Master Thesis

Estimating Discrete Stochastic Volatility Models & Application to Risk Measures Estimation

LOÏC CANTIN*,
loic.cantin@ensae.fr

This report provides a summary of the main methods used for the estimation of discrete stochastic volatility models after recalling their definition and their advantages, particularly with respect to the more commonly used traditional GARCH model. Among these methods, we can find the "frequentist" methods such as the estimation by QML or GMM and the more sophisticated methods based on simulations. For each of the two approaches, we propose at least one code and a basic Monte Carlo experiment to compare the efficiency of the models – we chose the QML, GMM and Indirect Inference methods. As an illustration, we propose a method for estimating risk measures and the construction of a measure of confidence in the markets, namely the volatility risk premium.

Keywords: Discrete Stochastic Volatility Models, GARCH Models, QML, GMM, Indirect Inference, Risk Measures, Volatility Risk Premium.

Contents

*ENSAE (Master in Economics)

1. Introduction

The ability to understand how financial prices evolve over time has always been one of the major concerns of the different actors operating in the financial markets. Whether it is for the personal or institutional investor wishing to make their capital grow, or for the regulator whose responsibility is to limit the systemic risks; understanding the dynamics of the market and its catalysts is of prime importance and remains a task of great complexity. Thus, the use of sophisticated tools and models has grown to improve the understanding of markets and get insights of the economic phenomena at stake for the theoretician; and to develop forecasting instruments for the practitioner. However, even if it is well known that there is no free lunch in the market – i.e. every rise in the expected return (ϵ_t) should be associated to a proportional increase in risk - and that it is therefore - according to classical financial theory and especially to the no arbitrage condition - almost impossible to know at date t the direction of returns at date *t+1*, it is possible to convincingly estimate what is called the "volatility" ($\sqrt{h_t}$) - i.e. a proxy for the risk, which measures the intensity of the movement of time series - at date t.

Volatility has traditionally been estimated using ARCH or GARCH (1) time series models introduced by Engle 1982 [1] and Bollerslev 1986 [2] respectively.

$$\text{(1) } \textbf{GARCH(1,1) Model} \begin{cases} \epsilon_t &= \sqrt{h_t}\,\eta_t \\ h_{t+1} &= \omega + \alpha\,\epsilon_t^2 + \beta\,h_t \end{cases} \quad \text{with } \eta_t \sim \mathcal{N}(0,1)$$

These models have the advantage of verifying stylized features such as volatility clustering - i.e. "large changes tend to be followed by large changes and small changes to be followed by small changes" (Mandelbrot 1963 [3]), resulting in positive auto-correlation of squared returns. The GARCH model also allows for larger distribution tails than in the standard Gaussian case, which have been observed empirically on stocks log-returns. The major advantage of the GARCH model is that the returns are not assumed to be independent, but only independent conditionally on the information available at date *t-1*. We will call hereafter $\mathcal{F}_{t-1} := \sigma(\epsilon_{-\infty}, ..., \epsilon_0, ..., \epsilon_{t-1})$ the filtration generated by the log-returns until date *t-1*.

The main alternative to the GARCH model is the stochastic volatility (SV) model, whose first introduction is generally attributed to Taylor 1982 [4]. The idea is that, unlike the GARCH model, the (discrete) stochastic volatility model postulates that a second stochastic process directly influences the volatility (2).

$$(2)\ \textbf{Stochastic Volatility Model}\ \begin{cases} \epsilon_t = \sqrt{h_t}\eta_t \\ log(h_t) = \omega + \beta log(h_{t-1}) + \sigma v_t \end{cases}$$

with $\eta_t \sim \mathcal{N}(0,1)$ and $v_t \sim \mathcal{N}(0,1)$

The major difference with the GARCH model is that conditional on $\mathcal{F}_{t-1}$, the volatility is a stochastic process and no longer a deterministic one, which obviously makes the model much more difficult to estimate. The reason why this model may be interesting to study however, is that it presents a few advantages over GARCH: **i)** the SV model offers a natural economic interpretation of volatility (based on the "mixture of distributions hypothesis" - which states that returns are driven by a mixture of two random variables; an independent noise term and a stochastic process representing the inflow of new information, **ii)** this model is a natural discretization of the continuous time theory (see appendices) and **iii)** it proves to be more flexible in the modeling of returns.

The main reason why GARCH models are still much more popular than the stochastic volatility model - especially in practice - is the high difficulty of estimating the stochastic volatility model and the problem of finding a simple and robust method even for estimating the most basic stochastic volatility model (2). One can indeed easily estimate the θ parameter of a GARCH model (i.e. $\theta = (\omega, \alpha, \beta)$ for a GARCH(1,1)) by QML as in Francq and Zakoïan (2019) [5], but this does not apply as easily to the case of the SV model. Indeed, the likelihood function of a SV model can be written as:

$$(3)\ \mathcal{L}(\theta, \underline{\epsilon}_T) \propto \int f(\underline{\epsilon}_T | \underline{h}_T; \theta) f(\underline{h}_T | \theta) d\underline{h}_T$$

- $\theta = (\omega, \beta, \sigma)$ the parameter
- $\underline{\epsilon}_T = (\epsilon_1, ..., \epsilon_T)$ refers to the (log-)returns

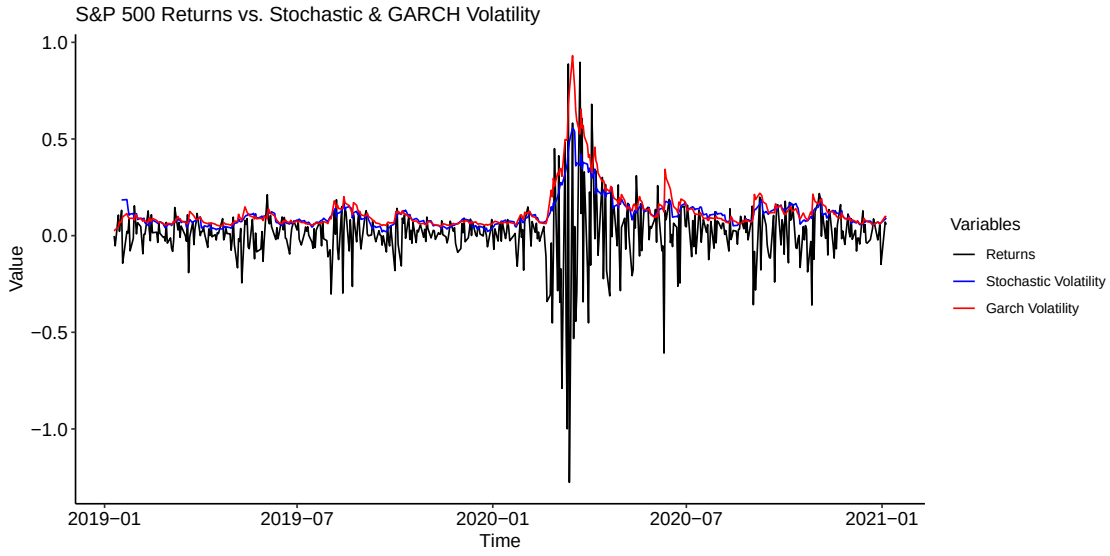
- $\underline{h}_T = (h_1, ..., h_T)$ refers to the squared volatility

By writing the ML function as in (3), we can see the difficulty of estimating the stochastic volatility model, since we are confronted to a T-dimensional integral that cannot be solved analytically and for which traditional numerical methods prove inefficient for a large sample (i.e. T large).

This report is therefore an attempt to propose a non-exhaustive synthesis of the main methods for estimating the basic model using methods that could be described as "frequentist" (QML, GMM) or more sophisticated methods based on simulations and bayesian statistics (indirect inference, importance sampling) **Section 2** consists in reviewing and explaining the functioning of these main methods. **Section 3** describes in details the implementation of three methods and show results of a Monte Carlo experiment. Finally, **section 4** presents the application of this model to the case of volatility risk premium and describes very simple method of risk estimation.

Figure 1 gives an illustration of the conditional volatility of a classical GARCH(1,1) model estimated by QML versus the stochastic volatility estimated according to the QML method of Francq and Zakoïan (2006) [6] on real data, namely the S&P 500 log-returns centered around the CoViD crisis.

Figure 1. S&P 500 Volatility - GARCH vs. Stochastic Volatility Model



2. Overview of the Existing Estimation Methods in the Literature

2.1. Traditional Approaches in Short

2.1.1. QML

The first and most common frequentist method for estimating stochastic volatility is the natural method used to estimate a GARCH model, i.e. the QML estimation. But for this method, modifications will have to be made in order to address the problems stated above. This method was first applied to this problem in the paper by Ruiz (1994) [7]. The main idea is to bypass the non-linearity of the model in order to apply the Kalman filter.

The first step is to linearize the equation:

$$y_t := log(\epsilon_t^2) = log(h_t) + log(\eta_t^2)$$

Let's introduce a few notations:

$$c_t \approx ln\epsilon_t^2 + 1.2704$$
$$d_t = \frac{1}{2}\pi$$

The author proposes to estimate the density of $log(\epsilon_t^2)$ by a normal density, with expectation and variance known to be ≈ -1.2704 and $\frac{1}{2}\pi^2$ respectively. We then look for a linear approximation of the density thanks to an evaluation using the Kalman filter.

The main idea is to use the following linear Gaussian state-space model to approximate (2):

$$c_t = log(h_t) + u_t,\ u_t \sim \mathcal{N}(0, d_t)$$

We will not give details in this section but will provide full analysis of the method in section 3, where we cover the paper of Francq and Zakoïan (2006) [6] and offer a numeric illustration.

2.1.2. GMM

The GMM estimation method is another classical method used to solve this kind of problem. We can quote mainly three papers dealing with this subject whose main difference is the choice of the number of moments to take into account: Taylor 1986 [8], Melino & Turnbull 1990 [9], and Andersen & Sorensen 1996 [10]. We will focus on the investigation of the latter in the next section. We will just present here the method in a few words. The idea is to compare the empirical moments to the theoretical moments of the model

based on a particular value of the parameter $\theta \in \Theta$ and find the one which minimizes a criteria of the following form:

$$\hat{\theta}_T = argmin_{\theta \in \Theta}(M_T(\theta) - A(\theta))' \Lambda_T^{-1} (M_T(\theta) - A(\theta))$$

with Λ_T a well-chosen weighting matrix, $M_T(\theta)$ a vector of empirical moments and $A(\theta)$ a vector of analytical moments. We will provide the full details of the implementation of the method in section 3.

2.2. Simulation-based Methods

A "simulation-based method" is an approach that relies on (a large number of) simulations to either approximate the likelihood using bayesian methods, or directly estimate the parameter of interest using a large number of draws of an auxiliary model. There are several types of simulation-based methods, we will present the main ones, namely: *Gaussian Mixture Sampling*, *Simulated Method of Moments*, *Importance Sampling* and *Indirect Inference*. This choice is motivated in particular by the presentation of the principal methods mentioned in Bauwens et al. (2012) [11]. Let's start with gaussian mixture sampling.

2.2.1. Gaussian Mixture Sampling

The method relying on Gaussian mixture sampling and developed by Carter and Kohn (1997) [12], Kim et al. (1998) [13] and Omori et al. (2007) [14] is based on the same observation that one can linearize the model as discussed above (see by QML). However, this time the density of $ln(e_t^2)$ is approximated using a Gaussian mixture (GM) density, as in:

$$ln(\eta_t^2) \approx u_t \sim \sum_{i=1}^{C} p_i \mathcal{N}(m_i, s_i^2)$$

- p_i are the weights of the mixture components
- m_i is the mean of the mixture components and,
- s_i^2 the variance.

We can already notice that the QML approach is just the same for C=1.

The full model is no longer an unconditionally linear and Gaussian state space model, as it depends on the discrete unobserved states i_t. Instead of optimizing the likelihood directly, a Bayesian Markov Chain Monte Carlo (MCMC) approach with data augmentation is used.

If one assumes, indices i_t indicating which element i of the mixture is used at time t, the model for $\underline{\epsilon}|\underline{i}$ is conditionally linear and Gaussian, where $\underline{i} = (i_1, ..., i_T)$. and let's

denote $\underline{h} = (log(h_1), ..., log(h_T))$ We have then:

$$c_t|i_t = ln(\epsilon_t^2) - m_{i_t}$$
$$d_t|i_t = s_{i_t}^2$$

Below the following steps to conduct the estimation:

1. We first need to choose a smoother (for example De Jong and Shephard (1995) presented a simulation smoother, which is a new multi-state Gibbs sampler for time series)
2. Let's denote $\mathbb{P}(i_t = i)$ the probability conditionally on the states $\underline{h}$, on the parameter θ and on data $\underline{\epsilon}$, which is specific to each mixture node and that we can approximate. From the resulting multinomial density, a sample of new indices i_t is drawn.
3. Then, we draw new parameter $\theta \in \Theta$ conditionally on the states $\underline{i} = \{i_1, ..., i_T\}$, $\underline{h}$ and the data $\underline{\epsilon}$.
4. A final sample from the parameters and states describes the posterior density of these, conditionally on the full set of data.
5. The posterior mean of the states is an estimate of the smoothed states.

2.2.2. Simulated Method of Moments (SMM)

The idea of this method is to deal with the cases where the log-likelihood $ln(P(\underline{\epsilon}; \theta))$ is not available in closed form (and thus can also be applied to the cases where we have a closed form but it is too complex to compute). This approach consists here in the papers of Gallant and Tauchen (1996) [15] and Andersen et al. (1999) [16] in finding the parameters that set the average score to zero, with score function:

$$s(\underline{\epsilon}; \theta) = \frac{1}{T} \frac{\partial ln(P(\underline{\epsilon}; \theta))}{\partial \theta}$$

But as we do not have an expression of such a score function, the idea of SMM is to use instead an auxiliary model, that delivers an auxiliary parameter estimate $\hat{\theta}_{Aux}$.

Clearly, at these estimates, the score of the log-likelihood $ln(P_{Aux}(\underline{\epsilon}; \hat{\theta}_{Aux}))$ of the auxiliary model for the current set of data equals zero.

Hence SMM evaluates:

$$s_{Aux}(\underline{\epsilon}^*(\theta), \hat{\theta}_{Aux}) = \frac{1}{T_{Aux}} \frac{\partial ln(P_{Aux}(\underline{\epsilon}^*(\theta); \theta_{Aux}))}{\partial \theta_{Aux}} \Bigg|_{\theta_{Aux} = \hat{\theta}_{Aux}}$$

, which is the score of the auxiliary model.

2.2.3. Methods based on Importance Sampling

The idea of importance sampling is to approximate the likelihood $\mathcal{L}(\underline{\epsilon};\theta)$ by simulations using:

$$\begin{aligned}\mathcal{L}(\underline{\epsilon};\theta) &= \int \frac{\mathbb{P}(\underline{\epsilon},\underline{h})}{G(\underline{h}|\underline{\epsilon}^*)} G(\underline{h}|\underline{\epsilon}^*)\partial\underline{h} \\ &\approx \frac{1}{M}\sum \frac{\mathbb{P}\left(\underline{\epsilon},\underline{h}^{(i)}\right)}{G\left(\underline{h}^{(i)}|\underline{\epsilon}^*\right)} \\ &= \frac{1}{M}\sum \omega_i := \overline{\omega} \\ log\mathcal{L}(\underline{\epsilon};\theta) &\approx log(\overline{\omega}) + \frac{s_\omega^2}{2M\overline{\omega}^2} \end{aligned} \tag{2.1}$$

- with $\underline{h}^{(i)} = \left(log(h_1)^{(i)}, ..., log(h_1)^{(T)}\right)$ sampled squared volatilities from the approximating importance density $G(\underline{h}|\underline{\epsilon}^*)$. This term depends on an auxiliary data $\underline{\epsilon}^*$ and must be a close approximation of $\mathbb{P}(\underline{\epsilon},\underline{h})$. We call $G(\underline{h}|\underline{\epsilon}^*)$ the importance density, from which we draw M series of $\underline{h}^{(i)}$ that we denote $(\underline{h}_1^{(i)}, ..., \underline{h}_M^{(i)})$
- $\overline{\omega}$ refers to the average weight and s_ω^2 the variance.

In the setup described by Durbin and Koopman (1997) [17], $G(\underline{h}^{(i)}\underline{\epsilon}^*)$ is the density of the linear states, dependent on the auxiliary data represented by $\underline{\epsilon}^*$. We can obtain a sample of this density by using a simulation smoother and the weights ω_i are evaluated using:

$$\begin{aligned} G\left(\underline{h}^{(i)}|\underline{\epsilon}^*\right) &= \frac{G\left(\underline{\epsilon}^*|\underline{h}^{(i)}\right) G(\underline{h}^{(i)})}{G(\underline{\epsilon}^*)} \\ P(\underline{\epsilon},\underline{h}^{(i)}) &= P(\underline{\epsilon}|\underline{h}^{(i)}P(h^{(i)}) \end{aligned}$$

such that,

$$\begin{aligned}\omega^{(i)} &= \frac{P(\underline{\epsilon},\underline{h}^{(i)})}{G\left(\underline{h}^{(i)}|\underline{\epsilon}^*\right)} = G(\underline{\epsilon}^*)\frac{P(\underline{\epsilon}|\underline{h}^{(i)})P(\underline{h}^{(i)})}{G\left(\underline{\epsilon}^*|\underline{h}^{(i)}\right)G(\underline{h}^{(i)})} \\ &= G(\underline{\epsilon}^*)\frac{P(\underline{\epsilon}|\underline{h}^{(i)})}{G\left(\underline{\epsilon}^*|\underline{h}^{(i)}\right)} = G(\underline{\epsilon}^*)\prod_t \frac{\mathbb{P}(\epsilon_t|h_t^i)}{G(\epsilon_t^*|h_t^i)}\end{aligned}$$

- $G(\underline{h}^{(i)})$ refers to the density of the states as defines in the transition equation (second line of the SV model (2))

- $G(\underline{\epsilon}^*|\underline{h}^{(i)})$ is the density corresponding with the "linearized" observation equation.
- $G(\underline{\epsilon}^*)$ represents the unconditional likelihood of the auxiliary data according to the approximating model and can be estimated using the Kalman Filter equations.

2.2.4. Indirect Inference

In a few words the indirect inference method allows to estimate the parameters of the true model by means of an auxiliary model with which we have a certain relation of "injectivity", and which can be simulated. Thanks to this injectivity relation, the knowledge of the true parameter of the auxiliary model θ_{aux} allows to know the true parameter of interest θ_0. It suffices to simulate the true model a large number of times and see which parameter $\theta \in \Theta$ allows to reduce to the minimum the distance between θ_{aux} and what will be $\hat{\theta}_{aux}$ and we find our estimator $\hat{\theta}$ of the true parameter θ_0. The theory and the implementation of this method will be detailed in section 3.

3. Detailed Description of the Chosen Methods and Implementation

3.1. Motivation

We have chosen the three following methods QML, GMM, Indirect Inference due to their relative simplicity and robustness. We wanted to have at least one method based on simulations (e.g. indirect inference) and also more tradional methods (QML and GMM). We have done some attempts to code some other methods based on Importance Sampling or MCMC but we did not manage to find convincing results.

3.2. The Methods in Details

3.2.1. QML

The method based on a QML estimation that we have chosen to investigate is the one developed in Francq, Zakoïan 2006 [6] and detailed in Francq, Zakoïan 2019 [5] which is based on the linear representation of the stochastic volatility model using an ARMA. The idea is to start from the stochastic volatility model (2), to use its linearization that we have already seen in section 1, and to apply the Kalman filter to the resulting state-space model:

$$\textbf{State-Space Model} \begin{cases} y_t := log(\epsilon_t^2) = log(h_t) + \mu_Z + u_t \\ log(h_t) = \beta log(h_{t-1}) + \omega + \sigma v_t \end{cases}$$

This method is not perfect as it is based on the assumption that $log(\eta_t^2)$ follows a Gaussian distribution, which seems to be empirically refuted. The marginal distribution of innovations has in principle larger tails even if the innovations would follow a normal

distributio, it would not be the case for $log(\eta_t^2)$. If this condition is indeed not satisfied the Kalman filter only offers an approximation of the model. To estimate the parameters we just have to implement the following algorithm which applies the Kalman filter to this problem and then to maximize the log-likelihood obtained.

Let's start with a few notations:

- $\alpha_{t|t-1} = E\left(log(h_t)|\epsilon_1^2, ..., \epsilon_{t-1}^2\right)$
- $P_{t|t-1} = V\left(log(h_t)|\epsilon_1^2, ..., \epsilon_{t-1}^2\right)$

And below is the algorithm that we just need to apply to the data (steps 2 and 3 can be done separately as they do not depend on the data):

1. $\alpha_{1|0} = \beta_0 a_0 + \omega, \qquad P_{1|0} = \beta^2 P_0 + \sigma^2$
2. $F_{t-1|t-2} = P_{t-1|t-2} + \sigma_Z^2, \qquad K_t = \beta P_{t-1|t-2} F_{t-1|t-2}^{-1}$
3. $P_{t|t-1} = \beta^2 P_{t-1|t-2} - K_t^2 F_{t-1|t-2} + \sigma^2$
4. $\alpha_{t|t-1} = \beta \alpha_{t-1|t-2} + K_t(y_{t-1} - \alpha_{t-1|t-2} - \mu_Z) + \omega$

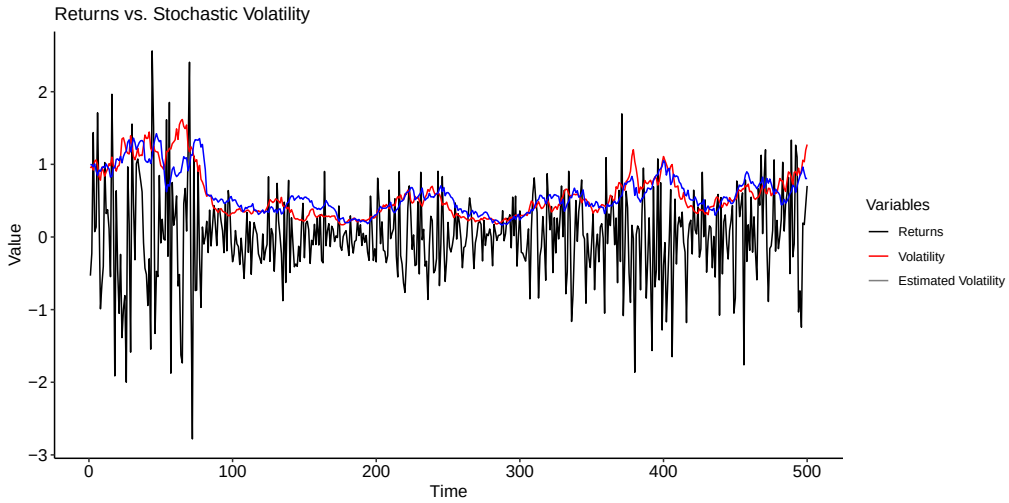
Once we have run this algorithm, we are then ready to minimize the following log-likelihood and get our estimator:

$$\widehat{\theta_T} \in argmin_{\theta \in \Theta} log\, \mathcal{L}(\epsilon_1, ..., \epsilon_T; \theta)$$

$$\text{with } log\, \mathcal{L}(\epsilon_1, ..., \epsilon_T; \theta) = -\frac{1}{T} log 2\pi - \frac{1}{2} \sum_{t=1}^{T} log F_{t|t-1} + \frac{\left(log(\epsilon_t^2) - \alpha_{t|t-1} - \mu_Z\right)^2}{F_{t|t-1}}$$

This was the algorithm that is implemented below. Figure 2 represents a series of log-returns simulated according to the canonical stochastic volatility model (1), the stochastic volatility of the model in red (which is computed with the true parameter θ_0) and the stochastic volatility estimated using the Kalman filter in blue. The estimation of the volatility is obtained using the $\alpha_{t|t-1}$ quantity calculated by the algorithm. Indeed, $\sqrt{exp(\alpha_{t|t-1})}$ is a $\mathcal{F}_{t-1}$-measurable approximation of the volatility $\sqrt{h_t}$

Figure 2. Illustration of a simulated log-returns series (according to the standard model) vs. discrete stochastic volatility model



We notice that the volatility of the model and its approximation by the Kalman filter seem to be quite close, given that the stochastic volatility of the model takes into account an innovation at date t, and which is not known at the time of the forecast (t-1).

3.2.2. GMM

We have already presented the general idea of the GMM method in the previous section, we will now go into the details of its implementation. First of all, we have to choose the right number of moments. This question is not obvious, the only condition is that the number of moments must exceed the size of the parameter. Without going into details, the three papers differ mainly on the number of moments they choose, we do not want to go into this level of detail so we choose to follow the most recent method (that of Andersen and Sorensen 1999 [16]) which is also the one with the most moments Q=24.

We define define the following moments as follows: $m_t(\theta) = (m_{1t}(\theta), ..., m_{Qt}(\theta))$

Selected moments:

- $m_{1t} := \mathbb{E}\left(|\epsilon_t|\right) = (\frac{1}{\pi})^{0.5}\mathbb{E}\left(\sqrt{h_t}\right)$
- $m_{2t} := \mathbb{E}\left(\epsilon_t^2\right) = \mathbb{E}\left(h_t\right)$
- $m_{3t} := \mathbb{E}\left(|\epsilon_t^3|\right) = 2\sqrt{\frac{2}{\pi}}\mathbb{E}(h_t^{3/2})$
- $m_{4t} := \mathbb{E}\left(\epsilon_t^4\right) = 3\mathbb{E}(h_t^2)$
- $m_{j+4,t} := \mathbb{E}\left(|\epsilon_t\epsilon_{t-j}|\right) = \frac{2}{\pi}\mathbb{E}(\sqrt{h_t}\sqrt{h_{t-j}})$
- $m_{j+14,t} := \mathbb{E}\left(\epsilon_t^2\epsilon_{t-j}^2\right) = \mathbb{E}(h_t h_{t-j})$

and their empirical counterparts: $M_T(\theta) = (M_{1T}(\theta), ..., M_{QT}(\theta))$

$M_{it}(\theta) = \sum_{t=j+1}^{T} \frac{m_{it}(\theta)}{T-j}$ for $i \in (1, ..., Q)$ and j is the maximum lag between the variables defining the sample moments.

The idea is to compare the empirical moments observed in the data to the "analytical moments" from which we have access to a closed form expression (from the hypotheses of the model):

$A(\theta) := (m_1(\theta), ..., m_Q(\theta))$

In order to compute the theoretical value of the moments we can rely on the following formulae:

$\mathbb{E}\left(\sqrt{h_t}^r\right) = exp\left(\frac{r\mu}{2} + \frac{r^2\sigma_h^2}{8}\right)$ for j a positive integer and r,s positive constants

$\mathbb{E}\left(\sqrt{h_t}^r \sqrt{h_t}^s\right) = \mathbb{E}\left(\sqrt{h_t}^r \sqrt{h_t}^s exp(\frac{rs\beta^j\sigma_h^2}{4}\right)$

$\mu = \frac{\omega}{1-\beta}$

$\sigma_h = \frac{\sigma^2}{1-\beta^2}$

The idea is then to minimize the following quantity and find the optimal parameter $\hat{\theta}_T$ doing so:

$\hat{\theta}_T = argmin_{\theta\in\Theta}(M_T(\theta) - A(\theta))'\Lambda_T^{-1}(M_T(\theta) - A(\theta))$

with Λ_T a positive definite random weighting matrix for which we will now detail the method of selection adopted. The optimal Λ is given by:

$$\Lambda = lim_{T\to\infty}\mathbb{E}\left(\sum_{t,\tau}^{T} \frac{(m_t - A(\theta_0)))\,(m_\tau - A(\theta_0)))'}{T}\right)$$

which can be approximated by:

$$\sum_{j=-T+1}^{T} k(j)\hat{\Gamma}_T(j)$$

- with k(j) weights that may become 0 for $|j| > L_T$ (we can select L_T=10 for instance - this is a lag truncation parameter)
- $\hat{\Gamma}_T(j) = \frac{1}{T}\sum_{t=j+1}^{T}\left(m_t(\hat{\theta}) - A(\hat{\theta})\right)\left(m_{t-j}(\hat{\theta}) - A(\hat{\theta})\right)'$

All that remains is to apply the algorithm to the data.

3.2.3. Indirect Inference

The indirect inference method has been introduced in Gouriéroux, Monfort and Renault 1993 [18] and is applied by Monfardini 1998 [19] to the issue of stochastic volatility model estimation.

Sometimes it is easier to rewrite the model this way:

$$\text{(4) } \textbf{Standard SV Model 2nd Form.} \quad \begin{cases} y_t &= exp\left(\frac{1}{2}h_t\right) u_t \\ h_t &= \mu + \rho h_{t-1} + v_t \end{cases}$$

where $u_t \sim I.I.N(0,1)$ and $v_t \sim I.I.N(0,\sigma^2)$ for $t = 1, ..., T$ and with parameter of interest $\theta' = (\mu, \rho, \sigma^2)$.

Consider an initial model (denoted M_θ) and its structural parameter $\theta \in \Omega \subset \mathbb{R}^{\text{I}}$ with true value θ_0. θ_0 is the parameter of interest. Assume that M_θ is too complex so that standard maximum likelihood techniques are not possible. Assume also that M_θ can be easily simulated for any value of θ.

Consider an approximated model (denoted M^a_β) of M_θ, and its parameter $\beta \in B \subset \mathbb{R}^{\text{II}}$ with true value β_0. We call M^a_β the auxiliary or instrumental model of M_θ, and β the auxiliary parameter.

If one constructs an estimator of the structural true parameter θ_0 using a criterion based on M^a_β, say the log-likelihood of M^a_β, this estimator would (generally) be inconsistent. Indeed, it would rely on the approximated (or misspecified) model M^a_β and not on the initial well-specified model M_θ. However, one could think in using this estimator in an indirect estimation procedure to get a consistent estimator of θ_0. This is the purpose of indirect inference.

Indirect inference principle consists in applying an estimation method both on observations and on simulations of M_θ (therefore, drawn as a function of θ), and then select the value of θ that puts as close as possible the estimator obtained from simulations to its counterpart obtained from observations. Formally, this method can be decomposed in two-steps.

In the first step, one can get an estimator of β_0, denoted $\hat{\beta}_T$, from the T observations $\underline{y_T} = (y_1, ..., y_T)$ using the auxiliary criterion Q_T (e.g. the log-likelihood of M^a_β) :

$$\hat{\beta}_T = \arg\max_\beta Q_T(\underline{y_T}, \beta).$$

The observations $\underline{y_T}$ are assumed to be generated by the initial true model M_{θ_0}.

Before continuing to the second step, it will be useful to introduce the *binding function* defined by:

$$b(\theta) = \arg\max_{\beta} Q_\infty(\theta, \beta).$$

If the conditions are met, $\hat{\beta}_T$ is a consistent estimator of β_0, and β_0 is the solution of the limit problem $\beta_0 = \arg\max_\beta Q_\infty(\theta_0, \beta)$, i.e. $\beta_0 = b(\theta_0)$. From here, one could define an estimator of θ_0 as the solution $\hat{\theta}_T$ of $\hat{\beta}_T = b(\hat{\theta}_T)$. Yet, the binding function may be either unknown or at least difficult to compute. The second step of the indirect estimation follows the previous idea by obtaining a functional estimator of $b(\cdot)$.

In the second step, one can simulate H times the initial model M_θ for a given value of θ and collect the corresponding simulated data $\{\underline{y_T^h}(\theta) = (y_1^h, ..., y_T^h), h = 1, ..., H\}$. From this TH simulated data, one can get H estimators of $b(\theta)$, denoted $\{\hat{\beta}_T^h(\theta), h = 1, ..., H\}$, using the same auxiliary criterion Q_T :

$$\hat{\beta}_T^h(\theta) = \arg\max_{\beta} Q_T(\underline{y_T^h}(\theta), \beta).$$

Then, one gets an estimator of $b(\theta)$, denoted $\hat{\beta}_{HT}(\theta)$, by averaging the H estimators $\hat{\beta}_T^h(\theta)$:

$$\hat{\beta}_{HT}(\theta) = \frac{1}{H} \sum_{h=1}^{H} \hat{\beta}_T^h(\theta).$$

If, for each θ, $\hat{\beta}_{HT}(\theta)$ is a consistent estimator of $b(\theta)$ then $\hat{\beta}_{HT}(\cdot)$ is a consistent functional estimator of $b(\cdot)$. In particular $\hat{\beta}_{HT}(\theta_0)$ is a consistent estimator of $b(\theta_0) = \beta_0$.

Intuitively, $\hat{\beta}_T$ and $\hat{\beta}_{HT}(\theta_0)$ should be close to each other for large T. Therefore, one can define the estimator of θ, $\hat{\theta}_{HT}$, as the value of θ that puts $\hat{\beta}_{HT}(\theta)$ as close as possible to $\hat{\beta}_T$. Formally, $\hat{\theta}_{HT}$ is defined as the solution of a minimum distance problem:

$$\hat{\theta}_{HT} = \arg\min_{\theta} [\hat{\beta}_T - \hat{\beta}_{HT}(\theta)]' \hat{\Omega}_T [\hat{\beta}_T - \hat{\beta}_{HT}(\theta)]$$

where $\hat{\Omega}_T$ is a positive definite matrix converging to a deterministic positive definite matrix Ω.

Among the two alternatives presented by the Monfardini, we focus on the one based on the $ARMA(1,1)$ representation (to be in a sense consistent with the approach chosen for the QML approach above):

$$x_t = \alpha_0^* + \alpha_1^* x_{t-1} + \omega_t - \alpha_2^* \omega_{t-1}, \omega_t \sim I.I.N(0, \nu^2).$$

This auxiliary model M_α^{ARMA} has a 4-dimensional parameter α_0 with true value $\alpha = (\alpha_0^*, \alpha_1^*, \alpha_2^*, \nu^2)'$, the auxiliary parameters of interest.

3.3. Monte Carlo Experiments

3.3.1. QML

We run a Monte Carlo experiment with M=1,000 independent draws to illustrate the consistence and the speed of convergence of the estimator. We represent for each coefficient ω, β, σ its asymptotic behavior: $\sqrt{T}\left(\hat{\theta}_T - \theta_0\right)$ for different sample size T=500, T=1,000, T=3,000, T=5,000.

Figure 3. $\sqrt{T}\left(\hat{\omega}_T - \omega_0\right)$

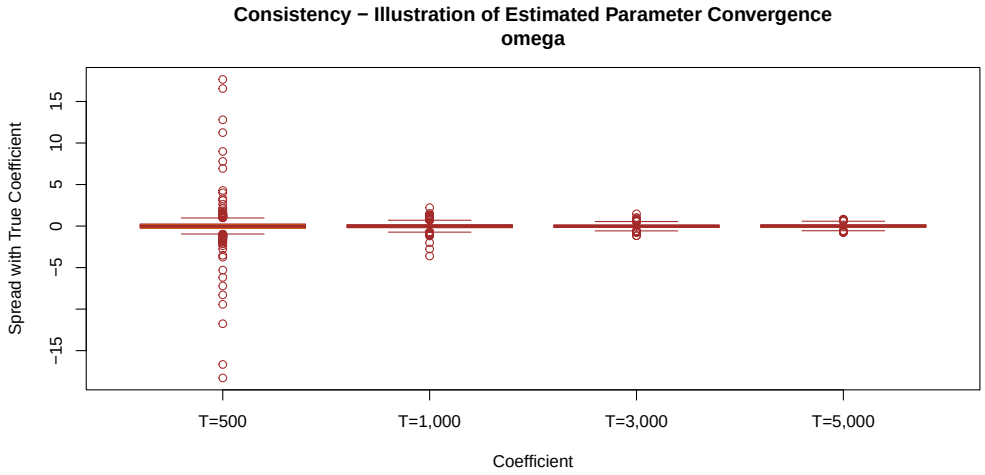


Figure 4. $\sqrt{T}\left(\hat{\beta}_T - \beta_0\right)$

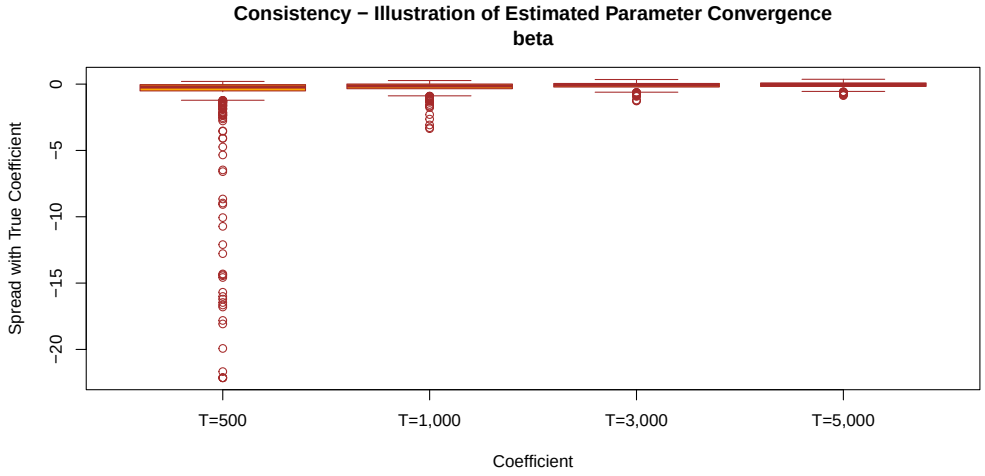
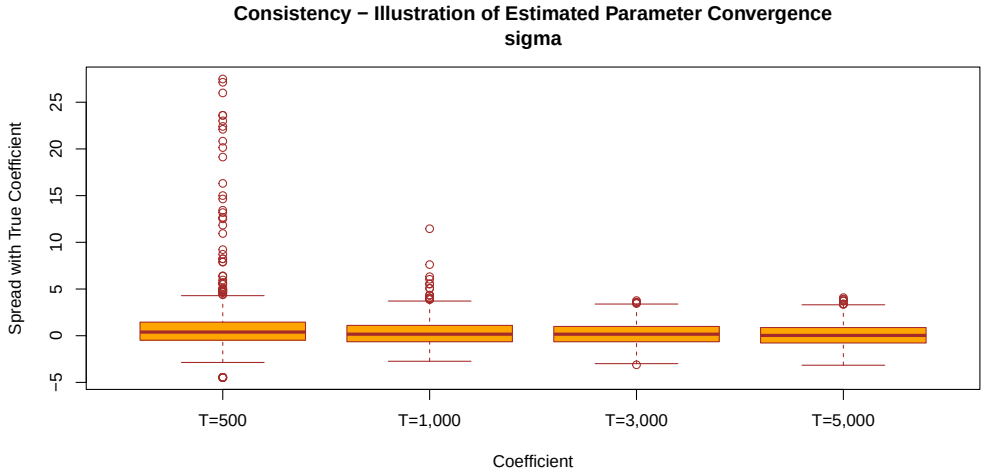


Figure 5. $\sqrt{T}\left(\hat{\sigma}_T - \sigma_0\right)$



Figures 3 to 5 seem to indicate the consistency of the estimator and a normal asymptotic speed of convergence, which proves the relatively good performance of this method of estimation, which we will therefore use in the application section.

3.3.2. GMM

For the GMM method, the convergence of the estimator to the true parameter seems less efficient than in the previous case with the QML approach. However we find similar results to those of Andersen and Sorensen (see table below). These results are obtained for T=2,000 and $theta_0 = (-0.736, 0.900, 0.363)$:

Table 1. Consistency Result

	Our Results	Results from Andersen & Sorensen
$\lvert\omega_0 - \hat{\omega_T}\rvert$	0.108	0.103
$\lvert\beta_0 - \hat{\beta_T}\rvert$	0.016	0.013
$\lvert\sigma_0 - \hat{\sigma_T}\rvert$	0.105	0.054

3.3.3. Indirect Inference

Work in progress

4. Applications

4.1. Risk Measures Estimation

4.1.1. Estimating Value at Risk

Estimating risk measures is one of the most natural applications of volatility models. The **Value at Risk** (VaR) is the most commonly used measure to account for the risk of a financial time series:

$$VaR(\alpha) = -inf\{\epsilon \in \Omega_\epsilon : F_\Upsilon(\epsilon) > \alpha\} = -F_\Upsilon^{-1}(\alpha)$$

- With ϵ a log-return in the set of taken values Ω_ϵ.
- Υ is the distribution of the ϵ_t for t $\in (1, ..., T)$.
- F_Υ is the cdf of the distribution Υ.
- F_Υ^{-1} is the is the inverse function of F_Υ.
- α is the risk level under consideration.

In order to estimate it, we will focus on what we call the **Conditional** Value at Risk at risk level α (conditional to the information set available at *t-1*) is defined as follows:

$$VaR_{t-1}(\alpha) = -q_\alpha\left(\epsilon_t | \mathcal{F}_{t-1}\right)$$

is minus the quantile of order α for the log-returns at date t knowing all the past information.

Its greatest advantage is that it does not require the existence of any moment to exist, which is not the case for some other risk measures such as the (Conditional) **Expected Shortfall** for instance:

$$ES_{t-1}(\alpha) \;=\; E_{t-1}\,\left(\epsilon_t \,|\, \epsilon_t \;<\; -VaR_{t-1}(\alpha)\right)$$

The estimation of the VaR is relatively easy with a GARCH model estimated by QML by applying, for example, the 2-step method presented by Francq and Zakoïan 2015 [20]. This method relies on the fact that the volatility $\sqrt{h_t}$ is $\mathcal{F}_{t-1}$ measurable and its positive homogenity property (see appendix). Indeed, it can be expressed in the form:

$$VaR_{t-1}(\alpha) \;=\; -\sqrt{h_t}(\theta_0)\,\xi_\alpha\,:$$

- ξ_α refers to the $\alpha - quantile$ of the innovations η_t.

Therefore, to estimate the conditional VaR we use the traditional two-step method consisting in obtaining a constant and asymptotically normal (CAN) estimator by Gaussian QML estimation in a first step. And then constructing estimations of the innovations (i.e. $\hat{\eta}_t$) as well as an estimator of the $\alpha - quantile$ (i.e. -$\hat{\xi}_T$). We will then consider the following estimator:

$$\widehat{VaR}_{t-1}(\alpha) \;=\; -\sqrt{h_t}(\hat{\theta}_T)\,\hat{\xi}_T\,:$$

- $\hat{\theta}_T = (\hat{\omega}, \hat{\alpha}, \hat{\beta})$ refers to the estimated paramater obtained by Gaussian QML.
- $\hat{\xi}_T$ is obtained taking the $\alpha - quantile$ of the residuals $\hat{\eta}_t$ defined as: $\hat{\eta}_t = \frac{\epsilon_t}{\sqrt{h_t}(\hat{\theta}_T)}$

However, this method cannot be applied as is to the stochastic volatility model since volatility $\sqrt{h_t}$ is not $\mathcal{F}_{t-1}$-measurable in this case. However, we have seen previously when applying the QML method to the state-space model and using the Kalman filter we could obtain an approximation of the volatility $\sqrt{exp(\alpha_{t|t-1})}$, which is $\mathcal{F}_{t-1}$-measurable.

We can then reconstruct an estimator of the innovations $(\hat{\eta_1}, ..., \hat{\eta_T})$ and thus apply the two-step method as in the case of a GARCH model. This method is not rigorous strictly speaking because in doing so we obtain approximations of the residuals $(\hat{\eta}_t)$, but we do not necessarily have consistency, since the $(\alpha_{t|t-1})$ are only an approximation of the volatility under the assumption of normality of $log(\eta_t^2)$, which is not verified (in the general case). However, we can assume that the error is quite small and we will backtest the model to be convinced. Below is an example of the VaR of the log-returns of the S&P500 at risk level, 1%, 5% and 20%, centered around the period of the "Covid crisis":

Figure 6. Illustration of the (Conditional) VaR(1%) of the log-returns of the S&P 500 during the Covid crisis

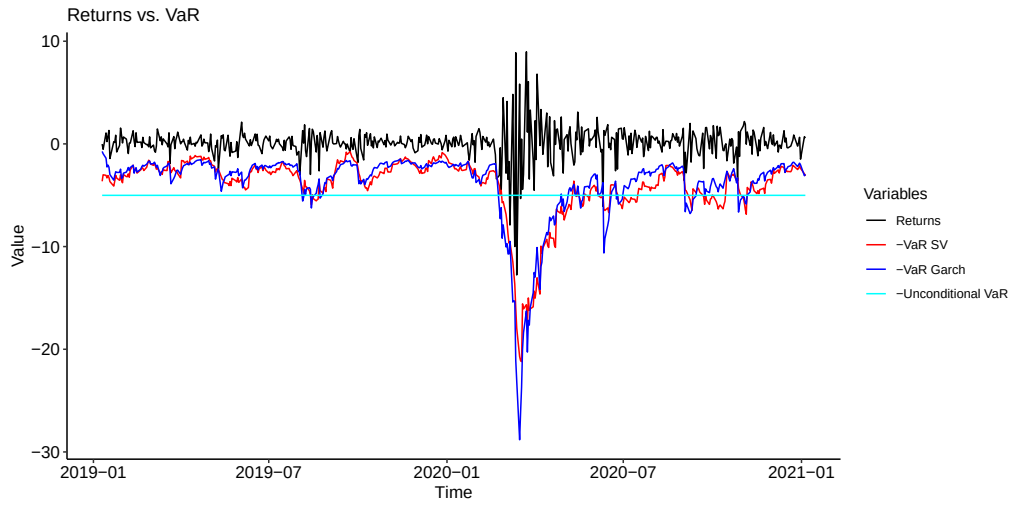


Figure 7. Illustration of the (Conditional) VaR(5%) of the log-returns of the S&P 500 during the Covid crisis

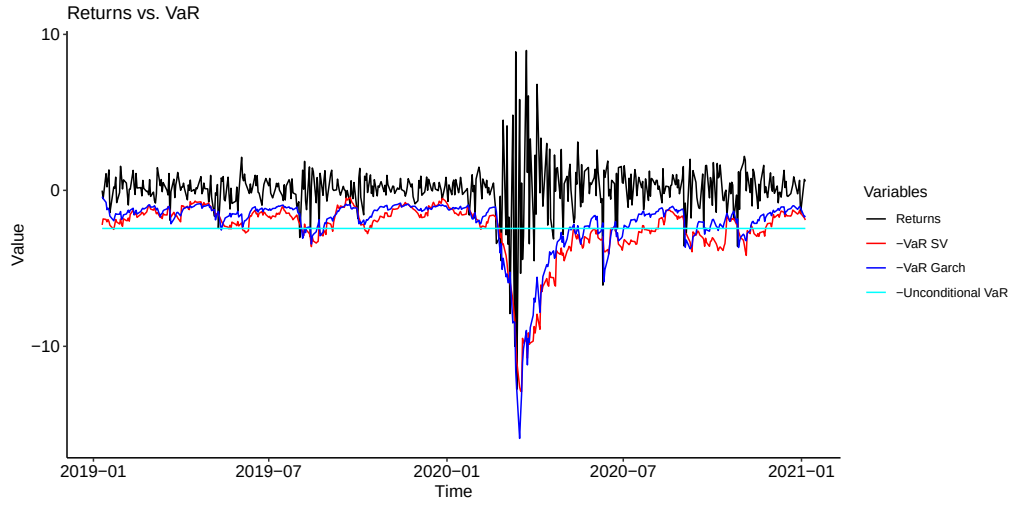
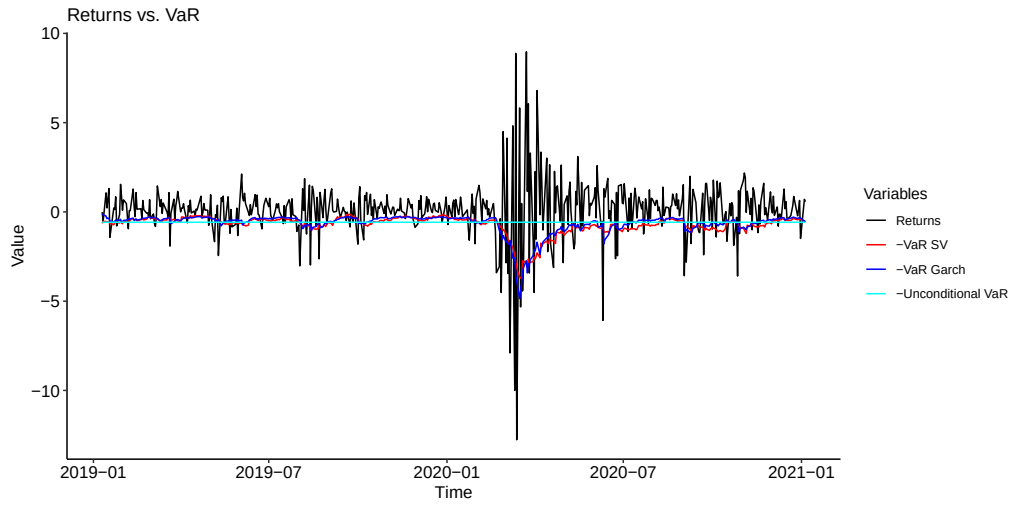


Figure 8. Illustration of the (Conditional) VaR(20%) of the log-returns of the S&P 500 during the Covid crisis



Let's mention the calculation of the (Conditional) Expected Shortfall, following the same methodology.

The estimation approach for the other conditional risk measures flows directly from the one used for VaR, the conditional Expected Shortfall is the first example and is expressed as follows:

$$\begin{aligned} ES_{t-1}(\alpha) &= E_{t-1}\left[\,\epsilon_t \,|\, \epsilon_t < -VaR_{t-1}(\alpha)\,\right] \\ &= E_{t-1}\left[\,\epsilon_t \,|\, \eta_t < -\xi_\alpha\,\right] \\ &= \sqrt{h_t}(\theta_0)\, E\left[\,\eta_t \,|\, \eta_t < -\xi_\alpha\,\right] \end{aligned}$$

We will then use the following empirical estimator:

$$\widehat{ES}_{t-1}(\alpha) = \frac{\sqrt{h_t}(\hat{\theta}_T)}{\tilde{T}} \sum_{t=1}^{T} \hat{\eta}_t \, 1\{\hat{\eta}_t \,|\, \hat{\eta}_t < -\hat{\xi}_T\}$$

- $\tilde{T} = \sum_{t=1}^{T} 1\{\hat{\eta}_t \,|\, \hat{\eta}_t < -\hat{\xi}_T\}$

Although the Expected Shortfall seems to be more satisfactory than VaR (as VaR is not sub-additive), it could be criticised for weighting uniformly all returns lower than the opposite of VaR. Indeed, if one assumes convex and not linear risk aversion - the

practitioner being more averse to the risk of larger losses compared to smaller losses - it may be interesting to consider distortion risk measures.

Figure 9. Illustration of the (Conditional) Expected Shortfall (1%) of the log-returns of the S&P 500 during the Covid crisis

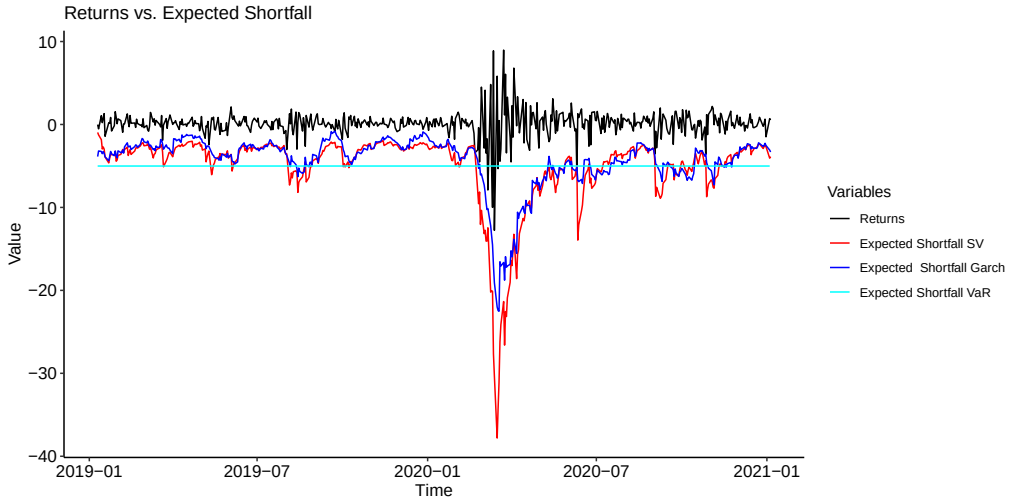


Figure 10. Illustration of the (Conditional) Expected Shortfall (5%) of the log-returns of the S&P 500 during the Covid crisis

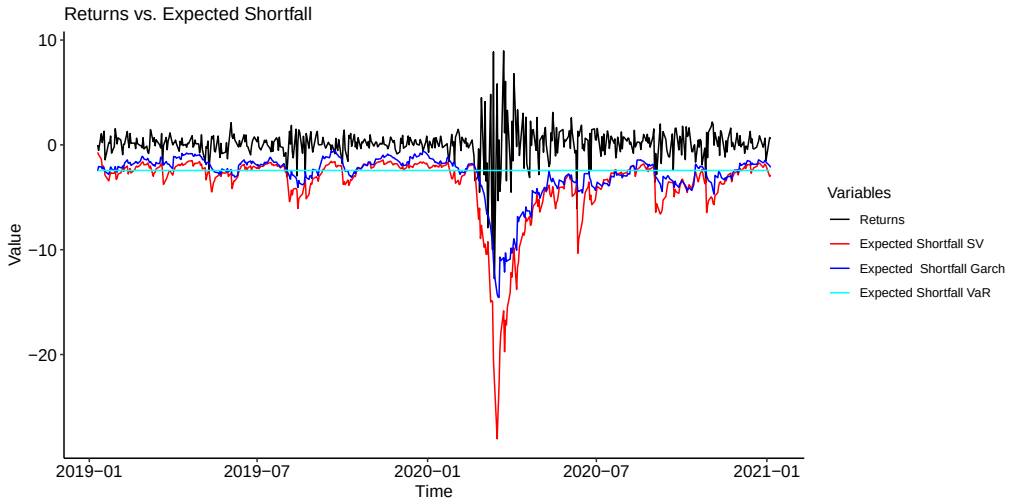
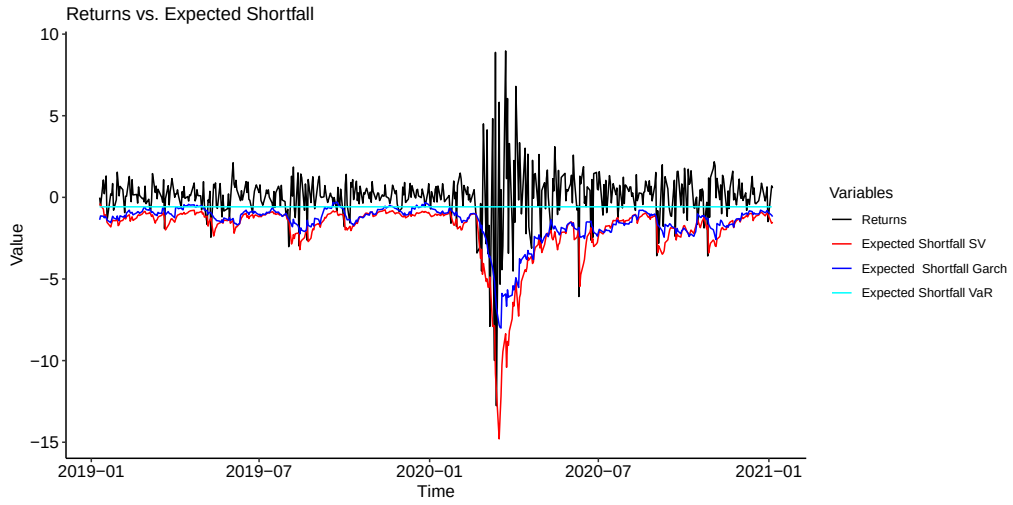


Figure 11. Illustration of the (Conditional) Expected Shortfall (20%) of the log-returns of the S&P 500 during the Covid crisis



We can see that both VaR and ES estimators are pretty close, therefore our method seems credible and we will backtest both methods on a large sample in order to determine, which method is more promising. Let's present quickly the traditional tests and the chosen backtesting methodology to evaluate the performance of our models.

4.1.2. Backtesting Methodology

Before we get to the core of the subject and present the various existing tests, let's have a brief presentation of the backtesting procedure. The backtest is performed on real data. The data set is composed of the log-returns of basket of 1,489 US stocks (that we have selected to have less than 1% missing data compared to the yearly market trading days). Each series is divided into two periods: an estimation period (called "in-sample") which is used, as its name suggests, to estimate the parameters of the model as well as to estimate the quantile of the innovations (here from 02/05/2009 to 01/05/2021), and a backtesting period (called "out-of-sample" - here from 01/06/2021 to 12/31/2021), which is used to evaluate the performance of the predictions resulting from the model - which has been calibrated on the in-sample estimates. The in-sample period, which is made up of observations $1,\ldots,T$ has been entirely observed on date T, whereas the out-of-sample period is experienced "day after day" and is made up of the n instants $T+1,\ldots,T+n$. We have chosen here to take a long estimation period $T=3,000$ and a backtesting period of 250 observations (i.e. 1 year of trading) as illustrated with the chronology below. It should be pointed out that the tests that will be presented do not take into account the estimation risk and consider that the true parameter θ_0 is observed. It is in order to

limit this estimation error that we take a large in-sample size *T=3,000* and the test is therefore applied only to the 250 data of the "out-of-sample" period.

4.1.3. Quick Overview of the Existing Tests

There are different types of tests, but the most frequent are those based on what we call the "Hit Variable" (defined below), which is simply a violation of the VaR. A violation is any moment t at which $\epsilon_t < -VaR_{t-1}(\alpha)$. Let's define the hit variable:

$$Hit_t(\alpha) = 1\{\epsilon_t < -VaR_{t-1}(\alpha)\}$$

The most commonly used tests are the Kupiec (1995) [21] - Christoffersen (1998) [22] tests, which - as we just mentioned - overlook the estimation error. The aim here is to check both that the number of violations corresponds to the risk level α and the independence of these violations.

4.1.4. Unconditional Coverage Test

The idea of the test relies on the fact that if $VaR_{t-1}(\alpha)$ really is a VaR measure at a α risk level, then we have:

$P[\epsilon_t < -VaR_{t-1}(\alpha)] = \alpha$ and $Hit_t \sim \mathcal{B}(\alpha)$.

- $\mathcal{B}(\alpha)$ refers to a Bernoulli distribution with paramater α.

Therefore we define the Null hypothesis: $H_0^{UC} : P[Hit_t = 1] = \alpha$ and the test statistics:

$LR_{UC} = 2log\frac{\pi_{exp}^{n_1}(1-\pi_{exp})^{n_0}}{\pi_{obs}^{n_1}(1-\pi_{obs})^{n_0}}$

- π_{exp} is the expected proportion of violations.
- π_{obs} is the observed proportion of violations.
- n_1 is the number of violations and $n_0 = n - n_1$ is the sample size.

$LR_{UC} \sim \chi_1^2$ under the null hypothesis H_0^{UC}. Hence the rejection area: $\{LR_{UC} > \chi_1^2(1-\underline{\alpha})\}$ at the confidence level $\underline{\alpha}$.

4.1.5. Independence Test

As $VaR_{t-1}(\alpha)$ is a conditional VaR for a risk level α, we have:

$$\begin{aligned} E[Hit_t\, Hit_{t+k}] &= E[E_{t+k-1}[Hit_t\, Hit_{t+k}]] \\ &= E[Hit_t E_{t+k-1}[Hit_{t+k}]] \\ &= \alpha\, E[Hit_t] \\ &= \alpha^2 \end{aligned}$$

Hence, $CoV(Hit_t, Hit_{t+k}) = 0$, which is equivalent to Hit_t and Hit_{t+k} are independents. Eventually, we have:

$(Hit_t)\, iid \sim \mathcal{B}(\alpha)$

Therefore we define the Null Hypothesis: $H_0^{Ind} : P[Hit_t = 1|Hit_{t-1} = 0] = P[Hit_t = 1|Hit_{t-1} = 1]$ and the test statistics:

$$LR_{Ind} = 2log\frac{\pi_{obs}^{n_1}(1-\pi_{obs})^{n_0}}{\pi_{01}^{n_{01}}(1-\pi_{01})^{n_{00}}\,\pi_{11}^{n_{11}}(1-\pi_{11})^{n_{10}}}$$

- n_{ij} is the number of indicator i followed by indicator j.
- $\pi_{01} = \frac{n_{01}}{(n_{00}+n_{01})}$ and $\pi_{11} = \frac{n_{11}}{(n_{10}+n_{11})}$

$LR_{Ind} \sim \chi_1^2$ under the null hypothesis H_0^{Ind}. Hence the rejection area: $\{LR_{Ind} > \chi_1^2(1-\underline{\alpha})\}$ at the confidence level $\underline{\alpha}$.

4.1.6. Conditional Coverage Test

The Conditional Coverage Test tests for both effects (consistency and independence). The test statistics take the following form:

$$LR_{cc} = 2log\frac{\pi_{exp}^{n_1}(1-\pi_{exp})^{n_0}}{\pi_{01}^{n_{01}}(1-\pi_{01})^{n_{00}}\,\pi_{11}^{n_{11}}(1-\pi_{11})^{n_{10}}}$$

Remark: LR_{cc} can easily been calculated if we already have LR_{UC} and LR_{Ind}, as $LR_{cc} = LR_{UC} + LR_{Ind}$.

$LR_{cc} \sim \chi_2^2$ under the null hypothesis. Hence the rejection area: $\{LR_{cc} > \chi_2^2(1-\underline{\alpha})\}$ at the confidence level $\underline{\alpha}$.

The three tests proposed by Christoffersen described above (Unconditional Coverage test, Independence Test and Conditional Coverage test) have the drawback that they only test the model for a particular α. However, the model could work for a well-chosen α and perform much worse for another α. This is why tests applied to multiple risk levels have been introduced and we will describe one of them in the following sub-section.

4.1.7. Test on Multiple Risk Levels

In terms of tests ran at different risks level we focused on the multivariate Portmanteau test proposed by Hurlin (2007) [23]. Let's define first what we call the "α-hit variable" as described in the quoted paper:

$$\tilde{Hit}_t(\alpha) = \left\{ \begin{array}{ll} 1-\alpha & \text{if } \epsilon_t < -VaR_{t-1}(\alpha) \\ -\alpha & \text{otherwise.} \end{array} \right\}$$

the Hit vector for several risk levels $\alpha_1, ..., \alpha_m$:

$\tilde{Hit}_t(\alpha) = [\tilde{Hit}_t(\alpha_1) : \tilde{Hit}_t(\alpha_2) : ... : \tilde{Hit}_t(\alpha_m)]$ and the empirical covariance matrix:

$\hat{C}_k = \sum_{t=k+1}^{T} \tilde{Hit}_t(\alpha).\tilde{Hit}'_{t-k}(\alpha)$

Eventually comes the multivariate Portmanteau test statistics first introduced by Hosking (1980) [24]:

$Q_m(K) = T^2.\sum_{k=1}^{K}(T-k)^{-1}tr(\hat{C}'_k\hat{C_k^{-1}}\hat{C}_k\hat{C_k^{-1}})$

- tr(.) designs the trace function

$Q_m(K) \sim \chi^2_{Km^2}$ when $d \to \infty$ under the null hypothesis. Hence the rejection area: $\{Q_m(K) > \chi^2_{Km^2}(1-\underline{\alpha})\}$ at the confidence level vector $\underline{\alpha} = (\underline{\alpha}_1, ..., \underline{\alpha}_m)$.

We therefore submitted the two VaR models (based on GARCH and SV models) to both the unconditional and independence tests.

In addition to the tests based on the number of violations, it may be interesting to look at the distance between the VaR and the log-return at each time t, thus providing a measure of the goodness of fit of the model. Indeed, a model that constantly overestimates the required reserves, even in quiet periods, would be a bad model - the money in reserve being a cost for the financial institution. The following measure, which we call the α-criterion, represents a measure of this distance between VaR and log-returns and allows us to compare two models (in this case the conditional model and the unconditional model).

4.1.8. α-Criterion

The idea of this criterion is to weight negative exceedances - where log-returns are below VaR - more strongly than positive exceedances (the former being more costly for the bank). Let's define the criterion:

$$C_\alpha = E[(1-\alpha).[\epsilon_t - VaR_{t-1}(\alpha)]^- + \alpha.[\epsilon_t - VaR_{t-1}(\alpha)]^+]$$

- $x^+ = max(x, 0) = (-x)^-$

and an estimator:

$$\widehat{C}_\alpha = T^{-1}.\sum_{t=1}^{T}(1-\alpha).[\epsilon_t - VaR_{t-1}(\alpha)]^- + \alpha[\epsilon_t - VaR_{t-1}(\alpha)]^+$$

4.1.9. Results

The following tables show the results of the backtests performed according to the procedure described above for risk levels of α=1%, 5% and 20% respectively. Two tests are performed, namely the unconditional coverage test and the independence test on nearly 1,500 stocks (note that their behavior is not independent and may even be strongly correlated). The numbers presented in these three tables represent the frequency for which the null hypothesis was rejected (i.e. the VaR does not satisfy the property in question). The tests are performed with a probability of Null rejection of 5%, so if the series were independent, one would expect a theoretical rejection frequency of $\underline{\alpha} = 5\%$. However, before interpreting it should be remembered that these tests do not take into account the estimation risk. We compare three VaR models: the model based on stochastic volatility as presented previously, the model based on a GARCH model (such as the one mentioned in Francq, Zakoïan (2015) [20]) and an unconditional VaR model that is computed as the quantile of the in-sample log-returns and therefore does not vary over the out-of-sample period.

As for the results, we observe that the GARCH model is rejected less often than the stochastic volatility model and therefore seems to perform better. This can be explained by two reasons, i) the GARCH model is indeed more efficient, notably because the method chosen for the SV is less rigorous; ii) the estimation risk for the stochastic volatility model may be higher. Nevertheless, the rejection rate of the stochastic volatility model turns out to be much better than that of the unconditional volatility and is therefore worth considering. Moreover, the model based on stochastic volatility seems to have a lower α-criterion than the GARCH model, which is a positive point for a company wishing to reduce its reserves, especially during a period of calm in the markets.

Table 2. Backtesting Results
Frequency of Null Rejection (over 1,489 stocks)

Risk level $\alpha = 1\%$	UC	Ind.	α-criterion
SV Model	0.208	0.017	0.171
GARCH Model	0.107	0.013	0.199
Unconditional Model	0.410	0.017	0.155

Table 3. Backtesting Results
Frequency of Null Rejection (over 1,489 stocks)

Risk level $\alpha = 5\%$	UC	Ind.	α-criterion
SV Model	0.187	0.056	0.467
GARCH Model	0.105	0.044	0.486
Unconditional Model	0.349	0.083	0.464

Table 4. Backtesting Results
Frequency of Null Rejection (over 1,489 stocks)

Risk level $\alpha = 20\%$	UC	Ind.	α-criterion
SV Model	0.148	0.093	0.960
GARCH Model	0.123	0.077	0.966
Unconditional Model	0.303	0.091	0.965

4.2. Volatility Risk Premium

To illustrate the potential use of the stochastic volatility model, we can also look at the so-called "volatility risk premium" - i.e. the difference between the implied volatility (definition below) and the "realized volatility":

$$(\textbf{Realized Volatility})\ RV_t = \sum_{i=t-N}^{t} \epsilon_i^2$$

where N is chosen by the practitioner (e.g. 252 for realized volatility computed over the past trading year, 21 when it is computed over the last month etc.).

This indicator does not give a good account of a short-term shock to volatility (as it is looking backward by definition). This is why we prefer to look at what we call the "conditional-stochastic volatility risk premium", which we define as the difference between the implied volatility and the conditional or stochastic volatility calculated with a GARCH or SV model, for example. The advantage of this measure is that it compares data calculated each day by overweighting the most recent events (with the decaying property of conditional and stochastic volatility models). The use of the stochastic volatility model is preferred to a GARCH model because the SV model is the direct discretization of the model used in the calculation of the implied volatility (through the Black Scholes formula).

The implied volatility is the volatility induced by the market option prices. Indeed, the calculation of the price of a large number of options is based on the Black-Scholes model, whose only "real" unknown is the volatility. By starting from the market price and the inverse relationship, we can indeed reconstruct the volatility induced by the price of the derivative product.

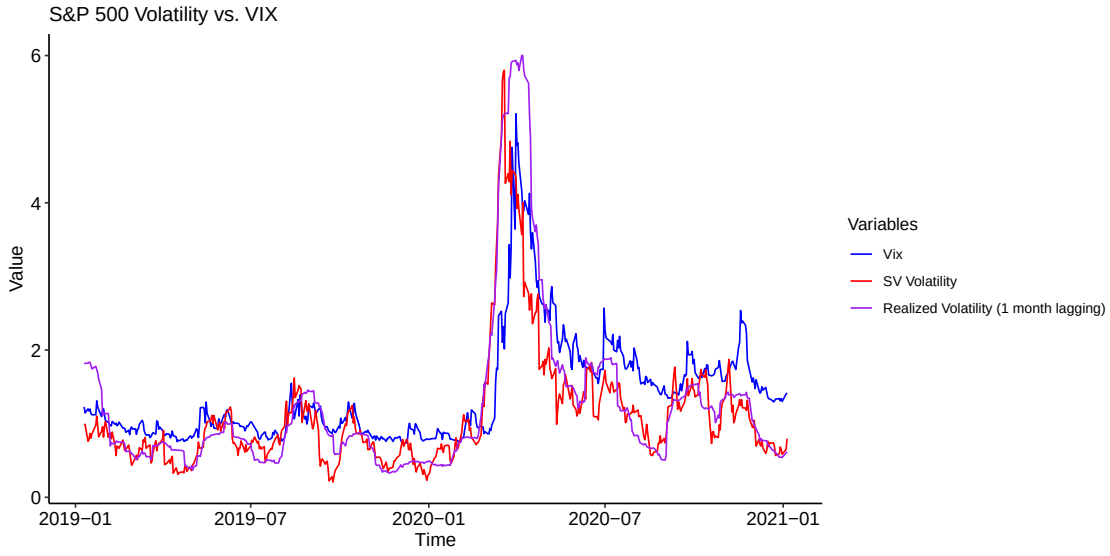
Example of the Black-Scholes formula for a call option:

$$
\begin{aligned}
C(S_t, t) &= N(d1)S_t - N(d_2)Ke^{-r(T-t)} \\
d_1 &= \frac{1}{\sigma\sqrt{T-t}}\left[ln\left(\frac{S_t}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)\right]] \\
d_2 &= d_1 - \sigma\sqrt{T-t}
\end{aligned}
$$

- With C the price of the call option depending on the underlying asset, whose price at date t is denoted S_t
- K is the strike price (price at which we can buy at maturity)
- σ refers to the volatility (our implied volatility of interest)
- r refers to the interest rate and T the time of the maturity
- N(.) refers to the standard normal cumulative distribution function $N(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{\frac{-z^2}{2}}dz$

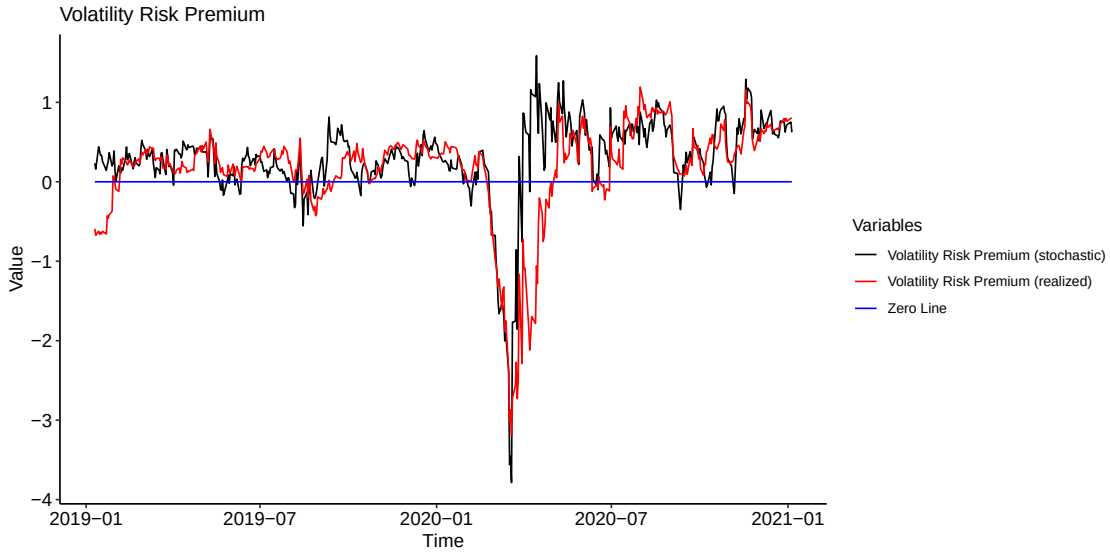
The (CBOE) VIX is the most common measure used to track the implied volatility of the market and is based on the price of options whose underlying is the S&P 500. An increase in the volatility risk premium indicates a sense of market anxiety, with investors tending to overprice relative to observed volatility. This may reflect an expectation of a downturn that has not yet been taken into account by the market. Conversely, if we believe traditional financial theory, which maintains that in a market without friction and with perfect information, prices immediately reflect all material and public information, the volatility risk premium would constitute an arbitrage opportunity for the speculator who, with the help of more or less sophisticated tools, can take a position on the increase or decrease of the volatility risk premium. To get an idea of the premium at date *t+1*, one can use volatility models such as GARCH or stochastic volatility models (with filtering). The two figures below represent respectively a comparison of the different volatilities and a representation of the volatility risk premium (and what we called the conditional/stochastic volatility risk premium).

Figure 12. Comparison of Different Volatilities: Implied (VIX), Stochastic and Realized (over past 1 month)



We notice that the implied volatility (VIX), which should be a forward indicator, as it refers to the future market volatility expected by investors, is in fact lagging the conditional volatility. We also note that the temporality of realized volatility calculated over the past month is about the same as that of implied volatility. This observation is shared by Adhikari et al (2014) [25] who argue that the VIX would indeed lag the market by one month. Thus, there would eventually be arbitrage opportunities in the volatility market. Let's look at the next chart which shows the evolution of VRPs during the Covid crisis period.

Figure 13. Volatility Risk Premia



We can see that, as expected, the VRP is generally positive because investors tend to keep a margin of risk in excess of the actual or predicted market risk. However, as we saw in the previous figure, implied volatility is lagging, which explains a negative VRP at the beginning of the health crisis, due to the fact that investors took time to realize the risk of the pandemic and to effectively price this risk. Interestingly, the VRP also tends to increase in the period following a crisis, which may indicate increased risk aversion on the part of investors following a crisis (i.e. after having potentially experiencing losses).

5. Conclusion

Through this paper, we wished to give some hints about the main existing methods for estimating a standard stochastic volatility model and have implemented some of them. We have mainly focused on the estimation method by QML with an ARMA representation developed by Francq and Zakoïan (2006) [6], which appears to be one of the simplest and most robust method we have seen (some methods may strongly depend on the inputs of the estimation algorithm as well as on the space considered Θ for $\hat{\theta}$). We have performed a Monte Carlo experiment which shows a reasonable convergence of the parameter estimated via this method towards the true parameter θ_0. It is thus on this method that we relied for our two applications. Namely, the proposal of the approximation of risk measures through a two-step approach similar to the one described by Francq and Zakoïan (2015) [20] initially developed for the GARCH model and modified to suit the case of stochastic volatility. We notice that this method (although if its application in this case is

not very rigorous) can be used in practice if we want to work with a stochastic volatility model. The backtest on real data seems to indicate an outperformance of the GARCH model though. The second application allows us to construct the volatility risk premium or what we have called the conditional/stochastic volatility risk premium that gives a measure of the risk aversion of an investor with respect to future market movements. However, in the time available, we were not able to convincingly run a model based on Bayesian statistics using MCMC construction as in Jacquier, Paulson and Rossi (2002) [26], which is a reference for this topic. However, we can conclude that current estimation methods are still too sophisticated and not sufficiently robust a priori to allow the stochastic volatility model to surpass the popularity of GARCH models for practitioners, even if it is interesting to look at it from a theoretical point of view. We have chosen to focus here on the simplest stochastic volatility model, since its estimation is already very complex. The extension to more complex models with memory or multivariate is surely a very nice theoretical challenge that some have had the courage to tackle but which goes far beyond the scope of this work.

Appendices

Link with continuous-time models

One of the motivations of preferring the discrete stochastic volatility model over the standard GARCH model is the fact that we can see the discrete SV model as a natural discretization of some very popular continuous processes. For instance:

$$\begin{cases} dlogS_t &= \mu dt + \sigma_t dW_{1t} \\ dlog\sigma_t^2 &= \{\omega + (\beta - 1)log\sigma_{t-1}^2\}dt + \sigma dW_{2t} \end{cases}$$

where (W_{1t}) and (W_{2t}) are two independent Brownian motions, is a popular model for representing asset prices S_t. Here the log-volatility follow an Orstein-Uhlenbeck process. If we add an intercept μ to the equation (1) of the standard SV of $\epsilon_t = log(\frac{S_t}{S_{t-1}})$ we find the natural discretization of this continuous model.

Positively Homogenity Property

A risk measure r(.) is called "Positively Homogene" if for every loss function L and every non-negative real number λ :

$r(\lambda L) \leq \lambda r(L)$

References

[1] Robert F Engle. "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the econometric society* (1982), pp. 987–1007.

[2] Tim Bollerslev. "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3 (1986), pp. 307–327.

[3] Benoit Mandelbrot. "The Variation of Certain Speculative Prices". In: *The Journal of Business* 36.4 (1963), pp. 394–419.

[4] Stephen John Taylor. "Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75". In: *Time series analysis: theory and practice* 1 (1982), pp. 203–226.

[5] Christian Francq and Jean-Michel Zakoian. *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons, 2019.

[6] Christian Francq and Jean-Michel Zakoıan. "Linear-representation based estimation of stochastic volatility models". In: *Scandinavian Journal of Statistics* 33.4 (2006), pp. 785–806.

[7] Andrew Harvey, Esther Ruiz, and Neil Shephard. "Multivariate stochastic variance models". In: *The Review of Economic Studies* 61.2 (1994), pp. 247–264.

[8] Stephen J Taylor. *Modelling financial time series*. world scientific, 1986.

[9] Angelo Melino and Stuart M Turnbull. "Pricing foreign currency options with stochastic volatility". In: *Journal of econometrics* 45.1-2 (1990), pp. 239–265.

[10] Torben G Andersen and Bent E Sørensen. "GMM estimation of a stochastic volatility model: A Monte Carlo study". In: *Journal of Business & Economic Statistics* 14.3 (1996), pp. 328–352.

[11] Luc Bauwens, Christian M Hafner, and Sébastien Laurent. *Handbook of volatility models and their applications*. Vol. 3. John Wiley & Sons, 2012.

[12] Christopher K Carter and Robert Kohn. "Semiparametric Bayesian inference for time series with mixed spectra". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.1 (1997), pp. 255–268.

[13] Sangjoon Kim, Neil Shephard, and Siddhartha Chib. "Stochastic volatility: likelihood inference and comparison with ARCH models". In: *The review of economic studies* 65.3 (1998), pp. 361–393.

[14] Yasuhiro Omori et al. "Stochastic volatility with leverage: Fast and efficient likelihood inference". In: *Journal of Econometrics* 140.2 (2007), pp. 425–449.

[15] A Ronald Gallant and George Tauchen. "Which moments to match?" In: *Econometric theory* 12.4 (1996), pp. 657–681.

[16] Torben G Andersen, Hyung-Jin Chung, and Bent E Sørensen. "Efficient method of moments estimation of a stochastic volatility model: A Monte Carlo study". In: *Journal of econometrics* 91.1 (1999), pp. 61–87.

[17] James Durbin and Siem Jan Koopman. "Monte Carlo maximum likelihood estimation for non-Gaussian state space models". In: *Biometrika* 84.3 (1997), pp. 669–684.

[18] Christian Gouriéroux, Alain Monfort, and Eric Renault. "Indirect inference". In: *Journal of applied econometrics* 8.S1 (1993), S85–S118.
[19] Chiara Monfardini. "Estimating stochastic volatility models through indirect inference". In: *The Econometrics Journal* 1.1 (1998), pp. 113–128.
[20] Christian Francq and Jean-Michel Zakoıan. "Risk-parameter estimation in volatility models". In: *Journal of Econometrics* 184.1 (2015), pp. 158–173.
[21] Paul Kupiec. "Techniques for verifying the accuracy of risk measurement models". In: *The J. of Derivatives* 3.2 (1995).
[22] Peter F Christoffersen. "Evaluating interval forecasts". In: *International economic review* (1998), pp. 841–862.
[23] Christophe Hurlin and Sessi Tokpavi. "Un test de validité de la Value at Risk". In: *Revue économique* 58.3 (2007), pp. 599–608.
[24] Jonathan RM Hosking. "The multivariate portmanteau statistic". In: *Journal of the American Statistical Association* 75.371 (1980), pp. 602–608.
[25] Binay K Adhikari and Jimmy E Hilliard. "The VIX, VXO and realised volatility: a test of lagged and contemporaneous relationships". In: *International Journal of Financial Markets and Derivatives* 3.3 (2014), pp. 222–240.
[26] Eric Jacquier, Nicholas G Polson, and Peter E Rossi. "Bayesian analysis of stochastic volatility models". In: *Journal of Business & Economic Statistics* 20.1 (2002), pp. 69–87.