Modelling Heteroscedasticity and Non-Normality

April 22, 2014

Contents

1. **Introduction**

In this chapter we concentrate on modelling heteroscedasticity and non-normality. By doing so we shall provide the reader with a number of alternative to the basic GARCH model used in the previous chapter to derive VaR of a given portfolio. The basic procedure which we have illustrated which uses a GARCH forecasting model for volatility and a simple specification for returns to derive by simulation the VaR of interest can then be used with alternative models for volatility, specifications of standardized returns that allow for deviations from normality and simpler methods than simulation to derive VaR.

2. **Computing Measures of Risk without simulation**

VaR simply answers the question "What percentage loss on a given portfolio PF is such that it will only be exceeded $p \times 100\%$ of the time in the next K trading periods (say, days)?" Formally:

$$VaR_{t,K} > 0 \text{ is such that } \Pr(R_{t,K}^{PF} < -VaR_{t,K}) = p,$$

where $R_{t,K}^{PF}$ is a continuously compounded portfolio return between time t and $t+K$, i.e., $R_{t,K}^{PF} \equiv \ln V_{t+K}^{PF} - \ln V_t^{PF}$, where V_t^{PF} is the portfolio value.

It is well known that even though it is widely reported and discussed, the key shortcoming of VaR is that it is concerned only with the range of the outcomes that exceed the VaR measure and not with the overall magnitude (for instance, as captured by an expectation) of these losses. This magnitude, however, should be of serious concern to a risk manager: large VaR exceedances—outcomes below the VaR threshold—are much more likely to cause financial distress, such as bankruptcy, than are small exceedances, and we therefore want to entertain a risk measure that accounts for the magnitude of large losses as well as their probability.[1] The challenge is to come up with a portfolio risk measure that retains the simplicity of the VaR but conveys information regarding the shape of the tail. Expected shortfall (ES), or TailVaR as it is sometimes called, does exactly this.[2] Expected shortfall (ES) is the expected value of tomorrow's return, conditional on it being worse than the VaR at given size p:

$$ES_{t+1}(p) = -E_t[R_{t+1}^{PF} | R_{t+1}^{PF} < -VaR_{t+1}(p)].$$

[1] Needless to say, the most complete measure of the probability and size of potential losses is the entire shape of the tail of the distribution of losses beyond the VaR. Reporting the entire tail of the return distribution corresponds to reporting VaRs for many different coverage rates, say p ranging from .001% to 1% in increments of .001%. It may, however, be less effective as a reporting tool to senior management than is a single VaR number, because visualizing and discussing a function is always more complex than a single number that answers a rather simple question such as "What's the loss so that only 1% of potential losses will be worse over the relevant horizon?"

[2] Additionally, Artzner et al. (1999) define the concept of a coherent risk measure and show that expected shortfall (ES) is coherent whereas VaR is not.

In essence, ES is just (the opposite of) a truncated conditional mean of portfolio returns, where the truncation is provided by VaR. In particular, the negative signs in front of the expectation and the VaR are needed because ES and VaR are defined as positive numbers.

In the previous chapter we have derived VaR via simulation, however the calculation of $VaR_{t,1} = VaR_{t+1}$ is trivial in the univariate case, when there is only one asset ($N = 1$) or one considers an entire portfolio, and $R^{PF}_{t,1}$ has a Gaussian density:[3]

$$p = \Pr(R^{PF}_{t+1} < -VaR_{t+1}) = \Pr\left(\frac{R^{PF}_{t+1} - \mu_{t+1}}{\sigma_{t+1}} < -\frac{VaR_{t+1} + \mu_{t+1}}{\sigma_{t+1}}\right)$$

$$= \Pr\left(z^{PF}_{t+1} < -\frac{VaR_{t+1}(p) + \mu_{t+1}}{\sigma_{t+1}}\right) = \Phi\left(-\frac{VaR_{t+1}(p) + \mu_{t+1}}{\sigma_{t+1}}\right),$$

where $\mu_{t+1} \equiv E_t[R^{PF}_{t+1}]$ is the conditional mean of portfolio returns predicted for time $t+1$ as of time t, $\sigma_{t+1} \equiv \sqrt{Var_t[R^{PF}_{t+1}]}$ is the conditional volatility of portfolio returns predicted for time $t+1$ as of time t (e.g., from some ARCH of GARCH model), and $\Phi(\cdot)$ is the standard normal CDF. Call now $\Phi^{-1}(p)$ the inverse Gaussian CDF, i.e., the value of z_p that solves $\Phi(z_p) = p \in (0,1)$; clearly, by construction, $\Phi^{-1}(\Phi(z_p)) = z_p$.[4] It is easy to see that from the expression above we have

$$\begin{aligned} \Phi^{-1}(p) &= \Phi^{-1}\left(\Phi\left(-\frac{VaR_{t+1}(p) - \mu_{t+1}}{\sigma_{t+1}}\right)\right) = -\frac{VaR_{t+1} + \mu_{t+1}}{\sigma_{t+1}} \\ &\Longrightarrow \quad VaR_{t+1}(p) = -\Phi^{-1}(p)\sigma_{t+1} - \mu_{t+1}. \end{aligned}$$

Note that $VaR_{t+1} > 0$ if $p < 0.5$ and when μ_{t+1} is small (better, zero); this follows from the fact that if $p < 0.5$ (as it is common; as you know typical VaR "levels" are 5 and 1 percent, i.e., 0.05 and 0.01), then $\Phi^{-1}(p) < 0$ so that $-\Phi^{-1}(p)\sigma_{t+1} > 0$ as $\sigma_{t+1} > 0$ by construction. μ_{t+1} is indeed small or even zero—as we have been assuming so far—for daily or weekly data, so that $VaR_{t+1} > 0$ typically obtains.[5] For example, if $\hat{\mu}_{t+1} = 0\%$, $\hat{\sigma}_{t+1} = 2.5\%$ (daily), then

$$\widehat{VaR}_{t+1}(1\%) = -0.025(-2.33) - 0 = 5.825\%,$$

which means that between now and the next period (tomorrow), there is a 1% probability of recording a percentage *loss* of 5.85 percent or larger.

[3] This chapter focusses on one-day-ahead distribution modeling and VaR calculations. Outside, the Gaussian benchmark, predicting multi-step distributions normally requires Monte Carlo simulation, which will be covered in chapter 8.

[4] The notation $z_p \ni \Phi(z_p) = p$ emphasizes that if you change $p \in (0,1)$, then $z_p \in (-\infty, +\infty)$ will change as well. Note that $\lim_{p\to 0^+} z_p = -\infty$ and $\lim_{p\to 1^-} z_p = +\infty$. Here the symbol '$\ni$' means "such that".

[5] What is the meaning of a negative VaR estimate between today and next period? Would it be illogical or mathematically incorrect to find and report such an estimate?

3. **Simple Models for Volatility**

In this section we discuss simpler specification for Volatility than the Benchmark GARCH. These specifications come with the benefit of easier computations and at the cost of potential mis-prediction of volatility.

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

$$\sigma_{t+1}^2 = \frac{1}{m}\sum_{\tau=1}^{T} R_{t+1-\tau}^2 = \sum_{\tau=1}^{T} \frac{1}{m} R_{t+1-\tau}^2. \tag{1}$$

This variance prediction function is simply a constant-weight sum of m past squared returns.[6] 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. 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. However, such extreme gyrations—especially the fact that predicted variance suddenly declines after m 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 (1). 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 clear that a high m will lead to an excessively smoothly evolving σ_{t+1}^2, and that a low m will lead to an excessively jagged pattern of σ_{t+1}^2. 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. A more interesting model that takes this evidence into account when

[6] 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.

computing forecasts of variance is JP Morgan's RiskMetrics system:

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

In this model, the weight on past squared returns declines exponentially as we move backward in time: $1, \lambda, \lambda^2, \ldots$[7] Because of this rather specific mathematical structure, the model is also called the exponential variance smoother. Exponential smoothers have a long tradition in econometrics 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.

(2) 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

$$\begin{aligned}
\sigma^2_{t+1} &= (1-\lambda)R^2_t + \frac{\lambda}{\lambda}(1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau}R^2_{t-\tau} \\
&= (1-\lambda)R^2_t + \lambda\left[\underbrace{\frac{1}{\lambda}(1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau}R^2_{t-\tau}}_{=\sigma^2_t}\right] \\
&= (1-\lambda)R^2_t + \lambda\sigma^2_t.
\end{aligned} \tag{3}$$

(3) 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.[8] In particular,

[7] 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 $\tau \geq 1$.

[8] 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.

notice that

$$\lim_{\lambda \to 1^-} \sigma_{t+1}^2 = \sigma_t^2,$$

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 $\sigma_{t+1}^2 = \sigma_t^2 = \sigma_{t-1}^2 = ... = \sigma_0^2$. 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 (3) presents a number of important advantages:

1. (2) 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 lag coefficient, τ, gets bigger. Figure 1 show the behavior of this weight as a function of τ.

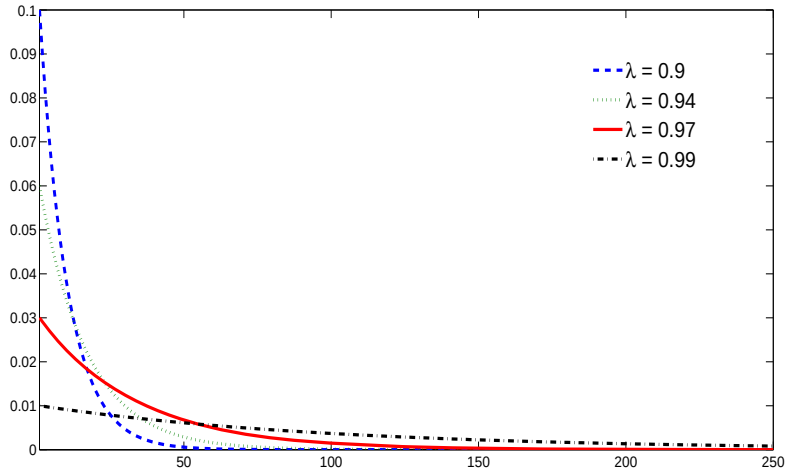

Figure 1 Weights of Past Observations as a function of τ

2. (3) 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.[9]

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

[9] 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 $\lambda = 0.94$.

beyond the current squared return R_t^2, are not needed. Figure 2 shows the behavior of the cumulative weight for a fixed number of past observations as a function of λ.

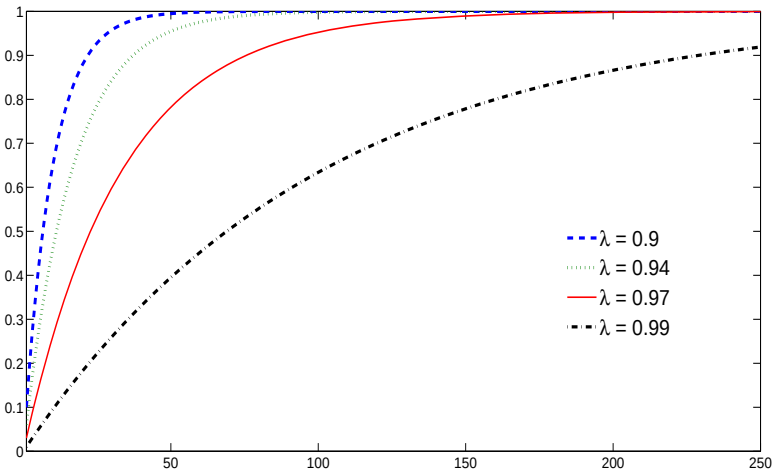

Figure 2 Cumulative Weight on Past Information as a Function of τ

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

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

On the one hand, RiskMetrics and GARCH are not that radically different: comparing (**??**) with (3) 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 $\sigma^2_{t+1|t} < \bar{\sigma}^2$.[10] 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. Consequently, the per-period long-run variance is:

$$\frac{(\sigma^2_{t+1:t+H})_{RMetrics}}{H} = (1-\lambda)R^2_t + \lambda\sigma^2_t = \sigma^2_{t+1|t}$$

Figure 3 illustrates this difference through a practical example in which for the RiskMetrics we set $\lambda = 0.94$.

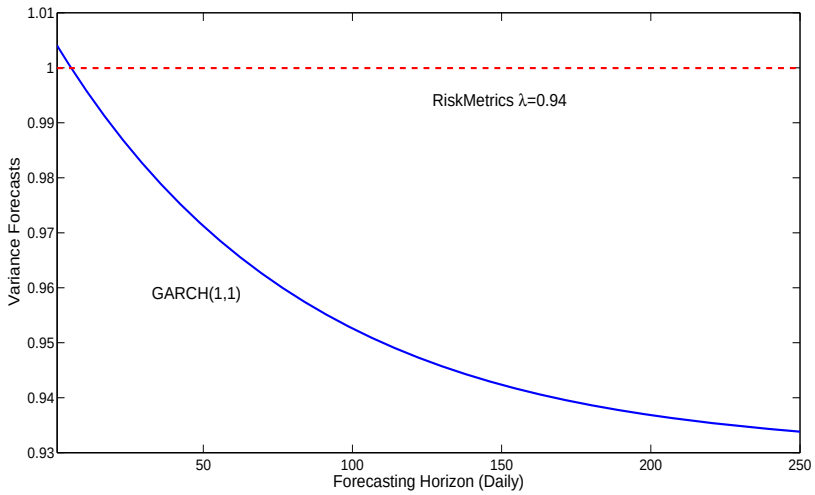

Figure 3: Per-period variance forecasts as a function of H under GARCH(1,1) vs. RiskMetrics

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

4. **Beyond GARCH**

4.1. *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.[11] 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.

Returns on most assets seem to be characterized by an 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.[12] 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

[11] 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.

[12] 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.

quadratic function of R_t^2 and therefore symmetric around 0 (with intercept A). Figure 4 shows such a symmetric NIC from a GARCH(1,1) model.

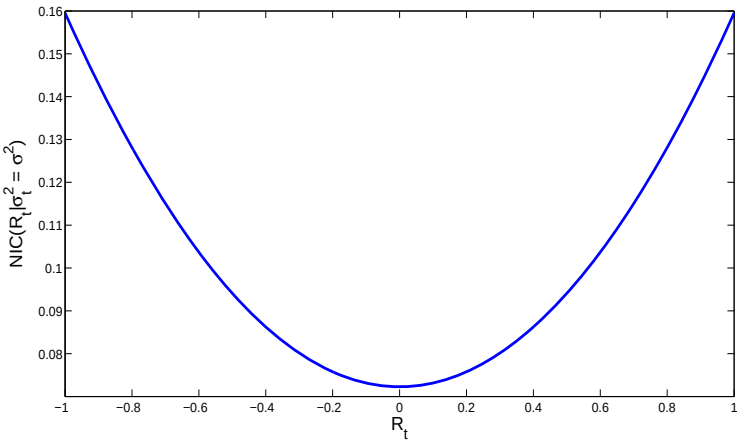

Figure 4: Symmetric NIC from a GARCH model

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.[13] Figure 5 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

[13]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.

regressions may be performed to assess whether the NIC is actually asymmetric.

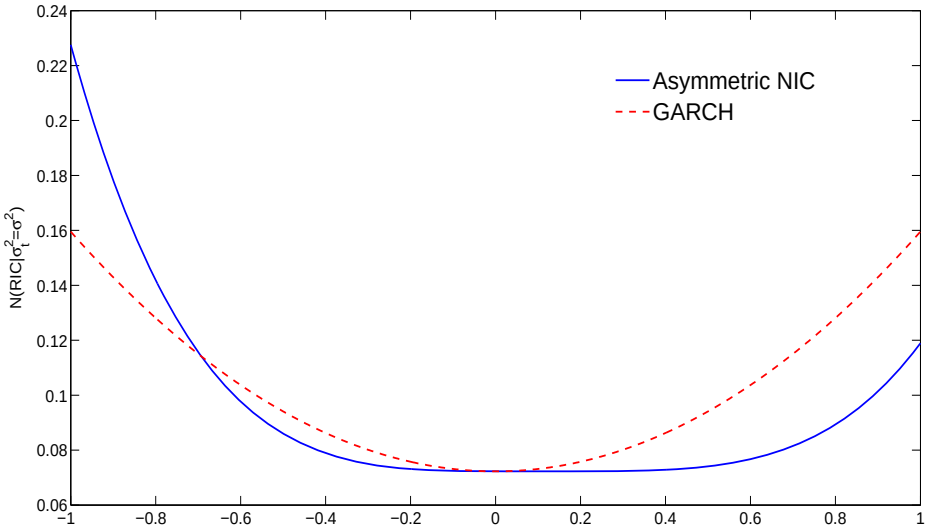

Figure 5: Symmetric and asymmetric NICs

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.

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. GARCH models can be cheaply modified to account for asymmetry, 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:

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

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.

4.2. *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 $f(\boldsymbol{\theta}) : \mathcal{R}^K \rightarrow \mathcal{R}$ 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:[14]

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

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

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

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.

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:

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

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.[15]

4.3. *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:

$$\begin{aligned} \sigma^2_{t+1} &= \omega + \alpha R^2_t + \alpha\theta I_t R^2_t + \beta\sigma^2_t \qquad I_t \equiv \begin{cases} 1 & \text{if } R_t < 0 \\ 0 & \text{if } R_t \geq 0 \end{cases} \quad \text{or} \\ \sigma^2_{t+1} &= \begin{cases} \omega + \alpha(1+\theta)R^2_t + \beta\sigma^2_t & \text{if } R_t < 0 \\ \omega + \alpha R^2_t + \beta\sigma^2_t & \text{if } R_t \geq 0 \end{cases} . \end{aligned} \tag{4}$$

A $\theta > 0$ will again capture the leverage effect. In fact, note that in (4) 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:

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

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

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:[16]

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

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

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

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 $R_t = 0$.

4.4. *NAGARCH model*

One simple choice of parameters in the generalized NIC in (**??**) 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 $\theta_1 > 0$,[17]

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

Written in extensive form that also includes the standard GARCH(1,1) component in (**??**), 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

[16] Obviously, this is the case in the model $R_{t+1} = \sigma_{t+1}z_{t+1}$, $z_{t+1} \sim$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 $z_{t+1} \sim$IID $\mathcal{D}(0,1)$ (e.g., think of a standard t-student). Also recall that $E[\sigma_{t+1}^2] = E[\sigma_t^2] = \bar{\sigma}^2$ by the definition of stationarity.

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

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 $(R_t - \delta\sigma_t)^2 > R_t^2$.[18]

- *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.[19] Here the claim of non-linearity follows from the fact that 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.[20]

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

[18] 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.

[19] 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, $\alpha\left(\sigma_t^2\right) \equiv \alpha\sigma_t^2 > 0$ 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.

[20] 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.

NAGARCH and the assumption of stationarity:[21]

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

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 $\alpha(1 + \delta^2) + \beta < 1$.

4.5. *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

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

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.[22] 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

[21] 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.

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

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[23]

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

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 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

[23] 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, $(1 + R_{t+1}) \in [0, +\infty)$.

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, $\beta > 0$.

4.5.1. **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:[24]

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

Because $E[\sigma_t^2] = \delta_0/(1-\delta_1)$ and also $E[\sigma_t^2] = \omega/(1-\alpha-\beta-\gamma)$, 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.

[24] For the asset pricing buffs, $E[VIX_t] = E[\sigma_t^2]$ may pose some problems, as VIX is normally calculated under the risk-neutral measure while $E[\sigma_t^2]$ 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.

4.6. *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,

$$h_{t+1} = q_{t+1} + \alpha_1(R_t^2 - h_t) + \beta_1(h_t - q_t), \tag{5}$$

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

$$q_{t+1} = \alpha_0 + \rho(q_t - \alpha_0) + \phi(R_t^2 - h_t). \tag{6}$$

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 (5), but that this process tends to evolve around (and mean-revert to) q_{t+1}, which follows itself the process in (6), 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 (5) and (6)—implicitly a (2,2) case has been used in many practical applications. To see that (5)-(6) 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 (6) in (5):

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

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 auto-correlations in squared returns . 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.

5. **Modelling Non-Normality**

So far we have emphasized that dynamic models of conditional heteroskedasticity imply (unconditional) return distributions that are non-normal. However, for most data sets and types of GARCH models, the latter do not seem to generate sufficiently strong non-normal features in asset returns to match the empirical properties of the data, i.e., the strength of deviations from normality that are commonly observed. Equivalently, this means that only a portion—sometimes well below their overall "amount"—of the non-normal behavior in asset returns may be simply explained by the times series models of conditional heteroskedasticity that we have introduced so far. For instance, most GARCH models fail to generate sufficient excess kurtosis in asset returns, when we compare the values they imply with those estimated in the data. This

can be seen from the fact that the standardized residuals from most GARCH models fail to be normally distributed. Starting from the most basic model

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

when one computes the standardized residuals from such typical conditional heteroskedastic framework, i.e.,

$$\hat{z}_{t+1} = \frac{R_{PF,t+1}}{\hat{\sigma}_{t+1}},$$

where $\hat{\sigma}_{t+1}$ is predicted volatility from some conditional variance model, $\hat{z}_{t+1}$ fails to be IID $\mathcal{N}(0,1)$, contrary to the assumption often adopted in estimation.[25] One empirical example can already be seen in Figure 6 where we assess over the sample of daily data January 2006-June 2008 the QQ plots of returns and on standardized (using GARCH and GJR-GARCH volatilities) returns on our portfolio.

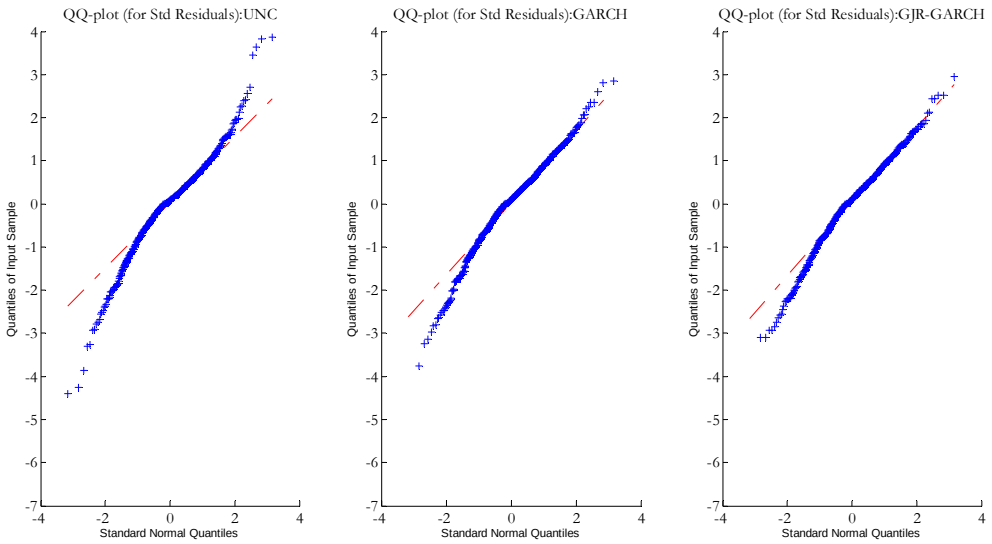

Figure 6: The non-normality of asset returns and standardized residuals from a GARCH model

The Figure illustrates that the standardized residuals originated from fitting a Gaussian GARCH(1,1) model and a GARCH-GJR : $\hat{z}^i_{t+1} = R_{t+1}/\hat{\sigma}^i_{t+1}$ still deviate from normality. If the Gaussian GARCH(1,1) model were correctly specified, then the hypothesis that $\hat{z}^{REIT,GC}_{t+1} \sim$ IID $\mathcal{N}(0,1)$ should not be rejected.

These results tends to be typical for most financial return series sampled at high (e.g., daily or weekly) and intermediate frequencies (monthly). For instance, stock markets exhibit occasional, very large drops but not equally large up moves. Consequently, the return distribution

[25] Some (better) textbooks carefully denote such prediction of volatility as $\sigma_{PF,t+1}$. To save space and paper (in case you print), we shall simply define $\sigma_{t+1} \equiv \sigma_{PF,t+1}$ and trust your memory to recall that we are dealing with a given, fixed-weight portfolio return series, as already explained above.

is asymmetric or negatively skewed. However, some markets such as that for foreign exchange tend to show less evidence of skewness. For most asset classes, in this case including exchange rates, return distributions exhibit fat tails, i.e., a higher probability of large losses (and gains) than the normal distribution would allow.

Note that Figure 6 is not only bad news: the improvement when one moves from the left to the right is obvious. Even though we lack at the moment a formal way to quantify this impression, it is immediate to observe that the "amount" of non-normalities declines when one goes from the raw (original) returns to the Gaussian GARCH-induced standardized residuals and the Gaussian GARCHGJR standardized residulas. Yet, the improvement is insufficient to make the standardized residuals normally distributed, as the model assumes. In the following sections , we also ask how the GARCH models can be extended and improved to deliver unconditional distributions that are distributed in the same way as their original assumptions imply.

5.1. *t-Student Distributions for Asset Returns*

An obvious question is then: if all (most) financial returns have non-normal distributions, what can we do about it? More importantly, this question can be re-phrased as: if most financial series yield non-normal standardized residuals even after fitting many (or all) of the GARCH models analyzed in chapter 4, that assume that such standardized residuals ought to have a Gaussian distribution, what can be done? Notice one first implication of these very questions: especially when high-frequency (daily or weekly) data are involved, we should stop pretending that asset returns "more or less" have a Gaussian distribution in many applications and conceptualizations that are commonly employed outside econometrics: unfortunately, it is rarely the case that financial returns do exhibit a normal distribution, especially if sampled at high frequencies (over short horizons).[26]

When it comes to find remedies to the fact that plain-vanilla, Gaussian GARCH models cannot quite capture the key properties of asset returns, there are two main possibilities that have been explored in the financial econometrics literature. First, to keep assuming that asset returns are IID, but with marginal, unconditional distributions different from the Normal; such marginal distributions will have to capture the fat tails and possibly also the presence of asymmetries. In this chapter we introduce the leading example of the t-Student distribution. Second, to stop

[26] One of the common explanations for the financial collpse of 2008-2009, is that many prop trading desks at major international banks had uncritically downplayed the probability of certain extreme, systematic events. One reason for why this may happen even when a quant is applying (seemingly) sophisticated techniques is that Gaussian shocks were too often assumed to represent a sensible specification, ignoring instead the evidence of jumps and non-normal shocks. Of course, this is just one aspect of why so many international institutions found themselves at a loss when faced with the events of the Fall and the Winter of 2008/09.

assuming that asset returns are IID and model instead the presence of rich—richer than it has been doneso far—dynamics/time-variation in their conditional densities. Indeed, it turns out that both approaches are needed by high frequency (e.g., daily) financial data, i.e., one needs ARCH and GARCH models extended to account for non-normal innovations (see e.g., Bollerslev, 1987).

Perhaps the most important type of deviation from a normal benchmark for $R_{PF,t}$ (or z_t) are the fatter tails and the more pronounced peak around the mean (or the model) for (standardized) returns distribution as compared with the normal one, see Figures 1, 2, and 4. Assume the instead that financial returns are generated by

$$R_{PF,t+1} = \sigma_{t+1} z_{t+1}, \quad z_{t+1} \sim \text{IID } t(d), \tag{7}$$

where σ_{t+1} follows some dynamic process that is left unspecified. The Student t distribution, $t(d)$ parameterized by d (stands for "degrees of freedom") is a relatively simple distribution that is well suited to deal with some of the features discussed above:[27]

$$f_{t(d)}(z;d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{\pi(d-2)}}\left[1+\frac{z^2}{d-2}\right]^{-\frac{d+1}{2}}, \tag{8}$$

where $d > 2$ and $\Gamma(\cdot)$ is the standard gamma function,

$$\Gamma(a) \equiv \int_0^{+\infty} e^{-t} t^{a-1} dt,$$

that is possible to compute not only by numerical integration, but also recursively (but Matlab® will take care of that, no worries). This expression for $f_{t(d)}(z;d)$ gives a non-standardized density, i.e., its mean is zero but its variance is not necessarily 1.[28] Note that while in principle the parameter d should be an integer, in practice quant users accept that in estimation d may turn out to be a real number. It can be shown that first d moments of $t(d)$ will exist, so that $d > 2$ is a way to guarantee that at least the variance exists, which appears to be crucial given our applications to financial data.[29] Another salient property of (8) is that it is only parameterized

[27] Even though in what follows we shall discuss the distribution of z, it is obvious that you can replace that with $R_{PF,t}$ and discuss instead of the distribution of asset returns and not of their standardized residuals.

[28] Christoffersen's book also defines a standardized Student t $f_{\tilde{t}(d)}(z;d)$ with unit variance. Because this may be confusing, we shall only work with the non-standardized case here. A standardized Student t has $Var[\tilde{z};d] = 1$ (note the presence of the tilda again). However, in subsequent VaR calculations, Christoffersen then uses the fact that

$$\Pr\left(z_t\sqrt{\frac{d}{d-2}} < t_p^{-1}(d)\right) = p$$

which means that the empirical variance must be taken into account.

[29] Technically, for the dth moment to exist, it is necessary that d equals d plus any small number, call it ϵ. This is important to understand a few claims that follow.

by d and one can prove (using a few tricks and notable limits from real analysis) that

$$\lim_{d\to\infty} f_{t(d)}(z;d) = f_{\mathcal{N}}(z),$$

as d diverges, the Student-t density becomes identical to a standard normal. This plays a practical role: even though you assume that (8) holds, if estimation delivers a rather large $\hat{d}$ (say, above 20, just to indicate a threshold), this will represent indication that either the data are approximately normal or that (8) is inadequate to capture the type of departure from normality that you are after. What could that be? This is easily seen from the fact that in the simple case of a constant variance, (8) is symmetric around zero, and its mean, variance, skewness (ζ_1), and excess kurtosis (ζ_2) are:

$$\begin{aligned} E[z;d] &= \mu = 0 \qquad Var[z;d] = \sigma^2 = \frac{d}{d-2} \\ Skew[z;d] &= \zeta_1 = 0 \qquad Ex.Kurtosis[z;d] = \zeta_2 = \frac{6}{d-4}. \end{aligned} \tag{9}$$

The skewness of (8) is zero (i.e., the t Student is symmetric around the mean), which makes it unfit to model asymmetric returns: this is the type of departure from normality that (8) cannot yet capture and no small d can be used to accomplish this.[30]

The key feature of the $t(d)$ density is that the random variable, z, is raised to a (negative) power, rather than a negative exponential, as in the standard normal distribution:

$$f_{\mathcal{N}}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

This allows $t(d)$ to have fatter tails than the normal, that is, higher values of the density $f_{t(d)}(z;d)$ when z is far from zero. This occurs because the negative exponential function is known to decline to zero (as the argument goes to infinity, in absolute value) faster than negative power functions may ever do. For instance, observe that for $z = 4$ (which may be interpreted as meaning four standard deviations away from the mean) while

$$e^{-\frac{1}{2}4^2} = 0.0003355,$$

under a negative power function with $d = 10$ (later you shall understand the reason of this choice),

$$\left[1 + \frac{4^2}{8}\right]^{-\frac{11}{2}} = 0.0023759.$$

[30] Let's play (as we shall in do in the class lectures): what is the excess kurtosis of the t-student if $d = 3$? Same question when $d = 4$. What if instead $d = 4.00001$ (which is 4 plus that small ϵ mentioned in a previous footnote)? Does the intution that as $d \to \infty$ the density becomes normal fit with the expression for ζ_2 reported above?

Notice that the second probability value is $(0.0023759/0.0003355) = 7.08$ times larger. If you repeat this experiment considering a really large, extreme realization, say some (standardized) return 12 times away from the sample mean (say a -9.5% return on a given day), then $\exp(-0.5 \cdot 12^2) = 5.3802e^{-32}$ which is basically zero (impossible, but how many -10% did we really see in the Fall of 2008?), while

$$\left[1 + \frac{12^2}{8}\right]^{-\frac{11}{2}} = 9.2652e^{-8}.$$

Although also the latter number is rather small,[31] the ratio between the two probability assessments $(9.2652e^{-8}/5.3802e^{-32})$ is now astronomical $(1.722e^{24})$: events that are impossible under a Gaussian distribution become rare but billions of times more likely under a fat-tailed, t-Student distribution. This result is interesting in the light of the comments we have expressed about the left tail of the density of standardized residuals in Figure 5.

In this section, we have introduced (8) as a way to take care of the fact that, even after fitting rather complex GARCH models, (standardized) returns often seemed not to conform to the properties—such as zero skewness and zero excess kurtosis—of a normal distribution. How do you now assess whether the new, non-normal distribution assumed for z_t actually comes from a Student t? In principle, one can easily deploy two of the methods reviewed in Section 3 and apply them to the case in which we want to test the null of z_t IID $t(d)$: first, extensions of Jarque-Bera exist to formally test whether a given sample has a distribution compatible with non-normal distributions, e.g., Kolmogorov-Smirnov's test (see Davis and Stephens, 1989, for an introduction); second, in the same way in which we have previously informally compared kernel density estimates with a benchmark Gaussian density for a series of interest, the same can be accomplished with reference to, say, a Student-t density. Finally, we can generalize Q-Q plots to assess the appropriateness of non-normal distributions. For instance, we would like to assess whether the same 500 daily returns standardized by a GARCH(1,1) model in Figure 5 may actually conform to a t(d) distribution in Figure 6. Because the quantiles of t(d) are usually not easily found, one uses a simple relationship with a standardized $\tilde{t}(d)$ distribution, where the tilde emphasizes that we are referring to a standardized t:

$$\Pr\left(z_t < t_p^{-1}(d)\sqrt{\frac{d-2}{d}}\right) = \Pr\left(z_t < \tilde{t}_p^{-1}(d)\right)$$

where the critical values of $\tilde{t}_p^{-1}(d)$ are tabulated. Figure 6 shows that assuming t-Student conditional distributions may often improve the fit of a GARCH model.

[31] Please verify that such probability increases becoming not really negligible if you lower the assumption of $d = 10$ towards $d = 2$.

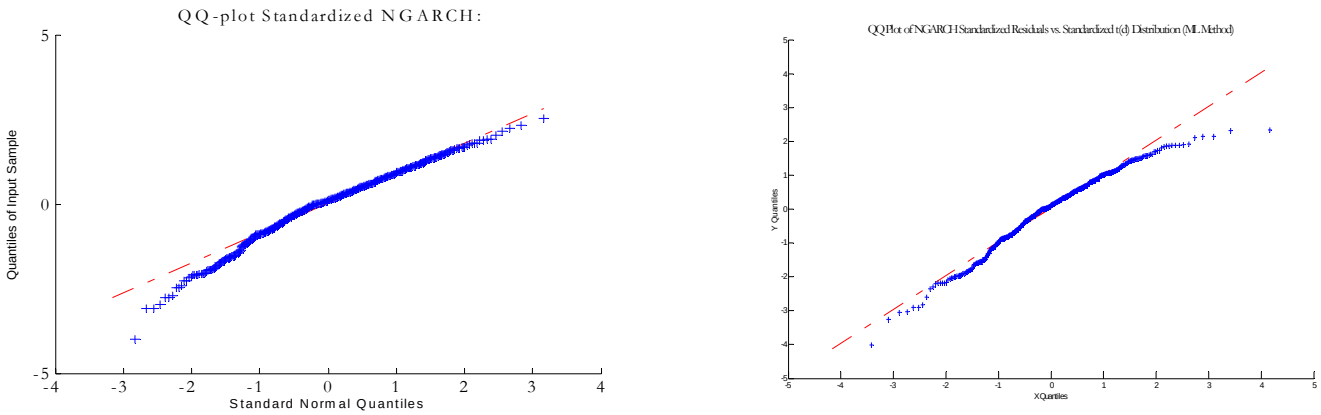

Figure 6: Q-Q plots of Gaussian vs. t-Student GARCH(1,1) standardized daily returns

Although some minor issues with the right tail of the standardized residuals remain, many users may actually judge the left-most QQ plot as completely satisfactory and favorable to a Student t GARCH(1,1) model capturing the salient features of daily returns.

5.2. *Estimation: method of moments vs. (Q)MLE*

We can estimate the parameters of (7)—when we estimate (8) directly on the standardized residuals, we can speak of d only—using MLE or the *method of moments* (MM). As you know from chapter 4, in the MLE case, we will exploit knowledge (real or assumed) of the density function of the (standardized) residuals. Nothing needs to be added to that, apart the fact that the functional form of the density function to be assumed is now given by (8). The method of moments relies instead on the idea of estimating any unknown parameters by simply matching the sample moments in the data with the theoretical (population) moments implied by a t-Student density. The intuition is simple: if the data at hand came from the Student-t family parameterized by d, μ, and σ^2 (say), then the best among the members of such a family will be characterized by a choice of $\hat{d}$, $\hat{\mu}$ and $\hat{\sigma}^2$ that generates population moments that are identical or at least close to the observed sample moments in the data.[32] Technically, if we define the non-central and central sample moments of order $i \geq 1$ (where i is a natural number) as[33]

$$\hat{m}_i \equiv \frac{1}{T}\sum_{t=1}^{T}(z_t)^i \qquad \hat{\widehat{m}}_i \equiv \frac{1}{T}\sum_{t=1}^{T}(z_t - \hat{m}_1)^i,$$

[32] In what follows, we will focus on the simple case in which σ is itself a constant and as such it directly becomes one of the parameters to be estimated. This means that (7) is really considered to be $R_{PF,t+1} = \mu + \sigma z_{t+1}$, $z_{t+1} \sim$IID $t(d)$ where a mean parameter is added, just in case.

[33] Notice that sample moments are sample statistics because they depend on a random sample and as such they are estimators. Instead the population moments are parameters that characterize the entire data generating process. Clearly, $\hat{m}_1 = \hat{\widehat{m}}_i = \hat{E}[z_t]$, while $\hat{\widehat{m}}_2 = \widehat{Var}[z_t]$. The expressions that follow still refer to z_t but there is little problem in extending them to raw portfolio returns ($R_{PF,t}$, as in the lectures) or to any other time series.

respectively, in the case of (7), it is by equating sample and theoretical moments that we get the following system to be solved with respect to the unknown parameters:

$$\begin{aligned} \mu &= \hat{m}_1 \text{ (population mean = sample mean)} \\ \sigma^2 \frac{d}{d-2} &= \widehat{\widehat{m}}_2 \text{ (population variance = sample variance)} \\ \zeta_2 &= \frac{6}{d-4} = \frac{\widehat{\widehat{m}}_4 - 3}{\widehat{\widehat{m}}_2^2} \text{ (population excess kurtosis = sample excess kurtosis).} \end{aligned}$$

Note that all quantities on the right-hand side of this system will turn into numbers when you are given a sample of data. Why these 3 moments? They make a lot of sense given our characterization of (7)-(8) and yet, these are selected, by us, rather arbitrarily (see below). This is a system of 3 equations in 3 unknown (with a recursive block structure) that is easy to solve to find:[34]

$$\hat{d}^{MM} = 4 + \frac{6}{\frac{\widehat{\widehat{m}}_4}{(\widehat{\widehat{m}}_2)^2} - 3} \qquad \hat{\sigma}^2_{MM} = \widehat{\widehat{m}}_2 \frac{\hat{d}^{MM} - 2}{\hat{d}^{MM}} \qquad \hat{\mu}^{MM} = \hat{m}_1.$$

In practice, one first goes from the sample excess kurtosis to estimate the number of degrees of freedom of the Student t, $\hat{d}^{MM}$; then to the estimate of the variance coefficient (also called diffusive coefficient), and finally as well as independently, to compute an estimate of the mean (which is just the sample mean). Interestingly, while under MLE we are used to the fact that one possible variance estimator is $\hat{\sigma}^2_{MLE} = \widehat{\widehat{m}}_2$, in the case of MM applied to the t-Student, we have

$$\hat{\sigma}^2_{MM} = \widehat{\widehat{m}}_2 \frac{\hat{d}^{MM} - 2}{\hat{d}^{MM}} < \hat{\sigma}^2_{MLE}$$

because $(\hat{d}^{MM} - 2)/\hat{d}^{MM} < 1$ for any $\hat{d}^{MM} > 2$. This makes intuitive sense because in the case of a t-Student, the variability of the data is not only explained by their "pure" variance, but also by the fact that their tails are thicker than under a normal: as $\hat{d}^{MM} \to 2$ (from the right), you see that $(\hat{d}^{MM} - 2)/\hat{d}^{MM}$ goes to zero, so that for given $\widehat{\widehat{m}}_2$, $\hat{\sigma}^2_{MM}$ can be much smaller than the sample variance; in that case, most of the variability in the data does come from the thick tails of the Student t. On the contrary, as $\hat{d}^{MM} \to \infty$, we know that this means that the Student t becomes indistinguishable from a normal density, and as such we have that $(\hat{d}^{MM} - 2)/\hat{d}^{MM} \to 1$ and $\hat{\sigma}^2_{MM} \to \widehat{\widehat{m}}_2 = \hat{\sigma}^2_{MLE}$.[35] Additionally, note that as intuition would suggest, as $\hat{\zeta}_2 \equiv (\widehat{\widehat{m}}_4/(\widehat{\widehat{m}}_2)^2) - 3$ gets larger and larger, then

$$\lim_{\hat{\zeta}_2 \to \infty} \hat{d}^{MM} = \lim_{\hat{\zeta}_2 \to \infty} 4 + \frac{6}{\hat{\zeta}_2} = 4,$$

[34] In the generalized MM case (called GMM) in which one has more moments than parameters to estimate, it will be possible to select weighting schemes across different moments that guarantee that GMM estimators may be as efficient as MLE ones. But this is an advanced topic, good for one of your electives.

[35] Even though at first glance it may look so, please do *not* use this example to convince yourself that MLE only works when the data are normally distributed. This is not true (under MLE one needs to know or assume the density of the data, and this can be also non-normal).

where 4 represents the limit of the minimal value for d that one may have with the fourth central moment remaining well-defined under a Student t. Moreover, based on our earlier discussion, we have that

$$\lim_{\hat{\zeta}_2 \to 0} \hat{d}^{MM} = \lim_{\hat{\zeta}_2 \to 0} 4 + \frac{6}{\hat{\zeta}_2} = +\infty,$$

which is a formal statement of the fact that a Student t distribution fitted on data that fail to exhibit fat tails, ought to simply become a normal distribution characterized by a diverging number of degrees of freedom, d. Finally, MM uses no information on the sample skewness of the data for a very simple reason: as we have seen, the Student t in (8) fails to accommodate any asymmetries.

Besides being very intuitive, is MM a good estimation method? Because MM does not exploit the entire empirical density of the data but only a few sample moments, it is clearly not as efficient as MLE. This means that the Cramer-Rao lower bound—the maximum efficiency (the smallest covariance matrix of the estimators) that any estimator may achieve—will not be attained. Practically, this means that in general MM tends to yield standard errors that are larger than those given by MLE. In some empirical applications, for instance when we are assessing models on the basis of tests of hypotheses of some of their parameter estimates, we shall care for standard errors. This result derives from the fact that while MLE exploits knowledge of the density of the data, MM does not, relying only on a few, selected moments (as a minimum, these must be in a number identical to the parameters that need to be estimated). Because while the density $f(z)$ (or the CDF $F(z)$) has implications for all the moments (an infinity of them), but the moments fail to pin down the density function—equivalently, $f(z) \Longrightarrow MGF(z)$, but the opposite does not hold so that it is NOT true that $f(z) \Longleftrightarrow MGF(z)$—MM potentially exploits much less information in the data than MLE does and as such it is less efficient.[36]

Given these remarks, we could of course estimate d also by MLE or QMLE. For instance, $\hat{d}$ could be derived from maximizing

$$\begin{aligned}\mathcal{L}_{1,t(d)}(z_1, z_2, ..., z_T; d) &= \sum_{t=1}^{T} \log f_{t(d)}(z_t; d) = T\left\{\log\Gamma\left(\frac{d+1}{2}\right) - \log\Gamma\left(\frac{d}{2}\right) - \log\frac{\pi}{2} - \log\frac{d-2}{2}\right\} + \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}(1+d)\log\left[1 + \frac{z_t^2}{d-2}\right].\end{aligned}$$

Given that we have already modeled and estimated the portfolio variance $\hat{\sigma}^2_{t+1}$ and taken it as given, we can maximize $\mathcal{L}_{1,t(d)}$ with respect to the parameter, d, only. This approach builds again on the quasi-maximum likelihood idea, and it is helpful in that we are only estimating

[36] Here $MGF(z)$ is the moment generating function of the process of z. Please review your statistics notes/textbooks on what a MGF is and does for you.

few parameters at a time, in this case only one.[37] The simplicity is potentially important as we are exploiting numerical optimization routines to get to $\hat{d} \equiv \arg\max_d \mathcal{L}_{1,t(d)}$. We could also estimate the variance parameters and the d parameter jointly. Section 4.2 details how one would proceed to estimate a model with t Student innovations by full MLE and its relationship with QMLE methods.

5.3. *ML vs. QML estimation of models with Student t innovations*

Consider a model in which portfolio returns, defined as $R_{PF,t} \equiv \sum_{i=1}^{n} \omega_i R_{i,t}$, follow the time series dynamics

$$R_{PF,t+1} = \sigma_{t+1} z_{t+1} \qquad z_{t+1} \sim \text{IID } t(d),$$

where $t(d)$ is a t-Student. As we know, if we assume that the process followed by σ_{t+1} is known and estimated without error, we can treat standardized returns as a random variable on which we have obtained sample data ($\{z_t\}_{t=1}^{T}$), calculated as $z_t = R_{PF,t}/\sigma_t$. The d parameter can then be estimated using MLE by choosing the d which maximizes:[38]

$$\begin{aligned}
\mathcal{L}_{1,t(d)}(z_1, z_2, ..., z_T; d) &= \sum_{t=1}^{T} \ln f(z_t; d) = \sum_{t=1}^{T} \ln \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{\pi(d-2)}} - \frac{1+d}{2} \sum_{t=1}^{T} \ln\left(1 + \frac{z_t^2}{d-2}\right) \\
&= T \ln \Gamma\left(\frac{d+1}{2}\right) - T \ln \Gamma\left(\frac{d}{2}\right) - \frac{1}{2} T \ln \pi - \frac{1}{2} T \ln(d-2) + \\
&\quad - \frac{1+d}{2} \sum_{t=1}^{T} \ln\left(1 + \frac{z_t^2}{d-2}\right).
\end{aligned}$$

On the contrary, if you ignored the estimate of either σ (if it were a constant) or of the process for σ_{t+1} (e.g., a GARCH(1,1) process) and yet you proceeded to apply the method illustrated above (incorrectly) taking some estimate of either σ or of the process for σ_{t+1} as given and free of estimation error, you would obtain a QMLE estimator of d. As already discussed in chapter 4, QML estimators have two important features. First, they are not as efficient as proper ML estimators because they ignore important information on the stochastic process followed by the estimator(s) of either σ or of the process followed by σ_{t+1}.[39] Second, QML estimators will be

[37] However, recall that also QMLE implies a loss of efficiency. Here one should assess whether it is either QMLE or MM that implies that mimimal loss of efficiency.

[38] Of course, Matlab® will happily do this for you. Please see the Matlab workout in Appendix B. See also the Excel estimation performed by Christoffersen (2012) in his book. Note that the constraint $d > 2$ will have to be imposed.

[39] In particular, you recognize that either σ or the process of σ_{t+1} will be estimated with (sometimes considerable) uncertainty (for instance, as captured by the estimate standard errors), but none of this uncertainty is taken into account by the QML maximization. Although the situation is clearly different, it is logically similar to have a sample of size T but to ignore a portion of the data available: that cannot be efficient. Here you would be

consistent and asymptotically normal only if we can assume that any dynamic process followed by σ_{t+1} has been correctly specified. Practically, this means that when one wants to use QML, extra care should be used in making sure that a "reasonable" model for σ_{t+1} has been estimated in the first step, although you see that what may be reasonable is obviously rather subjective.

If instead you do not want to ignore the estimated nature of the process for σ_{t+1} and proceed instead to full ML estimation, for instance when portfolio variance follows a GARCH(1,1) process,

$$\sigma^2_{PF,t} = \omega + \alpha R^2_{PF,t-1} + \beta \sigma^2_{PF,t-1},$$

the joint estimation of d, ω, α, and β implies that the density in the lectures,

$$f(z_t; d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{\pi(d-2)}}\left(1 + \frac{z_t^2}{d-2}\right)^{-\frac{1+d}{2}},$$

must be replaced by

$$f(R_{PF,t}; d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{\pi(d-2)\sigma_t^2}}\left(1 + \frac{(R_{PF,t}/\sigma_t)^2}{d-2}\right)^{-\frac{1+d}{2}}$$

where the σ_t^2 in

$$\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{\pi(d-2)\sigma_t^2}}$$

comes from $f(z_t; d) = t(d)$ so that $f(R_{PF,t}/\sigma_t; d) = t(d)/\sigma_t$ (this is called the Jacobian of the transformation, please review your Statistics notes or textbooks). Therefore, the ML estimates of d, ω, α, and β will maximize:

$$\mathcal{L}_{2,t(d)}(R_1, R_2, ..., R_T; d, \omega, \alpha, \beta) = \sum_{t=1}^{T} \log f(R_{PF,t}; d, \omega, \alpha, \beta) =$$

$$\sum_{t=1}^{T} \log\left\{\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{\pi(d-2)(\omega + \alpha R^2_{PF,t-1} + \beta\sigma^2_{t-1})}}\left(1 + \frac{R^2_{PF,t}}{(d-2)(\omega + \alpha R^2_{PF,t-1} + \beta\sigma^2_{t-1})}\right)^{-\frac{1+d}{2}}\right\}. \tag{10}$$

This looks very hard because the parameters enter in a highly non-linear fashion. Of course Matlab® can take care of it, but there is a way you can get smart about maximizing (10). Define $z_t^{GC} \equiv R_{PF,t}/\sqrt{\omega + \alpha R^2_{PF,t-1} + \beta\sigma^2_{t-1}}$. Call $\mathcal{L}^{GC}_{1,t(d)}(d)$ the likelihood function when the standardized residuals are the z_t^{GC}s and $\mathcal{L}^{GC}_{2,t(d)}(d, \omega, \alpha, \beta)$ the full log-likelihood function defined above. It turns out that $\mathcal{L}^{GC}_{2,t(d)}(d, \omega, \alpha, \beta)$ may be decomposed as

$$\mathcal{L}^{GC}_{2,t(d)}(d, \omega, \alpha, \beta) = \mathcal{L}^{GC}_{1,t(d)}(d) - \frac{1}{2}\sum_{t=1}^{T} \ln(\omega + \alpha R^2_{PF,t-1} + \beta\sigma^2_{t-1}).$$

potentially ignoring important sample information that the data are expressing through the sample distribution of either σ or the process of σ_{t+1}.

This derives from the fact that in (10),

$$\begin{aligned}\mathcal{L}^{GC}_{2,t(d)}(d,\omega,\alpha,\beta) &= T\ln\Gamma\left(\frac{d+1}{2}\right)-T\ln\Gamma\left(\frac{d}{2}\right)-\frac{1}{2}T\ln\pi-\frac{1}{2}T\ln(d-2)+\\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\ln(\omega+\alpha R^2_{PF,t-1}+\beta\sigma^2_{t-1})-\frac{1+d}{2}\sum_{t=1}^{T}\ln\left[1+\frac{(z^{GC}_t)^2}{d-2}\right]\\ &= \mathcal{L}^{GC}_{1,t(d)}(d)-\frac{1}{2}\sum_{t=1}^{T}\ln(\omega+\alpha R^2_{PF,t-1}+\beta\sigma^2_{t-1}).\end{aligned}$$

This decomposition helps us in two ways. First, it shows exactly in what way the estimation approach simply based on the maximization of $\mathcal{L}^{GC}_{1,t(d)}(d)$ is at best a QML one:

$$\arg\max_{d}\mathcal{L}^{GC}_{1,t(d)}(d)\leq \arg\max_{d,\omega,\alpha,\beta}\left[\mathcal{L}^{GC}_{1,t(d)}(d)-\frac{1}{2}\sum_{t=1}^{T}\ln(\omega+\alpha R^2_{PF,t-1}+\beta\sigma^2_{t-1})\right].$$

This follows from the fact that the maximization problem on the right-hand side also exploits the possibility to select the GARCH parameters ω, α, and β, while the one of the left-hand side does not. Second, it suggests a useful short-cut to perform ML estimation, especially under a limited computational power:

- Given some starting candidate values for $[\omega\ \alpha\ \beta]'$ maximize $\mathcal{L}^{GC}_{1,t(d)}(d)$ to obtain $\hat{d}_{(1)}$;
- Given $\hat{d}_{(1)}$, maximize $\mathcal{L}^{GC}_{1,t(d)}(d)-\frac{1}{2}\sum_{t=1}^{T}\ln(\omega+\alpha R^2_{PF,t-1}+\beta\sigma^2_{t-1})$ by selecting $[\hat{\omega}_{(1)}\ \hat{\alpha}_{(1)}\ \hat{\beta}_{(1)}]'$ and compute $\left\{z^{GC,(1)}_t\equiv R_{PF,t}/\sqrt{\hat{\omega}_{(1)}+\hat{\alpha}_{(1)}R^2_{PF,t-1}+\hat{\beta}_{(1)}\sigma^2_{t-1}}\right\}_{t=1}^{T}$;
- Given $[\hat{\omega}_{(1)}\ \hat{\alpha}_{(1)}\ \hat{\beta}_{(1)}]'$ maximize $\mathcal{L}^{GC}_{1,t(d)}(d)$ to obtain $\hat{d}_{(2)}$;
- Given $\hat{d}_{(2)}$, maximize $\mathcal{L}^{GC,(2)}_{1,t(d)}(\hat{d}_{(2)})-\frac{1}{2}\sum_{t=1}^{T}\ln(\omega+\alpha R^2_{PF,t-1}+\beta\sigma^2_{t-1})$ by selecting $[\hat{\omega}_{(2)}\ \hat{\alpha}_{(2)}\ \hat{\beta}_{(2)}]'$ and compute $\left\{z^{GC,(2)}_t\equiv R_{PF,t}/\sqrt{\hat{\omega}_{(1)}+\hat{\alpha}_{(1)}R^2_{PF,t-1}+\hat{\beta}_{(1)}\sigma^2_{t-1}}\right\}_{t=1}^{T}$.

At this point, proceed iterating following the steps above until convergence is reached on the parameter vector $[d\ \omega\ \alpha\ \beta]'$.[40] What is the advantage of proceeding in this fashion? Notice that you have replaced a (constrained) optimization in 4 control variables ($[d\ \omega\ \alpha\ \beta]'$) with an iterative process in which there is a constrained optimization in 1 control followed by a constrained optimization in 3 controls. These may seem small gains, but the general principle may find application to cases more complex than a t-Student marginal density of the shocks, in which more than one additional parameter (here d) may be featured.

[40] For instance, you could stop the algorithm when the Euclidean distance between $[\hat{d}_{(j+1)}\ \hat{\omega}_{(j+1)}\ \hat{\alpha}_{(j+1)}\ \hat{\beta}_{(j+1)}]'$ and $[\hat{d}_{(j)}\ \hat{\omega}_{(j)}\ \hat{\alpha}_{(j)}\ \hat{\beta}_{(j)}]'$ is below some arbitrarily small threshold ϵ (e.g., $\epsilon=1e-04$).

5.4. *A simple numerical example*

Consider extending the moment expressions in (9) to the simple time homogeneous dynamics

$$R_{PF,t} = \mu_{PF} + \sigma z_t \qquad z_t \sim \text{IID } t(d). \tag{11}$$

Because we know that if $z_t \sim$ IID $t(d)$, then $E[z_t] = 0$, $Var[z_t] = d/(d-2)$, $Skew[z_t] = 0$, and $Kurt[z_t] = 3 + 6/(d-4)$, it follows that

$$\begin{aligned}
E[R_{PF,t}] &= \mu_{PF} + \sigma E[z_t] = \mu_{PF} \\
Var[R_{PF,t}] &= \sigma^2 Var[z_t] = \frac{d}{d-2}\sigma^2 \\
E[(R_{PF,t} - E[R_{PF,t}])^3] &= \sigma^3 E[z_t^3] = 0 \\
Kurt(R_{PF,t}) &\equiv \frac{E[(R_{PF,t} - E[R_{PF,t}])^4]}{(Var[R_{PF,t}])^2} \\
&= \frac{\sigma^4}{\sigma^4 (Var[z_t])^2} E[z_t^4] = \frac{E[z_t^4]}{(Var[z_t])^2} = Kurt(z_t) = 3 + \frac{6}{d-4}.
\end{aligned}$$

Interestingly, while mean and variance are affected by the structure of (11), skewness and kurtosis, being standardized central moments, are not.

Clearly, if you had available sample estimates for mean, variance, and kurtosis from a data set of asset returns defined as

$$\begin{aligned}
\hat{m}_1 &\equiv \bar{m}_1 = \frac{1}{T}\sum_{t=1}^{T} R_{PF,t}, \quad \bar{m}_2 \equiv \frac{1}{T}\sum_{t=1}^{T}(R_{PF,t} - \hat{m}_1)^2, \quad \bar{m}_4 \equiv \frac{1}{T}\sum_{t=1}^{T}(R_{PF,t} - \hat{m}_1)^4 \\
\frac{\bar{m}_4}{(\bar{m}_2)^2} &= \frac{\sum_{t=1}^{T}(R_{PF,t} - \hat{m}_1)^4}{\left[\sum_{t=1}^{T}(R_{PF,t} - \hat{m}_1)^2\right]^2},
\end{aligned}$$

it would be easy to recover an estimate of d from sample kurtosis, an estimate of σ^2 from sample variance, and an estimate of μ_{PF} from the sample mean. Using the *method of moments*, we have also in this case 3 moments and 3 parameters to be estimated, which yields the just identified MM estimator (system of equations):

$$\begin{aligned}
\hat{E}[R_{PF,t}] &= \hat{\mu}_{PF} = \bar{m}_1 \\
\widehat{Var}[R_{PF,t}] &= \frac{d}{d-2}\hat{\sigma}^2 = \bar{m}_2 \Longrightarrow \hat{\sigma}^2 = \frac{d-2}{d}\bar{m}_2 \\
\widehat{Kurt}(R_{PF,t}) &= \frac{\bar{m}_4}{(\bar{m}_2)^2} = 3 + \frac{6}{d-4} \Longrightarrow \hat{d} = 4 + \frac{6}{[\bar{m}_4/(\bar{m}_2)^2] - 3}.
\end{aligned}$$

Suppose you are given the following sample moment information on monthly percentage returns on 4 different asset classes (sample period is 1972-2009):

Asset Class/Ptf.	Mean	Volatility	Skewness	Kurtosis
Stocks	0.890	4.657	-0.584	5.226
Real estate	1.052	4.991	-0.783	11.746
Government bonds	0.670	2.323	0.316	4.313
1m Treasury bills	0.465	0.257	0.818	4.334

Calculations are straightforward and lead to the following representations:

Asset/Ptf.	Mean	Vol.	Skew	Kurtosis	Process
Stocks	0.890	4.657	-0.584	5.226	$R_{stock,t}= 0.890 + 3.900 z_t^s \;\; z_t^s \sim t(6.70)$
Real estate	1.052	4.991	-0.783	11.746	$R_{RE,t}= 1.052 + 3.780 z_t^{RE} \;\; z_t^{RE} \sim t(4.69)$
Government bonds	0.670	2.323	0.316	4.313	$R_{bond,t}= 0.670 + 2.034_t^b \;\; z_t^b \sim t(8.57)$
1m Treasury bills	0.465	0.257	0.818	4.334	$R_{Tbill,t}= 0.465 + 0.225 z_t^{TB} \;\; z_t^{TB} \sim t(8.50)$

Clearly, the fit provided by this process cannot be considered completely satisfactory because $Skew[R_{PF,t}] = 0$ for any of the three return series, while sample skewness coefficients—in particular for real estate and 1-month Treasury bill—present evidence of large and statistically significant asymmetries. It is also remarkable that the estimates of d reported for all four asset classes are rather small and always below 10: this means that these monthly time series are indeed characterized by considerable departures from normality, in the form of thick tails. In particular, the $\hat{d}^{REIT} = 4.69$ illustrates how fat tails are for this return time series.

5.5. *A generalized, asymmetric version of the Student t*

The Student t distribution in (8) can accommodate for excess kurtosis in the (conditional) distribution of portfolio/asset returns but not for skewness. It is possible to develop a generalized, asymmetric version of the Student t distribution that accomplishes this important goal. The price to be paid is some degree of additional complexity, i.e., the loss of the simplicity that characterizes the implementation and estimation of (8) analyzed early on this Section. Such an asymmetric t Student is defined by pasting together two distributions at a point $-\psi/\varrho$ on the horizontal axis. The density function is defined by:

$$f_{asyt(d)}(z; d_1, d_2) = \begin{cases} \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1-2)}} \varrho \left[1 + \frac{(\varrho z+\psi)^2}{(1-d_2)^2(d_1-2)}\right]^{-\frac{d_1+1}{2}} & \text{if } z < -\psi/\varrho \\ \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1-2)}} \varrho \left[1 + \frac{(\varrho z+\psi)^2}{(1+d_2)^2(d_1-2)}\right]^{-\frac{d_1+1}{2}} & \text{if } z \geq -\psi/\varrho \end{cases} \tag{12}$$

$$\text{where } \psi \equiv 4d_2 \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1-2)}} \frac{d_1-2}{d_1-1} \qquad \varrho \equiv \sqrt{1 + 3d_2^2 - \psi^2},$$

$d_1 > 2$, and $-1 < d_2 < 1$.[41] Because when $d_2 = 0$, $\psi = 0$ and $\varrho \equiv \sqrt{1 + 3 \times 0 - 0} = 1$, so that

$$f_{asyt(d)}(z; d_1, d_2) = \begin{cases} \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1-2)}} \left[1 + \frac{z^2}{(d_1-2)}\right]^{-\frac{d_1+1}{2}} & \text{if } z < 0 \\ \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1-2)}} \left[1 + \frac{z^2}{(d_1-2)}\right]^{-\frac{d_1+1}{2}} & \text{if } z \geq 0 \end{cases}$$

$$= \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1-2)}} \left[1 + \frac{z^2}{(d_1-2)}\right]^{-\frac{d_1+1}{2}} = f_{t(d)}(z; d),$$

we have that in this case, the asymmetry disappears and we recover the expression for (8) with $d = d_1$. Yes, (12) does not represent a simple extension, as the number of parameters to be estimated in addition to a Gaussian benchmark goes now from one (only d) to two, both d_1 and d_2, and the functional form takes a piece-wise nature. Although also the expression for the (population) excess kurtosis implied by (12) gets rather complicated, for our purposes it is important to emphasize that (12) yields (for $d_1 > 3$, which implies that existence of the third central moment depends on the parameter d_1 only):[42]

$$\zeta_1 = \frac{E[z^3]}{\sigma^3} = \frac{1}{\sqrt[3]{1 + 3d_2^2 - \psi^2}} \left[16d_2 \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1-2)}} (1 + d_2^2) \frac{(d_1-2)^2}{(d_1-1)(d_1-3)} + \right.$$

$$\left. -34d_2 \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1\text{-}2)}} \frac{d_1-2}{d_1-1} (1 + 3d_2^2) + 128d_2^3 \left(\frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\sqrt{\pi(d_1\text{-}2)}} \frac{d_1-2}{d_1-1} \right)^3 \right] \neq 0.$$

It is easy to check that skewness is zero if $d_2 = 0$ is zero.[43] Moreover, skewness is a highly nonlinear functions of both d_1 and d_2, even though it can be verified (but this is hard, do not try unless you are under medical care), that $\zeta_1 \lesseqgtr 0$ if $d_2 \lesseqgtr 0$, i.e., the sign of d_2 determines the sign of skewness. The asymmetric t distribution is therefore capable of generating a wide range of skewness and kurtosis levels.

While in Section 4.1, MM offered a convenient and easy-to-implement estimation approach, this is no longer the case when either returns or innovations are assumed to be generated by (12). The reason is that the moment conditions (say, 4 conditions including skewness to estimate 4 parameters, μ, σ^2, d_1, and d_2) are highly non-linear in the parameters and solving the resulting system of equations will anyway require that numerical methods be deployed. Moreover, the existence of an exact solution may become problematic, given the strict relationship between ζ_1

[41] Christoffersen's book (p. 133) shows a picture illustrating how the asymmetry in this density function depends on the combined signs of d_1 and d_2. It would be a good time to take a look.

[42] The expression for ζ_2 is complicated enough to advise us to omit it. It can be found in Christoffersen (2012).

[43] This is obvious: when $d_2 = 0$, then the generalized asymmetric t Student reduces to the standard, symmetric one.

and ζ_2 implied by (12). In this case, it is common to estimate the parameters by either (full) MLE or at least QMLE (limited to d_1, and d_2).

5.6. *Cornish-Fisher Approximations to Non-Normal Distributions*

The t(d) distributions are among the most frequently used tools in applied time series analysis that allow for conditional non-normality in portfolio returns. However, they build on only few (or one) parameters and in their simplest implementation in (8) they do not allow for conditional skewness in either returns or standardized residuals. As we have seen in Section 2, time-varying asymmetries are instead typical in finance applications. Density approximations represent a simple alternative in risk management that allow for *both* non-zero skewness and excess kurtosis and that remain simple to apply and memorize. Here, one of the easiest to remember and therefore widely applied tools is represented by Cornish-Fisher approximations (see Jaschke, 2002):[44]

$$\begin{aligned} VaR_{t+1}^{CF}(p) &= -CF_p^{-1}\sigma_{t+1} - \mu_{t+1} \\ CF_p^{-1} &\equiv \Phi_p^{-1} + \frac{\zeta_1}{6}\left[(\Phi_p^{-1})^2 - 1\right] + \frac{\zeta_2}{24}\left[(\Phi_p^{-1})^3 - 3\Phi_p^{-1}\right] - \frac{\zeta_1^2}{36}\left[2(\Phi_p^{-1})^3 - 5\Phi_p^{-1}\right], \end{aligned}$$

where $\Phi_p^{-1} \equiv \Phi^{-1}(p)$ to save space and ζ_1, ζ_2 are population skewness and excess kurtosis, respectively. The Cornish-Fisher quantile, CF_p^{-1}, can be viewed as a Taylor expansion around a normal, baseline distribution. This can be easily seen from the fact that if we have neither skewness nor excess kurtosis so that $\zeta_1 = \zeta_2 = 0$, then we simply get the quantile of the normal distribution back, $CF_p^{-1} = \Phi_p^{-1}$, and $VaR_{t+1}^{CF}(p) = VaR_{t+1}(p)$.

For instance, for our monthly data set on U.S. stock portfolio returns, $\hat{\mu}_{t+1} = 0.89\%$, $\hat{\sigma}_{t+1} = 4.66\%$, $\hat{\zeta}_1 = -0.584$, and $\hat{\zeta}_2 = 2.226$. Because $\Phi_p^{-1} = -2.326$, we have:

$$\frac{\hat{\zeta}_1}{6}\left[(\Phi_p^{-1})^2 - 1\right] = -0.423 \qquad \frac{\hat{\zeta}_2}{24}\left[(\Phi_p^{-1})^3 - 3\Phi_p^{-1}\right] = -0.520 \qquad -\frac{\zeta_1^2}{36}\left[2(\Phi_p^{-1})^3 - 5\Phi_p^{-1}\right] = 0.128.$$

Therefore $CF_{0.01}^{-1} = -3.148$ and $\widehat{VaR}_{t+1}^{CF}(1\%) = 13.77\%$ per month. You can use the difference between $\widehat{VaR}_{t+1}^{CF}(1\%) = 13.77\%$ and $\widehat{VaR}_{t+1}^{t}(1\%) = 10.95\%$ to quantify the importance of negative skewness for monthly risk management (2.82% per month).[45] Figure 8 plots 1% VaR

[44] This way of presenting CF approximations takes as a given that many other types of approximations exist in the statistics literature. For instance, the Gram-Charlier's approach to return distribution modeling is rather popular in option pricing. However, CF approximations are often viewed as the basis for an approximation to the value-at-risk from a wide range of conditionally non-normal distributions.

[45] Needless to say, our earlier Gaussian VaR estimate of $\widehat{VaR}_{t+1}(1\%) = 9.94\%$ looks increasingly dangerous, as in a single day it may come to under-estimate the VaR of the U.S. index by a stunning 400 basis points!

for monthly US stock returns data (i.e., again $\hat{\mu}_{t+1} = 0.89\%$, $\hat{\sigma}_{t+1} = 4.66\%$) when one changes sample estimates of skewness ($\hat{\zeta}_1$) and excess kurtosis ($\hat{\zeta}_2$), keeping in mind that $\hat{\zeta}_2 > -3$.

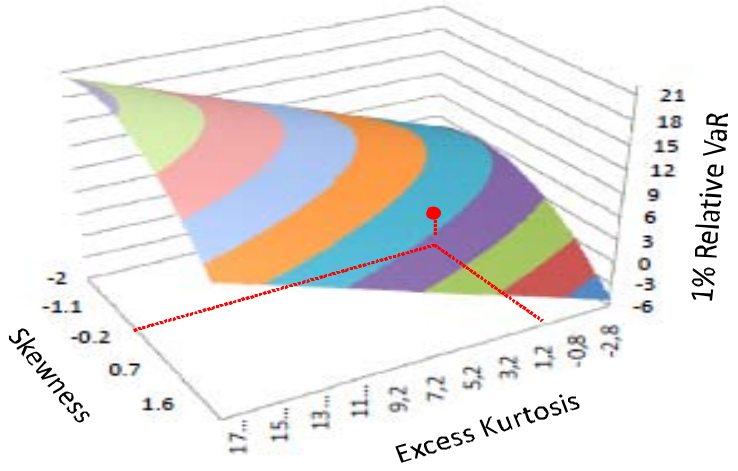

Figure 8: 1% Value-at-Risk estimates as a function of skewness and excess kurtosis

The dot tries to represent in the three-dimensional space the Gaussian benchmark. On the one hand, Figure 8 shows that is easy for a CF VaR to exceed the normal estimate. In particular, this occurs for all combinations of negative sample skewness and non-negative excess kurtosis. On the other hand, and this is rather interesting as many risk managers normally think that accommodating for departures from normality will always increase capital charges, Figure 8 also shows the existence of combinations that yield estimates of VaR that are below the Gaussian estimate. In particular, this occurs when skewness is positive and rather large and for small or negative excess kurtosis, which is of course what we would expect.

5.7. *A numerical example*

Consider the main statistical features of the daily time series of S&P 500 index returns over the sample period 1926-2009. These are characterized by a daily mean of 0.0413% and a daily standard deviation of 1.1521%. Their skewness is -0.00074 and their excess kurtosis is 17.1563. Figure 9 computes the 5% VaR exploiting the CF approximation on a grid of values for daily skewness built as [-2 -1.9 -1.8 ... 1.8 1.9 2] and on a grid of values for excess kurtosis built as

[-2.8 -2.6 -2.4 ... 17.6 17.8 18].

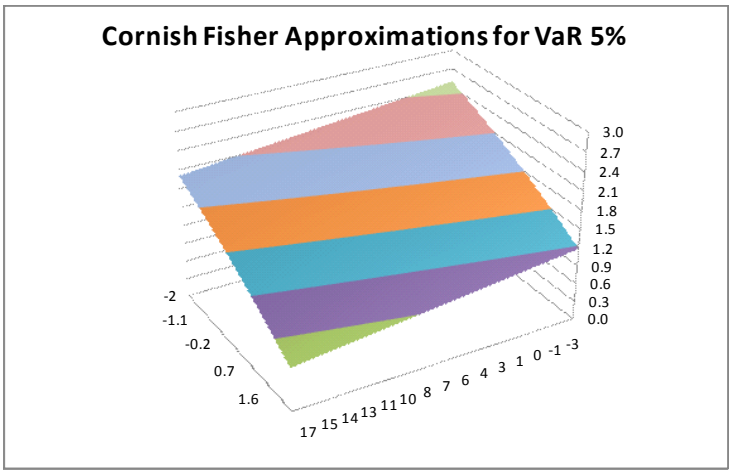

Figure 9: 5% Value-at-Risk estimates as a function of skewness and excess kurtosis

Let's now calculate a standard Gaussian 5% VaR assessment for S&P 500 daily returns: this can be derived from the two-dimensional Cornish-Fisher approximation setting skewness to 0 and excess kurtosis to 0: $\text{VaR}_{0.05} = 1.85\%$. This implies that a standard Gaussian 5% VaR will *over*-estimate the $\text{VaR}_{0.05}$: because S&P500 skewness is -0.00074 and excess kurtosis is 17.1563, your two-dimensional array should reveal an approximate $\text{VaR}_{0.05}$ of 1.46%. Two comments are in order. First, the mistake is obvious but not as bad as you may have expected (the difference is 0.39% which even at a daily frequency may seem moderate). Second, to your shock the mistake does not have the sign you expect: this depends on the fact that while in the lectures, the 1% VaR surface is steeply monotonic increasing in excess kurtosis, for a 5% VaR surface, the shape is (weakly) monotone *decreasing*. Why this may be, it is easy to see, as the term

$$\frac{\zeta_2}{24}[(\Phi_{0.05}^{-1})^3 - 3\Phi_{0.05}^{-1}] \simeq 0.484\frac{\zeta_2}{24} > 0$$

Because $VaR_{t+1}^{CF}(p) = -\sigma_{SP500}CF_{0.05}^{-1}$, i.e., the Cornish-Fisher percentile is multiplied by a -1 coefficient, a positive $\frac{\zeta_2}{24}[(\Phi_{0.05}^{-1})^3 - 3\Phi_{0.05}^{-1}]$ term means that the higher excess kurtosis is, the lower the $\text{VaR}_{0.05}$ is. Now, the daily S&P 500 data present an enormous excess kurtosis of 17.2. This lowers $\text{VaR}_{0.05}$ below the Gaussian $\text{VaR}_{0.05}$ benchmark of 1.85%. Finally,

$$\begin{aligned} VaR_{t+1}^{t}(0.05) &= -\sigma_{SP500}[(\hat{d}-2)/\hat{d}]^{1/2}t_p^{-1}(\hat{d}) \\ &= -1.1521[2.35/4.35]^{1/2}(-2.0835) = 1.764\% \end{aligned}$$

where $\hat{d}$ comes from the method of moment estimation equation

$$\hat{d} = 4 + \frac{6}{\widehat{Kurt}(R_{PF,t}) - 3} = 4 + \frac{6}{20.156 - 3} = 4.35.$$

Notice that also the t-Student estimate of $\text{VaR}_{0.05}$ (1.76%) is lower than the Gaussian VaR estimate, although the two are in this case rather close.

If you repeat this exercise for the case of $p = 0.1\%$, you get Figure 10:

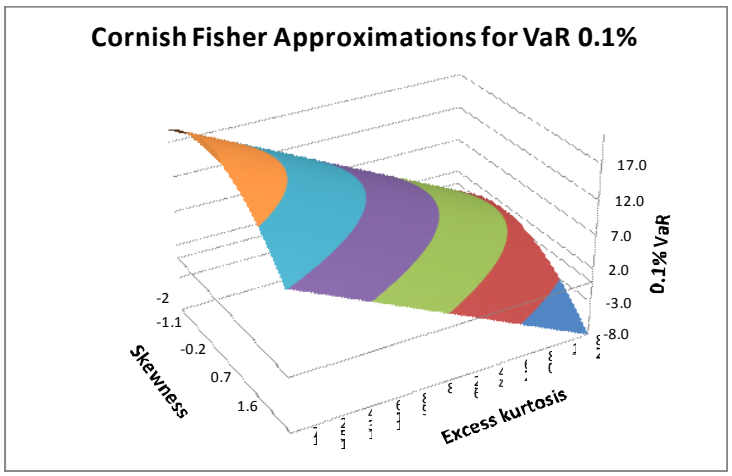

Figure 10: 0.1% Value-at-Risk estimates as a function of skewness and excess kurtosis

Let's now calculate a standard Gaussian 0.1% VaR assessment for S&P 500 daily returns: this can be derived from the two-dimensional Cornish-Fisher approximation setting skewness to 0 and excess kurtosis to 0: $\text{VaR}_{0.001} = 3.52\%$. This implies that a standard Gaussian 5% VaR will severely *under*-estimate the $\text{VaR}_{0.01}$: because S&P500 skewness is -0.00074 and excess kurtosis is 17.1563, your two-dimensional array should reveal an approximate $\text{VaR}_{0.05}$ of 20.50%. Both the three-dimensional plot and the comparison between the CF and the Gaussian $\text{VaR}_{0.001}$ conform with your expectations. First, a Gaussian $\text{VaR}_{0.001}$ gives a massive underestimation of the S&P 500 $\text{VaR}_{0.001}$, which is as large as 20.5% as a result of a huge excess kurtosis. Second, in the diagram, the CF $\text{VaR}_{0.001}$ increases in excess kurtosis and decreases in skewness. In the case of excess kurtosis, this occurs because the term

$$\frac{\zeta_2}{24}[(\Phi_{0.001}^{-1})^3 - 3\Phi_{0.001}^{-1}] \simeq -20.24\frac{\zeta_2}{24} < 0$$

which implies that the higher excess kurtosis is, the higher is $\text{VaR}_{0.001}$. Now, the daily S&P 500 data present an enormous excess kurtosis of 17.2. This increases $\text{VaR}_{0.001}$ well above the Gaussian $\text{VaR}_{0.001}$ benchmark of 3.67%. Finally,

$$\begin{aligned} VaR_{t+1}^t(0.001) &= -\sigma_{SP500}[(\hat{d}-2)/\hat{d}]^{1/2} t_p^{-1}(\hat{d}) \\ &= -1.1521[2.35/4.35]^{1/2}(-6.618) = 5.604\%, \end{aligned}$$

where $\hat{d} = 4.65$. Even though such estimate certainly exceeds the 3.52% obtained under a Gaussian benchmark, this $VaR_{t+1}^t(0.001)$ pales when compared to the 20.50% full CF VaR.

Finally, some useful insight may be derived from fixing the first four moments of S&P 500 daily returns to be: mean of 0.0413%, standard deviation of 1.1521%, skewness of -0.00074, excess kurtosis of 17.1563. Figure 11 plots the VaR(p) measure as a function of p ranging on the grid [0.05% 0.1% 0.15%... 4.9% 4.95% 5%] for four statistical models: (i) a standard Gaussian VaR$_p$; (ii) a Cornish-Fisher VaR$_p$ with CF expansion arrested to the second order, i.e.,

$$VaR_p^{CF,2} = -\sigma_{PF}\left[\Phi_p^{-1} + \frac{\zeta_1}{6}\left(\Phi_p^{-1}\right)^2 - \frac{\zeta_1}{6}\right];$$

(iii) a standard four-moment Cornish-Fisher VaR$_p$ as presented above; (iv) a t-Student VaR$_p$.

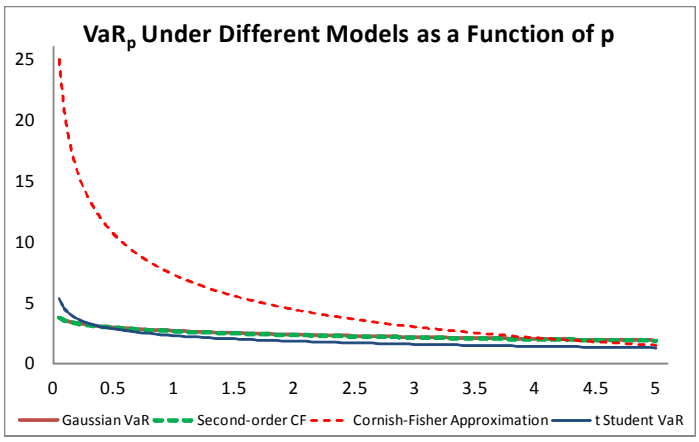

Figure 11: VaR for different coverage probabilities p and alternative econometric models

For high p, there are only small differences among different VaR measures, and a Gaussian VaR may even be higher than VaRs computed under different models. For low values of p, the Cornish-Fisher VaR largely exceeds any other measure because of the large excess kurtosis of daily S&P 500 data. Finally, as one should expect, S&P 500 returns have a skewness that is so small, that the differences between Gaussian VaR and Cornish-Fisher VaR measures computed from a second-order Taylor expansion (i.e., that reflects only skewness) are almost impossible to detect in the plot (if you pay attention, we plotted four curves, but you can detect only three of them).

It is also possible to use the results in Figure 11 to propose *one* measure of the contribution of *skewness* to the calculation of VaR$_p$ and *two* measures of the contribution of *excess kurtosis* to the calculation of VaR$_p$. This is what Figure 12 does. Note that different types of contributions

are measured on different axis/scales, to make the plot readable.

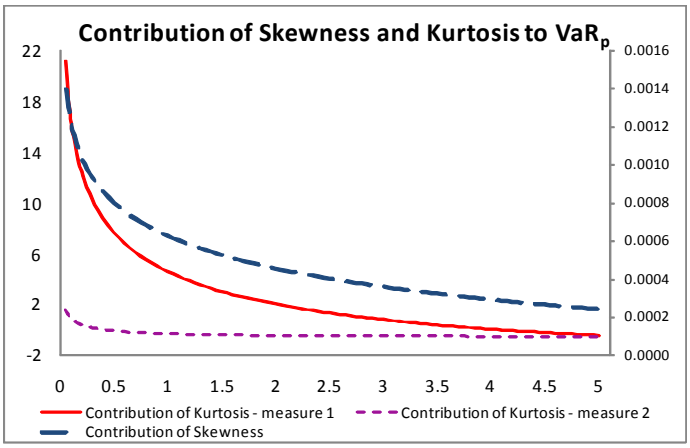

Figure 12: Measures of contributions of skewness and excess kurtosis to VaR

The measure of skewness is obvious, the difference between the second-order CF VaR and the Gaussian VaR measure. On the opposite, for kurtosis we have two possible measures: the difference between the standard CF VaR and the Gaussian VaR, net of the effect of skewness (as determined above); the difference between the symmetric t-Student VaR and the Gaussian VaR, because in the case of t-Student, any asymmetries cannot be captured. Figure 12 shows such measures, with the skewness contribution plotted on the right axis. Clearly, the contribution of skewness is very small, because S&P 500 returns present very modest asymmetries. The contribution of kurtosis is instead massive, especially when measured using CF VaR measures.

6. **Direct Estimation of Tail Risk: A Quick Introduction to Extreme Value Theory**

The approach to risk management followed so far was a bit odd: we are keen to model and obtain accurate estimates of the left tail of the density of portfolio returns; however, to accomplish this goal, we have used time series methods to (mostly, parametrically) model the time-variation in the entire density of returns. For instance, if you care for getting a precise estimate of $\widehat{VaR}_{t+1}(1\%)$ and use a t-Student GARCH(1,1) model (see Teräsvirta, 2009),

$$R_{t+1}^{S\&P} = (\sqrt{\omega + \alpha(R_t^{S\&P})^2 + \beta\sigma_t^2})z_{t+1} \qquad z_{t+1} \sim \text{IID } t(d),$$

you are clearly modelling the dynamics—as driven by changes in σ_t^2 induced by the GARCH—over the entire density over time. But given that your interest is in $\widehat{VaR}_{t+1}(1\%)$, one wonders when and how it can be optimal for you to deal with all the data in the sample and their distribution. Can we do any differently? This is what *extreme value theory* (EVT) accomplishes for you (see McNeil, 1998).

Typically, the biggest risks to a portfolio are represented by the unexpected occurrence of a single large negative return. Having an as-precise-as-possible knowledge of the probabilities of such extremes is therefore essential. One assumption typically employed by EVT greatly simplifies this task: an appropriately scaled version of asset returns—for instance, standardized returns from some GARCH model—must be IID according to some distribution, it is not important the exact parametric nature of such a distribution:[46]

$$z_{t+1} = \frac{R_{PF,t+1}}{\hat{\sigma}_{t+1}} \text{ IID } \mathcal{D}(0,1)$$

Although early on this will appear to be odd, EVT studies the probability that, conditioning that they exceed a threshold u, the standardized returns z less a threshold u are below a value x:

$$F_u(x) \equiv \Pr\{z - u \leq x | z > u\}, \tag{13}$$

where $x > 0$. Admittedly, the probabilistic object in (13) has no straightforward meaning and it does trigger the question: why should a risk or portfolio manager care for computing and reporting it? Figure 13 represents (13) and clarifies that this represents the probability of a "slice" of the support for z. Figure 13 marks a progress in our understanding for the fascination of EVT experts for (13). However, in Figure 13, what remains odd is that we apparently care for a probability slice from the right tail of the distribution of standardized returns.

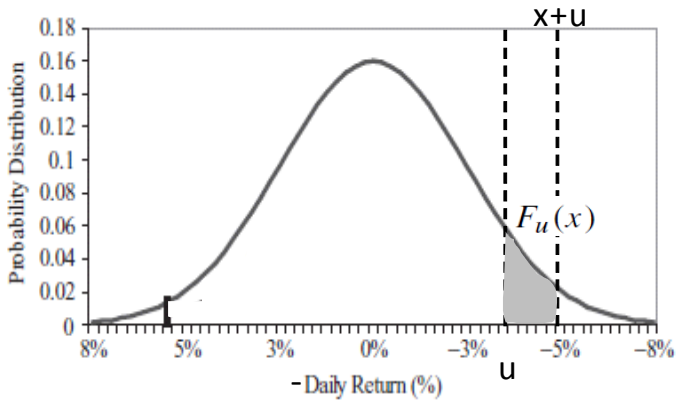

Figure 13: Graphical representation of $F_u(x) \equiv \Pr\{z - u \leq x | z > u\}$

Yet, if you instead of conditioning on some positive value of z, you condition on $-z$, the negative

[46] Unfortunately, the IID assumption is usually inappropriate at short horizons due to the time-varying variance patterns of high-frequency returns. We therefore need to get rid of the variance dynamics before applying EVT, which is what we have assumed above.

of a given standardized return, then, given u, $x > 0$,

$$\begin{aligned} 1 - F_u(x) &\equiv 1 - \Pr\{-z - u \leq x | -z > u\} \\ &= 1 - \Pr\{-z \leq x + u | z < -u\} \\ &= 1 - \Pr\{z > -(x+u) | z < -u\} \\ &= \Pr\{z \leq -(x+u) | z < -u\}, \end{aligned}$$

where we have repeatedly exploited the fact that if $-z > u$ then $-1 \cdot (-z) < -1 \cdot u$ or $z < -u$, and that that $1 - \Pr\{A > B|C\} = \Pr\{A \leq B|C\}$. At this point, the finding that

$$F_u(x) = 1 - \Pr\{z \leq -(x+u) | z < -u\}$$

is of extreme interest: $F_u(x)$ represents the complement to 1 of $\Pr\{z \leq -(x+u)|z < -u\}$, which is the probability that the standardized return does not exceed a negative value $-(x+u) < 0$, conditioning on the fact that such a standardized return is below a threshold $-u < 0$. For instance, if you set $u = 0$ and x to be some large positive value, $1 - F_u(x)$ equals the probability that standardized portfolio returns are below $-x$, conditioning on the fact that these returns are negative and hence in the left tail: this quantity is clearly relevant to all portfolio and risk managers. Interestingly then, while x is the analog to defining the tail of interest through a point in the empirical support of z, u acts as a truncation parameter: it defines how far in the (left) tail our modelling effort ought to go.

In practice, how do we compute $F_u(x)$? On the one hand, this is all we have been doing in this set of lecture notes: any (parametric or even non-parametric) time series model will lead to an estimate of the PDF and hence (say, by simple numerical integration) to an estimate of the CDF $F(x; \hat{\boldsymbol{\theta}})$ from which $F_u(x; \hat{\boldsymbol{\theta}})$ can always be computed as

$$F_u(x) = \frac{\Pr\{u < z \leq x + u\}}{\Pr\{z > u\}} = \frac{F(x+u) - F(u)}{1 - F(u)}, \tag{14}$$

that derives from the fact that for two generic events A and B,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) > 0$$

and the fact that over the real line, $\Pr\{a < z < b\} = F(b) - F(a)$. In principle, as many of our models have implied, such an estimate of the CDF may even be a conditional one, i.e., $F_{u,t+1}(x; \hat{\boldsymbol{\theta}}|\mathcal{F}_t)$. However, as we have commented already, this seems rather counter-intuitive: if we just need an estimate of $F_{u,t+1}(x; \hat{\boldsymbol{\theta}}|\mathcal{F}_t)$, it seems a waste of energies and computational power to first estimate the entire conditional CDF, $F_{t+1}(x; \hat{\boldsymbol{\theta}}|\mathcal{F}_t)$, to then compute $F_{u,t+1}(x; \hat{\boldsymbol{\theta}}|\mathcal{F}_t)$ which may be of interest to a risk manager. In fact, EVT relies one very interesting—once more,

almost "magical"—statistical result: if the series z is independently and identically distributed over time (IID), as you let the threshold, u, get large ($u \to \infty$ so that one is looking at the extreme tail of the CDF), almost any CDF distribution, $F_u(x)$, *for observations beyond the threshold* converges to the *generalized Pareto (GP) distribution*, $G(x;\xi,\beta)$, where $\beta > 0$ and[47]

$$F_u(x) \overset{pointwise}{\to} G(x;\xi,\beta) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right) & \text{if } \xi = 0 \end{cases} \quad \text{where} \quad \begin{cases} x \geq u & \text{if } \xi \geq 0 \\ u \leq x \leq u - \frac{\beta}{\xi} & \text{if } \xi < 0 \end{cases}.$$

ξ is the key parameter of the GPD. It is also called the *tail-index parameter* and it controls the shape of the distribution tail and in particular how quickly the tail goes to zero when the extreme, x, goes to infinity. $\xi > 0$ implies a thick-tailed distribution such as the t-Student; $\xi = 0$ leads to a Gaussian density; $\xi < 0$ to a thin-tailed distribution. The fact that for $\xi = 0$ one obtains a Gaussian distribution should be no surprise: when tails decay exponentially, the advantages of using a negative power function (see our discussion in Section 4) disappear.

At this point, even though for any CDF we have that $F_u(x) \to G(x;\xi,\beta)$, it remains the fact that the expression in (14) is unwieldy to use in practice. Therefore, let's re-write it instead as (for $y \equiv x + u$, a change of variable that helps in what follows):

$$\begin{aligned} F_u(y-u) &= \frac{F(y) - F(u)}{1 - F(u)} \Longrightarrow [1 - F(u)]F_u(y-u) = F(y) - F(u) \\ &\Longrightarrow F(y) = F(u) + [1 - F(u)]F_u(y-u) = 1 - 1 + F(u) + [1 - F(u)]F_u(y-u) \\ &= 1 - [1 - F(u)] + [1 - F(u)]F_u(y-u) = 1 - [1 - F(u)][1 - F_u(y-u)]. \end{aligned}$$

Now let T denote the total sample size and let T_u denote the number of observations beyond the threshold, u: $T_u \equiv \sum_{t=1}^{T} I(z_t > u)$. The term $1 - F(u)$ can then be estimated simply by the proportion of data points beyond the threshold, u, call it

$$1 - \hat{F}(u) = \frac{T_u}{T}.$$

$F_u(y-u)$ can be estimated by MLE on the standardized observations in excess of the chosen threshold u. In practice, assuming $\xi \neq 0$, suppose we have somehow obtained ML estimates of ξ and β in

$$G(x;\xi,\beta) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right) & \text{if } \xi = 0 \end{cases},$$

which we know to hold as $u \to \infty$. Then the resulting ML estimator of the CDF $F(y)$ is:

$$\hat{F}(y) = 1 - \frac{T_u}{T}[1 - \hat{F}_u(y-u)] = 1 - \frac{T_u}{T}\left[1 - 1 + \left(1 + \frac{\hat{\xi} x}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}\right] = 1 - \frac{T_u}{T}\left[\left(1 + \frac{\hat{\xi} x}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}\right]$$

[47] Read carefully: $G(x;\xi,\beta)$ approximates the truncated CDF beyond the threshold u as $u \to \infty$.

so that

$$\lim_{u\to\infty}\hat{F}_u(x)=\frac{1-\frac{T_u}{T}\left[\left(1+\frac{\hat{\xi}x}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}\right]-1+\frac{T_u}{T}}{\frac{T_u}{T}}=1-\left[\left(1+\frac{\hat{\xi}x}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}\right].$$

This way of proceeding represents the "high" way because it is based on MLE plus an application of the GPD approximation result for IID series (see e.g., Huisman, Koedijk, Kool, and Palm, 2001). However, in the practice of applications of EVT to risk management, this is not the most common approach: when $\xi > 0$ (the case of fat tails is obviously the most common in finance, as we have seen in Sections 2 and 3 of this chapter), then a very easy-to-compute estimator exists, namely *Hill's estimator*. The idea is that a rather complex ML estimation that exploits the asymptotic GPD result may be approximated in the following way (for $y > u$):

$$\Pr\{z > y\} = 1 - F(y) = B(y)y^{-\frac{1}{\xi}} \approx cy^{-\frac{1}{\xi}},$$

where $B(y)$ is an appropriately chosen, slowly varying function of y that works for most distributions and is thus (because it is approximately constant as a function of y) set to a constant, c.[48] Of course, in practice, both the constant c and the parameter ξ will have to be estimated. We start by writing the log-likelihood function for the approximate conditional density for all observations y_t as:

$$L(c,\xi)=\prod_{t=1}^{T}f(y_t|y_t>u)=\prod_{i=1}^{T_u}\frac{f(y_i)}{1-F(u)}=-\prod_{i=1}^{T_u}\frac{1}{\xi}cy_i^{-\frac{1}{\xi}-1}\frac{1}{cu^{-\frac{1}{\xi}}}.$$

The expression $f(y_i)/1-F(u)$ in the product involving only observations to the right of the u threshold derives from the fact that

$$f(y_t|y_t>u)=\frac{f(y_t)}{\Pr(y_t>u)}=\frac{f(y_i)}{1-F(u)}$$

for $y_t > u$. Moreover,

$$f(y_i)=\frac{\partial F(y_i)}{\partial y_i}=\frac{\partial\left[1-cy_i^{-\frac{1}{\xi}}\right]}{\partial y_i}=\frac{1}{\xi}cy_i^{-\frac{1}{\xi}-1}.$$

Therefore the log-likelihood function is

$$\mathcal{L}(c,\xi)=\log L(c,\xi)=-\sum_{i=1}^{T_u}\left\{-\log\xi-(\frac{1}{\xi}+1)\log y_t+\frac{1}{\xi}\log u\right\}.$$

Taking first-order conditions and solving, delivers a simple estimator for ξ:[49]

$$\hat{\xi}^{Hill}=\frac{1}{T_u}\sum_{i=1}^{T_u}\ln\left(\frac{y_i}{u}\right)\qquad y_i>u,$$

[48] Formally, this can be obtained by developing in a Taylor expansion $B(y)y^{-1/\xi}$ and absorbing the parameter β into the constant c (which will non-linearly depend on β).

[49] In practice, the Hill's estimator $\hat{\xi}^{Hill}$ is an approximate MLE in the sense that it is derived from taking an approximation of the conditional PDF under the EVT (as $u \to \infty$) and developing and solving FOCs of the corresponding approximate log-likelihood function.

which is easy to implement and remember. At this point, we can also estimate the parameter c by ensuring that the fraction of observations beyond the threshold u is accurately captured by the density as in $\hat{F}(u) = 1 - T_u/T$:

$$1 - \hat{c}u^{-\frac{1}{\hat{\xi}^{Hill}}} = 1 - T_u/T \Longrightarrow \hat{c} = \frac{T_u}{T} u^{\frac{1}{\hat{\xi}^{Hill}}},$$

from the fact that we have approximated $F(u)$ as $1 - cu^{-1/\xi}$. At this point, collecting all these approximation/estimation results we have that

$$\begin{aligned} \hat{F}(y) &= 1 - \hat{c}y^{-\frac{1}{\hat{\xi}^{Hill}}} = 1 - \frac{T_u}{T} u^{\frac{1}{\hat{\xi}^{Hill}}} y^{-\frac{1}{\hat{\xi}^{Hill}}} \\ &= 1 - \frac{T_u}{T}\left(\frac{y}{u}\right)^{-\frac{1}{\hat{\xi}^{Hill}}} = 1 - \frac{T_u}{T}\left(\frac{y}{u}\right)^{-\left[\frac{1}{T_u}\sum_{i=1}^{T_u}\ln\left(\frac{y_i}{u}\right)\right]^{-1}} \end{aligned}$$

where the first line follows from $F(y) \approx 1 - cy^{-\frac{1}{\xi}}$ and the remaining steps have simply plugged estimates in the original equations. Because we had defined $y \equiv x + u$, equivalently we have:

$$\hat{F}^{Hill}(x+u) = 1 - \frac{T_u}{T}\left(1 + \frac{x}{u}\right)^{-\left[\frac{1}{T_u}\sum_{i=1}^{T_u}\ln\left(1+\frac{x_i}{u}\right)\right]^{-1}},$$

which is a Hill/ETV estimator of the CDF when $u \to \infty$, i.e., of the extreme right tail of distribution of (the negative of) standardized returns. This seems rather messy, but the pay-off has been quite formidable: we now have a closed-form expression for the shape of the very far CDF of portfolio percentage losses which does not require numerical optimization within ML estimation. Such an estimate is therefore easy to calculate and to apply within (14), knowing that if $\hat{F}^{Hill}(x+u)$ is available, then

$$\hat{F}_u^{Hill}(x) = \frac{\hat{F}^{Hill}(x+u) - \hat{F}^{Hill}(u)}{1 - \hat{F}^{Hill}(u)}.$$

Obviously, and by construction, such an approximation is increasingly good as $u \to \infty$.

How do you know whether and how your EVT (Hill's) estimator is fitting the data well enough? Typically, portfolio and risk managers use our traditional tool to judge of this achievement, i.e., a (partial) QQ plots. A partial QQ plot consists of a standard QQ plot derived and presented only for (standardized) returns below some threshold loss $-u < 0$. It can be shown that the partial QQ plot from EVT can be built representing in a classical Cartesian diagram the relationship

$$\{X_i, Y_i\} = \left\{ u\left[\frac{i - 0.5}{T} \cdot \frac{T}{T_u}\right]^{-\hat{\xi}}, y_i \right\},$$

where y_i is the ith standardized loss sorted in descending order (i.e., for negative standardized returns). The first and basic logical step consists in taking a time series of portfolio returns and analyzing their (standardized) opposite, i.e., $y_t \equiv -R_{PF,t}/\sigma_t$. This way, one formally looks

at the right-tail conditioning on some threshold $u > 0$, even though the standard logical VaR meanings obtain. In a statistical perspective, the first and initial step is to set the estimated cumulative probability function equal to $1 - p$ so that there is only a p probability of getting a standardized loss worse than the quantile, $(\hat{F}^{-1}_{1-p})$, which is implicitly defined by $F_u(\hat{F}^{-1}_{1-p}) = 1-p$ or

$$1 - \frac{T_u}{T}\left(\frac{\hat{F}^{-1}_{1-p}}{u}\right)^{-1/\hat{\xi}} = 1 - p \Longrightarrow \frac{\hat{F}^{-1}_{1-p}}{u} = \left[p\frac{T}{T_u}\right]^{-\hat{\xi}} \Longrightarrow \hat{F}^{-1}_{1-p} = u\left[p\frac{T}{T_u}\right]^{-\hat{\xi}}.$$

At this point, the Q-Q plot can be constructed as follows: First, sort all standardized returns, y_t, in ascending order, and call the ith sorted value $y_i > u$. Second, calculate the empirical probability of getting a value below the actual as $(i - .5)/T$, where T is the total number of observations.[50] We can then scatter plot the standardized and sorted returns on the Y-axis against the implied ETV quantiles on the X-axis as follows:

$$\{X_i, Y_i\} = \left\{ u\left[\underbrace{\frac{(i-0.5)}{T}}_{\hat{p}\text{ matching }i\text{s quantile}}\frac{T}{T_u}\right]^{-\hat{\xi}}, y_i \right\}.$$

If the data were distributed according to the assumed EVT distribution for $y_i > u$, then the scatter plot should conform roughly to the the 45-degree line.

Because they are representations of partial CDF estimators—limited to the right tail of negative standardized returns, that is the left tail of actual standardized portfolio returns—ETV-based QQ plots are frequently excellent, which fully reflects the power of EVT methods to capture in extremely accurate ways the features of the (extreme) tails of the financial data, see the example in Figure 14. Clearly, everything works in Figure 14, as shown by the fact that all the percentiles practically fall on the left-most branch of the 45-degree line. However, not all is as good as it seems: as we shall see in the worked-out Matlab® session at the end of this chapter, these EVT-induced partial QQ plots obviously suffer from consistency issues, as the same quantile may strongly vary with the threshold u. In fact, and *with reference to the same identical quantiles*, if one changes u, plots that are very different (i.e., much less comforting) than Figure 14 might be obtained and this is logically problematic, as it means that the same method and estimator (Hill's approximate MLE) may give different results as a function of the

[50] The subtraction of .5 is an adjustment allowing for a continuous distribution.

nuisance parameter represented by u.

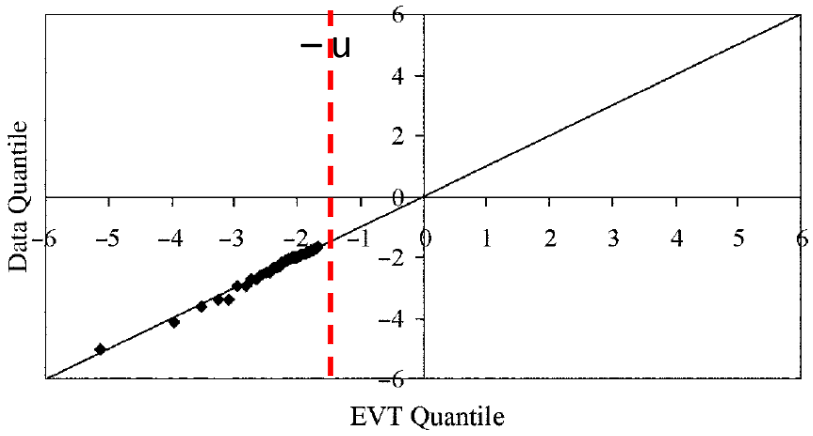

Figure 14: Partial QQ plot for an EVT tail model of $F_u(x) \equiv \Pr\{z - u \leq x | z > u\}$

In itself, the choice of u appears problematic because a researcher must balance a delicate trade-off between bias and variance. If u is set too large, then only very few observations are left in the tail and the estimate of the tail parameter, ξ, will be very uncertain because it is based on a small sample. If on the other hand u is set to be too small, then the EVT key result that all CDFs may be approximated by a GPD may fail, simply because this result held as $u \to \infty$; this means that the data to the right of the threshold do not conform sufficiently well to the generalized Pareto distribution to generate unbiased estimates of ξ. For samples of around 1,000 observations, corresponding to about 5 years of daily data, a good rule of thumb (as shown by a number of simulation studies) is to set the threshold so as to keep the largest 5% of the observations for estimating ξ—that is, we set $T_u = 50$. The threshold u will then simply be the 95th percentile of the data.

In a similar fashion, Hill's p-percent VaR can be computed as (in the simple case of the one-step ahead VaR estimate):

$$VaR_{t+1}^{Hill}(p;u) = F_{1-p,u}^{-1}\sigma_{t+1} + \mu_{t+1}^{y} = u\left[p\frac{T}{T_u}\right]^{-\xi}\sigma_{t+1} + \mu_{t+1}^{y},$$

where $\mu_{t+1}^{y} = -\mu_{t+1}$ represents the conditional mean not for returns but for the negative of returns, $y_t \equiv -R_t$.[51] The reason for using the $(1-p)$th quantile from the EVT loss distribution in the VaR with coverage rate p is that the quantile such that $(1-p) \times 100\%$ of losses are smaller than it is the same as minus the quantile such that $p \times 100\%$ of returns are smaller than it. Note that the VaR expression remains conditional on the threshold u; this an additional parameter that tells the algorithm how specific (tailored) to the tail you want your VaR estimate to be. However, as already commented above with reference to the partial QQ plots, this

[51] The use of the negative of returns explains the absence of negative signs in the expression.

may be a source of problems: for instance one may find that $VaR_{t+1}^{Hill}(1\%; 2\%) = 4.56\%$ but $VaR_{t+1}^{Hill}(1\%; 3\%) = 5.04\%$: even though they are both sensible (as $VaR_{t+1}^{Hill} > u$ which is a minimal consistency requirement), which one should we pick to calculate portfolio and risk management capital requirements?

In the practice of risk management, it is well known that normal and EVT distributions often lead to similar 1% VaRs but to very different 0.1% VaRs due to the different tail shapes that the two methods imply, i.e., the fact that Gaussian models often lead to excessively thin estimates of the left tail. Figure 15 represents one such case: even though the 1% VaR under normal and EVT tail estimates are identical, the left tail behavior is sufficiently different to potentially cause VaR estimates obtained for $p << 1\%$ to differ considerably. The tail of the normal distribution very quickly converges to zero, whereas the EVT distribution has a long and fat tail.

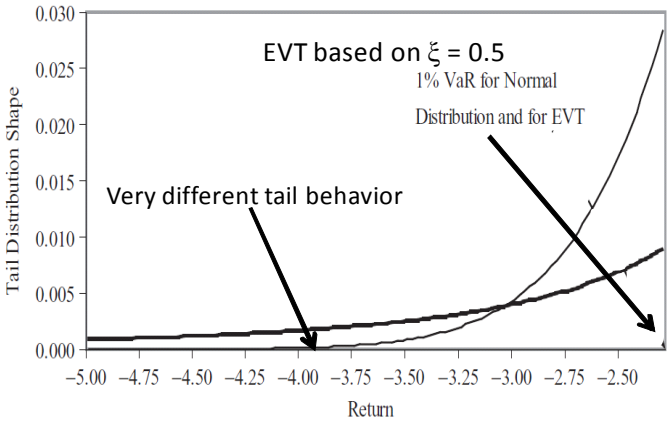

Figure 15: Different tail behavior of normal vs. EVT distribution models

Visually, this is due to the existence of a crossing point in the far left tail of the two different distributions. Therefore standard Basel-style VaR calculations based on a 1% coverage rate may conceal the fact that the tail shape of the distribution does not conform to the normal distribution: in Figure 15, VaRs below 1% will differ by a factor as large as 1 million! In this example, the portfolio with the EVT distribution is much riskier than the portfolio with the normal distribution in that it implies non-negligible probabilities of very large losses. What can we do about it? The answer is to supplement VaR measures with other measures such as plots in which VaR is represented as a function of p (i.e., one goes from seeing VaR as an estimate of an unknown parameter to consider VaR as an estimate of a function of p, to assess the behavior of the tails) or to switch to alternative risk management criteria, for instance the *Expected Shortfall* (also called TailVaR), see Appendix A for a quick review of the concept.

How can you compute ES in practice? For the remainder of this Section, assume $\mu_{t+1} = 0\%$. Let's start with the bad news: it is more complex than in the case of the plain-vanilla VaR

because ES actually conditions on VaR. In fact, usually one has to perform simulations under the null of a given econometric model to be able to compute an estimate of ES. Now it is time for the good news: at least in the Gaussian case, one can find a (sort of) closed form expression:

$$ES_{t+1}(p) = -E_t[R_{t+1}^{PF}|R_{t+1}^{PF} < -VaR_{t+1}(p)] = \sigma_{t+1}\frac{\phi\left(-\frac{VaR_{t+1}(p)}{\sigma_{t+1}}\right)}{\Phi\left(-\frac{VaR_{t+1}(p)}{\sigma_{t+1}}\right)} = \sigma_{t+1}\frac{\phi\left(\Phi_p^{-1}\right)}{p}$$

where the last equality follows from $VaR_{t+1}(p) = -\sigma_{t+1}\Phi_p^{-1}$ and $\Phi\left(-\Phi_p^{-1}\right) = p$. Here $\phi(\cdot)$ denotes the standard normal PDF, while $\Phi(\cdot)$ is, as before, the standard normal CDF. For instance, if $\sigma_{t+1} = 1.2\%$, $ES_{t+1}(p) = 0.012\{[(-2\pi)^{-1/2}\exp(-(-2.33)^2/2)]/0.01\} = 3.17\%$ from

$$\phi(z) = (-2\pi)^{-1/2}\exp\left(-\frac{z^2}{2}\right).$$

Interestingly, the ratio between $ES_{t+1}(p)$ and $VaR_{t+1}(p)$ possesses two key properties. First, under Gaussian portfolio returns, as $p \to 0^+$, $ES_{t+1}(p)/VaR_{t+1}(p) \to 1$ and so there is little difference between the two measures. This makes intuitive sense: the ES for a very extreme value of p basically reduces to the VaR estimate itself as there is very little probability mass left to the left of VaR. In general, however, the ratio of ES to VaR for fat-tailed distribution will be higher than 1, which was already the intuitive point of Figure 15 above. Second, for EVT distributions, when p goes to zero, the ES to VaR ratio converges to

$$\lim_{p\to 0^+}\frac{ES_{t+1}(p)}{VaR_{t+1}(p)} = \frac{1}{1-\xi},$$

so that as $\xi \to 1$ (which is revealing of fat tails, as claimed above), $ES_{t+1}(p)/VaR_{t+1}(p) \to +\infty$.[52] Moreover, the larger (closer to 1) is $\xi < 1$, the larger is $ES_{t+1}(p)$ for given $VaR_{t+1}(p)$.

Appendix 1 — A Matlab® Workout on Modelling Volatility

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.[53]

[52] For instance, in Figure 15, where $\xi = 0.5$, the ES to VaR ratio is roughly 2, even though the 1% VaR is the same in the two distributions. Thus, the ES measure is more revealing than the VaR about the magnitude of losses larger than the VaR.

[53] 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.

1. 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).

2. 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.

3. 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.

4. 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?

5. 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.

6. 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

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

$\alpha = 0$ 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; 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)?

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:

1. 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.

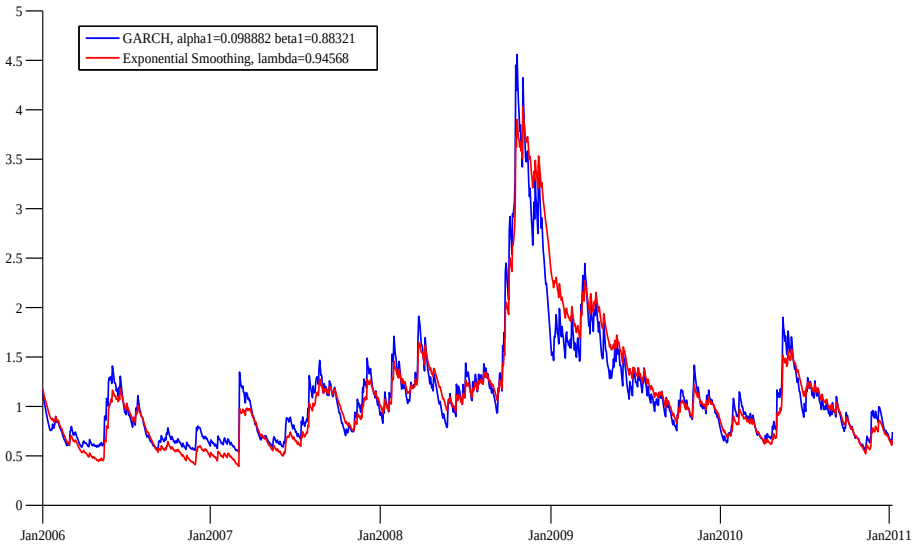

Figure A5:Comparing in-sample predictions of conditional volatility from GARCH vs. RiskMetrics

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)^2;

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.

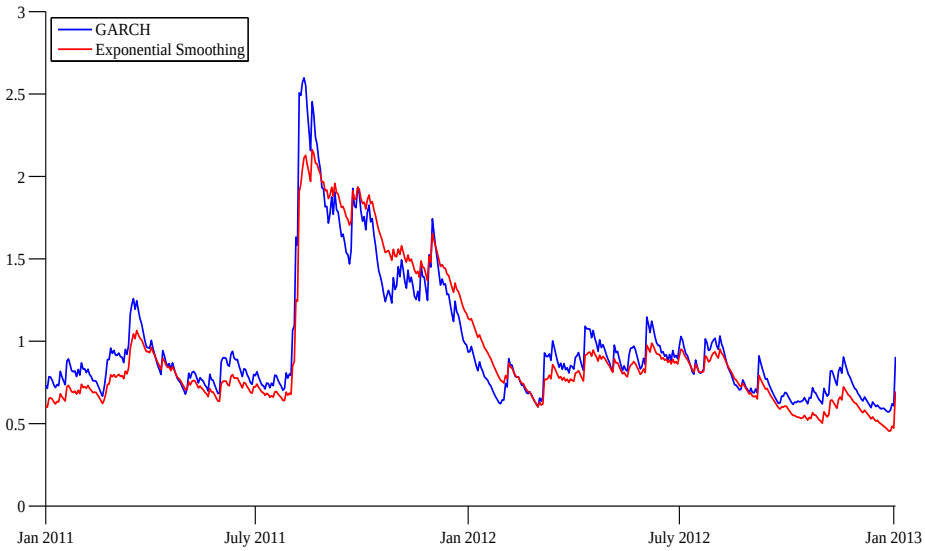

Figure A6:Comparing out–of-sample predictions of conditional volatility from GARCH vs. RiskMetrics

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)*epsilonˆ2+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)*epsilonˆ2+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.

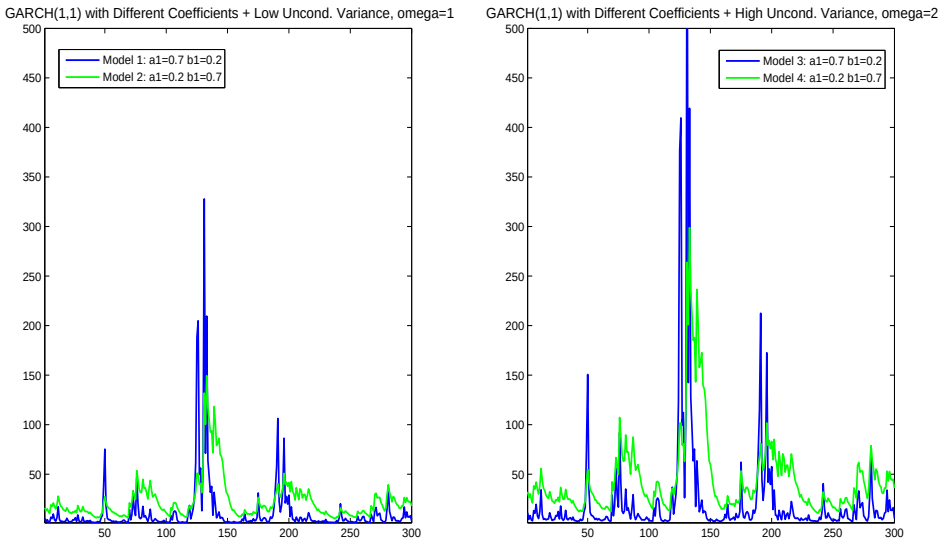

Figure A7: Simulating 4 alternative GARCH models

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 $H \geq 1$.

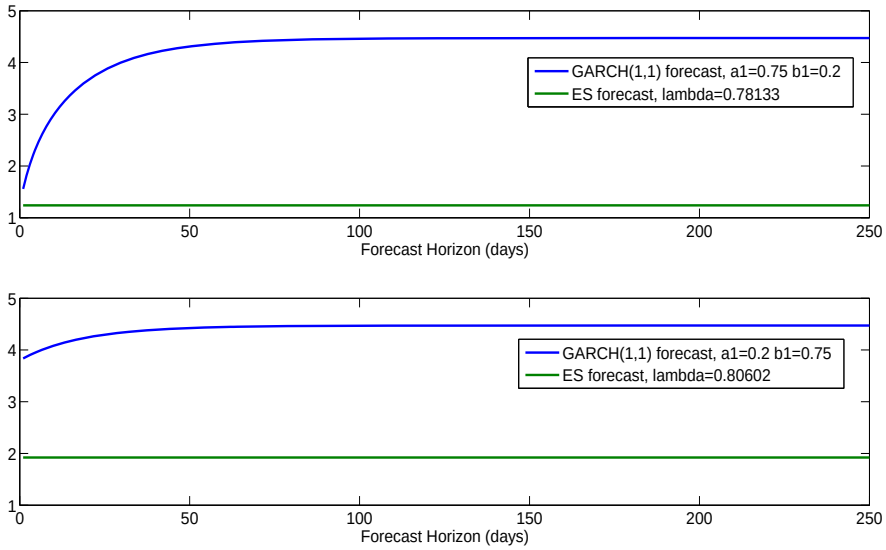

Figure A8: Variance forecasts (250 daily) from two alternative GARCH models vs. RiskMetrics

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.[54] 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

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

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.

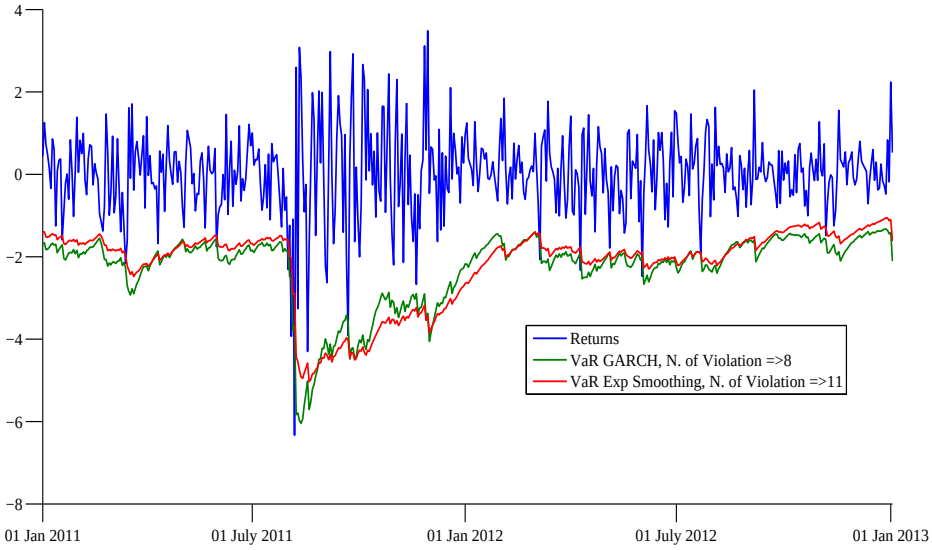

Figure A9: Daily 1% VaR bounds from GARCH vs. RiskMetrics

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, 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):[55]

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    0.023148       0.027168         0.8520
           K    0.022564       0.003262         6.9173
    GARCH(1)    0.9055         0.013017        69.5627
     ARCH(1)    0             0.012043         0.0000
 Leverage(1)    0.14436        0.019254         7.4980

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         1.1973
            K    0.0057138      0.0032902        1.7366
     GARCH(1)    0.97518        0.0034576      282.0408
      ARCH(1)    0.12183        0.016455         7.4042
  Leverage(1)    -0.10597       0.011775        -8.9999

 Stationarity measure 1.0440
```

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

993All 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 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.

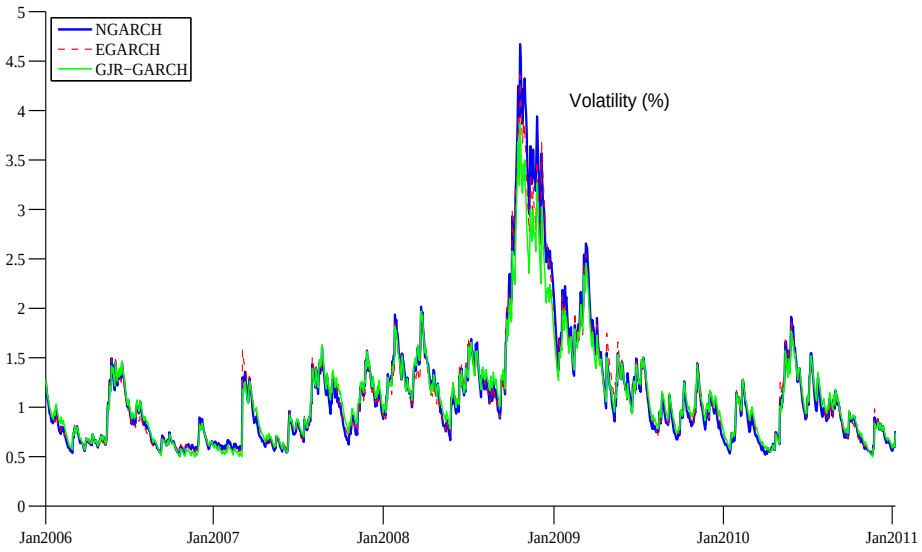

Figure A10: Comparing in-sample fitted volatility dynamics under GJR, EGARCH, and NAGARCH

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:
    0.1763    0.8855

With t-stats for the null of alpha=0 and beta=1 of:
    1.0662   -2.0001


and an R-square of:
    0.1552

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.

Appendix B — A Matlab® Workout on Modelling Non-Normality

Suppose you are a European investor and your reference currency is the Euro. You evaluate the properties and risk of your *equally weighted* portfolio on a daily basis. Using daily data in STOCKINT2013.XLS, construct daily returns (*in Euros*) using the three price indices **DS Market-PRICE Indexes** for three national stock markets, Germany, the US, and the UK.

1. For the sample period of 03/01/2000- 31/12/2011, plot the returns on each of the three individual indices and for the equally weighted portfolio *denominated in Euros.* Just to make sure you have correctly applied the exchange rate transformations, also proceed to plot the exchange rates derived from your data set.

2. Assess the normality of your portfolio returns by computing and charting a QQ plot, a Gaussian Kernel density estimator of the empirical distribution of data, and by performing a Jarque-Bera test using daily portfolio data for the sample period 03/01/2000-31/12/2011. Perform these exercises both with reference to the raw portfolio returns (in euros) and with reference to portfolio returns standardized using the unconditional sample mean standard deviation over your sample. In the case of the QQ plots, observe any differences between the plot for raw vs. standardized returns and make sure to understand the source of any differences. In the case of the Kernel density estimates, produce two plots, one comparing a Gaussian density with the empirical kernel for portfolio returns and the other comparing a Gaussian density with the empirical kernel for portfolio returns standardized using the unconditional sample mean and standard deviation over your sample. In the case of the Jarque-Bera tests, comment on the fact that the test results seem not to depend on whether raw or standardized portfolio returns are employed. Are either the raw portfolio or the standardized returns normally distributed?

3. Estimate a GARCH with leverage model over the same period and assess the normality of the resulting standardized returns. You are free to shop among the asymmetric GARCH models with Gaussian innovations that are offered by Matlab and the ones that have been presented during the lectures. In any event make sure to verify that the estimates that you have obtained are compatible with the stationarity of the variance process. Here it would be useful if you were to estimate at least two different leverage GARCH models and compare the normality of the resulting standardized residuals. Can you find any evidence that either of the two volatility models induces standardized residuals that are consistent with the assumed model, i.e., $R_{t+1} = \sigma_{t+1} z_{t+1}$ with z_{t+1} IID $N(0,1)$?

4. Simulate returns for your sample using *at least* one GARCH with leverage model, calibrated on the basis of the estimation obtained under the previous point with normally

distributed residuals. Evaluate the normality properties of returns and standardized returns using QQ plots and a Kernel density fit of the data.

5. Compute the 5% Value at Risk measure of the portfolio for each day of January 2012 (in the Excel file, January 2012 has 20 days) using, respectively, a Normal quantile when variance is constant (homoskedastic), a Normal quantile when conditional variance follows a GJR process, a t-Sstudent quantile with the appropriately estimated number of degrees of freedom and a Cornish-Fisher quantile and compare the results. Estimate the number of degrees of freedom by maximum likelihood. In the case of a conditional t-Student density and of the Cornish-Fisher approximation, use a conditional variance process calibrated on the filtered conditional GJR variance in order to define standardized returns. The number of degrees of freedom for the t-Student process should be estimated by QML.

6. Using QML, estimate a $t(d)$-NGARCH(1,1) model. Fix the variance parameters at their values from question 3. If you have not estimated a (Gaussian) NGARCH(1,1) in question 3, it is now time to estimate one. Set the starting value of d equal to 10. Construct a QQ plot for the standardized returns using the standardized $t(d)$ distribution under the QML estimate for d. Estimate again the $t(d)$-NGARCH(1,1) model using now full ML methods, i.e., estimating jointly the t-Student d parameter as well as the four parameters in the nonlinear GARCH written as

$$\sigma_t^2 = \omega + \alpha(R_{t-1} - \theta\sigma_{t-1})^2 + \beta\sigma_{t-1}^2.$$

   Is the resulting GARCH process stationary? Are the estimates of the coefficients d different across QML and ML methods and why? Construct a QQ plot for the standardized returns using the standardized $t(d)$ distribution under the ML estimate for d. Finally, plot and compare the conditional volatilities resulting from your QML (two-step) and ML estimates of the $t(d)$-NGARCH(1,1) model.

7. Estimate the EVT model on the standardized portfolio returns from a Gaussian NGARCH(1,1) model using the Hill estimator. Use the 4% largest losses to estimate EVT. Calculate the 0.01% standardized return quantile implied by each of the following models: Normal, $t(d)$, Hill/EVT, and Cornish-Fisher. Notice how different the 0.01% VaRs would be under these alternative four models. Construct the QQ plot using the EVT distribution for the 4% largest losses. Repeat the calculations and re-plot the QQ graph when the threshold is increased to be 8%. Can you notice any differences? If so, why are these problematic?

8. Perform a simple asset allocation exercise under three alternative econometric specifica-

tions using a Markowitz model, under a utility function of the type

$$U(\mu_t, \sigma_t^2) = \mu_t - \frac{1}{2\gamma}\sigma_t^2,$$

with $\gamma = 0.5$, in order to determine optimal weights. Impose no short sale constraints on the stock portfolios and no borrowing at the riskless rate. The alternative specifications are:

(a) Constant mean and a GARCH (1,1) model for conditional variance, assuming normally distributed innovations.

(b) Constant mean and an EGARCH (1,1) model for conditional variance, assuming normally distributed innovations.

(c) Constant mean and an EGARCH (1,1) model for conditional variance, assuming t-Student$t$ distributed innovations.

Perform the estimation of the model parameters using a full sample of data until 02/01/2013. Note that, just for simplicity (we shall relax this assumption later on) all models assume a constant correlation among different asset classes, equal to sample estimate of their correlations in pairs. Plot optimal weights and the resulting *in-sample*, realized Sharpe ratios of your optimal portfolio under each of the three different frameworks. Comment the results. [IMPORTANT: Use the toolboxes *regression_tool_1.m* and *mean_variance_multiperiod.m* that have been made available with this exercise set]

Solution

This solution is a commented version of the MATLAB code Ex_CondDist_VaRs_2013.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:

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

The above two lines import only the numbers, not the strings, from a .txt file.[56] The following lines of the codes take care of the strings:

[56]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.

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 a set of instructions that is not be commented in detail because their structure closely resembles other plots proposed in Lab 1, see worked-out exercise in chapter 4. Figure A1 shows the euro-denominated returns on each of the four indices.

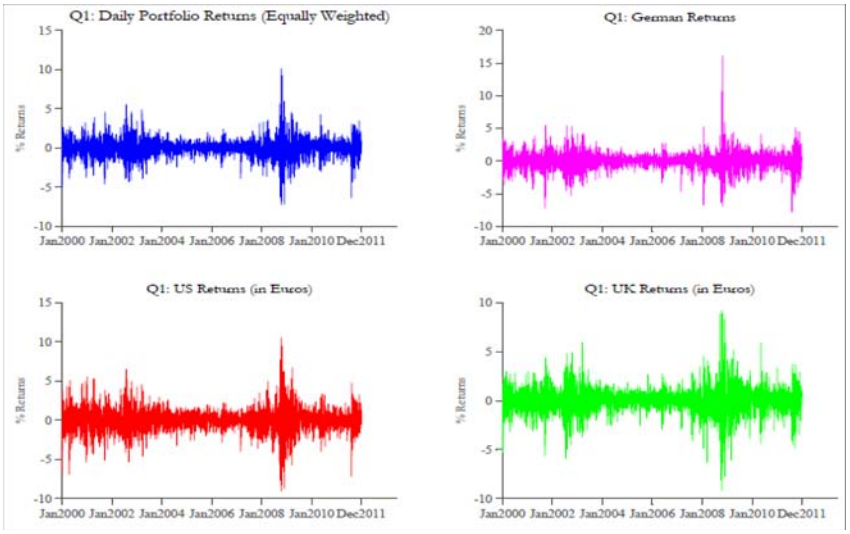

Figure A1:Daily portfolio returns on four national stock market indices

Even though these plots are affected by the movements of the €/$ and £/$ exchange rates, the volatility bursts recorded in early 2002 (Enron and Worldcom scandal and insolvency), the

Summer of 2011 (European sovereign debt crisis), and especially the North-American phase of the great financial crisis in 2008-2009 are well-visible.

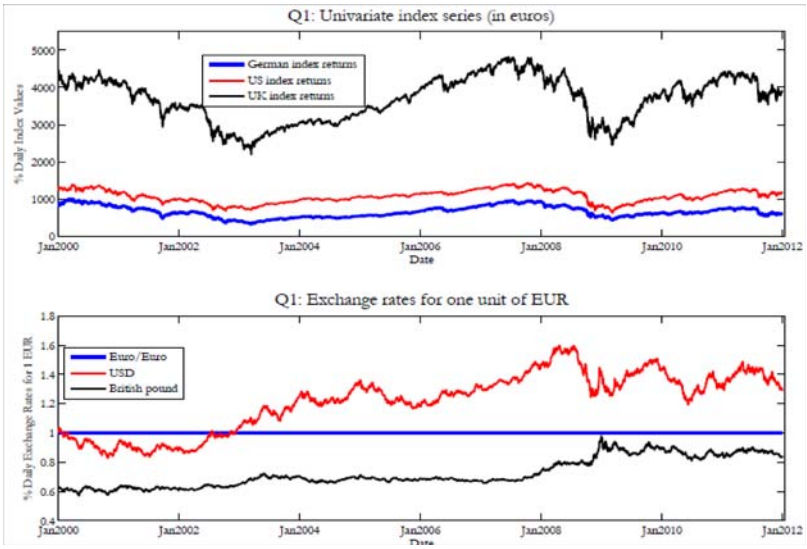

Figure A2:Daily portfolio indices and exchange rates

As requested, Figure A2 plots the values of both indices and implied exchange rates, mostly to make sure that the currency conversions have not introduced any anomalies.

2. The calculation of the unconditional sample standard deviation and the standardization of portfolio returns is simply performed by the lines of code:

unc_std=std(port_ret(ind(1):ind(2)));

std_portret=(port_ret(ind(1):ind(2))-mean(port_ret(ind(1):ind(2))))./unc_std;

Note that standardizing by the unconditional standard deviation is equivalent to divide by a constant, which is important in what follows. The set of instructions that produces QQ plots and displays them horizontally to allow a comparison of the plots of raw vs. standardized returns iterates on the simple function:

qqplot(RET(:,i));

where **qqplot** displays a quantile-quantile plot of the sample quantiles of X versus theoretical quantiles from a normal distribution. If the distribution of X is normal, the plot will be close to linear. The plot has the sample data displayed with the plot symbol '+'.[57] Figure A3

[57]Superimposed on the plot is a line joining the first and third quartiles of each distribution (this is a robust linear fit of the order statistics of the two samples). This line is extrapolated out to the ends of the sample to help evaluate the linearity of the data. Note that 'qqplot(X,PD)' would create instead an empirical quantile-quantile plot of the quantiles of the data in the vector X versus the quantiles of the distribution specified by PD.

displays the two QQ plots and emphasizes the strong, obvious non-normality of both raw and standardized data.

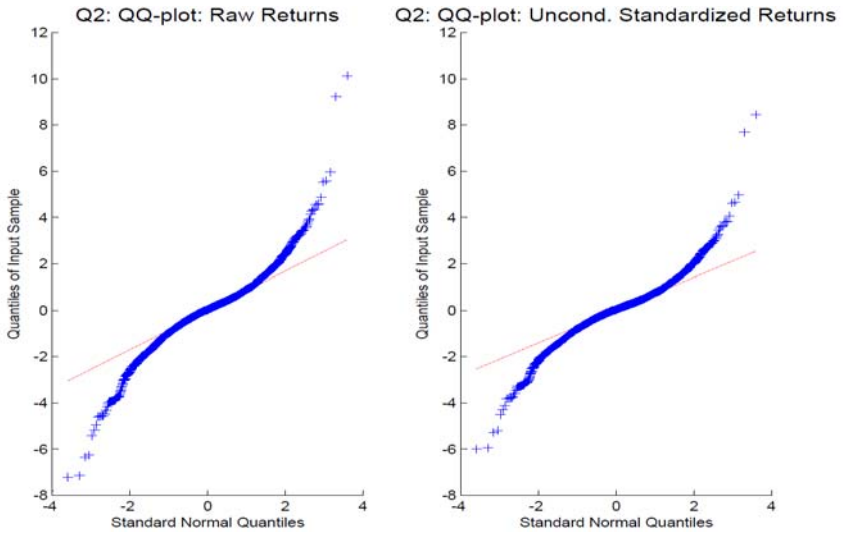

Figure A3:Quantile-quantile plots for raw vs. standardized returns (under constant variance)

The kernel density fit comparisons occur between a normal distribution, that is simply represented by a simulation performed by the lines of codes

norm=randn(1000*rows(RET(:,1)),1);
norm1=mean(RET(:,1))+std(RET(:,1)).*norm;
norm2=mean(RET(:,2))+std(RET(:,2)).*norm;
[Fnorm1,XInorm1]=ksdensity(norm1,'kernel','normal');
[Fnorm2,XInorm2]=ksdensity(norm2,'kernel','normal');

To obtain a smooth Gaussian bell-shaped curve, you should generate a large number of values, while the second and third lines ensure that the Gaussian random numbers will have the same mean and variance as raw portfolio returns (however, by construction **std(RET(:,2)) = 1)**. **[f,xi] = ksdensity(x)** computes a probability density estimate of the sample in the vector x. f is the vector of density values evaluated at the points in xi. The estimate is based on a normal kernel function, using a window parameter (bandwidth) that is a function of the number of points in x. The density is evaluated at 100 equally spaced points that cover the range of the data in x. 'kernel' specifies the type of kernel smoother to use. The possibilities are 'normal' (the default), 'box', 'triangle', 'epanechnikov'. The following lines of codes perform the normal kernel density estimation with reference to the actual data, both raw and standardized:

[F1,XI1]=ksdensity(RET(:,1),'kernel','normal');
[F2,XI2]=ksdensity(RET(:,2),'kernel','normal');

Figure A4 shows the results of this exercise. Clearly, both raw and standardized data deviate from a Gaussian benchmark in the same ways commented early on: tails are fatter (especially the left one); "bumps" in probability in the tails; less probability mass than the normal around $\pm 1/1.5$ standard deviations from the normal, but a more peaked density around the mean.

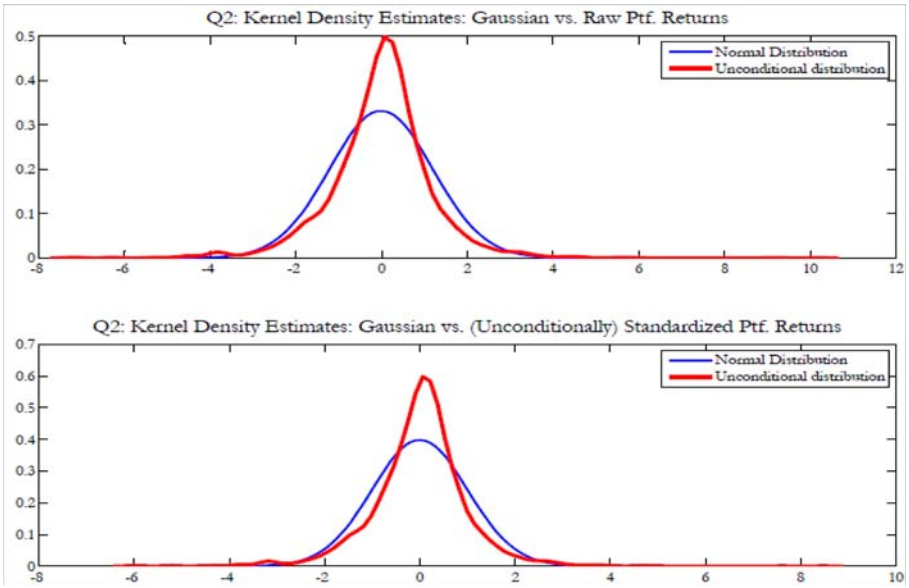

Figure A4:Kernel density estimates: raw and standardized data vs. Normal kernel

Finally, formal Jarque-Bera tests are performed and displayed in Matlab using the following lines of code:

```
[h,p_val,jbstat,critval] = jbtest(port_ret(ind(1):ind(2),1));
[h_std,p_val_std,jbstat_std,critval_std] = jbtest(std_portret);
col1=strvcat(' ','JB statistic: ','Critical val:','P-value:','Reject H0?');
col2=strvcat('RETURNS
',num2str(jbstat),num2str(critval),num2str(p_val),num2str(h));
col3=strvcat('STD. RETURNS',num2str(jbstat_std), ...
...num2str(critval_std),num2str(p_val_std),num2str(h_std));
mat=[col1,col2,col3];
disp(['Jarque-Bera test for normality (5%)']);
```

This gives the following results that, as you would expect, reject normality with a p-value that

is very close to zero (i.e., simple bad luck cannot be responsible for deviations from normality:

```
================ Q2: Test for normality of raw portfolio returns ==============
Jarque-Bera test for normality (5%)
               RETURNS     STD. RETURNS
JB statistic:  4456.6819   4456.6819
Critical val:  5.9709      5.9709
P-value:       0.001       0.001
Reject H0?     1           1
```

3. In our case we have selected GJR-GARCH and NAGARCH with Gaussian innovations as our models. Both are estimated with lines of codes that are similar or identical to those already employed in Lab 1 (second part of the course) and chapter 4. he standardized GJR GARCH standardized returns are computed as:[58]

z_gjr= port_ret(ind(1):ind(2),:)./sigmas_gjr;

The estimate of the two models lead to the following printed outputs:

```
  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    0.0012998      0.016009         0.0812
           K    0.017561       0.0018332        9.5793
    GARCH(1)    0.91313        0.0081071      112.6334
     ARCH(1)    0              0.0074764        0.0000
 Leverage(1)    0.13813        0.012404        11.1357
Stationarity measure 0.9131

NGARCH PARAMETERS
omega  0.0196
alpha  0.0575
theta  1.1277
beta   0.8534
MaxLik 4382.9125
Stationarity measure 0.9840
```

These give no surprises compared to the ones reported in chapter 4, for instance. Figure A5 compares the standardized returns from the GJR and NAGARCH models. Clearly, there are

[58] You could compute standardized residuals, but with an estimate of the mean equal to 0.0013, that will make hardly any difference.

differences, but these seem to be modest at best.

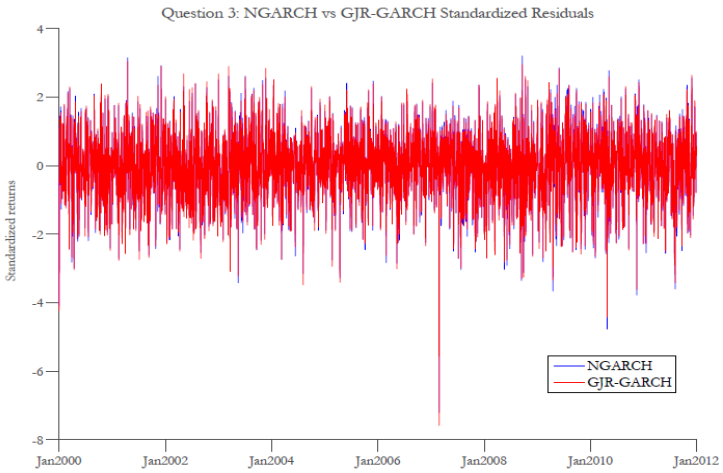

Figure A5: Standardized returns from GJR(1,1) vs. NAGARCH(1,1)

In Figure A6, the QQ plots for both series of standardized returns are compared. While both models seem to fit rather well the right tail of the data, as the standardized returns imply high-order percentiles that are very similar to the normal ones, in the left tail—in fact this concerns at least the first, left-most 25 percentiles of the distribution—the issues emphasized by Figure A3 remain. Also, there is no major difference between the two alternative asymmetric conditional heteroskedastic models.

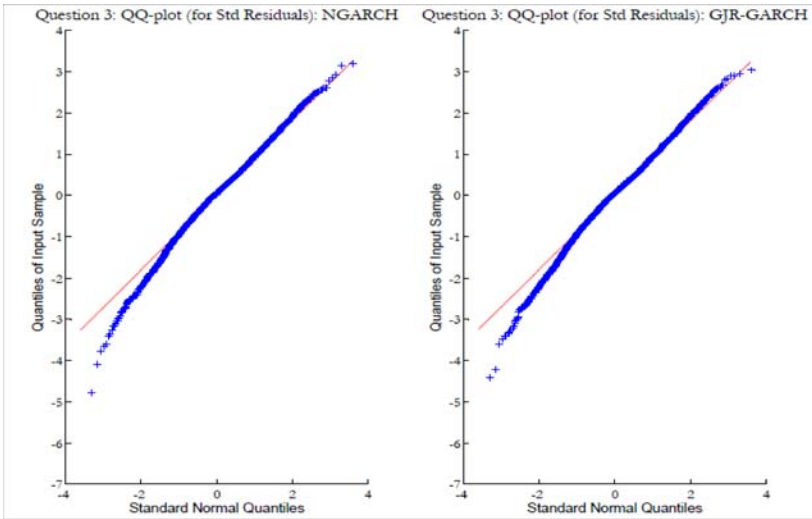

Figure A6: QQ plots for standardized returns of GJR vs. NAGARCH models

Figure A7 shows the same result using kernel density estimators. The improvement vs.

Figure A4 is obvious, but this does not seem to be sufficient yet.

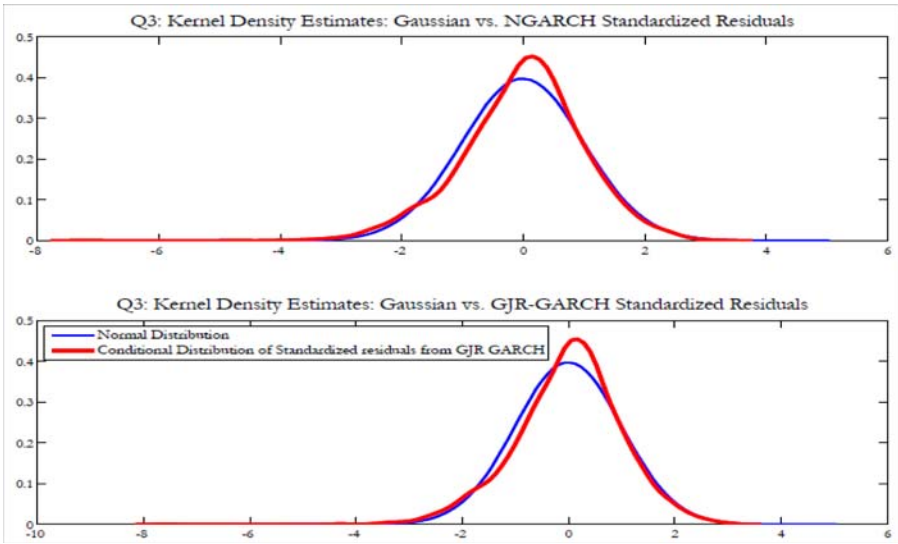

Figure A7: Kernel density estimates of GJR vs. NAGARCH standardized returns

Finally, formal Jarque-Bera tests still lead to rejections of the null of normality of standardized returns, with p-values that remain essentially nil.

```
Jarque-Bera test for normality (5%)
                NGARCH    GJR-Garch
JB statistic:   306.7869  362.0346
Critical val:   5.9709    5.9709
P-value:        0.001     0.001
Reject H0?      1         1
```

4. The point of this question is for you to stop and visualize how "things should look like" if you were to discover the true model that has generated the data. In this sense, the point represents a sort of a break, I believe a useful one, in the flow of the exercise. The goal is to show that if returns actually came from an assumed asymmetric GARCH model with Gaussian innovations such as the ones estimated above, then the resulting (also simulated) standardized returns would be normally distributed. Interestingly, Matlab provides a specific garch-related function to perform simulations given the parameter estimates of a given model:

spec_sim=garchset('Distribution','Gaussian','C',0,'VarianceModel','GJR','P',param_gjr.P,
...'Q',param_gjr.Q,'K',param_gjr.K,'GARCH',param_gjr.GARCH,'ARCH',param_gjr.ARCH,
...'Leverage',param_gjr.Leverage);
[ret_sim, sigma2_sim]=garchsim(spec_sim,length(z_ng),[]);
z_sim=ret_sim./sigma2_sim;

Using **[Innovations,Sigmas,Series] = garchsim(Spec,NumSamples,NumPaths)**, each simulated path is sampled at a length of NumSamples observations. The output consists of the NumSamples × NumPaths matrix 'Innovations' (in which the rows are sequential observations, the columns are alternative paths), representing a mean zero, discrete-time stochastic process that follows the conditional variance specification defined in Spec. The simulations from the NAGARCH model are obtained using:

zt=random('Normal',0,1,length(z_ng),1);
[r_sim,s_sim]=ngarch_sim(param_ng,var(port_ret(ind(1):ind(2),:)),zt);

where 'random' is the general purpose random number generator in Matlab and 'ngarch_sim(par,sig2_0,innov)' is our customized procedure that takes the NGARCH 4x1 parameter vector (omega; alpha; theta; beta), initial variance (sig2_0), and a vector of innovations to generate a number ind(1)-ind(2) of simulations. Figure A8 shows the QQ plots for both returns and standardized returns generated from the GJR GARCH(1,1) model.

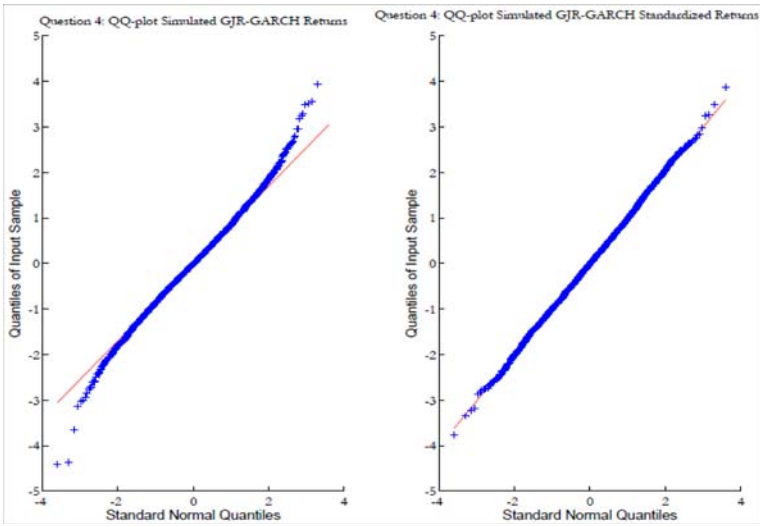

Figure A8: QQ Plots for raw and standardized GJR GARCH(1,1) simulated returns

The left-most plot concerns the raw returns and makes a point already discussed in chapter 4: if the model is

$$R_{t+1} = \left(\sqrt{\omega + \alpha R_t^2 + \theta I_{\{R_t<0\}} + \beta\sigma_t^2}\right) z_{t+1} \qquad z_{t+1} \text{ IID } \mathcal{N}(0,1),$$

then you know that even though z_{t+1} IID $\mathcal{N}(0,1)$, R_{t+1} will not be normally distributed, as shown to the left of Figure A8. The righ-most plot concerns instead

$$z_{t+1} \equiv \frac{R_{t+1}}{\sqrt{\omega + \alpha R_t^2 + \theta I_{\{R_t<0\}} + \beta\sigma_t^2}} \text{IID } \mathcal{N}(0,1),$$

and shows that normality approximately obtains.[59] Figure A9 makes the same point using not QQ plots, but normal kernel density estimates.

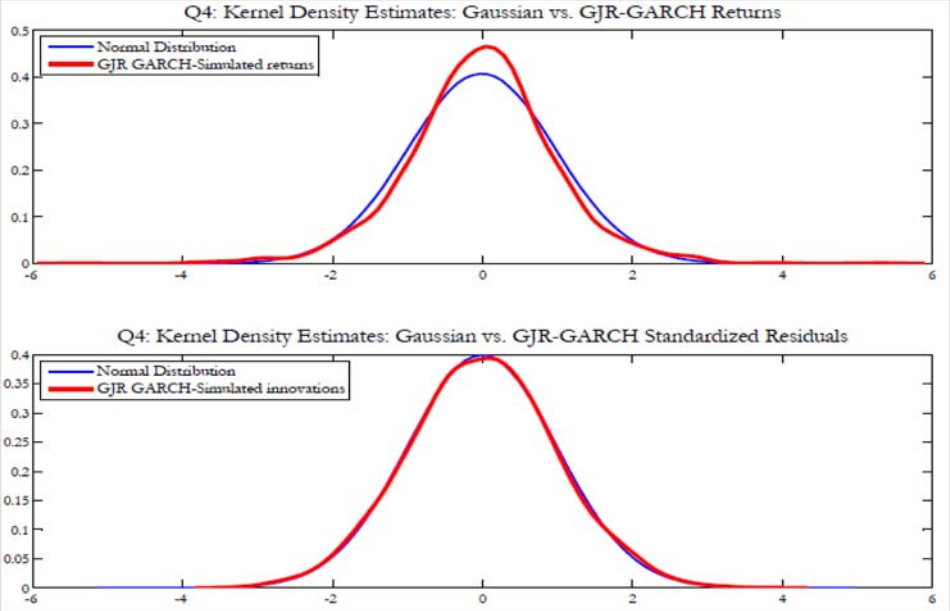

Figure A9: Normal kernel density estimates applied to raw and standardized GJR simulated returns

Figures A10 and A11 repeat the experiment in Figures A8 and A9 with reference to simulated returns and hence standardized returns from the other asymmetric model, a NAGARCH. The lesson they teach is identical to Figures A8 and A9.

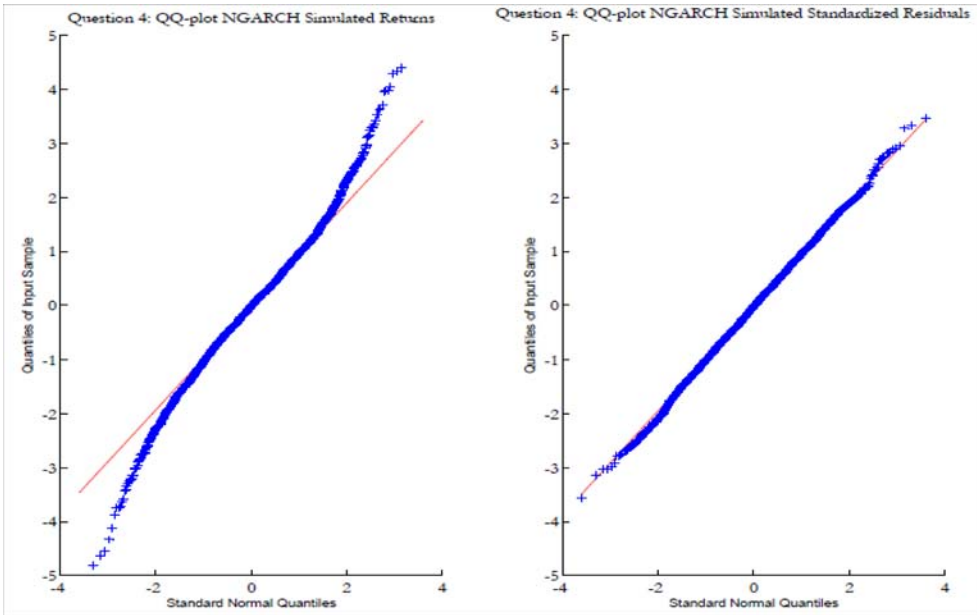

Figure A10: QQ Plots for raw and standardized NAGARCH(1,1) simulated returns

[59] Why only approximately? Think about it.

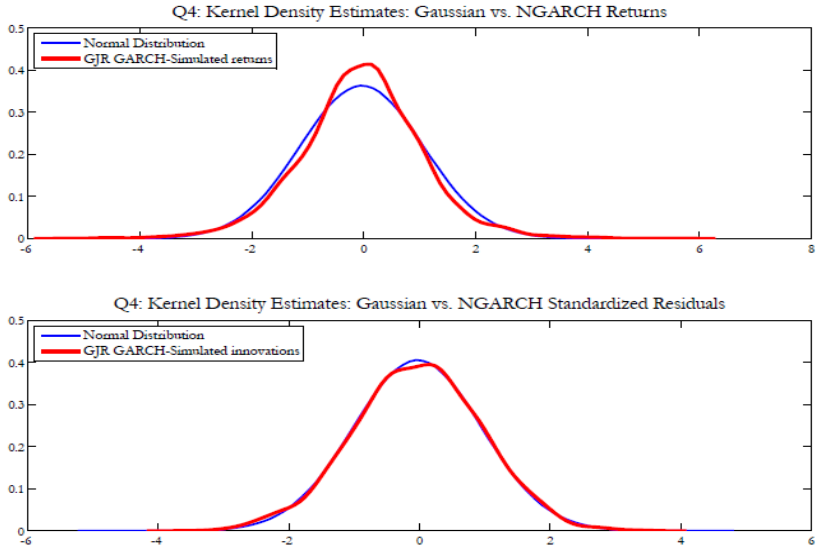

Figure A11: Normal kernel density estimates applied to raw and standardized NAGARCH simulated returns

Formal Jarque-Bera tests confirm that while simulated portfolio returns cannot be normal under an asymmetric GARCH model, they are—and by construction, of course—after these are standardized.

```
Jarque-Bera test for normality of GJR-GARCH(5%)
                RETURNS (SIM)    STD RET (SIM)
JB statistic:   123.0307         2.6723
Critical val:   5.9709           5.9709
P-value:        0.001            0.25744
Reject H0?      1                0


Jarque-Bera test for normality of NGARCH(5%)
                RETURNS (SIM)    STD RET (SIM)
JB statistic:   335.5308         1.4493
Critical val:   5.9708           5.9709
P-value:        0.001            0.48018
Reject H0?      1                0
```

5. Although the objective of this question is to compute and compare VaRs computed under a variety of methods, this question implies a variety of estimation and calculation steps. First, the estimation of the degrees of freedom for a standardized t-Student is performed via quasi maximum likelihood (i.e., taking the GJR standardized residuals as given, which means that the estimation is split in two sequential steps):

cond_std=sigmas_gjr;

df_init=4; %This is just an initial condition

[df,qmle]=fminsearch('logL1',df_init,[],port_ret(ind(1):ind(2),:),cond_std);

VaR_tstud=-for_cond_std_gjr'.*q_tstud;

where **df_init** is just an initial condition, and the QMLE estimation performed with **fminsearch** calling the used-defined objective function **logL1_asym** that takes as an input **df**, the number of degrees of freedom, the vector of returns **ret**, and **sigma**, the vector of filtered time-varying standard deviations. You will see that Matlab prints on your screen an estimate of the number of degrees of freedom that equals 10.342 which marks a non-negligible departure from a Gaussian benchmark. The VaR is then computed as:

q_norm=inv;

q_tstud=sqrt((df-2)/df)*tinv((p_VaR),df);

Note that the standardization adjustment discussed during the lectures, $Var(z) = df/(df-2)$, which means that z is not standardized; it is then obvious that if you produce inverse t-value critical points from a standardized t-Student—as **tinv((p_VaR))** does—then you have to adjust the critical value by de-standardizing it, which is done dividing it by $sqrt(df/(df-2))$, that is multiplying by $sqrt((df-2)/df)$.

The estimation of the Cornish-Fisher expansion parameters and the computation of VaR is performed by the following portion of code:

zeta_1=skewness(z_gjr);

zeta_2=kurtosis(z_gjr)-3;

inv=norminv(p_VaR,0,1);

q_CF=inv+(zeta_1/6)*(invˆ2-1)+(zeta_2/24)*(invˆ3-3*inv)-(zeta_1ˆ2/36)*(2*(invˆ3)-5*inv);

VaR_CF=-for_cond_std_gjr'.*q_CF;

Figure A12 plots the behavior of 5 percent VaR under the four alternative models featured

by this question.

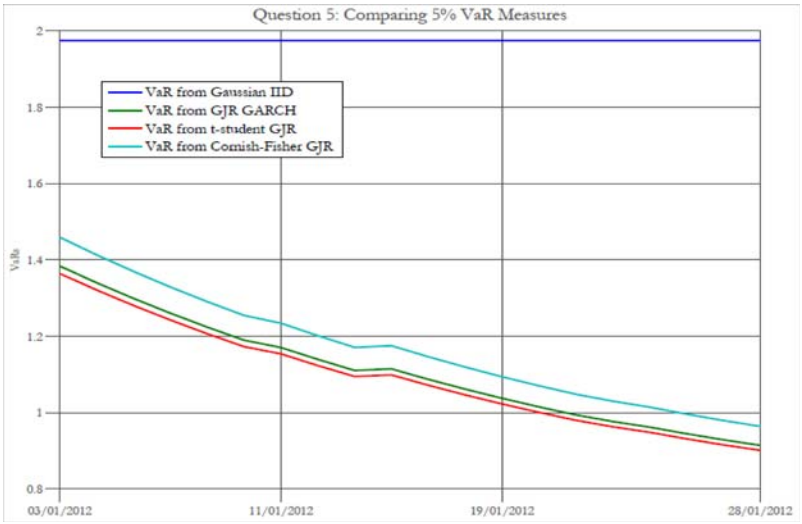

Figure A12: 5% VaR under alternative econometric models

Clearly, VaR is constant under a homoskedastic, constant variance model. It is instead time-varying under the remaining models, although these all change in similar directions. The highest VaR estimates are yielded by the GJR GARCH(1,1) models, quite independently of the assumption made on the distribution of the innovations (normal or t-Student). The small differences between the normal and t-Student VaR estimates indicate that at a 5% level, the type of non-normalities that a t-Student assumption may actually pick up remain limited, when the estimated number of degrees of freedom is about 10.[60] Finally, the VaR computed under a CF approximation is considerably higher than the GJR GARCH VaR estimates: this is an indication of the presence of negative skewness in portfolio returns that only a CF approximation may capture. Figure A12 emphasizes once more the fact that adopting more complex, dynamic time series models is not always leading to higher VaR estimates and more prudent risk management: in this example—also because volatility has been declining during early 2012, after the Great Financial crisis and European sovereign debt fears—constant variance models imply higher VaR estimates than richer models do.[61]

6. Starting from an initial condition **df_init=10**, QML estimates of a NAGARCH with standardized t(d) innovations is performed by:

[df,qmle]=fminsearch('logL1',df_init,[],port_ret(ind(1):ind(2),:),sqrt(cond_var_ng));

[60] This also derives from the fact that a 5 percent VaR is not really determined by the behavior of the density of portfolio returns in the deep end of the left tail. Try and perform calculations afresh for a 1 percent VaR and you will find interesting differences.

[61] Of course, lower VaR, lower capital charges and capital requirements.

where cond_var_ng is taken as given from question 3 above. The QML estimate of the number of degrees of freedom is 10.342. The resulting QQ plot is shown in Figure A13: interestingly, compared to Figure A6 where the NAGARCH innovations were normally distributed, marks a strong improvement in the left tail, although the quality of the fit in the right tail appears inferior to Figure A6.

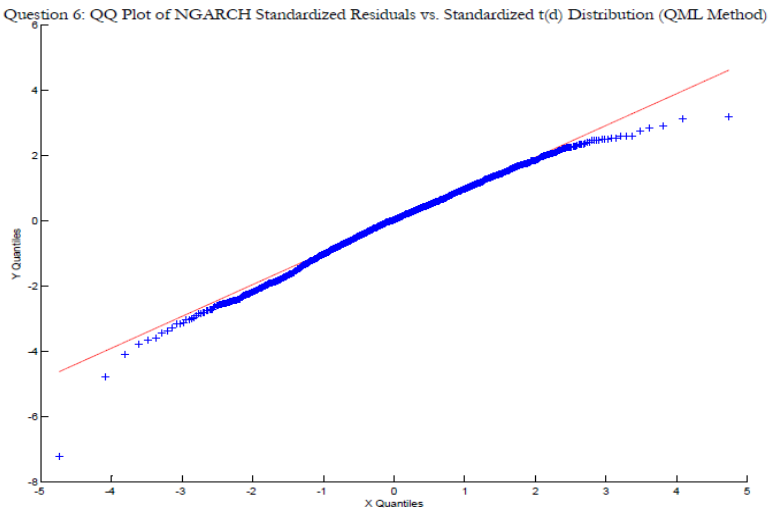

Figure A13: QQ plot of QML estimate of t-Student NAGARCH(1,1) model

Interestingly, Figure A13 displays a QQ plot built from scratch and not using the Matlab function, using the following code:

```
z_ngarch=sort(z_ng);
z=sort(port_ret(ind(1)-1:ind(2)-1,:));
[R,C]=size(z);
rank=(1:R)';
n=length(z);
quant_tstud=tinv(((rank-0.5)/n),df);
cond_var_qmle=cond_var_ng;

qqplot(sqrt((df-2)/df)*quant_tstud,z_ngarch);
set(gcf,'color','w');
title('Question 6: QQ Plot of NGARCH Standardized Residuals vs. Standardized
t(d) Distribution (QML Method)','fontname','garamond','fontsize',15);
```

The full ML estimation is performed in ways similar to what we have already described above.

The results are:

```
NGARCH estimated parameters (assuming std. t innovations):
omega  0.016
alpha  0.058
theta  1.145
beta   0.854
t~ d.f 10.169
The implied persistence of the ML estimate of the t(d)-NGARCH(1,1) model is:
Persistence:  0.989
```

and shows that the full ML estimation yields a 10.17 estimate that does not differ very much from the QML estimate of 10.34 commented above.[62] The corresponding QQ plot is in Figure A14 and is not materially different from Figure A13, showing that often—at least for practical purposes—QMLE gives results that are comparable to MLE.

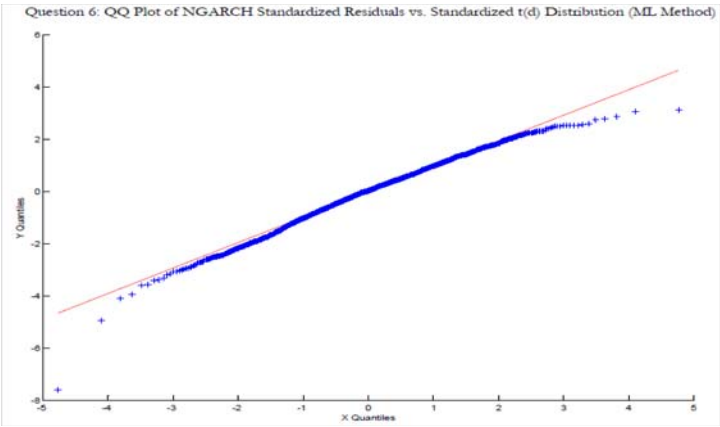

Figure A14: QQ plot of ML estimate of t-Student NAGARCH(1,1) model

Figures A15 and A16 perform the comparison between the filtered (in-sample) conditional volatilities from the two sets of estimates—QML vs. ML—of the t-Student NAGARCH (A15)

[62] No big shock: although these are numerically different, you know that the real diffence between QMLE and MLE consists in the lack of the efficiency of the former when compared to the latter. However, in this case we have not computed and reported the corresponding standard errors.

and among the t-Student NAGARCH and a classical NAGARCH with normal innovations.

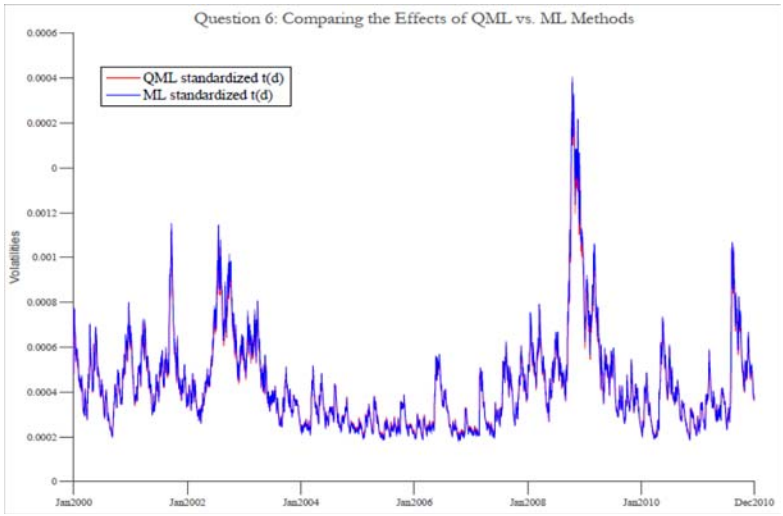

Figure A15: Comparing filtered conditional volatilities across QML and ML t-Student NAGARCH

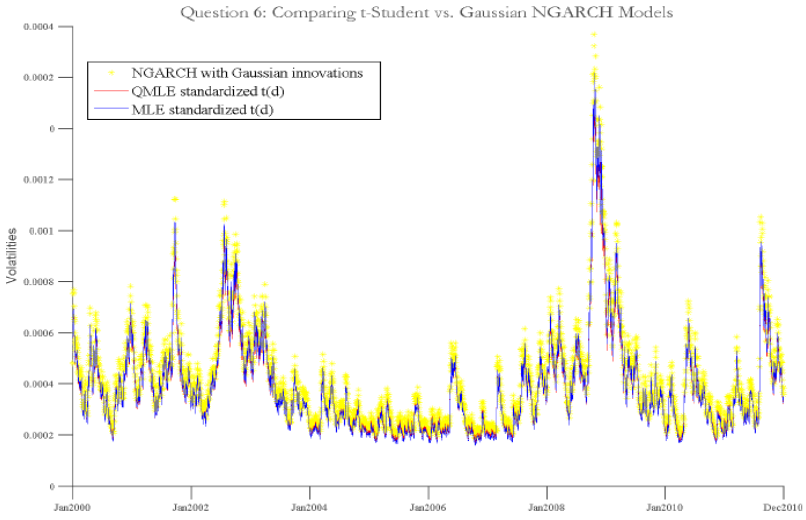

Figure A16: Comparing conditional volatilities across QML and ML t-Student vs. Gaussian NAGARCH

Interestingly, specifying t-Student errors within the NAGARCH model systematically reduces conditional variance estimates, vs. the Gaussian case. Given our result in Section 4 that

$$\hat{\sigma}^2 = \widehat{\widehat{m}}_2 \frac{\hat{d}-2}{\hat{d}},$$

when $\hat{d}$ is relatively small, $\hat{\sigma}^2$ tends to be smaller than a pure, ML-type sample-induced estimate of σ^2.

7. The lines of code that implement the EVT quantile estimation through Hill's estimation are:

```
p_VaR=0.0001;
std_loss=-z_ng;
[sorted_loss I]=sort(std_loss,'descend');
u=quantile(sorted_loss,0.96);      % This is the critical threshold choice
tail=sorted_loss(sorted_loss>u);
Tu=length(tail);
T=length(std_loss);
xi=(1/Tu)*sum(log(tail./u));
% Quantiles
q_EVT=u*(p_VaR./(Tu/T)).^(-xi);
```

The results are:

```
============  Exercise 7: Extreme Value Theory (EVT) VaR Estimates vs. MLE
Estimated VaRs, p= 0.0001
Normal NGARCH          3.342
Std-T  NGARCH          4.523
Cornish Fisher         5.712
Extreme Value (Hill)   6.756
```

and at such a small probability size of the VaR estimation, the largest estimate is given by the EVT, followed by the Cornish-Fisher approximation. The partial EVT QQ plot is shown in Figure A17 and shows excellent fit in the very far left tail.

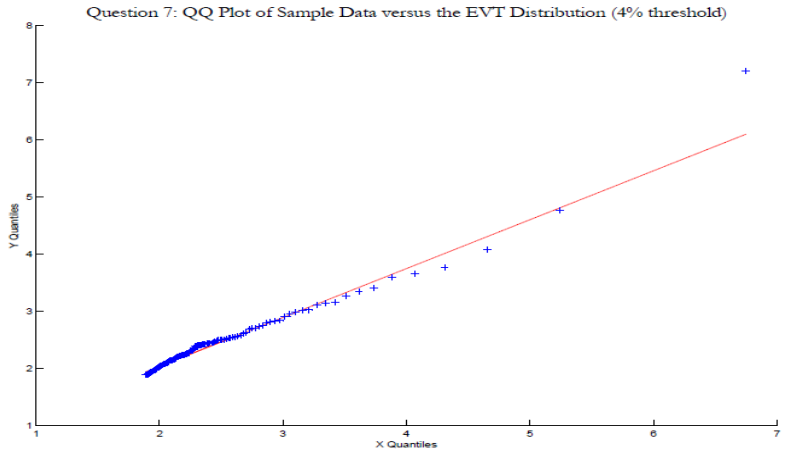

Figure A17: Partial QQ plot (4% u threshold)

However, if we double to 8% the u threshold used in the Hill-type estimation, the partial QQ plot results in Figure A18 are much less impressive. The potential inconsistency of the density fit provided by the EVT approach in dependence of a choice of the parameter u has been discussed in Chapter 6.

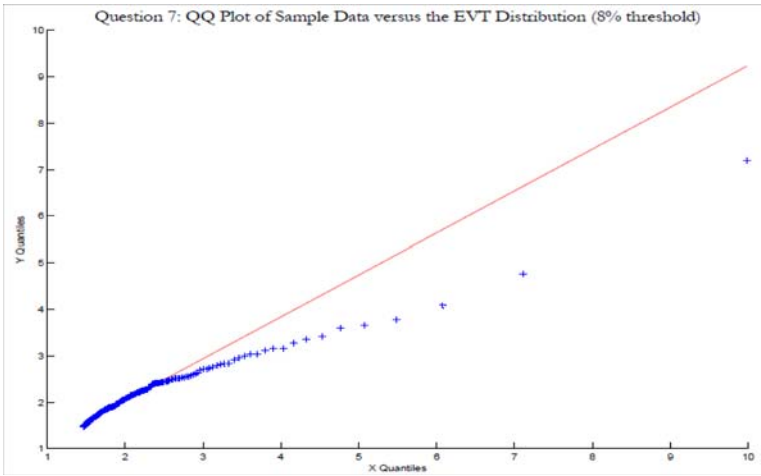

Figure A18: Partial QQ plot (8% u threshold)

8. The estimation of conditional mean and variance under model 8.a (Constant mean and GARCH (1,1) assuming normally distributed innovations) are performed using

[coeff_us1,errors_us1,sigma_us1,resid_us1,Rsqr_us1,miu_us1]=

regres-

sion_tool_1('GARCH','Gaussian',ret1(2:end,1),[ones(size(ret1(2:end,1)))],1,1,n);

[coeff_uk1,errors_uk1,sigma_uk1,resid_uk1,Rsqr_uk1,miu_uk1]=

regres-

sion_tool_1('GARCH','Gaussian',ret1(2:end,2),[ones(size(ret1(2:end,2)))],1,1,n);

[coeff_ger1,errors_ger1,sigma_ger1,resid_ger1,Rsqr_ger1,miu_ger1]=

regres-

sion_tool_1('GARCH','Gaussian',ret1(2:end,3),[ones(size(ret1(2:end,3)))],1,1,n);

The estimation of conditional mean and variance under model 8.b (Constant mean and EGARCH (1,1) assuming normally distributed innovations) is similar (please see the code). Finally, conditional mean and variance estimation for model 8.c (constant mean and EGARCH (1,1) model assuming Student-t distributed innovations) are performed with the code:

[coeff_us3,errors_us3,sigma_us3,resid_us3,Rsqr_us3,miu_us3]=

regression_tool_1('EGARCH','T',ret1(2:end,1),[ones(size(ret1(2:end,1)))],1,1,n);

[coeff_uk3,errors_uk3,sigma_uk3,resid_uk3,Rsqr_uk3,miu_uk3]=

regression_tool_1('EGARCH','T',ret1(2:end,2),[ones(size(ret1(2:end,2)))],1,1,n);

[coeff_ger3,errors_ger3,sigma_ger3,resid_ger3,Rsqr_ger3,miu_ger3]=

regression_tool_1('EGARCH','T',ret1(2:end,3),[ones(size(ret1(2:end,3)))],1,1,n);

regression_tool_1 is used to perform recursive estimation of simple GARCH models (please check out its structure by opening the corresponding procedure). The unconditional correlations are estimated and appropriate covariance matrices are built using:

```
corr_un1=corr(std_resid1); %Unconditional correlation of returns for model under 8.a
corr_un2=corr(std_resid2); %Unconditional correlation of residuals from model under 8.b
corr_un3=corr(std_resid3);

T=size(ret1(2:end,:),1);
cov_mat_con1=NaN(3,3,T); %variances and covariances
cov_mat_con2=NaN(3,3,T);
cov_mat_con3=NaN(3,3,T);
for i=2:T
cov_mat_con1(:,:,i)=diag(sigma1(i-1,:))*corr_un1*diag(sigma1(i-1,:));
cov_mat_con2(:,:,i)=diag(sigma2(i-1,:))*corr_un2*diag(sigma2(i-1,:));
cov_mat_con3(:,:,i)=diag(sigma3(i-1,:))*corr_un3*diag(sigma3(i-1,:));
end
```

The asset allocation (with no short sales and limited to risky assets only) is performed for each of the three models using the function **mean_variance_multiperiod** that we have used already in chapter 4. Figure A19 shows the corresponding results.

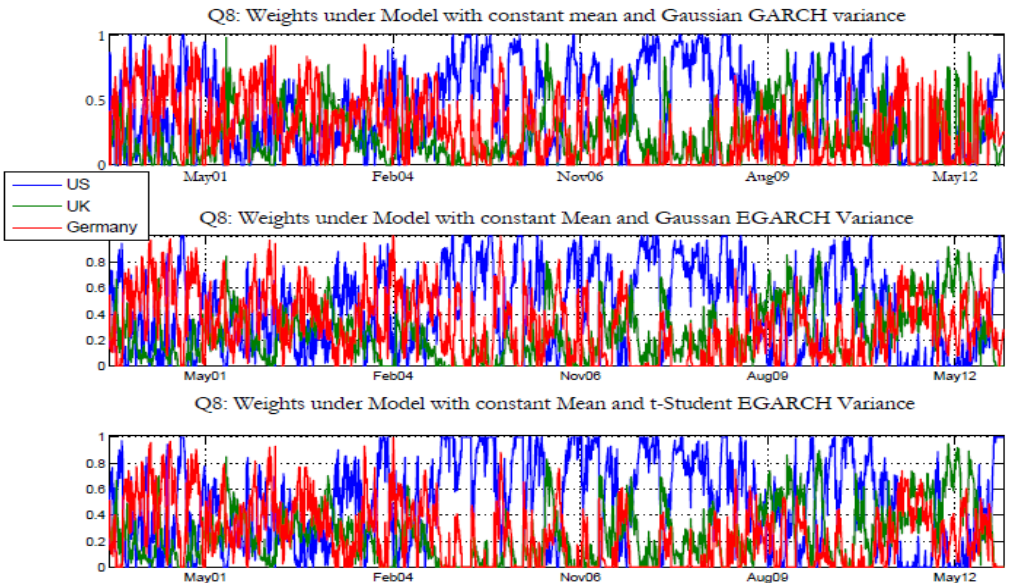

Figure A19: Recursive mean-variance portfolio weights ($\gamma = 0.5$) from three alternative models

Clearly, there is considerable variation over time in the weights that—although different if one carefully inspects them—are eventually characterized by similar dynamics over time, with an average prevalence of U.S. stocks. Figure A20 shows the resulting, in-sample realized Sharpe ratios using a procedure similar to the one already followed in chapter 4.

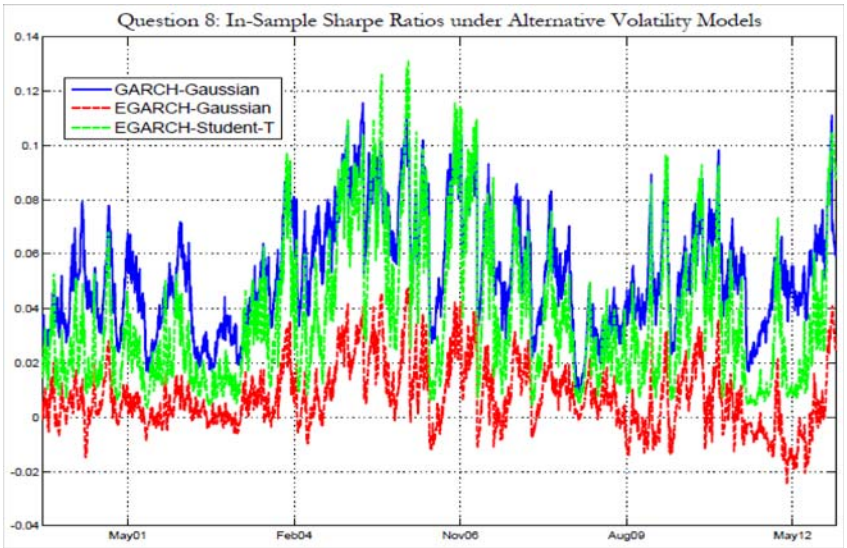

Figure A20: Recursive realized Sharpe ratios from mean-variance portfolio weights ($\gamma = 0.5$) from three models

References

[1] Artzner, P., Delbaen, F., Eber, J., and Heath, D., 1999. Coherent measures of risk. *Mathematical Finance* 9, 203-228.

[2] Bollerslev, T., 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. *Review of Economic Statistics* 69, 542-547.

[3] Davis, C., and Stephens, M., 1989. Empirical distribution function goodness-of-fit tests. *Applied Statistics* 38, 535-582.

[4] Huisman, R., Koedijk, K., Kool, C., Palm, F., 2001. Tail-index estimates in small samples. *Journal of Business and Economic Statistics* 19, 208-216.

[5] Jaschke, S. 2002. The Cornish-Fisher-Expansion in the context of Delta-Gamma-Normal approximations. *Journal of Risk*, Summer 2002.

[6] McNeil, A. 1998. Calculating quantile risk measures for financial return series using Extreme Value Theory, ETH Zentrum, working paper.

[7] Teräsvirta T., 2009. An Introduction to Univariate GARCH Models, in Andersen, T., Davis, R., Kreiß, J.-P., and Mikosch, T., *Handbook of Financial Time Series*, Springer.

Errata Corrige

(30/04/2013, p. 8) The sentence in the second equation from top of the page should read as "Fraction of your data equal to x.", not x_i.

(30/04/2013, p. 10) Towards the end of the page, the sentence should read as "This means that the right tail of the empirical distribution of S&P 500 returns is *thicker* than the normal tail".

(30/04/2013, p. 14) A new footnote 21 has been added to explain what the model of reference is at pp. 14-16.

(30/04/2013, p. 15) A -3 has been added in the equation providing the moment matching condition for ζ_2 and one spurious equal sign removed from $\sigma^2 \frac{d}{d-2} = \widehat{\widehat{m}}_2$.

(30/04/2013, p. 46 and workout Matlab code posted on the web) The formula $\alpha(1+0.5\theta)+\beta$ has been now used to compute the GJR stationarity measure (there would be reasons not to, but it is easier this way; thanks M. Fiorani-Gallotta for pointing out the insidious inconsistency). In this case, $0(1 + 0.5 \times 0.1381) + 0.9131 = 0.9131$, of course.

(07/05/2013, p. 8) In equation (4) the pedices labelling the two kernel densities as "Box" and "Triangular" have been switched.

(07/05/2013, p. 23) $\tilde{t}_p^{-1}(6.70)$ should be $t_p^{-1}(6.70)$.

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] Enders, W., 2004. *Applied Econometric Time Series*. John Wiley, New York.

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

[7] 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.

[8] 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.

[9] 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.

[10] Hall, P., and P., Yao, 2003. "Inference in ARCH and GARCH Models with Heavy-Tailed Errors", *Econometrica*, 71, 285-317.

[11] Lee, S., W., and B., E., Hansen, 1994. "Asymptotic Properties of the Maximum Likelihood Estimator and Test of the Stability of Parameters of the GARCH and IGARCH Models", *Econometric Theory*, 10, 29-52.

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

Errata Corrige

(30/04/2013, p. 35) One incorrect notation $l_t \equiv \Pr(z_t; \boldsymbol{\theta})$ has been replaced with the correct $l_t \equiv \Pr(R_t; \boldsymbol{\theta})$.

(30/04/2013, p. 36) 2 incorrect occurrence of $\alpha\beta$ that should have been simply β have been fixed.

(30/04/2013, p. 48) Equation (5) should read as

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

i.e., the last q_t should be such and not q_{t+1}.

(18/05/2013, p. 24) Figure 7 refers to per-period forecasts of variance as a function of H, not to total variance between $t+1$ and $t+H$.