ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Bachelor Thesis Econometrics and Operations Research

Stochastic volatility models for density forecasting and option pricing

Jasper Grootscholten
479426

Supervisor: M.D. Zaharieva
Second assessor: B. van Os

Date: 5-7-2020

Abstract

This paper investigates popular stochastic volatility models for the purpose of density forecasting and option pricing. Using data of the S&P 500 index and its options for the period 1990-2014, I evaluate the performance of the different stochastic volatility models. Evaluation of density forecasts is done through the simulation-based dynamic probability integral transform, introduced by Yun (2018), while option pricing is evaluated through standard econometric methods. I show that important features to include in the stochastic volatility model are the leverage effect and the fat tails of the asset returns. The asymmetric GARCH model with a skewed t-distribution, which includes both of these features, performs best for density forecasting. For option pricing, the best performing model is the two-factor diffusion model with jumps in mean, spot variance and long-term variance. The jumps allow for extra flexibility, which turns out to improve option pricing performance.

Contents

1 Introduction

Forecasting of financial asset returns has a central place in financial econometrics. Cont (2001) outlined the stylized facts that describe most financial assets, which form a basis for building a model:

1. Absence of autocorrelations, leading to the famous paradigm: “Past performance is not indicative of future results”
2. Non-normal distribution with negative skewness and excess kurtosis.
3. Small, declining autocorrelations in squared returns, with periods of ‘volatility clustering’.

Although these stylized facts do not always hold (see for example Jegadeesh and Titman (1993)), a model should capture these facts to correctly describe most financial assets. In the past, financial institutions would mostly focus on point forecasts, however due to the first stylized fact, these rarely beat out the simple random walk benchmark, as they do not contain much more information. Nowadays, interval forecasts are preferred, as these also capture the uncertainty around the point forecast. However, an interval forecast still does not describe the full conditional probability distribution, leaving important information out of the forecast. Therefore, this paper focuses on density forecasts, where the full conditional probability distribution is given by the forecast, encompassing both the point and interval forecast.

The GARCH model (Engle (1982), Bollerslev (1986)) is the most popular for modelling asset returns, and has been widely studied. However, the GARCH model is not flexible enough to include all financial phenomena researchers come across. Therefore, recent research, especially in the field of option pricing, has been focused on modelling an asset return and its volatility as separate, possibly correlated, diffusion processes. This class of models is referred to as stochastic volatility models. This paper analyses these models and their evaluation. A stochastic volatility model can be widely applied from density forecasting all the way to for example derivative pricing. Therefore, this research is relevant for companies across the financial industry, from more traditional investors such as pension funds to complex traders like hedge funds. The purposefully ambiguous central question is: “Which stochastic volatility model is most accurate in modelling asset returns?”. How to evaluate the accuracy depends on the goal of the model. First I look at density forecasting, where accuracy is not straightforward to evaluate, as the true density is unknown. This is solved by using the simulation-based dynamic probability integral transform introduced by Yun (2018). For option pricing, the market prices are observed, so that evaluation can be done through standard econometric methods.

Yun (2018) has shown that the out-of-sample forecasting performance of one-factor diffusion models that include jumps is promising, using his simulation-based dynamic probability integral transform. This makes these models a strong alternative to the popular GARCH type models. In this paper I replicate the analysis performed in Yun (2018), but I also include two-factor diffusion models with jumps, which are not evaluated in Yun (2018). Moreover, I briefly examine how dynamic parameter estimation leads to further improvement of the density forecasts. Furthermore, I investigate whether the results found using the evaluation methods by Yun (2018) also translate over to option pricing. For this purpose, I use the stochastic volatility models to price options using Monte Carlo simulation.

Using data from the S&P 500 index and its options for the period 1990-2014, I evaluate a wide range of stochastic volatility models. The period 1990-2000 is used for estimation, while the period 2001-2014 is saved for out-of-sample evaluation. I show that the GJR-ST model, a GARCH type model, has the best out-of-sample forecasting performance. The one- and two-factor diffusion models with a jump in mean are relatively close in performance. Furthermore, I show that stochastic volatility models outperform the standard Black-Scholes model for option pricing, with the two-factor diffusion model with jumps in mean, spot variance and long-term variance being the best performing. This shows the potential stochastic volatility models have, not only for forecasting, but also for option pricing.

The rest of this paper is structured as follows. The remainder of this Section is devoted to discussing literature regarding stochastic volatility models, density forecast evaluation and option pricing. In Section 2 the data used for the analysis is outlined. Next, in Section 3 the different models and methods are further specified and elaborated. In Section 4 the results of a simulation study as well as the empirical results are presented. Lastly, in Section 5 a discussion of the results as well as a conclusion are given.

1.1 Stochastic volatility models

Engle (1982) and Bollerslev (1986) introduced the GARCH model, which is still the standard for volatility modelling in the financial industry. This discrete time model gives an explicit value for the model implied volatility, given the past asset returns. Due to the closed form likelihood of the model being available, estimation is done through maximum likelihood estimation (MLE). The model is able to describe the stylized facts from asset returns outlined in Section 1. The GARCH model has been widely studied and numerous extensions (see Bollerslev (2008) for an extensive list) to the model have been suggested to allow for additional features. An example is assuming a different underlying distribution for the error, such as the (skewed) t-distribution, to deal with excess kurtosis.

A prominent extensions captures the 'leverage effect', which states that volatility increases more because of a large negative return than a large positive return. This is done through the GJR (or threshold/asymmetric GARCH) model, suggested by Glosten et al. (1993), which includes positive and negative shocks separately. Again different distributional assumptions can be introduced to this model. I include both the standard GARCH and GJR model in my analysis.

Another recent approach to volatility modelling is the usage of a stochastic process in the variance specification, allowing for more flexibility. The idea is to model the asset return and volatility as separate, possibly correlated, diffusion processes. This class of models is referred to as stochastic volatility models, and these models are typically time continuous, making estimation challenging. The stochastic volatility model is able to incorporate a wide range of financial phenomena. As an example, Heston (1993) showed that this type of model can produce a closed form option pricing formula, which is able to describe the 'volatility smile' in option prices. This makes it a strong alternative for the well-known Black-Scholes formula (based on constant volatility), which does not capture this anomaly observed in option prices. The first stochastic volatility model described is the log stochastic volatility (LSV) model, proposed by Taylor (1982). A popular form of the stochastic volatility model is to use a CIR (Cox et al. (1985)) specification for the variance and to incorporate sudden moves in the asset price and/or volatility through a jump process (see for example Duffie et al. (2000), Broadie and Kaya (2006)). Similarly to Yun (2018), the LSV type model as well as one-factor diffusion models, with a CIR specification and possible jumps, are examined in this paper.

To include the difference between short and long-term volatility, two-factor diffusion models have been studied by Bates (2000), Chernov et al. (2003) and Kaeck and Alexander (2012). These models allow even more flexibility, again with for example the inclusion of jump processes in the different variables. Kaeck and Alexander (2012) conclude that these models are superior to more traditional models in the field of option pricing. As noted earlier, only two-factor models without jumps have been evaluated in Yun (2018), whereas I also include two-factor models with jumps.

Estimation of the stochastic volatility models is commonly done using Bayesian Markov Chain Monte Carlo (MCMC) estimation, introduced by Jacquier et al. (1994). They argued that this method of estimation is superior to for example GMM or QMLE. Eraker et al. (2003), Yu (2005) and Li et al. (2008) have shown that Bayesian MCMC accurately estimate parameters for diffusion type stochastic volatility models, possibly including jumps. Therefore, Bayesian MCMC is also the estimation method I use.

1.2 Density forecast evaluation

Pioneering work for density forecast evaluation has been done by Rosenblatt (1952), who introduced the probability integral transform (PIT). Through the PIT a series of generalized residuals is obtained. If the model is correctly specified, these generalized residuals follow a uniform distribution. This fact forms the basis for testing the forecasting ability of a model. Because the stochastic volatility models do not have a closed form PIT available, the simulation-based dynamic PIT method based on particle filters proposed by Yun (2018) is used. The testing of the distribution of the generalized residuals can be done through informal analysis of the histogram and autocorrelation function as in Diebold et al. (1998). Formal statistical tests for distributional assumptions can also be used, such as the Kolmogorov-Smirnov (KS) (1948), Jarque and Bera (JB) (1980), Berkowitz (B) (2001) or Hong and Li (HL) (2005) test. All these tests will be employed in my analysis.

Another way to evaluate density forecasts is by looking at the coverage of predicted intervals (or specifically the value at risk (VaR)) as in Christoffersen (1998) or Engle and Manganelli (2004). Lastly the log-likelihood can be used for evaluation as well. This is done by using a scoring rule, such as the one suggested by Gneiting and Raftery (2007), and comparing these scores through the test developed by Diebold and Mariano (1995). Both the model implied VaR violations as well as the log-likelihood are available from the simulation-based dynamic PIT by Yun (2018). These methods will be used for evaluation of the models in addition to the distribution based tests.

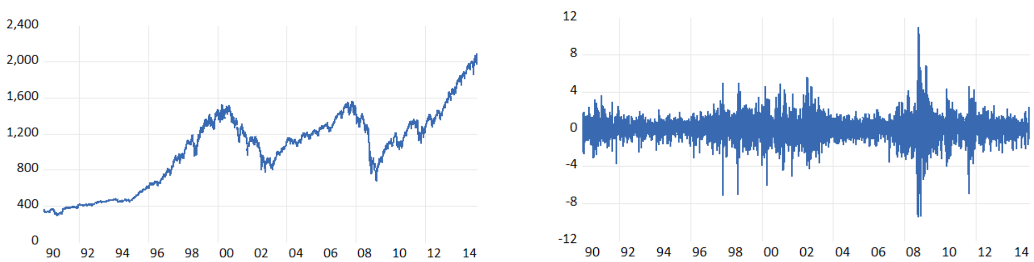
1.3 Option pricing

Stochastic volatility models are most prominent in the field of option pricing. In their seminal paper, Black and Scholes (1973) model an asset using a geometric Brownian motion with constant volatility. Combined with the assumption of a perfect market, they derive a closed form formula for the price of a European call option. This now famous Black-Scholes formula is still considered the standard for option pricing. Other popular model choices are the one-factor diffusion model (Heston (1993)) and one-factor diffusion model with jumps in mean (Bates (1996)), which also allow for a closed form option price. These models, as well as the other stochastic volatility models discussed in Section 1.1 will be used for option pricing in this paper.

2 Data

For empirical analysis, daily closing prices in dollars of the S&P 500 index[1] are used. These are transformed into continuously compounded daily returns through $r_t = 100\log(\frac{S_t}{S_{t-1}})$, where S_t is the asset price at time t. I use data ranging from 1990 until 2014, similarly to Yun (2018). The period 1990-2000 (2780 observations) is used for estimation. I consider two different out-of-sample periods: 2001-2007 (1758 observations) is used as a smaller out-of-sample period and 2001-2014 (3521 observations) is used as a second, larger out-of-sample period, which includes the financial crisis.

The S&P 500 index and daily return are shown in Figure 1. Here the financial crisis of 2008 is clearly visible, leading to a drop in the index, as well as a larger volatility. This pattern is in accordance with what one would expect from the leverage effect. In Figure 1b, volatility clustering is also present, as there are extended periods of continuous small and large volatility. A histogram and summary statistics of the returns, as well as the autocorrelation function are shown in Figure 2. Here the stylized facts as discussed in Section 1 are visible. The non-normality is clear from Figure 2a, with a negative skewness of -0.24 and excess kurtosis of 11.76. Figure 2b shows the autocorrelation function of the (squared) returns. The patterns described by the stylized facts are again visible, namely nearly insignificant autocorrelation for the returns, and significant, slowly decreasing autocorrelation for the squared returns.

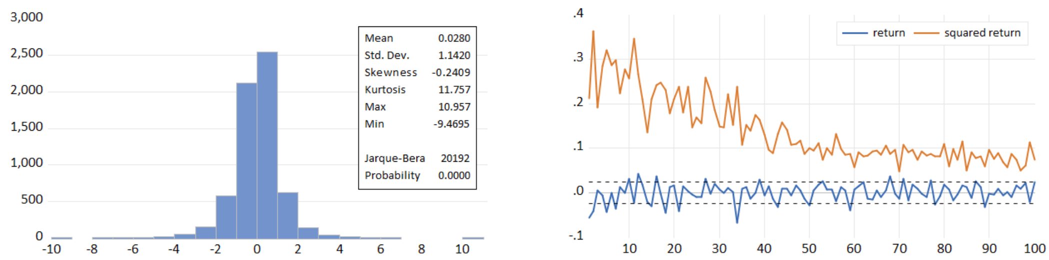


(a) S&P 500 index over 1990-2014

(b) S&P 500 daily returns over 1990-2014

Figure 1: S&P 500 index and daily returns over 1990-2014

[1] Obtained from YahooFinance: `https://finance.yahoo.com/quote/%5EGSPC/history/`



(a) Histogram of returns over 1990-2014

(b) Autocorrelation of returns over 1990-2014

Figure 2: Histogram and autocorrelation of returns over 1990-2014

For option pricing, I use SPX (based on S&P 500 index) option data[2]. I work with European call options, and only consider nearly at-the-money (meaning the strike price is within 2.5% of the value of the underlying) options that have been traded on the first 7 trading days of 2001, giving a total of 112 options. Summary statistics of these options are given in Table 1. I define the market value of the option as the average between the highest ask and the lowest bid. The maturity of the options ranges from 7 days to almost 1 year. An important problem with the option dataset is that the exact time at which the option was traded is unknown, meaning the exact value of the underlying at which the option was traded cannot be determined.

Table 1: Summary statistics of the option data

	Mean	**Std. dev.**	**Min**	**Max**
Market value (dollars)	50.71	32.37	8.250	155.4
Strike price (dollars)	1315.5	29.46	1275.0	1380.0
Maturity (trading days)	44.88	54.22	7	243

3 Methodology

3.1 Stochastic volatility models

3.1.1 GARCH

The standard GARCH model as described by Engle (1982) and Bollerslev (1986) takes the following form:

$$r_t = \mu + \sigma_t \varepsilon_t \tag{1}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \alpha_2 \varepsilon_{t-1}^2 \tag{2}$$

where μ is the mean return, σ_t^2 is the conditional variance and $\varepsilon_t \sim N(0,1)$ (GARCH-N) or $\varepsilon_t \sim \sqrt{\frac{\nu-2}{\nu}} t(\nu)$ (GARCH-T).

Glosten et al. (1993) introduced the GJR model to capture the leverage effect. This is done by changing the conditional variance specification in Equation 2 to:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \alpha_2 \varepsilon_{t-1}^2 + \alpha_3 \varepsilon_{t-1}^2 I_{\varepsilon_{t-1}<0} \tag{3}$$

[2]Obtained from OptionMetrics page of WRDS: `https://wrds-www.wharton.upenn.edu/`

where I is the indicator function. α_3 is the leverage effect parameter, a significantly positive α_3 indicates the presence of the leverage effect. Again $\varepsilon_t \sim N(0,1)$ (GJR-N) or $\varepsilon_t \sim \sqrt{\frac{\nu-2}{\nu}}t(\nu)$ (GJR-T). Lastly the skewed t-distribution (Fernández and Steel (1998)) (GJR-ST) defined as:

$$p(x|\xi) = \begin{cases} \frac{2}{\xi+\xi^{-1}} f(x\xi), & \text{if } x < 0 \\ \frac{2}{\xi+\xi^{-1}} f(x\xi^{-1}), & \text{if } x \geq 0 \end{cases} \tag{4}$$

is also used. Here ξ is the skewness parameter and $f(\nu)$ the density of the t-distribution with ν degrees of freedom. All GARCH-type models are straightforward to estimate using maximum likelihood estimation (MLE) in R (using the package rugarch).

3.1.2 LSV

The log stochastic volatility model, first suggested by Taylor (1982), is given by the stochastic differential equations (SDE):

$$r(t) = \mu dt + \sigma(t) dW^r(t) \tag{5}$$

$$d\text{log}(\sigma^2(t)) = \alpha + \beta\text{log}(\sigma^2(t))dt + \sigma_v dW^\sigma(t) \tag{6}$$

where $W^r(t)$ and $W^\sigma(t)$ are Brownian motions with $\text{corr}(W^r(t), W^\sigma(t)) = \rho$.

An Euler-Maruyama discretization is necessary for estimation, the discretized model is given by:

$$r_t = \mu + e^{\frac{1}{2}\sigma_t^2}\varepsilon_t \tag{7}$$

$$\Delta\sigma_t^2 = \alpha + \beta\sigma_{t-1}^2 + \sigma_v\rho\varepsilon_{t-1} + \sigma_v\sqrt{1-\rho^2}\eta_t \tag{8}$$

where $(\varepsilon_{t-1}, \eta_t) \sim N(0, I_2)$. A negative parameter ρ indicates the presence of the leverage effect. Namely when the correlation between the return and variance is negative, a negative return leads to higher variance. A restricted version of the LSV model with $\rho = 0$ (LSV0) is also considered. Estimation is done through Bayesian MCMC, using the specialized software JAGS in R. I use initialization and priors similar to Meyer and Yu (2000) and Yu (2005), given by $\mu \sim N(0, 25)$, $\alpha^* \sim N(0, 25)$, $\alpha = \frac{\alpha^*}{1-\beta}$, $\beta^* \sim \beta(20, 1.5)$, $\beta = 2\beta^* - 1$, $\sigma_v^2 \sim IG(2.5, 0.025)$, $\rho \sim U(-1, 1)$ and $\sigma_0^2 \sim N(\alpha, \sigma_v^2)$.

3.1.3 Diffusion models with jumps

Diffusion models allow the incorporation of jumps in the stochastic processes, as well as multiple factors that describe the asset return. The two-factor diffusive model considers a separate process for the long-term mean M of the spot variance σ^2, and is given in its most general form by the SDE (based on ideas from Bates (2000) and Kaeck and Alexander (2012)):

$$r(t) = \mu dt + \sigma(t)dW^r(t) + dJ^r(t) \tag{9}$$

$$dM(t) = \kappa_M(\theta_M - M(t))dt + \sigma_M\sqrt{M(t)}dW^M(t) + dJ^M(t) \tag{10}$$

$$d\sigma^2(t) = \kappa(M(t) - \sigma^2(t))dt + \sigma_v\sigma(t)dW^\sigma(t) + dJ^\sigma(t) \tag{11}$$

where $W^r(t)$, $W^M(t)$ and $W^\sigma(t)$ are Brownian motions with $\text{corr}(W^r(t), W^\sigma(t)) = \rho$ and $W^M(t)$ independent of both $W^r(t)$ and $W^\sigma(t)$. The process $dJ^i(t) = Z^i(t)dN(t)$ with $i \in \{r, M, \sigma\}$ gives the jumps, with $Z^r(t) \sim N(\mu_r, \sigma_r^2)$ the size of the jump in return, $Z^M(t) \sim \exp(\mu_M)$ the size of the jump in long-term variance, $Z^\sigma(t) \sim \exp(\mu_\sigma)$ the size of the jump in spot variance and $dN(t)$ a Poisson process with intensity λ. Both the long-term mean of the variance and the spot variance are modelled using the CIR process (Cox et al. (1985)). Note that the jumps are assumed to be contemporaneous, which is reasonable as an unexpected event should influence both the mean and variance at the same time.

Now the Euler-Maruyama discretization of the model is given by:

$$r_t = \mu + \sigma_{t-1}\varepsilon_t + J_t Z_t^r \tag{12}$$

$$\Delta M_t = \kappa_M(\theta_M - M_{t-1}) + \sigma_M \sqrt{M_{t-1}} u_t + J_t Z_t^M \tag{13}$$

$$\Delta \sigma_t^2 = \kappa(M_t - \sigma_{t-1}^2) + \sigma_v \rho \sigma_{t-1}\varepsilon_{t-1} + \sigma_v \sqrt{1-\rho^2}\sigma_{t-1}\eta_t + J_t Z_t^\sigma \tag{14}$$

where $(\varepsilon_{t-1}, u_t, \eta_t,) \sim N(0, I_3)$ and $J_t \sim \text{Ber}(\lambda)$ (a Bernoulli distribution is used to discretize the Poisson process). The models considered are: One-factor models (just using the spot variance σ^2, where M_t in Equation 14 is replaced by θ) without jumps (SV), with jumps in the return (SVJ) and with jumps in both the return and variance (SV2J). And two-factor models, without jumps (2-SV), with jumps in return (2-SVJ), jumps in return and spot variance (2-SV2J) and jumps in return, spot variance and long-term variance (2-SV3J). The models are again estimated through Bayesian MCMC using JAGS. Priors and initialization is based on Eraker et al. (2003): $\mu \sim N(1, 25)$, $\mu_r \sim N(0, 100)$, $\sigma_r^2 \sim IG(5, 20)$, $\kappa_M \sim N(0, 1)$, $\theta_M \sim N(1, 1)$, $\sigma_M^2 \sim 1/16IG(2.5, 0.1)$, $\mu_M \sim G(20, 10)$, $\kappa \sim N(0, 1)$, $\kappa\theta \sim N(0, 1)$, $\sigma_v^2 \sim IG(2.5, 0.1)$, $\mu_\sigma \sim G(20, 10)$, $\lambda \sim \beta(2, 40)$ and $\rho \sim U(-1, 1)$. Initialization is done through $M_0 = \theta_M$ and $\sigma_0^2 = \theta$ (one-factor) or M_0 (two-factor).

3.2 Density forecast evaluation

3.2.1 Probability integral transform

Central for the evaluation of density forecasts is the probability integral transform (PIT), first introduced by Rosenblatt (1952). Given a model implied density function $f_t()$ for the asset return at time t, the PIT is defined as:

$$z_t = \int_{-\infty}^{r_t} f_t(u)du \tag{15}$$

When the model is correct, the series $\{z_t\}$ of generalized residuals is i.i.d. $U(0, 1)$ distributed, which forms a basis for testing the model specification.

3.2.2 Simulation-based dynamic probability integral transform

However, as discussed in Yun (2018), evaluation of the PIT is troublesome for some of the complex models described in Section 3.1, because a closed form PIT is not available. Yun (2018) gives the LSV model as an example, where multiple complex integrals are necessary to evaluate the PIT (the same holds for log-likelihood values):

$$
\begin{aligned}
z_t &= \int_{-\infty}^{r_t} f(r|r_{t-1})dr \\
&= \int_{-\infty}^{r_t}\int_0^{\infty} f(r|\sigma_t)f(\sigma_t|r_{t-1})d\sigma_t dr \\
&= \int_{-\infty}^{r_t}\int_0^{\infty}\int_0^{\infty} f(r|\sigma_t)f(\sigma_t|\sigma_{t-1})f(\sigma_{t-1}|r_{t-1})d\sigma_{t-1}d\sigma_t dr
\end{aligned} \tag{16}
$$

This problem is solved by Yun (2018) through a simulation-based approach using particle filters. All the steps of the algorithm for the different models are shown in the Appendix (Sections 6.2 and 6.3), and for a more detailed explanation I refer to the paper by Yun (2018).

3.2.3 Distribution based testing

In their paper regarding the evaluation of density forecasts, Diebold et al. (1998) advocate an informal investigation of the histogram as well as the empirical autocorrelation function of the generalized residuals. Through this analysis potential problems such as non-uniformity or serial dependence are identified.

The distribution of the generalized residuals can also be tested formally. There is a plethora of formal tests for distributional assumptions, in this paper some of the most popular ones are used. The KS-test proposed by Kolmogorov and Smirnov (1948), tests the null hypothesis that a sample is drawn from a reference distribution. The test statistic is given by:

$$KS = \max_{0<x<1} |\tilde{F}(x) - F(x)| \tag{17}$$

where $\tilde{F}()$ is the empirical CDF, and $F()$ is the CDF of the reference distribution. The test statistic follows the non-standard Kolmogorov distribution, which is used to find critical values. Due to its generality, this test is low in power. Moreover, the test assumes that the independence of the generalized residuals is satisfied.

The uniformity of the generalized residuals allows for a transformation into the normal distribution, which forms the basis for the JB-test by Jarque and Bera (1980) and the B-test by Berkowitz (2001). The transformation is given by:

$$x_t = \Phi^{-1}(z_t) \tag{18}$$

where Φ is the standard normal CDF. The JB-test tests the null hypothesis of normality by looking at the skewness and kurtosis, which are 0 and 3 respectively for a standard normal distribution. The test statistic is given by:

$$JB = n(\frac{SK^2}{6} + \frac{(K-3)^2}{24}) \tag{19}$$

where SK is the skewness and K is the kurtosis of the series $\{x_t\}$. The statistic follows the standard Chi-square distribution with 2 degrees of freedom. Again the JB-test does not consider dependence, and has relatively low power due to its generality.

The B-test tests the null hypothesis of normality versus the alternative of an AR(1) process given by:

$$x_t - \mu = \phi(x_{t-1} - \mu) + \varepsilon_t \text{ with } \varepsilon_t \sim N(0, \sigma^2) \tag{20}$$

Under the null hypothesis, the parameters in this process are restricted to $\mu = 0$, $\sigma^2 = 1$ and $\phi = 0$, leading to a likelihood ratio test statistic given by:

$$B = -2(\mathcal{L}_0(\mu = 0, \sigma^2 = 1, \phi = 0) - \mathcal{L}_a(\mu, \sigma^2, \phi)) \tag{21}$$

where $\mathcal{L}$ is the log-likelihood of the model. This statistic follows a Chi-square distribution with 3 degrees of freedom, and tests for both normality as well as independence simultaneously.

The HL-test by Hong and Li (2005) is most geared towards the testing of density forecasts. The test compares a kernel estimator $\hat{g}(u_1, u_2)$ for the joint density of $\{z_t, z_{t-j}\}$ with the product of two uniform densities (which is given by 1), with j a chosen lag order. Hong et al. (2007) suggest combining the different statistics into one single Portmanteau type statistic. I choose to work with this statistic, with lag truncation orders $\rho = 5, 10$ and 20. All steps of the test are given in the Appendix (Section 6.4), for more details I again refer to the paper by Yun (2018) as well as Hong and Li (2005) and Hong et al. (2007). Because of its specificity, this test has the highest power for density forecast evaluation. All these tests are performed using the implementations in R (using the packages tseries and rugarch).

3.2.4 VaR based testing

The value at risk (VaR) is one of the most popular risk measures in finance. Through the PIT, VaR violations are identified without the need for explicit VaR calculations, due to the fact that $I(r_t < -\text{VaR}_t(p)) = I(z_t < p)$ where p is the chosen probability level. These VaR violations are used for formal testing. The test by Christoffersen (1998) provides a likelihood ratio statistic for the unconditional coverage (which indicates whether the amount of violations corresponds with the chosen probability level), for the independence of the violations and for the conditional coverage (which combines the unconditional coverage and independence). The DQ test by Engle and Manganelli (2004) also tests for unconditional coverage and independence simultaneously. Details of both tests are given in the Appendix (Sections 6.5 and 6.6).

3.2.5 Log-likelihood based testing

The methods suggested by Yun (2018) in addition allow for the calculation of the log-likelihood $\mathcal{L}^{(m)}$ for a model m. This allows for the calculation of a series of differences in log-likelihood between two competing models m and n given by:

$$d_t = \mathcal{L}^{(m)}(r_t) - \mathcal{L}^{(n)}(r_t) \tag{22}$$

The significance of these differences can be formally tested through a Diebold and Mariano (1995) type test given by:

$$DM = \frac{\bar{d}}{\sqrt{\text{VAR}(d)/T}} \tag{23}$$

where $\bar{d}$ and $\text{VAR}(d)$ are the sample mean and variance of $\{d_t\}$ and T is the sample size. Under the null hypothesis of equal model performance, this statistic follows a standard normal distribution.

3.3 Monte Carlo method for option pricing

I now move to using the stochastic volatility models to price European call options. Although stochastic volatility models allow for closed form option pricing formulas, this involves solving complex systems of differential equations. Instead I use Monte Carlo simulation to find option prices, based on option pricing theory by Cox and Ross (1976). Monte Carlo simulation can more easily adopt different stochastic volatility models for the underlying, as well as price more exotic options. The disadvantage of simulation is that it is more time consuming, but speed is not of concern in this paper. Boyle (1977) has shown that Monte Carlo simulation gives similar results to analytical pricing for the Black-Scholes model.

The Monte Carlo simulation approach, as proposed by Boyle (1977), is described by the following steps:

1. Generate a large amount N sample paths for the underlying security S, based on the chosen model. Because the models are formulated in terms of the daily return r_t, this means the formula $S_{t+1} = S_t e^{r_t}$ is used to calculate the value of S at time $t+1$.

2. Calculate the option payoff for each simulated underlying, given by $\pi_i = \max(S_{T,i} - K, 0)$ for a European call option. Here $S_{T,i}$ is the value of the underlying at maturity for simulation i, and K is the strike price.

3. Average the payoffs $\bar{\pi} = \frac{1}{N}\sum_{i=1}^{N}\pi_i$.

4. Discount the average payoff, to find the simulated present value of the option $\hat{V} = e^{-rT}\bar{\pi}$, where r is the risk-free rate, and T the time to maturity in years.

The Black-Scholes model is used as a benchmark. Here the underlying is modeled using the simple formula $r_t = \mu + \sigma\varepsilon_t$ with $\varepsilon_t \sim N(0,1)$, corresponding to the geometric Brownian motion. Evaluation of the models is done through the mean squared error $\text{MSE} = \frac{1}{K}\sum_{k=1}^{K}(V_k - \hat{V}_k)^2$ and mean absolute percentage error $\text{MAPE} = \frac{1}{K}\sum_{k=1}^{K}\frac{|V_k - \hat{V}_k|}{V_k}$, where V_k and $\hat{V}_k$ are respectively the true and simulated price of option $k = 1, ..., K$. The difference in these measures between the models is formally tested using a Diebold-Mariano type test similarly to Section 3.2.5.

Quite substantial biases could occur in the predicted option prices for two reasons. First of all problems with the data, such as the asynchronization between the option prices and price of the underlying mentioned in Section 2, can lead to a bias. Especially because returns are continuously compounded, meaning that an incorrect value of the underlying gets further amplified. Secondly, the static parameter estimates can lead to a discrepancy between the model and the real world underlying prices. To solve these biases, I introduce a simple correction. Namely for every day, I subtract the average error made by the model on that day from the modeled option prices. This results in a mean error of zero, which could be interpreted as the bias no longer existing.

Note that this correction cannot be done in practice, as it requires knowledge of the real option prices to calculate the modeled option prices. However this correction can be justified, as it helps the theoretical comparison of the models, and in practice a trader can avoid the issues that lead to these biases in the first place.

4 Results

4.1 Simulation based test performance

Before looking at empirical results, I check whether the different statistical tests from Section 3.2 give correct size performance. This check is done by simulating sample paths generated by the different models and then applying the different statistical tests. Note that this approach ignores the additional uncertainty from parameter estimation and discretization. Ideally, this would be taken into account by using Bayesian MCMC estimation during the simulations, but this induces a too high computational cost. If the size performance of a test is correct, the rejection rate of the test should be similar to the chosen significance level. Table 8 in the Appendix, shows the results for the different models (the GARCH type models are excluded, as these do not require a simulation-based PIT). This table shows that indeed the rejection rates of the different tests for the different models are very close to the chosen significance levels of 10%, 5% and 1%, for different sample lengths. This confirms that these evaluation methods are a reliable way too evaluate the accuracy of density forecasts from the different models.

4.2 Application to the empirical S&P 500 data

All models from Section 3.1 are estimated on daily returns of the S&P 500 over the period 1990-2000. Estimated coefficients are shown in Table 9 in the Appendix. The estimated coefficients are comparable to those found in related studies. Notably, the coefficient related to the leverage effect, which is α_3 for GARCH type models, and ρ for all other models, is always found to be significant, suggesting it is important to incorporate this phenomenon into the model. Another interesting observation is that the estimated jump intensity lies around 1% for jumps in mean (for example $\lambda = 0.0116$ for SVJ), and 0.4% for jumps in mean and variance. On a side note, this shows that the discretization bias is not that influential for jumps as the probability of for example two or more jumps in mean during a day, is only 0.01%[3].

Given the parameter estimates, the accuracy of the different density forecasts can be evaluated for the two out-of-sample periods 2001-2007 and 2001-2014. Note that static parameter estimates are used for forecasting, meaning that the parameters are not updated after every forecasts, as this would induce too high computational costs. This means that the performance of the models is not be optimal, especially during and after the crisis, where parameter updates would be most necessary. The usage of dynamic parameter estimates is discussed in the following section. Full results of the different tests are given in Tables 10 and 11, however for clarity I will only show results of the best performing model within each class[4] in this section. Looking

[3]Simply sum the probability $\sum_{n=2}^{\infty} 0.01^n = 0.0001$

[4]I consider GARCH type, LSV, one-factor diffusion and two-factor diffusion as 4 different classes

at the best performing model within each class also helps distinguish important features that should be incorporated in the model.

From the test statistics in Tables 10 and 11, I conclude that for GARCH type models, the best performing model is the GJR-ST model. As noted before the leverage effect seems important to incorporate, which is confirmed by the test results, as the GJR models perform better than their GARCH counterparts. Moreover, the (skewed) t-distribution strongly outperforms the normal distribution on all test statistics, because it is able to incorporate the fat tails usually observed in asset returns. Although the estimated skew coefficient ξ is not significantly different from 1 (which indicates a standard t-distribution), the skewed t-distribution performs slightly better. This is presumably because it is more flexible in allowing asymmetry in the distribution. The unrestricted LSV model outperforms the restricted version, again showing the importance of including the leverage effect. For one-factor diffusion models, results are very close, but the SVJ model with jumps in mean is slightly better. For two-factor diffusion models, results are again close, but the 2-SVJ model seems to have an edge. In both cases the inclusion of a jump in mean appear to help deal with spikes and regime switches in the data, but adding a contemporaneous jump in (long-term) variance, does not lead to further improvement. In summary, the best performing model within each class are the GJR-ST, LSV, SVJ and 2-SVJ model. The rest of this section will be devoted to the results of these four models.

Firstly, I do an informal analysis of the histograms and autocorrelation functions of the generalized residuals as suggested by Diebold et al. (1998). Only results for the larger out-of-sample period are used. The histograms of the generalized residuals for the different models are shown in Figure 3. It can be observed that the generalized residuals are reasonably uniformly distributed, mostly falling within the 95% confidence interval. The histograms display what is called a (one-winged) butterfly shape in Diebold et al. (1998), with a hump in the middle, and a wing at the left tail. This indicates that especially large losses are a bit more common than forecasted by the models, which is plausible as the financial crisis is included in this time period.

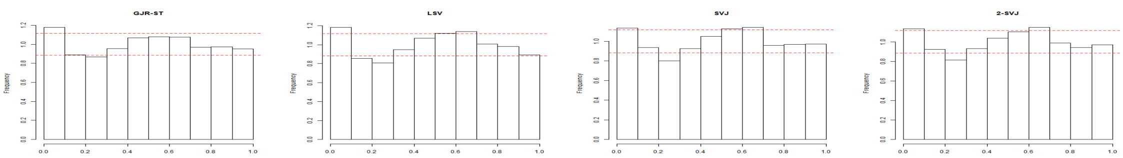


Figure 3: Histograms of generalized residuals for the period 2001-2014, dotted lines give 95% confidence interval

The autocorrelation function of the (squared) generalized residuals are shown in Figures 4 and 5 respectively. From these figures it can be seen that the models did not fully capture the conditional volatility dynamics of the data, as there is still significant autocorrelation for the first lag in both the normal and squared residuals. So from the informal analysis I conclude that the specifications cannot be considered fully correct, as they violate both uniformity and independence.

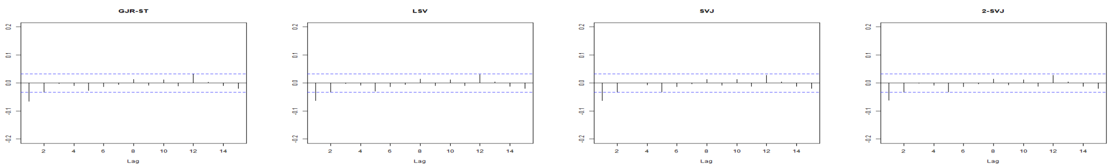


Figure 4: Autocorrelation function of generalized residuals for the period 2001-2014, dotted lines give 95% confidence interval

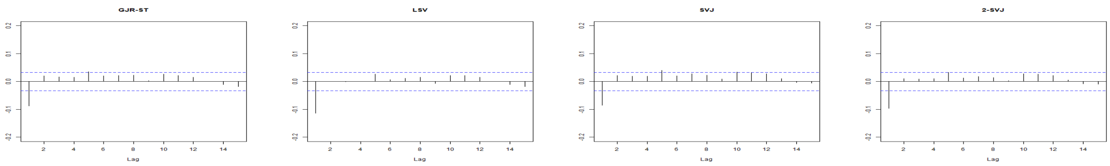


Figure 5: Autocorrelation function of squared generalized residuals for the period 2001-2014, dotted lines give 95% confidence interval

Secondly, I discuss the distribution based tests, shown in Table 2. For the first out-of-sample period, none of the models are rejected by the Kolmogorov-Smirnov test at 5% significance level. Only the GJR-ST and LSV model are rejected by the Jarque-Bera test, however all models are rejected by the Berkowitz and Hong-Li tests. As discussed in Section 3.2, these tests are more geared towards specifically evaluating density forecasts, and therefore have the highest statistical power. This means they are more likely to reject a false null hypothesis. Additionally, these two tests also look at serial dependence, which I already identified as problematic in the informal analysis. For these reasons I consider these two tests, especially the Hong-Li test, most relevant. Based on these results it is hard to call the model specifications correct, as they are still rejected by the relevant Berkowitz and Hong-Li tests. Finally, the difference in log-likelihood between the GJR-ST model (-2311.6) and the others (the second highest log-likelihood is -2319.7) is quite substantial in favor of the GJR-ST model. However, this comparison is not completely fair, as the GJR-ST model is estimated using maximum likelihood. As suggested by its name, MLE is solely focused on maximizing the likelihood, whereas the Bayesian MCMC estimation is not. Moreover, the evaluation of the GJR-ST log-likelihood is direct, whereas the other log-likelihoods are evaluated using the simulation-based dynamic PIT.

Next, I look at the second out-of-sample period, also shown in Table 2. Here the performance is generally worse, as all models are rejected by the Jarque-Bera, Berkowitz and Hong-Li test. Only the Kolmogorov-Smirnov test does not reject the GJR-ST and 2-SVJ model. There are two main reasons for this worse performance. First and most obvious, this sample period includes the financial crisis, during which asset prices are much harder to forecast, leading to worse forecasting performance. Second, because this sample period is longer, the usage of static parameter estimates is even more problematic, as the longer timeframe gives more room for the parameters to change. This effect gets magnified by the crisis, which makes it even more likely that parameters change. Even though the inclusion of jumps should theoretically help deal with the sudden moves in the asset price due to the financial crisis, it appears from these results that it does not help in practice.

Table 2: Distribution based test results

2001-2007							
	KS	**JB**	**B**	**HL(5)**	**HL(10)**	**HL(20)**	**log-likelihood**
GJR-ST	0.030*(0.083)	7.55 (0.023)	9.1 (0.028)	**8.1**	**10.0**	**13.6**	**-2311.6**
LSV	0.029*(0.096)	14.3 (0.001)	8.8 (0.032)	10.0	13.1	17.8	-2326.8
SVJ	0.027*(0.148)	**3.07***(0.215)	**8.1** (0.043)	9.4	12.4	17.0	-2319.7
2-SVJ	**0.026***(0.194)	4.35*(0.114)	8.4 (0.038)	9.2	12.2	16.3	-2321.9
2001-2014							
	KS	**JB**	**B**	**HL(5)**	**HL(10)**	**HL(20)**	**log-likelihood**
GJR-ST	**0.018***(0.189)	16.1 (0.000)	**15.2** (0.002)	**12.8**	**14.0**	**17.7**	**-4870.6**
LSV	0.025 (0.026)	56.1 (0.000)	17.7 (0.000)	18.8	23.0	31.3	-4893.3
SVJ	0.024 (0.037)	17.4 (0.000)	20.1 (0.000)	18.0	22.1	29.2	-4911.7
2-SVJ	0.022*(0.064)	**13.8** (0.001)	17.6 (0.001)	17.2	20.5	27.4	-4902.0

Test results of the Kolmogorov-Smirnov, Jarque-Bera, Berkowitz and Hong-Li tests, as well as the log-likelihood for both out-of-sample periods, with p-values in parenthesis. For the HL test, 5% critical values obtained from simulation are 3.40 for 2001-2007 and 3.58 for 2001-2014. * indicates the null hypothesis of correct model specification is not rejected at 5% significance level. Bold gives the model with the best result on the corresponding statistic.

Thirdly, I turn to the VaR based tests, shown in Table 3. For the first out-of-sample period, all models have correct conditional coverage, based on both the LR_{cc} and DQ test statistics. This shows that VaR forecasts based on the models are very accurate for this period. Similarly to the distribution based tests, the performance is worse for the second out-of-sample period. Although the independence of VaR violations is correct for all models, the unconditional coverage is not (with the exception of the GJR-ST and 2-SVJ model for 1% VaR). This can be attributed to abnormally large losses that happened during the crisis. Only the GJR-ST model still performance reasonably for this timeframe, specifically correct conditional coverage is not rejected by the LR_{cc} for 5% VaR and not rejected by both the LR_{cc} and DQ for 1% VaR.

Table 3: VaR based test results

2001-2007								
	5% VaR				1% VaR			
	$\boldsymbol{LR_{uc}}$	$\boldsymbol{LR_{id}}$	$\boldsymbol{LR_{cc}}$	**DQ**	$\boldsymbol{LR_{uc}}$	$\boldsymbol{LR_{id}}$	$\boldsymbol{LR_{cc}}$	**DQ**
GJR-ST	**0.656***	0.581*	**0.778***	0.891*	0.156*	**0.687***	0.337*	**0.846***
LSV	0.382*	**0.732***	0.644*	0.361*	0.426*	0.250*	0.376*	0.231*
SVJ	0.382*	**0.732***	0.644*	0.908*	**0.920***	0.470*	**0.827***	0.507*
2-SVJ	0.382*	**0.732***	0.644*	**0.925***	0.701*	0.590*	0.803*	0.398*
2001-2014								
	5% VaR				1% VaR			
	$\boldsymbol{LR_{uc}}$	$\boldsymbol{LR_{id}}$	$\boldsymbol{LR_{cc}}$	**DQ**	$\boldsymbol{LR_{uc}}$	$\boldsymbol{LR_{id}}$	$\boldsymbol{LR_{cc}}$	**DQ**
GJR-ST	**0.029**	0.348*	**0.059***	**0.016**	**0.641***	**0.365***	**0.595***	**0.775***
LSV	0.000	0.233*	0.000	0.000	0.000	0.098*	0.000	0.000
SVJ	0.000	**0.544***	0.000	0.000	0.000	0.150*	0.000	0.000
2-SVJ	0.000	0.260*	0.000	0.000	0.057*	0.261*	0.088*	0.000

P-values of the Christoffersen test for unconditional coverage, independence and conditional coverage as well as the DQ test for both out-of-sample periods. * indicates correct VaR's at 5% significance level.

Lastly, I formally test the difference in log-likelihood through the Diebold-Mariano type test. As discussed earlier the likelihood comparison is not completely fair due to the different estimation methods. The results are given in Table 4, where a positive number indicates the model in the row has a higher log-likelihood than the one in the column. Only the GJR-ST model achieves a log-likelihood significantly higher than the other models for the second out-of-sample period. All other differences in log-likelihood are not significant.

Table 4: Likelihood based Diebold-Mariano test results

	GJR-ST	LSV	SVJ	2-SVJ
GJR-ST	-	2.341*	2.905*	2.376*
LSV	-1.522	-	1.269	0.641
SVJ	-1.290	0.173	-	-2.370*
2-SVJ	-1.484	0.022	-0.834	-

DM-test statistic between the different models. Bottom left results are for the period 2001-2007, top right for 2001-2014. A positive number indicates the model in the row has a higher likelihood than the one in the column. * indicates a significant difference at 5% significance level.

Combining all these results, I conclude that none of the models are fully satisfactory. The informal analysis by Diebold et al. (1998) shows that uniformity and dependence are problematic, and all models are rejected by the relevant Hong-Li test. However, if one model had to be chosen, my preference would go to the GJR-ST model, as it has slightly better test statistics compared to its competitors, especially during the second out-of-sample period. Specifically, the model scores best on the relevant HL test during both periods, and has the best VaR coverage. Moreover, the GJR-ST is also less computationally expensive to estimate.

4.2.1 Dynamic parameter estimates

As discussed before, the static parameters might lead to sub optimal results. To investigate the amount of improvement possible, I estimate the GJR-ST model dynamically. This is done by using a moving window; a window length is chosen, and then the parameters are estimated using data within the window length only. Through this approach, parameters are (slowly) adjusted for different financial circumstances. For the GJR-ST model, this is computationally expensive, but still feasible. However, for the stochastic volatility models this is not feasible, due to too high computational times of repeatedly estimating the model. Results of the distribution based tests and the log-likelihood are given in Table 2, for the window lengths of 0.5, 1, 3 and 10 years (the results of the statically estimated GJR-ST model are also included for comparison).

From the table, it can be seen that dynamic estimation leads to quite substantial improvement. A window length of about 0.5 to 1 years seems optimal from these results. With this window length, the test statistics improve, to the point where even the Berkowitz test does not reject correct model specification during the first out-of-sample period (all static models were rejected). Notably, log-likelihood value as well as the relevant Hong-Li statistic also get improved, although correct model specification is still rejected by the Hong-Li test. Overall this shows that dynamic estimation indeed improves the results substantially. For the stochastic volatility models, similar improvement is expected, if the computational resources would be available.

Table 5: Distribution based test results for dynamic estimation

2001-2007							
	KS	**JB**	**B**	**HL(5)**	**HL(10)**	**HL(20)**	**log-likelihood**
GJR-ST	0.030*(0.083)	7.55 (0.023)	9.1 (0.028)	8.1	10.0	13.6	-2311.6
GJR-ST, 0.5 year	**0.026***(0.183)	5.25*(0.072)	**4.3***(0.232)	**6.0**	**6.2**	**8.0**	**-2238.8**
GJR-ST, 1 year	0.028*(0.117)	2.37*(0.306)	5.6* (0.132)	6.7	7.2	9.5	-2268.1
GJR-ST, 3 years	0.035 (0.025)	1.56*(0.459)	7.0* (0.073)	8.6	9.3	11.0	-2292.9
GJR-ST, 10 years	0.034 (0.038)	**1.27***(0.530)	11.9 (0.001)	6.4	8.7	12.2	-2309.2
2001-2014							
	KS	**JB**	**B**	**HL(5)**	**HL(10)**	**HL(20)**	**log-likelihood**
GJR-ST	**0.018***(0.189)	16.1 (0.000)	15.2 (0.002)	12.8	14.0	17.7	-4870.6
GJR-ST, 0.5 year	0.021*(0.087)	5.63*(0.060)	**9.0** (0.030)	7.8	8.8	10.6	**-4705.6**
GJR-ST, 1 year	0.019*(0.141)	2.81*(0.244)	10.4 (0.015)	**6.8**	**7.2**	**8.6**	-4768.7
GJR-ST, 3 years	0.026 (0.019)	**2.03***(0.361)	13.9 (0.003)	11.9	12.4	13.7	-4815.5
GJR-ST, 10 years	0.025 (0.027)	4.64*(0.098)	15.2 (0.002)	10.9	12.4	15.6	-4844.0

Test results of the Kolmogorov-Smirnov, Jarque-Bera, Berkowitz and Hong-Li tests, as well as the log-likelihood for both out-of-sample periods, with p-values in parenthesis. The window length in years is indicated after the model. For the HL test, 5% critical values obtained from simulation are 3.40 for 2001-2007 and 3.58 for 2001-2014. * indicates the null hypothesis of correct model specification is not rejected at 5% significance level. Bold gives the model with the best result on the corresponding statistic.

4.3 Application to SPX option pricing

Given the estimated parameters from Table 9, I employ the models to price European call options using Monte Carlo simulation as described in Section 3.3. The Black-Scholes (BS) model is used as a benchmark, with estimated parameters[5] $\mu = 0.0469$ and $\sigma = 0.9457$. For simplicity I use a risk-free rate of 3% (which is close to the federal funds rate in 2001). Full results are shown in Table 12 in the Appendix. Most models have a quite substantial bias, indicated by the average error being substantially different from zero. Therefore, I will focus my analysis on the errors that have been corrected for this bias as described in Section 3.3.

For clarity I will again show results from the best performing models within each class, which for option pricing are the GJR-N, LSV, SVJ and 2-SV3J model (which can be seen from Table 12). The GJR-N model performing well is a bit surprising, since GARCH type models with a normal distribution did not perform well for density forecasts. However, this result cannot be attributed to coincidence, as the GARCH-N model also outperforms the GARCH-T model, suggesting that a normal is more appropriate in the context of option pricing.

Results of the best performing models within each class are shown in Table 6. All 4 models are able to outperform the Black-Scholes model on both corrected MSE and MAPE. The SVJ and 2-SV3J models perform especially well on corrected MSE, mainly because these models price the options with a long maturity more accurately, as shown in Figure 6, where the corrected errors are plotted against the option maturity. This can be explained by the fact that the likelihood of a jump occurring in the actual data is higher during a longer time to maturity, meaning that incorporating this jump into the model specification becomes more relevant.

Table 6: Accuracy of option prices given by the different models

	mean error	**MSE**	**MAPE**	**corrected MSE**	**corrected MAPE**
BS	**-4.1113**	122.4	**0.2137**	97.81	0.1154
GJR-N	-6.4470	82.22	0.2222	35.72	0.0671
LSV	-8.4358	136.6	0.2912	58.57	0.0843
SVJ	-9.9970	115.3	0.2560	**9.899**	0.0762
2-SV3J	-8.0801	**81.09**	0.2188	10.90	**0.0542**

The mean error, mean squared error and mean absolute percentage error and their corrected (see Section 3.3) version, for at the money options traded on the first 7 trading days of 2001. 10.000 simulated sample paths are used to calculate the option price. Bold gives the model with the best result on the corresponding statistic.

Furthermore, Table 7 shows that the differences in corrected MSE and MAPE are mostly significant. Especially the difference between the Black-Scholes model and all other models is always significant in favor of the stochastic volatility models. Also, the SVJ and 2-SV3J models score significantly better than all other models on corrected MSE. Between these two, there is no significant difference in MSE, but a very significant difference in MAPE. Meaning that from an absolute point of view, the models make similar errors, but from a relative point of view the

[5]Parameters are estimated similarly to the other models; using Bayesian MCMC on the period 1990-2000

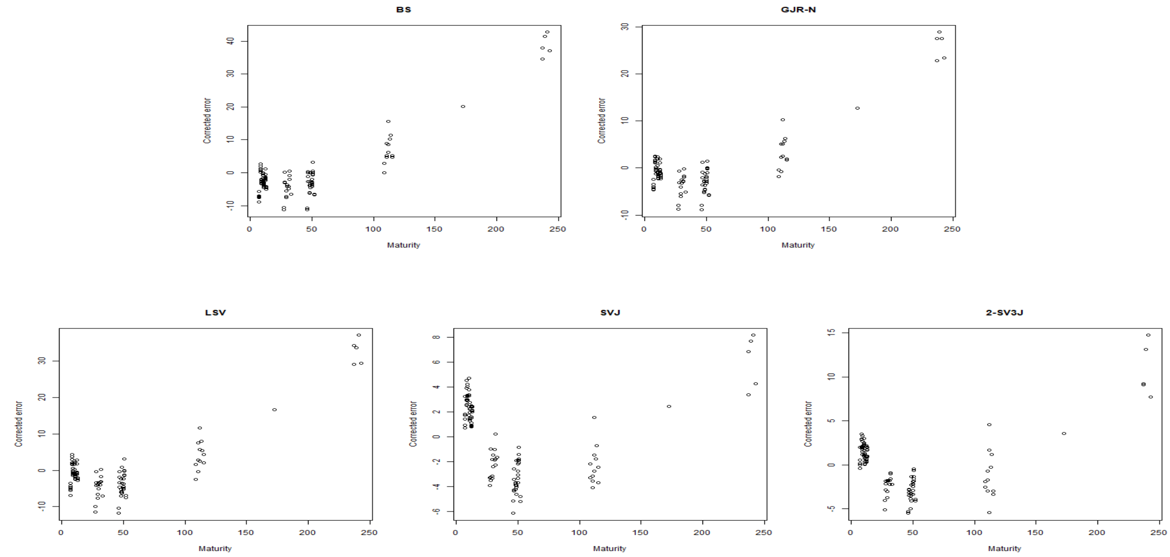


Figure 6: Corrected error plotted against the maturity of the option

2-SV3J model is strongly preferred. The reason that the test statistics are not as large as one could expect from the results of Table 6, is that there is also large variance in the size of the errors made.

These results show that the stochastic volatility models can be effectively applied to option pricing. The models lead to smaller errors than the benchmark Black-Scholes model. Looking at the absolute performance, which corresponds to the MSE, the SVJ and 2-SV3J model are the best performing. Notably, both of these models include jumps. For relative performance, corresponding to the MAPE, the 2-SV3J model performs best. This is in contrast with the results for density forecasting, where the GJR-ST model is the best performing. This shows that the extra flexibility allowed by the inclusion of jumps turns out to be more relevant for option pricing.

Table 7: Diebold-Mariano test results for corrected MSE and MAPE

	BS	GJR-N	LSV	SVJ	2-SV3J
BS	-	7.244*	4.672*	4.038*	5.894*
GJR-N	-3.088*	-	-8.017*	-1.293	2.235*
LSV	-2.884*	3.333*	-	1.449	4.604*
SVJ	-2.814*	-2.313*	-2.741*	-	6.755*
2-SV3J	-2.938*	-2.586*	-2.937*	-0.387	-

DM-test statistic between the different models. Bottom left results are for the corrected MSE, top right for the corrected MAPE. A positive number indicates the model in the row has higher errors than the one in the column. * indicates a significant difference at 5% significance level.

5 Conclusion

In this paper I have examined a wide range of stochastic volatility models, focusing on density forecasting and option pricing. First off all I have shown that it is important to incorporate the fat tails and the leverage effect in the model. As the models that include these phenomena are the best performing within their respective model class.

For density forecasts, none of the models are fully satisfactory, but the GJR-ST model performs best. It scores best on the relevant distribution based test statistics, specifically the Hong-Li test, during both out-of-sample periods. Furthermore, the GJR-ST model is the only model that gives correct conditional coverage for both 5% and 1% value at risk, measured by both the Christoffersen and DQ tests, during both sample periods (with the exception of the DQ test for sample period 2 for 5% VaR). The model also has a higher likelihood than the competing models, with this difference being significant for the second out-of-sample period, although the likelihood based comparison is not completely fair. Results for the other models are also reasonable, for example the LSV, SVJ and 2-SVJ model have correct conditional coverage for both 5% and 1% VaR for the first out-of-sample period. I have in addition shown that dynamic estimation leads to substantial further improvement of the GJR-ST model. Dynamic estimation is not yet feasible for the stochastic volatility models.

For option pricing however, the GJR-ST is not even the best performing among the GARCH type models, being outperformed by the GJR-N model. The best performing models within each class significantly outperform the standard Black-Scholes model. In contrast to density forecasting, the best performing models are the SVJ and 2-SV3J model. The increased flexibility, induced by the introduction of jumps, seems to be especially effective for the pricing of (long-maturity) options. Looking at the errors from an absolute perspective, the SVJ and 2-SV3J model are similar, but from a relative perspective, the 2-SV3J model makes significantly smaller errors.

All in all this shows that the question "Which stochastic volatility model is most accurate?", can have a different answer depending on the type of application the stochastic volatility model is used for. For making density forecasts of the S&P 500 I have shown that the GJR-ST is the best of the models considered, while the 2-SV3J model is preferred for option pricing. With the field of quantitative finance constantly evolving, and new, more sophisticated models being introduced constantly, a lot of research is still necessary to give a definite answer, if one even exists.

For future research one of the most obvious suggestions is to look into making dynamic parameter estimation feasible, as it can lead to substantial improvement of the models. Dynamic estimation helps improve the flexibility of the models, as it allows the models to pick up on changes in parameters, for example due to regime switches, in the data. Of course developing and evaluating different stochastic volatility models, as well as looking into different data sets or financial applications is also very useful. Lastly the usage of not only the daily returns, but also different exogenous variables that potentially contain information to improve the models is interesting.

References

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *The Review of Financial Studies*, 9(1):69–107.

Bates, D. S. (2000). Post-'87 crash fears in the sp 500 futures option market. *Journal of Econometrics*, 94(1-2):181–238.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–474.

Black, F. and Scholes, M. S. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3):637–654.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T. (2008). Glossary to arch (garch). *CREATES Research Paper*, 49.

Boyle, P. P. (1977). Options: A monte carlo approach. *Journal of Financial Economics*, 4(3):323–338.

Broadie, M. and Kaya, Ö. (2006). Exact simulation of stochastic volatility and other affine jump diffusion processes. *Operations Research*, 54(2):217–231.

Chernov, M., Gallant, A. R., Ghysels, E., and Tauchen, G. (2003). Alternative models for stock price dynamics. *Journal of Econometrics*, 116(1-2):225–257.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4):841–862.

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

Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407.

Cox, J. C. and Ross, S. A. (1976). The valuation of options for alternative stochastic processes. *Journal of financial economics*, 3(1-2):145–166.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Duffie, D., Pan, J., and Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6):1343–1376.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

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

Eraker, B., Johannes, M., and Polson, N. (2003). The impact of jumps in volatility and returns. *The Journal of Finance*, 58(3):1269–1300.

Fernández, C. and Steel, M. F. J. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

Hong, Y. and Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *The Review of Financial Studies*, 18(1):37–84.

Hong, Y., Li, H., and Zhao, F. (2007). Can the random walk model be beaten in out-of-sample density forecasts? evidence from intraday foreign exchange rates. *Journal of Econometrics*, 141(2):736–776.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business Economic Statistics*, 12(4):371–389.

JAGS (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. Available at: `http://mcmc-jags.sourceforge.net/`.

Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255 – 259.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Kaeck, A. and Alexander, C. (2012). Volatility dynamics for the s&p 500: Further evidence from non-affine, multi-factor jump diffusions. *Journal of Banking & Finance*, 36(11):3110–3121.

Kolmogorov, A. (1933). Sulla determinazione empirica di uma legge di distribuzione.

Li, H., Wells, M. T., and Yu, C. L. (2008). A bayesian analysis of return dynamics with lévy jumps. *The Review of Financial Studies*, 21(5):2345–2378.

Meyer, R. and Yu, J. (2000). Bugs for a bayesian analysis of stochastic volatility models. *The Econometrics Journal*, 3(2):198–215.

R (2013). R: A language and environment for statistical computing. Available at: `http://www.R-project.org/`.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470–472.

rugarch (2020). *rugarch: Univariate GARCH models.* R package version 1.4-2.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19(2):279–281.

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75. *Time series analysis: theory and practice*, 1:203–226.

tseries (2019). *tseries: Time Series Analysis and Computational Finance.* R package version 0.10-47.

Yu, J. (2005). On leverage in a stochastic volatility model. *Journal of Econometrics*, 127(2):165–178.

Yun, J. (2018). Density Forecast Evaluations via a Simulation-Based Dynamic Probability Integral Transformation. *Journal of Financial Econometrics*, 18(1):24–58.

6 Appendix

6.1 Tables

Table 8: Performance of different tests on simulated data

		LSV0			**LSV**			**SV**		
		10%	5%	1%	10%	5%	1%	10%	5%	1%
T=250	**KS**	0.096	0.048	0.013	0.084	0.030	0.003	0.095	0.048	0.013
	JB	0.081	0.039	0.014	0.083	0.047	0.017	0.078	0.041	0.016
	B	0.090	0.045	0.008	0.087	0.038	0.004	0.098	0.041	0.010
	HL(5)	0.092	0.043	0.011	0.087	0.036	0.006	0.099	0.051	0.007
T=500	**KS**	0.106	0.050	0.003	0.093	0.042	0.006	0.107	0.051	0.008
	JB	0.098	0.040	0.013	0.079	0.041	0.014	0.094	0.051	0.019
	B	0.119	0.064	0.014	0.094	0.049	0.008	0.118	0.060	0.010
	HL(5)	0.097	0.057	0.009	0.086	0.041	0.013	0.109	0.055	0.009
T=1000	**KS**	0.115	0.058	0.010	0.102	0.049	0.010	0.100	0.048	0.011
	JB	0.097	0.052	0.017	0.098	0.062	0.021	0.102	0.056	0.018
	B	0.094	0.053	0.015	0.097	0.045	0.012	0.105	0.057	0.013
	HL(5)	0.096	0.042	0.010	0.122	0.061	0.008	0.105	0.049	0.007
		SVJ			**SV2J**			**2-SV**		
		10%	5%	1%	10%	5%	1%	10%	5%	1%
T=250	**KS**	0.106	0.057	0.013	0.097	0.052	0.013	0.096	0.055	0.008
	JB	0.080	0.043	0.017	0.082	0.049	0.018	0.083	0.045	0.020
	B	0.103	0.060	0.008	0.097	0.047	0.013	0.090	0.040	0.008
	HL(5)	0.105	0.055	0.017	0.113	0.055	0.011	0.091	0.042	0.007
T=500	**KS**	0.106	0.054	0.008	0.109	0.049	0.007	0.104	0.045	0.005
	JB	0.094	0.053	0.016	0.087	0.044	0.012	0.094	0.036	0.008
	B	0.111	0.060	0.008	0.105	0.062	0.012	0.092	0.051	0.005
	HL(5)	0.103	0.047	0.011	0.111	0.054	0.013	0.088	0.039	0.009
T=1000	**KS**	0.106	0.056	0.011	0.109	0.048	0.011	0.090	0.043	0.009
	JB	0.087	0.044	0.014	0.100	0.049	0.016	0.084	0.049	0.010
	B	0.101	0.046	0.007	0.107	0.058	0.013	0.113	0.051	0.009
	HL(5)	0.107	0.052	0.009	0.111	0.057	0.012	0.092	0.042	0.006
		2-SVJ			**2-SV2J**			**2-SV3J**		
		10%	5%	1%	10%	5%	1%	10%	5%	1%
T=250	**KS**	0.102	0.048	0.011	0.114	0.058	0.008	0.104	0.046	0.011
	JB	0.069	0.044	0.020	0.087	0.049	0.022	0.095	0.058	0.023
	B	0.110	0.052	0.005	0.103	0.058	0.010	0.127	0.068	0.016
	HL(5)	0.109	0.056	0.016	0.092	0.046	0.004	0.102	0.053	0.008
T=500	**KS**	0.115	0.065	0.013	0.114	0.061	0.011	0.100	0.048	0.010
	JB	0.081	0.041	0.011	0.093	0.051	0.020	0.108	0.060	0.027
	B	0.115	0.062	0.019	0.115	0.056	0.007	0.124	0.075	0.018
	HL(5)	0.116	0.065	0.013	0.123	0.064	0.006	0.108	0.057	0.010
T=1000	**KS**	0.093	0.047	0.013	0.099	0.046	0.010	0.093	0.049	0.014
	JB	0.098	0.055	0.017	0.115	0.060	0.018	0.121	0.069	0.023
	B	0.101	0.045	0.009	0.090	0.052	0.008	0.124	0.066	0.015
	HL(5)	0.111	0.065	0.011	0.100	0.052	0.013	0.107	0.058	0.015

Fraction rejected based on 1000 simulated sample paths of length of the return, using the estimated coefficients from Table 9. The sample paths have length $T = 250, 500$ and 1000. For the HL test, critical values for 10%, 5% and 1% are given by 2.52, 3.56 and 5.73 for $T = 250$, 2.52, 3.43 and 5.41 for $T = 500$ and 2.54, 3.43 and 5.605 for $T = 1000$ respectively (Yun (2018)).

Table 9: Estimated coefficients for density forecast models

	GARCH-N	**GARCH-T**	**GJR-N**	**GJR-T**	**GJR-ST**
μ	0.0548*(0.0142)	0.0608*(0.0131)	0.0383*(0.0136)	0.0492*(0.0132)	0.0411*(0.0140)
α_0	0.0047*(0.0017)	0.0029*(0.0013)	0.0100 (0.0063)	0.0063*(0.0024)	0.0067*(0.0025)
α_1	0.9439*(0.0088)	0.9538*(0.0142)	0.9291*(0.0290)	0.9403*(0.0115)	0.9391*(0.0116)
α_2	0.0525*(0.0082)	0.0447*(0.0048)	0.0136 (0.0089)	0.0119 (0.0079)	0.0117 (0.0079)
α_3	-	-	0.0938*(0.0427)	0.0829*(0.0196)	0.0855*(0.0197)
ν	-	6.1474*(0.6990)	-	6.6636*(0.8127)	6.8472*(0.8649)
ξ	-	-	-	-	0.9547 (0.0252)

	LSV0	**LSV**
μ	0.0641*(0.0132)	0.0416*(0.0135)
α	-0.0059*(0.0017)	-0.0087*(0.0039)
β	-0.0133*(0.0046)	-0.0196*(0.0054)
ρ	-	-0.5629*(0.0597)
σ_v^2	0.1339*(0.0175)	0.1671*(0.0198)

	SV	**SVJ**	**SV2J**
μ	0.0421*(0.0141)	0.0436*(0.0136)	0.0436*(0.0138)
μ_r	-	-1.2948 (0.7446)	-4.4274*(0.8419)
σ_r	-	2.0144*(0.3563)	2.1500*(0.4291)
θ	0.9458*(0.1385)	0.9261*(0.1714)	0.7585*(0.1219)
κ	0.0166*(0.0041)	0.0142*(0.0042)	0.0206*(0.0046)
σ_v	0.1237*(0.0110)	0.1117*(0.0106)	0.1172*(0.0098)
ρ	-0.4175*(0.0884)	-0.5407*(0.0689)	-0.4729*(0.0779)
μ_σ	-	-	0.6206*(0.1474)
λ	-	0.0116*(0.0050)	0.0034*(0.0013)

	2-SV	**2-SVJ**	**2-SV2J**	**2-SV3J**
μ	0.0367*(0.0131)	0.0451*(0.0128)	0.0300*(0.0131)	0.0327*(0.0133)
μ_r	-	-1.5395 (0.8919)	-2.6110*(1.1445)	-2.7016*(1.0796)
σ_r	-	2.0001*(0.3368)	2.2159*(0.4176)	2.2040*(0.3955)
κ_M	0.4879 (0.3036)	0.8050 (0.4669)	0.9025*(0.0142)	0.9994*(0.1509)
θ_M	1.0324*(0.0294)	0.9603*(0.0272)	0.9983*(0.0368)	1.1027*(0.0508)
σ_M	0.0781*(0.0319)	0.0545*(0.0180)	0.0475*(0.0099)	0.0631*(0.0241)
μ_M	-	-	-	0.4542*(0.1081)
κ	0.0150*(0.0031)	0.0125*(0.0028)	0.0156*(0.0041)	0.0126*(0.0026)
σ_v	0.1245*(0.0092)	0.1105*(0.0089)	0.1204*(0.0150)	0.1183*(0.0102)
ρ	-0.4448*(0.0628)	-0.5233*(0.0700)	-0.5277*(0.0802)	-0.4872*(0.0769)
μ_σ	-	-	0.6441*(0.1585)	0.6873*(0.1631)
λ	-	0.0104*(0.0049)	0.0043*(0.0015)	0.0034*(0.0014)

Estimated coefficients of daily returns of the S&P 500 index for the period 1990-2000. The parameters of the models are of the models in Section 3, with standard errors in parenthesis. * indicates significance at 5% significance level. The estimates for the GARCH type models are the ML estimators. For all other models, the estimators are the mean and standard deviation of the posterior distribution, given by the Bayesian MCMC. The Bayesian MCMC uses 11.000 iterations and 4 chains, 1000 iterations are burn-in.

Table 10: Distribution based test results

2001-2007							
	KS	**JB**	**B**	**HL(5)**	**HL(10)**	**HL(20)**	**log-likelihood**
GARCH-N	0.047 (0.001)	264 (0.000)	9.8 (0.021)	16.6	21.8	30.4	-2357.4
GARCH-T	0.040 (0.007)	8.7 (0.013)	11.6 (0.009)	12.5	15.3	20.7	-2341.1
GJR-N	0.042 (0.004)	175 (0.000)	9.1 (0.028)	15.2	20.7	28.9	-2325.7
GJR-T	0.034 (0.037)	11.3 (0.004)	9.7 (0.021)	10.5	13.1	18.1	-2315.3
GJR-ST	0.030*(0.083)	7.55 (0.023)	9.1 (0.028)	**8.1**	**10.0**	**13.6**	**-2311.6**
LSV0	0.036 (0.020)	9.65 (0.008)	12.3 (0.006)	13.0	16.5	22.4	-2341.5
LSV	0.029*(0.096)	14.3 (0.001)	8.8 (0.032)	10.0	13.1	17.8	-2326.8
SV	0.030*(0.087)	12.1 (0.002)	8.0 (0.046)	9.3	12.3	16.8	-2319.6
SVJ	0.027*(0.148)	**3.07***(0.215)	8.1 (0.043)	9.4	12.4	17.0	-2319.7
SV2J	0.028*(0.128)	4.09*(0.130)	7.9 (0.047)	10.8	14.0	18.7	-2320.9
2-SV	0.029*(0.109)	13.1 (0.001)	**7.7***(0.053)	10.8	14.2	19.0	-2325.0
2-SVJ	**0.026***(0.194)	4.35*(0.114)	8.4 (0.038)	9.2	12.2	16.3	-2321.9
2-SV2J	0.027*(0.151)	4.95*(0.084)	11.6 (0.009)	11.4	14.7	19.9	-2317.6
2-SV3J	0.029*(0.107)	4.91*(0.444)	11.7 (0.008)	11.4	14.9	20.3	-2320.4
2001-2014							
	KS	**JB**	**B**	**HL(5)**	**HL(10)**	**HL(20)**	**log-likelihood**
GARCH-N	0.040 (0.000)	425 (0.000)	18.4 (0.000)	35.8	47.5	64.4	-4993.7
GARCH-T	0.027 (0.014)	22.0 (0.000)	19.1 (0.000)	21.0	24.9	30.4	-4944.8
GJR-N	0.039 (0.000)	280 (0.000)	**14.5** (0.002)	33.9	43.9	59.8	-4913.5
GJR-T	0.022*(0.065)	29.3 (0.000)	15.7 (0.001)	17.4	20.4	26.7	-4881.5
GJR-ST	**0.018***(0.189)	16.1 (0.000)	15.2 (0.002)	**12.8**	**14.0**	**17.7**	**-4870.6**
LSV0	0.026 (0.014)	47.0 (0.000)	25.4 (0.000)	22.0	27.5	36.0	-4943.5
LSV	0.025 (0.026)	56.1 (0.000)	17.7 (0.000)	18.8	23.0	31.3	-4893.3
SV	0.023 (0.047)	65.7 (0.000)	24.0 (0.000)	17.6	21.6	28.6	-4913.7
SVJ	0.024 (0.037)	17.4 (0.000)	20.1 (0.000)	18.0	22.1	29.2	-4911.7
SV2J	0.024 (0.032)	14.1 (0.001)	14.7 (0.002)	18.9	22.6	30.1	-4883.4
2-SV	0.026 (0.019)	67.1 (0.000)	19.7 (0.000)	20.0	24.5	32.8	-4910.5
2-SVJ	0.022*(0.064)	**13.8** (0.001)	17.6 (0.001)	17.2	20.5	27.4	-4902.0
2-SV2J	0.033 (0.001)	20.4 (0.000)	14.7 (0.002)	20.6	24.2	32.1	-4876.6
2-SV3J	0.032 (0.002)	22.3 (0.000)	15.0 (0.001)	20.9	24.7	32.8	-4883.7

Test results of the Kolmogorov-Smirnov, Jarque-Bera, Berkowitz and Hong-Li tests, as well as the log-likelihood for both out-of-sample periods, with p-values in parenthesis. For the HL test, 5% critical values obtained from simulation are 3.40 for 2001-2007 and 3.58 for 2001-2014. * indicates the null hypothesis of correct model specification is not rejected at 5% significance level. Bold gives the model with the best result on the corresponding statistic.

Table 11: VaR based test results

2001-2007	5% VaR				1% VaR			
	LR_{uc}	LR_{id}	LR_{cc}	DQ	LR_{uc}	LR_{id}	LR_{cc}	DQ
GARCH-N	0.233*	0.154*	0.178*	0.150*	0.004	0.293*	0.008	0.000
GARCH-T	0.069*	0.066*	0.035	0.026	0.701*	0.590*	0.803*	0.102*
GJR-N	0.580*	0.618*	0.758*	0.693*	0.145*	0.417*	0.248*	0.001
GJR-T	0.382*	0.437*	0.505*	0.732*	0.250*	0.662*	0.469*	0.168*
GJR-ST	0.656*	0.581*	0.778*	0.891*	0.156*	**0.687***	0.337*	**0.846***
LSV0	0.069*	0.141*	0.065*	0.149*	0.145*	0.336*	0.217*	0.119*
LSV	0.382*	0.732*	0.644*	0.361*	0.426*	0.250*	0.376*	0.231*
SV	0.233*	0.303*	0.290*	0.433*	0.308*	0.457*	0.451*	0.003
SVJ	0.382*	0.732*	0.644*	0.908*	**0.920***	0.470*	**0.827***	0.507*
SV2J	0.327*	0.469*	0.476*	0.433*	0.526*	0.613*	0.720*	0.323*
2-SV	0.277*	0.279*	0.308*	0.519*	0.308*	0.617*	0.451*	0.003
2-SVJ	0.382*	0.732*	0.644*	**0.925***	0.701*	0.590*	0.803*	0.398*
2-SV2J	**0.991***	**0.837***	**0.979***	0.770*	0.374*	0.638*	0.602*	0.242*
2-SV3J	0.904*	0.809*	0.964*	0.751*	0.374*	0.638*	0.602*	0.242*
2001-2014	5% VaR				1% VaR			
	LR_{uc}	LR_{id}	LR_{cc}	DQ	LR_{uc}	LR_{id}	LR_{cc}	DQ
GARCH-N	0.009	**0.846***	0.031	0.000	0.000	0.519*	0.000	0.000
GARCH-T	0.001	0.777*	0.003	0.000	0.112*	0.282*	0.158*	0.000
GJR-N	0.016	0.298*	0.032	0.012	0.000	0.113*	0.000	0.000
GJR-T	0.007	0.394*	0.018	0.002	0.264*	0.316*	0.324*	0.024
GJR-ST	**0.029**	0.348*	**0.059***	**0.016**	**0.641***	0.365*	**0.595***	**0.775***
LSV0	0.000	0.803*	0.000	0.000	0.000	**0.746***	0.000	0.000
LSV	0.000	0.233*	0.000	0.000	0.000	0.098*	0.000	0.000
SV	0.000	0.643*	0.000	0.000	0.000	0.088*	0.000	0.000
SVJ	0.000	0.544*	0.000	0.000	0.000	0.150*	0.000	0.000
SV2J	0.000	0.389*	0.000	0.000	0.012	0.222*	0.021	0.001
2-SV	0.000	0.457*	0.001	0.000	0.000	0.103*	0.000	0.001
2-SVJ	0.000	0.260*	0.000	0.000	0.057*	0.261*	0.088*	0.000
2-SV2J	0.013	0.033	0.005	0.003	0.202*	0.304*	0.262*	0.000
2-SV3J	0.024	0.042	0.010	0.009	0.152*	0.293*	0.206*	0.001

P-values of the Christoffersen test for unconditional coverage, independence and conditional coverage as well as the DQ test for both out-of-sample periods. * indicates correct VaR's at 5% significance level. Bold gives the model with the best result on the corresponding statistic.

Table 12: Accuracy of option prices given by the different models

	mean error	**MSE**	**MAPE**	**corrected MSE**	**corrected MAPE**
BS	-4.1113	122.4	0.2137	97.81	0.1154
GARCH-N	2.9569	267.4	**0.1386**	244.3	0.2080
GARCH-T	22.390	1351	0.3542	812.0	0.4620
GJR-N	-6.4470	82.22	0.2222	35.72	0.0671
GJR-T	9.0989	532.8	0.1685	429.1	0.3253
GJR-ST	-4.5447	90.92	0.2040	64.10	0.0962
LSV0	**1.0689**	473.3	0.2526	443.7	0.2777
LSV	-8.4358	136.6	0.2912	58.57	0.0843
SV	-5.3649	99.20	0.2174	64.28	0.0907
SVJ	-9.9970	115.3	0.2560	**9.899**	0.0762
SV2J	-11.146	143.3	0.2955	13.70	0.0778
2-SV	-6.2390	**78.26**	0.2111	34.24	0.0613
2-SVJ	-9.2444	101.6	0.2428	10.61	0.0664
2-SV2J	-11.003	138.9	0.2686	11.22	0.0905
2-SV3J	-8.0801	81.09	0.2188	10.90	**0.0542**

The mean error, mean squared error and mean absolute percentage error and their corrected (see Section 3.3) version, for at the money options traded on the first 7 trading days of 2001. 10.000 simulated sample paths are used to calculate the option price. Bold gives the model with the best result on the corresponding statistic.

6.2 Dynamic probability integral transform for LSV

The simulation-based dynamic probability integral transform algorithm for the LSV (Section 3.1.2) model as described in Yun (2018) consists of the following steps, given an amount of particles P (Like Yun (2018), I work with $P = 25000$):

1. **Auxiliary variable**: Suppose a set of P particles for conditional variances, $\{\sigma^2_{t-1}(p)\}_{p=1}^P$, is given, drawn from $f(\sigma^2_{t-1}|r_{t-1})$. Draw the initial particles according to the marginal distribution given by the model: $\{\sigma^2_0(p)\}_{p=1}^P \sim N(\frac{\alpha}{-\beta}, \frac{\sigma_v^2}{1-(1+\beta)^2})$. Then compute the auxiliary variable $\hat{\sigma}^2_t(p)$ for each $p = 1, ..., P$. For each p, $\hat{\sigma}^2_t(p)$ is the conditional expectation of σ^2_t given $\sigma^2_{t-1}(p)$ and r_{t-1}. This auxiliary variable is used to improve efficiency and is given by:

$$\hat{\sigma}^2_t(p) = \exp(\log(\sigma^2_{t-1}(p)) + \alpha + \beta\log(\sigma^2_{t-1}(p)) + \frac{\rho\sigma_v(r_{t-1} - \mu)}{\sigma_{t-1}(p)} + \frac{1}{2}\sigma_v^2(1-\rho^2) \quad (24)$$

   The last term $\frac{1}{2}\sigma_v^2(1-\rho^2)$ comes from Jensen's inequality adjustment.

2. **First-stage resampling**: Using $\hat{\sigma}^2_t(p)$ from the previous step, compute the first-stage weights for each p according to:

$$\tilde{w}_t(p) \propto \phi(r_t|\mu, \hat{\sigma}^2_t(p)) \quad (25)$$

   where ϕ is the normal distribution PDF. Standardize the weights to $w_t(p) = \frac{\tilde{w}_t(p)}{\sum_{p=1}^P \tilde{w}_t(p)}$. Using these weights, resample the particles $\hat{\sigma}^2_t(p)$ (meaning draw the particles with replacement, with the probability of drawing the particle $\hat{\sigma}^2_t(p)$ given by $w_t(p)$). Let $\{\sigma^2_{t-1}(p, w)\}_{p=1}^P$ denote the set of resampled particles.

3. **Second-stage resampling**: Simulate the log conditional variance at time t, i.e. $\sigma^2_t(p, w)$, by drawing $\tilde{\eta}_t(p, w) \sim N(0, 1)$ and computing:

$$\log(\sigma^2_t(p, w)) = \log(\sigma^2_{t-1}(p, w)) + \alpha + \beta\log(\sigma^2_{t-1}(p, w)) + \frac{\rho\sigma_v(r_{t-1} - \mu)}{\sigma_{t-1}(p, w)} + \sigma_v\sqrt{1-\rho^2}\tilde{\eta}_t(p, w) \quad (26)$$

   Then compute the second-stage weight for each (p, w):

$$\tilde{\pi}_t(p, w) \propto \frac{f(r_t|\sigma^2_t(p, w))}{f(r_t|\hat{\sigma}^2_t(p, w))} = \frac{\phi(r_t|\mu, \sigma^2_t(p, w))}{\phi(r_t|\mu, \hat{\sigma}^2_t(p, w))} \quad (27)$$

   Again standardize the weights $\pi_t(p) = \frac{\tilde{\pi}_t(p)}{\sum_{p=1}^P \tilde{\pi}_t(p)}$. Use these weights to resample the particles $\sigma^2_t(p, w)$. Let $\{\sigma^2_t(p)\}_{p=1}^P$ denote these final particles, which are now equivalent to a sample from $f(\sigma^2_t|r_t)$

4. **Simulation-based dynamic probability integral transform and log-likelihood evaluation**: For each particle $\sigma^2_t(p)$, draw $\tilde{\sigma}^2_{t+1}(p)$ from the transition density of the model, given in Equation 8. Using the drawn $\{\tilde{\sigma}^2_{t+1}(p)\}_{p=1}^P$, approximate the generalized

residual z_{t+1} through the PIT as follows:

$$\begin{aligned} z_{t+1} &\approx \frac{1}{P}\sum_{p=1}^{P}\int_{-\infty}^{r_t+1} f(r|\tilde{\sigma}^2_{t+1}(p))dr \\ &= \frac{1}{P}\sum_{p=1}^{P}\int_{-\infty}^{r_t+1} \phi(r|\mu,\tilde{\sigma}^2_{t+1}(p))dr \end{aligned} \tag{28}$$

Then calculate the log-likelihood as follows:

$$\begin{aligned} \mathcal{L}_{t+1} &= \log f(r_{t+1}|r_t) \\ &\approx \log[\frac{1}{P}\sum_{p=1}^{P}\phi(r_{t+1}|\mu,\tilde{\sigma}^2_{t+1}(p))] \end{aligned} \tag{29}$$

Lastly discard $\{\tilde{\sigma}^2_{t+1}(p)\}_{p=1}^P$

5. **Iteration**: Go back to step 1, and start to update the particles from $\{\sigma^2_t(p)\}_{p=1}^P$ to $\{\sigma^2_{t+1}(p)\}_{p=1}^P$ with r_{t+1} until $t = T-1$.

6.3 Dynamic probability integral transform for 2-SV3J

The simulation-based dynamic probability integral transform algorithm for the 2-SV3J model (Section 3.1.3) as described in Yun (2018) consists of the following steps, given an amount of particles P (Like Yun (2018), I work with $P = 25000$):

1. **Auxiliary variable**: Suppose a set of P particles at time $t-1$, $\{L_{t-1}(p)\}_{p=1}^P$, is given, with $L_{t-1}(p) = (M_{t-2}(p), \sigma^2_{t-2}(p), J_{t-1}(p), Z^r_{t-1}(p), Z^M_{t-2}(p), Z^\sigma_{t-2}(p))$ generated from $f(L_{t-1}|r_{t-1})$. Draw the initial particles according to the marginal distribution given by the model $\{L_0(p)\}_{p=1}^P$. Compute the auxiliary variables, which are the conditional expectation of the long-term and spot variance on $L_{t-1}(p)$ and r_{t-1}:

$$\hat{M}_{t-1}(p) = M_{t-2}(p) + \kappa_M(\theta_M - M_{t-2}(p)) + J_{t-1}(p)\mu_M \tag{30}$$

$$\hat{\sigma}^2_{t-1}(p) = \sigma_{t-2}(p) + \kappa(\hat{M}_{t-1}(p) - \sigma^2_{t-2}(p)) + \rho\sigma_v(r_{t-1} - \mu - J_{t-1}(p)Z^r_{t-1}(p)) + J_{t-1}(p)\mu_\sigma \tag{31}$$

2. **First-stage resampling**: Using $\hat{\sigma}^2_{t-1}(p)$ from step 1, compute the first-stage weights for each p:

$$\tilde{w}_t(p) \propto \lambda\phi(r_t|\mu+\mu_r, \hat{\sigma}^2_{t-1}(p) + \sigma^2_r) + (1-\lambda)\phi(r_t|\mu, \hat{\sigma}^2_{t-1}(p)) \tag{32}$$

Then standardize the weights $w_t(p) = \frac{\tilde{w}_t(p)}{\sum_{p=1}^P \tilde{w}_t(p)}$, and resample the particles $\{L_{t-1}(p)\}_{p=1}^P$ according to the weights $w_t(p)$, obtaining the set of resampled particles denoted $\{L_{t-1}(p,w)\}_{p=1}^P$. Also resample the weights to obtain $w_t(p,w)$, which is necessary for step 7.

3. **Generating a jump occurrence**: For each (p,w), draw an occurrence of the jump $J_t(p,w)$ from a Bernoulli distribution with jump probability $q_J(p,w)$ given by:

$$q_J(p,w) = Pr(J_t(p,w) = 1|\hat{\sigma}^2_{t-1}(p,w), r_t) \tag{33}$$

where

$$\begin{aligned} Pr(J_t(p,w)=1|\hat{\sigma}^2_{t-1},r_t) &\propto Pr(J_t(p,w)=j)Pr(r_t|\hat{\sigma}^2_{t-1}(p,w),J_t(p,w)=j) \\ &= \lambda^j(1-\lambda)^{1-j}\phi(r_t|\mu+j\mu_r,\hat{\sigma}^2_{t-1}(p,w)+j\sigma_r^2) \end{aligned} \tag{34}$$

for $j=0,1$.

4. **Generating a jump in return size**: For (p,w) with $J_t(p,w)=1$, draw a jump-in-return size $Z_t^r(p,w)$ from a normal distribution with mean $\bar{\mu}(p,w)$ and variance $\bar{\sigma}^2(p,w)$ given by:

$$\bar{\mu}(p,w)=\bar{\sigma}^2(p,w)(\frac{r_t-\mu}{\hat{\sigma}^2_{t-1}(p,w)}+\frac{\mu_r}{\sigma_r^2}) \tag{35}$$

$$\bar{\sigma}^2(p,w)=(\frac{1}{\hat{\sigma}^2_{t-1}(p,w)}+\frac{1}{\sigma_r^2})^{-1} \tag{36}$$

5. **Simulating long-term variance**: For (p,w) draw an innovation $u_{t-1}(p,w)$ and if $J_{t-1}(p,w)=1$, draw a jump in long-term variance size $Z^M_{t-1}(p,w)$ from $\exp(\mu_M)$. Then $M_{t-1}(p,w)$ can be generated through:

$$M_{t-1}(p,w)=M_{t-2}(p,w)+\kappa_M(\theta_M-M_{t-2}(p,w))+\sigma_M\sqrt{M_{t-2}(p,w)}u_{t-1}(p,w)+J_{t-1}(p,w)Z^M_{t-1}(p,w) \tag{37}$$

6. **Simulating spot variance**: For (p,w) draw an innovation $\eta_{t-1}(p,w)$ and if $J_{t-1}(p,w)=1$, draw a jump in spot variance size $Z^\sigma_{t-1}(p,w)$ from $\exp(\mu_\sigma)$. Then $\sigma^2_{t-1}(p,w)$ can be generated through:

$$\begin{aligned} \sigma^2_{t-1}(p,w) &= \sigma^2_{t-2}(p,w)+\kappa(M_{t-1}(p,w)-\sigma^2_{t-2}(p,w))+\sigma_v\rho(r_{t-1}-\mu-J_{t-1}(p,w)Z^r_{t-1}(p,w)) \\ &\quad +\sigma_v\sqrt{1-\rho^2}\sqrt{\sigma^2_{t-2}(p,w)}\eta_{t-1}(p,w)+J_{t-1}(p,w)Z^\sigma_{t-1}(p,w) \end{aligned} \tag{38}$$

The particles $\{L_t(p,w)\}_{p=1}^P$ have now been obtained.

7. **Second-stage resampling**: Calculate weights according to:

$$\tilde{\pi}_t(p,w)\propto\frac{\lambda^{J_t(p,w)}(1-\lambda)^{1-J_t(p,w)}\phi(Z_t^r(p,w)|\mu_r,\sigma_r^2)^{J_t(p,w)}\phi(r_t|\mu+J_t(p,w)Z_t^r(p,w),\sigma_t^2(p,w))}{q_J(p,w)^{J_t(p,w)}(1-q_J(p,w))^{1-J_t(p,w)}\phi(Z_t^r(p,w)|\bar{\mu}(p,w),\bar{\sigma}^2(p,w))w_t(p,w)} \tag{39}$$

Then standardize the weights $\pi_t(p)=\frac{\tilde{\pi}_t(p)}{\sum_{p=1}^P\tilde{\pi}_t(p)}$. Resample the particles $L_t(p,w)$ using these weights. Let $\{L_t(p)\}_{p=1}^P$ denote these final particles, which are now equivalent to a sample from $f(L_t|r_t)$.

8. **Simulation-based dynamic probability integral transform and log-likelihood evaluation**: For each particle $L_t(p)$, draw $\tilde{\eta}_t(p)$ from $N(0,1)$, $\tilde{J}_{t+1}(p)$ from Ber(λ) and $\tilde{Z}_t^\sigma(p)$ from $\exp(\mu_\sigma)$. $\{\tilde{\sigma}_t^2(p)\}_{p=1}^P$ can now be obtained from the transition density of the model, given in Equation 14. Using the simulated $\{\tilde{\sigma}_t^2(p),\tilde{J}_{t+1}(p)\}_{p=1}^P$, approximate the

generalized residual z_{t+1} through the PIT as follows:

$$
\begin{aligned}
z_{t+1} &= \int_{-\infty}^{r_{t+1}} f(r|r_t)dr \\
&= \int_{-\infty}^{r_{t+1}} \int_0^{\infty} \int f(r|\sigma_t^2) f(\sigma_t^2|L_t, r_t) f(L_t|r_t) dL_t d\sigma_t^2 dr \\
&\approx \frac{1}{P} \sum_{p=1}^{P} \int_{-\infty}^{r_{t+1}} f(r|\tilde{\sigma}_t^2(p), \tilde{J}_{t+1}(p))dr \\
&= \frac{1}{P} \sum_{p=1}^{P} \int_{-\infty}^{r_{t+1}} I_{\tilde{J}_{t+1}(p)=1} \phi(r|\mu+\mu_r, \tilde{\sigma}_t^2(p)+\sigma_r^2) + I_{\tilde{J}_{t+1}(p)=0} \phi(r|\mu, \tilde{\sigma}_t^2(p))dr
\end{aligned}
\tag{40}
$$

Then calculate the log-likelihood as follows:

$$
\begin{aligned}
\mathcal{L}_{t+1} &= \log f(r_{t+1}|r_t) \\
&\approx \log[\frac{1}{P} \sum_{p=1}^{P} I_{\tilde{J}_{t+1}(p)=1} \phi(r_{t+1}|\mu+\mu_r, \tilde{\sigma}_t^2(p)+\sigma_r^2) + I_{\tilde{J}_{t+1}(p)=0} \phi(r_{t+1}|\mu, \tilde{\sigma}_t^2(p))]
\end{aligned}
\tag{41}
$$

Lastly discard $\{\tilde{V}_t(p), \tilde{J}_{t+1}(p)\}_{p=1}^P$

9. **Iteration**: Go back to step 1, and start to update the particles from $\{L_t(p)\}_{p=1}^P$ to $\{L_{t+1}(p)\}_{p=1}^P$ until $t = T-1$.

This method can be easily modified for the other diffusion type models, described in Section 3.1.3.

6.4 HL test

For the HL test by Hong and Li (2005), the kernel estimator of the joint densities for lag order j is given by:

$$
\hat{g}(u_1, u_2) = \frac{1}{T-j} \sum_{t=j+1}^{T} K_h(u_1, z_t) K_h(u_2, z_{t-j}) \tag{42}
$$

where z_t is evaluated at any $\sqrt{T}$-consistent estimator for the true model parameter and $K_h(u, z_t)$ for $x \in [0,1]$ is given by

$$
K_h(x,y) = \begin{cases} \frac{k(\frac{x-y}{h})}{h\int_{-x/h}^{1} k(u)du}, & \text{if } x \in [0,h) \\ k(\frac{x-y}{h}), & \text{if } x \in [h, 1-h] \\ \frac{k(\frac{x-y}{h})}{h\int_{-1}^{(1-x)/h} k(u)du}, & \text{if } x \in (1-h, 1) \end{cases} \tag{43}
$$

with $k()$ a pre-specified symmetric probability density, for example given by $k(u) = \frac{15}{16}(1-u^2)^2 I_{|u|\le 1}$ and $h = \hat{S}_z T^{-\frac{1}{6}}$ with $\hat{S}_z$ the sample standard deviation of $\{z_t\}$ as in Yun (2018).

The HL statistic is based on a properly standardized version of the quadratic form between

$\hat{g}_j(u1, u2)$ and 1:

$$\hat{Q}(j) = \frac{(T-j)h \int_0^1 \int_0^1 (\hat{g}_j(u_1, u_2) - 1)^2 du_1 du_2 - hA_h^0}{\sqrt{V^0}} \tag{44}$$

with

$$A_h^0 = [(h^{-1} - 2) \int_{-1}^1 k^2(u)du + 2 \int_0^1 \int_{-1}^1 k_b^2(u)dudb]^2 - 1 \tag{45}$$

$$V^0 = 2[\int_{-1}^1 [\int_{-1}^1 k(u+v)k(v)dv]^2 du]^2 \tag{46}$$

with $k_b() = \frac{k()}{\int_{-1}^b k(v)dv}$. Such that under correct model specification, $\hat{Q}(j) \to N(0, 1)\ \forall j$.

Hong et al. (2007) suggest a Portmanteau type statistic to combine different lag orders into a single test-statistic, given by:

$$HL(\rho) = \frac{1}{\sqrt{\rho}} \sum_{j=1}^{\rho} \hat{Q}(j) \tag{47}$$

where ρ is a chosen lag truncation order, in this paper I work with $\rho = 5, 10$ and 20. This statistic also asymptotically follows a $N(0, 1)$ distribution. However, Hong et al. (2007) have shown that using the asymptotic test statistic over rejects the null hypothesis. Instead they suggest simulation to obtain critical values. I use the simulated critical values obtained by Yun (2018) in this paper.

6.5 Christoffersen test

Christoffersen (1998) developed a test based on VaR violations. First the test for unconditional coverage, which should be equal to the chosen probability level p. Let T_1 denote the amount of VaR violations, i.e. $T_1 = \sum_{t=1}^T I_{r_t < \text{VaR}_t(p)}$ and $T_0 = T - T_1$ the amount of non-violations. The null hypothesis is tested that the fraction of VaR violations $\pi = \frac{T_1}{T_0+T_1}$ is equal to the chosen probability level p for the VaR forecasts. Under this null hypothesis, the test statistic is given by:

$$LR_{uc} = -2(\log(p^{T_1}(1-p)^{T_0}) - \log(\pi^{T_1}(1-\pi)^{T_0})) \sim X^2(1) \tag{48}$$

Next the VaR violations should be independent. This comes down to testing the null hypothesis of independence against a first-order Markov chain. Let T_{ij} denote the number of transitions from state $i = 0, 1$ to $j = 0, 1$, where 0 indicates non-violation and 1 indicates violation. For example $T_{00} = \sum_{t=2}^T I_{r_{t-1} \geq \text{VaR}_{t-1}(p)} I_{r_t \geq \text{VaR}_t(p)}$. Under the null hypothesis the transition probability is given by $\pi = \frac{T_{01}+T_{11}}{T_{00}+T_{01}+T_{10}+T_{11}}$. Under the alternative hypothesis the transition matrix is given by $\Pi = \begin{pmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{pmatrix}$ with $\pi_{01} = \frac{T_{01}}{T_{00}+T_{01}}$ and $\pi_{11} = \frac{T_{11}}{T_{10}+T_{11}}$. The test statistic is given by:

$$LR_{id} = -2(\log(\pi^{(T_{01}+T_{11})}(1-\pi)^{(T_{00}+T_{10})}) - \log((1-\pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}})) \sim X^2(1) \tag{49}$$

Finally these test can be combined into a joint test statistic for conditional coverage, the test statistic is given by:

$$LR_{cc} = LR_{uc} + LR_{id} \sim X^2(2) \tag{50}$$

6.6 DQ test

Engle and Manganelli (2004) proposed a joint test for unconditional coverage and independence as well. Define the adjusted hit rate: $\tilde{hit}_t = I_{r_t<\text{VaR}_t(p)} - p$, with p the chosen probability level. Now consider the linear regression:

$$\tilde{hit}_t = \delta_0 + \delta_1 \tilde{hit}_{t-1} + ... + \delta_l \tilde{hit}_{t-l} + u_t \tag{51}$$

with l a chosen lag length (I use $l = 4$ like Yun (2018)). Unconditional coverage would imply $\delta_0 = 0$ and independence would imply $\delta_i = 0$ for $i = 1, ..., l$. So testing conditional coverage is similar to testing whether all estimated coefficients are jointly equal to 0. Hence, the test statistic is given by:

$$DQ = \frac{\hat{\delta}'_{OLS}(X'X)^{-1}\hat{\delta}_{OLS}}{p(1-p)} \sim X^2(l+1) \tag{52}$$

where $\hat{\delta}_{OLS}$ is a vector containing the coefficients estimated by OLS, and X is a matrix containing all regressors.

6.7 Description of code

Code has been uploaded in the file Thesis code.zip. The code consists of the following folders, with the program name indicating which model the file is for.

- Simulation: Here the size performance of the different tests is measured using simulation
- Estimation JAGS: Here all the stochastic volatility models are estimated using Bayesian MCMC (the GARCH type models are directly estimated in the empirical analysis)
- Empirical: Here all the empirical analysis is done, the dynamic parameter estimation script is included in this folder as well, named GJRST - dynamic.R
- Options: Here option pricing is done

If data is used or figures are generated, these are also included in the folder. Most programs should be able to run, once the relevant packages are installed. However the estimation requires a correctly installed version of JAGS. Also, there are 2 Eviews files for the Diebold-Mariano test, as this was simpler to implement in Eviews. These of course require Eviews to be installed, and can be used by opening the corresponding file of LikelihoodsDMtest.xlsx or correctedmsemapeDMtest.xlsx.

Four important things to note:

- Notation used in the code is not completely the same as in the paper, as I changed notation in the paper for the sake of clarity
- I did not use fixed seeds, so running the code again gives slightly different results, although results should of course be similar

- Comments are a bit scarce, due to the amount of coding that was necessary
- I did not use a master script to run the different models, instead there is a different R script for every model