Back-testing conditional volatility models

By Céline Detilleux, Antoine Soetewey and Vincent Starck*

Data from the Swiss Stock Exchange's returns have been analyzed to describe the prediction accuracy of Value-at-Risk and Expected Shortfall when they are estimated using EGARCH and GARCH models. Tests suggest that VaR and ES forecasts based on these conditional volatility models are valid in most cases. The independence of violations, when tested by the means of Christoffersen's test, seems to be satisfied for all models while the number of violations is usually a bit high but not enough to lead to a definitive rejection. The results of back-testing procedures appear to depend on the choice of estimation and forecast windows, leading sometimes to different model selections. Moreover, whether it is VaR or ES that is backtested may have an influence on the model chosen.

I. Introduction

"In the late 1970s and 1980s, a number of major financial institutions started work on internal models to measure and aggregate risks across the institution as a whole" (Dowd, 2007). Their works led to several risk measures, the most popular and commonly used being the Value-at-Risk (VaR). Despite its popularity, this risk measure has been widely critized recently. As a response, the Expected Shortfall (ES) has been proposed as a better alternative to the VaR. Constructing estimates of the VaR and the ES can be done in various ways. A common choice is to use conditional volatility models such as GARCH to predict the volatility of the market.

Once a risk measure system has been developed, an important issue is that of assessing the quality of its forecast. Danielsson et al. (2001) discuss inaccuracy of these risk measures. A classical way to judge the validity of the models is the backtesting procedure, which is the application of statistical methods where actual profits and losses are systematically compared to corresponding VaR estimates.

This paper discusses the back-testing of VaR and ES estimates based on several popular conditional volatility models. The influence of the sample on backtesting performances is evaluated by taking different subsamples of the returns of the

* Céline Detilleux (I6051088): Maastricht University, c.detilleux@student.maastrichtuniversity.nl. Antoine Soetewey (I6083256): Maastricht University, a.soetewey@student.maastrichtuniversity.nl. Vincent Starck (I6122010): Maastricht University, v.starck@student.maastrichtuniversity.nl.

Swiss Stock Exchange during the last two decades.

The paper is organized as follows. Section II describes the theoretical framework that we will use for empirical analysis. Section III will briefly introduce conditional volatility models. In section IV the data, the econometric methodology and the results are provided. Finally, section V concludes.

II. Theoretical Background

A. Measuring risk

Financial risk is the prospect of financial loss - or gain - due to unforeseen changes in under-lying risk factors (Dowd, 2007). Usually, risk is summarized by a number called a risk measure. To elaborate, let the value of the portfolio at the end of day be denoted by P_{t-1} and the one at the end of tomorrow by P_t. Then, obviously, a P_t greater than P_{t-1}, leads to profits equal to the difference between P_t and P_{t-1} while a P_t lower than P_{t-1} leads to losses equal to the difference between P_{t-1} and P_t. Since P_t is uncertain then so is the profit or loss (P/L). The next-period P/L is risky; risk measures are framework to measure this risk.

The purpose of the study is to assess the validity of risk measures based on conditional volatility model. The evaluation is done using back-testing methods, which are used to determine whether the forecasts of a conditional volatility model are consistent with the assumptions on which the model is based (Dowd, 2007).

B. Coherent risk measures

A set of properties that a suitable risk measure should possess have been proposed by Artzner et al. (1999). Risk measures that satisfy the following four axioms are called *coherent*:

Axiom 1. Subadditivity. *For all* X_1 *and* $X_2 \in \mathcal{G}$, $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$.

Subadditivity embodies the reduction of risks associated with diversification. This axioms means that the risk related to two different instruments taken together should be less than or equal to the sum of the risk associated with each of the two instruments taken individually. This is particularly relevant for instruments that are negatively correlated since the risk associated to one of the two assets is, to some extent, canceled out by the other.

Axiom 2. Positive homogeneity. *For all* $\lambda \geq 0$ *and all* $X \in \mathcal{G}$, $\rho(\lambda X) = \lambda\rho(X)$.

This axiom suggests that the risk of a position is proportional to its size.

Axiom 3. Monotonicity. *For all X and $Y \in \mathcal{G}$ with $X \leq Y$, we have $\rho(Y) \leq \rho(X)$.*

This axiom indicates that if portfolio Y has better values than portfolio X then the risk associated to portfolio Y should be less than or equal to the risk related to portfolio X. This also entails the fact that a position with greater future returns is less risky.

Axiom 4. Translation invariance. *For all $X \in \mathcal{G}$ and all real numbers α, we have $\rho(X + \alpha \cdot r) = \rho(X) - \alpha$.*

This axiom implies that the risk measure decreases by α when adding a sure amount of α invested in the reference instrument, whose return is denoted by r, to the initial portfolio.

Any measure of risk which violates one of the axioms will generate inaccurate results, leading to an erroneous assessment of relative risks.

C. Popular risk measures

As explained above, the paper looks at two very popular risk measures, the VaR and the ES, which are briefly described below. Since both measure different things, their value can differ substantially and they have different theoretical properties. For both of them, the computation hinges on the stochastic process that is assumed for the returns.

Value-at-Risk

First, let's look at the VaR, which was popularized by JP Morgan in 1994. Given a certain portfolio and time horizon, the VaR is defined as the maximum amount that can be lost given a certain confidence level. It basically looks at the quantile of the return distribution, that is, for a continuous distribution of the returns:

$$Pr[r_t < VaR_t(\alpha)] = \alpha \tag{1}$$

where r_t are the returns at time $t = 1, \ldots, T$ and α represents the worst $\alpha\%$ of outcomes (confidence level). VaR can also be represented by Figure 1, where we chose an α equal to 5%. VaR has two main attractions; it provides common consistent measure of risk across different positions and risk factors and it takes account of the correlations between different risk factors. Common consistent measure of risk means that risks associated with a fixed-income position is measured in a way that is comparable to and consistent with the way risks associated with equity positions are measured. On the other hand, the correlations between risk factors must be taken into account because if two risks offset each other, the

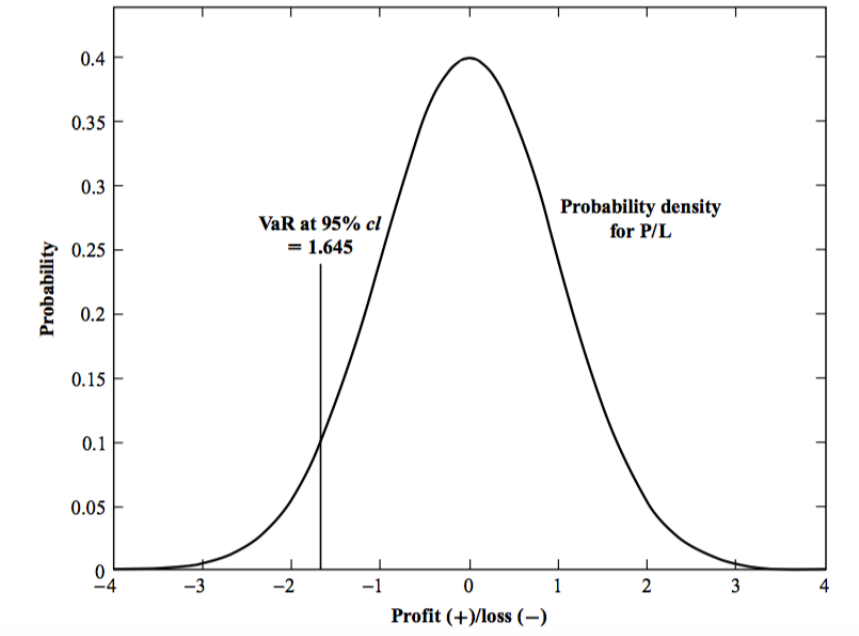

Figure 1. : Value-at-Risk

VaR allows for this offset.

However, VaR is also subject to criticisms, mainly due to its lack of coherence and its incapacity to capture tail risk.

Danielsson et al. (2001) describe several pitfalls of the model. In particular, they mention that VaR assumes elliptically distributed returns while existing databases show that the distribution of operational risks are heavy-tailed. Due to non-elliptical distribution of the returns, VaR fails to satisfy the subadditivity axiom (Embrechts, 2000). This implies that the VaR of a portfolio might be larger than the sum of VaRs of the individuals assets, in contradiction with the intuition behind diversification (Acerbi, Nordio and Sirtori, 2001; Artzner et al., 1997, 1999; Rootzen and Klüppelberg, 1999). Therefore, VaR is not a coherent risk measure outside the class of elliptical distributions.

VaR's other major limitation, mentioned by both Danielsson et al. (2001) and Dowd (2007), is its incapacity to describe the shape of the tail beyond the lower $\alpha\%$ quantile. The VaR only provides the point estimate of the loss distribution, however, the interest is on the distribution of the loss given that a certain extremal threshold has been breached. By looking at Figure 1, it can be seen that VaR says nothing about the shape of the probability density for P/L at

its left. Firms are concerned by this pitfalls because 'spike-the-firm' events (low probability, high loss) are very difficult to capture with VaR methods. Moreover, since the tail shape can differ substantially for a given quantile, firms could alter their risk profile while keeping their VaR constant, implying that setting capital requirements based on VaR may not be very wise.

Expected Shortfall

Considering the various VaR shortcomings, especially its inability to capture 'tail risk' and its unfitness to determine capital requirement, the Basel Committee proposed in May 2012 to replace VaR with ES (Basel Committee on Banking Supervision, 2013). The ES is the expected loss from a portfolio in the worst $\alpha\%$ of cases:

$$ES_t(\alpha) = E[r_t | r_t < VaR_t(\alpha)] \tag{2}$$

which in turn can be expressed as a function of VaR (Acerbi and Tasche, 2002):

$$ES_t(\alpha) = \alpha^{-1} \int_0^{\alpha} VaR_t(u)du. \tag{3}$$

Furthermore, the integral above can be approximated as the average of VaRs at different levels:

$$ES_t(\alpha) \approx [VaR_t(\alpha) + VaR_t(0.75\alpha) + VaR_t(0.5\alpha) + VaR_t(0.25\alpha)] \tag{4}$$

ES can be represented graphically, as in Figure 2 (similar to Figure 1) where the x-axis has been multiplied by (-1). As mentioned before, a major pitfall of the VaR is its incapacity to capture 'tail risk', therefore the ES has been computed to determine the loss in a tail event. Moreover, ES is a coherent risk measure, meaning that it satisfies the sub-additivity, positive homogeneity, monotonicity and translation invariance axioms. In contrast to the VaR, it does not discourage risk diversification.

ES has thus some theoretical advantages over VaR. However, it still has some drawbacks because it may be cumbersome to evaluate and some issues related to back-testing have been reported in the literature. These are detailed in the next subsection.

D. Backtesting

According to Jorion et al. (2007), "backtesting is a statistical testing framework that consists of checking whether actual trading losses are in line with forecasts." Backtesting is used to check the performance of the models. The Basel Committee, which had identified a number of weakness measurement under the models-based approach, decided on a set of quantitative tools to measure the performance

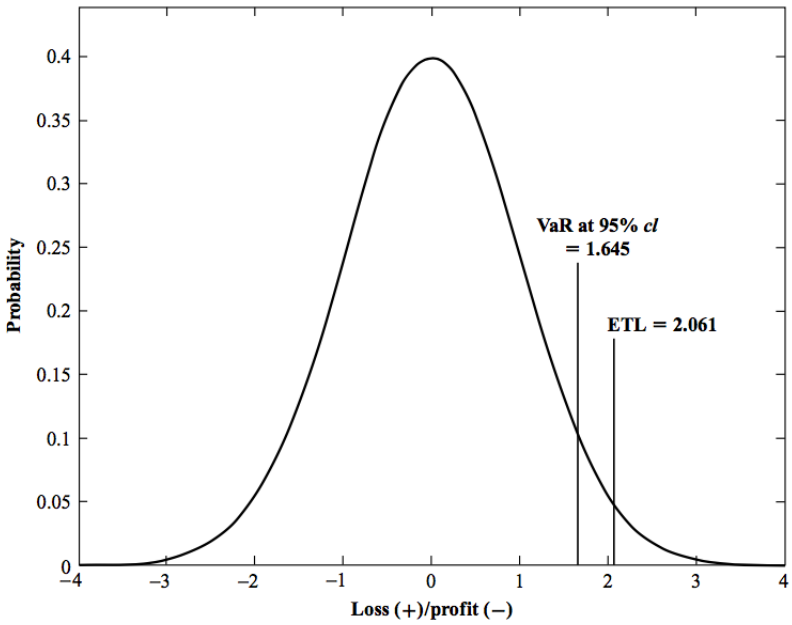

Figure 2. : Expected Shortfall

of the models (Basel Committee on Banking Supervision, 2013). One of them is the backtesting framework which compares forecasted losses with actual ones.

For the VaR, various tests have been proposed in the literature, some of which are detailed in the section. However, due to absence of elicitability, the ES cannot be directly back-tested (Carver, 2013; Gneiting, 2012), which led to many criticism in 2011. A statistics $\psi(Y)$ of a random variable Y is said to be *elicitable* if it minimizes the expected value of a scoring function S:

$$\psi = \operatorname*{argmin}_x E[S(x, Y)] \tag{5}$$

Bellini and Rosazza Gianin (2013) introduced expectiles, which are the minimizers of an asymmetric quadratic loss functions. This risk measure is both coherent and elicitable.

A few month later, Emmer, Kratz and Tasche (2013) argue that evidence is not strong enough to justify a replacement of ES by expectiles and suggest an alternative way to back-test ES. Noting that ES can be expressed as an integral of VaRs (equation 1), they argue that the ES can be considered successfully back-tested to some extent if the VaR is successfully back-tested at different fractions of the initial level. Note that the ES's back-test may require a higher number of

observations since it amounts to back-testing the VaR at fractions of the initial level. As Acerbi and Szekely (2014) point out, elicitability has actually to do with model selection and not with model testing, and is therefore irrelevant for the choice of a regulatory risk standard.

Kupiec and Christoffersen's tests

The idea behind the Kupiec (1995) test is to test whether the observed frequency of tail losses (or frequency of losses that exceed VaR) is consistent with the frequency of tail losses predicted by the model. Christoffersen (1998) looks in addition at the independence of violations. A good model delivers both the right frequency of failures and failures independent of each others.

Consider a sample path $\{y_t\}_{t=1}^T$ of the time series y_t. Let$\{[L_{t|t-1}(p); U_{t|t-1}(p)]\}_{t=1}^T$ where $L_{t|t-1}(p)$ and $U_{t|t-1}(p)$ are the lower and upper limits of the ex ante interval forecast for time t made at time $t-1$ for the coverage probability p.

Define the indicator variable I_t by

$$I_t = \begin{cases} 1, & \text{if } y_t \in [L_{t|t-1}(p); U_{t|t-1}(p)] \\ 0, & \text{if } y_t \notin [L_{t|t-1}(p); U_{t|t-1}(p)] \end{cases} \tag{6}$$

Note that VaR estimates have been mentioned as an application of interval forecasting where the intervals are one-sided, that is $[L_{t|t-1}(p); +\infty]$ or $[-\infty; U_{t|t-1}(p)]$ (Christoffersen, 1998). As mentioned before we are first interested in the accuracy of the frequency of failures. Namely, we say that a sequence of interval forecasts, $\{[L_{t|t-1}(p); U_{t|t-1}(p)]\}_{t=1}^T$ has *correct conditional coverage* if $\{I_t\} \overset{iid}{\sim} \text{Bern(p)}$, $\forall t$. Conditional coverage implies that VaR satisfies both unconditional coverage and independence.

LEMMA II.1: *Testing $E[I_t|\Psi_{t-1} = E[I_t|I_{t-1}; I_{t-2}; \ldots; I_1] = p$, for all t, is equivalent to testing that the sequence $\{I_t\}$ is identically and independently distributed Bernoulli with parameter p. Write $\{I_t\} \overset{iid}{\sim}$ Bern(p).*

The idea of Kupiec (1995) (which is similar to the first part in Christoffersen (1998)) is to test the i.i.d Bern(p) hypothesis for the sequence of interval forecasts by looking at how close the actual coverage is to the correct conditional coverage. This assesses the unconditional coverage assumption. It can be done using a likelihood ratio test where $H_0 : E[I_t] = p$ while $H_A : E[I_t] = \pi$ where $\pi \neq p$.

The test can be formulated as a standard likelihood ratio test by:

$$LR_{uc} = -2 \cdot log\Big[\frac{L(p; I_1, I_2, \ldots, I_T)}{L(\hat{\pi}; I_1, I_2, \ldots, I_T)}\Big] = -2 \cdot ln\Big[\frac{(1-p)^{n_0} p^{n_1}}{(1-\pi)^{n_0} \pi^{n_1}}\Big] \overset{asy}{\sim} \chi^2(s-1) = \chi^2(1) \tag{7}$$

where $\hat{\pi} = n_1/(n_0 + n_1)$ is the maximum likelihood estimate of π, n_0 and n_1 represent the number of zeros and ones, respectively, of the indicator variable. Note that the degree of freedom of the χ^2 distribution is 1 because the number of possible outcomes of the sequence is 2 ($s = 2$).

In addition to this test, Christoffersen (1998) designed a Markov test to check whether violations occur independently. To do so, consider a binary first-order Markov chain, $\{I_t\}$, with transition matrix:

$$\Pi_1 = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix} \tag{8}$$

where $\pi_{ij} = Pr(I_t = j | I_{t-1} = i)$. The hypothesis of independence can be tested by noting that

$$\Pi_2 = \begin{bmatrix} 1 - \pi_2 & \pi_2 \\ 1 - \pi_2 & \pi_2 \end{bmatrix} \tag{9}$$

corresponds to independence, i.e. past violation outcome has no influence on the current probabilities. The likelihood ratio test of independence can be represented by:

$$LR_{ind} = -2 \cdot log\Big[\frac{L(\hat{\Pi}_2; I_1, I_2, \ldots, I_T)}{L(\hat{\Pi}_1; I_1, I_2, \ldots, I_T)}\Big] \overset{asy}{\sim} \chi^2[(s-1)^2] = \chi^2(1) \tag{10}$$

where

$$\hat{\Pi}_1 = \begin{bmatrix} \frac{n_{00}}{n_{00}+n_{01}} & \frac{n_{01}}{n_{00}+n_{01}} \\ \frac{n_{10}}{n_{10}+n_{11}} & \frac{n_{11}}{n_{10}+n_{11}} \end{bmatrix} \text{ and } \hat{\Pi}_2 = \hat{\pi_2} = \frac{n_{01} + n_{11}}{n_{00} + n_{10} + n_{01} + n_{11}} \tag{11}$$

Note that the independence assumption is tested in a particular way, for instance dependence of order greater than one are not considered. Other tests that do not share this drawback exist, see for instance Engle and Manganelli (2004).

Finally, a joint test that consider both unconditional coverage and independence simultaneously can be constructed. Christoffersen (1998) forms a joint test of coverage and independence that is given by:

$$LR_{ind} = -2 \cdot log\Big[\frac{L(p; I_1, I_2, \ldots, I_T)}{L(\hat{\Pi}_1; I_1, I_2, \ldots, I_T)}\Big] \overset{asy}{\sim} \chi^2[s(s-1)] = \chi^2(2) \tag{12}$$

The empirical literature suggests that these test may lack power: they tend to validate models quite easily (Campbell, 2006; Christoffersen and Pelletier, 2004). Hurlin and Tokpavi (2008) showed that predictions of the most common VaR models were usually all validated by back-testing procedures even though their estimates and violation patterns differ substantially.

III. Conditional volatility model

A. GARCH

Mandelbrot (1963) was the first one to discover clustering phenomenon in the volatility of financial asset returns, which were assumed to be normally distributed before. This phenomenon implies that large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes. Hence, we are led to a conditional model where the past explains the current variance.

Engle (1982) tried to solve this problem by introducing the ARCH(1) model to obtain more realistic forecast variances. Soon after, Engle and Bollerslev (1986) extended the ARCH model, in which the next period's variance only depends the squared residual of the last periods, to the Generalized ARCH (GARCH) model, in which the variance is also explained by its own lags. The GARCH model is given by the following equations:

$$y_t = \mu + \varepsilon_t \tag{13a}$$

$$\varepsilon_t = z_t {h_t}^{1/2} \tag{13b}$$

$$V_{t-1}(\varepsilon_t) = h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1} \tag{13c}$$

where $\omega > 0$ and $\alpha; \beta \geq 0$. z_t forms an i.i.d. sequence with zero mean and unit variance. Classical choices are normal or student-distribution.

B. EGARCH

As it has been empirically observed in many markets, the impact of negative price moves on future volatility is different from that of positive price moves (Reider, 2009). That is, bad news causes more violent fluctuations than good news does. This phenomenon is called *Leverage Effect*. A model that allows for the asymmetry between up and down moves of future volatility is the EGARCH model (Nelson, 1991) in which the variance of ε_t is given by:

$$ln(h_t) = \omega + \frac{\alpha\varepsilon_{t-1} + \gamma|\varepsilon_{t-1}|}{\sqrt{h_{t-1}}} + \beta ln(h_{t-1}) \tag{14}$$

Another advantage of the EGARCH is that no constraint on the coefficients has to be imposed to ensure a positive variance.

According to Ding (2011), however, both GARCH and EGARCH have some drawbacks as they do not account for long memory in volatility. That is, they do not control for historical events which have a long and lasting effect on volatility. Several researchers attempt to simulate the long memory by introducing new models such as IGARCH, FIGARCH, FIEGARCH or LM-ARCH models.

Another interesting model is the APARCH model, which allows for asymmetric error term. Giot and Laurent (2003) proved that models that rely on a symmetric density distribution for the error term underperform compare to those which use a skewed density. Hence, they suggest using an APARCH model with a skewed Student-t conditionally distributed innovations.

Andersen and Bollerlev (1998) has shown that these conditional volatility models do provide good volatility forecasts. They showed that the way to judge the accuracy of the forecast, *i.e.* comparing the prediction to squared returns, was inadequate and that comparisons with realized volatility suggest a strong predictive power. Thereafter, VaR and ES estimates are based on GARCH and EGARCH models.

IV. Empirical analysis

A. Data

Our analysis will be based on a data set from the Swiss Market Index (SMI). The SMI is the most important index of the country and is made up of 20 of the largest and most liquid large- and mid-cap stocks. The data set consists of daily observations of its historical prices from November 9, 1990 to April 29, 2016, providing 6,433 observations of prices. The descriptive statistics are summarized in Table 1.

Mean	0.0002717
Std. Deviation	0.0116161
Skewness	-0.2149894
Kurtosis	9.271999
Obs.	6,432

Table 1—: Summary statistics of the data

Different sub-samples are considered in the following.

B. Econometric Methodology

Since the returns exhibit the typical volatility clustering, conditional volatility models are applied to the data. The GARCH(1,1) (Bollerslev, 1986) and EGARCH(1,1,1) (Nelson, 1991) are considered because they are both well-performing and quite popular, ensuring the results are representative. Regarding the lag length, a single lag leads to a more parsimonious model is usually enough to obtain a satisfying fit. Moreover, information criteria usually selects a lag length of one in our sample. In a first step, logarithm of the prices are taken and returns are computed as first differences. The sample is then split into an estimation period and a forecast period. Note that the forecast period had to be large enough for back-testing test statistics, whose properties rely on large samples, to be effective. We use two years of data for forecast.

Regarding the estimation period, it should be large enough to deliver accurate estimates and contain tumultuous periods as well, so that the model can capture features of the crisis. We use four years of data for estimation.

Since one of our purposes is to see whether the model selection depends on the estimation and forecast windows, different choices are considered.

The performance of the model is assessed during three different periods. A pre- and a post-crisis period have been selected: 2002-2003 and 2014-2015. Moreover, we test the model during the particularly stormy period of 2008-2009.

A rolling window is used to estimate the parameters of the models. That is, after the day following the estimation period has been forecasted, the estimation period is shifted by one day and the procedure is iterated again until the end of the sample. As the distribution appears unstable, regular updating of the estimated parameters is particularly important.

Conditional volatility models are estimated by maximum of likelihood, which requires numerical maximization methods. As a first approximation, the error term is assumed to be normally distributed. However, it is likely that even the conditional distribution exhibits fat tails and a t-distribution whose degrees of freedom are estimated as an additional parameter is also considered. VaR is then estimated in a straightforward way as the quantile of the return distribution. For the ES, analytical formulae are also available given the distribution considered.

Then, both risk-measures are back-tested. For the VaR, the Kupiec (1995) likelihood-ratio test, which checks whether the number of violations is in line with the one implied by the confidence level, and the Christoffersen (1998) test, which assesses whether violations truly occur independently, are computed. Then, the joint test that assesses the validity of the model is conducted by summing the two test statistics.

Following Emmer, Kratz and Tasche (2013), we assess the quality of the ES model by back-testing the VaR at different levels. More precisely, we consider

the approximation of the ES by $1/5(VaR_t(\alpha) + VaR_t(0.8\alpha) + VaR_t(0.6\alpha) + VaR_t(0.4\alpha) + VaR_t(0.2\alpha))$ and back-test the VaR at these 5 confidence levels to judge the ES's reliability. Note that we end up with fractions of the initial confidence level, so that the confidence levels lead to a small number of violations to deal with, requiring more data. This motivates our choice of $\alpha = 0.05$ in the following: it implies that we back-test VaR at 5%, 4%, 3%, 2%, 1% to backtest the Expected Shortfall; it also ensure that the lowest confidence level at which the VaR is back-tested is not too low (with 500 days, it is expected to have about 5 violations at the 1 % level).

C. Results

Estimation				
Dates	Mean	Std. Dev.	Skewness	Kurtosis
Sub-sample 1				
[1998; 2001]	0.00004	0.013	-0.218	6.256
Sub-sample 2				
[2004; 2007]	0.00042	0.008	-0.545	4.578
Sub-sample 3				
[2010; 2013]	0.00020	0.010	-0.282	6.413
Prediction				
Dates	Mean	Std. Dev.	Skewness	Kurtosis
Sub-sample 1				
[2002; 2003]	-0.00025	0.016	0.213	4.945
Sub-sample 2				
[2008; 2009]	-0.00047	0.018	0.227	7.673
Sub-sample 3				
[2014; 2015]	0.00008	0.010	-1.709	16.644

Table 2—: Descriptive statistics of the three sub-samples for both estimation and forecast window

Table 2 displays the descriptive statistics for the estimation and forecast window of the three sub-samples. It can be seen that some features of the market differ substantially among the different periods. The distribution of the returns appears skewed, but the sign of the asymmetry changes from one sub-sample to another. The kurtosis indicates fat tails in all sub-samples but it is particularly high in the last period, reaching a value of 16.64 which is more than twice as large as in the other periods.

Therefore, descriptive statistics taken over 2-3 years vary over time. It is possible that the data generating process is fundamentally time-varying or that events

such as the crisis have caused some structural breaks. Since estimation and forecast windows differ in some aspects, it would be possible that even conditional volatility model fail to capture some features of the market if a rolling window was not used.

Pre-Crisis 2002-2003

Model			Obs.	Mean	Std. Dev.
VaR	GARCH	Normal	511	-0.024	0.011
	GARCH	t	511	-0.024	0.011
	EGARCH	Normal	487	-0.023	0.010
	EGARCH	t	456	-0.023	0.010
ES	GARCH	Normal	511	-0.030	0.013
	GARCH	t	511	-0.032	0.014
	EGARCH	Normal	487	-0.029	0.012
	EGARCH	t	456	-0.085	1.166

Table 3—: Summary statistics for the VaR and ES in the different models for the period 2002-2003

		GARCH(n)	GARCH(t)	EGARCH(n)	EGARCH(t)
VaR	*Kupiec:*				
	t-statistics	4.014	4.773	6.931	5.065
	p-value	0.045	0.029	0.008	0.024
	Christoffersen:				
	t-statistics	1.362	1.563	0.000	1.776
	p-value	0.243	0.211	0.991	0.183
	Joint test:				
	t-statistics	5.376	6.336	6.931	6.841
	p-value	0.068	0.042	0.031	0.033

Table 4—: Back-test of the VaR for the period 2002-2003

We start discussing the GARCH and EGARCH model with the assumption of normality. On average, the VaR estimated by the means of the GARCH model reaches -2.42%. This is slightly higher than the average prediction of the EGARCH model, which is -2.30%. Standard deviations of the Value-at-Risk estimates are quite close, both round up at 0.01. The sequence of VaRs and ESs predicted by the models is plotted in Figure A1 and A2. The conditional volatility models seem to perform relatively well: VaR predictions adjust as the market

	GARCH(n)	GARCH(t)	EGARCH(n)	EGARCH(t)
Kupiec:				
p-value	0.045	0.029	0.008	0.024
p-value 80	0.043	0.026	0.014	0.005
p-value 60	0.037	0.064	0.067	0.010
p-value 40	0.158	0.399	0.114	0.130
p-value 20	0.118	0.961	0.192	0.838
Christoffersen:				
p-value	0.243	0.211	0.991	0.183
p-value 80	0.509	0.456	0.456	0.278
p-value 60	0.124	0.140	0.159	0.082
p-value 40	0.340	0.410	0.340	0.340
p-value 20	0.570	0.753	0.614	0.753
Joint test:				
p-value	0.068	0.042	0.031	0.033
p-value 80	0.104	0.064	0.037	0.012
p-value 60	0.035	0.060	0.070	0.008
p-value 40	0.234	0.499	0.183	0.202
p-value 20	0.251	0.951	0.376	0.932

Table 5—: Back-test of the ES for the period 2002-2003

becomes more volatile; the models seem to correctly anticipate the volatility.

Consider now a student distribution for the error term. The estimated degrees of freedom are about 8 or 9 in general, suggesting fatter tails than those implied by the normal distribution in the conditional distribution. Previsions are depicted in Figure A3 and A4. The same issues as in the normal case seem to be present.

The VaR estimates are in general slightly higher (-2.40% on average for the GARCH model, -2.31% on average for the EGARCH model) than those of the conditional volatility models with the normality assumption.

The results of Kupiec's test suggest that the number of violations may not be in line with the confidence level of the VaR. The number of violations is higher than expected and the likelihood ratio test indicates that the difference is significant at the 5% level. However, only the normal EGARCH model is rejected at a 1% level so the evidence against a correct coverage of the VaR models are not extremely compelling.

Results are more favorable to the models regarding Christoffersen's test. There is no evidence against the hypothesis of independence of violations Note that only independence of violations is only tested at the first-order with this test so that more subtle forms of dependence may not be detected.

The joint test leads to a non-rejection of the hypothesis of unconditional coverage for all models at a 1% size and at a 5% size for the normal GARCH. There is weak evidence against the validity of the models.

The back-test of the ES, which has been reduced to back-testing VaR at fractions of the initial confidence level, gives similar results than the VaR about the validity of the models. While the assumption of independence, clearly, look reasonable for any confidence level of the VaR according to Christoffersen's test, the Kupiec test indicates that there is still some doubt (at a 5% size) about the quality of the models in the first three confidence level. There is no evidence of an incoherent number of rejection in the others confidence level of the VaR. Finally, the joint test suggests that everything is fine at a 1% level but some doubts subsist at a 5% level for the highest fraction of α, for all models.

Note that the quality of the approximation by the VaRs must be good for the indirect backtest to be relevant. The square of the error made by approximating the ES by the mean of these 5 VaRs is 310^{-6}, 3.3110^{-6}, 6.0210^{-6} and 8.3410^{-6} for the GARCH, EGARCH, GARCH-t and EGARCH-t, respectively. On average, the approximation tends to overestimate the ES by about 0.002. An illustration of the fit is provided in Figure A13 for the EGARCH model; it can be seen that the approximation follows reasonably well the true pattern of the ES. Other models yield very similar figures.

Crisis 2008-2009

Model			Obs.	Mean	Std. Dev.
VaR	GARCH	Normal	511	-0.025	0.014
	GARCH	t	511	-0.024	0.014
	EGARCH	Normal	494	-0.023	0.011
	EGARCH	t	428	-0.023	0.019
ES	GARCH	Normal	511	-0.031	0.017
	GARCH	t	511	-0.033	0.018
	EGARCH	Normal	494	-0.029	0.016
	EGARCH	t	428	-0.031	0.027

Table 6—: Summary statistics for the VaR and ES in the different models for the period 2008-2009

The period seems more tumultuous but the VaR estimates seem to adjust to this: the lowest value taken by the VaR estimates varies from -9.26% (EGARCH with t-distribution) to -8.1% (GARCH with normal distribution). For the first

		GARCH(n)	GARCH(t)	EGARCH(n)	EGARCH(t)
VaR	*Kupiec:*				
	t-statistics	4.014	6.464	11.801	4.829
	p-value	0.045	0.011	0.001	0.028
	Christoffersen:				
	t-statistics	0.140	0.000	0.306	0.172
	p-value	0.708	0.988	0.580	0.679
	Joint test:				
	t-statistics	4.154	6.464	12.107	5.000
	p-value	0.125	0.039	0.002	0.082

Table 7—: Back-test of the VaR for the period 2008-2009

	GARCH(n)	GARCH(t)	EGARCH(n)	EGARCH(t)
Kupiec:				
p-value	0.045	0.011	0.001	0.028
p-value 80	0.016	0.026	0.000	0.007
p-value 60	0.021	0.021	0.000	0.002
p-value 40	0.003	0.006	0.000	0.004
p-value 20	0.023	0.118	0.000	0.001
Christoffersen:				
p-value	0.708	0.988	0.580	0.679
p-value 80	0.407	0.457	0.936	0.778
p-value 60	0.827	0.827	0.319	0.628
p-value 40	0.881	0.806	0.095	0.201
p-value 20	0.487	0.570	0.251	0.374
Joint test:				
p-value	0.125	0.039	0.002	0.082
p-value 80	0.038	0.064	0.001	0.027
p-value 60	0.069	0.069	0.000	0.008
p-value 40	0.011	0.023	0.000	0.008
p-value 20	0.060	0.251	0.000	0.002

Table 8—: Back-test of the ES for the period 2008-2009

sub-sample, the lowest estimated VaR was only -6.09%. The average VaR ranges from -2.46 to -2.25%. Figures A5 and A6 displays VaR and ES estimates along with the returns. It can be seen that VaR and ES forecasts moves in line with the volatility of the market. Nevertheless, a cluster of violations seem to be present in the middle of the forecast: this period corresponds to October 2008. VaR models adjust afterwards, but some violations occur in a relatively short time.

Assuming now a student distribution, VaR and ES forecasts are represented in Figures A7 and A8

All models satisfy the Kupiec test at the 1% level except for the normal-EGARCH. However, they all fail the test at the 5 % level. Again, this suggests that the number of violation is a bit too high and that the hypothesis of correct conditional coverage is not completely valid.

All models satisfy clearly the hypothesis of independence of the violations. As a result, the joint hypothesis tend not to be rejected; the normal-GARCH is validated at a 10 % level, the t-EGARCH at a 5% level and the t-GARCH at a 1% level.

The GARCH-n and GARCH-t's VaR are successfully back-tested at all fractions of α that are considered when a 1% level is selected for the joint test and only a few are rejected at a 5% level. Once again, the Kupiec test show p-values are above 10% except for the VaR at (0.4 α), suggesting that GARCH models perform well. On the other hand, the EGARCH models show surprising bad performance, compared to the other periods. Once again, independency hypothesis cannot be rejected; p-values are all above 10% for Christoffersen's test.

Current state 2014-2015

Estimates are plotted with the returns in Figures A9 and A10 and A11 and A12.

No model can be rejected by the Kupiec test at the 10% level. In this sample, there is no evidence that the number of violations is not in line with the 5%-confidence level of the VaR. Moreover, all models satisfy the independence of violations assumption at the 10%-size, except for the GARCH-t which passes the test at the % level.

Consequently, the joint test lead to a non-rejection of the validity of the VaR prediction regardless of the model at a 5% level.

Regarding the Expected Shortfall's backtest, results are also quite good. The independence of violations is not rejected in any case. The Kupiec test leads to a non-rejection of the model for all fraction of α at a 2% level for the normal GARCH, at a 9% level for the GARCH-t, at a 1% level for the normal EGARCH and at a 5% level for the EGARCH-t. Note the tendency to be a bit pulled toward rejection for lower fractions of α here.

D. Further research

Since the VaR fails the backtesting procedures most of the time, further research is required to develop a suitable model. The apparently unstable data

Model			Obs.	Mean	Std. Dev.
VaR	GARCH	Normal	511	-0.015	0.007
	GARCH	t	511	-0.014	0.006
	EGARCH	Normal	476	-0.015	0.006
	EGARCH	t	481	-0.014	0.007
ES	GARCH	Normal	511	-0.019	0.009
	GARCH	t	511	-0.020	0.009
	EGARCH	Normal	476	-0.019	0.008
	EGARCH	t	481	-0.021	0.010

Table 9—: Summary statistics for the VaR and ES in the different models for the period 2014-2015

		GARCH(n)	GARCH(t)	EGARCH(n)	EGARCH(t)
VaR	*Kupiec:*				
	t-statistics	0.774	2.677	0.147	0.039
	p-value	0.379	0.102	0.702	0.843
	Christoffersen:				
	t-statistics	0.826	2.946	0.788	0.335
	p-value	0.363	0.086	0.375	0.563
	Joint test:				
	t-statistics	1.601	5.624	0.935	0.374
	p-value	0.449	0.060	0.626	0.830

Table 10—: Back-test of the VaR for the period 2014-2015

generating process has to be apprehended in a different way than above. A first diagnostic could be derived from the use of realized volatility to assess the quality of the GARCH forecasts.

There are several possibilities to be explored. A more complex form for the mean equation of the GARCH model, *e.g.* an ARMA form instead of the simple constant, could be used. This is motivated by the atypical significant correlation in Returns until 5 lags. Other extensions of the GARCH model could also be used in order to obtain more relevant volatility forecasts. Finally, additional explanatory variables for the conditional variance could improve the quality of the forecasts. The following have been suggested in the literature:

- trading volume, macroeconomic news announcements (Lamoureux and Lastrapes, 1990; Flannery and Protopapadakis, 2002; Bomfim, 2003)
- implied volatility from option prices and realized volatility (Blair, Poon and

	GARCH(n)	GARCH(t)	EGARCH(n)	EGARCH(t)
Kupiec:				
p-value	0.379	0.102	0.702	0.843
p-value 80	0.433	0.228	0.993	0.686
p-value 60	0.359	0.163	0.650	0.244
p-value 40	0.050	0.091	0.170	0.058
p-value 20	0.023	0.118	0.014	0.182
Christoffersen:				
p-value	0.363	0.086	0.375	0.563
p-value 80	0.434	0.181	0.225	0.320
p-value 60	0.185	0.055	0.118	0.225
p-value 40	0.118	0.092	0.453	0.118
p-value 20	0.486	0.570	0.447	0.614
Joint test:				
p-value	0.449	0.060	0.626	0.830
p-value 80	0.542	0.197	0.479	0.562
p-value 60	0.272	0.060	0.266	0.243
p-value 40	0.043	0.058	0.295	0.049
p-value 20	0.060	0.251	0.037	0.361

Table 11—: Back-test of the ES for the period 2014-2015

Taylor, 2010)

- overnight returns (Martens, 2002; Gallo and Pacini, 1998)

Lastly, it may be that the entire framework in which VaR and ES estimates have been constructed is flawed. Danielsson et al. (2001) point out the potential endogeneity of risk, particularly in stormy periods. Considering the instability of the Swiss Stock Exchange and the necessity to assess the performance of the models during the crisis, this is a major concern. In the VaR and ES framework, it is assumed that one's actions, based on a volatility forecast, have no effect on future volatility. Nevertheless, volatility is determined within the market, making risk endogenous. During peaceful periods, failure to recognize the endogeneity of risk does not matter much. However, things are very different during a crisis. In the case a price drop for instance, incentive for market participants to sell their asset is reinforced by the sales of the other individuals. This externality is not taken into account by the models that try to forecast risk. In time of crisis, those effects can be so damaging that both ES and VaR may no longer be justified. This phenomenon leads to a structural break in the data generating process at the beginning of crises, violating the stationarity assumption. Hence, prior data is no more useful for estimating risk.

V. Conclusion

From a theoretical point of view, Basel committee's decision to replace the VaR by the ES appears judicious. While back-testing the Expected Shortfall may require alternative techniques, the actual properties of this risk measure are better suited for management purposes. Both the ability to assess the loss within the tail and the satisfaction of the subadditivity axiom are compelling arguments for ES's supremacy over VaR.

This study assessed the predictions of VaR and ES using different conditional volatility models, noting that back-testing ES amounts to back-test VaR at different levels. Christoffersen's test of independence suggests that violations occur indeed at random, regardless of the model considered. However, in some cases, the Kupiec test casts some doubt about the validity of the conditional coverage; the number of violation is often slightly too high but not so much as to lead to a definitive rejection.

Of the four models, none is rejected in all subsamples. There is a tendency not to reject VaR predictions for various models even if their estimates differ, as observed by Hurlin and Tokpavi (2008).

Different subsamples may lead to a different model selection. There is no clear choice of a model when considering all subsamples although it can be seen that the normal GARCH model usually performs quite well in general. Hence, the need to use more complex models is not always apparent but, as Hurlin and Tokpavi (2008) point out, VaR predictions tend to be too easily validated so that sophisticated methods that take into account for instance covariances between assets (mutivariate GARCH, DCC models, etc.) should not be necessarily thrown away in favor of simpler models: it might just be that these tests lack power and fail to reject invalid VaR predictions. Note that the EGARCH-t usually performs well too.

Backtesting the ES brought us to back-test the VaR for different values of α. Note that VaR should be successfully back-tested at all levels, which is not always the case but, at least for 1%-sized tests, ES can be considered successfully back-tested in most cases.

Whether the risk measure is the VaR or the ES does not lead to different assessments: both of their back-testing suggest that the conditional volatility models are appropriate. However, model selection can differ a bit since results of a VaR back-testing appear to differ a bit for different confidence level. Moreover, the selection of a conditional volatility model may very well depend on the estimation and forecast windows since performances of the models seem to vary across samples. The results showed that the validity of a model in one period does not

always guarantee its validity in another period.

REFERENCES

Acerbi, Carlo, and Balazs Szekely. 2014. "Bactesting Expected Shorfall: Introducing three model-independent, non-parametric back-test methodologies for Expected Shortfall." *MSCI.*

Acerbi, Carlo, and Dirk Tasche. 2002. "On the coherence of expected shortfall." *Journal of Banking & Finance*, 26(7): 1487–1503.

Acerbi, Carlo, Claudio Nordio, and Carlo Sirtori. 2001. "Expected shortfall as a tool for financial risk management." *arXiv preprint cond-mat/0102304.*

Andersen, Torben G., and Tim Bollerlev. 1998. "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts." *International Economic Review*, 39(4).

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, and David Heath. 1997. "Thinking coherently." *Risk*, 10: 68–71.

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, and David Heath. 1999. "Coherent measures of risk." *Mathematical finance*, 9(3): 203–228.

Basel Committee on Banking Supervision. 2013. "Fundamental review of the trading book: A revised market risk framework."

Bellini, F., Klar B. Müller A., and A. Rosazza Gianin. 2013. "Generalized quantiles as risk measures." *Journal of the American Statistical Association.*

Blair, Bevan J, Ser-Huang Poon, and Stephen J Taylor. 2010. "Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns." In *Handbook of Quantitative Finance and Risk Management.* 1333–1344. Springer.

Bollerslev, Tim. 1986. "Generalized autoregressive conditional heteroskedasticity." *Journal of econometrics*, 31(3): 307–327.

Bomfim, Antulio N. 2003. "Pre-announcement effects, news effects, and volatility: Monetary policy and the stock market." *Journal of Banking & Finance*, 27(1): 133–151.

Campbell, Sean D. 2006. "A review of backtesting and backtesting procedures." *The Journal of Risk*, 9(2): 1.

Carver, L. 2013. "Mooted VAR substitute cannot be back-tested, says top quant." *RISK.*

Christoffersen, Peter, and Denis Pelletier. 2004. "Backtesting value-at-risk: A duration-based approach." *Journal of Financial Econometrics*, 2(1): 84–108.

Christoffersen, Peter F. 1998. "Evaluating interval forecasts." *International economic review*, 841–862.

Danielsson, Jon, Paul Embrechts, Charles Goodhart, Con Keating, Felix Muennich, Olivier Renault, Hyun Song Shin, et al. 2001. "An academic response to Basel II."

Ding, Ding. 2011. "Modeling of Market Volatility with APARCH Model." *Department of Mathematics Uppsala University*.

Dowd, Kevin. 2007. *Measuring market risk.* John Wiley & Sons.

Embrechts, P. 2000. "Extreme Value Theory: Potentials and Limitations as an Integrated Risk Management Tool. Manuscript, Zurich, Switzerland: Department of athematics, ETH, Swiss Federal Technical University. M."

Emmer, Susanne, Marie Kratz, and Dirk Tasche. 2013. "What is the best risk measure in practice? A comparison of standard measures."

Engle, F. Robert. 1982. "Autoregressive Conditional Heteroskedasticity with estimates of the variance of U.K. inflation." *Econometrica*, 50(4).

Engle, F. Robert, and Tim Bollerslev. 1986. "Modelling the persistence of Conditional Variances." *Econometric Review*, 50(1).

Engle, Robert F, and Simone Manganelli. 2004. "CAViaR: Conditional autoregressive value at risk by regression quantiles." *Journal of Business & Economic Statistics*, 22(4): 367–381.

Flannery, Mark J, and Aris A Protopapadakis. 2002. "Macroeconomic factors do influence aggregate stock returns." *Review of Financial Studies*, 15(3): 751–782.

Gallo, Giampiero M, and Barbara Pacini. 1998. "Early news is good news: the effects of market opening on market volatility." *Studies in Nonlinear Dynamics & Econometrics*, 2(4).

Giot, Pierre, and Sébastien Laurent. 2003. "Value-at-risk for long and short trading positions." *Journal of Applied Econometrics*, 18(6): 641–663.

Gneiting, T. 2012. "Making and evaluating point forecasts." *Journal of the American Statistical Association*, 106(494).

Hurlin, Christophe, and Sessi Tokpavi. 2008. "Une évaluation des procédures de Backtesting." *Finance*, 29(1): 53–80.

Jorion, Philippe, et al. 2007. *Financial risk manager handbook.* Vol. 406, John Wiley & Sons.

Kupiec, Paul H. 1995. "Techniques for verifying the accuracy of risk measurement models." *The J. of Derivatives*, 3(2).

Lamoureux, Christopher G, and William D Lastrapes. 1990. "Heteroskedasticity in stock return data: volume versus GARCH effects." *The Journal of Finance*, 45(1): 221–229.

Mandelbrot, B. 1963. "The variation of certain speculative prices." *Journal of Business*, 36.

Martens, Martin. 2002. "Measuring and forecasting S&P 500 index-futures volatility using high-frequency data." *Journal of Futures Markets*, 22(6): 497–518.

Nelson, Daniel B. 1991. "Conditional heteroskedasticity in asset returns: A new approach." *Econometrica: Journal of the Econometric Society*, 347–370.

Reider, Rob. 2009. "Volatility Forecasting I: GARCH Models."

Rootzen, Holger, and Claudia Klüppelberg. 1999. "A single number can't hedge against economic catastrophes." *AMBIO-STOCKHOLM-*, 28: 550–555.

Appendix

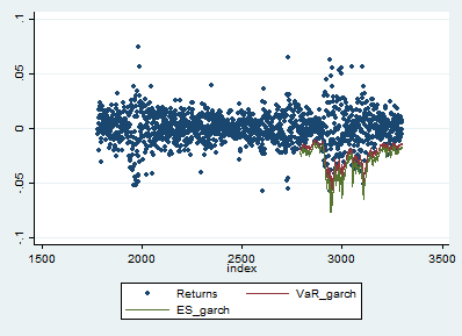

Figure A1. : VaR and ES in the normal-GARCH model in the first subsample

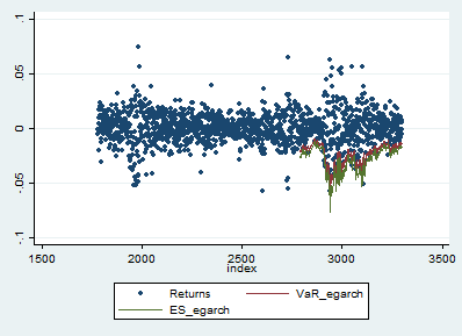

Figure A2. : VaR and ES in the normal-EGARCH model in the first subsample

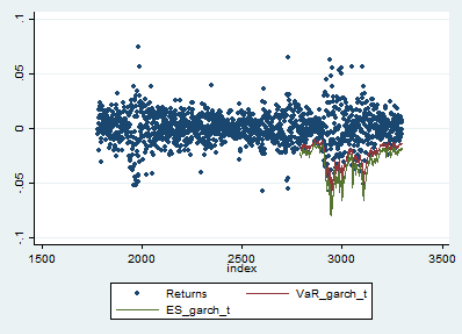

Figure A3. : VaR and ES in the t-GARCH model in the first subsample

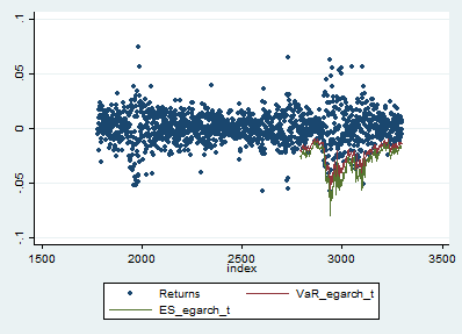

Figure A4. : VaR and ES in the t-EGARCH model in the first subsample

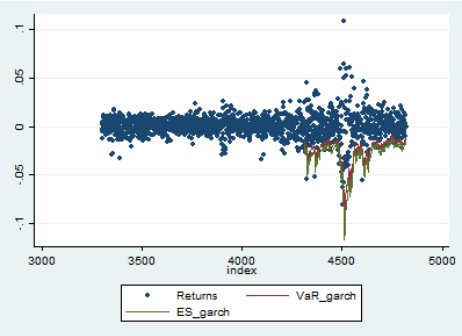

Figure A5. : VaR and ES in the normal-GARCH model in the second subsample

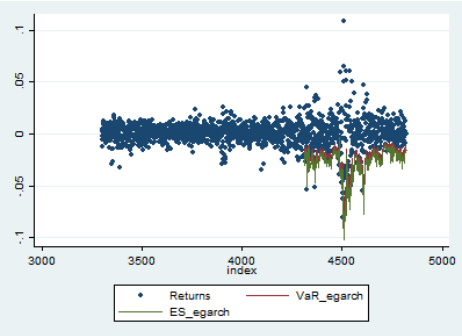

Figure A6. : VaR and ES in the normal-EGARCH model in the second subsample

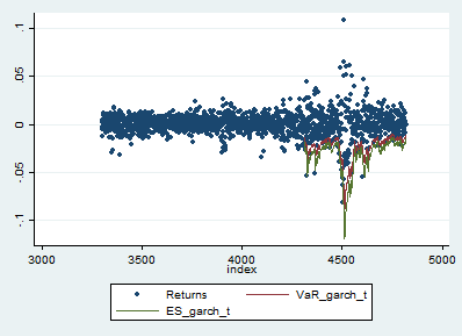

Figure A7. : VaR and ES in the t-GARCH model in the second subsample

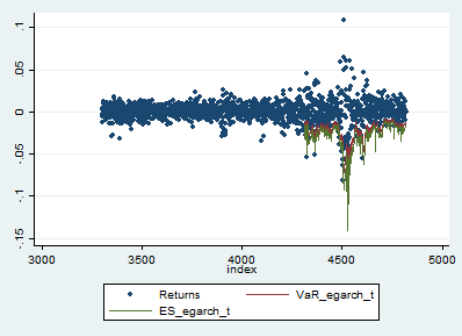

Figure A8. : VaR and ES in the t-EGARCH model in the second subsample

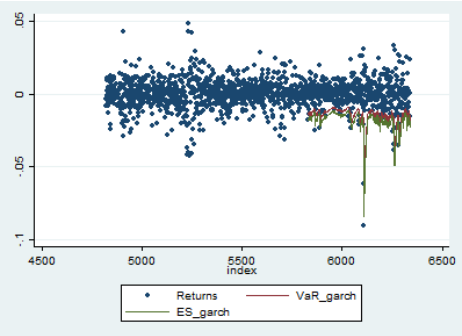

Figure A9. : VaR and ES in the normal-GARCH model in the third subsample

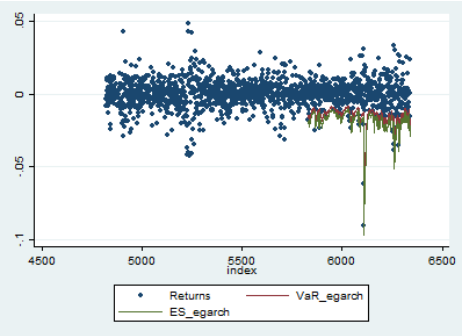

Figure A10. : VaR and ES in the normal-EGARCH model in the third subsample

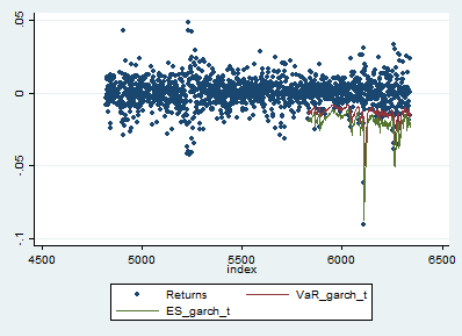

Figure A11. : VaR and ES in the t-GARCH model in the third subsample

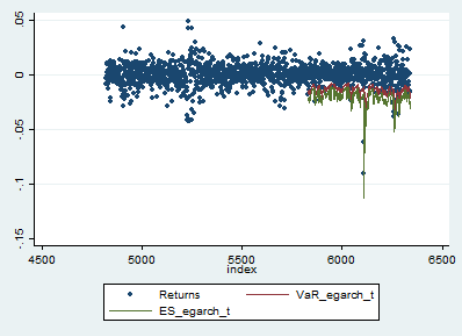

Figure A12. : VaR and ES in the t-EGARCH model in the third subsample

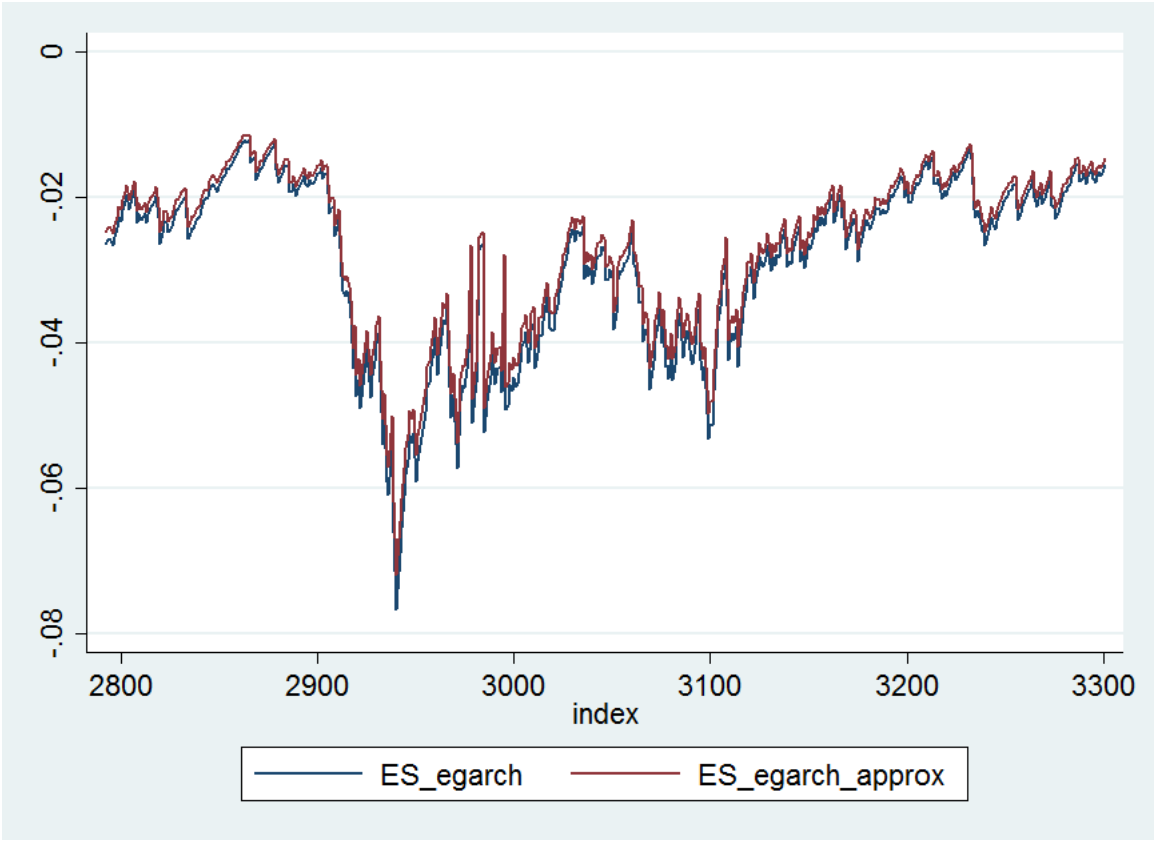

Figure A13. : Comparison of approximation of EGARCH with true pattern of the ES