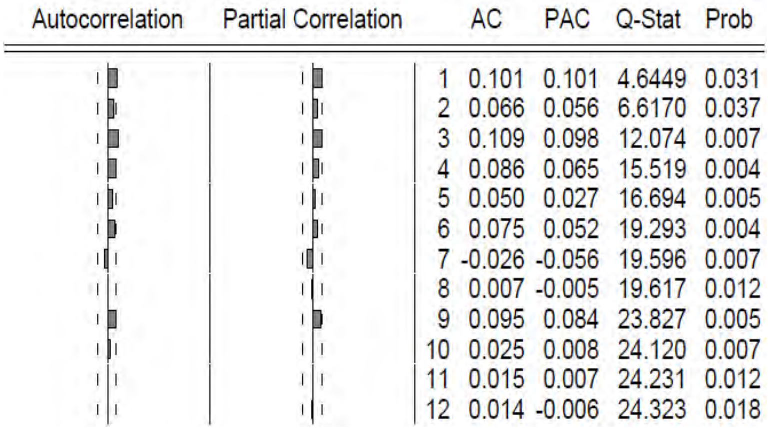
[8] We shall explain later the exact meaning of denoting portfolio returns as R_t.

[9] It is not surprising that the distribution of the test statistic ($\hat{Q}_k$) is derived assuming the null hypothesis holds: the goal is indeed to find sample evidence in the data to reject such a null hypothesis. Therefore the logical background is: are the data providing evidence inconsistent with the statistical properties that $\hat{Q}_k$ should possess under the null?

[10] Technically, one could even state that $Cov[g(R_t), h(R_{t+\tau})] = 0$ for any choice of sufficiently "smooth" functions $g(\cdot)$ and $h(\cdot)$ and $\forall\tau \neq 0$.

and $g(\cdot)$ is any (measurable) function that satisfies appropriate "smoothness" conditions. For instance, one may consider $g(x) = x^d$, where d is any positive integer and where $d > 1$ is admissible. Another alternative mentioned above is the case of the function $g(x) = |x|$, the absolute value transformation that turns negative real numbers into positive ones (and leaves positive real numbers unchanged). In practice, independence implies not only the absence of any serial correlation in the *level* of returns—i.e., in the first power of returns, $\hat{Q}_k \simeq 0$ for all $k \geq 1$—but it is equivalent to the absence of any serial correlations in all possible functions of returns, $g(R_t)$.

The high dependence in series of square and absolute returns proves that the returns process is not made up of IID random variables: appropriate functions of past returns do give information on appropriate functions of current and future returns. For instance, table 2 concerns the squares of value-weighted monthly U.S. CRSP stock returns and shows that in this case the sample autocorrelation coefficients of the squares are statistically significant (i.e., the null that these are zero can be rejected) for $\tau = 1$, 3, 4, and 9.[11]



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.101	0.101	4.6449	0.031
		2	0.066	0.056	6.6170	0.037
		3	0.109	0.098	12.074	0.007
		4	0.086	0.065	15.519	0.004
		5	0.050	0.027	16.694	0.005
		6	0.075	0.052	19.293	0.004
		7	-0.026	-0.056	19.596	0.007
		8	0.007	-0.005	19.617	0.012
		9	0.095	0.084	23.827	0.005
		10	0.025	0.008	24.120	0.007
		11	0.015	0.007	24.231	0.012
		12	0.014	-0.006	24.323	0.018

Table 2: Serial correlation properties of value-weighted squared U.S. monthly stock returns

Indeed $\hat{Q}_k^{sq}$ implies p-values below 0.05 (and often below 0.01, indicating strong statistical significance) for all values of k, and especially for $k \geq 3$ due to the large $\hat{\rho}_3^{sq} = 0.11$ (here the acronym '*sq*' refers to the fact that we are considering $g(x) = x^2$). The evidence in table 2 implies that large squared returns are more likely to be followed by large squared returns than small squared returns are. The fact that past *squared* returns predict subsequent *squared* returns—for instance, this is the meaning of $\hat{\rho}_3^{sq}$ being high and statistically significant (as it exceeds the 95% confidence bound threshold of 0.094)

$$\hat{\rho}_3^{sq} \equiv \frac{\sum_{t=1}^{T-\tau}(R_t^2 - \overline{R_t^2})(R_{t+\tau}^2 - \overline{R_t^2})}{\sum_{t=1}^{T-\tau}(R_t^2 - \overline{R_t^2})} \qquad (1)$$

—does not imply that past returns may predict subsequent returns: clearly, (1) may give a large

[11] The asymptotic distribution of the Box-Pierce $\hat{Q}_k^{sq}$ statistic applies if and only if the returns themselves are serially uncorrelated (in levels), i.e., if the null of $Q_k = 0$ cannot be rejected. This means that if one were to be uncertain about the fact that the zero mean assumption is correctly specified in $R_{t+1} = \sigma_{t+1}z_{t+1}$, this may imply that residuals are not serially uncorrelated so that one cannot simply apply portmanteau tests to test for the presence of ARCH. As stated, for most daily data series this does not represent a problem.

value even though

$$\hat{\rho}_3 \equiv \frac{\sum_{t=1}^{T-\tau}(R_t - \overline{R})(R_{t+\tau} - \overline{R})}{\sum_{t=1}^{T-\tau}(R_t - \overline{R})}$$

may be identically zero. This relates to a phenomenon that we have already commented in chapter 1: at (relatively) high frequencies, it is possible that higher-order moments—in this case, the second—may be strongly predictable even when the level of asset returns is not, so that they are well approximated by the simple model

$$R_{t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{D}(0,1),$$

where the fact that σ_{t+1} changes over time captures the predictability in squared returns that we have just illustrated.

At this point we face two challenges. First, and this is a challenge we are not about to pursue, one wonders what type of economic phenomenon may cause the predictability in squares (or more generally, in higher-order moments, as parameterized by a choice of $m \geq 3$ in $g(x) = x^m$), commonly referred to as *volatility clustering*, the fact that periods of high (low) squared returns tend to be followed by other periods of high (low) squared returns. Providing an answer to such a question is the subject of an exciting subfield of financial economics called asset pricing theory. In short, the general consensus in this field is that changes in the speed of flow of relevant—concerning either the exposures to risks or their prices—information to the market causes changes in price volatility that creates clusters of high and low volatility. However, this just moves the question of what may trigger such changes in the speed of information flows elsewhere. Although a range of explanations have been proposed (among them, the effects of transaction costs when trading securities, the fact that investors must learn the process of the fundamentals underlying asset prices in a complex and uncertain world, special features of investors' preferences such as habit formation and loss aversion, etc.) we will drop the issue for the time being. Second, given this evidence of volatility clustering, one feels a need to develop models in which volatility follows a stochastic process where today's volatility is positively correlated with the volatility of subsequent returns. This is what ARCH and GARCH models are for and what we set out to present in the following section.

3. Naive Models for Variance Forecasting

Consider the simple model for one asset (or portfolio) returns:[12]

$$R_{t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1). \tag{2}$$

Note that if we compare this claim to Section 2, we have specified the generic distribution $\mathcal{D}$ to be a normal distribution. We shall relax this assumption in the next chapter, but for the time being this will do for our goals. Here R_{t+1} is a continuously compounded return: the notation R is to be opposed to the lowercase r notation for returns that has appeared early on because we want to emphasize that R is generated by a model in which the expected return is zero: $E[R_{t+1}] = \sigma_{t+1}E[z_{t+1}] = \sigma_{t+1} \times 0 = 0$. Equivalently, at high frequency, we can safely assume that the mean value of R_{t+1} is zero as it is dominated by the standard deviation of returns. In fact, not only z_{t+1} is a pure random "shock" to returns, but z_{t+1} also has another interesting interpretation that will turn out to be useful later on:

$$z_{t+1} = \frac{R_{t+1}}{\sigma_{t+1}},$$

which implies that z_{t+1} is also a *standardized* return.[13] Note that in (**??**), z_{t+1} and σ^2_{t+1} are assumed to be statistically independent: this derives from the fact that σ^2_{t+1} is a conditional variance function that—at least in our treatment—only depends on past information, i.e., $\sigma^2_{t+1} \equiv Var[R_{t+1}|\mathcal{F}_t]$.

A model in which $E[R_{t+1}] = 0$ is an acceptable approximation when applied to daily data. Absent this assumption, a more realistic model would be instead

$$R_{t+1} = \mu_{t+1} + \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1),$$

where $\mu_{t+1} \equiv E_t[R_{t+1}]$. In this case, $z_{t+1} = (R_{t+1} - \mu_{t+1})/\sigma_{t+1}$. This model will reappear in our concerns in later chapters. How do you test whether μ_{t+1} or, more concretely, $\mu = 0$ or not? This is a standard test of a mean, see your notes from any undergraduate statistics sequence.[14]

[12] We shall be modelling asset or portfolio *returns*, and never prices! This is important, because the absence of serial correlation in returns means that a good model for returns is indeed (ignoring the mean and any dividends or interim cash flows) $R_{t+1} = \log(P_{t+1}) - \log(P_t) = \sigma_{t+1}z_{t+1}$, which implies that $\log(P_{t+1}) = \log(P_t) + \epsilon_{t+1}$, i.e., (the log of) prices tend to follow a random walk. Because (log-)asset prices are I(1) process, they contain a stochastic trend, to analyze them without first removing the trend is always unwieldy and often plainly incorrect. Incorrect here means that most of the tests and inferential procedures you have been taught apply only—except for major and complicated corrections, if any—to stationary series, not to I(1) series. This also means that in most cases there is only one type of econometrics that can be applied to the *prices* of assets or portfolios, the wrong one—the one we should never hear about in MSc. theses, for instance.

[13] You will recall that if X_t is characterized by an expectation of $E[X_{t+1}]$ and a variance of $Var[X_{t+1}]$, the standardized version of the variable is:

$$\frac{X_{t+1} - E[X_{t+1}]}{\sqrt{Var[X_{t+1}]}}.$$

Clearly, if $E[X_{t+1}] = 0$ the standardization simply involves scaling X_{t+1} by its standard deviation. Note that standardization may also apply in *conditional* terms: if $E_t[X_{t+1}] \equiv E[X_{t+1}|\mathcal{F}_t]$ and $Var_t[X_{t+1}] \equiv Var[X_{t+1}|\mathcal{F}_t]$, where $\mathcal{F}_t$ is the information set at time t, then the conditional standardized variable is: $(X_{t+1} - E_t[X_{t+1}])/\sqrt{Var_t[X_{t+1}]}$.

[14] Right, you cannot find your notes or textbooks now. OK then: the null hypothesis is $\mu = 0$ and the test statistic

3.1. *Rolling window variance model*

The easiest way to capture volatility clustering is by letting tomorrow's variance be the simple average of the most recent m squared observations, as in

$$\sigma^2_{t+1} = \frac{1}{m}\sum_{\tau=1}^{T} R^2_{t+1-\tau} = \sum_{\tau=1}^{T}\frac{1}{m}R^2_{t+1-\tau}. \tag{3}$$

This variance prediction function is simply a constant-weight sum of m past squared returns.[15] This is called a *rolling window variance forecast model.* However, the fact that the model puts equal weights (equal to $1/m$) on the past m observations often yields unwarranted and hard to justify results. Figure 2 offers a snapshot of the problems associated with rolling window variance models. The figure concerns S&P 500 daily data and uses a rolling window of 25 observations, $m = 25$. The figure emphasizes that, when plotted over time, predicted rolling window variance exhibits box-shaped patterns: An extreme return (positive or negative) today will bump up variance by $1/m$ times the return squared for exactly m periods after which variance immediately drops back down.

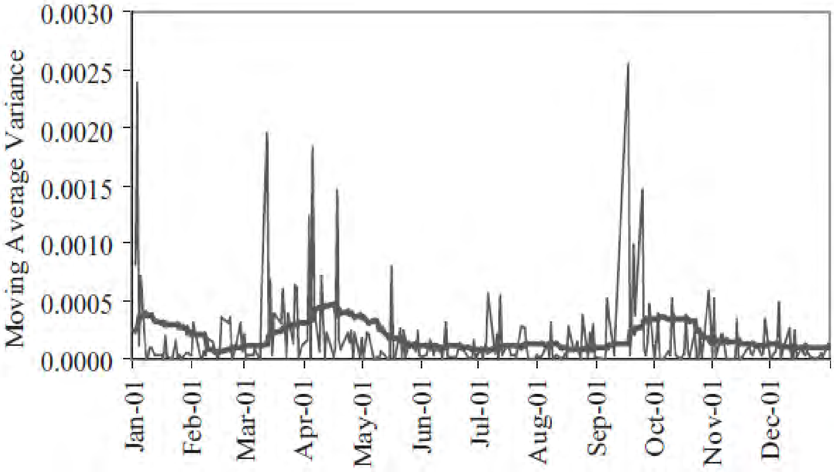


Figure 2: Squared S&P500 returns with moving average variance estimate (bold), $m = 25$

However, such extreme gyrations—especially the fact that predicted variance suddenly declines after 25 periods—does not reflect the economics of the underlying financial market. It is instead just caused by the mechanics of the volatility model postulated in (3). This brings us to the next issue: given that m has such a large impact on the dynamics of predicted variance, one wonders how m should be selected and whether any optimal choice may be hoped for. In particular, it is

(when the variance is unknown) is:

$$t_\mu = \frac{\bar{R}}{\sqrt{S^2/T}} \sim t_{T-1},$$

where $\bar{R}$ is the sample mean and S^2 is the sample variance. Alternatively, simply estimate a regression of returns on just an intercept and test whether the constant coefficient is statistically significant at a given, chosen size of the test.

[15] Because we have assumed that returns have zero mean, note that when predicting variance we do not need to worry about summing or weighing squared deviations from the mean, as in general the definition of variance would require.

clear that a high m will lead to an excessively smoothly evolving σ^2_{t+1}, and that a low m will lead to an excessively jagged pattern of σ^2_{t+1}. Unfortunately, in the financial econometrics literature no compelling or persuasive answer has been yet reported.

3.2. *Exponential variance smoothing: the RiskMetrics model*

Another reason for dissatisfaction is that typically the sample autocorrelation plots/functions of squared returns suggest that a more gradual decline is warranted in the effect of past returns on today's variance, see table 2.

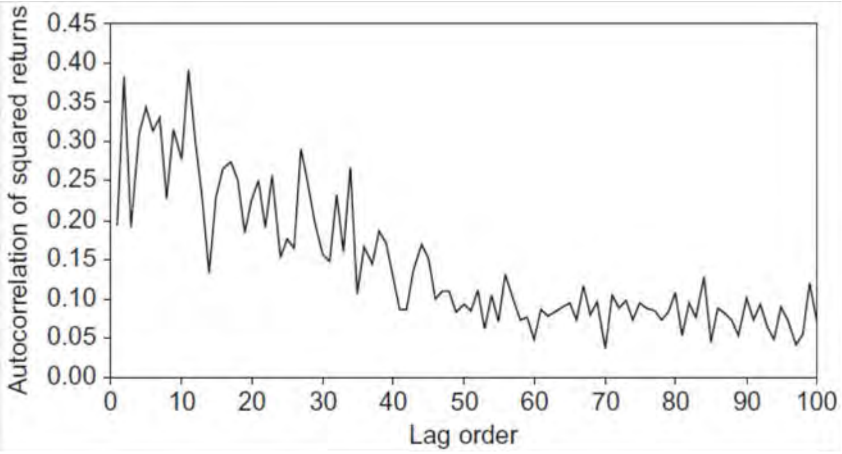


Figure 3: Autocorrelation of squared daily S&P 500 returns

To make this point more persuasively (and waiting for our own evidence from the Matlab sessions), observe now figure 3, concerning *daily* S&P 500 returns (table 2 concerned instead monthly value-weighted U.S. stock returns). The sample underlying the sample calculations in the figure is January 1, 2010–December 31, 2010. Clearly, in the figure sample autocorrelations decline rather slowly (in spite the inevitable sample variation of all estimators) from initial levels of $\hat{\rho}^{sq}_{\tau}$ of 0.25-0.30 for small values of τ to values below 0.10 when τ exceeds 50. A more interesting model that takes this evidence into account when computing forecasts of variance is JP Morgan's RiskMetrics system:

$$\sigma^2_{t+1} = (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau-1}R^2_{t+1-\tau} \qquad \lambda \in (0,1). \tag{4}$$

In this model, the weight on past squared returns declines exponentially as we move backward in time: 1, λ, λ^2, ...[16] Because of this rather specific mathematical structure, the model is also called the exponential variance smoother. Exponential smoothers have a long tradition in econometrics

[16] However, the weights do sum to 1, as you would expect them to do. In fact, this is the role played by the factor $(1-\lambda)$ that multiplies the infinite sum $\sum_{\tau=1}^{\infty}\lambda^{\tau-1}R^2_{t+1-\tau}$. Noting that because the sum of a geometric series is $\sum_{\tau=0}^{\infty}\lambda^{\tau} = 1/(1-\lambda)$, we have

$$\sum_{\tau=1}^{\infty}\kappa_{\tau} = \sum_{\tau=1}^{\infty}(1-\lambda)\lambda^{\tau-1} = (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau-1} = (1-\lambda)\sum_{\tau=0}^{\infty}\lambda^{\tau} = (1-\lambda)\frac{1}{(1-\lambda)} = 1,$$

where $\kappa_{\tau} \equiv (1-\lambda)\lambda^{\tau-1}$ for $\tau \geq 1$.

and applied forecasting because they are known to provide rather accurate forecasts of the level of time series. JP Morgan's RiskMetrics desk was however rather innovative in thinking that such a model could also provide good predictive accuracy when applied to second moments of financial time series.

(4) does not represent either the most useful or the most common way in which the RiskMetrics model is presented and used. Because for $\tau = 1$ we have $\lambda^0 = 1$, it is possible to re-write it as:

$$\sigma^2_{t+1} = (1-\lambda)R^2_t + (1-\lambda)\sum_{\tau=2}^{\infty}\lambda^{\tau-1}R^2_{t+1-\tau} = (1-\lambda)R^2_t + (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau}R^2_{t-\tau}.$$

Yet it is clear that

$$\sigma^2_t = (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau-1}R^2_{t-\tau} = \frac{1}{\lambda}(1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau}R^2_{t-\tau}.$$

Substituting this expression into $\sigma^2_{t+1} = (1-\lambda)R^2_t + (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau}R^2_{t-\tau}$, gives

$$\begin{aligned}\sigma^2_{t+1} &= (1-\lambda)R^2_t + \frac{\lambda}{\lambda}(1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau}R^2_{t-\tau} \\ &= (1-\lambda)R^2_t + \lambda\left[\underbrace{\frac{1}{\lambda}(1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau}R^2_{t-\tau}}_{=\sigma^2_t}\right] \\ &= (1-\lambda)R^2_t + \lambda\sigma^2_t. \qquad (5)\end{aligned}$$

(5) implies that forecasts of time $t+1$ variance are obtained as a weighted average of today's variance and of today's squared return, with weights λ and $1-\lambda$, respectively.[17] In particular, notice that

$$\lim_{\lambda\to 1^-}\sigma^2_{t+1} = \sigma^2_t,$$

i.e., as $\lambda \to 1^-$ (a limit from the left, given that we have imposed the restriction that $\lambda \in (0,1)$) the process followed by conditional variance becomes a constant, in the sense that $\sigma^2_{t+1} = \sigma^2_t = \sigma^2_{t-1} = ... = \sigma^2_0$. The naive idea that one can simply identify the forecast of time $t+1$ variance as the squared return of R_t, corresponds instead to the case of $\lambda \to 0^+$.

The RiskMetrics model in (5) presents a number of important advantages:

1. (4) is a sensible formula as it implies that recent returns matter more for predicting tomorrow's variance than distant returns do; this derives from $\lambda \in (0,1)$ so that λ^τ gets smaller when the

[17] One of your TAs has demanded that also the following, equivalent formulation be reported: $\sigma^2_{t+1|t} = (1-\lambda)R^2_t + \lambda\sigma^2_t$, where $\sigma^2_{t+1|t}$ emphasizes that this is the forecast of time $t+1$ variance given the time t information set. This notation will also appear later on in the chapter.

lag coefficient, τ, gets bigger. Figure 4 show the behavior of this weight as a function of τ.

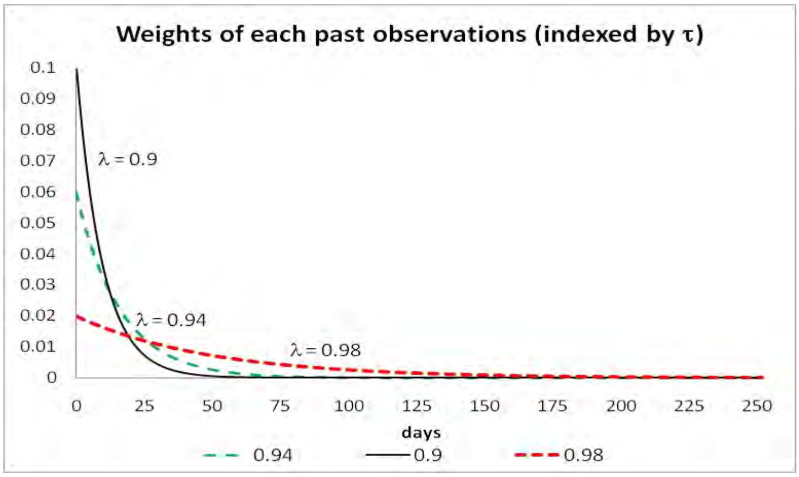


Figure 4

2. (5) only contains one unknown parameter, λ, that we will have to estimate. In fact, after estimating λ on a large number of assets, RiskMetrics found that the estimates were quite similar across assets, and therefore suggested to simply set λ for every asset and daily data sets to a typical value of 0.94. In this case, no estimation is necessary.[18]

3. Little data need to be stored in order to calculate and forecast tomorrow's variance; in fact, for values of λ close to the 0.94 originally suggested by RiskMetrics, it is the case that after including 100 lags of squared returns, the cumulated weight is already close to 100%. This is of course due to the fact that, once σ_t^2 has been computed, past returns beyond the current squared return R_t^2, are not needed. Figure 5 shows the behavior of the cumulative weight for a fixed number of past observations as a function of λ.

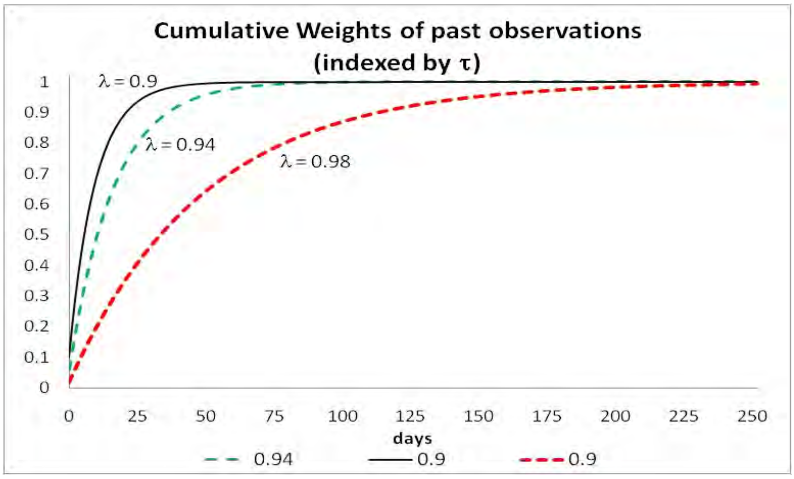


Figure 5

Given all these advantages of the RiskMetrics model, why not simply end the discussion on variance forecasting here?

[18] We shall see later in this chapter that maximum likelihood estimation of λ tends to provide estimates that hardly fall very far from the classical RiskMetrics $\lambda = 0.94$.

4. Generalized Autoregressive Conditional Heteroskedastic (GARCH) Variance Models

The RiskMetrics model has a number of shortcomings, but these can be understood only after introducing ARCH(q) models, where ARCH is the celebrated acronyms for Autoregressive Conditional Heteroskedastic. Historically, ARCH models were the first-line alternative developed to compete with exponential smoothers and one quick glance at their functional form reveals why. In the zero-mean return case, their structure is very simple:

$$\sigma_{t+1}^2 = \omega + \alpha R_t^2.$$

In particular, this is a simple, plain-vanilla ARCH(1) process and it implies that

$$R_{t+1} = \left(\sqrt{\omega + \alpha R_t^2}\right) z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1).$$

The intuition of this model is immediate: the appearance of $\alpha R_t^2 > 0$ (if $\alpha > 0$, as we shall impose later) is what captures the clustering intuition that large movements in asset prices tend to follow large movements, of either sign (as the square function only produces positive contributions). The impact of past large movements in prices will be large if α is large. In fact, as $\alpha \to 1^-$ (from the left, as we will see that $\alpha < 1$), any return (shock) will cause an impact on subsequent variances that is nearly permanent.

The differences vs. (5), $\sigma_{t+1}^2 = (1-\lambda)R_t^2 + \lambda\sigma_t^2$, are obvious. On the one hand, RiskMetrics can be taken as a special case of ARCH(1) in which $\omega = 0$; on the other hand, it is clear that an exponential smoother does not only attach a weight $(1-\lambda)$ to current squared return, but also a weight λ on current variance, σ_t^2. The fact that the good performance of RiskMetrics mentioned above is based on both R_t^2 and σ_t^2 makes it less than surprising the fact that, historically, it became soon obvious that just using one lag of past squared returns would not be sufficient to produce accurate forecasts of variance: for most assets and sample periods there is indeed evidence that one needs to use a large number $q > 1$ of lags on the right-hand side (RHS) of the ARCH(q) representation:

$$\sigma_{t+1}^2 = \omega + \sum_{i=1}^{q} \alpha_i R_{t+1-i}^2. \qquad (6)$$

Yet, even though it is simple, in statistical terms ARCH(q) is not as innocuous as it may seem: maximum likelihood estimation of models of the type (6) implies nonlinear parameter estimation, on which some details will be provided later. It is easy to find the unconditional, long-run variance under (6). Because (20) implies that $E[R_{t+1}^2] = E[\sigma_{t+1}^2 z_{t+1}^2] = E[\sigma_{t+1}^2]E[z_{t+1}^2] = E[\sigma_{t+1}^2] \times 1 =$

$E[\sigma^2_{t+1}]$, setting $\bar{\sigma}^2 \equiv E[\sigma^2_{t+1}] = E[R^2_{t+1-i}]$ $\forall i$:[19]

$$\begin{aligned}\bar{\sigma}^2 &= E[\sigma^2_{t+1}] = \omega + \sum_{i=1}^{q} \alpha_i E[R^2_{t+1-i}] = \omega + \sum_{i=1}^{q} \alpha_i \bar{\sigma}^2 \\ &= \omega + \bar{\sigma}^2 \sum_{i=1}^{q} \alpha_i \Longrightarrow \bar{\sigma}^2 = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i}. \qquad (7)\end{aligned}$$

Because unconditional variance makes sense (technically, we say that it exists, i.e., it is defined) only when $\bar{\sigma}^2 > 0$, (7) implies that when $\omega > 0$, the condition

$$1 - \sum_{i=1}^{q} \alpha_i > 0 \Longrightarrow \sum_{i=1}^{q} \alpha_i < 1$$

must hold. When the long-run, unconditional variance of a ARCH process exists, because in a ARCH model the only source of time-variation in conditional moments comes from the variance, we say that the ARCH process is stationary and we also refer to the condition $\sum_{i=1}^{q} \alpha_i < 1$ as a stationarity condition. Moreover, because also existence of conditional variances requires that $\sigma^2_{t+1} > 0$, the additional restrictions that $\omega > 0$ and α_1, α_2, ..., $\alpha_q > 0$ are usually added both in theoretical work and in applied estimation.

4.1. *Inside the box: basic statistical properties of a simple AR(1)-ARCH(1) model*

To get a concrete grip of the statistical implications of ARCH modelling and of the possible interactions between conditional mean and conditional variance functions, consider the simplest possible ARCH model with some structure in its conditional mean function, i.e., a Gaussian AR(1)-ARCH(1) model:

$$R_{t+1} = [\phi_0 + \phi_1 R_t] + \left[\omega + \alpha \epsilon_t^2\right]^{1/2} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1),$$

where $|\phi_1| < 1$, $0 < \alpha < 1$, while $\omega > 0$ keeps variance well-defined and

$$\epsilon_t \equiv \left[\omega + \alpha \epsilon_{t-1}^2\right]^{1/2} z_t.$$

Notice that in this model we are temporarily removing the assumption that $\mu_{t+1} = 0$. In a way, this is to show you why this assumption had been introduced in the first place: if $\mu_{t+1} \neq 0$, even with very simple conditional heteroskedastic models, things get considerably complicated. For instance, the ARCH process is no longer simply defined in terms of one lag of returns, R^2_{t-1}, but instead in terms of ϵ^2_{t-1}. The Gaussian AR(1)-ARCH(1) model has to be compared with the homoskedastic Gaussian AR(1)process

$$R_{t+1} = [\phi_0 + \phi_1 R_t] + [\omega]^{1/2} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1),$$

[19] $E[\sigma^2_{t+1} z^2_{t+1}] = E[\sigma^2_{t+1}]E[z^2_{t+1}]$ derives from the fact that z_{t+1} and σ^2_{t+1} are statistically independent. On its turn, this derives from the fact that σ^2_{t+1} is a conditional variance function that only depends on past information, i.e., $\sigma^2_{t+1} \equiv Var[R_{t+1}|\mathcal{F}_t]$. $E[\sigma^2_{t+1}]E[z^2_{t+1}] = E[\sigma^2_{t+1}]$ comes then from the fact that if z_{t+1} $\sim$IID $\mathcal{N}(0,1)$, then $E[z^2_{t+1}] = Var[z_{t+1}] = 1$.

you are already familiar with from the first part of the course. Assume that z_t is independent of $\epsilon_{t-1}, \epsilon_{t-2}, .., \epsilon_0$.

Consider first the total residual of the process, i.e., $\epsilon_t \equiv \sigma_t z_t = \left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2} z_t$. We show that the process for the total residuals, denoted $\{\epsilon_t\}$, has zero mean and is serially uncorrelated at all lags $j \geqslant 1$. This can be seen from

$$\begin{aligned}
E[\epsilon_t] &= E\left[\left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2} z_t\right] = \overbrace{E\left[\left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2}\right]\underbrace{E[z_t]}_{=0}}^{\text{from independence of } z_t \text{ from } \epsilon_{t-1},\epsilon_{t-2},..,\epsilon_0} \\
&= E\left[\left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2}\right]0 = 0 \\
E[\epsilon_t\epsilon_{t-j}] &= E\left[\left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2}\left[\omega + \alpha\epsilon_{t-1-j}^2\right]^{1/2} z_t z_{t-j}\right] \\
&= \overbrace{E\left[\left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2}\left[\omega + \alpha\epsilon_{t-1-j}^2\right]^{1/2}\right] \quad \underbrace{E[z_t z_{t-j}]}_{=0 \text{ b/c } z_{t+1}\sim IID\ N(0,1)}}^{\text{from independence of } z_t \text{ from } \epsilon_{t-1},\epsilon_{t-2},..,\epsilon_0} \\
&= E\left[\left[\omega + \epsilon_{t-1}^2\right]^{1/2}\left[\omega + \alpha\epsilon_{t-1-j}^2\right]^{1/2}\right]0 = 0 \qquad (j \geqslant 1).
\end{aligned}$$

This property is important because it provides guarantees (necessary and sometimes sufficient conditions) to proceed to the estimation of the conditional mean function using standard methods, such as OLS. Yet, $\{\epsilon_t\}$ has a finite unconditional variance of $\omega/(1-\alpha)$. This can be seen from

$$\begin{aligned}
E\left[\epsilon_t^2\right] &= E\left[(\omega + \alpha\epsilon_{t-1}^2)z_t^2\right] = E\left[\omega + \alpha\epsilon_{t-1}^2\right]\underbrace{E\left[z_t^2\right]}_{=1} \\
&= \omega + \alpha E\left[\epsilon_{t-1}^2\right] = \omega + \alpha E\left[\epsilon_t^2\right] \\
E\left[\epsilon_t^2\right] &= Var[\epsilon_t] = \omega/(1-\alpha).
\end{aligned}$$

This iterates a point made above already: ARCH does not imply non-stationarity, and in fact a finite long-run, average, unconditional variance exists, although it diverges to $+\infty$ as $\alpha \to 1^-$. It is also easy to prove that the conditional process for total residuals, $\{\epsilon_t | \epsilon_{t-1}, \epsilon_{t-2}, ...\}$, has a zero conditional mean and a conditional variance of $\omega + \alpha\epsilon_{t-1}^2$:

$$\begin{aligned}
E[\epsilon_t|\epsilon_{t-1},...] &= E_{t-1}\left[\left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2} z_t\right] = \overbrace{E_{t-1}\left[\left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2}\right]E_{t-1}[z_t]}^{\text{from independence of } z_t \text{ from } \epsilon_{t-1},\epsilon_{t-2},..,\epsilon_0} \\
&= \left[\omega + \alpha\epsilon_{t-1}^2\right]^{1/2} 0 = 0 \\
E\left[\epsilon_t^2|\epsilon_{t-1},...\right] &= E_{t-1}\left[\left[\omega + \alpha\epsilon_{t-1}^2\right]z_t^2\right] = \overbrace{\left[\omega + \alpha\epsilon_{t-1}^2\right]\underbrace{E_{t-1}\left[z_t^2\right]}_{=1}}^{\text{from independence of } z_t \text{ from } \epsilon_{t-1},\epsilon_{t-2},..,\epsilon_0} \\
&= \left[\omega + \alpha\epsilon_{t-1}^2\right]1 = \omega + \alpha\epsilon_{t-1}^2 = Var_{t-1}[\epsilon_t].
\end{aligned}$$

This confirms what we have stated early on about the typical properties of financial data: under ARCH, shocks may be serially uncorrelated as $E[\epsilon_t\epsilon_{t-j}] = 0$ but they are not independent because $E\left[\epsilon_t^2|\epsilon_{t-1},...\right] = \omega + \alpha\epsilon_{t-1}^2$.

Finally, let's verify that the famous Wold's representation theorem that you have encountered in the first part of this course—by which any AR(q) process can be represented as an infinite MA process—also applies to ARCH(1) models.[20] By a process of recursive substitution, we have:

$$\epsilon_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \eta_t = \omega + \alpha \overbrace{\left[\omega + \alpha\epsilon_{t-2}^2 + \eta_{t-1}\right]}^{\text{from } \epsilon_{t-1}^2 = \omega + \alpha\epsilon_{t-1}^2 + \eta_t} + \eta_t = \omega\left(1+\alpha\right) + \alpha^2\epsilon_{t-2}^2 + \left[\eta_t + \alpha\eta_{t-1}\right]$$

$$\begin{aligned} &= \omega\left(1+\alpha+\alpha^2\right) + \alpha^3\epsilon_{t-3}^2 + \left[\eta_t + \alpha\eta_{t-1} + \alpha^2\eta_{t-2}\right] = ... \\ &= ... = \omega\sum_{j=0}^{t-1}\alpha^j + \alpha^t\epsilon_0^2 + \sum_{j=0}^{t-1}\alpha^j\eta_{t-j}. \end{aligned}$$

This means that if the return series had started in the sufficiently "distant" past or, equivalently, as $t \rightharpoonup +\infty$ this is indeed an $MA(\infty)$ process, $\epsilon_t^2 = [\omega/(1-\alpha)] + \eta_t + \alpha\eta_{t-1} + \alpha^2\eta_{t-2} + \alpha^3\eta_{t-3} + ...$ Note that $\lim_{t\to\infty}\omega\sum_{j=0}^{t-1}\alpha^j = \omega/(1-\alpha)$ because for $\alpha < 1$, $\sum_{j=0}^{\infty}\alpha^j$ is a convergent geometric series.

4.2. *GARCH(q, p) models*

Although you may not see that yet, (6) has the typical structure of a AR(q) model. To see this note two simple facts. First given any random variable X_{t+1}, notice that the variable can always be decomposed as the sum of its conditional expectation plus a zero-mean white noise shock:

$$X_{t+1} = E_t[X_{t+1}] + \epsilon_{t+1}.$$

Hence applying this principle to square asset returns, one has $R_{t+1}^2 = E_t[R_{t+1}^2] + \eta_{t+1}$. Second, from the definition of conditional variance and the fact that $E_t[R_{t+1}] = 0$, we have that $\sigma_{t+1}^2 \equiv Var_t[R_{t+1}]$ $= E_t[R_{t+1}^2]$. Therefore, putting these two simple facts together, we have:

$$\begin{aligned} R_{t+1}^2 &= E_t[R_{t+1}^2] + \eta_{t+1} = \sigma_{t+1}^2 + \eta_{t+1} \\ &= \omega + \sum_{i=1}^{q}\alpha_i R_{t+1-i}^2 + \eta_{t+1}. \end{aligned}$$

Surprise: this is a standard AR(q) model for squared asset returns! At this point, if you have paid *some* attention to what has happened in the last 5 weeks, you know where to look for when it comes to generalize and improve the predictive performance of an AR(q) model: ARMA(q, p) models.

Before proceeding to that, we dig a bit deeper on this AR(q) characterization of ARCH by showing—at least for the simple case of AR(1)-ARCH(1), when the algebra is relatively simple—that the autocorrelogram of the series of squared shocks $\left\{\epsilon_t^2\right\}$ implied by an ARCH(1) decays at

[20] Here we use a property that $\epsilon_t^2 = \sigma_t^2 + \eta_t$ so that $\epsilon_t^2 = \omega + \alpha_1\epsilon_{t-1}^2 + \eta_t$ derived in next subsection. This just means that in a ARCH model, squared shocks follow an AR(1) process (hence the "AR" in ARCH). Apologies for running ahead, just take this property as a fact for the time being.

speed $(\alpha)^j$. Note that under a ARCH(1), the forecast error when predicting squared residuals is (note that $\epsilon_t = R_t$ when the conditional mean is zero, i.e., $\phi_0 = \phi_1 = 0$):

$$\eta_t = \epsilon_t^2 - E_{t-1}\left[\epsilon_t^2\right] = \epsilon_t^2 - \sigma_t^2 \qquad \sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2.$$

Therefore $\epsilon_t^2 = \sigma_t^2 + \eta_t$ or $\epsilon_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \eta_t$ which is an AR(1) process for squared innovations to financial returns. This implies that the autocorrelogram for the series of squared shocks $\{\epsilon_t^2\}$ from an ARCH(1) decays at speed $(\alpha)^\tau$, because of the properties of autoregressive processes seen in the first part of the course. Here τ is the order of the autocorrelogram, i.e., the lag in $Cov\left[\epsilon_t^2, \epsilon_{t-\tau}^2\right]/Var\left[\epsilon_t^2\right]$, the implication is that unless $\alpha \rightharpoonup 1$, the autocorrelogram of a ARCH(1) will decay very quickly. See for instance the simulations below in figure 6.

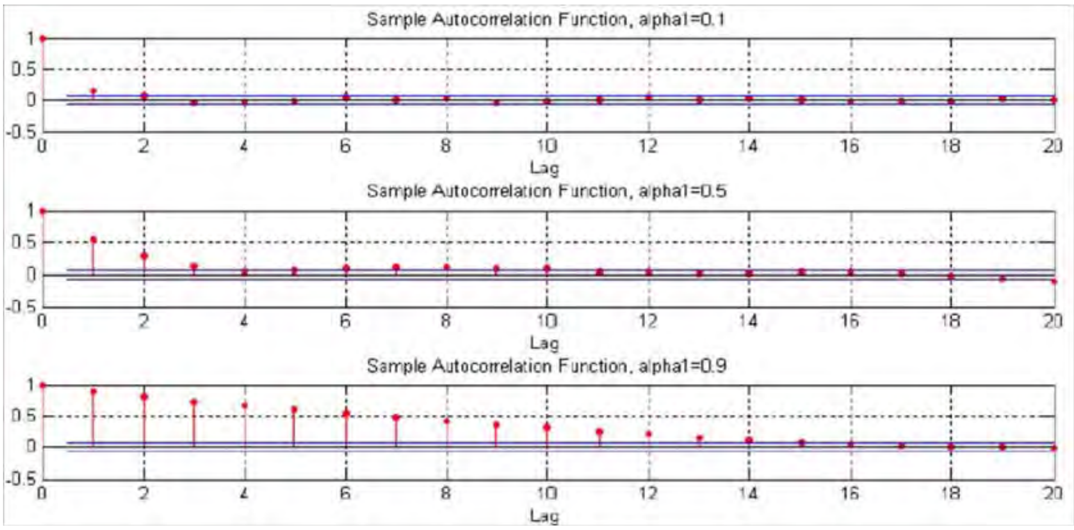


Figure 6: Simulated sample autorecorrelation function for alternative choices of α (0.1, 0.5, 0.9)

As far as the ARMA extensions are concerned, the simplest generalized autoregressive conditional heteroskedasticity (GARCH(1,1)) model is:

$$\sigma_{t+1}^2 = \omega + \alpha R_t^2 + \beta\sigma_t^2. \tag{8}$$

which yields a model for returns given by $R_{t+1} = (\sqrt{\omega + \alpha R_t^2 + \beta\sigma_t^2} z_{t+1})$, where $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$. More generally, in the ARMA(q,p) case, we have:

$$\sigma_{t+1}^2 = \omega + \sum_{i=1}^{q} \alpha_i R_{t+1-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t+1-j}^2. \tag{9}$$

Similarly to the steps followed in the ARCH(q) case, setting $\bar{\sigma}^2 \equiv E[\sigma_{t+1}^2]$:[21]

$$\begin{aligned}\bar{\sigma}^2 &= E[\sigma_{t+1}^2] = \omega + \sum_{i=1}^{q}\alpha_i E[R_{t+1-i}^2] + \sum_{j=1}^{p}\beta_j E[\sigma_{t+1-j}^2] = \omega + \sum_{i=1}^{q}\alpha_i\bar{\sigma}^2 + \sum_{j=1}^{p}\beta_j\bar{\sigma}^2 \\ &= \omega + \bar{\sigma}^2\left(\sum_{i=1}^{q}\alpha_i + \sum_{j=1}^{p}\beta_j\right) \Longrightarrow \bar{\sigma}^2 = \frac{\omega}{1 - \sum_{i=1}^{q}\alpha_i - \sum_{j=1}^{p}\beta_j}.\end{aligned}$$

[21] The following derivation exploits the fact that $\bar{\sigma}^2 = E[\sigma_{t+j}^2]$ $\forall j \geq 0$. This is true of any stationary process: its properties do not depend on the exact indexing of the time series under investigation.

Because unconditional variance exists only if $\bar{\sigma}^2 > 0$, the equation above implies that when $\omega > 0$, the condition

$$1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j > 0 \Longrightarrow \sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$$

must hold. When the long-run (i.e. ergodic) variance of a GARCH process exists, because in a GARCH model the only source of time-variation in conditional moments comes from the variance, we say that the GARCH process is stationary and we also refer to the condition $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$ as a stationarity condition. Moreover, because also existence of conditional variances requires that $\sigma_{t+1}^2 > 0$, the additional restrictions that $\omega > 0$, α_1, α_2, ..., $\alpha_q > 0$, β_1, β_2, ..., $\beta_p > 0$ are usually added both in theoretical work and in applied estimation. Of course in the $q = p = 1$ case, such restrictions are simply $\omega > 0$, $\alpha > 0$, $\beta > 0$ and $\alpha + \beta < 1$.

Even though they are straightforward logical extensions of GARCH(1,1), rich GARCH(q, p) models with q and p exceeding 1 are rarely encountered in practice (but see section 8 for one important exception). This occurs not only because most data sets do not seem to strongly need the specification of higher-order lags q and p in GARCH models, but also because in practical estimation so many constraints have to be imposed to ensure that variance is positive and the process stationary, that numerical optimization may often be problematic. It is natural to ask why can it be that a simple GARCH(1,1) is so popular and successful? This is partly surprising because one of the problems with the early ARCH literature in the 1980s, consisted of the need to pick relatively large values of q with all the estimation and numerical problem that often ensued. The reason for the success of simple GARCH(1,1) models is that these can be shown to be equivalent to an ARCH(∞) model! Notice that by recursive substitution,

$$\begin{aligned}
\sigma_{t+1}^2 &= \omega + \alpha R_t^2 + \beta \sigma_t^2 = \omega + \alpha R_t^2 + \beta [\underbrace{\omega + \alpha R_{t-1}^2 + \beta \sigma_{t-1}^2}_{\sigma_t^2}] = \omega(1+\beta) + \alpha(1+\beta) R_{t-1}^2 + \beta^2 \sigma_{t-1}^2 \\
&= \omega(1+\beta) + \alpha(1+\beta) R_{t-1}^2 + \beta^2 [\underbrace{\omega + \alpha R_{t-2}^2 + \beta \sigma_{t-2}^2}_{\sigma_{t-1}^2}] \\
&= \omega(1+\beta+\beta^2) + \alpha R_t^2 + \alpha\beta R_{t-1}^2 + \alpha\beta^2 R_{t-2}^2 + \beta^3 \sigma_{t-2}^2 \\
&= ... = \omega \sum_{j=0}^{\infty} \beta^j + \alpha \sum_{j=0}^{\infty} \beta^j R_{t-j}^2 + \lim_{j \to +\infty} \beta^j \sigma_{t-j}^2. \qquad (10)
\end{aligned}$$

If the return series had started in the sufficiently "distant" past or, equivalently, when $t \to \infty$, so that

$$\lim_{j \to +\infty} \beta^j \sigma_{t-j}^2 = 0$$

which is implied by $\alpha + \beta < 1$ or $\beta < 1 - \alpha < 1$ (as $\alpha > 0$), (10) is an ARCH(∞) with a particular structure of decaying power weights, given by $\alpha \sum_{j=0}^{\infty} \beta^j$. Because $0 < \beta < 1$ implies that

$$\omega \sum_{j=0}^{\infty} \beta^j = \frac{\omega}{1 - \beta},$$

(10) is then equivalent to

$$\sigma_{t+1}^2 = \frac{\omega}{1-\beta} + \text{ARCH}(\infty).$$

Therefore, because a seemingly innocuous GARCH(1,1) is in fact equivalent to a ARCH(∞), its empirical power should be a little less than surprising.

There is another, useful way to re-write the GARCH(1,1) model (something similar applies to the general (q,p) case but the algebra is tedious) that becomes useful when it comes to investigate variance predictions under GARCH. Because

$$\bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta} \Longrightarrow \omega = (1-\alpha-\beta)\bar{\sigma}^2,$$

substituting this expression into (8), we have:

$$\begin{aligned} \sigma_{t+1}^2 &= \omega + \alpha R_t^2 + \beta\sigma_t^2 = (1-\alpha-\beta)\bar{\sigma}^2 + \alpha R_t^2 + \beta\sigma_t^2 \\ &= \bar{\sigma}^2 + \alpha(R_t^2 - \bar{\sigma}^2) + \beta(\sigma_t^2 - \bar{\sigma}^2), \end{aligned} \tag{11}$$

which means that under a GARCH(1,1), the forecast of tomorrow's variance is the long-run average variance, adjusted by:

- adding (subtracting) a term that measures whether today's squared return is above (below) its long-run average, and
- adding (subtracting) a term that measures whether today's variance is above (below) its long-run average.

4.3. *A formal (G)ARCH test*

A more formal (Lagrange multiplier) test for (G)ARCH in returns/disturbances vs. the sample autocorrelogram ones, has been proposed by Engle (1982). The methodology involves the following two steps: First, use simple OLS to estimate the most appropriate regression equation or ARMA model on asset returns and let $\{\hat{z}_t^2\}$ denote the squares of the standardized returns (residuals), for instance coming from a homoskedastic model, $\hat{z}_t^2 = R_t^2/\hat{\sigma}$; Second, regress these squared residuals on a constant and on q lagged values $\hat{z}_{t-1}^2$, $\hat{z}_{t+2}^2$, ..., $\hat{z}_{t-q}^2$ (e_t is a white noise shock):

$$\hat{z}_t^2 = \xi_0 + \xi_1\hat{z}_{t-1}^2 + \xi_2\hat{z}_{t-2}^2 + ... + \xi_q\hat{z}_{t-q}^2 + e_t. \tag{12}$$

If there are no ARCH effects, the estimated values of ξ_1 through ξ_q should be zero, $\xi_1 = \xi_2 = ... = \xi_q$. Hence, this regression will have little explanatory power so that the coefficient of determination (i.e., the usual R^2) will be quite low. Using a sample of T standardized returns, under the null hypothesis of no ARCH errors, the test statistic TR^2 converges to a χ_q^2. If TR^2 is sufficiently large, rejection of the null hypothesis that ξ_1 through ξ_q are jointly equal to zero is equivalent to rejection of the null hypothesis of no ARCH errors. On the other hand, if TR^2 is sufficiently low, it is possible to conclude that there are no ARCH effects.[22]

A straightforward extension of (12) can also be used to test alternative specifications of (G)ARCH models. For instance, to test for ARCH(q_1) against ARCH(q_2), with $q_2 > q_1$, you simply estimate (12) by regressing the standardized squared residuals from the ARCH(q_1) model on q_2 lags of the same squared residuals and then use an F-test for the null hypothesis that $\xi_{q_1} = \xi_{q_1+1} = ... = \xi_{q_2}$ in:

$$\hat{z}_t^2 = \xi_0 + \xi_{q_1}\hat{z}_{t-q_1-1}^2 + \xi_{q_1+1}\hat{z}_{t-q_1-2}^2 + ... + \xi_{q_2}\hat{z}_{t-q_2}^2 + e_t.$$

Note that these tests will be valid in small samples only if all the competing ARCH models have been estimated on the same data sets, in the sense that the total number of observations should be identical even though $q_2 > q_1$.

It is also possible to specifically test for GARCH effects by performing a Lagrange multiplier regression-based test. For instance, if one has initially estimated a ARCH(q) model and wants to test for p generalized ARCH terms, then the needed auxiliary regression is:

$$\hat{z}_t^2 = \varsigma_0 + \varsigma_1\hat{\sigma}_{t-1}^{2,ARCH(q)} + \varsigma_2\hat{\sigma}_{t-2}^{2,ARCH(q)} + ... + \varsigma_p\hat{\sigma}_{t-p}^{2,ARCH(q)} + e_t,$$

where $\hat{\sigma}_t^{2,ARCH(q)}$ is the time series of filtered, in-sample ARCH(q) conditional variances obtained in the first-stage estimation. Also in this case, if there are no GARCH effects, the estimated values of ς_1 through ς_p should be zero, $\varsigma_1 = \varsigma_2 = ... = \varsigma_p$. Hence, this regression will have little explanatory power so that the coefficient of determination (i.e., the usual R^2) will be quite low. Using a sample of T standardized returns, under the null hypothesis of no ARCH errors, the test statistic TR^2 converges to a χ_q^2. As before, in small samples, an F test may have superior power.

4.4. *Forecasting with GARCH models*

We have emphasized on several occasions that the point of GARCH models is more proposing forecasts of subsequent future variance than telling or supporting some economic story for why variance may be time-varying. It is therefore natural to ask how does one forecast conditional variance with a GARCH model.[23] At one level, the answer is very simple because the one-step (one-day) ahead forecast of variance, $\sigma_{t+1|t}^2$, is given directly by the model in (8):

$$\sigma_{t+1|t}^2 = \omega + \alpha R_t^2 + \beta\sigma_t^2,$$

[22] With the small samples typically used in applied work, an F-test for the null hypothesis $\xi_1 = \xi_2 = ... = \xi_q$ has been shown to be superior to a χ_q^2 test. In this case, we compare the sample value of F to the values in an F-table with q degrees of freedom in the numerator and $T - q$ degrees of freedom in the denominator.

[23] For concreteness, in what follows we focus on the case of a simple GARCH(1,1) model. All these results, at the cost of tedious algebra, may be generalized to the GARCH(q, p) case. This may represent a useful (possibly, boring) exercise.

where the notation $\sigma^2_{t+1|t} \equiv E_t[\sigma^2_{t+1}]$ now stresses that such a prediction for time $t+1$ is obtained on the basis of information up to time t, i.e., that $\sigma^2_{t+1|t}$ is a short-hand for $Var[R_t|\mathcal{F}_t] = E[R_t^2|\mathcal{F}_t]$, where the equality derives from the fact that we have assumed $\mu_{t+1} = 0$.

However we are rarely interested in just forecasting one-step ahead. Consider a generic forecast horizon, $H \geq 1$. In this case, it is easy to show that from (11),

$$\begin{aligned} \sigma^2_{t+H|t} - \bar{\sigma}^2 &= E_t[\sigma^2_{t+H}] - \bar{\sigma}^2 = \alpha E_t[R^2_{t+H-1} - \bar{\sigma}^2] + \beta E_t[\sigma^2_{t+H-1} - \bar{\sigma}^2] \\ &= \alpha(E_t[R^2_{t+H-1}] - \bar{\sigma}^2) + \beta(E_t[\sigma^2_{t+H-1}] - \bar{\sigma}^2) \\ &= \alpha(\sigma^2_{t+H-1|t} - \bar{\sigma}^2) + \beta(\sigma^2_{t+H-1|t} - \bar{\sigma}^2) = (\alpha+\beta)(\sigma^2_{t+H-1|t} - \bar{\sigma}^2). \end{aligned}$$

This establishes a recursive relationship: the predicted deviations of $t+H$ forecasts from the unconditional, long-run variance on the left-hand side equal $(\alpha+\beta) < 1$ times the predicted deviations of $t+H-1$ forecasts from the unconditional, long-run variance. All the forecasts are computed conditioning on time t information. However, we know from the recursion that $\sigma^2_{t+H-1|t} - \bar{\sigma}^2 = (\alpha+\beta)(\sigma^2_{t+H-2|t} - \bar{\sigma}^2)$, and

$$\sigma^2_{t+H|t} - \bar{\sigma}^2 = (\alpha+\beta)\left[\underbrace{(\alpha+\beta)(\sigma^2_{t+H-2|t} - \bar{\sigma}^2)}_{\sigma^2_{t+H-1|t}-\bar{\sigma}^2}\right] = (\alpha+\beta)^2(\sigma^2_{t+H-2|t} - \bar{\sigma}^2).$$

Working backwards this way $H-1$ times, it is easy to see that

$$\sigma^2_{t+H|t} - \bar{\sigma}^2 = (\alpha+\beta)^{H-1}(\sigma^2_{t+1|t} - \bar{\sigma}^2) \tag{13}$$

or

$$\sigma^2_{t+H|t} = \bar{\sigma}^2 + (\alpha+\beta)^{H-1}(\sigma^2_{t+1} - \bar{\sigma}^2) = \bar{\sigma}^2 + (\alpha+\beta)^{H-1}[\alpha(R_t^2 - \bar{\sigma}^2) + \beta(\sigma_t^2 - \bar{\sigma}^2)].$$

This expression implies that as the forecast horizon H grows, because for $(\alpha+\beta) < 1$ the limit of $(\alpha+\beta)^{H-1}$ is 0, we obtain

$$\lim_{H\to\infty} \sigma^2_{t+H|t} = \bar{\sigma}^2,$$

i.e., the very long horizon forecast from a stationary GARCH(1,1) model is the long-run variance itself. Practically, this means that because stationary GARCH models are mean-reverting, any long-run forecast will simply exploit this fact, i.e., use $\bar{\sigma}^2$ as the prediction. Of course, for finite but large H it is easy to see that when $(\alpha+\beta)$ is relatively small, then $\sigma^2_{t+H|t}$ will be close to $\bar{\sigma}^2$ for relatively modest values of H; when $(\alpha+\beta)$ is instead close to 1, $\sigma^2_{t+H|t}$ will depart from $\bar{\sigma}^2$ even for large values of H. (13) has another key implication: because in a GARCH we also restrict both α and β to be positive, $(\alpha+\beta) \in (0,1)$ implies that $(\alpha+\beta)^{H-1} > 0$ for all values of the horizon $H \geq 1$. Therefore it is clear that $\sigma^2_{t+H|t} > \bar{\sigma}^2$ when $\sigma^2_{t+1|t} > \bar{\sigma}^2$, and vice-versa. This means that H-step ahead forecasts of the variance will exceed long-run variance if 1-step ahead forecasts exceed long-run variance, and vice-versa. As you have understood at this point, the coefficient sum $(\alpha+\beta)$

plays a crucial role in all matters concerning forecasting with GARCH models and is commonly called the *persistence level/index* of the model: a high persistence, $(\alpha+\beta)$ close to 1, implies that shocks which push variance away from its long-run average will persist for a long time, even though eventually the long-horizon forecast will be the long-run average variance, $\bar{\sigma}^2$.

In asset allocation problems, we sometimes care for the variance of long-horizon returns,

$$R_{t+1:t+H} \equiv \sum_{h=1}^{H} R_{t+h}.$$

Chapter 1 has already extensively discussed the properties of long-horizon returns, emphasizing how simple sums make sense in the case of continuously compounded returns.[24] Here we specifically investigate conditional forecasts (expectations) of the variance of long-horizon returns. Because the model $R_{t+1} = \sigma_{t+1}z_{t+1}$, z_{t+1} ~IID $\mathcal{N}(0,1)$, implies that financial returns have zero autocorrelations, the variance of the cumulative H-day returns is:

$$\begin{aligned}\sigma^2_{t+1:t+H} &\equiv Var_t\left[\sum_{h=1}^{H} R_{t+h}\right] = E_t\left[\left(\sum_{h=1}^{H} R_{t+h}\right)^2\right] = E_t\left[\sum_{h=1}^{H} R^2_{t+h}\right] \\ &= \sum_{h=1}^{H} E_t[R^2_{t+h}] = \sum_{h=1}^{H} \sigma^2_{t+h|t}.\end{aligned}$$

Note that $Var_t\left[\sum_{h=1}^{H} R_{t+h}\right] = E_t\left[\sum_{h=1}^{H} R^2_{t+h}\right]$ because $E_t\left[\sum_{h=1}^{H} R_{t+h}\right] = \sum_{h=1}^{H} E_t[R_{t+h}] = 0$. Moreover, $E_t[(\sum_{h=1}^{H} R_{t+h})^2] = E_t\left[\sum_{h=1}^{H} R^2_{t+h}\right]$ because the absence of autocorrelation in returns leads to all the conditional expectations of the cross-products, $E_t\left[R_{t+\tau}R^2_{t+\tau+k}\right]$ $(k \neq 0)$ to vanish by construction. Solving in the GARCH(1,1) case, we have:

$$\begin{aligned}\sigma^2_{t+1:t+H} &= \sum_{h=1}^{H} \bar{\sigma}^2 + \sum_{h=1}^{H} (\alpha+\beta)^{h-1}(\sigma^2_{t+1|t} - \bar{\sigma}^2) \\ &= H\bar{\sigma}^2 + \sum_{h=1}^{H} (\alpha+\beta)^{h-1}(\sigma^2_{t+1|t} - \bar{\sigma}^2) \neq H\bar{\sigma}^2.\end{aligned}$$

In particular, $\sigma^2_{t+1:t+H} \gtrless H\bar{\sigma}^2$ when $\sum_{h=1}^{H}(\alpha+\beta)^{h-1}(\sigma^2_{t+1|t} - \bar{\sigma}^2)$, which requires that $\sigma^2_{t+1|t} \gtrless \bar{\sigma}^2$. More importantly, note that the variance of the (log-) long horizon returns is not simply H times their unconditional, long-run variance: the term $H\bar{\sigma}^2$ needs to be adjusted to take into account transitory effects, concerning each of the R_{t+h} contributing to $R_{t+1:t+H}$.

4.5. *Are GARCH(1,1) and RiskMetrics different?*

This is a key question: in section 3.2 we have mentioned that the RiskMetrics model has been rather successful in practice. Do we need to bother with learning and (this is harder) estimating a GARCH(q,p) model? This leads to ask whether RiskMetrics and GARCH are really that different:

[24] The notation $R_{t+1:t+H}$ may be new, but it is also rather self-evident.

as we shall see, they are indeed quite different statistical objects because they imply divergent unconditional, long-run properties, even though in a small sample of data you cannot rule out the possibility that their performance may be similar. Yet, especially in long-horizon forecasting applications, the structural differences between the two ought to be kept in mind.

On the one hand, RiskMetrics and GARCH are not that radically different: comparing (8) with (5) you can see that RiskMetrics is just a special case of GARCH(1,1) in which $\omega = 0$ and $\alpha = 1 - \beta$ so that, equivalently, $(\alpha + \beta) = 1$. On the other hand, this simple fact has a number of important implications:

1. Because $\omega = 0$ and $\alpha + \beta = 1$, under RiskMetrics the long-run variance does not exist as gives an indeterminate ratio "0/0":

$$\bar{\sigma}^2_{RMetrics} = \frac{0}{1 - \alpha - \beta} = \frac{0}{0}.$$

   Therefore while RiskMetrics ignores the fact that the long-run, average variance tends to be relatively stable over time, a GARCH model with $(\alpha + \beta) < 1$ does not. Equivalently, while a GARCH with $(\alpha + \beta) < 1$ is a stationary process, a RiskMetrics model is not. This can be seen from the fact that $\bar{\sigma}^2_{RMetrics}$ does not even exist (do not spend much time trying to figure out the value of 0/0).

2. Because under RiskMetrics $(\alpha + \beta) = 1$, it follows that

$$(\sigma^2_{t+H|t})_{RMetrics} - \bar{\sigma}^2 = (1)^{H-1}(\sigma^2_{t+1|t} - \bar{\sigma}^2) = \sigma^2_{t+1|t} - \bar{\sigma}^2 \Longrightarrow (\sigma^2_{t+H|t})_{RMetrics} = \sigma^2_{t+1|t},$$

   which means that any shock to current variance is destined to persist forever: If today is a high (low)-variance day, then the RiskMetrics model predicts that all future days will be high (low)- variance days, which is clearly rather unrealistic. In fact, this can be dangerous: assuming the RiskMetrics model holds despite the data truly look more like GARCH will give risk managers a false sense of the calmness of the market in the future, when the market is calm today and $\sigma^2_{t+1|t} < \bar{\sigma}^2$.[25] A GARCH more realistically assumes that eventually, in the future, variance will revert to the average value $\bar{\sigma}^2$.

3. Under RiskMetrics, the variance of long-horizon returns is:

$$\begin{aligned}(\sigma^2_{t+1:t+H})_{RMetrics} &= \sum_{h=1}^{H} \sigma^2_{t+h|t} = \sum_{h=1}^{H} \sigma^2_{t+1|t} = H\sigma^2_{t+1} \\ &= H(1-\lambda)R^2_t + H\lambda\sigma^2_t,\end{aligned}$$

   which is just H times the most recent forecast of future variance. Figure 7 illustrates this

[25] Clearly this point cannot be appreciated by such a risk-manager: under RiskMetrics $\bar{\sigma}^2$ does not exist.

difference through a practical example in which for the RiskMetrics we set $\lambda = 0.94$.

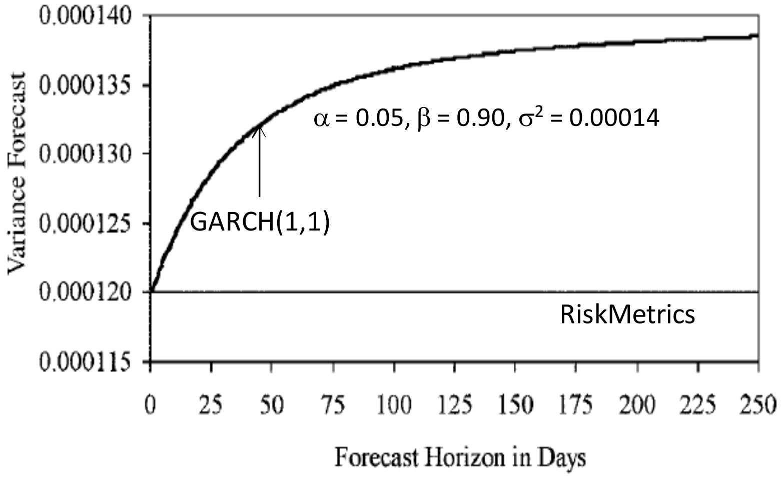


Figure 7: Variance forecasts as a function of horizon (H) under a GARCH(1,1) vs. RiskMetrics

5. Asymmetric GARCH Models (with Leverage) and Predetermined Variance Factors

A number of empirical papers have emphasized that for many assets and sample periods, a negative return increases conditional variance by more than a positive return of the same magnitude does, the so-called *leverage effect*. Although empirical evidence exists that has shown that speaking of a leverage effect with reference to corporate leverage may be slightly abusive of what the data show, the underlying idea is that because, in the case of stocks, a negative equity return implies a drop in the equity value of the company, this implies that the company becomes more highly levered and thus riskier (assuming the level of debt stays constant). Assuming that on average conditional variance represents an appropriate measure of risk—which, as we shall discuss, requires rather precise assumptions within a formal asset pricing framework—the logical flow of ideas implies that negative (shocks to) stock returns ought to be followed by an increase in conditional variance, or at least that negative returns ought to affect subsequent conditional variance more than positive returns do.[26] More generally, even though a leverage-related story remains suggestive and a few researchers in asset pricing have indeed tested this linkage directly, in what follows we shall write about an *asymmetric effect* in conditional volatility dynamics, regardless of whether this may actually be a leverage effect or not.

To quant experts, what matters is that returns on most assets seem to be characterized by an

[26] These claims are subject to a number of qualifications. First, this story for the existence of asymmetric effects in conditional volatility only works in the case of stock returns, as it is difficult to imagine how leverage may enter the picture in the case of bond, real estate, and commodities' returns, not to mention currency log-changes. Second, the story becomes fuzzy when one has to specify the time lag that would separate the negative shock to equity returns and hence the capital structure and the (subsequent?) reaction of conditional volatility. Third, as acknowledged in the main text, there are potential issue with identifying the (idiosyncratic) capital structure-induced risk of a company with forecasts of conditional variance.

asymmetric *news impact curve* (NIC). The NIC measures how new information is incorporated into volatility, i.e., it shows the relationship between the current return R_t and conditional variance one period ahead σ^2_{t+1}, holding constant all other past and current information.[27] Formally, $\sigma^2_{t+1} = NIC(R_t|\sigma^2_t = \sigma^2)$ means that one investigates the behavior of σ^2_{t+1} as a function of the current return, taking past variance as given. For instance, in the case of a GARCH(1,1) model we have:

$$NIC(R_t|\sigma^2_t = \sigma^2) = \omega + \alpha R^2_t + \beta\sigma^2 = A + \alpha R^2_t$$

where the constant $A \equiv \omega + \beta\sigma^2$ and $\alpha > 0$ is the convexity parameter. This function is a quadratic function of R^2_t and therefore symmetric around 0 (with intercept A). Figure 8 shows such a symmetric NIC from a GARCH(1,1) model.

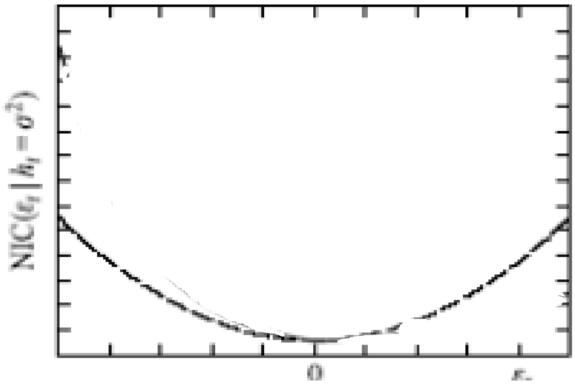


Figure 8: Symmetric NIC from a GARCH model

However, from empirical work, we know that for most return series, the empirical NIC fails to be symmetric. As already hinted at, there is now massive evidence that negative news increase conditional volatility much more than positive news do.[28] Figure 9 compares a symmetric GARCH-induced NIC with an asymmetric one.

How do you actually test whether there are asymmetric effects in conditional heteroskedasticity? The simplest and most common way consists of using (Lagrange multiplier) ARCH-type tests similar to those introduced before. After having fitted to returns data either a ARCH or GARCH model, call $\{\hat{z}_t\}$ the corresponding time series of standardized residuals. Then simple regressions may be

[27] In principle the NIC should be defined and estimated with reference to shocks to returns, i.e., *news*. In general terms, news are defined as the unexpected component of returns. However, in this chapter we are working under the assumption that $\mu_{t+1} = 0$ so that in our view, returns and news are the same. However, some of the language in the text will still refer to news as this is the correct thing to do.

[28] Intuitively, both negative and positive news should increase conditional volatility because they trigger trades by market operators. This is another flaw of our earlier presentation of asymmetries in the NIC as leverage effects: in this story, positive news ought to reduce company leverage, reduce risk, and volatility. In practice, all kinds of news tend to generate trading and hence volatility, even though negative news often bump variance up more than positive news do.

performed to assess whether the NIC is actually asymmetric.

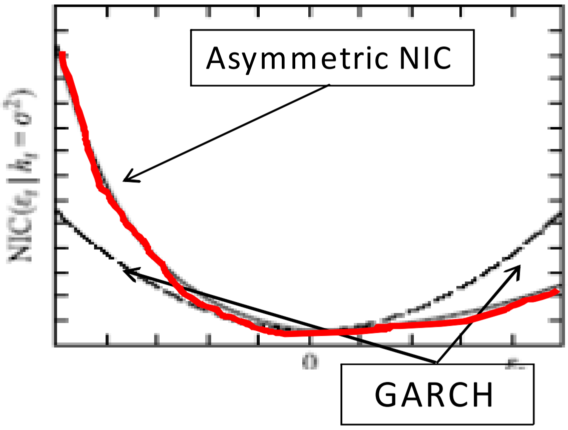


Figure 9: Symmetric and asymmetric NICs

If tests of the null hypothesis that the coefficients γ_1, γ_2, ..., γ_d, φ_1, φ_2, ..., φ_d are all equal to zero (jointly or individually) in the regressions ($1_{\hat{z}_t<0}$ is the notation for a dummy variable that takes a value of 1 when the condition $z_t < 0$ is satisfied, and zero otherwise)

$$\hat{z}_t^2 = \gamma_0 + \gamma_1 \hat{z}_{t-1} + \gamma_2 \hat{z}_{t-2} + ... + \gamma_d \hat{z}_{t-d} + e_t$$

or

$$\hat{z}_t^2 = \gamma_0 + \gamma_1 1_{\hat{z}_{t-1}<0} + ... + \gamma_d 1_{\hat{z}_{t-2}<0} + \varphi_1 1_{\hat{z}_{t-1}<0} \hat{z}_{t-1} + ... + \varphi_d 1_{\hat{z}_{t-d}<0} \hat{z}_{t-d} + e_t$$

lead to rejections, then this is evidence of the need of modelling asymmetric conditional variance effects. This occurs because either the signed level of past estimated shocks ($\hat{z}_{t-1}$, $\hat{z}_{t-2}$, ..., $\hat{z}_{t-d}$), dummies that capture such signs, or the interaction between their signed level and dummies that capture theirs signs, provide significant explanation for subsequent squared standardized returns.

Let's keep in mind that this is not just semantics or a not better specified need to fit the data by some geeky econometrician: market operators will care of the presence of any asymmetric effects because this may massively impact their forecasts of volatility, depending on whether recent market news have been positive or negative. Here the good news (to us) are that we can cheaply modify the GARCH models introduced in section 4 so that the weight given to current returns when forecasting conditional variance depends on whether past returns were positive or negative. In fact, this can be done in some many effective ways to have sparked a proliferation of alternative asymmetric GARCH models currently entertained by a voluminous econometrics literature. In the rest of this section we briefly present some of these models, even though a Reader must be warned that several dozens of them have been proposed and estimated on all kinds of financial data, often affecting applications, such as option pricing.

The general idea is that—given that the NIC is asymmetric or displays other features of interest—we may directly incorporate the empirical NIC as part of an extended GARCH model specification according to the following logic:

Standard GARCH model + asymmetric NIC component.

where the NIC under GARCH (i.e., the standard component) is $NIC(z_t|\sigma_t^2 = \sigma^2) = A + \alpha R_t^2 = A + \alpha\sigma^2 z_t^2$. In fact, there is an entire family of volatility models parameterized by θ_1, θ_2, and θ_3 that can be written as follows:

$$NIC(z_t) = [|z_t - \theta_1| - \theta_2(z_t - \theta_1)]^{2\theta_3}. \tag{14}$$

One retrieves a standard, plain vanilla GARCH(1,1) by setting $\theta_1 = 0$, $\theta_2 = 0$, and $\theta_3 = 1$. In principle the game becomes then to empirically estimate θ_1, θ_2, and θ_3 from the data.

5.1. *Exponential GARCH*

EGARCH is probably the most prominent case of an asymmetric GARCH model. Moreover, the use of EGARCH—where the "E" stands for exponential—is predicated upon the fact that while in standard ARCH and GARCH estimation the need to impose non-negativity constraints on the parameters often creates numerical as well as statistical (inferential, when the estimated parameters fall on a boundary of the constraints) difficulties in estimation, EGARCH solves these problems by construction in a very clever way: even though $f(\boldsymbol{\theta}) : \mathcal{R}^K \to \mathcal{R}$ can take any real value (here $\boldsymbol{\theta}$ is a vector of parameters to be estimated and $f(\cdot)$ some function, for instance predicted variance), it is obviously the case that

$$\exp(f(\boldsymbol{\theta})) > 0 \ \forall \boldsymbol{\theta} \in \mathcal{R}^K,$$

i.e., "exponentiating" any real number gives a positive real. Equivalently, one ought to model not $f(\boldsymbol{\theta})$ but directly $\log f(\boldsymbol{\theta})$, knowing that $f(\boldsymbol{\theta}) = \exp(\log f(\boldsymbol{\theta}))$: the model is written in *log-linear form.*

Nelson (1990) has proposed such a EGARCH in which positivity of the conditional variance is ensured by the fact that $\log \sigma_{t+1}^2$ is modeled directly:[29]

$$\log \sigma_{t+1}^2 = \omega + \beta \log \sigma_t^2 + g(z_t) \qquad g(z_t) = \theta z_t + \alpha(|z_t| - E|z_t|),$$

and recall that $z_t \equiv R_t/\sigma_t$. The sequence of random variables $\{g(z_t)\}$ is a zero-mean, IID stochastic process with the following features: (i) if $z_t \geq 0$, as $g(z_t) = \theta z_t + \alpha(z_t - E|z_t|) = -\alpha E|z_t| + (\theta + \alpha)z_t$, $g(z_t)$ is linear in z_t with slope $\theta + \alpha$; (ii) if $z_t < 0$, as $g(z_t) = \theta z_t + \alpha(-z_t - E[-z_t]) = -\alpha E|z_t| + (\theta - \alpha)z_t$, $g(z_t)$ is linear in z_t with slope $\theta - \alpha$. Thus, $g(z_t)$ is a function of both the magnitude and the sign of z_t and it allows the conditional variance process to respond asymmetrically to rises and falls in stock prices. Indeed, $g(z_t)$ can be re-written as:

$$g(z_t) = -\alpha E|z_t| + (\theta + \alpha)z_t 1_{z_t \geq 0} + (\theta - \alpha)z_t 1_{z_t < 0},$$

[29] This EGARCH(1,1) model may be naturally extended to a general EGARCH(q,p) case:

$$\log \sigma_{t+1}^2 = \omega + \sum_{j=1}^{p} \beta_j \log \sigma_{t+1-j}^2 + g(z_t, z_{t-1}, ..., z_{t-q}) \quad g(z_t, z_{t-1}, ..., z_t - q) = \sum_{i=1}^{q} [\theta_i z_{t+1-i} + \alpha_i(|z_{t+1-i}| - E|z_{t+1-i}|)].$$

However on a very few occasions these extended EGARCH(q,p) models have been estimated in the literature, although their usefulness in applied forecasting cannot be ruled out on an ex-ante basis.

where $1_{z_t \geq 0}$ is a standard dummy variable. The term $\alpha(|z_t| - E|z_t|)$ represents a magnitude effect:

- If $\alpha > 0$ and $\theta = 0$, innovations in the conditional variance are positive (negative) when the magnitude of z_t is larger (smaller) than its expected value;
- If $\alpha = 0$ and $\theta < 0$, innovations in the conditional variance are positive (negative) when returns innovations are negative (positive), in accordance with empirical evidence for stock returns.[30]

5.2. *Threshold (GJR) GARCH model*

Another way of capturing the leverage effect is to directly build a model that exploits the possibility to define an indicator variable, I_t , to take on the value 1 if on day t the return is negative and zero otherwise. For concreteness, in the simple (1,1) case, variance dynamics can now be specified as:

$$\begin{aligned} \sigma^2_{t+1} &= \omega + \alpha R^2_t + \alpha\theta I_t R^2_t + \beta\sigma^2_t \qquad I_t \equiv \begin{cases} 1 & \text{if } R_t < 0 \\ 0 & \text{if } R_t \geq 0 \end{cases} \quad \text{or} \\ \sigma^2_{t+1} &= \begin{cases} \omega + \alpha(1+\theta)R^2_t + \beta\sigma^2_t & \text{if } R_t < 0 \\ \omega + \alpha R^2_t + \beta\sigma^2_t & \text{if } R_t \geq 0 \end{cases} . \end{aligned} \tag{15}$$

A $\theta > 0$ will again capture the leverage effect. In fact, note that in (15) while the coefficient on the current positive return is simply α, i.e., identical to a plain-vanilla GARCH(1,1) model when $R_t \geq 0$, this becomes $\alpha(1+\theta) > \alpha$ when $R_t < 0$, just a simple and yet powerful way to capture asymmetries in the NIC. This model is sometimes referred to as the GJR-GARCH model—from Glosten, Jagannathan, and Runkle's (1993) paper—or threshold GARCH (TGARCH) model. Also in this case, extending the model to encompass the general (q, p) case is straightforward:

$$\sigma^2_{t+1} = \omega + \sum_{i=1}^{q} \alpha_i(1 + \theta_i I_t)R^2_{t+1-i} + \sum_{j=1}^{p} \beta_j \sigma^2_{t+1-j}.$$

In this model, because when 50% of the shocks are assumed to be negative and the other 50% positive, so that $E[I_t] = 1/2$, the long-run variance equals:[31]

$$\begin{aligned} \bar{\sigma}^2 &\equiv E[\sigma^2_{t+1}] = \omega + \alpha E[R^2_t] + \alpha\theta E[I_t R^2_t] + \beta E[\sigma^2_t] = \omega + \alpha\bar{\sigma}^2 + \alpha\theta E[I_t]\bar{\sigma}^2 + \beta\bar{\sigma}^2 \\ &= \omega + \alpha\bar{\sigma}^2 + \frac{1}{2}\alpha\theta\bar{\sigma}^2 + \beta\bar{\sigma}^2 \Longrightarrow \bar{\sigma}^2 = \frac{\omega}{1 - \alpha(1 + 0.5\theta) - \beta}. \end{aligned}$$

Visibly, in this case the persistence index is $\alpha(1 + 0.5\theta) + \beta$. Formally, the NIC of a threshold GARCH model is:

$$NIC(R_t|\sigma^2_t = \sigma^2) = \omega + \alpha R^2_t + \alpha\theta I_t R^2_t + \beta\sigma^2 = A + \alpha(1 + \theta I_t)R^2_t$$

[30] $g(z_t) = \theta z_t < 0$ when $z_t < 0$ represents no problem thanks to the exponential transformation.

[31] Obviously, this is the case in the model $R_{t+1} = \sigma_{t+1}z_{t+1}$, z_{t+1} ∼IID $\mathcal{N}(0, 1)$ as the density of the shocks is normal and therefore symmetric around zero (the mean) by construction. However, this will also apply to any symmetric distribution z_{t+1} ∼IID $\mathcal{D}(0, 1)$ (e.g., think of a standard t-student). Also recall that $E[\sigma^2_{t+1}] = E[\sigma^2_t] = \bar{\sigma}^2$ by the definition of stationarity.

where the constant $A \equiv \omega + \beta\sigma^2$ and $\alpha > 0$ is a convexity parameter that is increased to $\alpha(1+\theta)$ for negative returns. This means that the NIC will be a parabola with a steeper left branch, to the left of $R_t = 0$.

5.3. *NAGARCH model*

One simple choice of parameters in the generalized NIC in (14) yields an increasingly common asymmetric GARCH model: when $\theta_2 = 0$ and $\theta_3 = 1$, the NIC becomes $NIC(z_t) = (|z_t - \theta_1|)^2 = (z_t - \theta_1)^2$ and an asymmetry derives from the fact that when $\theta_1 > 0$,[32]

$$(z_t - \theta_1)^2 = \begin{cases} (z_t - \theta_1)^2 < z_t^2 & \text{if } z_t \geq 0 \\ (z_t - \theta_1)^2 > z_t^2 & \text{if } z_t < 0 \end{cases}.$$

Written in extensive form that also includes the standard GARCH(1,1) component in (14), such a model is called a Nonlinear (Asymmetric) GARCH, or N(A)GARCH:

$$\begin{aligned} \sigma_{t+1}^2 &= \omega + \alpha(R_t - \delta\sigma_t)^2 + \beta\sigma_t^2 = \omega + \alpha\sigma_t^2(z_t - \delta)^2 + \beta\sigma_t^2 \\ &= \omega + \alpha\sigma_t^2 z_t^2 + \alpha\delta^2\sigma_t^2 - 2\alpha\delta\sigma_t^2 z_t + \beta\sigma_t^2 \\ &= \omega + \alpha R_t^2 + (\beta + \alpha\delta^2 - 2\alpha\delta z_t)\sigma_t^2 = \omega + \alpha R_t^2 + \beta'\sigma_t^2 - 2\alpha\delta z_t\sigma_t^2, \end{aligned}$$

where $\beta' \equiv \beta + \alpha\delta^2 > \beta'$ if $\alpha > 0$. As you can see, NAGARCH(1,1) is:

- *Asymmetric*, because if $\delta \neq 0$, then the NIC (for given $\sigma_t^2 = \sigma^2$) is: $A + \alpha\sigma^2 z_t^2 - 2\alpha\delta\sigma^2 z_t$ which is no longer a simple, symmetric quadratic function of standardized residuals, as under a plain-vanilla GARCH(1,1); equivalently, and assuming $\delta > 0$, while $R_t \geq 0$ impacts conditional variance only in the measure $(R_t - \delta\sigma_t)^2 < R_t^2$, $R_t < 0$ impacts conditional variance in the measure $(R_t - \delta\sigma_t)^2 > R_t^2$.[33]

- *Non-linear*, because NAGARCH may be written in the following way:

  $$\sigma_{t+1}^2 = \omega + \alpha R_t^2 + [\beta' - 2\alpha\delta z_t]\sigma_t^2 = \omega + \alpha R_t^2 + \beta(z_t)\sigma_t^2$$

  where $\beta(z_t) \equiv \beta' - 2\alpha\delta z_t$ is a function that makes the beta coefficient of a GARCH depend on a lagged standardized residual.[34] Here the claim of non-linearity follows from the fact that

[32] $(|z_t - \theta_1|)^2 = (z_t - \theta_1)^2$ because squaring an absolute value makes the absolute value operator irrelavant, i.e., $|f(x)|^2 = (f(x))^2$.

[33] When $\delta < 0$ the asymmetry remains, but in words it is stated as: while $R_t < 0$ impacts conditional variance only in the measure $(R_t - \delta\sigma_t)^2 < R_t^2$, $R_t \geq 0$ impacts conditional variance in the measure $(R_t - \delta\sigma_t)^2 > R_t^2$. This means that $\delta > 0$ captures a "left" asymmetry consistent with a leverage effect and in which negative returns increase variance more than positive returns do; $\delta < 0$ captures instead a "right" asymmetry that is sometimes observed for some commodities, like precious metals.

[34] Some textbooks emphasize non-linearity in a different way: a NAGARCH implies

$$\sigma_{t+1}^2 = \omega + \alpha\sigma_t^2(z_t - \delta)^2 + \beta\sigma_t^2 = \omega + \alpha\left(\sigma_t^2\right)[z_t - \delta]^2 + \beta\sigma_t^2,$$

where it is the alpha coefficient that now becomes a function of the last filtered conditional variance, $\alpha\left(\sigma_t^2\right) \equiv \alpha\sigma_t^2 > 0$

all models that are written under a linear *functional form* (i.e., $f(x) = a + bx$) but in which some or all coefficients depend on their turn on the conditioning variables or information (i.e., $f(x) = a_x + b_x x$, in the sense that $a_x = a(x)$ and/or $b_x = b(x)$) is also a non-linear model.[35]

NAGARCH plays key role in option pricing with stochastic volatility because, as we shall see later on, NAGARCH allows you to derive closed-form expressions for European option prices in spite of the rich volatility dynamics. Because a NAGARCH may be written as

$$\sigma_{t+1}^2 = \omega + \alpha\sigma_t^2(z_t - \delta)^2 + \beta\sigma_t^2$$

and, if $z_t \sim$IID $\mathcal{N}(0,1)$, z_t is independent of σ_t^2 as σ_t^2 is only a function of an infinite number of past squared returns, it is possible to easily derive the long-run, unconditional variance under NAGARCH and the assumption of stationarity:[36]

$$\begin{aligned} E[\sigma_{t+1}^2] &= \bar{\sigma}^2 = \omega + \alpha E[\sigma_t^2(z_t-\delta)^2] + \beta E[\sigma_t^2] \\ &= \omega + \alpha E[\sigma_t^2]E[z_t^2 + \delta^2 - 2\delta z_t] + \beta E[\sigma_t^2] = \omega + \alpha\bar{\sigma}^2(1+\delta^2) + \beta\bar{\sigma}^2, \end{aligned}$$

where $\bar{\sigma}^2 = E[\sigma_t^2]$ and $E[\sigma_t^2] = E[\sigma_{t+1}^2]$ because of stationarity. Therefore

$$\bar{\sigma}^2[1 - \alpha(1+\delta^2) - \beta] = \omega \Longrightarrow \bar{\sigma}^2 = \frac{\omega}{1 - \alpha(1+\delta^2) - \beta}$$

which is exists and positive if and only if $\alpha(1+\delta^2) + \beta < 1$. This has two implications: (i) the persistence index of a NAGARCH(1,1) is $\alpha(1+\delta^2)+\beta$ and not simply $\alpha+\beta$; (ii) a NAGARCH(1,1) model is stationary if and only if $\alpha(1+\delta^2) + \beta < 1$.

5.4. *GARCH with exogenous (predetermined) factors*

There is also a smaller literature that has connected time-varying volatility as well asymmetric NICs not only to pure time series features, but to observable economic phenomena, especially at daily frequencies. For instance, days where no trading takes place will affect the forecast of variance for the days when trading resumes, i.e., days that follow a weekend or a holiday. In particular, because information flows to markets even when trading is halted during weekends or holidays, a rather popular model is

$$\sigma_{t+1}^2 = \omega + \alpha R_t^2 + \beta\sigma_t^2 + \gamma IT_{t+1} = \omega + \alpha\sigma_t^2 z_t^2 + \beta\sigma_t^2 + \gamma IT_{t+1},$$

if $\alpha > 0$. It is rather immaterial whether you want to see a NAGARCH as a time-varying coefficient model in which α' depends on σ_t^2 or in which β' depends on z_t, although the latter view is more helpful in defining the NIC of the model.

[35] Technically, this is called a time-varying coefficient model. You can see that easily by thinking of what you expect of a derivative to be in a linear model: $df(x)/dx = b$, i.e., a constant indenpendent of x. In a time-varying coefficient model this is potentially not the case as $df(x)/dx = [da(x)/dx] + [db(x)/dx] \cdot x + b(x)$ which is not a constant, at least not in general. NAGARCH is otherwise called a time-varying coefficient GARCH model, with a special structure of time-variation.

[36] The claim that σ_t^2 is a function of an infinite number of past squared returns derives from the fact that under GARCH, we know that the process of squared returns follows (under appropriate conditions) a stationary ARMA. You know from the first part of your econometrics sequence that any ARMA has an autoregressive representation.

where IT is a dummy that takes a unit value in correspondence of a day that follows a weekend. Note that in this model, the plain-vanilla GARCH(1,1) portion (i.e., $\omega + \alpha R_t^2 + \beta \sigma_t^2$) has been re-written in a different but completely equivalent way, exploiting the fact that $R_t^2 = \sigma_t^2 z_t^2$ by definition. Moreover, this variance model implies that it is IT_{t+1} that affects σ_{t+1}^2, which is sensible because IT is deterministic (we know the calendar of open business days on financial markets well in advance) and hence clearly pre-determined. Obviously, many alternative models including predetermined variables different from IT could have been proposed. Other predetermined variables could be yesterday's trading volume or pre-scheduled news announcement dates such as company earnings and FOMC (Federal Open Market Committee at the U.S. Federal Reserve) meeting dates.[37] For example, suppose that you want to detect whether the terrorist attacks of September 11, 2001, increased the volatility of asset returns. One way to accomplish the task would be to create a dummy variable $D_t^{09/11}$ that equals 0 before September 11 and equals 1 thereafter. Consider the following modification of the GARCH(1,1) specification:

$$\sigma_{t+1}^2 = \omega + \alpha R_t^2 + \beta \sigma_t^2 + \gamma D_t^{09/11}.$$

If it is found that $\gamma > 0$, it is possible to conclude that the terrorist attacks increased the mean of conditional volatility.

More generally, consider the model

$$R_{t+1} = x_t z_{t+1},$$

where z_{t+1} is IID $\mathcal{D}(0,1)$ and x_{t+1} is a random variable observable at time t. Note that while if $x_t = x_0 > 0 \ \forall t \geq 1$, then $Var_t[R_{t+1}] = x_0^2 Var_t[z_{t+1}] = x_0^2 \cdot 1 = x_0^2$ and R_{t+1} is also $\mathcal{D}(0, x_0^2)$ so that returns are homoskedastic, when the realizations of the $\{x_t\}$ process are random, then $Var_t[R_{t+1}] = x_t^2$; because we can observe x_t at time t, one can forecast the variance of returns conditioning on the realized value of x_t. Furthermore, if $\{x_t\}$ is positively serially correlated, then the conditional variance of returns will exhibit positive serial correlation. The issue is what variable(s) may enter the model with the role envisioned above. One approach is to try and empirically discover what such a variable may be using standard regression analysis: you might want to modify the basic model by introducing the coefficients a_0 and a_1 and estimate the regression equation in logarithmic form as[38]

$$\log(1 + R_{t+1}) = a_0 + a_1 \log x_t + e_{t+1}.$$

This procedure is simple to implement since the logarithmic transformation results in a linear regression equation; OLS can be used to estimate a_0 and a_1 directly. A major difficulty with this strategy is that it assumes a specific cause for the changing variance. The empirical literature has

[37] See also the Spline-GARCH model with a deterministic volatility component in Engle and Rangel (2008).

[38] Here $e_{t+1} = \ln z_{t+1}$ which will require however $z_{t+1} > 0$. Moreover, note that the left-hand side is now the log of $(1 + R_{t+1})$ to keep the logarithm well defined. If R_{t+1} is a *net* returns (i.e., $R_{t+1} \in [-1, +\infty)$), then $(1 + R_{t+1})$ is a *gross* returns, $(1 + R_{t+1}) \in [0, +\infty)$.

had a hard time coming up with convincing choices of variables capable to affect the conditional variance of returns. For instance, was it the oil price shocks, a change in the conduct of monetary policy, and/or the breakdown of the Bretton-Woods system that was responsible for the volatile exchange rate dynamics during the 1970s?

Among the large number of predetermined variables that have been proposed in the empirical finance literature, one (family) of them has recently acquired considerable importance in exercises aimed at forecasting variance: option implied volatilities, and in particular the (square of the) CBOE's (Chicago Board Options Exchange) VIX as well as other functions and transformations of the VIX. In general, models that use explanatory variables to capture time-variation in variance are represented as:

$$\sigma_{t+1}^2 = \omega + g(\mathbf{X}_t) + \alpha\sigma_t^2 z_t^2 + \beta\sigma_t^2,$$

where $\mathbf{X}_t$ is a vector of predetermined variables that may as well include VIX. Note that because this volatility model is not written in log-exponential form, it is important to ensure that the model always generates a positive variance forecast, which will require that restrictions—either of an economic type or to be numerically imposed during estimation—must be satisfied, such as $g(\mathbf{X}_t) > 0$ for all possible values of $\mathbf{X}_t$, besides the usual ω, α, $\beta > 0$.

5.4.1. **One example with VIX predicting variances**

Consider the model

$$\begin{aligned} R_{t+1} &= \sigma_{t+1} z_{t+1} \quad \text{with} \quad z_{t+1} \sim \text{IID } \mathcal{N}(0,1) \\ \sigma_{t+1}^2 &= \omega + \alpha R_t^2 + \beta\sigma_t^2 + \gamma VIX_t \end{aligned}$$

where VIX_t follows a stationary autoregressive process, $VIX_t = \delta_0 + \delta_1 VIX_{t-1} + \zeta_t$ with $E[\zeta_t] = 0$. The expression for the unconditional variance remains easy to derive: if the process for VIX_t is stationary, we know that $|\delta_1| < 1$. Moreover, from

$$E[VIX_t] = \delta_0 + \delta_1 E[VIX_{t-1}] \Longrightarrow E[VIX_t] = E[VIX_{t-1}] = \frac{\delta_0}{1-\delta_1}$$

which is finite because $|\delta_1| < 1$. Now

$$\begin{aligned} E[\sigma_{t+1}^2] &= \omega + \alpha E[R_t^2] + \beta E[\sigma_t^2] + \gamma E[VIX_t] \\ &= \omega + (\alpha+\beta)E[\sigma_t^2] + \gamma\frac{\delta_0}{1-\delta_1} \Longrightarrow E[\sigma_t^2] = \frac{\omega + \gamma\frac{\delta_0}{1-\delta_1}}{1-\alpha-\beta}. \end{aligned}$$

One may actually make more progress by imposing economic restrictions. For instance, taking into account that, if the options markets are efficient, then $E[VIX_t] = E[\sigma_t^2]$ may obtain, one can establish a further connection between the parameters δ_0 and δ_1 and ω, α, and β:[39]

$$\begin{aligned} E[\sigma^2_{t+1}] &= \omega + \alpha E[r^2_t] + \beta E[\sigma^2_t] + \gamma E[VIX_t] \\ &= \omega + (\alpha+\beta)E[\sigma^2_t] + \gamma E[\sigma^2_t] \Longrightarrow E[\sigma^2_t] = \frac{\omega}{1-\alpha-\beta-\gamma}. \end{aligned}$$

Because $E[\sigma^2_t] = \delta_0/(1-\delta_1)$ and also $E[\sigma^2_t] = \omega/(1-\alpha-\beta-\gamma)$, we derive the restriction that

$$\delta_0/(1-\delta_1) = \frac{\omega}{(1-\alpha-\beta-\gamma)}$$

should hold, which is an interesting and testable restriction.

In case you want to get "your hands dirty" with the data, we did that for you. We have asked whether the VIX index, more precisely the logarithm of $VIX^2/252$ may be driving the variance of US stock returns over the sample period February 1990 - February 2012. We have estimated an OLS regression of log square returns on the scaled squared VIX to find (standard errors are reported in parenthesis):

$$\log(1+R_{t+1})^2 = \underset{(0.7193)}{7.7125} + \underset{(0.0501)}{1.3097}\log(\frac{VIX^2}{252}).$$

Even though the coefficients are highly significant, the R-square of the regression is 10.6%, i.e., VIX plays a role in determining the variance of returns (what a surprise!), it is clearly unable alone to capture all the variance. Graphical results are plotted below in figure 10.

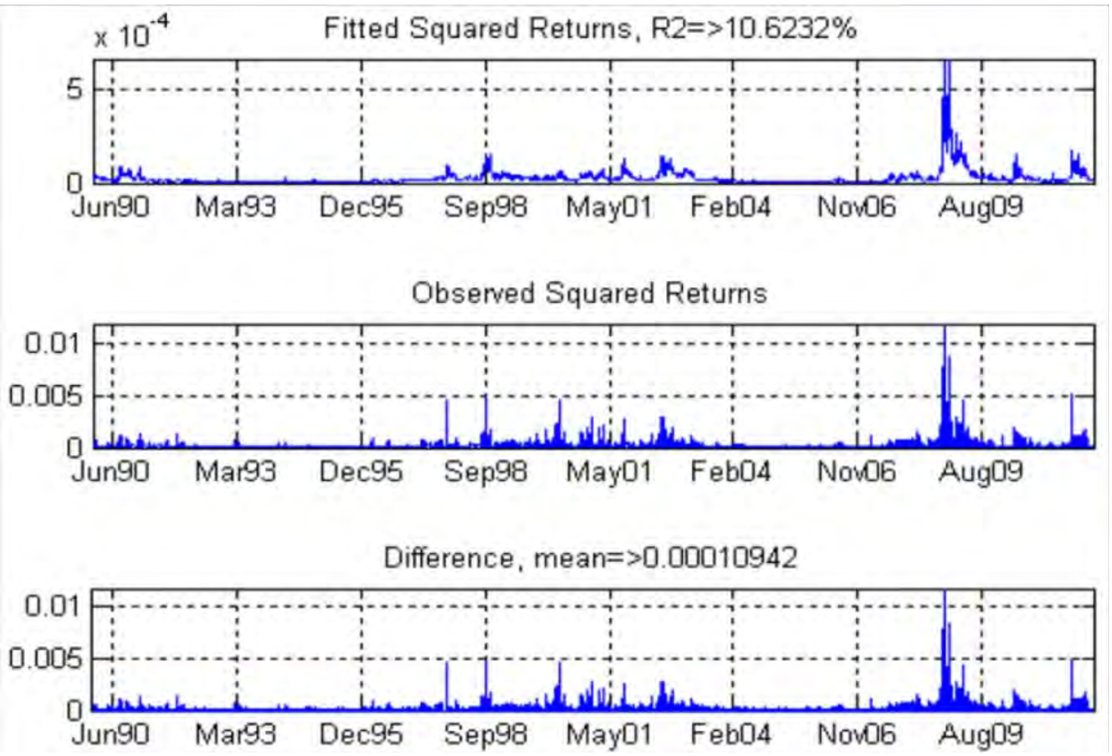


Figure 10: Estimation output from regression of squared log (gross) returns on the CBOE VIX index

[39] For the asset pricing buffs, $E[VIX_t] = E[\sigma^2_t]$ may pose some problems, as VIX is normally calculated under the risk-neutral measure while $E[\sigma^2_t]$ refers to the physical measure. If this bothers you, please assume the two measures are the same, which means you are assuming local risk-neutrality.

6. **Estimation of GARCH Models**

In a frequentist statistical perspective—which is the one adopted in most of your financial econometrics sequence—to estimate the parameters of a GARCH model means that, given a random sample of data on asset returns, one wants to select a vector of parameters $\boldsymbol{\theta} \in \mathcal{R}^K$ in a way that maximizes some criterion function that measures how *plausible* each possible value for $\boldsymbol{\theta}$ is relative to the recorded sample.[40] The parameters collected in $\boldsymbol{\theta} \in \mathcal{R}^K$ are fixed, but they are also unknown. K is the number of parameters to be estimated. Frequentist inferential methods aim at recovering $\boldsymbol{\theta}$ from some sample of data, randomly obtained from an underlying population, the true data generating process. The choice of the criterion and of a method to maximize it, defines the estimation method and as such one specific type of *estimator*. This general principle will be made clear later on. However, to gain some intuition, consider two examples. First, we may look for a unique $\hat{\boldsymbol{\theta}} \in \mathcal{R}^K$ such that the probability that the observed data sample has been generated by the assumed stochastic process is maximized when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. One such estimator will be the maximum likelihood estimator. Second, we may look for a unique $\tilde{\boldsymbol{\theta}} \in \mathcal{R}^K$ such that some features implied by the data generating process—for instance, some interesting moments, such as unconditional means and variances—are the same when computed from the assumed stochastic process when $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ as in the observed data; one such estimator, based on matching sample with population moments, is the method-of-moments estimator that we shall encounter in the following chapters. For the time being, we focus on maximum likelihood estimators of $\boldsymbol{\theta}$. Here $\boldsymbol{\theta}$ collects all the parameters of interest, for instance ω, α, and β in the case of a plain-vanilla GARCH(1,1) Gaussian model. In this case, $K = 3$, in principle $\boldsymbol{\theta} \equiv [\omega\ \alpha\ \beta]' \in \mathcal{R}^3$, but we know already that positivity constraints will be imposed so that in fact $\boldsymbol{\theta} \in \mathcal{R}^3_+$, where $\mathcal{R}^3_+$ is just the sub-set of strictly positive real numbers.[41]

As you may recall from your statistics sequence, given the need to choose some criterion function to be "optimized" (often, maximized) and the fact that many alternative criteria can be proposed (see our earlier example of two different types of criteria), to perform point estimation, you will need not only to propose one estimator (or method of estimation) but also this estimator should better have "good" properties.[42] For GARCH models, maximum likelihood estimation (MLE) is

[40] Recall that in a frequentist framework, the data are fixed but are considered a realization (say, $(R_1, R_2, ..., R_T)$) of a random sample from the stochastic process $\{R_t\}_{t=1}^T$. Because in practice estimators will yield estimates that are a function of the data $(R_1, R_2, ..., R_T)$ and these are from a random sample, the estimator will be a function of the random sample, and as such itself a random variable (also called, a statistic). For instance, you will recall that $\hat{\mathbf{b}}^{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$; because $\mathbf{y}$ collects realizations from a random sample, $\hat{\mathbf{b}}^{OLS}$ itself is a random vector. Let's add that in fact, you have encountered a few occasional exceptions to the frequentist approach, for instance Black and Littermmann's methods in portfolio choice use a Bayesian approach to inference that differs from the frequentist one.

[41] Of course, additional constraints, such as the stationarity restriction $\alpha + \beta < 1$, will impose further limitations to the set to which $\boldsymbol{\theta}$ may belong to, in which case we write $\boldsymbol{\theta} \in \Theta \subseteq \mathcal{R}^3_+$.

[42] The typical properties of an estimator $\hat{\boldsymbol{\theta}}_T$ that are examined in a standard statistics course are: unibiasedness, $E[\hat{\boldsymbol{\theta}}_T] = \boldsymbol{\theta}$; consistency, in heuristic terms the fact that as $T \to \infty$, $\hat{\boldsymbol{\theta}}_T$ converges to the true but unknown $\boldsymbol{\theta}$; efficiency, the fact that among the estimators that are asymptotically unbiased, $\hat{\boldsymbol{\theta}}_T$ has the smallest possible (co)variance. Notice that several alternative models of convergence may be employed to define consistency. Moreover, ruling out pathological cases, it is clear that if $E[\hat{\boldsymbol{\theta}}_T] = \boldsymbol{\theta}$, it will be easy to establish that as $T \to \infty$, $\hat{\boldsymbol{\theta}}_T$ converges to the true

such a method.

6.1. *Maximum likelihood estimation*

MLE is based on knowledge of the likelihood function of the sample of data, which is affine (i.e., it is not always identical, but for all practical purposes, it is) to the joint probability density function (PDF) of the same data. In general, models that are estimated by maximum likelihood must be fully specified *parametric* models, in the sense that once the parameter values are known, all necessary information is available to simulate the (dependent) variable(s) of interest; yet, if one can simulate the process of returns, this means that their PDF must be known, both for each observation as a scalar random variable, and for the full sample as a vector random variable. The intuition of ML estimation has been already illustrated above: to look for a unique $\hat{\boldsymbol{\theta}} \in \Theta$ (Θ is the space of possible values of the parameters, to accommodate any restrictions or constraints) such that the joint, total probability that the observed data sample has been generated by the assumed stochastic process parameterized by $\boldsymbol{\theta}$ is maximized when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. In what follows, for concreteness, we refer to the MLE for a standard GARCH(1,1) model, when $\boldsymbol{\theta} \equiv [\omega \ \alpha \ \beta]'$. However, it will be clear that these concepts easily generalize to all conditional heteroskedastic models covered in this chapter and therefore to any possible structure for $\boldsymbol{\theta} \in \Theta$.

The assumption of IID normal shocks (z_t),

$$R_{t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1),$$

implies (from normality and identical distribution of z_{t+1}) that the density of the time t observation is:

$$l_t \equiv \Pr(z_t; \boldsymbol{\theta}) = \frac{1}{\sigma_t(\boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{R_t^2}{\sigma_t^2(\boldsymbol{\theta})}\right),$$

where the notation $\sigma_t^2(\boldsymbol{\theta})$ emphasizes that conditional variance depends on $\boldsymbol{\theta} \in \Theta$. Because each shock is independent of the others (from independence over time of z_{t+1}), the total probability density function (PDF) of the entire sample is then the product of T such densities:

$$L(R_1, R_2, ..., R_T; \boldsymbol{\theta}) \equiv \prod_{t=1}^{T} l_t = \prod_{t=1}^{T} \frac{1}{\sigma_t(\boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{R_t^2}{\sigma_t^2(\boldsymbol{\theta})}\right). \tag{16}$$

This is called the *likelihood function.* However, because it is more convenient to work—especially when we are about to take the derivatives required by first-order conditions, and also to avoid numerical problems when computers are involved—with sums than with products, we usually consider the natural logarithm of the likelihood function,

$$\mathcal{L}(R_1, R_2, ..., R_T; \boldsymbol{\theta}) \equiv \log L(R_1, R_2, ..., R_T; \boldsymbol{\theta}) = \log \prod_{t=1}^{T} l_t = \sum_{t=1}^{T} \log l_t$$

but unknown $\boldsymbol{\theta}$ (e.g., a law of large numbers will be sufficient because in this case as as $T \to \infty$, $E[\hat{\boldsymbol{\theta}}_T] \to \boldsymbol{\theta}$).

$$= \sum_{t=1}^{T}\left[-\log\sigma_t(\boldsymbol{\theta}) - \log\sqrt{2\pi} - \frac{1}{2}\frac{R_t^2}{\sigma_t^2(\boldsymbol{\theta})}\right]$$

$$= -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\sigma_t^2(\boldsymbol{\theta}) - \frac{1}{2}\sum_{t=1}^{T}\frac{R_t^2}{\sigma_t^2(\boldsymbol{\theta})}, \qquad (17)$$

where we have used several obvious properties of natural logarithms, including the fact that $\log\sqrt{x} = \log x^{1/2} = 0.5\log x$ and $\log\sigma_t(\boldsymbol{\theta}) = \log\sqrt{\sigma_t^2(\boldsymbol{\theta})} = 0.5\log\sigma_t^2(\boldsymbol{\theta})$. $\mathcal{L}(R_1, R_2, ..., R_T; \boldsymbol{\theta})$ is also called *log-likelihood function* and the notation employed emphasizes that it is the log joint probability of the sample of data, given a choice for the parameter vector $\boldsymbol{\theta} \in\Theta$. However, nothing prevents you from seeing the log-likelihood as a function that simply depends on the unknown parameters in (say) $\boldsymbol{\theta} \equiv [\omega\ \alpha\ \beta]'$. Note that whatever value of $\boldsymbol{\theta} \in\Theta$ maximizes(17) will also maximize the likelihood function (16), because $\mathcal{L}(R_1, R_2, ..., R_T; \boldsymbol{\theta})$ is just a monotonic transformation of $L(R_1, R_2, ..., R_T; \boldsymbol{\theta})$. Therefore MLE is simply based on the idea that once the functional form of (17) has been written down, for instance

$$\mathcal{L}(R_1, R_2, ..., R_T; \boldsymbol{\theta}) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\left[\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2\right] - \frac{1}{2}\sum_{t=1}^{T}\frac{R_t^2}{\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2},$$

and initialized at

$$\sigma_0^2 = \frac{\omega}{1-\alpha-\beta},$$

simply maximizing the log-likelihood to select the unknown parameters,

$$\max_{\boldsymbol{\theta}\in\Theta}\left\{-\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\left[\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2\right] - \frac{1}{2}\sum_{t=1}^{T}\frac{R_t^2}{\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2}\right\}$$

will deliver the MLE, denoted as $\hat{\boldsymbol{\theta}}_T^{ML}$, or

$$\hat{\boldsymbol{\theta}}_T^{ML} = \arg\max_{\boldsymbol{\theta}\in\Theta}\left\{-\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\left[\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2\right] - \frac{1}{2}\sum_{t=1}^{T}\frac{R_t^2}{\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2}\right\}.$$

Here the reference to some need to "initialize" σ_0^2 refers to the fact that the log-likelihood function has a clear recursive structure: given σ_0^2, $\omega + \alpha\beta\sigma_0^2$ can be evaluated and therefore the $t = 1$ term of $\mathcal{L}$ can be numerically assessed for a given choice of ω and α;[43] at this point, given the value of σ_1^2, $\omega + \alpha R_1^2 + \alpha\beta\sigma_1^2$ can be evaluated and therefore the $t = 2$ term of $\mathcal{L}$ can be numerically assessed for a given choice of ω, α, and β. The algorithm proceeds now iteratively until time T, when given the value of σ_{T-1}^2, $\omega + \alpha R_{T-1}^2 + \alpha\beta\sigma_{T-1}^2$ can be evaluated and therefore the $t = T$ term of $\mathcal{L}$ can be numerically assessed for a given choice of ω α, and β.

Another aspect needs some care: note that $\hat{\boldsymbol{\theta}}_T^{ML}$ is the maximizer of the log-likelihood function for $\boldsymbol{\theta} \in\Theta$. As already mentioned, this is a compact way to state that ML estimation may be performed subject to a number of constraints, such as positivity restrictions on the parameters and the

[43] R_0^2 does not appear because it is not available and it is implicitly set to zero, which in this corresponds to the unconditional mean of the process. You know from your ML estimation theory for AR(q) models, that this is not an innocent choice. However, asymptotically, for $T \to \infty$ as it is frequently assumed in finance, such a short-cut will not matter.

stationarity condition by which $\alpha+\beta<1$. How do you do all this amazing amount of calculations? Surely enough, not using paper and pencil. Note that even in our short description of the recursive structure of the log-likelihood function calculation, that was done only for a given choice of the parameters $\boldsymbol{\theta} \in \Theta$: infinite such choices remain possible. Therefore, at least in principle, to maximize $\mathcal{L}$ you will then need to repeat this operation an infinite number of times, to span all the vectors of parameters in Θ. Needless to say, it takes an infinite amount time to span all of Θ. Therefore, appropriate methods of numerical, constrained optimization need to be implemented: this is what packages such as Matlab, Gauss or Stata are for.[44]

What about the desired good properties of the estimator? ML estimators have very strong theoretical properties:

- They are *consistent* estimators: this means that as the sample size $T \to \infty$, the probability that the estimator $\hat{\boldsymbol{\theta}}_T^{ML}$ (in repeated samples) shows a large divergence from the true (unfortunately unknown) parameter values $\boldsymbol{\theta}$, goes to 0.

- They are the *most efficient* estimators (i.e., those that give estimates with the smallest standard errors, in repeated samples) among all the (asymptotically) unbiased estimators.[45]

The concept of efficiency begs the question of how does one compute standard errors for ML estimates, in particular with reference to GARCH estimation. If the econometric model is correctly specified, such an operation is based on the concept of *information matrix*, that under correct model specification is given by:

$$\mathcal{I}(\boldsymbol{\theta}) = \lim_{T\to\infty} -E\left[\frac{1}{T}\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]. \tag{18}$$

Correct specification means that the conditional mean and variance functions (i.e., μ_{t+1} and σ_{t+1}^2) should be correct and that the parametric distribution of the shocks (here, so far it was $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$) is also correct. Visibly, the information matrix is based on the Hessian of the MLE problem.[46] In fact, under the assumption of correct specification, the result in (18) is called *information*

[44] For instance, Newton's method makes use of the Hessian, which is a $K\times K$ matrix $\mathcal{H}(\boldsymbol{\theta}) \equiv \partial^2\mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ that collects second partial derivatives of the log-likelihood function with respect to each of the parameters in $\boldsymbol{\theta}$. Similarly the gradient $\partial\mathcal{L}(\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}$ collects the first partial derivatives of the log-likelihood function with respect to each of the elements in $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_j$ denote the value of the vector of estimates at step j of the algorithm, and let $\partial\mathcal{L}(\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}$ and $\mathcal{H}(\hat{\boldsymbol{\theta}}_j)$ denote, respectively, the gradient and the Hessian evaluated at $\hat{\boldsymbol{\theta}}_j$. Then the fundamental equation for Newton's Method is $\hat{\boldsymbol{\theta}}_{j+1} = \hat{\boldsymbol{\theta}}_j - \mathcal{H}^{-1}(\hat{\boldsymbol{\theta}}_j)[\partial\mathcal{L}(\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}]$. Because the log-likelihood function is to be maximized, the Hessian should be negative definite, at least when $\hat{\boldsymbol{\theta}}_j$ is sufficiently near $\hat{\boldsymbol{\theta}}_T$. This ensures that this step is in an uphill direction.

[45] What does asymptotically unbiased mean? Something related to consistency (not exactly the same, but the same for most cases) and for the time being, you may ignore the details of the technical differences between the two concepts. One indirect but equivalent way to state that the MLE is the most efficient estimator is to state that "it achieves the Crámer-Rao lower bound" for the variance of the estimator. Such famous bound represents the least possible covariance matrix among all possible estimators, $\hat{\boldsymbol{\theta}}$.

[46] Wow, big words flying here... The Hessian is simply the matrix of second partial derivatives of the objective function—here the log-likelihood function—and the vector of parameters $\boldsymbol{\theta} \in \Theta$. Let's quickly review it with one

matrix equality (to the Hessian). In particular, it is the inverse of the information matrix, $\mathcal{I}^{-1}(\boldsymbol{\theta})$ that will provide the asymptotic covariance of the estimates:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T^{ML} - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta})\right),$$

where $\xrightarrow{D}$ denotes convergence in distribution. Obviously, this result implies that $\hat{\boldsymbol{\theta}}_T^{ML} \xrightarrow{a} \boldsymbol{\theta}$.[47] Consistent estimates of the information matrix may be calculated from T sample observations as:[48]

$$\mathcal{I}_T(\hat{\boldsymbol{\theta}}_T^{ML}) = -\frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right],$$

where, for instance, in the GARCH(1,1) case the log-likelihood contribution $\mathcal{L}(R_t;\boldsymbol{\theta})$ is:

$$\mathcal{L}(R_t;\boldsymbol{\theta}) \equiv -\log 2\pi - \frac{1}{2}\log\left[\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2\right] - \frac{1}{2}\frac{R_t^2}{\omega + \alpha R_{t-1}^2 + \beta\sigma_{t-1}^2}.$$

The information matrix measures the average amount of information about the parameters that is contained in the observations of the sample. As $T \to \infty$, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_T^{ML}$ allows us to approximate its variance as:

$$Var[\hat{\boldsymbol{\theta}}_T^{ML}] \simeq \left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right\}^{-1}. \tag{19}$$

The inverse of this matrix can be used for hypothesis testing by constructing the usual z-ratio statistic. As usual, asymptotically valid tests of hypothesis are built as z ratios that have a structure similar to t-ratios, although their normal distribution obtains only asymptotically, as $T \to \infty$. For instance, consider testing the null hypothesis that the parameter $\alpha = \alpha^*$ (α^* is not necessarily zero, but $\alpha^* = 0$ is very common) from a GARCH(1,1), i.e., $H_0 : \alpha = \alpha^*$. The first step is to find the MLE estimate $\hat{\alpha}_T^{ML}$. Second, we compute an estimate of the covariance matrix, i.e.

$$\mathbf{e}_2'\left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right\}^{-1}\mathbf{e}_2,$$

example: given the function $\mathcal{L}(\theta_1, \theta_2)$, the Hessian is:

$$\frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2\mathcal{L}(\theta_1,\theta_2)}{\partial\theta_1^2} & \frac{\partial^2\mathcal{L}(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2} \\ \frac{\partial^2\mathcal{L}(\theta_1,\theta_2)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2\mathcal{L}(\theta_1,\theta_2)}{\partial\theta_2^2} \end{bmatrix}.$$

Clearly, the Hessian is a symmetric matrix because $\frac{\partial^2\mathcal{L}(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2} = \frac{\partial^2\mathcal{L}(\theta_1,\theta_2)}{\partial\theta_2\partial\theta_1}$. Also note that the main diagonal of the Hessian collects second partial derivatives vs. the same variable (here, parameter), while the off-diagonal elements collect the cross-partial derivatives.

[47] Technically, under adequate assumptions, this may be stated as $\hat{\boldsymbol{\theta}}_T^{ML}$ converging to $\boldsymbol{\theta}$ almost surely (a.s.), meaning that the event in which asymptotically $\hat{\boldsymbol{\theta}}_T^{ML} \neq \boldsymbol{\theta}$ has probability zero.

[48] Probably you are wondering about the origin of the negative sign in the definition of the Hessian. Just think about it: heuristically, you are maximizing the log-likelihood function, which is a function from $\Theta\subseteq\mathcal{R}^K$ into $\mathcal{R}$, $K \geq 1$; at any (also local) maximum a function that is being maximized will be concave; hence, in correspondence to $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, the second derivative should be negative; but for a function from $\Theta\subseteq\mathcal{R}^K$ into $\mathcal{R}$ such a second derivative is in fact the Hessian; hence the Hessian is expected to be negative at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$; only taking the opposite of the negative definite Hessian, one obtaines a positive definite covariance matrix, and we know that covariance matrix ought to be positive definite by construction.

where $\mathbf{e}_2 = [0\ 1\ 0]'$ (because α is the second element in $\boldsymbol{\theta} \equiv [\omega\ \alpha\ \beta]' \in \mathcal{R}^3_+$). Third, we define the ratio

$$z(\hat{\alpha}_T^{ML}; \alpha^*) \equiv \frac{\hat{\alpha}_T^{ML} - \alpha^*}{\mathbf{e}_2' \left\{ -\frac{1}{T} \sum_{t=1}^T \left[\left. \frac{\partial^2 \mathcal{L}(R_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right] \right\}^{-1} \mathbf{e}_2}$$

and compare it with a chosen critical value under a $\mathcal{N}(0, 1)$, assuming α^* belongs to the feasible set, $\Theta \subseteq \mathcal{R}^K$.[49]

6.2. *Quasi maximum likelihood (QML) estimation*

One key aspect needs to be further emphasized: although the idea of trying and finding a unique $\hat{\boldsymbol{\theta}}_T^{ML} \in \Theta$ that maximizes the joint probability that the sample of data actually came from the process parameterized by $\boldsymbol{\theta} \in \Theta$ is highly intuitive—it answers the layman question "let's rig the assumed model (e.g., a GARCH) to make it as consistent as possible to what we see out there in real life and real financial markets"—one detail should not go unnoticed: the fact that MLE requires knowledge of

$$R_{t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0, 1). \tag{20}$$

In fact, as we have seen, both the IID nature of z_{t+1} and the fact that $z_{t+1} \sim \mathcal{N}(0, 1)$ has been repeatedly exploited in building the log-likelihood function. What if you are not positive about the fact that (20) actually adequately describes the data? For instance, what if all you can say is that

$$R_{t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{D}(0, 1),$$

but it looks rather unlikely that $\mathcal{D}(0, 1)$ may actually turn out to be a $\mathcal{N}(0, 1)$?[50] Can we still somehow do what we have described above and enjoy *some* of the good properties of MLE? The answer is a qualified—i.e., that will hold subject to specific but possibly verifiable conditions—"yes" and the resulting estimator is called a *quasi (or pseudo) maximum likelihood estimator* (QMLE). Interestingly, the corresponding statistical result is one of the most useful and frequently exploited finding in modern econometrics—in a way, as close to "magic" as econometrics can go.

The key finding concerning the QMLE estimator is that even though the conditional distribution of the shocks z_t is **not** normal (i.e., $z_{t+1} \sim$IID $\mathcal{D}(0, 1)$ and $\mathcal{D}$ does not reduce to a $\mathcal{N}$), *under some conditions*, an application of MLE based on $z_{t+1} \sim$IID $\mathcal{N}(0, 1)$ will yield estimators of the mean and variance parameters which converge to the true parameters as the sample gets infinitely large, i.e. that are *consistent*.[51] What are the conditions mentioned above? You will need that:

[49] For instance, if the test is based on a type I error of 5%, then if $|z(\hat{\alpha}_T^{ML}; \alpha^*)| \gtrapprox 1.96$, the null hypothesis of $\hat{\alpha}_T^{ML} = \alpha^*$ is rejected; if instead $|z(\hat{\alpha}_T^{ML}; \alpha^*)| < 1.96$, the null cannot be rejected. $\mathbf{e}_2' \left\{ -\frac{1}{T} \sum_{t=1}^T \left[\left. \frac{\partial^2 \mathcal{L}(R_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right] \right\}^{-1} \mathbf{e}_2$ is simply the matrix algebra operation that selects the second element on the diagonal of the approximate covariance matrix of $\hat{\boldsymbol{\theta}}$. You may find quicker ways to refer to this element of the main diagonal of the covariance matrix.

[50] For instance, you may feel that in fact $z_{t+1} \sim$IID t-student$(0, 1)$ may be more sensible. We will deal with this case extensively in the next chapter.

[51] Such conditions and technical details are presented in Bollerslev and Wooldridge (1992).

- The conditional variance function, σ^2_{t+1} seen as a function of the information at time t, $\mathcal{F}_t$, must be correctly specified.
- The conditional mean function, μ_{t+1} seen as a function of the information at time t, $\mathcal{F}_t$, must be correctly specified.

Two issues need to be clarified. First, "correctly specified" means that the mathematical, functional specification of the models for the conditional mean and variance are "right". In practice, most of this chapter may be taken as a survey of alternative and increasingly complex conditional variance functions. One example of what it means to mis-specify a model will help understanding what correct specification means. Suppose the world as we know it, is actually ruled—as far conditional variance of the market portfolio (say)— by a EGARCH(1,1) process:

$$\log \sigma^2_{t+1} = \omega + \beta \log \sigma^2_t + g\left(z_t\right) \qquad g\left(z_t\right) = \theta z_t + \alpha(|z_t| - E|z_t|).$$

However, you do not know it (how could you, given that until a few hours ago you were healthy and never heard of such a EGARCH gadget before?) and just out of sheer laziness you proceed to estimate a plain-vanilla, off-the-shelf GARCH(1,1) model,

$$\sigma^2_{t+1} = \omega + \alpha R^2_t + \beta \sigma^2_t.$$

Therefore the very functional form that you use, not to mention the fact that you should be paying attention to 4 parameters (ω, β, θ, and α in the EGARCH) and not 3 (ω, β, and α in the GARCH) will be a source of a violation of the needed assumptions to operationalize the QMLE. How would you know in practice that you are making a mistake and using the wrong model for the conditional variance? It is not easy and we shall return to this point, but one useful experiment would be: simulate a long time series of returns from (20) under some EGARCH(1,1). Instead of estimating such a EGARCH(1,1) model on the simulated data, estimate mistakenly a GARCH(1,1) model and look at the resulting standardized residuals, $\hat{z}_{t+1} = R_{t+1}/\hat{\sigma}^{GARCH}_{t+1}$, where the hat alludes to the fact that the GARCH standard deviations have been computed (filtered) under the estimated GARCH model. Because the data came from (20), you know that in a long sample you should never reject the (joint) null hypothesis that $\hat{z}_{t+1}$ $\sim$IID $\mathcal{N}(0,1)$. Trust me: if you performed this experiment, because you have incorrectly estimated a GARCH in place of a EGARCH, $\hat{z}_{t+1}$ $\sim$IID $\mathcal{N}(0,1)$ will be instead rejected in most long samples of data.[52] Second, note that the set of assumptions needed for the properties of QMLE to obtain include the correct specification of the conditional mean function, μ_{t+1}. Although technically this necessary and sufficient for the key QMLE result to obtain, clearly in this chapter this is not strictly relevant because we have assumed from the very beginning that

[52] One good reason for that is that the data are simulated to include asymmetric effects that you would be instead completely ignoring under a simpler, incorrect GARCH. Therefore $\hat{z}_{t+1}$ $\sim$IID $\mathcal{N}(0,1)$ will be rejected because the filtered standard residuals will have an asymmetric distribution, which is inconsistent with the null of $\mathcal{N}(0,1)$.

$\mu_{t+1} = 0$. However, more generally, also the assumption that μ_{t+1} has been correctly specified will have to be tested.[53]

This may feel as the classical case of "Too good to be true", and you would be right in your instincts: QMLE methods do imply a precise cost, in a statistical sense as they will in general be less efficient than ML estimators are. By using QMLE, we trade-off theoretical asymptotic parameter efficiency for practicality.[54]

In short, the QMLE result says that we can still use MLE estimation *based on normality assumptions* even when the shocks are not normally distributed, if our choices of conditional mean and variance function are defendable, at least in empirical terms. However, because the maintained model still has that $R_{t+1} = \sigma_{t+1} z_{t+1}$ with $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$, the shocks will have to be anyway IID: you can just do without normality, but the convenience of $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$ needs to be preserved. In practice, QMLE buys us the freedom to worry about the conditional distribution later on, and we shall, in the next chapter.

In this case, you will have to take our world for good, but it can be shown that although QMLE yields an estimator that is as consistent as the true MLE one (i.e., they both converge to the same, true $\boldsymbol{\theta} \in \Theta$), the covariance estimator of the QMLE needs to be adjusted with respect to (19). In the QMLE, the optimal estimator of $Var[\hat{\boldsymbol{\theta}}_T^{QML}]$ becomes:

$$\begin{aligned} Var[\hat{\boldsymbol{\theta}}_T^{QML}] \;\simeq\; & \left\{ -\frac{1}{T}\sum_{t=1}^{T} \left[\left. \frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\}^{-1} \left\{ -\frac{1}{T}\sum_{t=1}^{T} \left[\left. \frac{\partial \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\} \times \\ & \left\{ -\frac{1}{T}\sum_{t=1}^{T} \left[\left. \frac{\partial \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\}' \left\{ -\frac{1}{T}\sum_{t=1}^{T} \left[\left. \frac{\partial^2 \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\}^{-1}, \end{aligned}$$

where the $K \times 1$ vector $-\frac{1}{T}\sum_{t=1}^{T} \left[\left. \frac{\partial \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right]$ is called the sum of the sample gradients of the log-likelihood function, i.e., the first-partial derivative of the log-likelihood evaluated in correspondence to $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{QML}$. Such a vector is also called the sample *score* of the log-likelihood function.[55]

6.3. *Sequential estimators as QMLEs*

There is one special case in which we may indulge into QMLE estimation even though our key problem is not really the correct specification of the joint density of the shocks to returns, i.e., we

[53] Notice that all misspecification tests that you have encountered in your econometrics sequence so far concerned indeed tests of the correct specification of the conditional mean function, for instance when μ_{t+1} was a simple regression.

[54] Equivalently, a QMLE fails to "achieve the Crámer-Rao lower bound" for the variance among all possible estimators. Such lower bound is in fact attained by the MLE, which however requires that you can both correctly specify the joint density of the data and that shocks are IID.

[55] The elements of such a vector are K because $\boldsymbol{\theta}$ has K elements and therefore the same holds for $\partial\mathcal{L}(R_t;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Moreover,

$$\left\{ -\frac{1}{T}\sum_{t=1}^{T} \left[\left. \frac{\partial \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\} \left\{ -\frac{1}{T}\sum_{t=1}^{T} \left[\left. \frac{\partial \mathcal{L}(R_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \right\}'$$

is a $K \times K$ square, symmetric matrix.

may need to invoke the QMLE result even though (20) actually holds. This occurs when estimation of some vector of parameters $\boldsymbol{\theta} \in\Theta\subseteq\mathcal{R}^K$ is conveniently—this is only reason why we would do that, because we now understand that QMLE implies costs—split up in a number of *sequential estimation* stages. For instance, if $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']' \in \Theta$, the idea is that one would first estimate by full MLE $\boldsymbol{\theta}_1$ and then, conditional on the $\hat{\boldsymbol{\theta}}_1$ obtained during the first stage, estimate—again, at least in principle by full MLE—$\boldsymbol{\theta}_2$. Why would we do that? Sometimes because of practicality, because estimation would be otherwise much harder; in other occasions, to avoid numerical optimization.

The problem with sequential estimation is simply defined: successive waves of (seemingly) partial MLE that may even, at least on the surface, fully exploit (20) will not deliver the optimal statistical properties and characterization of the MLE. On the contrary, a sequential ML-based estimator may be characterized as a QMLE and as such it will be subject to the same limitations as all QMLEs are: loss of asymptotic efficiency. Intuitively, this is due to the fact that when we split $\boldsymbol{\theta}$ down into $[\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']'$ to separately estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, this very separation in a sequential estimator will imply that for all $\hat{\theta}_{1i} \in \hat{\boldsymbol{\theta}}_1$ and $\hat{\theta}_{2j} \in \hat{\boldsymbol{\theta}}_2$, $Cov[\hat{\theta}_{1i}, \hat{\theta}_{2j}] = 0$ even though empirically there is no presumption that this should or might be the case. A few examples will help to clarify this point but also to appreciate the potential advantages from sequential estimation.

6.3.1. **Example 1 (OLS estimation of ARCH models)**

Let's go back to our AR(1)-ARCH(1) example. We know what the right estimation approach is: MLE applied to full log-likelihood function, that in this case will take the form

$$\mathcal{L}(R_1, R_2, ..., R_T; \phi_0, \phi_1, \omega, \alpha) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\left[\omega + \alpha\epsilon_{t-1}^2\right] - \frac{1}{2}\sum_{t=1}^{T}\frac{(R_t - \phi_0 - \phi_1 R_{t-1})^2}{\omega + \alpha\epsilon_{t-1}^2}, \tag{21}$$

where $\epsilon_{t-1} \equiv R_{t-1} - \phi_0 - \phi_1 R_{t-2}$. Note that $\mathcal{L}(R_1, R_2, ..., R_T; \phi_0, \phi_1, \omega, \alpha, \beta)$ jointly and simultaneously depends on all the 4 parameters that characterize our AR(1)-ARCH(1) model. Yet, many of you have been subject to a temptation that has started many pages ago (so I am afraid): why not obtain the estimated OLS residuals from a simple regression as

$$\hat{\epsilon}_t = R_t - \hat{\phi}_0 - \hat{\phi}_1 R_{t-1}$$

(which incidentally already gives estimates for ϕ_0 and ϕ_1) and then separately estimate ω and α from maximization of

$$\mathcal{L}_2(\hat{\epsilon}_1, \hat{\epsilon}_2, ..., \hat{\epsilon}_T; \omega, \alpha) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log\left[\omega + \alpha\hat{\epsilon}_{t-1}^2\right] - \frac{1}{2}\sum_{t=1}^{T}\frac{\hat{\epsilon}_t^2}{\omega + \alpha\hat{\epsilon}_{t-1}^2},$$

where the $\{\hat{\epsilon}_t\}_{t=1}^T$ are considered as if they were data even though these are obtained conditional on the OLS estimates of $\hat{\phi}_0$ and $\hat{\phi}_1$. In this case, given $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2']'$, we have $\boldsymbol{\theta}_1 \equiv [\phi_0 \ \phi_1]'$ and $\boldsymbol{\theta}_2 \equiv [\omega \ \alpha]'$. Clearly, there is no illusion: this is a QMLE and the loss of efficiency vs. maximization of (21) may be dramatic. In fact, you even suspect that the very estimation of ϕ_0 and ϕ_1 by OLS in the first stage may be problematic, as in the case of an AR(q) process, MLE does not correspond to OLS. In short, OLS estimation of GARCH models should be avoided in favor of MLE.

6.3.2. Example 2 (variance targeting)

This is another common example of sequential estimation that frequently appears in practice. Because we know that the long-run (ergodic) variance from a GARCH(1,1) is $\bar{\sigma}^2 = \omega/(1-\alpha-\beta)$, instead of jointly estimating ω, α, and β, you simply set

$$\tilde{\omega} = (1-\alpha-\beta)\left[\frac{1}{T}\sum_{t=1}^{T} R_t^2\right]$$

for whatever values of α and β, where the term in square brackets is simply the sample variance of financial returns to be estimate beforehand, on the basis of the data. In this case, given $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1' \; \theta_2]'$, we have $\boldsymbol{\theta}_1 \equiv [\alpha \; \beta]'$ and $\theta_2 \equiv \omega$. Here the sample variance estimator for $\bar{\sigma}^2$, $\hat{S}^2 \equiv T^{-1}\sum_{t=1}^{T} R_t^2$, is itself a first-step MLE. Of course, the fact that a pre-MLE run of estimation concerning the sample variance to scale down the dimension of $\boldsymbol{\theta}$ makes the resulting estimates of $\hat{\boldsymbol{\theta}}_T$ a QMLE. There are, as usual, two obvious advantages from this approach: (i) you impose the long-run variance estimate on the GARCH model directly and avoid that the model may yield nonsensical estimates;[56] (ii) you have reduced the number of parameters to be estimated in the model by one. These benefits must be carefully contrasted with the well-known costs, the loss of efficiency caused by QMLE.

6.3.3. Example 3 (TARCH estimation in two steps or iteratively)

This is an academic example on which we shall follow through in our Matlab exercises. Given a GJR model,

$$\sigma_{t+1}^2 = \omega + \alpha R_t^2 + \alpha\delta I_t R_t^2 + \beta\sigma_t^2 \qquad I_t \equiv \begin{cases} 1 & \text{if } R_t < 0 \\ 0 & \text{if } R_t \geq 0 \end{cases},$$

the idea is to first perform a round of plain-vanilla GARCH estimation via MLE, by setting $\delta = 0$, thus obtaining estimates of ω, α, and β.[57] This also gives a filtered time series of GARCH variances,

$$\tilde{\sigma}_{t+1}^2 = \tilde{\omega} + \tilde{\alpha}R_t^2 + \tilde{\beta}\tilde{\sigma}_t^2,$$

where $\tilde{\omega}$, $\tilde{\alpha}$, and $\tilde{\beta}$ are first-round estimates.[58] In the second step, one simply estimates a regression

$$\frac{\tilde{\sigma}_{t+1}^2 - (\tilde{\omega} + \tilde{\alpha}R_t^2 + \tilde{\beta}\tilde{\sigma}_t^2)}{\tilde{\alpha}} = \delta(I_t R_t^2) + v_{t+1}$$

[56] Note that MLE is not set up to match the sample moments of the data: this means that once $\hat{\boldsymbol{\theta}}_T^{ML}$ is obtained, if the implied moments of the process—for instance, mean and variance—were computed, this may differ from those in the data because of the structure of the log-likelihood function that in general weighs means and variances in a highly non-linear fashion. We shall return on this distinction between MLE and method-of-moment estimators in the next chapter.

[57] We have changed the notation of the TARCH parameter that had been previously called θ to avoid confusion with the new meaning that the vector $\boldsymbol{\theta}$ has acquired in the meantime.

[58] We call *filtered* GARCH variances those that are obtained from a conditional variance model when the estimates of the parameters involved are plugged in the model and, given some starting condition σ_0^2, σ_t^2 is computed given the information in the sample: $\tilde{\sigma}_1^2 = \tilde{\omega} + \tilde{\beta}\sigma_0^2$; $\tilde{\sigma}_2^2 = \tilde{\omega} + \tilde{\alpha}R_1^2 + \tilde{\beta}\tilde{\sigma}_1^2$; ... $\tilde{\sigma}_T^2 = \tilde{\omega} + \tilde{\alpha}R_{T-1}^2 + \tilde{\beta}\tilde{\sigma}_{T-1}^2$.

to obtain a second-step estimate of δ, $\tilde{\delta}$. In this case, given $\boldsymbol{\theta} \equiv [\boldsymbol{\theta}_1' \; \theta_2]'$, we have $\boldsymbol{\theta}_1 \equiv [\omega \; \alpha \; \beta]'$ and $\theta_2 \equiv \delta$. One interesting idea is that the sequential estimation process does not stop at this stage: instead, the algorithm proceeds now to re-apply MLE to estimate a modified GARCH(1,1) model written as

$$(\sigma_{t+1}^2 - \tilde{\alpha}\tilde{\delta} I_t R_t^2) = \omega + \alpha R_t^2 + \beta \sigma_t^2,$$

to obtain new (Q)MLE estimates $\tilde{\omega}'$, $\tilde{\alpha}'$, and $\tilde{\beta}'$, to be followed by a new regression estimate $\tilde{\delta}'$. The algorithm may in principle be iterated until convergence, although this is rather rare in practice. Clearly, the iterative nature hardly affects the fact that we are facing another QMLE.

7. Evaluating Conditional Variance Models

Let's now move where the money is (or not): how can you tell whether a (univariate) volatility model works in practice? A number of methods—called diagnostic or misspecification checks—exist. In this concluding section, we simply discuss four among the many possible methods, even though a few more ideas on how to test whether conditional variance models are correctly specified will emerge in later chapters.

The first, rather simple (and already mentioned, to some extent) method consists of applying standard *univariate tests of normality*, that aim at checking whether data from a given stochastic process $\{X_t\}$ may have been generated by a normal random variable. In practice, if you have estimated the parameters of a conditional volatility model by MLE and exploited the assumption that $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$ in (20), then this implies that the standardized model residuals defined as $\hat{z}_{t+1} \equiv R_{t+1}/\hat{\sigma}_{t+1}$ should have a normal distribution with zero mean and unit variance, where $\hat{\sigma}_{t+1}$ denotes the time series of filtered standard deviations derived from the estimated volatility model. Moreover, because a standard normal distribution is symmetric around 0 and the thickness of its tails are used as benchmarks to measure tail thickness of all distributions (i.e., the excess kurtosis of a normal is set to 0 by construction), the empirical (unconditional, overall) distribution of $\hat{z}_{t+1}$ should be characterized by zero skewness and zero excess kurtosis. At this point, a typical approach consists of using Jarque and Bera's (JB) test : JB proposed a test that measures departures from normality in terms of the skewness and kurtosis of standardized residuals. Under the null hypothesis of normally distributed errors, the JB statistic has a known asymptotic distribution:[59]

$$\widehat{JB}(z) \equiv \frac{T}{6}\left[\underbrace{\widehat{Skew}(z)}_{=0 \text{ under } N(0,1)}\right]^2 + \frac{T}{24}\left[\underbrace{\widehat{Kurt}(z) - 3}_{=0 \text{ under } N(0,1)}\right]^2 \overset{a}{\sim} \chi_2^2,$$

[59] In the expression that follows, we define:

$$\widehat{Skew}(z) \equiv \frac{\sum_{t=1}^T \hat{z}_t^3}{\left(\sum_{t=1}^T \hat{z}_t^2\right)^{3/2}} \quad \widehat{Kurt}(z) \equiv \frac{\sum_{t=1}^T \hat{z}_t^4}{\left(\sum_{t=1}^T \hat{z}_t^2\right)^2}.$$

The intuition behind these scaled unconditional sample moments will be further explained in the next chapter.

where "hats" denote samples estimates of the moments under investigation. Clearly, $\widehat{JB}(z) = 0$ under the null of normality; a large value of $\widehat{JB}(z)$ denotes a departure from normality, and JB tests will formally reject the null of normality when $\widehat{JB}(z)$ exceeds the critical value under a χ_2^2. This means that when the null of normality is rejected, then there is evidence against $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$, which is an indication of model misspecification.

A second method echoes our earlier tests of time series independence of z_{t+1}: this derives from the fact that even though normality has not been assumed (this is the case of QMLE) so that the assumed model for returns is $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$ and $\mathcal{D}(0,1)$ is not $\mathcal{N}(0,1)$, a correctly specified anyway implies

$$z_{t+1} \sim \text{IID}.$$

As we know, independence implies that $\hat{Q}_k^g(z) \simeq 0$ for all $k \geq 1$ where

$$\hat{Q}_k^g(z) \equiv T \sum_{\tau=1}^{k} (\hat{\rho}_\tau^g)^2 \overset{a}{\sim} \chi_k^2 \qquad \hat{\rho}_\tau^g \equiv \frac{\sum_{t=1}^{T-\tau}(g(z_t) - \overline{g(z_t)})(g(z_{t+\tau}) - \overline{g(z_t)})}{\sum_{t=1}^{T-\tau}(g(z_t) - \overline{g(z_t)})}$$

and $g(\cdot)$ is any (measurable) function. Because we are testing the correct specification of a conditional volatility model, it is typical to set $g(x) = x^2$, i.e., we test whether the squared standardized residuals, $\hat{z}_{t+1}^2 \equiv R_{t+1}^2/\hat{\sigma}_{t+1}^2$, display any systematic autocorrelation patterns. As it is now clear, one often simply uses sample autocorrelations to test the null of IID standardized residuals, possibly with tests based on the Bartlett's asymptotic standard errors. For instance, figure 11 shows a case in which there is little or no serial correlation in the levels of z_t, but there is some serial correlation left in the squares, at low orders: probably this means that one should build a different/better volatility model.

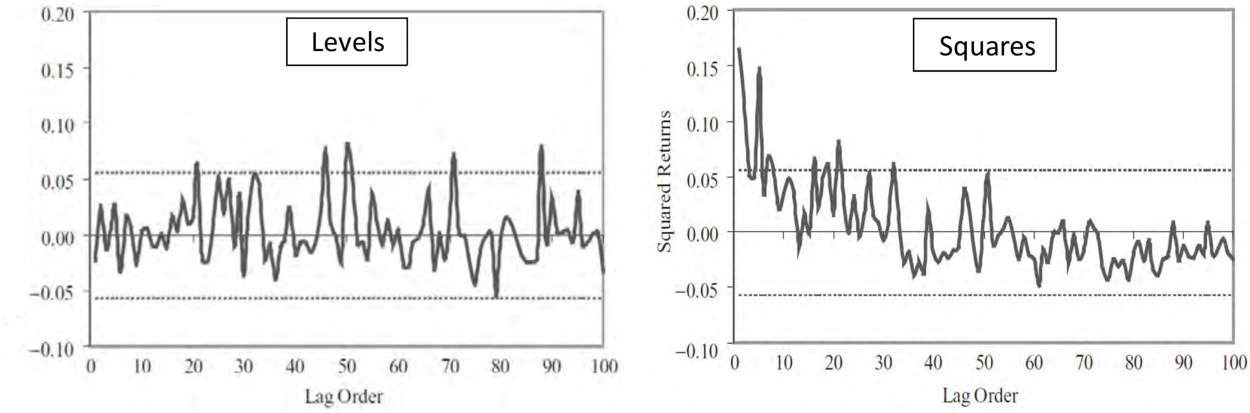


Figure 11: Sample autocorrelations for standardized residuals from a GARCH(1,1) for S&P 500 returns

However, the more informative way in which conditional volatility models are typically tested for misspecification is by a smart use of so-called "variance regressions". The idea is simply to regress squared returns computed over a forecast period on the forecasts derived from the conditional variance model under examination:[60]

$$R^2_{t+1} = b_0 + b_1\hat{\sigma}^2_{t+1} + e_{t+1},$$

where e_{t+1} follows a white noise process, i.e., $e_{t+1} \sim \mathcal{D}(0,1)$. Estimation may be simply performed using OLS, no sweat. Let's first state how one proceeds to use such a regression to test whether the conditional variance forecasts obtained from the model, $\hat{\sigma}^2_{t+1}$, are consistent with the null hypothesis of correct specification: in this case, $b_0 = 0$ and $b_1 = 1$. When $b_0 = 0$, we say that the variance model yields *unbiased* forecasts; $b_1 = 1$ implies that the variance model is *efficient*. Our goal is then to use standard OLS inference (as you have learned it from the first part of the Financial Econometrics sequence) to test whether $b_0 = 0$ and $b_1 = 1$. The reason for why correct specification is equivalent to $b_0 = 0$ and $b_1 = 1$ is that under these restrictions

$$R^2_{t+1} = \hat{\sigma}^2_{t+1} + e_{t+1} \Longleftrightarrow E_t[R^2_{t+1}] = \hat{\sigma}^2_{t+1}, \tag{22}$$

which is indeed what we expect of an unbiased and efficient forecast.

This variance forecast regression has however one problem: the squared returns are used as a proxy (technically, estimator) for the true but unobserved variance in period $t+1$; one wonders, whether this proxy for squared returns is any good. On the one hand, in principle we are fine because from our model $R_{t+1} = \sigma_{t+1}z_{t+1}$ with $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$, so that $E_t[R^2_{t+1}] = \sigma^2_{t+1}E_t[z^2_{t+1}] = \sigma^2_{t+1}$ because $Var_t[z^2_{t+1}] = 1 = E_t[z^2_{t+1}] - \{E_t[z_{t+1}]\}^2 = E_t[z^2_{t+1}]$ by assumption. $E_t[R^2_{t+1}] = \sigma^2_{t+1}$ means that R^2_{t+1} is an unbiased estimator of conditional variance. On the other hand, you know better than assessing estimators just on the basis of their being unbiased: the optimal estimator ought to be also the most efficient one. Therefore one wonders what the variance of R^2_{t+1} as an estimator of σ^2_{t+1} is:

$$\begin{aligned}
Var_t[R^2_{t+1}] &= E_t[(R^2_{t+1} - E_t[R^2_{t+1}])^2] = E_t[(R^2_{t+1} - \sigma^2_{t+1})^2] = E_t[(\sigma^2_{t+1}z^2_{t+1} - \sigma^2_{t+1})^2] \\
&= E_t[\sigma^2_{t+1}(z^2_{t+1} - 1)^2] = \sigma^4_{t+1}E_t[z^4_{t+1} - 2z^2_{t+1} + 1] \\
&= \sigma^4_{t+1}\{\underbrace{E_t[z^4_{t+1}]}_{=\kappa} - 2\underbrace{E_t[z^2_{t+1}]}_{=1} + 1\} = \sigma^4_{t+1}(\kappa - 1),
\end{aligned}$$

where κ is the kurtosis coefficient of z_{t+1}.[61] Because κ for typical (especially, daily) empirical standardized residuals tends to be much higher than 3, the variance of the square return proxy for realized variance is often very poor (i.e., imprecisely estimated), in the sense that $Var_t[R^2_{t+1}]$

[60] It just occured to me: R^2_{t+1} has nothing to do with the OLS coefficient of determination, R^2, often also called "R-square"!

[61] Note that there is no contradiction between $E_t[z^4_{t+1}] = \kappa$ and our general assumptions that $R_{t+1} = \sigma_{t+1}z_{t+1}$ with $z_{t+1} \sim$ IID $\mathcal{D}(0,1)$. Naturally, when $R_{t+1} = \sigma_{t+1}z_{t+1}$ with $z_{t+1} \sim$ IID $\mathcal{N}(0,1)$, then $E_t[z^4_{t+1}] = 3$ and $Var_t[R^2_{t+1}] = 2\sigma^4_{t+1}$. As for the fact that $E_t[z^4_{t+1}] = \kappa$ is the kurtosis coefficient, note that

$$Kurt_t(z_{t+1}) \equiv \frac{E_t[z^4_{t+1}]}{\{E_t[z^2_{t+1}]\}^2} = \frac{E_t[z^4_{t+1}]}{\{1\}^2} = E_t[z^4_{t+1}].$$

in excess of 10 times σ^4_{t+1} emerges not infrequently. More generally, if we take the coefficient of variation (defined as $E[\hat{\theta}]/\sqrt{Var[\hat{\theta}]}$) as a measure of the variability of an estimator, then

$$\frac{E_t[R^2_{t+1}]}{Var_t[R^2_{t+1}]} = \frac{\sigma^2_{t+1}}{\sqrt{\sigma^4_{t+1}(\kappa - 1)}} = \frac{1}{\sqrt{\kappa - 1}}$$

and this coefficient declines as κ increases. Due to the high degree of noise in squared financial returns, the fit of the variance regression as measured by the regression R^2 (coefficient of determination) is typically very low, typically around 5 to 10%, even if the variance model used to forecast is indeed the correctly specified one. Thus obtaining a low R^2 in such regressions should not lead one to reject a variance model even though the fact that variance regressions lead to a poor fit is certainly not something that can be completely dismissed. What can be done about the fact that (22) is based on an estimator of realized variance, R^2_{t+1}, that is extremely inefficient? Simply enough, to replace the estimator with a better estimator. How can that be done, will be analyzed in later chapters.

Finally, alternative conditional heteroskedastic models can also be compared using *penalized measures of fit* which trade-off in-sample fit with parsimony, i.e., whose value increases as the fit to the data improves but also decreases as the number of estimated parameters increase. Since your early age you have been familiar with one such measure, the adjusted R^2 (often denoted as $\bar{R}^2$) which, indeed, penalizes the standard R^2 with a measure of the parameter vector dimension to prevent that big models have an unfair advantage over smaller, tightly parameterized ones. Why do we value parsimony? Because in general terms the forecasting performance of a model improves as the number of parameters used to fit the data in sample declines—i.e., smaller models tend to perform better than bigger ones do. For instance, the general empirical finding is that, given an identical in-sample fit, e.g., a GARCH(1,1) model will perform better than a GARCH(2,2) when it comes to actual, out-of-sample volatility prediction because the latter implies two additional parameters to be estimated. This is of course the forecasting analog of Occam's razor. In a maximum likelihood set up, the traditional concept of $\bar{R}^2$ is generalized to *information criteria*: in the same way in which the $\bar{R}^2$ is based on the application of penalties to the classical coefficient of determination (R^2), information criteria are based on the concept of applying additional penalty terms to the maximized log-likelihood. Their general structure is:

$$-(\text{Maximized Log-Lik}) + l(dim(\hat{\boldsymbol{\theta}})),$$

where $l(\cdot)$ is a penalty function, and $dim(\hat{\boldsymbol{\theta}})$ is the notation for a counter of the number of different parameters in to be estimated in $\boldsymbol{\theta} \in\Theta$ (this was K in our early treatment). You may wonder way the maximized log-likelihood function enters information criteria with a negative sign: this is due to the fact that, as we have seen, most numerical optimization software actually minimize the negative of the log-likelihood function. Because the maximized log-likelihood is multiplied by -1 while the penalty has been added, it is clear that empirically we shall select models that actually *minimize* information criteria, not maximize them. Three information criteria are widely employed:

- The Bayes-Schwartz information criterion (BIC): $-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + (dim(\hat{\boldsymbol{\theta}})ln(T)/T)$; this criterion is known to select rather parsimonious models and it appears to be very popular in the applied literature.
- The Akaike information criterion (AIC): $-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2(dim(\hat{\boldsymbol{\theta}})/T)$; this criterion is also popular because it has optimal asymptotic properties (it is consistent, according to an appropriate definition), although it is also known to select too large non-linear models in small samples (GARCH are non-linear models).
- The Hannan-Quinn information criterion (H-Q):$-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2[dim(\hat{\boldsymbol{\theta}})\log(\log(T))/T]$; this criterion has been shown to perform very strongly in small samples and for non-linear models; numerically, it can be shown that it represents a compromise between BIC and AIC.

8. **Component GARCH Models: Short- vs. Long Run Variance Dynamics**

Engle and Lee (1999) have proposed a novel component GARCH model that expands the previously presented volatility models in ways that have proven very promising in applied option pricing (see e.g., Christoffersen, Jacobs, Ornthanalai, and Wang, 2008). Consider a model in which there is a distinction between the short-run variance of the process, h_t, that is assumed to follow a GARCH(1,1) process,

$$h_{t+1} = q_{t+1} + \alpha_1(R_t^2 - h_t) + \beta_1(h_t - q_{t+1}), \tag{23}$$

and the time-varying long-run variance, q_t, which also follows a GARCH(1,1) process

$$q_{t+1} = \alpha_0 + \rho(q_t - \alpha_0) + \phi(R_t^2 - h_t). \tag{24}$$

The distinction between h_{t+1} and q_{t+1} has been introduced to avoid any confusion with σ_{t+1}^2, when there is only one variance scale (you can of course impose $h_{t+1} = \sigma_{t+1}^2$ without loss of generality). This process implies that there is one conditional variance process for the short-run, as shown by (23), but that this process tends to evolve around (and mean-revert to) q_{t+1}, which follows itself the process in (24), which is another GARCH(1,1).

One interesting feature of this component GARCH model is it can re-written (and it is often estimated) as a GARCH(2,2) process. This interesting because as you may have been wondering about the actual use of GARCH(q,p) when $q \geq 2$ and $p \geq 2$. In fact, higher-order GARCH models are rarely used in practice, and this GARCH(2,2) case represents one of the few cases in which—even though it will be subject to constraints coming from the structure of (23) and (24)—implicitly a (2,2) case has been used in many practical applications. To see that (23)-(24) can be re-written as a GARCH(2,2), note first that the process for long-run variance may be written as $q_{t+1} = (1-\rho)\alpha_0 + \rho q_t + \phi(R_t^2 - h_t)$. At this point, plug the expression of q_{t+1} from (24) in (23):

$$\begin{aligned}
h_{t+1} &= (1-\beta_1)q_{t+1} + \alpha_1 R_t^2 + (\beta_1 - \alpha_1)h_t \\
&= (1-\beta_1)(1-\rho)\alpha_0 + (1-\beta_1)\rho q_t + (1-\beta_1)\phi(R_t^2 - h_t) + \alpha_1 R_t^2 + (\beta_1 - \alpha_1)h_t \\
&= (1-\beta_1)(1-\rho)\alpha_0 + (1-\beta_1)\rho q_t + [(1-\beta_1)\phi + \alpha_1]R_t^2 + \\
&\quad +[\beta_1 - \alpha_1 - (1-\beta_1)\phi]h_t \\
&= (1-\beta_1)(1-\rho^2)\alpha_0 + (1-\beta_1)\rho^2 q_{t-1} + [(1-\beta_1)\phi + \alpha_1]R_t^2 + (1-\beta_1)\rho\phi R_{t-1}^2 + \\
&\quad +[\beta_1 - \alpha_1 - (1-\beta_1)\phi]h_t - (1-\beta_1)\rho\phi h_{t-1} \\
&= \varpi + \alpha_1' R_t^2 + \alpha_2' R_{t-1}^2 + \beta_1' h_t + \beta_2' h_{t-1}
\end{aligned}$$

where we have exploited the fact that $E[q_{t-1}] = \alpha_0$ and set

$$\varpi = (1-\beta_1)\alpha_0 \qquad \alpha_1' = (1-\beta_1)\phi + \alpha_1$$
$$\alpha_2' = (1-\beta_1)\rho\phi \qquad \beta_1' = [\beta_1 - \alpha_1 - (1-\beta_1)\phi]$$
$$\beta_2' = -(1-\beta_1)\rho\phi.$$

One example may help you familiarize with this new, strange econometric model. Suppose that at time t, the long-run variance is 0.01 above short-run variance, it is equal to $(0.15)^2$ and is predicted to equal $(0.16)^2$ at time t. Yet, at time t returns are subject to a large shock, $R_t = -0.2$ (i.e., a massive -20%). Can you find values for $\alpha_1 \geq 0$ and $\beta_1 \geq 0$ such that you will forecast at time t short-run variance of zero? Because we know that $h_t - q_t = -0.01$, $q_{t+1} = 0.0225$, and $R_t^2 = 0.04$,

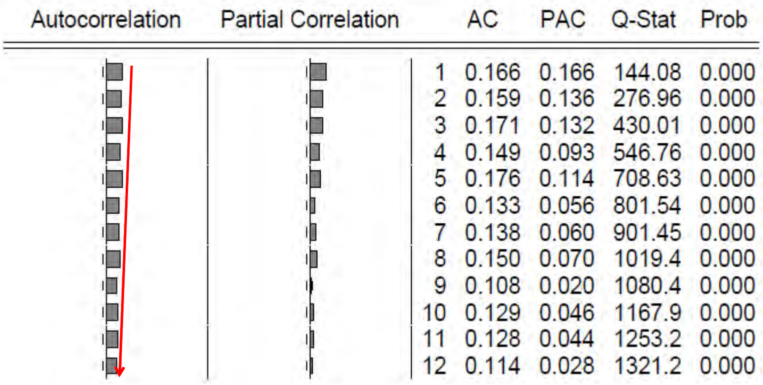
$$h_{t+1} = 0.0225 + \alpha_1(0.04 - 0.0125) + \beta_1(-0.01) = 0.0225 + 0.0275\alpha_1 - 0.01\beta_1$$

and we want to find a combination of $\alpha_1 \geq 0$ and $\beta_1 \geq 0$ that solves

$$0.0225 + 0.0275\alpha_1 - 0.01\beta_1 = 0 \qquad \text{or} \qquad \beta_1 = 2.25 + 2.75\alpha_1.$$

This means that such a value in principle exists but for $\alpha_1 \geq 0$ this implies that $\beta_1 \geq 2.25$.

Empirical, component GARCH models are useful because they capture the slow decay of autocorrelations in squared returns that we found in section 2 and that we reinforce here (as well as in the Matlab workout that follows). Consider for instance, the sample autocorrelogram obtained from a long 1926-2009 daily data set on S&P 500 returns in Figure 12. Clearly, the rate of decay in the level and significance of squared daily returns is very slow (technically, the literature often writes about volatility processes with a *long memory*, in the sense that shocks take a very long time to be re-absorbed). Component GARCH(1,1) models—also because of their (constrained) GARCH(2,2) equivalence—have been shown to provide an excellent fit to data that imply long memory in the variance process.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.166	0.166	144.08	0.000
		2	0.159	0.136	276.96	0.000
		3	0.171	0.132	430.01	0.000
		4	0.149	0.093	546.76	0.000
		5	0.176	0.114	708.63	0.000
		6	0.133	0.056	801.54	0.000
		7	0.138	0.060	901.45	0.000
		8	0.150	0.070	1019.4	0.000
		9	0.108	0.020	1080.4	0.000
		10	0.129	0.046	1167.9	0.000
		11	0.128	0.044	1253.2	0.000
		12	0.114	0.028	1321.2	0.000

Figure 12: Sample autocorrelations for squared daily S&P 500 returns

Appendix — A Matlab Workout

Suppose you are a German investor. Unless it is otherwise specified, you evaluate the properties and risk of your *equally weighted* stock portfolio on a daily basis. Using daily data in the file "data_daily.txt", construct daily portfolio returns. Please pay attention to the exchange rate transformations required by the fact that you are a German investor who measures portfolio payoffs in euros.[62]

1. For the sample period of 02/01/2006 - 31/12/2010 plot the time series of daily returns. Notice that in what follows, you will use this sample until otherwise instructed. [*Note*: When you run the code, you need to select the first time the file "data_daily.txt", and the second time the file "data_daily_string.txt" that will import strings to identify the series.]

2. Compute and plot the autocorrelogram functions (for up to 60 lags) for the (i) level, (ii) the square, and (iii) the absolute value of the equally weighted portfolio returns.

3. Plot the unconditional distribution of daily returns against a Gaussian (normal) density with the same empirical moments, i.e., with the same mean and variance.

4. Estimate a GARCH(1,1) model for the daily returns and plot the fitted values for *volatility*: (i) using the Matlab command *garchfit*,[63] and (ii) computing $\hat{\sigma}_t^2$ directly from

$$\hat{\sigma}_t^2 = \hat{\omega} + \hat{\alpha}R_{t-1}^2 + \hat{\beta}\sigma_{t-1}^2.$$

Compare the two series and verify whether these are identical; if they are not, explain why they are different.

[62] In case there is any residual confusion: a portfolio is just a choice of weights (in this case, a 3×1 vector) summing to one. 3×1 implies that you should be investing 100% in stocks. Equivalently, we are dealing with an equity diversification problem and not with a strategic asset allocation one. You can pick any real values, but it may be wise, to keep the current lab session sufficiently informative, to restrict weights to $(0, 1)$, possibly avoiding zeroes.

[63] Notice that the fitted volatility series automatically generated by this command corresponds to to the *sigma* output as defined in the function help

5. Estimate a RiskMetrics exponential smoother (i.e., estimate the RiskMetrics parameter λ) and plot the fitted conditional volatility series against those obtained from the GARCH(1,1).

6. Compute and plot daily one-day ahead recursive forecasts for the period 01/01/2011-31/01/2013 given the ML estimates for the parameters of the models in questions 4 and 5.

7. To better realize what the differences among GARCH(1,1) and RiskMetrics are when it comes to forecast variances in the long term, proceed to a 300-day long simulation exercise for four alternative GARCH(1,1) models: (i) with $\omega = 1$, $\alpha = 0.75$, $\beta = 0.2$; (ii) with $\omega = 1$, $\alpha = 0.2$, $\beta = 0.75$; (iii) with $\omega = 2$, $\alpha = 0.75$, $\beta = 0.2$; (iv) with $\omega = 2$, $\alpha = 0.2$, $\beta = 0.75$. Plot the process of the conditional variance under these alternative four models. In the case of models 1 and 2 ((i) and (ii)), compare the behavior of volatility forecasts between forecast horizons between 1- and 250-days ahead with the behavior of volatility forecasts derived from a RiskMetrics exponential smoother.

8. Estimate the 1% Value-at-Risk under the alternative GARCH(1,1) and RiskMetrics models with reference to the OOS period 01/01/2011-31/01/2013, given the ML estimates for the parameters of the models in questions 4 and 5. Compute the number of violations of the VaR measure. Which of the two models performed best and why?

9. Using the usual sample of daily portfolio returns, proceed to estimate the following three "more advanced" and asymmetric GARCH models: NGARCH(1,1), GJR-GARCH(1,1), and EGARCH(1,1). In all cases, assume that the standardized innovations follow an IID $N(0,1)$ distribution. Notice that in the case of the NGARCH model, it is not implemented in the Matlab *garchfit* toolbox and as a result you will have to develop and write the log-likelihood function in one appropriate procedure. After you have performed the required print on the Matlab screen all the estimates you have obtained and think about the economic and statistical strength of the evidence of asymmetries that you have found. Comment on the stationarity measure found for different volatility models. Finally, plot the dynamics of volatility over the estimation sample implied by the three alternative volatility models.

10. For the sample used in questions 4, 5, and 9, use the fitted variances from GARCH(1,1), RiskMetrics' exponential smoothed, and a GJR-GARCH(1,1) to perform an out-of-sample test for the three variance models inspired by the classical test that in the regression

$$R_t^2 = \alpha + \beta\widehat{\sigma}^{2,m}_{t,t-1} + \epsilon_t^m$$

$\alpha = 0$ and $\beta = 1$ to imply that $E_{t-1}[R_t^2] = \sigma_t^2 = \widehat{\sigma}^{2,m}_{t,t-1}$, where $\widehat{\sigma}^{2,m}_{t,t-1}$ is the the time $t-1$ conditional forecast of the variance from model m; moreover, as explained in the lectures, we would expect the R^2 of this regression to be high if model m explains a large portion of realized stock variance. In your opinion, which model performs best in explaining observed variance (assuming that the proxies for observed variances are squared returns)?

11. Assume now you are a German investor. Perform an asset allocation exercise using a simple Markowitz model using quarterly *excess* stock returns on three country price-indices: UK, US, and Germany. Starting from March 1976 until the end of the available data set, compute optimal weights, predicted (mean) returns and variances of your portfolio. Impose no short sale constraints on the stock portfolios and no borrowing at the riskless rate. Notice that this requires that you re-select your input data files: the first time the file "data_quarterly.txt", and the second time the file "data_quarterly _string.txt" that will import strings to identify the series. In particular, you are asked to compare three different mean and variance frameworks, to be able to appreciate how and whether volatility models affect financial decisions:

    (a) Variances: constant for all three indices; correlations: equal to the unconditional, constant sample correlations for all three pairs of indices; means: constant. This is of course a rather classical, standard Gaussian IID model in which means, variances, and covariances are all constant.

    (b) Variances: modeled as a GJR-GARCH(1,1) for all three indices; correlations: equal to the unconditional, constant sample correlations for all three pairs of indices; mean: constant.

    (c) Variances: modeled as a GJR-GARCH(1,1) for all three indices; correlations: equal to the unconditional, constant sample correlations for all three pairs of indices; mean: assume a model of the type

    $$r^j_{t+1} = \delta^j_0 + \delta^j_1 dp^j_t + \epsilon^j_{t+1} \qquad j = 1, 2, 3.$$

    where $r^j_{t+1}$is the log excess return on country j's stock index, and dp^j_t is the log dividend yield of country j.

    Notice that, just for simplicity (we shall relax this assumption later on), all models assume a constant correlation among different asset classes. Plot optimal weights and the resulting in-sample, realized Sharpe ratios of your optimal portfolio under each of the three different frameworks. What is, in your opinion, the best-performing framework given a risk aversion coefficient $\gamma = 10$ under a utility function of the type

    $$U(\mu_t, \sigma^2_t) = \mu_t - \frac{1}{2\gamma}\sigma^2_t \text{ ?}$$

    [IMPORTANT: Use the toolboxes *regression_tool_1.m* and *mean_variance_multiperiod.m* that have been made available with this exercise set]

12. Compute the Value-at-Risk with a 95% confidence level and the resulting number of violations for the optimal Markowitz portfolio derived under 11.c above, i.e., when both the mean and the variance are predictable. Comment the results, and think about a better model to track VaR. How could the model under 11.c be improved?

Solution

This solution is a commented version of the MATLAB code Ex_GARCH_2012.m posted on the course web site. Please make sure to use a “Save Path” to include *jplv7* among the directories that Matlab reads looking for usable functions. The loading of the data is performed by the lines of code:

```
filename=uigetfile(‘*.txt’);
data=dlmread(filename);
```

The above two lines import only the numbers, not the strings, from a .txt file.[64] The following lines of the codes take care of the strings:

```
filename=uigetfile(‘*.txt’);
fid = fopen(filename);
labels = textscan(fid, ‘%s %s %s %s %s %s %s %s %s %s’);
fclose(fid);
```

1. The plot requires that the data are read in and transformed in euros using appropriate exchange rate log-changes, that need to be computed from the raw data, see the posted code for details on these operations. The following lines proceed to convert Excel serial date numbers into MATLAB serial date numbers (the function *x2mdate(·)*), set the dates to correspond to the beginning and the end of the sample, while the third and final dates are the beginning and the end of the out-of-sample (OOS) period:

```
date=datenum(data(:,1));
date=x2mdate(date);
f=[‘02/01/2006’;‘31/12/2010’; ‘03/01/2013’];
date_find=datenum(f,‘dd/mm/yyyy’);
ind=datefind(date_find,date);
```

The figure is then produced using the following set of instructions, that shall not be commented in detail because the structure of the plot should closely resemble many other plots proposed in the first part of the course.[65]

```
figure(1);
t=ind(1):ind(2);
```

[64] The reason for loading from a .txt file in place of the usual Excel is to favor usage from Mac computers that sometimes have issues with reading directly from Excel, because of copyright issues with shareware spreadsheets.

[65] Those ‘...’ that are featured below represent the way in which you go to a new line in the text editor of a Matlab code without actually breaking the line in the perspective of the compiler.

```
plot(t’, port_ret(ind(1):ind(2),:),‘b’);
title(’Daily Portfolio Returns’,’fontname’,’Garamond’,’fontsize’,14);
set(gca,’fontname’,’garamond’,’fontsize’,12);
set(gca,’xtick’,index-1+ind(1)+5);
set(gca,’xticklabel’,’Jan2006|Jan2007|Jan2008|Jan2009|Jan2010|Jan2011||Dec2011’);
ylabel(’% Returns’);
set(gcf,’color’,’w’);
set(gca,’Box’, ’off’, ’TickDir’, ’out’, ’TickLength’, [.02 .02],’XMinorTick’,
’off’,’YMinorTick’, ... ’off’,’XColor’,[.3 .3 .3],’YColor’,[.3 .3 .3],’LineWidth’, 1,
’FontName’, ’Times’);
```

The resulting plot looks as follows and clearly shows the volatility outburst during the financial crisis of the Summer 2008-Spring 2009, plus some further sovereign crisis jitters during the Summer of 2010:

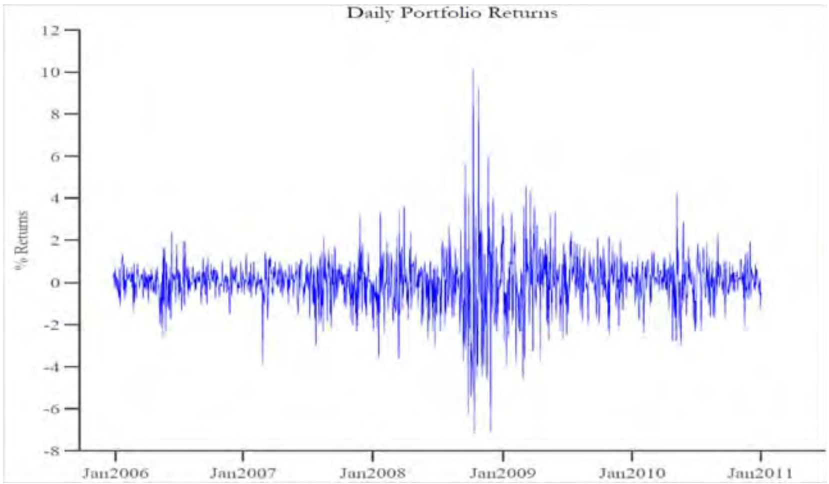


Figure A1:Daily Portfolio Returns

2. As already seen in the first part of the course, the Matlab functions that compute and plot the autocorrelogram functions (for up to 60 lags) for the levels, squares, absolute values of portfolio returns are:

```
sq_port_ret=port_ret.^2; %Squared daily returns
abs_port_ret=abs(port_ret);
figure(2);
subplot(3,1,1)
autocorr(port_ret(ind(1):ind(2),:),60,[],2);
title(‘ACF: Daily Returns’,‘fontname’,‘garamond’,‘fontsize’,16);
set(gcf,‘color’,‘w’);
subplot(3,1,2)
```

```
autocorr(sq_port_ret(ind(1):ind(2),:),60,[],2);
title(‘ACF: Daily Squared Returns’,‘fontname’,‘garamond’,‘fontsize’,16);
set(gcf,‘color’,’w’);
subplot(3,1,3)
autocorr(abs_port_ret(ind(1):ind(2),:),60,[],2);
title(‘ACF: Daily Absolute Returns’,‘fontname’,‘garamond’,‘fontsize’,16);
set(gcf,‘color’,‘w’);
```

The *autocorr(Series,nLags,M,nSTDs)* function computes and plots the sample ACF of a univariate time series already complete with confidence bounds; the input argument *nLags* is a positive scalar integer that indicates the number of lags of the ACF to compute;[66] M is a nonnegative integer scalar indicating the number of lags beyond which the theoretical ACF is effectively 0; *autocorr* assumes the underlying Series is an MA(M) process, and uses Bartlett's approximation to compute the large-lag standard error for lags greater than M;[67] finally, *nSTDs* is a positive scalar indicating the number of standard deviations of the sample ACF estimation error to compute; if *nSTDs* = [] or is unspecified, the default is 2 (that is, approximate 95 percent confidence interval). Note that by using the command *subplot* divides the current figure into rectangular panes that are numbered row-wise; each pane contains an axes object. Subsequent plots are output to the current pane. In particular, *subplot(m,n,p)* breaks the figure window into an $m \times n$ matrix of small axes, selects the pth axes object for the current plot, and returns the axes handle. The axes are counted along the top row of the figure window, then the second row, etc.

The resulting set of 3 plots in figure A2 shows the typical result already commented in section 2 of this chapter: while the level of financial returns hardly features any significant autocorrelations (not even at the shortest lags), other functions $g(R_t)$—such as $g(x) = x^2$ and $g(x) = |x|$—are characterized by many statistically significant, when the Bartlett's standard errors are used to form confidence intervals, values and these values tend to decay rather slowly as the lag order τ increases towards the bound of 60 that we have imposed. This is particularly visible in the case of the absolute value of returns, and this is typical of all the literature. As commented in section 2, this evidence allows us to conclude that our portfolio returns are not independently distributed over time and that there is evidence of conditional heteroskedasticity because large past squared returns forecast

[66] If *nLags* = [] or is unspecified, the default is to compute the ACF at lags 0, 1, 2, ..., T, where $T = min([20, length(Series) - 1])$.

[67] If M = [] or is unspecified, the default is 0, and autocorr assumes that the *Series* is Gaussian white noise. If *Series* is a Gaussian white noise process of length T, the standard error is approximately $1/\sqrt{T}$. M must be less than *nLags*.

subsequently large squared returns.

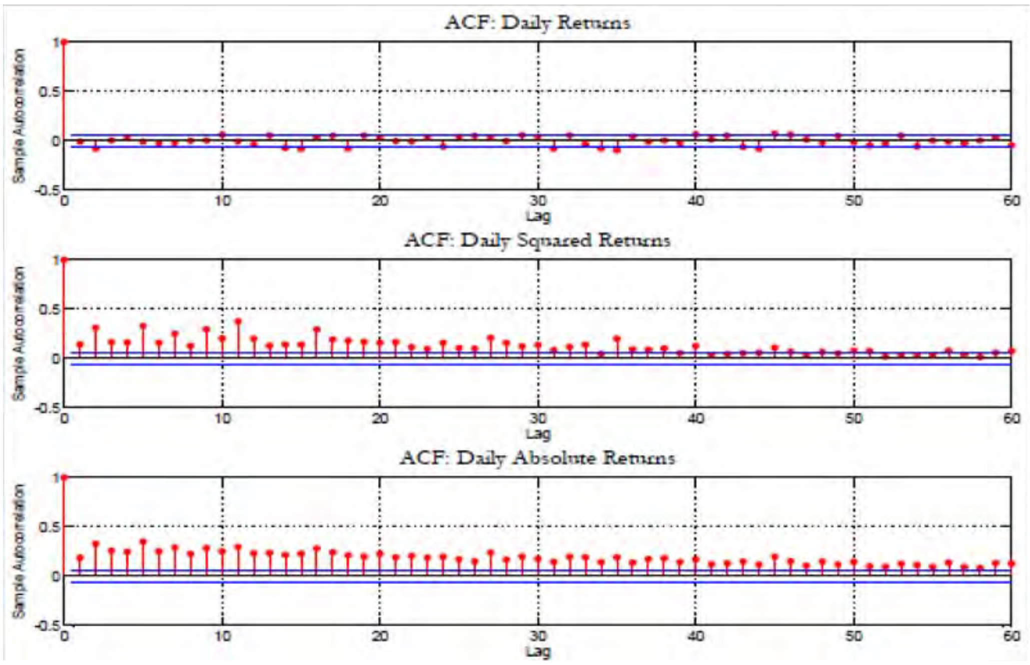


Figure A2:Sample Autocorrelations of a Range of Functions of Portfolio Returns

3. We use the lines of code

histfit(port_ret(ind(1):ind(2),:),100);

to plot a histogram with Gaussian fit that matches the empirical moments, i.e., with the same mean and variance. Note that here the 100 refers to the number of bins for the histogram.[68] The resulting histogram is shown in figure A3. The figure clearly shows that our portfolio returns are highly non-normal. In particular, they are leptokurtic, in the sense that when compared to a Gaussian density, there is excessive probability mass allocated to a neighborhood of the sample mean and to both tails (in particular the left tail), while insufficient probability mass is allocated to intermediate values in the support of the empirical distribution of returns, approximately around ± one empirical standard deviation.

4. We use the Matlab function *garchfit* to estimate a GARCH(1,1) model for daily returns,

spec=garchset('P',1,'Q',1);
[coeff, errors,llf,innovation,sigma,summary]=garchfit(spec,port_ret(ind(1):ind(2),:));
garchdisp(coeff,errors);

[68] *histfit(data,nbins,dist)* would instead plot a histogram with a density from the distribution specified by dist, one of the following strings: 'beta', 'birnbaumsaunders', 'exponential', 'extreme value' or 'ev', 'gamma', 'generalized extreme value' or 'gev', 'generalized pareto' or 'gp', 'inversegaussian', 'logistic', 'loglogistic', 'lognormal', 'nakagami', 'negative binomial' or 'nbin', 'normal' (default), 'poisson', 'rayleigh', 'rician', 'tlocationscale', 'weibull' or 'wbl'. The normal distribution represents the default, used in the absence of other indications.

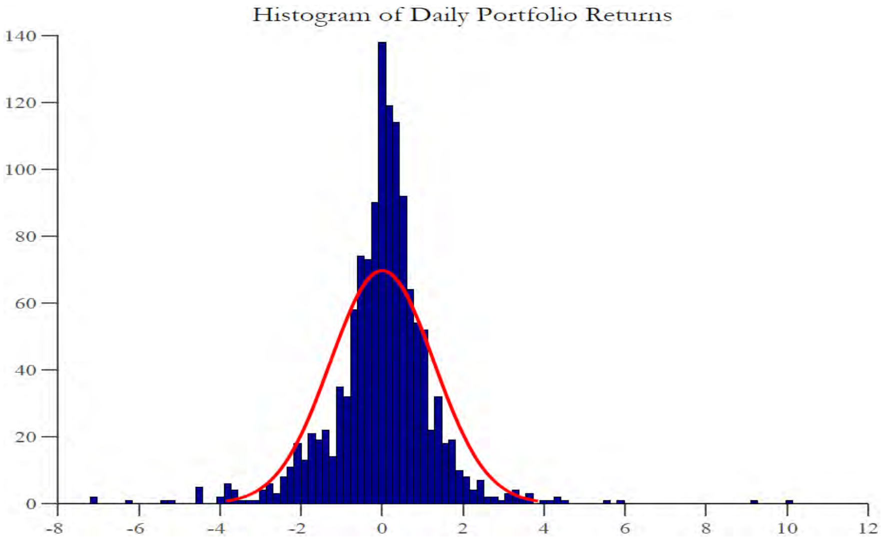


Figure A3:Unconditional distribution (histogram) of daily returns vs. matching Gaussian density

1. while the following computes the vector *sigma* step-by-step:

```
param(1:4,1)=[coeff.C;coeff.K;coeff.GARCH;coeff.ARCH];
init=param(2)/(1-param(3)-param(4));
cond_var_garch=zeros(rows(port_ret(ind(1):ind(2))),1);
cond_var_garch(1)=init;
for i=1:ind(2)-ind(1)
cond_var_garch(i+1)=param(2)+param(3)*cond_var_garch(i)+param(4)*(innovation(i)^2);
end
cond_std_garch=sqrt(cond_var_garch);
```

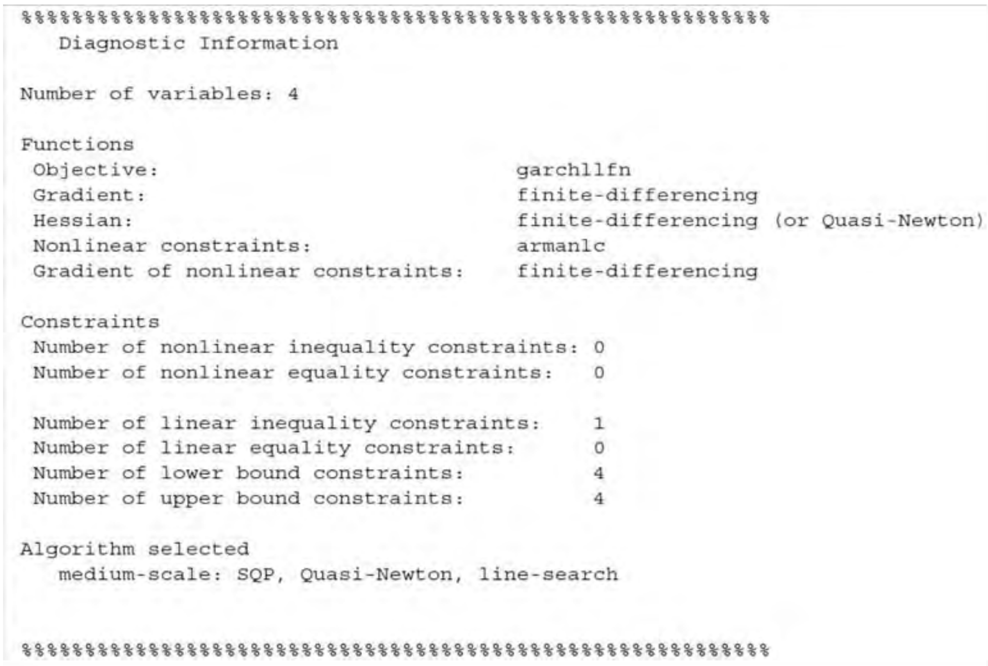
Here, '*init=param(2)/(1-param(3)-param(4));*' initializes the first value of sigma to equal the unconditional variance, which is a necessary choice. *garchset* sets the ARMAX/GARCH model specification parameters; a GARCH specification structure includes these parameters: General Parameters, Conditional Mean Parameters, Conditional Variance Parameters, Equality Constraint Parameters, and Optimization Parameters, even though garchset sets all parameters you do not specify to their respective defaults; among the Conditional Variance Parameters, there is the type of model: 'GARCH', 'EGARCH', 'GJR', or 'Constant'. The default is 'GARCH'.

garchfit estimates the parameters of a conditional mean specification of ARMAX form, and conditional variance specification of GARCH, EGARCH, or GJR form. The estimation process infers the innovations (that is, residuals) from the return series. It then fits the model specification to the return series by *constrained* maximum likelihood.[69] The outputs include a GARCH specification structure containing the estimated coefficients, where *Coeff* is of the same form as the *Spec* input

[69] *garchfit* performs the optimization using the Optimization Toolbox fmincon function. The constraints on the parameters are the ones discussed in Sections 4-6.

structure given by *garchset*; *errors* is a structure containing the estimation errors (that is, the standard errors) of the coefficients with the same form as the *Spec* and *Coeff* structures; *LLF* is the optimized loglikelihood objective function value associated with the parameter estimates found in Coeff; *Innovations* containts the residual time series column vector inferred from the data; *Sigmas* collects the conditional standard deviation vector corresponding to *Innovations*; *Summary* includes '*covMatrix*', the Covariance matrix of the parameter estimates computed using the outer-product method. Finally, *garchdisp* displays ARMAX/GARCH model parameters and statistics. The tabular display includes parameter estimates, standard errors, and t-statistics for each parameter in the conditional mean and variance models.

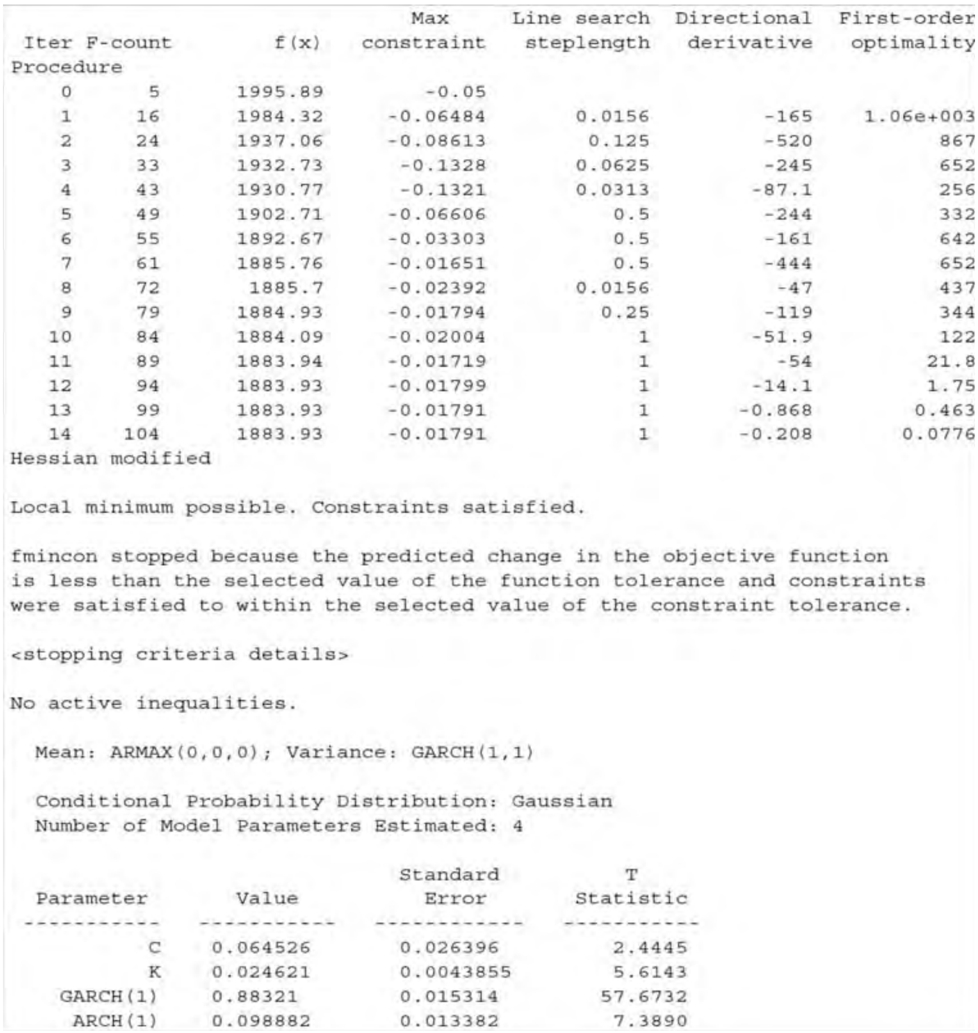
Matlab prints at the screen the following information concerning estimation (we select the information to be printed to save space):



```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
   Diagnostic Information

Number of variables: 4

Functions
 Objective:                            garchllfn
 Gradient:                             finite-differencing
 Hessian:                              finite-differencing (or Quasi-Newton)
 Nonlinear constraints:                armanlc
 Gradient of nonlinear constraints:    finite-differencing

Constraints
 Number of nonlinear inequality constraints: 0
 Number of nonlinear equality constraints:   0

 Number of linear inequality constraints:    1
 Number of linear equality constraints:      0
 Number of lower bound constraints:          4
 Number of upper bound constraints:          4

Algorithm selected
   medium-scale: SQP, Quasi-Newton, line-search


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

The first panel gives a number of technical information on the numerical optimization that Matlab has performed. Although not directly useful, by clicking this information when possible, you will get to know more about what Matlab is doing in the background of the numerical optimization it is performing for you. The second panel details the 14 iterative steps followed by Matlab to reach the optimum and how the value function $f(x) = \mathcal{L}(\boldsymbol{\theta})$—in our case it is the log-likelihood function—changes across different iterations. Notice that $f(x)$ obviously *declines* across iterations. This is due to the fact that Matlab actually minimizes the opposite (i.e., $\times -1$) of the log-likelihood function, because

$$\hat{\boldsymbol{\theta}}^{ML} \equiv \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}}(-\mathcal{L}(\boldsymbol{\theta})).$$



```
                                   Max     Line search  Directional  First-order
 Iter F-count          f(x)    constraint   steplength   derivative   optimality
Procedure
    0       5       1995.89         -0.05
    1      16       1984.32      -0.06484       0.0156         -165    1.06e+003
    2      24       1937.06      -0.08613        0.125         -520          867
    3      33       1932.73       -0.1328       0.0625         -245          652
    4      43       1930.77       -0.1321       0.0313        -87.1          256
    5      49       1902.71      -0.06606          0.5         -244          332
    6      55       1892.67      -0.03303          0.5         -161          642
    7      61       1885.76      -0.01651          0.5         -444          652
    8      72        1885.7      -0.02392       0.0156          -47          437
    9      79       1884.93      -0.01794         0.25         -119          344
   10      84       1884.09      -0.02004            1        -51.9          122
   11      89       1883.94      -0.01719            1          -54         21.8
   12      94       1883.93      -0.01799            1        -14.1         1.75
   13      99       1883.93      -0.01791            1       -0.868        0.463
   14     104       1883.93      -0.01791            1       -0.208       0.0776
Hessian modified

Local minimum possible. Constraints satisfied.

fmincon stopped because the predicted change in the objective function
is less than the selected value of the function tolerance and constraints
were satisfied to within the selected value of the constraint tolerance.

<stopping criteria details>

No active inequalities.

  Mean: ARMAX(0,0,0); Variance: GARCH(1,1)

  Conditional Probability Distribution: Gaussian
  Number of Model Parameters Estimated: 4

                                Standard          T
  Parameter       Value           Error       Statistic
 -----------   -----------   ------------   -----------
           C    0.064526       0.026396          2.4445
           K    0.024621       0.0043855         5.6143
    GARCH(1)    0.88321        0.015314         57.6732
     ARCH(1)    0.098882       0.013382          7.3890
```

Visibly, after the 12th iteration, $-\mathcal{L}(\boldsymbol{\theta})$ stabilizes to 1883.93 (i.e., $\mathcal{L}(\boldsymbol{\theta})$ stabilizes to -1883.93) and this represents the optimum, in the sense that the objective function seems to have converged to a stationary point (as signalled by "fmincon stopped because the predicted change in the objective function is less than the selected value of the function tolerance and constraints were satisfied to within the selected value of the constraint tolerance."), even though Matlab warns you that "Local minimum possible. Constraints satisfied." In the case of parameter estimates, *garchfit* yields point estimates ("Value"), the corresponding standard error, and the t-ratio. Obviously, $t_{\hat{\theta}_k} \equiv \hat{\theta}_k/\sqrt{Var[\hat{\theta}_k]}$ so that once you know the standard error, you could have derived $t_{\hat{\theta}_k}$ yourself; for instance, $2.4445 \simeq 0.064526/0.026396$. Note that given a non-zero mean model

$$\begin{aligned} R_{t+1} &= \mu + \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } \mathcal{N}(0,1) \\ \sigma_t^2 &= \omega + \alpha(R_{t-1} - \mu)^2 + \beta\sigma_{t-1}^2, \end{aligned}$$

Matlab calls C the parameter μ and K the parameter ω of the GARCH, i.e., $R_{t+1} = C + \sigma_{t+1} z_{t+1}$ and $\sigma_t^2 = K + \alpha(R_{t-1} - C)^2 + \beta\sigma_{t-1}^2$. The estimated GARCH model is clearly stationary as $\hat{\alpha} + \hat{\beta} \simeq 0.9821 < 1$ and it implies a long-run, unconditional variance $\bar{\sigma}^2 = 0.02462/(1 - 0.9821)$

$\simeq 1.3754$, which implies a standard deviation of $\sqrt{1.3754}$ of 1.1728 percent per day.[70] The resulting plots (plural because we have re-done calculations manually but also used the *sigma* series that *garchfit* yields) of the (in-sample) forecasts of variance, also called filtered variances, from the estimated GARCH,

$$\hat{\sigma}_t^2 = 0.0246 + 0.0989 R_{t-1}^2 + 0.8832 \sigma_{t-1}^2$$

are shown in Figure A4.

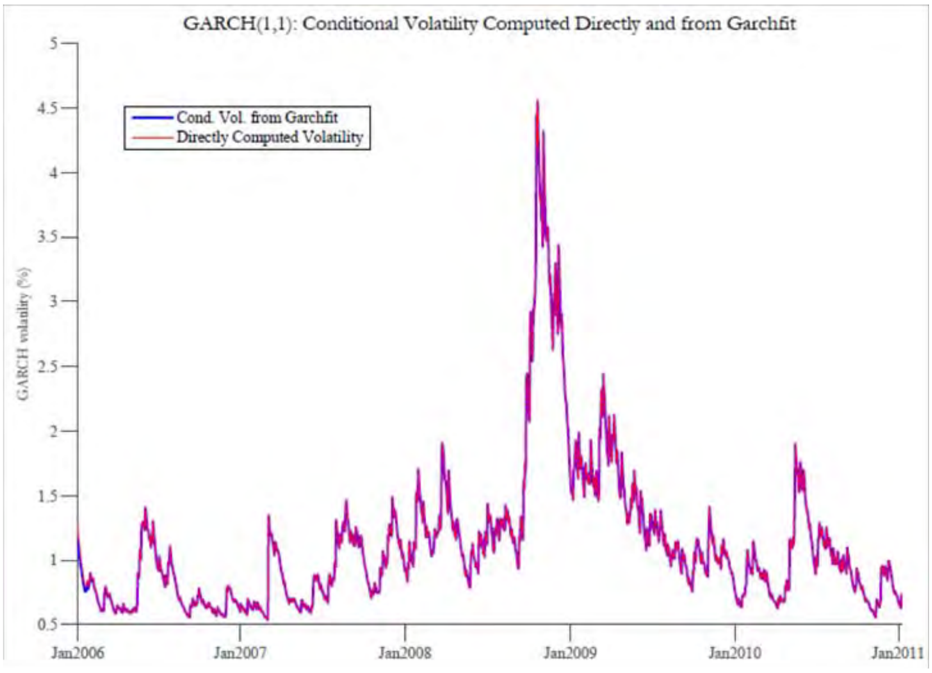


Figure A4:(In-sample predictions) of conditional volatility of daily returns

The plot shows that (apart from the very few initial observations, say until the end of Jan. 2006), manual and automatic calculations give identical results.[71] Not surprisingly, the daily conditional volatility spikes up during the early Fall of 2008, after Lehmann's demise, and remains elevated until March 2009; a further spike, although it less pronounced, occurs in the Summer 2010 and is

[70] A quick (but not necessarily correct, because what follows assumes that variance follow a unit root process, which is clearly not the case under a stationary GARCH(1,1)) back-of-the-envelope calculations transforms that into an *annualized* volatility of approximately $\sqrt{252} \times 1.1728 \simeq 18.62$ percent, which is what you would expect of an equity portfolio mostly sampled during the financial crisis.

[71] The difference is due to the fact that we have manually initialized the loop that computes in-sample forecasts of GARCH variances from the unconditional variance $\bar{\sigma}^2 = 0.02462/\ (1-0.9821) \simeq 1.3754$ that depends on the estimates of K, α, and β. On the opposite, if you provide no explicit pre-sample data (here it is simply the intial return R_0, which is difficult to sample because the time scale is $t = 1, ..., T$), Matlab derives the necessary pre-sample observations using time series techniques: the idea is to generate output with reference to an (approximate) steady state by attempting to eliminate transients in the data. Matlab first estimates the number of observations needed for the transients to decay to some arbitrarily small value, subject to a 10000-observation maximum; it then generates a number of observations equal to the sum of this estimated value and the number of observations you request to be simulated or filtered (here they are just equal to T). Matlab then ignores the earlier estimated number of initial observations needed for the transients to decay sufficiently, and returns only the requested number of later observations. To do this, garchsim interprets a GARCH(q, p) conditional variance process as an ARMA$(\max(q, p),\ p)$ model for the squared innovations. Further and increasingly tedious details on this algorithm can be found in Matlab's help.

probably related to the initial PIIGS sovereign debt crisis jitters. If you had any doubts volatility is actually time-varying, this GARCH model shows that given a long-run average level that we know to be at just above 1% per day, volatility rather often doubles up to almost touch 2% per day, while spikes in excess of 3% per day may occur.

5. Here we proceed to estimate a RiskMetrics exponential smoother (i.e., estimate the RiskMetrics parameter λ) by ML. Note that this is different from the simple approach mentioned in the lectures where λ was fixed at the level suggested by RiskMetrics.

```
parm=0.1;
logL= maxlik(‘objfunction’,parm,[],port_ret(ind(1):ind(2)+1));
lambda=logL.b;
disp(‘The estimated RiskMetrics smoothing coefficient is:’)
disp(lambda)
```

parm=0.1 sets an initial condition for the estimation (a weird one, indeed, but the point is to show that in this case the data have such a strong opinion for what is the appropriate level of λ that such an initial condition hardly matters; try to change it and see what happens). This *maxlik* call is based on the maximization of the log-likelihood given in *objfunction.* That procedure reads as

```
ret=y;
R=rows(ret);
C=cols(ret);
conditional_var=NaN(R,C);
conditional_var(1,1)=var(ret);
for i=2:R
conditional_var(i,1)=(1-lambda)*ret(i-1,1).^2+lambda*conditional_var(i-1,1);
end
z=ret./sqrt(conditional_var);
y=-sum(-0.5*log(2*pi)-0.5*log(conditional_var)-0.5*(z.^2));
```

In figure A5 we plot the fitted (also called in-sample filtered) conditional volatility series and compare it to that obtained from the GARCH(1,1) in the earlier question. Clearly, the two models behave rather differently and such divergencies were substantial during the financial crisis. This may have mattered to financial institutions and their volatility traders and risk managers.

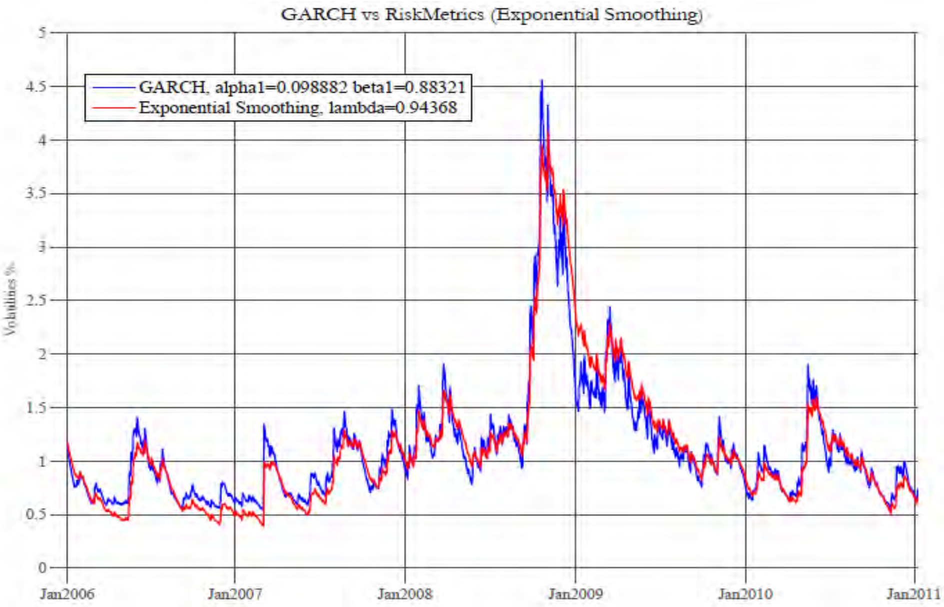


Figure A5:Comparing in-sample predictions of conditional volatility from GARCH vs. RiskMetrics

6. Using the following lines of code, we compute and plot daily *one-day ahead*, recursive out-of-sample forecasts for the period 01/01/2011-01/01/2013 given the ML estimates for the parameters of the models in questions 4,

```
spec_pred=garchset(‘C’,coeff.C,‘K’,coeff.K,‘ARCH’,coeff.ARCH,‘GARCH’,coeff.GARCH);
garch_pred=NaN(ind(3)-ind(2),1);
for i=1:(ind(3)-ind(2))
[SigmaForecast,MeanForecast,SigmaTotal,MeanRMSE] = ...
garchpred(spec_pred,port_ret(ind(1):ind(2)+i-1),1);
garch_pred(i)=SigmaForecast(1);
end
```

and 5, using

```
for i=1:(ind(3)-ind(2)-1)
es_pred(i+1)=lambda*es_pred(i)+(1-lambda)*port_ret(ind(2)+i)ˆ2;
end
es_std_pred=sqrt(es_pred);
```

Here *garchpred* forecasts the conditional mean of the univariate return series and the standard deviation of the innovations *ind(3)-ind(2)* into the future, a positive scalar integer representing the forecast horizon of interest. It uses specifications for the conditional mean and variance of an observed univariate return series as input. In both cases, note that actual returns realized between 2011 and early 2013 is fed into the models, in the form of series $\{(R_{t-1} - C)^2\}$ sampled over time.

Figure A6 shows the results of this *recursive* prediction exercises and emphasizes once more the existence of some difference across GARCH and RiskMetrics during the Summer 2011 sovereign debt crisis.

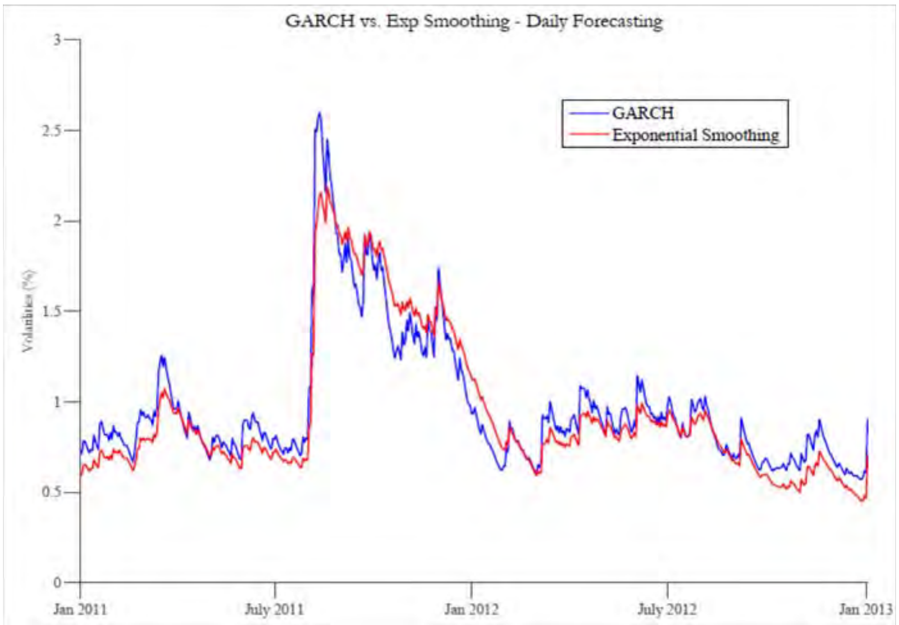


Figure A6:Comparing out–of-sample predictions of conditional volatility from GARCH vs. RiskMetrics

7. To better realize what the differences among GARCH(1,1) and RiskMetrics are when it comes to forecast variances in the long term, we proceed to a 300-day long *simulation* exercise for four alternative GARCH(1,1) models, when the parameters are set by us instead of being estimated: (i) $\omega = 1$, $\alpha = 0.75$, $\beta = 0.2$; (ii) $\omega = 1$, $\alpha = 0.2$, $\beta = 0.75$; (iii) with $\omega = 2$, $\alpha = 0.75$, $\beta = 0.2$; (iv) with $\omega = 2$, $\alpha = 0.2$, $\beta = 0.75$. Importantly, forecasts under RiskMetrics are performed using a value of λ that makes it consistent with the first variance forecast from GARCH. For all parameterizations, this is done by the following lines of code:

```
for j=1:length(alpha)
for i=2:dim
epsilon=sqrt(garch(i-1,j))*ut(i);
garch(i,j)=omega(1)+alpha(j)*epsilon^2+beta(j)*garch(i-1,j);
end
end
for j=3:length(alpha)+length(omega)
for i=2:dim
epsilon=sqrt(garch(i-1,j))*ut(i);
garch(i,j)=omega(2)+alpha(j-2)*epsilon^2+beta(j-2)*garch(i-1,j);
end
end
```

Figure A7 presents simulation results. Clearly, the blue models imply generally low variance but frequent and large spikes, while the green models imply considerably more conditional persistence of past variance, but a smoother temporal path. Try and meditate on these two plots in relation to the meaning of your MLE optimization setting the “best possible” values of α and β to fit the data.

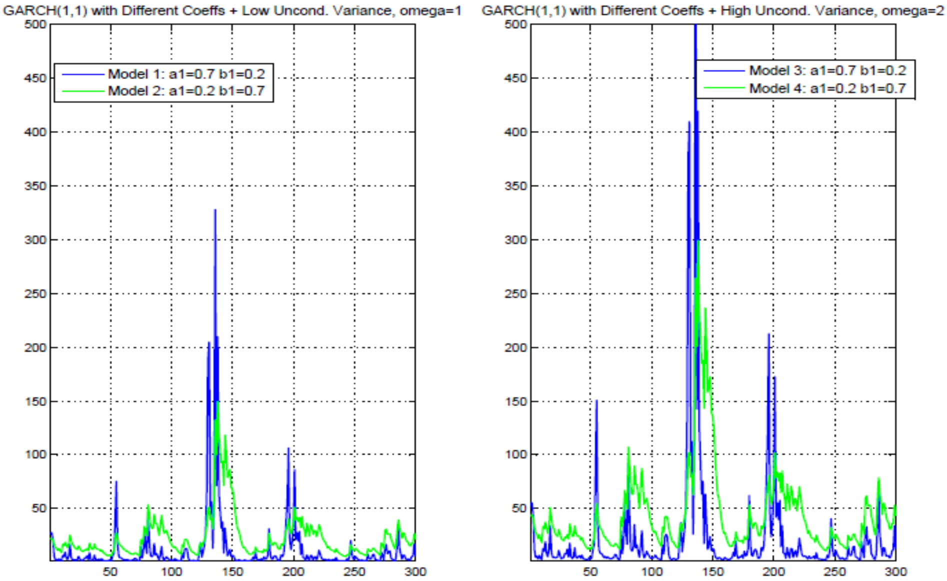


Figure A7: Simulating 4 alternative GARCH models

The following code computes insteads true out-of-sample forecasts 250 periods ahead. Notice that these forecasts are no long recursive, i.e., you do not feed the actual returns realized over the out-of-sample periods, and this occurs for a trivial reason: you do not know them because this is a truly out-of-sample exercise. Initialization is done with reference to the last shock obtained in the previous run of simulations:

```
horz=250;
A=NaN(horz,1);
garch_sigma_sq_t_plus_one_a=omega(1)+alpha(1)*epsilon^2+beta(1)*garch(end,1);
garch_sigma_sq_t_plus_one_b=omega(1)+alpha(2)*epsilon^2+beta(2)*garch(end,2);
(%Derives forecasts under Model 1)
A(1)=garch_sigma_sq_t_plus_one_a;
uncond_var=omega(1)/(1-alpha(1)-beta(1));
for i=2:horz
A(i)=uncond_var+((alpha(1)+beta(1))^(i-1))*(garch_sigma_sq_t_plus_one_a-uncond_var);
end
garch_forecast_a=sqrt(A);
lambda_a=(garch_sigma_sq_t_plus_one_a-epsilon^2)/(garch(end,1)-epsilon^2);
es_forecast_a=lambda*garch_forecast_a(1)+(1-lambda)*epsilon^2;
```

```
es_forecast_a=sqrt(es_forecast_a).*ones(horz,1);
```

Here the initial value for the variance in the GARCH model is set to be equal to the unconditional variance. The expression for *lambda_a* sets a value for λ that makes it consistent with the first variance forecast from GARCH. Figure A8 plots the forecasts between 1- and 250-periods ahead obtained under models (i) and (ii) when the RiskMetrics λ is set in the way explained above. As commented in the lectures, it is clear that while GARCH forecasts converge in the long-run to a steady, unconditional variance value that by construction is common and equal to 4.5 in both cases, RiskMetrics implies that the forecast is equal to the most recent variance estimate for all horizons $H \geq 1$.

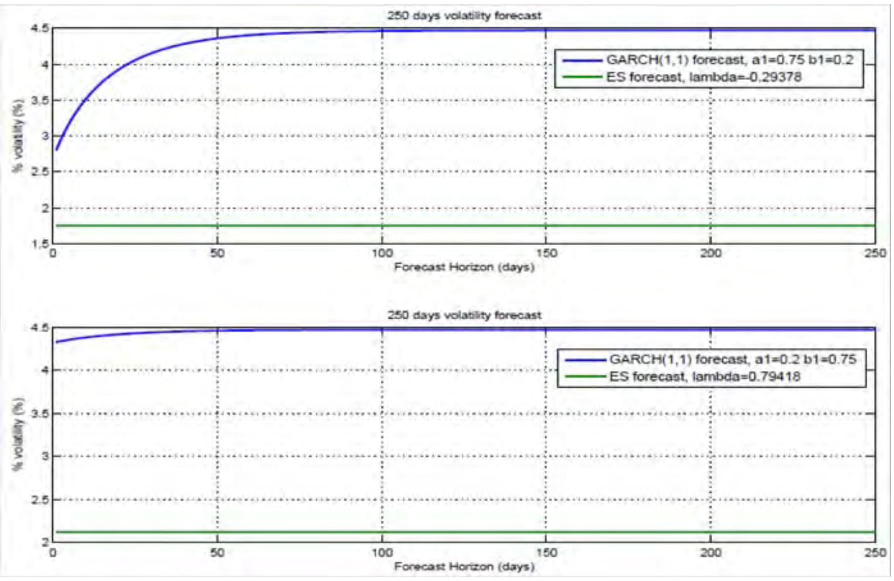


Figure A8: Variance forecasts from two alternative GARCH models vs. RiskMetrics

8. We now estimate the 1% Value-at-Risk under the alternative GARCH(1,1) and RiskMetrics models with reference to the OOS period 01/01/2011-31/01/2013, given the ML estimates for the parameters of the models in questions 4 and 5. This is accomplished through the following lines of code:

```
alpha=0.01;
Var_garch=norminv(alpha,0,garch_pred);
Var_es=norminv(alpha,0,es_std_pred);
index_garch=(port_ret(ind(2)+1:ind(3))<Var_garch);
viol_garch=sum(index_garch);
index_es=(port_ret(ind(2)+1:ind(3))<Var_es);
viol_es=sum(index_es);
```

Figure A9 shows the results: because during parts of the Summer 2011 crisis, the RiskMetrics one-step ahead variance forecast was below the GARCH(1,1), there are more violations of the 1%

VaR bound under the former model than under the second, 11 and 8, respectively.[72] Also note that if a volatility model is correctly specified, then we should find that in a recursive back testing period of 524 days (which is the number of trading days between Jan. 1, 2011 and Jan. 31, 2013), one ought to approximately observe $0.01 \times 524 =$ roughly 5 violations. Here we have instead 8 and 11, and especially the latter number represents more than the double than the total number one expects to see. This is an indication of misspecification of RiskMetrics and probably of the GARCH model too. Even worse, most violations do occur in early August 2011, exactly when you would have needed a more accurate forecasts of risk and hence of the needed capital reserves! However, RiskMetrics also features occasional violations of the VaR bound in the Summer of 2012.

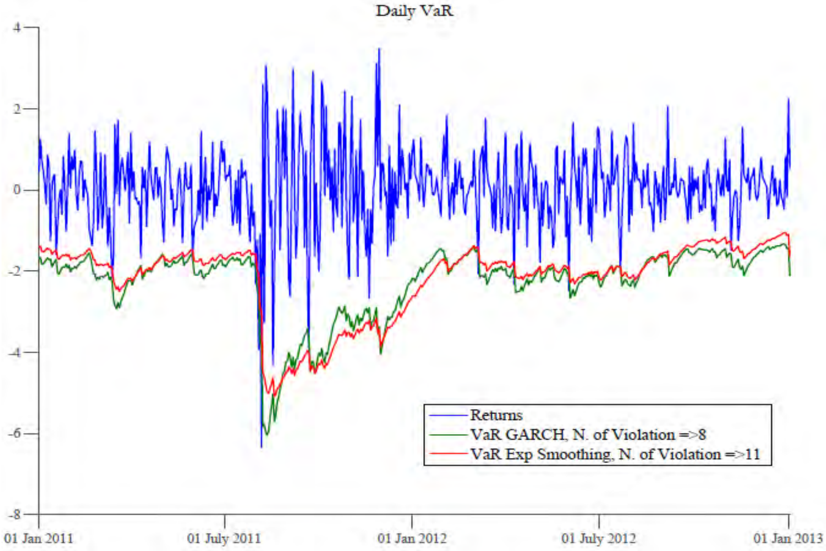


Figure A9: Daily 1% VaR bounds from GARCH vs. RiskMetrics

9. Next, we proceed to estimate three “more advanced” and asymmetric GARCH models: NGARCH (1,1), GJR-GARCH(1,1), and EGARCH(1,1). While for GJR and EGARCH estimation proceeds again using the Matlab *garchfit* toolbox in the same way we have seen above, the GJR(1,1) (also called threshold GARCH) model is estimated by MLE, using

```
GJRspec=garchset(‘VarianceModel’,‘GJR’,‘Distribution’,‘Gaussian’,‘P’,1,‘Q’,1);
[GJRcoeff, GJRerrors,GJRllf,GJRinnovation,GJRsigma,GJRsummary]=...
garchfit(GJRspec,port_ret(ind(1):ind(2),:));
garchdisp(GJRcoeff,GJRerrors);
EGARCHspec=garchset(‘VarianceModel’,‘EGARCH’,‘Distribution’,‘Gaussian’,‘P’,1,‘Q’,1);
[EGARCHcoeff, EGARCHerrors,EGARCHllf,EGARCHinnovation,EGARCHsigma,EGARCHsummary]=...
garchfit(EGARCHspec,port_ret(ind(1):ind(2),:));
garchdisp(EGARCHcoeff,EGARCHerrors);
```

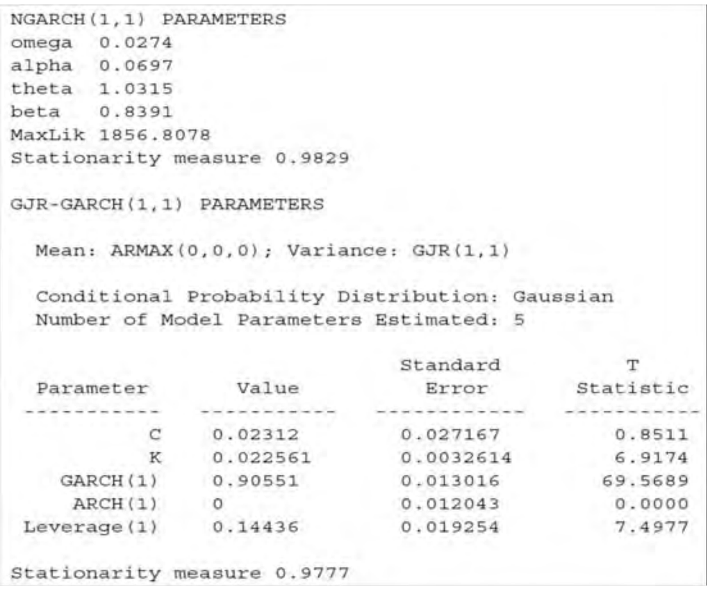
In the case of the NGARCH model, estimation is not implemented through *garchfit* and as a result you will have to develop and write the log-likelihood function in one appropriate procedure,

[72] These are easily computed simply using *sum(viol_garch)* and *sum(viol_es)* in Matlab.

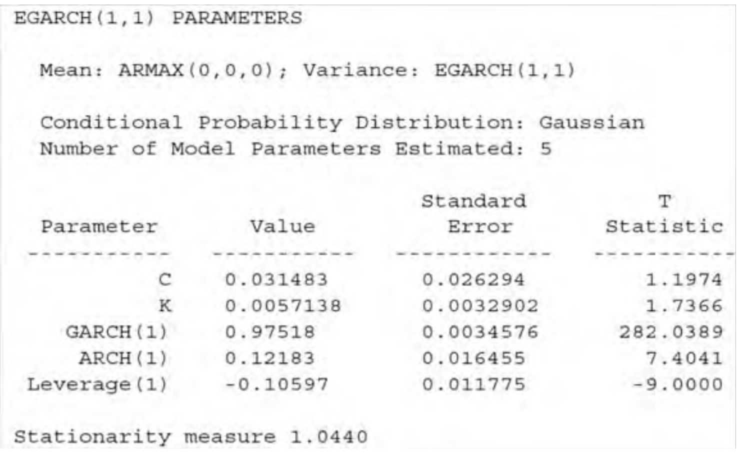
which is the appropriate function *ngarch*, initialized at *par_initial(1:4,1)=[0.05;0.1;0.05;0.85]*. This procedure uses Matlab unconstrained optimization *fminsearch* (please press F1 over *fminsearch* and read up on what this is):[73]

par_initial(1:4,1)=[0.05;0.1;0.05;0.85];
function [sumloglik,z,cond_var] = ngarch(par,y);
[mle,z_ng,cond_var_ng]=ngarch(param_ng,port_ret(ind(1):ind(2),:));

ngarch takes as an input the 4x1 vector of NGARCH parameters (ω, α, β, and θ) and the vector y of returns and yields as an output *sumloglik*, the (scalar) value of likelihood function (under a normal distribution), the vector of standardized returns *z*, and the conditional variance (note) *cond_var*. The various points requested by the exercise have been printed directly on the screen:



```
NGARCH(1,1) PARAMETERS
omega  0.0274
alpha  0.0697
theta  1.0315
beta   0.8391
MaxLik 1856.8078
Stationarity measure 0.9829

GJR-GARCH(1,1) PARAMETERS

  Mean: ARMAX(0,0,0); Variance: GJR(1,1)

  Conditional Probability Distribution: Gaussian
  Number of Model Parameters Estimated: 5

                                 Standard          T
  Parameter         Value         Error        Statistic
 -----------     -----------   ------------   -----------
          C      0.02312       0.027167          0.8511
          K      0.022561      0.0032614         6.9174
    GARCH(1)     0.90551       0.013016         69.5689
     ARCH(1)     0             0.012043          0.0000
 Leverage(1)     0.14436       0.019254          7.4977

Stationarity measure 0.9777
```



```
EGARCH(1,1) PARAMETERS

  Mean: ARMAX(0,0,0); Variance: EGARCH(1,1)

  Conditional Probability Distribution: Gaussian
  Number of Model Parameters Estimated: 5

                                 Standard          T
  Parameter         Value         Error        Statistic
 -----------     -----------   ------------   -----------
          C      0.031483      0.026294          1.1974
          K      0.0057138     0.0032902         1.7366
    GARCH(1)     0.97518       0.0034576       282.0389
     ARCH(1)     0.12183       0.016455          7.4041
 Leverage(1)     -0.10597      0.011775         -9.0000

Stationarity measure 1.0440
```

All volatility models imply a starionarity index of approximately 0.98, which is indeed typical of daily data. The asymmetry index θ is large (but note that we have not yet derived standard errors, which would not be trivial in this case) at 1.03 in the NAGARCH case, it is 0.14 with a t-stat of

[73] *fminsearch* finds the minimum of an unconstrained multi-variable function using derivative-free methods and starting at a user-provided initial estimate.

7.5 in the GJR case, and it is -0.11 with a t-stat 9 in the EGARCH case: therefore in all cases we know or we can easily presume that the evidence of asymmetries in these portfolio returns is strong. Figure A10 plots the dynamics of volatility over the estimation sample implied by the three alternative volatility models. As you can see, the dynamics of volatility models tends to be rather homogeneous, apart from the Fall of 2008 when NAGARCH tends to be above the others while simple GJR GARCH is instead below. At this stage, we have not computed VaR measures, but you can easily figure out (say, under a simple Gaussian VaR such as the one presented in chapter 1) what these different forecasts would imply in risk management applications.

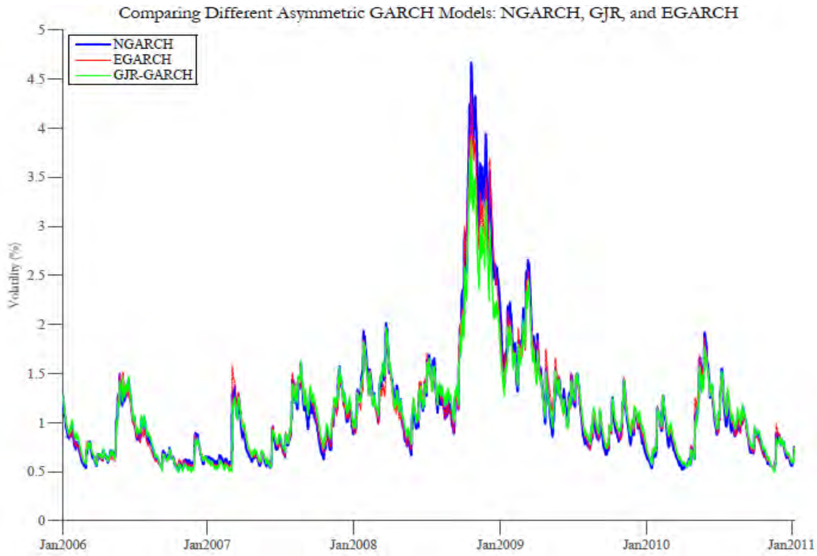


Figure A10: Comparing in-sample fitted volatility dynamics under GJR, EGARCH, and NAGARCH

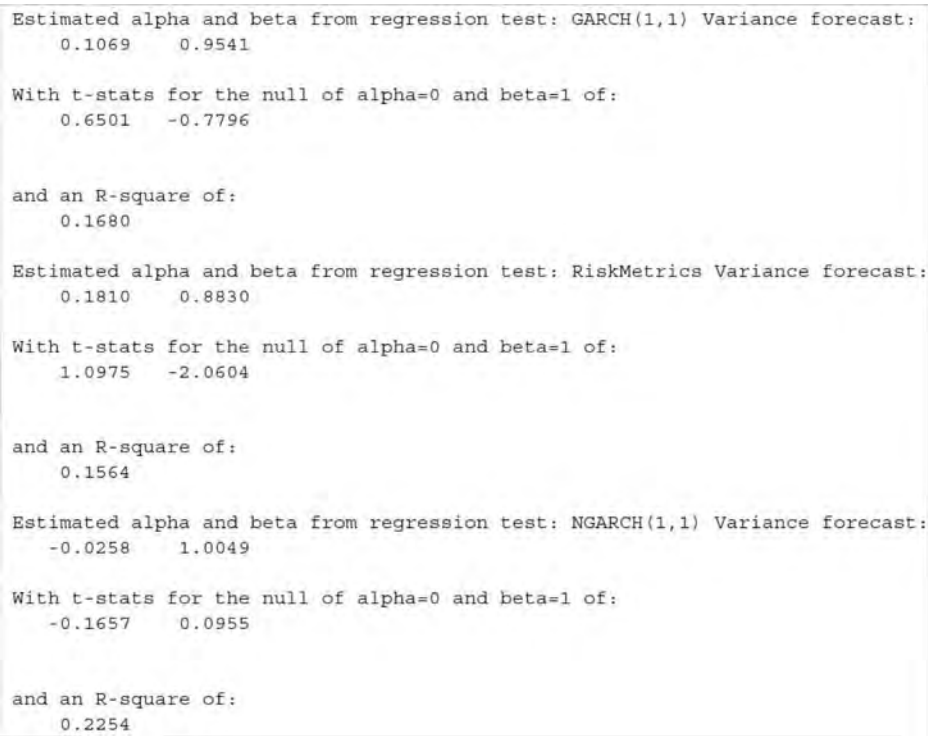
10. We now compare the accuracy of the forecasts given by different volatility models. We use the fitted/in-sample filtered variances from GARCH(1,1), RiskMetrics' exponential smoother, and a GJR-GARCH(1,1) to perform the out-of-sample test that is based on the classical test that in the regression

$$R_t^2 = \alpha + \beta \widehat{\sigma}_{t,t-1}^{2,m} + \epsilon_t^m$$

$\alpha = 0$ and $\beta = 1$ to imply that $E_{t-1}[R_t^2] = \sigma_t^2 = \widehat{\sigma}_{t,t-1}^{2,m}$, where $\widehat{\sigma}_{t,t-1}^{2,m}$ is the the time $t-1$ conditional forecast of the variance from model m. For instance, in the case of GARCH, the lines of codes estimating such a regression and printing the relevant outputs are:

```
result = ols((port_ret(ind(1):ind(2),:).^2),[ones(ind(2)-ind(1)+1,1) (cond_var_garch)]);
disp('Estimated alpha and beta from regression test: GARCH(1,1) Variance forecast:');
disp(result.beta');
disp('With t-stats for the null of alpha=0 and beta=1 of:');
disp([result.tstat(1) ((result.beta(2)-1)/result.bstd(2))]); fprintf('\n');
disp('and an R-square of:');
disp(result.rsqr)
```

The regression is estimated using the Matlab function *ols* that you are invited to review from your first course in the Econometrics sequence. The results displayed on your screen are:

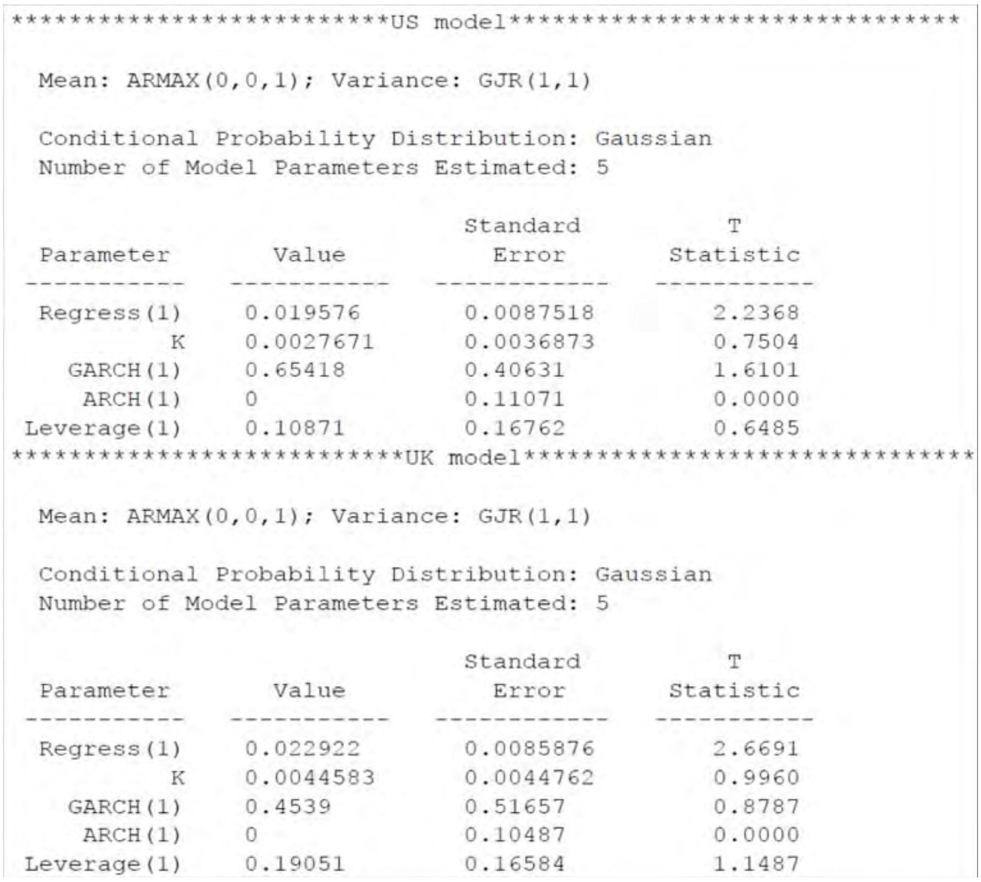


```
Estimated alpha and beta from regression test: GARCH(1,1) Variance forecast:
    0.1069    0.9541

With t-stats for the null of alpha=0 and beta=1 of:
    0.6501   -0.7796


and an R-square of:
    0.1680

Estimated alpha and beta from regression test: RiskMetrics Variance forecast:
    0.1810    0.8830

With t-stats for the null of alpha=0 and beta=1 of:
    1.0975   -2.0604


and an R-square of:
    0.1564

Estimated alpha and beta from regression test: NGARCH(1,1) Variance forecast:
   -0.0258    1.0049

With t-stats for the null of alpha=0 and beta=1 of:
   -0.1657    0.0955


and an R-square of:
    0.2254
```

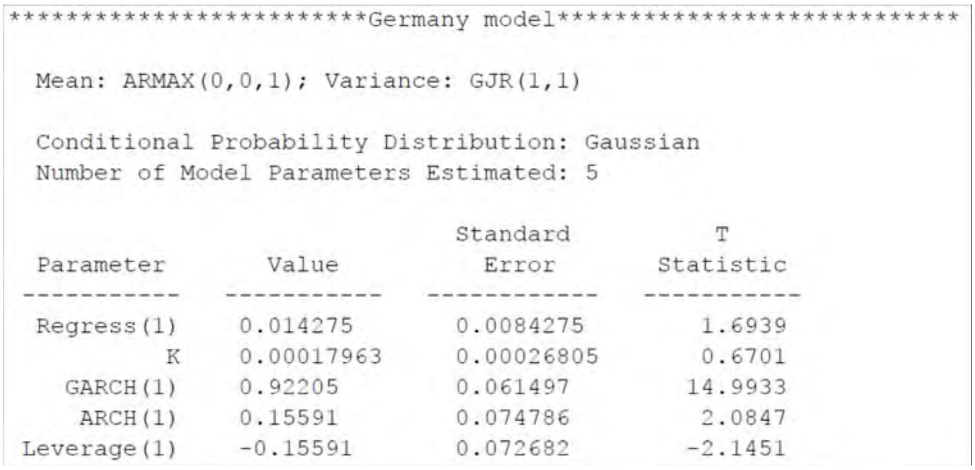
In a way, the winner is the NAGARCH(1,1) model: the null of $\alpha = 0$ and $\beta = 1$ cannot be rejected and the R^2, considering that we are using noisy, daily data is an interesting 22.5%; also GARCH gives good results, in the sense that $\alpha = 0$ and $\beta = 1$ but the R^2 is "only" 17%. Not good news instead for RiskMetrics, because the null of $\beta = 1$ can be rejected: $\hat{\beta} = 0.88 < 1$ implies a t-stat of -2.06 (=(0.88-1)/std.err($\hat{\beta}$)). Note that these comments assume that the proxy for observed variances are squared returns, which—as seen in the lectures—may be a questionable choice.

11. We now perform some Markowitz asset allocation back-testing workout: we start from March 31 1976 and until the end of the available data, we compute optimal weights based on predicted mean returns and variances of the three risky indices at *quarterly frequency.* We emphasize this recourse to quarterly data for two reasons. First, this represents a rejoinder with the work that you have performed in earlier chapters, when low frequency time series had been used. Second, you will note that GARCH models will not work perfectly in this example: this is due to the fact that–as emphasized during the lectures—conditional heteroskedasticity is the dominant phenomenon at relatively or very high frequencies, such as daily or weekly (possibly also monthly, that depends a lot on the specific data). In chapters that will follow we shall specialize instead on monthly and daily data and see that in that case GARCH models perform much better. Note that this exercise requires you to re-load new, quarterly data and to apply new exchange rate transformations, which is done at the very beginning of the portion of code.

In the case of the standard Gaussian IID model in which means, variances, and covariances are all constant, the estimates are obtained with *regression_tool_1* which performs recursive estimation.[74] In the case in which the means are constant but the individual variances follow a GJR-GARCH(1,1) and correlations are equal to the unconditional, constant sample correlations for all three pairs of indices, the estimates displayed on the Matlab screen are:[75]



```
**************************US model*******************************

 Mean: ARMAX(0,0,1); Variance: GJR(1,1)

 Conditional Probability Distribution: Gaussian
 Number of Model Parameters Estimated: 5

                                Standard          T
 Parameter        Value           Error       Statistic
-----------    -----------    ------------   -----------
 Regress(1)     0.019576       0.0087518        2.2368
          K     0.0027671      0.0036873        0.7504
   GARCH(1)     0.65418        0.40631          1.6101
    ARCH(1)     0              0.11071          0.0000
Leverage(1)     0.10871        0.16762          0.6485
**************************UK model********************************

 Mean: ARMAX(0,0,1); Variance: GJR(1,1)

 Conditional Probability Distribution: Gaussian
 Number of Model Parameters Estimated: 5

                                Standard          T
 Parameter        Value           Error       Statistic
-----------    -----------    ------------   -----------
 Regress(1)     0.022922       0.0085876        2.6691
          K     0.0044583      0.0044762        0.9960
   GARCH(1)     0.4539         0.51657          0.8787
    ARCH(1)     0              0.10487          0.0000
Leverage(1)     0.19051        0.16584          1.1487
```



```
************************Germany model****************************

 Mean: ARMAX(0,0,1); Variance: GJR(1,1)

 Conditional Probability Distribution: Gaussian
 Number of Model Parameters Estimated: 5

                               Standard          T
 Parameter       Value           Error       Statistic
-----------   -----------    ------------   -----------
 Regress(1)    0.014275       0.0084275        1.6939
          K    0.00017963     0.00026805       0.6701
   GARCH(1)    0.92205        0.061497        14.9933
    ARCH(1)    0.15591        0.074786         2.0847
Leverage(1)    -0.15591       0.072682        -2.1451
```

[74] The unconditional correlations are 0.73 between US and UK returns, 0.64 between US and German returns, and 0.60 between UK and German returns. You are also invited to inspect the structure and capabilities of *regression_tool_1*, which is provided for your use.

[75] Such unconditional, constant correlations are 0.73 between US and UK returns, 0.60 between US and German returns, and 0.57 between UK and German returns. As we shall see in due time, a multivariate model in which the conditional variances follow a GARCH process but correlations are assumed to be constant over time is called a Constant Conditional Correlation model, CCC.

> Here please note that the user-provided toolbox *regression_tool_1* prints the constant unconditional mean parameter (previously called C) as *Regress(1)* because it is well known that the estimate of a constant in a regression model can be obtained from a regressor that is the unit vector of ones, as in this case. The estimates displayed are the ones corresponding to last quarter in the sample, September 2012. Interestingly, there is little evidence of GARCH and no evidence of an asymmetric in quarterly US and UK data (see comments made in the lecture slides); there is instead strong evidence of GARCH as well as of asymmetric effects in quarterly German stock returns, even though the leverage effect has a negative sign, differently from what one would expect ($\hat{\theta} = 0.16$ with a t-stat of -2.2).[76] The estimated values of C are positive as expected and also generally statistically significant.

We also estimate a fully conditional model in which both the conditional mean and conditional variances are specified to be time-varying,:

$$r^j_{t+1} = \delta^j_0 + \delta^j_1 dp^j_t + \epsilon^j_{t+1} \qquad j = 1, 2, 3.$$

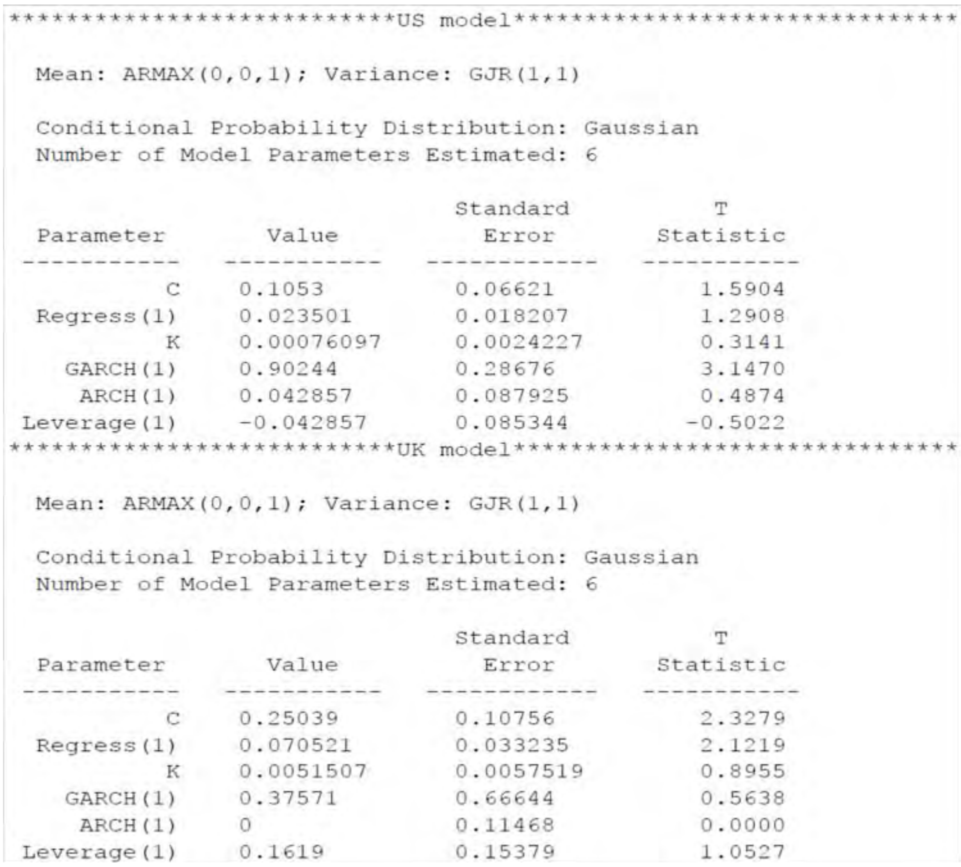
where $r^j_{t+1}$is the log excess return on country j's stock index, and dp^j_t is the log dividend yield of country j. However—just because at this point we do not know what else could be done—we still assume that all correlations equal the unconditional, constant sample correlations for all three pairs of indices. In this case, the unconditional, constant correlations are 0.73 between US and UK returns, 0.60 between US and German returns, and 0.57 between UK and German returns.[77] As we shall see in due time, a multivariate model in which the conditional variances follow a GARCH process but correlations are assumed to be constant over time is called a Constant Conditional Correlation model, CCC.

In this case, a difference between the regression constant C (i.e., δ^j_0, $j = 1, 2, 3$) and the coefficient attached to *Regress(1)*, in this case the dividend-price ratio (i.e., δ^j_1, $j = 1, 2, 3$) appears in the way Matlab prints the estimated coefficients. There is now evidence of GARCH in US stock returns, even though for this time series lagged dividend yields fail to forecast subsequent stock returns; in the case of UK returns, it remains the case that the variance is homoskedastic, but there is evidence that a high dividend yield ratio forecasts higher subsequent returns; finally, in the case of German data, it remains the case that GARCH is strong (but with an odd negative leverage effect), but the dividend-price ratio is not a strong predictor of subsequent returns. Therefore also in this third and more complex model we are probably over-parameterizing the exercise: we are imposing GJR GARCH on UK data when there seems to be evidence of homoskedasticity; we are also forcing a predictive model from past dividend yields multiples to stock returns, when in the case of German

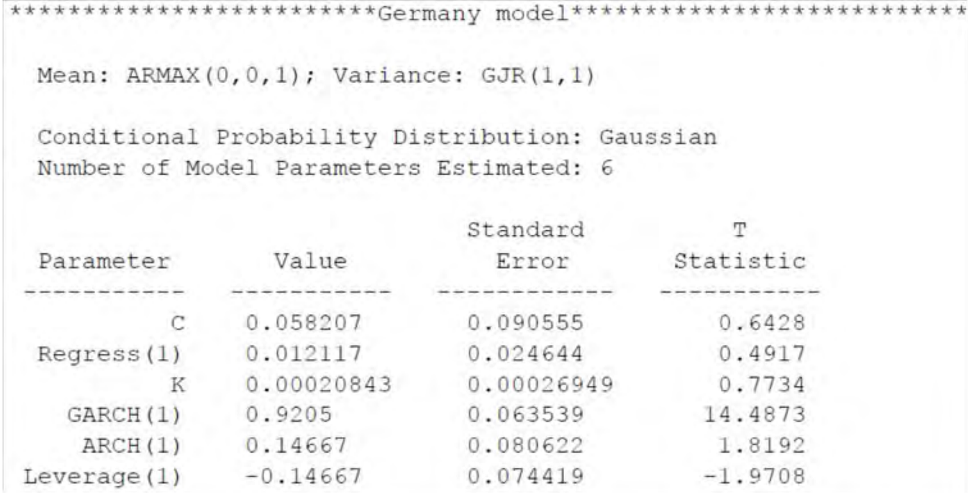
[76] Here it is clear that a decision to estimate a GJR GARCH(1,1) is either arbitrary or triggered by a need to at least accommodate GARCH in German data. We leave it as an exercise to see what happens to optimal weights when GJR GARCH is modelled only for German returns, while UK and US returns are simply taken to be homoskedastic.

[77] I know, these seem to be the same estimates as in a previous footnote, but this is just because of the rounding, see for yourself the differences between **corr_un1=corr(std_resid1)** and **corr_un2=corr(std_resid2)** in the code.

and possibly US data there is weak evidence of such a predictability pattern.



```
****************************US model*******************************

 Mean: ARMAX(0,0,1); Variance: GJR(1,1)

 Conditional Probability Distribution: Gaussian
 Number of Model Parameters Estimated: 6

                             Standard         T
 Parameter        Value        Error       Statistic
-----------    -----------  ------------  -----------
          C     0.1053       0.06621          1.5904
 Regress(1)     0.023501     0.018207         1.2908
          K     0.00076097   0.0024227        0.3141
   GARCH(1)     0.90244      0.28676          3.1470
    ARCH(1)     0.042857     0.087925         0.4874
Leverage(1)    -0.042857     0.085344        -0.5022
****************************UK model*******************************

 Mean: ARMAX(0,0,1); Variance: GJR(1,1)

 Conditional Probability Distribution: Gaussian
 Number of Model Parameters Estimated: 6

                             Standard         T
 Parameter        Value        Error       Statistic
-----------    -----------  ------------  -----------
          C     0.25039      0.10756          2.3279
 Regress(1)     0.070521     0.033235         2.1219
          K     0.0051507    0.0057519        0.8955
   GARCH(1)     0.37571      0.66644          0.5638
    ARCH(1)     0            0.11468          0.0000
Leverage(1)     0.1619       0.15379          1.0527
```



```
*************************Germany model***************************

 Mean: ARMAX(0,0,1); Variance: GJR(1,1)

 Conditional Probability Distribution: Gaussian
 Number of Model Parameters Estimated: 6

                             Standard         T
 Parameter        Value        Error       Statistic
-----------    -----------  ------------  -----------
          C     0.058207     0.090555         0.6428
 Regress(1)     0.012117     0.024644         0.4917
          K     0.00020843   0.00026949       0.7734
   GARCH(1)     0.9205       0.063539        14.4873
    ARCH(1)     0.14667      0.080622         1.8192
Leverage(1)    -0.14667      0.074419        -1.9708
```

At this point, asset allocation is computed by the following lines of code that use the user-provided procedure *mean_variance_multiperiod.* Risk aversion is assumed to be high, 10. Because the weights computed are the weights of the *risky* assets, they might not sum up to 1, in which case what is left of your wealth is invested in the risk-free asset.

```
gamma=10;
lower=0;
upper=10;
rskfree_shortselling=0;
```

```
%Portfolio allocation with GARCH modelling and conditional mean (model 11.c)[78]
[w_11c,miu_portf11c,sigma_portf11c,act_ret_portf11c]=
mean_variance_multiperiod(cov_mat_con1,miu_con1',ret2,gamma,lower,upper,rskfree_shortselling);
%Portfolio allocation with GARCH modelling and constant mean (model 11.b)
[w_11b,miu_portf11b,sigma_portf11b,act_ret_portf11b]=
mean_variance_multiperiod(cov_mat_con2,miu_uncon1',ret2,gamma,lower,upper,rskfree_shortselling);
%Portfolio allocation without GARCH modelling and with constant mean (Gaussian IID model, 11.a)
[w_11a,miu_portf11a,sigma_portf11a,act_ret_portf11a]=
mean_variance_multiperiod(cov_mat_uncon,miu_uncon1',ret2,gamma,lower,upper,rskfree_shortselling);
```

Here *lower=0* and *upper=10* are the lower and upper bounds on the weights of risky assets. *rskfree_shortselling=0* indicates the minimum weight of the risk-free asset and in this case the zero derives from the requirement of no short-selling. Figure A11 plots the resulting portfolio weights. Clearly, the Gaussian IID model implies constant weights over time, because there is no predictability.[79] Visibly, such a model favors UK stocks over US ones and especially German stocks, which are never demanded. Under the remaining two models, recursive optimal portfolio weights become time-varying because as the one-quarter ahead forecasts of variances (the second plot) and of both variances and means (the third plot) change over time, the optimal Markowitz portfolio changes too. Such a variation seems to be substantial and to come more from time-variation in variances than in the means: this is visible from the fact that the second and third plots are somewhat similar (but not identical, of course). This is not surprising also because the predictability from the dividend-price ratio to subsequent stock returns is rather weak, over this sample. In both the second and third plots, the weight attributed to UK stocks remains dominant, but there are now occasional spells (in particular from early 1981 to 1982) in which the weight to be assigned to German stocks is even larger. Moreover, the overall weight to stocks increases somewhat when only predictability in variance is taken into account—it is on average in excess of 40% vs. just less than 40% when both conditional mean and variance are time-varying. Investors may now time periods of favorable predicted moments (i.e., higher than average mean returns and

[78] Just a reminder: in Matlab, lines of codes preceded by a % are simply comments.

[79] One could also have recursively re-estimated sample means and variances, but that would have been spurious because their very variation over time indicates that the IID model should be rejected.

below-average variance, for all or most stock indices).

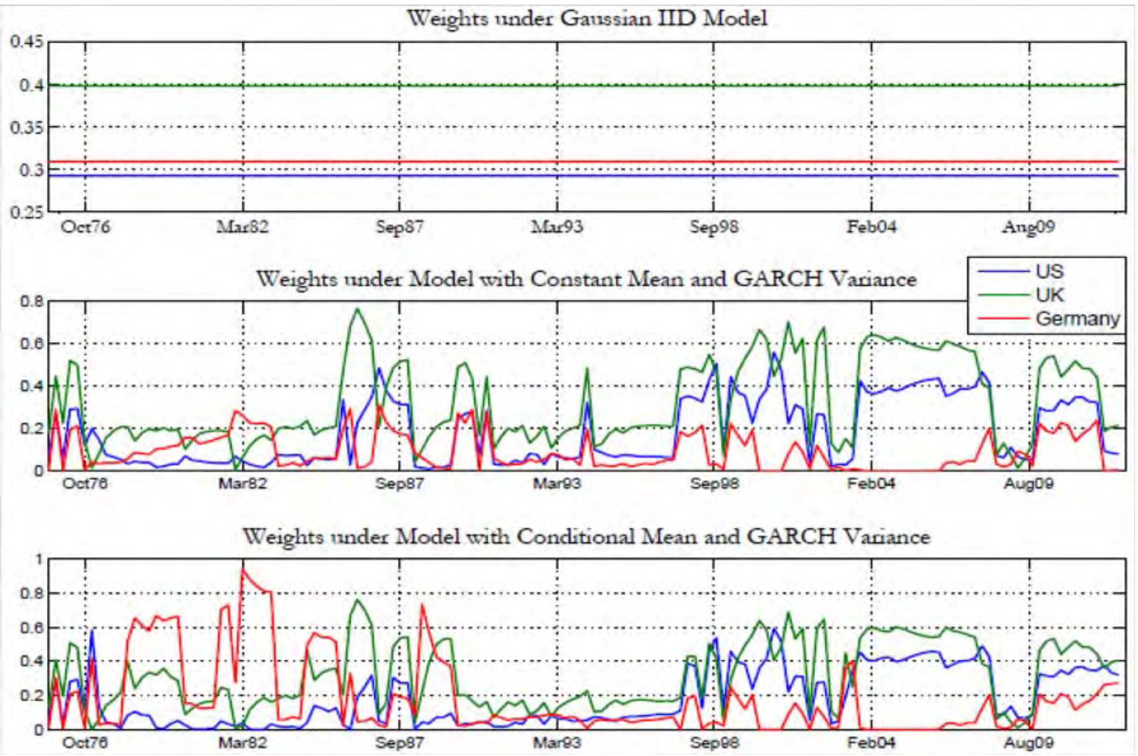


Figure A11: Recursive mean-variance portfolio weights under alternative econometric models

As far as the in-sample Sharpe ratios, because *mean_variance_multiperiod* automatically computes and reports the mean and variance of the optimal portfolios over the asset allocation backtesting sample, in the code it will be sufficient to use

sharpe_ratios=[miu_portf11c./sigma_portf11c miu_portf11b./sigma_portf11b miu_portf11a./sigma_portf11a];

to obtain the Sharpe ratios. Figure A11 plots such in-sample Sharpe ratios showing interesting results. GARCH-based strategies have been most of the time better than classical IID strategies that ignore predictability in variance between 1976 and the late 1990s (these produce a constant Sharpe ratio just in excess of 0.23 that is rather typical at quarterly level). However, the variability has also been substantial. Moreover, in this first part of the sample, to try and predict mean stock returns besides predicting the variance, would have led to a loss in the Sharpe ratio. Starting in 1999, the predictability-based strategies wildly fluctuate according to an easily interpretable pattern: during good times, bull market states (as we known and interpret them ex-post), the predictability-based strategies out-perform the IID strategy; however during bear markets (as identified ex-post), such strategies are inferior vs. the IID one. For instance, during 2004-mid 2007, a strategy that "times" both the conditional mean and the conditional variances, achieves a Sharpe ratio of 0.28 vs. 0.23 for the IID case; however in 2008-2009, the realized Sharpe ratios decline to 0.15-0.18, with a strategy just exploiting time variance in variance performing not as poorly as one instead based on predicting the mean. Keep in mind however that in this exercise we have fixed all pairwise correlations to correspond to their unconditional, full-sample estimates. Moreover, we have used a rather simple GJR GARCH model. It remains to be seen whether jointly modelling all the stock indices—and hence also trying to forecast their correlations—or by complicating the heteroskedastic model may yield superior or more stable in-sample Sharpe ratios.

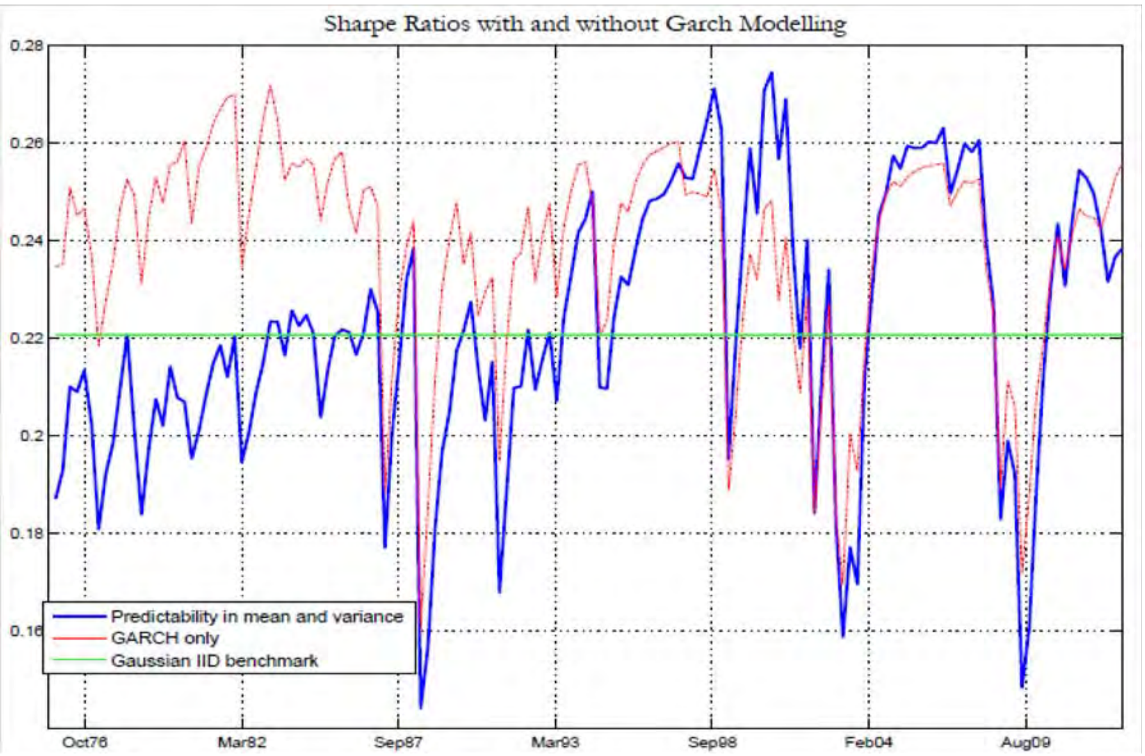


Figure A12: Recursive mean-variance portfolio weights under alternative econometric models

12. Finally, to close this long ride through the application of simple, univariate conditional heteroskedastic models in finance, we have computed the Value-at-Risk with a 95% confidence level and the resulting number of violations for the optimal Markowitz portfolio derived under question 11 above, when both the mean and the variance are predictable. This is performed using the user-provided function *VaR_compute(confidence_level, miu, sigma)* that has the following structure:

```
VaR=NaN(size(miu));
for i=1:rows(VaR)
VaR(i)=norminv(1-confidence_level,miu(i),sigma(i));
end
```

Figure A13 shows the results. As you can see there are several violations, although in the case of a 95% VaR, 5% of them ought to be expected. Yet, we record 6.6% such violations and these are once more frequent and rather painful in two quarters during the financial crisis. How to improve the model to avoid this 1.6% excess of VaR violations is the objective of the next chapter.

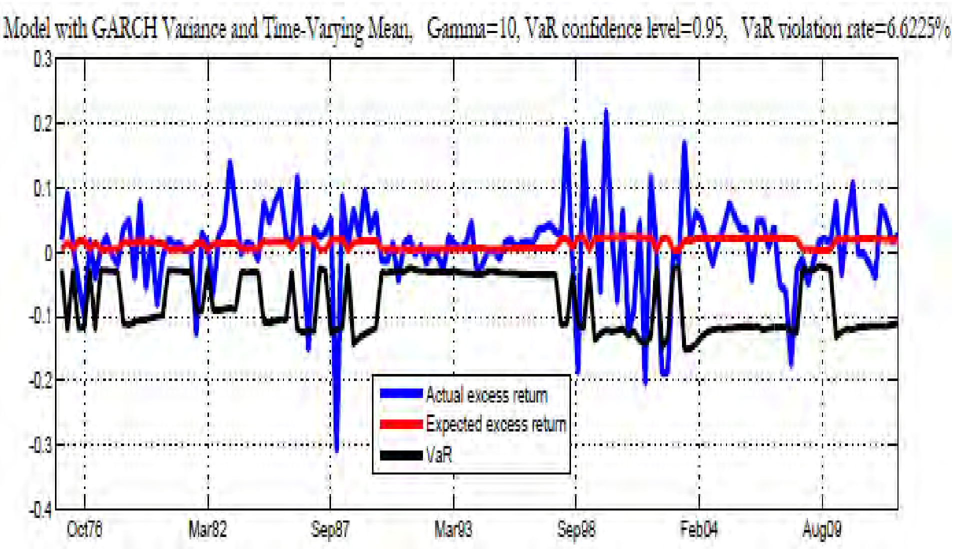


Figure A13: 95% VaR for quarterly returns from optimal mean-variance portfolio

References

[1] Black, F., and Scholes, M., 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637-654.

[2] Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307-327.

[3] Bollerslev, T., and Wooldridge, J., 1992. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews* 11, 143-172.

[4] Christoffersen, P., Jacobs, K., Ornthanalai, C., Wang, Y., 2008. Option valuation with long-run and short-run volatility components. *Journal of Financial Economics* 90, 272-297.

[5] Engle, R., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50 , 987-1007.

[6] Engle, R., Lee, G., 1999. A permanent and transitory component model of stock return volatility. In: Engle, R., White, H. (Eds.), Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger, Oxford University Press, New York, NY, pp. 475-497.

[7] Engle, R., Rangel, J., 2008. The spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *Review of Financial Studies* 21, 1187-1222.

[8] Glosten, L., 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.

[9] Nelson, D., 1990. Conditional heteroskedasticity in asset pricing: A new approach. *Econometrica* 59, 347-370.

List of Errors of Previous Versions and Revisions

(May 20, 2013) Figure A11 has changed because a bug in *mean_variance_multiperiod.m* has been fixed (thanks to Daniele Bianchi for finding it). Qualitative results are similar but the weigth of stocks increase.