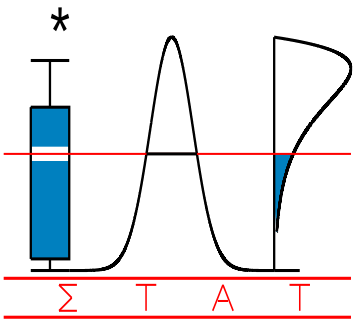
T E C H N I C A L

R E P O R T

0648

DECIDING BETWEEN GARCH AND STOCHASTIC VOLATILITY VIA STRONG DECISION RULES

PREMINGER A., and C. HAFNER



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Deciding between GARCH and stochastic volatility via strong decision rules

Arie Preminger[1] Christian M. Hafner[2]

June 2006

Abstract

The GARCH and stochastic volatility (SV) models are two competing, well-known and often used models to explain the volatility of financial series. In this paper, we consider a closed form estimator for a stochastic volatility model and derive its asymptotic properties. We confirm our theoretical results by a simulation study. In addition, we propose a set of simple, strongly consistent decision rules to compare the ability of the GARCH and the SV model to fit the characteristic features observed in high frequency financial data such as high kurtosis and slowly decaying autocorrelation function of the squared observations. These rules are based on a number of moment conditions that is allowed to increase with sample size. We show that our selection procedure leads to choosing the best and simple model with probability one as the sample size increases. The finite sample size behaviour of our procedure is analyzed via simulations. Finally, we provide an application to stocks in the Dow Jones industrial average index.

Keywords: GARCH, stochastic volatility, model selection
JEL Classification: C13, C22, C53

[1]CORE, Université catholique de Louvain, Voie du Roman Pays 20, B-1348, Louvain-la-Neuve, Belgium preminger@core.ucl.ac.be

[2]Institut de statistique and CORE, Université catholique de Louvain, Voie du Roman Pays 20, B-1348, Louvain-la-Neuve, Belgium hafner@stat.ucl.ac.be

1 Introduction

During the last decade there has been an increasing interest in modeling the volatility of high frequency financial data using GARCH type models introduced by Bollerslev (1986), and stochastic volatility (SV) models, see Taylor (1986), among others. For surveys on GARCH models we refer to Bollerslev et al. (1994). Some detailed reviews of SV models are given by Shephard (1996), Ghysels et al. (1996) and Broto and Ruiz (2004). Note that for these models the most widely studied specification in this literature consists of GARCH (1,1) and an SV model of order one with Gaussian log-volatility and zero (or constant) mean. Basically the SV model assumes two error processes, while the GARCH model allows for a single error term. This implies that the SV models can be more flexible than GARCH model in fitting the data. Carnero et al. (2004) find that in the GARCH model, the parameters explaining persistence and kurtosis are closely linked whereas these features can be modelled independently in SV models, so the latter can better represent the empirical regularities often observed in financial time series, see also Danielsson (1994) and Kim et al. (1998). However the GARCH model is often preferred in empirical applications. This is mainly due to the problems which arise as a consequence of the intractability of the likelihood function of the SV model which prohibits its direct evaluation.

Financial time series are mainly characterized by high persistence in the autocorrelation of squared observations and leptokurtosis, see e.g. Taylor (1986). Therefore, the ability of GARCH and SV models to reproduce such stylized facts is a desirable feature and failure to do so is most often a criterion to dismiss one specification. The ability of these models to accommodate and explain the empirical properties observed in real time series have been analyzed separately by Teräsvirta (1996) and Liesenfeld and Jung (2000). Kim et al. (1998) tested the models using a simulated likelihood ratio test. Bai et al. (2003) and Carnero et al. (2004) compared the SV and the GARCH models using the kurtosis autocorrelation relationship as their benchmark. Gerlach and Tuyl (2006) compare the empirical fit of both models using Markov Chain Monte Carlo (MCMC) methods. However, to the best of our knowledge no formal selection procedure between the GARCH and the SV model based on their ability to represent the observed properties of the data has been suggested. It should be noted that the models are non-nested, and this can complicate model comparison. Furthermore, testing two models against each other does not necessarily lead to a unique choice of a model, since both models may be rejected or accepted against each other, see Granger el al. (1995).

In this paper a set of strongly consistent decision rules for selecting between the models

is suggested. More specifically, we assign to each model a goodness of fit statistic which measures the ability to fit interesting features or moments of the data such as persistence, variance or kurtosis. The difference between the statistics of the GARCH and the SV model is denoted by $\Delta\hat{Q}_n$ and its population counterpart is defined by Δ_n. We can decompose $\Delta\hat{Q}_n$ as $\Delta_n + (\Delta\hat{Q}_n - \Delta_n)$, so, information about the best model is given by the sign of Δ_n, whereas the component $\Delta\hat{Q}_n - \Delta_n$ involves a noise that disturbs this information. We may therefore regard Δ_n as the signal and $\Delta\hat{Q}_n - \Delta_n$ as the noise of $\Delta\hat{Q}_n$. We would like to select the model that best fits the data, regardless of whether or not the models considered are correctly specified. If Δ_n converges to zero, we will prefer the GARCH model because of its relative simplicity. The reason is that volatility in the GARCH model is a measurable function of observables, whereas it is latent in the SV model. Thus, even with closed form estimators for the parameters of the SV model, additional effort has to be spent in practice to estimate the volatility. Therefore, we suggest the following decision rule: choose the SV model if $SC_n > 0$, otherwise select the GARCH model, where $SC_n = \sqrt{n}\Delta\hat{Q}_n - c_n$ and $c_n > 0$ is a penalty term which encourages the selection of the GARCH model when both models are equivalent. We show that under mild regularity conditions the use of the decision rule leads to selection of the best or simplest model with probability one (strong consistency of selection) as the sample size increases. The strong consistency implies that these rules have type I and type II errors approaching zero asymptotically, or equivalently, the size approaches zero and the power approaches one.

The consistency property of the selection procedure is obtained by providing an almost sure bound for $\Delta\hat{Q}_n$ via the law of iterated logarithm (LIL). Thus, the sequence of penalty terms should grow with the sample size at an appropriate rate. However, the choice of the penalty term can be problematic. Two researchers using the same data and models but different penalty terms can arrive at differing conclusions in finite samples. One way to avoid such difficulties is to use a strongly consistent estimator of the bound. We suggest such an estimator which is based on the Newey-West variance estimator of the noise and show its strong consistency, where in this case we strengthen slightly our regularity conditions. Decision rules which are based on almost sure bounds to some test statistics have been suggested by Stinchcombe and White (1998), Corradi (1999), Altissimo and Corradi (2002, 2003) and Preminger and Wettstein (2005) among others as an alternative to hypothesis testing. It should be noted, however, that the usage of such selection criteria is not necessarily optimal in finite samples.

In addition, given the slow decay of autocorrelation of squared data, we allow the number of moment conditions q_n to increase with the sample size. However, as q_n increases,

variation over the moment conditions accumulates and might alter the true signal coming from $\Delta\hat{Q}_n$ and the strong consistency of the bound estimate may fail. We note that similar problems occur in GMM estimation when we allow the number of orthogonality conditions to increase with the sample size (see e.g. Stock and Yogo (2005) and Han and Phillips (2006)). We shall show that if the order of q_n and the signal is controlled appropriately, the strong consistency of the selection procedure is preserved.

Unfortunately, the estimation of GARCH and especially SV models can be quite difficult. One of the most important limitations of SV models is that the exact likelihood function is difficult to evaluate. A variety of estimation methods for the SV model have been proposed, the simplest estimator being the method of moments used by, for example, Taylor (1986), Chesney and Scott (1989) and Andersen and Sorensen (1996). Melino and Turnbull (1990) used the generalized method of moments (GMM). Duffie and Singleton (1993) proposed the simulated method of moments (SMM). However, these procedures have poor finite sample properties and their efficiency is suboptimal with respect to maximum likelihood (ML) type estimation methods. Jacquier et al. (1994) find substantial bias for these estimates and show that the performance of the technique worsens when there is high persistence and a low coefficient of variation. Furthermore, the GMM criterion surface for the SV model is highly irregular. Therefore, optimization fails to converge, especially for small sample sizes. Another literature focuses on likelihood-based estimation using importance sampling or MCMC methods, either in a Bayesian setting (see e.g., Jacquier et al. (1994) and Wong (2002)) or in a classical setting (see e.g., Danielsson and Richard (1993), Danielsson 1994; Kim et al. (1998); Liesenfeld and Richard (2003), Sandmann and Koopman (1998), among others). A good overview of these methods is provided by Bauwens and Rombouts (2004). Other estimation procedures based on an auxiliary model also have been applied such as the indirect inference (see Engle and Lee (1996) and Calzolari et al. (2001)) and the efficient method of moments (EMM) methods (see Gallant et al. (1997) and Jiang and van der Sluis (2000)). In particular, all these simulation methods are not widely adopted since they are computationally intensive and rely on assumptions that are hard to check in practice, such as the convergence of simulated Markov chains to their steady state. Further, it is difficult to establish limit theorems for these estimators e.g. the LIL. The calculation of the GARCH model requires the use of numerical optimization procedures which are sensitive to initialization and the optimization procedure used, see Brooks et al. (2001).

Motivated by the difficulties associated with estimating the SV model, we provide a simple closed-form estimator for a Gaussian first order SV model and derive its asymptotic properties. The estimator is based on the second order properties of the logarithm of

the squared data. Note that for the SV models, other estimators, based on similar transformation have been suggested see e.g., Harvey et al. (1994), Ruiz (1994). Another closed form estimator has been recently proposed independently by Dufour and Valéry (2006). We show that in scenarios where the kurtosis is very large and persistence is high, our estimator is more efficient than that of Dufour and Valéry (2006). For the GARCH model we use a closed-form estimator that has recently been proposed by Kristensen and Linton (2006) which can be easily implemented and neither requires the use of any numerical optimization procedures nor relies on initial conditions.

The paper is organized as follows. In Section 2 the closed-form estimators of the GARCH and SV model are described. For the estimator of the SV model, the asymptotic properties are given. Section 3 introduces the basic set-up for our model selection procedure and establishes its strong consistency. Section 4 reports the results of a simulation to study on the performance of our procedure. Section 5 applies the procedure to stocks in Dow Jones industrial index. We conclude in section 6. All proofs are gathered in the Appendix.

2 Closed form estimators

This section considers closed form estimators both for the SV and the GARCH model. Note that these estimators are not efficient, but in the usually large samples in financial applications this should be of minor importance. As our simulation study will demonstrate, standard errors of parameter estimates are small.

2.1 The stochastic volatility model

Consider the following stochastic volatility (SV) model for the time series $\{y_t\}_{t\in\mathrm{N}}$

$$y_t = \sqrt{h_t}\xi_t \tag{1}$$

$$\log h_t = \phi + \delta \log h_{t-1} + \sigma\varepsilon_t \tag{2}$$

where ξ_t and ε_t are i.i.d. error terms with joint distribution given by $(\xi_t, \varepsilon_t)' \sim N(0, I_2)$. Due to the normality assumption, the mean and variance of $\log(\xi_t)^2$, -1.27 and $\pi^2/2$, are known.

Assumption 1 *In (2),* $|\delta| < 1$ *and* $\{h_t\}_{t\in\mathrm{N}}$ *is initialized from its stationary distribution.*

Denoting $\eta_t = \log \xi_t^2 + 1.27$ and $z_t = \log y_t^2$, we obtain the following linear state space representation.

$$z_t = -1.27 + \log h_t + \eta_t \tag{3}$$

$$\log h_t = \phi + \delta \log h_{t-1} + \sigma\varepsilon_t \tag{4}$$

The model is characterized by the parameter vector $\theta^1 = (\phi, \delta, \sigma^2)' \in \Theta_1$.

Now, we can express the autocorrelation function (ACF) of y_t^2 and the kurtosis of y_t as a function of the model parameters. Denote by $\rho(\tau, \theta^1)$,$\psi(\theta^1)$ and $\kappa(\theta^1)$ the autocorrelation function (ACF) the variance and kurtosis of y_t^2, respectively. For the Gaussian SV model it can be shown that

$$\rho(\tau, \theta^1) = \frac{\exp\left(\frac{\sigma^2}{1-\delta^2}\delta^\tau\right) - 1}{\kappa - 1}, \quad \tau \geq 1 \tag{5}$$

$$\psi(\theta^1) = \exp\left(\frac{\phi}{1-\delta} + \frac{\sigma^2}{2(1-\delta^2)}\right) \tag{6}$$

$$\kappa(\theta^1) = 3\exp\left(\frac{\sigma^2}{1-\delta^2}\right) \tag{7}$$

The closed form estimator $\hat{\theta}_n^1 = (\hat{\delta}_n, \hat{\phi}_n, \hat{\sigma}_n)'$ is given by

$$\hat{\delta}_n = \frac{\hat{\vartheta}_{1n}(2) - \hat{\vartheta}_{1n}^2}{\hat{\vartheta}_{1n}(1) - \hat{\vartheta}_{1n}^2} \tag{8}$$

$$\hat{\phi}_n = (\hat{\vartheta}_{1n} - c)(1 - \hat{\delta}_n) \tag{9}$$

$$\hat{\sigma}_n^2 = \left(\hat{\vartheta}_{2n} - \hat{\vartheta}_{1n}^2 - \frac{\pi^2}{2}\right)(1 - \hat{\delta}_n^2) \tag{10}$$

where, $c = -1.27$, $\hat{\vartheta}_{1n} = n^{-1}\sum_{t=1}^n z_t$, $\hat{\vartheta}_{2n} = n^{-1}\sum_{t=1}^n z_t^2$ and $\hat{\vartheta}_{1n}(k) = n^{-1}\sum_{t=1}^n z_t z_{t-k}$. Let $\hat{\lambda}_n^1 \equiv (\hat{\vartheta}_{1n}, \hat{\vartheta}_{1n}(1), \hat{\vartheta}_{1n}(2), \hat{\vartheta}_{2n})$, $\lambda_0^1 = E(\hat{\lambda}_n^1)$ and $\tilde{V} = \lim_{n\to\infty} var\sqrt{n}(\hat{\lambda}_n^1 - \lambda_0^1)$. From (8)-(10) we can define a continuously differentiable function $F_1 : \Re^4 \to \Re^3$ such that $\hat{\theta}_n^1 = F_1(\hat{\lambda}_n^1)$ and $\theta_0^1 = F_1(\lambda_0^1)$. After some tedious algebra and by applying the continuous mapping theorem, we obtain the following result.

Theorem 1 *Under Assumption 1,*

$$\hat{\theta}_n^1 \to_{a.s.} \theta_0^1 \quad \textit{and} \quad \sqrt{n}(\hat{\theta}_n^1 - \theta_0^1) \to_d N(0, V_1),$$

where $V_1 = \left(\frac{\partial F_1(\lambda_1^0)}{\partial \lambda_1}\right)\tilde{V}\left(\frac{\partial F_1(\lambda_1^0)}{\partial \lambda_1}\right)'$.

Note that this estimator converges to the true parameter values when the SV model is the true process. Otherwise, $\hat{\theta}_n^1$ is the "quasi-true" parameter estimate of an SV model, which is used to approximate some of the data moments .

The closed form for the SV model which was proposed by Dufour and Valéry (2006) is given as follows.

$$\tilde{\delta}_n \;=\; \frac{\log\hat{\mu}_{2n}(1) - \log 3 - 4\log\hat{\mu}_{2n} + \log\hat{\mu}_{4n}}{\log\hat{\mu}_{4n} - \log 3 - 2\log\hat{\mu}_{2n}} - 1$$

$$\begin{aligned}\tilde{\phi}_n &= \left(\frac{1}{2}\log 3 + 2\log\hat{\mu}_{2n} - \frac{1}{2}\log\hat{\mu}_{4n}\right)(1-\tilde{\delta}_n)\\ \tilde{\sigma}_n^2 &= \left(\log\hat{\mu}_{4n} - \log 3 - 2\log\hat{\mu}_{2n}\right)(1-\tilde{\delta}_n^2)\end{aligned}$$

where $\hat{\mu}_{2n} = n^{-1}\sum_{t=1}^{n} y_t^2$, $\hat{\mu}_{4n} = n^{-1}\sum_{t=1}^{n} y_t^4$ and $\hat{\mu}_{2n}(1) = n^{-1}\sum_{t=1}^{n} y_t^2 y_{t-1}^2$.

To see how the two estimators behave we do a small Monte Carlo experiment. We follow Sandmann and Koopman (1998) and Hafner and Herwartz (2000) in specifying the parameters. Defining the coefficient of variation, $CV = var(h_t)/E[h_t]^2$, one obtains the expression $CV = \exp(\sigma^2/(1-\delta^2)) - 1$. The coefficient of variation for this model is directly related to the kurtosis of y_t, which is given by $\kappa = 3(CV+1)$. Here, ϕ is an irrelevant scaling parameter, but Sandmann and Koopman (1998) determine ϕ such that $E[h_t] = 0.0009$, which gives a realistic annualized standard deviation of 22%. Now, we fix δ at 0.90, 0.95, and 0.98 and CVat 10 and 1. The corresponding kurtosis coefficients are 33 and 6, respectively, where the former could represent a very high frequency (e.g. hourly samples) and the latter a lower frequency (e.g. daily or weekly) financial time series. This gives 6 different parameterizations. As sample sizes we use $n = 2,500$, and 5,000. Each process is simulated $k = 10,000$ times. Table 1 reports the means of estimates of δ across the k simulations, together with their standard deviation divided by $\sqrt{k}$. To economize on space we do not report the results for the other parameters, but they are very similar. The results suggest that $\hat{\delta}$ is more efficient than $\tilde{\delta}$ when the kurtosis κ is large, but worse if κ is small.

2.2 The GARCH model

The second volatility model we consider is the standard GARCH (1,1) model as proposed by Bollerslev (1986). It reads

$$y_t = \sqrt{h_t}\xi_t \tag{11}$$

$$h_t = \omega + \alpha y_{t-1}^2 + \beta h_{t-1} \tag{12}$$

where $\xi_t \sim i..i.N(0,1)$ and the model parameters are $\theta^2 = (\omega, \alpha, \beta)' \in \Theta_2$. The main difference to SV models is that here, volatility h_t is measurable w.r.t. the information set at time $t-1$. To ensure finiteness of fourth moments we shall throughout assume that $3\alpha^2 + 2\alpha\beta + \beta^2 < 1$. The ACF, variance and kurtosis of y_t are given as follows

$$\rho(1,\theta^2) = \frac{\alpha(1-\beta(\alpha+\beta))}{1-\beta^2-2\alpha\beta} \tag{13}$$

$$\rho(\tau,\theta^2) = (\alpha+\beta)\rho(\tau-1), \quad \tau \geq 2 \tag{14}$$

$$\psi(\theta^2) = \omega/(1-\alpha-\beta) \tag{15}$$

$n = 2,500$					
κ	δ	$\hat{\delta}$	$SE(\hat{\delta})$	$\tilde{\delta}$	$SE(\tilde{\delta})$
33	0.9	0.9001	(0.0007)	0.8962	(0.0016)
33	0.95	0.9502	(0.0006)	0.9467	(0.0015)
33	0.98	0.9816	(0.0006)	0.9776	(0.0014)
6	0.9	0.9268	(0.0041)	0.9113	(0.0019)
6	0.95	0.9819	(0.0029)	0.9603	(0.0020)
6	0.98	1.0206	(0.0037)	0.9926	(0.0020)
$n = 5,000$					
κ	δ	$\hat{\delta}$	$SE(\hat{\delta})$	$\tilde{\delta}$	$SE(\tilde{\delta})$
33	0.9	0.8988	(0.0004)	0.8994	(0.0013)
33	0.95	0.9501	(0.0004)	0.947	(0.0013)
33	0.98	0.9809	(0.0004)	0.979	(0.0012)
6	0.9	0.907	(0.0016)	0.9048	(0.0015)
6	0.95	0.9627	(0.0016)	0.9527	(0.0015)
6	0.98	0.9927	(0.0017)	0.983	(0.0014)

Table 1: *Monte Carlo estimation results of the SV model based on 10,000 replications, using the Dufour-Valéry estimator $\tilde{\delta}$ and our estimator $\hat{\delta}$.*

$$\kappa(\theta^2) = 3 + \frac{6\alpha^2}{1 - 3\alpha^2 - 2\alpha\beta - \beta^2} \tag{16}$$

Estimation of GARCH models is usually done by maximum likelihood, which is much less complicated than for the SV model. Recently, a closed form estimator for the GARCH model based on the ACF of y_t^2 has been proposed by Kristensen and Linton (2006), which is strongly consistent but less efficient than the ML estimator.

Let $\hat{\rho}_n(k)$and $\hat{\psi}_n$ be the sample autocorrelation function and average of y_t^2 (the sample variance) respectively, $\hat{\rho}_n(k) = \hat{v}_n(k)/\hat{v}_n(0)$ with

$$\hat{v}_n(k) = \frac{1}{n}\sum_{t=1}^{n}(y_t^2 - \hat{\psi}_n)(y_{t-k}^2 - \hat{\psi}_n), \tag{17}$$

$$\hat{\psi}_n = \frac{1}{n}\sum_{t=1}^{n} y_t^2. \tag{18}$$

Given (17)-(18) let $\hat{\zeta}_n = \hat{v}_n(2)/\hat{v}_n(1)$, $\hat{b}_n \equiv \frac{\hat{\zeta}_n + 1 - 2\hat{v}_n(1)\hat{\zeta}_n}{\hat{\zeta}_n - \hat{v}_n(1)}$ and $\hat{\theta}_n = \frac{-\hat{b}_n + \sqrt{\hat{b}_n^2 - 4}}{2}$ assuming that $\hat{b}_n > 2$. The closed form estimator $\hat{\theta}_n^2 = (\hat{\omega}_n, \hat{\alpha}_n, \hat{\beta}_n)'$ is given by

$$\hat{\alpha}_n = \hat{\theta}_n + \hat{\zeta}_n, \quad \hat{\beta}_n = -\hat{\theta}_n, \quad \hat{\omega}_n = \hat{\psi}_n(1 - \hat{\zeta}_n). \tag{19}$$

The parameter $\hat{\zeta}_n$ is constrained within the unit interval. However in practice it may occur that $\hat{\zeta}_n < 0$ or $\hat{\zeta}_n > 1$. To deal with this problem, we set $\hat{\zeta}_n = \max(\varepsilon, \min(1-\varepsilon, \hat{\zeta}_n))$ using a small $\varepsilon > 0$. Furthermore, the parameter can be estimated by $\hat{\zeta}_n = \sum_{j=1}^{m} \omega_j \frac{\hat{v}_n(j+1)}{\hat{v}_n(j)}$ for any ω_j sequence with $\sum_{j=1}^{m} \omega_j = 1$, so that a more general class of estimators can be defined. Kristensen and Linton (2006) show that if h_t is strictly stationary and $E[(\beta + \alpha\xi_t^2)^4] < 1$, these estimators are $\sqrt{n}$-consistent and are asymptotically normal. To obtain this result, the finiteness of eight moments of innovation ξ_t is needed, while a slower rate is obtained if only finite fourth moment is assumed. Since the closed form estimator is not efficient, it was suggested by Kristensen and Linton (2006) to use these estimates as a starting point in the numerical optimization of the QMLE or the Whittle estimation procedure.

3 Model Selection

We consider selection between the SV (model 1) model and the GARCH (model 2). Let $r_i(\theta^i) \equiv (\psi(\theta^i), \kappa(\theta^i), \rho(1, \theta^i), \ldots, \rho(q_n, \theta^i))$ be variance, kurtosis and the autocorrelations of y_t^2 which are implied by the i-th model and let $\hat{\gamma}_n \equiv (\hat{\psi}_n, \hat{\kappa}_n, \hat{\rho}_n(1), \ldots, \hat{\rho}_n(q_n))$ be the corresponding vector of empirical moments and $\gamma_n \equiv (\psi_0, \kappa_0, \rho(1), \ldots, \rho(q_n))$ be its

population counterpart. The kurtosis is estimated by $\hat{\kappa}_n \equiv \frac{1}{n}\sum_{t=1}^{n} y_t^4/\hat{\psi}_n^2$ and $v_0(k) = E(\hat{v}_n(k))$. The distance between γ_0 and $r_i(\theta_0^i)$ is given by

$$Q_{in} = (r_i(\theta_0^i) - \gamma_0)' W (r_i(\theta_0^i) - \gamma_0), \quad i = 1, 2 \tag{20}$$

where $W = diag(w_1, \ldots, w_{q_n+2})$ and $W \geq 0$ (where the inequality means that it holds for each of the elements). This distance measures the ability of the model to fit well some interesting features of the data such as persistence, variance or kurtosis. The weights in the matrix W reflect the importance that the researcher assigns to the moments used. For example $w_j = 0$ for some j suggest that some moments are negligible. Also we can set $w_j = 1/j$ which indicate that more weights are given to the variance, kurtosis and the first elements of the ACF and so on. The number of weights\moment conditions $q_n + 2$ is non-decreasing in n. We throughout assume that the Q_{in}'s are finite, thus the models are *comparable* (see Rivers and Vuong (2002)).

Our goal is to select the model that fits the data best according to the goodness of fit measure given in equation (20). In case that both models fit the data equally well, we will prefer the GARCH model because of its relative simplicity. Since we provide an asymptotic theory below, we now introduce the precise notion of better and equal fits accordingly. Let $\Delta_n = Q_{2n} - Q_{1n}$, if $\liminf_{n\to\infty} \Delta_n > 0$, we say that the SV model is *asymptotically better* than the GARCH model and vice versa when $\limsup_{n\to\infty} \Delta_n < 0$. If instead $\lim_{n\to\infty} \Delta_n = 0$, we say that the two models are *asymptotically equivalent.* When $\Delta_n = 0$ for each sample size, the two models are called *equivalent.* For simplicity and without loss of generality we assume that the series is strictly stationary. If the number of moments does not depend on the sample size ($q_n = q$), we have that $\Delta_n = \Delta$ and the notion of the asymptotic equivalence reduces to the equivalence.

We construct the following selection criteria (SC).

$$SC_n = \sqrt{n}\Delta\hat{Q}_n - c_n \tag{21}$$

where $\Delta\hat{Q}_n = \hat{Q}_{2n} - \hat{Q}_{1n}$, $\hat{Q}_{in} = (r_i(\hat{\theta}_n^i) - \hat{\gamma}_n)' W (r_i(\hat{\theta}_n^i) - \hat{\gamma}_n)$ for $i = 1, 2$ and $c_n > 0$ imposes a penalty which encourages the selection of the GARCH model when both models are asymptotically equivalent. The penalty terms can be a sequence of non-stochastic numbers. The selection procedure works as follows: select the GARCH model if $SC_n \leq 0$, otherwise select the SV model.

Next, we need to introduce some notation. Let $\hat{\lambda}_n^2 \equiv (\hat{\psi}_n, \hat{\psi}_n^2\hat{\kappa}_n, \hat{v}_n(1), \ldots, \hat{v}_n(q_n))$, $\hat{\tau}_n \equiv diag(\,1,\, 1, 1, 1, 1,\, \hat{\psi}_n^{-2}, \hat{v}_n(0)^{-1}, \ldots \hat{v}_n(0)^{-1})$, $\hat{\lambda}_n \equiv (\hat{\lambda}_n^1, \hat{\lambda}_n^2)$ and λ_0^2, λ_0,τ_0 be their population counterparts, respectively. Given (19), we can define a continuously differentiable function $F_2 : \Re^3 \to \Re^3$ such that $\hat{\theta}_n^2 = F_2(\hat{\gamma}_n^1, \hat{\gamma}_n^3, \hat{\gamma}_n^4)$, and $\theta_0^2 = F_2(\gamma_0^1, \gamma_0^3, \gamma_0^4)$, where $\hat{\gamma}_n^j$ and γ_0^j

denote the j-th element of $\hat{\gamma}_n$ and γ_0, respectively. Let $F_3 : \Re^{q_n+2} \to \Re^{q_n+2}$ be the identity function (that is, $F_3(\gamma) \equiv \gamma$). We can now define the function $F : \Re^{q_n+6} \to \Re^{q_n+8}$, where $F(Z) \equiv [F_1(\cdot)',\ F_2(\cdot)',\ F_3(\cdot)']\ '$ and $Z \equiv (\lambda_0^{1'},\ \gamma')'$.

If $q_n = q < \infty$, using a first order Taylor expansion of (20) we have that

$$\hat{Q}_{2n} - \hat{Q}_{1n} = \Delta + 2A(\bar{\theta}_n^1, \bar{\theta}_n^2, \bar{\gamma}_n) \left[\begin{array}{c} \hat{\theta}_n^1 - \theta_0^1 \\ \hat{\theta}_n^2 - \theta_0^2 \\ \hat{\gamma}_n - \gamma_0 \end{array} \right]_{(q_n+8)\times 1} \tag{22}$$

where $\bar{\theta}_n^1, \bar{\theta}_n^2, \bar{\gamma}_n$ lie on the chord between $\hat{\theta}_n^1, \hat{\theta}_n^2, \hat{\gamma}_n$ and $\theta_0^1, \theta_0^2, \gamma_0$ and

$$A(\theta^1, \theta^2, \gamma) \equiv \left[(\gamma - r_1(\theta^1))' W \frac{\partial r_1}{\partial \theta^{1'}}(\theta^1),\ (r_2(\theta^2) - \gamma)' W \frac{\partial r_2}{\partial \theta^{2'}}(\theta^2),\ (r_1(\theta^1) - r_2(\theta^2))' W \right]$$

Using the smoothness of $F(\cdot)$, under mild regularity conditions, we further have

$$\Delta\hat{Q}_n = \Delta + B_0 \cdot (\hat{\lambda}_n - \lambda_0) + o_{a.s.}(1) = \Delta + o_{a.s.}(1) \tag{23}$$

where, $B_0 = 2A(\theta_0^1, \theta_0^2, \gamma_0)\ \frac{\partial F}{\partial Z}\big|_{Z=(\lambda_0^1, \gamma_0)}\ \tau_0$ and

$$\frac{\partial F}{\partial Z} = \left[\begin{array}{ccccccc} \partial F_1/\partial \lambda^{1'} & 0_{3\times 1} & 0_{3\times 1} & 0_{3\times 1} & 0_{3\times 1} & \dots & 0_{3\times 1} \\ 0_{3\times 4} & \partial F_2/\partial \gamma^1 & 0_{3\times 1} & \partial F_2/\partial \gamma^3 & \partial F_2/\partial \gamma^4 & \dots & 0_{3\times 1} \\ 0_{1\times 4} & & & & & & \\ \vdots & & & I(q_n+2) & & & \\ 0_{1\times 4} & & & & & & \end{array} \right]$$

and where ($I(q_n + 2)$ is $(q_n + 2) \times (q_n + 2)$ identity matrix).

Therefore, if $\Delta > 0$, and $c_n = o_{a.s.}(\sqrt{n})$, then the SV model will be selected almost surely. If the models are equivalent, under more regularity conditions which allow us to apply the law of iterated logarithm (LIL) we have that

$$\sqrt{n}\Delta\hat{Q}_n \leq a_n \sqrt{2B_0' V_0 B_0} + o_{a.s.}(1) \tag{24}$$

where $a_n \equiv \sqrt{\log\log n}$ and $V_0 = \mathrm{Var}\sqrt{n}(\hat{\lambda}_n - \lambda_0)$, so in order to ensure that the GARCH model will be selected it is sufficient to impose that $c_n \geq a_n \sqrt{2\hat{B}_n' \hat{V}_n \hat{B}_n}$, where $\hat{B}_n$ is a strongly consistent estimator of B_0 and is given by $\hat{B}_n = 2A(\hat{\theta}_n^1, \hat{\theta}_n^2, \hat{\gamma}_n)\ \frac{\partial F}{\partial Z}\big|_{Z=(\hat{\lambda}_n^1, \hat{\gamma}_n)}\ \hat{\tau}_n$. Furthermore, $\hat{V}_n$ is a strongly consistent estimate of V_0 which is given below.

Let x_t denote the summands of $\sqrt{n}(\hat{\lambda}_n - \lambda_0)$ assuming that $\hat{\psi}_n$ is replaced by ψ_0 in $\hat{\lambda}_n$, that is, $x_t = x_t(\lambda_0, \psi_0)$, and we seek to estimate $V_0 = E(x_t x_t') + 2\sum_{s=1}^{\infty}(1 - \frac{s}{n})E(x_t x_s')$.

However, in practice we will have to use $\hat{x}_t = \hat{x}_t(\hat{\lambda}_n, \hat{\psi}_n)$, in which λ_0 and ψ_0 are replaced by $\hat{\lambda}_n$ and $\hat{\psi}_n$. We apply the Newey and West (1987) statistic,

$$\hat{V}_n = \frac{1}{n}\sum_{t=1}^{n} \hat{x}_t \hat{x}_t' + \frac{2}{n}\sum_{\tau=1}^{\ell_n} \varpi(\tau) \sum_{t=i+1}^{n} \hat{x}_t \hat{x}_{t-\tau}{}' \tag{25}$$

with $\varpi(\tau) = 1 - \frac{\tau}{1+\ell_n}$ and ℓ_n is non decreasing in the sample size n and $\ell_n \to \infty$ as the sample size grows. Let $\hat{V}_n^i$ and V_0^i be the true and the estimated variance of the i-element of $\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t$. In order to formally establish the strong consistency of our selection criteria when $q_n = q < \infty$, we will need the following assumption.

Assumption 2.

1. *The observed process is strictly stationary, ergodic and strong mixing with mixing coefficients $\alpha(m)$ satisfying that $\sum_{m=0}^{\infty} \alpha(m)^{2(1/r-1/p)} < \infty$, where $r \in (2, 4]$, $p \geq r$ and $|\Delta_n| < \infty$.*

2. *$E(|y_t^{2p}|) < \infty$ for i) $p \geq 2$ ii) $p \geq 4$ iii) $E(|y_t^{4p}|) < \infty$ for $p \geq r > 2$*

3. *$\ell_n = O(n^{1/2-1/r}(\log n)^{-1/r-\eta})$ for some $\eta > 0$.*

Assumptions 2.1 and 2.2 are standard memory and moments conditions. In 2.1 the positive constants $\alpha(m)$ specify the decay rates for the mixing coefficients; the condition is implied if $\alpha(m) = -rp/2(p-r)$ and it is automatically satisfied if the mixing coefficients decay at geometric rates. This assumption implies that the ACF sequence is absolutely summable and under sufficient restrictions on the parameter sets Θ_1 and Θ_2 (see section 2-3) we can show the same for the ACF in (5) and (14). Hence, the models are comparable and $|\Delta_n| < \infty$. Using Carrasco and Chen (2002) results for GARCH and SV we can derive primitive conditions underlying Assumptions 2.1-2.2. For example, for the lognormal stochastic volatility, which is widely used among the stochastic volatility models, Assumption 1 above implies that the process is β-mixing with geometric decay and that the moment requirements in Assumption 2.2 are satisfied. The existence of fourth moments of y_t is sufficient to show that the best model will be selected with probability one by the law of large numbers (LLN). However, when the models are equivalent we need the eighth moment of y_t in order to establish (24) via the LIL. To estimate the almost sure sharper bound we need to show that $\hat{V}_n \to_{a.s.} V_0$, hence, we strengthen slightly the moment requirements and add Assumption 2.3. In this assumption, we note the usual trade-off between the sufficient moment restrictions and the rate of divergence for the bandwidth parameter (see Hansen (1992)). The sequence of weights used in our Newey-West estimator is related to the Bartlett (1950) sequence. However, other choices for the

weighting function $\varpi(\tau)$ in (25) are possible (see Assumptions 1.1 and 1.4 in De Jong (2000)).

Theorem 2 *Suppose that $q_n = q < \infty$ and Assumption 2.1 holds.*

1. *If $\Delta > 0, c_n = o_{a.s.}(\sqrt{n})$ and 2.2(i) hold, then $P(\lim_{n\to\infty} SC_n > 0) = 1$.*
2. *Suppose that the models are equivalent, $c_n/a_n \to \infty$, and 2.2(ii) hold, then $P(\lim_{n\to\infty} SC_n \leq 0) = 1$.*
3. *Suppose that the models are equivalent, $c_n \geq a_n\sqrt{2\hat{B}_n'\hat{V}_n\hat{B}_n}$ 2.2(iii) and 2.3 hold, then $P(\lim_{n\to\infty} SC_n \leq 0) = 1$.*

Theorem 2.1 shows that when the penalty terms grows slower than $\sqrt{n}$ the SC_n tends to select the best model, once the sample size becomes large. For the case that the models are equivalent, the conditions imposed in Theorem 2.2 and 2.3 ensure that$\sqrt{n}\Delta\hat{Q}_n = O_{a.s.}(a_n)$, which is asymptotically dominated by the penalty term, so that the GARCH model is selected with probability one.

We note that in order to obtain the strong consistent decision rule we require the existence of the eighth moments. This assumption is relatively strong for the GARCH model. However, similar moment conditions are often required in order to derive the asymptotic properties of the GARCH estimator, see e.g. Baillie and Chung (2001), Comte and Lieberman (2003) and Storti (2006). On the other hand, for the SV model the existence of higher moments does not pose any problem. Furthermore, we should note that the models compared are more likely to be misspecified and are used just to fit some features of the data.

We note that instead of weighting the moments, our selection procedure can be decomposed into q different decision rules. For example, we can check if the SV model fits the kurtosis of the data better than the GARCH and we can then examine if the same is implied for the first order autocorrelation and so on. Since each decision rule is strongly consistent, the proposed procedure is also strongly consistent. This is equivalent to performing a joint test with q different hypotheses on the ability of the GARCH and SV to fit different moments of the data. However, in this case, standard tests cannot be applied in the usual way because repeated application of such tests yields a procedure that rejects the true null hypothesis with probability approaching one as the number of applications grows. Therefore, the critical values have to be allowed to grow with the sample size in order to control the asymptotic size as the test is repeated (see e.g. Bai (1999) and Altissimo and Corradi (2003)).

Theorem 2 implies that if the number of moment conditions employed is fixed, the noise is eliminated asymptotically by the action LLN, and the LIL can be used to bound the noise in (24) if the models are equivalent. So our decision rules are strongly consistent. However empirical evidence suggests that in financial data sample autocorrelations of the squared data decay more slowly than the exponential rate prescribed by the GARCH or the SV models. It is desirable to allow the number of moment conditions $(q_n + 2)$ to increase with the sample size. As q_n increases, the second term on the RHS of (23) may dominate Δ_n since the number of elements in $\{\hat{\lambda}_n - \lambda_0\}_{n\in\mathrm{N}}$ is allowed to grow with the sample size. Also, there exist no triangular LIL which can be used in this case; hence the LIL is used for x_t^i , $i = 1, \ldots, q_n$, changing the a.s. bounds in 3(iii) slightly. In order to establish the strong consistency of selection in this case we add the following assumptions. Hence, in order to establish the strong consistency of selection in this case we add the following assumptions.

Assumption 3.

1. $E(|y_t^{2p}|) < \infty$ *for* $p \geq r$
2. $q_n = O(n^{1/2-1/r}(\log n)^{-1/r-\eta})$ *for some* $\eta > 0$.

Theorem 3 *Suppose that Assumption 2.1 holds.*

1. *If* $\liminf_{n\to\infty} \Delta_n > 0$, $c_n = o_{a.s.}(\sqrt{n})$, *Assumptions 3.1 and 3.2 hold, then* $P(\lim_{n\to\infty} SC_n > 0) = 1$.
2. *Suppose that* $\sqrt{n}\Delta_n = O(a_n q_n), q_n = o(\sqrt{n}/a_n), c_n/(a_n q_n) \to \infty$, *and Assumption 2.2(ii) hold, then* $P(\lim_{n\to\infty} SC_n \leq 0) = 1$.
3. *Suppose that* $\sqrt{n}\Delta_n = o(a_n q_n)$, $q_n = o(\sqrt{n}/a_n)$, *and* $c_n \geq a_n \sum_{i=1}^{q_n} \sqrt{\hat{V}_n^i}|\hat{B}_n^i|$, *where* $\hat{B}_n^i$ *is the i-th element of* $\hat{B}_n$. *If Assumptions 2.2(iii) and 2.3 hold, then* $P(\lim_{n\to\infty} SC_n \leq 0) = 1$.

The theorem is based on providing mixingale-type bounds for the products $x_t^i x_{t-j}^i$ and $y_t y_{t-j}$ and applying results of De Jong (2000). We see that when the SV model is asymptotically better than the GARCH model, the growth rate of q_n is restricted. In Theorem 3.2 and 3.3, the models are asymptotically equivalent as long as the order of q_n is carefully determined.

The penalty term can be interpreted as a critical value, i.e. the associated size of the implicit test that the GARCH and the SV fit the data equally well under the null

hypothesis. Furthermore, from (24), given mild regularity conditions which allow us to use the LIL, we see that under the null hypothesis $\limsup_{n\to\infty} d_n(\hat{B}_n'\hat{V}_n\hat{B}_n)^{-0.5}\Delta\hat{Q}_n \leq 1$ a.s., where $d_n = a_n\sqrt{0.5n}$. Note that in the standard framework of hypothesis testing, d_n is replaced by $\sqrt{n}$ (see e.g. Gallant and White (1988), White (1994)). However, in this case the inference procedure is consistent in the sense that asymptotically it has zero size and unit power (see also Altissimo and Corradi (1997, 2002)).

We note that although the paper concentrates on statistics of the form (21), it is straightforward to apply our results to statistics of the form $g(\sqrt{n}\Delta\hat{Q}_n) - g(c_n)$ where $g(\cdot)$ is a sign preserving continuous (a.s) function. Such statistics may have some advantages in terms of power or size in small samples; this is left for future research. In this paper we use estimators that are based on data moments. However, our results can be generalized to any estimator of the form $\hat{\theta}_n^i - \theta_0^i = B_n \times H_n$, where B_n is $(k \times q)$ with $B_n \to_{a.s.} B_0$, B_0 is a matrix of rank k, $H_n = \frac{1}{n}\sum_{t=1}^n h_t(\theta_0^i)$ for $(q \times 1)$ orthogonality conditions $h_t(\theta_0^i)$ and a random diagonal matrix, $\hat{W}_n$, if $\hat{W}_n \to_{a.s.} W$.

4 A simulation study

We now show in a Monte Carlo simulation study how the derived model selection procedure works for reasonable sample sizes. We use the criterion (21) with $c_n = a_n\sqrt{2\hat{B}_n'\hat{V}_n\hat{B}_n}$. One has to determine the number of moments q that influence the selection criterion. As we use moderate to large sample sizes, we decided to fix $q = 10$. This seems to be a common value, also considering for example the common use of ten lags in Portmanteau-type statistics. However, in very large samples one should probably increase this value.

To investigate the selection performance we consider the following three possible scenarios: the first is given by a true SV model, the second by equivalence between SV and GARCH (in which case we prefer GARCH), and the third one by a true GARCH model.

4.1 SV is true model

To consider a realistic SV model for which we know that our closed form estimator works well, we let $\phi = -0.411$, $\delta = 0.95$, $\sigma = 0.484$. This is one of the processes investigated above in the simulation comparison between two closed form estimators of the SV model. We saw there that our estimator is more efficient than the one of Dufour and Valéry for this parameterization. Recall that this process is characterized by high persistence (i.e. δ is close to one) and a high kurtosis of 33. These are typical values for high frequency financial returns. We set the weight matrix to the identity matrix, which puts enough

weight on the kurtosis such that the true model, SV will be picked asymptotically. The SV process was generated k times for sample size n ranging from 100 to 10,000. Table 2 reports the number of times the SV and GARCH models were selected. To shed light on the average magnitude of the selection criterion, we also report the median over the k replications.

n	k	SV	GARCH	med(SC_n)
100	10,000	327	9673	-671.26
200	10,000	256	9744	-786.10
500	1,000	102	898	-264.02
750	1,000	274	726	-466.69
1,000	1,000	390	610	-225.95
2,000	1,000	667	333	576.47
5,000	1,000	786	214	3078.63
10,000	1,000	848	152	7600.80

Table 2: *Monte Carlo results. Reported are the number of times the GARCH or SV models are selected, out of k replications of a generated SV process of length n. The weight matrix is $W = I_q$, such that the SV model should be preferred. The last column reports the median of the selection criterion over the k replications.*

As a result one can state that the model selection procedure is selecting the correct model in most cases when the sample size is moderate to large, although the convergence seems to be relatively slow. The reason is the penalty term that encourages selection of the GARCH model and that is dominated by the goodness-of-fit term for sufficiently large samples. Recall also that the penalty term depends on the estimation of moments of order eight, which is highly erratic in small samples. This can explain the positive finite sample bias of the estimated penalty term, which implies the slow convergence of the selection procedure if the SV model is true.

4.2 SV and GARCH are equivalent

The same process is generated: SV with $\phi = -0.411$, $\delta = 0.95$, $\sigma = 0.484$, but now with W equal to zero except for: $W_{11} = 1$, $W_{33} = 1$ and $W_{44} = 1$. As the variance, the first and second order autocorrelations are used in the estimation of GARCH, these moments will asymptotically be the same as the true ones, given by the SV model. Thus, we have

equivalence. The pseudo-true parameters of the GARCH model are $\omega = 0.00010749$, $\alpha = 0.1834$ and $\beta = 0.6965$. Results are reported in Table 3.

n	k	SV	GARCH	med(SC_n)
100	10,000	4	9996	-2.15
200	10,000	16	9984	-1.98
500	1,000	2	998	-1.75
1,000	1,000	3	997	-1.66
2,000	1,000	5	995	-1.58
5,000	1,000	3	997	-1.74

Table 3: *Monte Carlo results. Reported are the number of times the GARCH or SV models are selected, out of k replications of a generated SV process of length n. The weight matrix is chosen such that both models are equivalent, so that we should prefer the GARCH model. The last column reports the median of the selection criterion over the k replications.*

We conclude that the GARCH model is selected for all sample sizes in almost all cases.

4.3 GARCH is true model

We generate a GARCH(1,1) model with $\alpha = 0.1$, $\beta = 0.85$ and $\omega = 0.0009(1 - \alpha - \beta)$. This generates the same persistence ($\alpha + \beta = 0.95$) as the previous SV model. Again we use $W = I_q$. Table 4 shows the results.

To conclude, if the GARCH model is true, then using our selection criterion gives a clear preference for the GARCH model, even in small samples. Thus, our simulation study confirms that the proposed model selection criterion is consistent and working well in moderate to large sample sizes. The imposed preference for GARCH under equivalence implies that convergence of the selection if the SV model is true is slower than convergence if the GARCH model is true.

5 Empirical Application

We now apply our selection procedure to 26 daily stock returns of the New York Stock Exchange (NYSE), contained in the Dow Jones Industrial Average index. The investigated period is January 2, 1973 to May 13, 2003. Cases where the estimated GARCH or SV models implied nonstationarity or infinite fourth moments were discarded. The sample

n	k	SV	GARCH	med(SC_n)
100	10,000	1	9999	-17453
200	10,000	1	9999	-18786
500	1,000	1	999	-14945
1,000	1,000	1	999	-12778
2,000	1,000	0	1000	-10280
5,000	1,000	0	1000	-7736
10,000	1,000	0	1000	-7696
20,000	1,000	0	1000	-2281

Table 4: *Monte Carlo results. Reported are the number of times the GARCH or SV models are selected, out of k replications of a generated GARCH process of length n. The weight matrix is $W = I_q$, so that we should prefer the GARCH model. The last column reports the median of the selection criterion over the k replications.*

size in most cases is $n = 7667$. Again we fix $q = 10$ and set the weight matrix to identity, $W = I_q$. Table 5 reports the results, where the selection criterion SC_n is decomposed into the goodness-of-fit term and the penalty term.

As a result, the GARCH model is selected in all cases. Note however that only looking at the goodness-of-fit term would give a preference for the SV model in all cases, except for one (Merck). A possible interpretation is that the advantage of the SV model in terms of goodness-of-fit is not precise enough, so that the penalty term that depends on the variation of sample moments takes large values. As a consequence, one would prefer the GARCH model because the risk of model equivalence (or even misspecification) would otherwise be too large. It seems that the SV model, although being more flexible and providing a better fit to the data, suffers from the high variability of sample moments, especially the second moments of $\log y_t^2$, and that a preference for SV models would only become apparent for even larger sample sizes. Recall also from the simulation study that even in the case of a true SV model and $n = 5000$, there is still a misclassification rate of 21% in favor of GARCH.

6 Conclusions

In this paper we proposed a simple model selection procedure to select between a GARCH and a stochastic volatility (SV) model. We have shown that this procedure is strongly con-

Stock	$\sqrt{n}(\hat{Q}_{2n} - \hat{Q}_{1n})$	c_n	SC_n
Caterpillar	3418	7103	-3684
Citibank	8107	22411	-14303
Cocacola	14486	37137	-22651
Dupont	1402	5924	-4521
Ekodak	53775	128455	-74679
Exxon	1932	42908	-40975
GenMotors	2151	3891	-1740
Homedepot	6120	17530	-11410
Honeywell	4120	40507	-36387
HP	2490	2698	-208
IBM	1923	46362	-44439
Johnson	2969	9116	-6146
JP Morgan	3623	15248	-11625
McDonalds	4920	12709	-7789
Merck	-406	4848	-5255
MMM	1451	34521	-33070
PhMorris	2280	52706	-50426
ProcterGamble	13686	193295	-179609
SBC	315	4711	-4395

Table 5: *Empirical application of the model selection to stocks in the Dow Jones index at NYSE.*

sistent, meaning that we find the correct model asymptotically with probability one. Under model equivalence we prefer the GARCH model due to its simplicity and widespread availability in software packages. The selection is based on closed form estimators that avoid typical problems of numerical optimization and dependence on initial values.

Finally, we would like to emphasize that the proposed model selection procedure is of course more generally applicable to the selection of non-nested models. We have focused in this paper on the GARCH and SV models mainly for the sake of illustration and for practical relevance, as it is a well-known problem in financial econometrics. For example, the same model selection procedure could be applied to the selection between the autoregressive conditional duration (ACD) model of Engle and Russell (1998) and the stochastic conditional duration model of Bauwens and Veredas (2003).

Acknowledgements

Arie Preminger acknowledges the financial support provided through the European Community's Human Potential Programme under contract HPRN-CT-2002-00232 , MICFINMA and Ernest foundation. Both authors are grateful for helpful discussions and comments by Luc Bauwens, Robert De Jong, Oliver Linton, Sharon Rubin, and Giuseppe Storti.

Appendix

Proof of Theorem 1

By Proposition 15 of Carrasco and Chen (2002), Assumption 1 implies that $\{h_t\}_{t\in\mathrm{N}}$ and $\{y_t\}_{t\in\mathrm{N}}$ is strictly stationary and β-mixing with exponential decay. As $\{z_t\}_{t\in\mathrm{N}}$ is a measurable function of $\{y_t\}_{t\in\mathrm{N}}$, this implies by Theorem 14.1 of Davidson (1994) that it is also strictly stationary, ergodic and exponential β-mixing. By the ergodic theorem (e.g. theorem 13.12 of Davidson, 1994), $\bar{z} \rightarrow_{a.s.} E(z_t) = c_0 + \phi_0/(1-\delta_0)$. Thus, the numerator of (8) can be approximated almost surely by $\sum_{t=3}^{n} z_{t-2}(z_t - c - \phi_0/(1-\delta_0))$. Substituting z_t by (3), straightforward calculations show that

$$\hat{\delta}_n = \delta_0 + \frac{n^{-1}\sum_{t=3}^{n} z_{t-2}(\eta_t - \delta_0\eta_{t-1} + \sigma_0\varepsilon_t)}{n^{-1}\sum_{t=3}^{n} z_{t-2}(z_{t-1} - \bar{z})} + o_{a.s.}(1) \tag{26}$$

Note that z_{t-2} and $\eta_t - \delta\eta_{t-1} + \varepsilon_t$ are independent with mean zero, so that the numerator of the second term of (26) converges to zero almost surely. It remains to show that the denominator converges a.s. to a positive limit. Note that $\{z_t z_{t-1}\}_{t\in\mathrm{N}}$ is strictly stationary and ergodic by Theorem 14.1 of Davidson (1994). Now, $E(z_t z_{t-1}) = (c + \phi_0/(1-\delta_0))^2 +$

$\sigma_0^2/(1-\delta_0^2) > 0$, so that the denominator of the second term of (26) converges a.s. to $\sigma_0^2/(1-\delta_0^2) > 0$. This proves that the second term of (26) converges a.s. to zero, and that $\hat{\delta}_n \to \delta_0$ a.s. Second, note that $\hat{\phi}_n$ is a continuous function of two strongly consistent estimators, so that $\hat{\phi}_n \to_{a.s.} (E[z_t] - c)(1-\delta_0)$. Since $E[z_t] = c + \frac{\phi_0}{1-\delta_0}$, this means that $\hat{\phi}_n \to_{a.s.} \phi_0$ as stated. Finally, $var(z_t) = var(\log h_t) + var(\eta_t)$ due to the independence of η_t and h_t, and $var(\log h_t) = \sigma_0^2/(1-\delta_0^2)$. Again using Theorem 14.1 of Davidson (1994), $\{z_t^2\}_{t\in\mathrm{N}}$ is strictly stationary and ergodic, which implies that a strong law of large numbers holds for the mean of z_t^2. Together with strong consistency of $\hat{\delta}_n$, $\bar{z}$, it follows that $\hat{\sigma}_n^2 \to_{a.s.} \sigma_0^2$. Now, since $\{z_t\}_{t\in\mathrm{N}}$ is β-mixing with exponential decay that is bounded in L_2-norm, it is also an adapted L_2-mixingale of size -1 with respect to the subfields $\mathrm{G}_t = \sigma(z_t, z_{t-1}, \ldots)$, see Davidson (1994), Theorem 14.2. Therefore, we can apply the CLT of Scott (1973) to show that $\sqrt{n}(\hat{\lambda}_n^1 - \lambda_0^1) \sim N(0, \tilde{V})$ and the desired result follows by the continuous mapping theorem.

Proof of Theorem 2

1. Assumptions 2.1 and 2.2, allow us to show that $\hat{\lambda}_n \to_{a.s.} \lambda_0$, $\hat{\theta}_n^1 \to_{a.s.} \theta_0^1$ and $\hat{\theta}_n^2 \to_{a.s.} \theta_0^2$ by Slutsky's theorem and the ergodic theorem. Further, the smoothness of F and A, Slutsky's theorem and (22) imply that

$$SC_n = \sqrt{n}(\Delta + o_{a.s.}(1))$$

and the desired result therefore follows.

2. Similar to Theorem 2.1 we can show that $\hat{\tau}_n \to_{a.s.} \tau_0$ and by using a mean value expansion, we have that

$$\sqrt{n}(\hat{Q}_{2n} - \hat{Q}_{1n}) = 2\left[A(\theta_0^1, \theta_0^2, \gamma_0) + o_{a.s.}(1)\right]\left[\frac{\partial F}{\partial Z} + o_{a.s.}(1)\right](\tau_0 + o_{a.s.}(1))\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t(\lambda_0, \hat{\psi}_n),$$

where $\partial F/\partial Z$ is evaluated at $Z = (\lambda_0^1, \gamma_0)$. Assumptions 2.1, 2.2(ii), McLeish α-mixing inequality (McLeish (1975)), and Cauchy-Schwartz inequality imply that the centred series $\{x_t(\lambda_0, \psi_0)\}_{t\in\mathrm{N}}$ is an adapted L_2-mixingale of size $-\frac{1}{2}$ with respect to the subfields $\mathrm{F}_t = \sigma(y_t, y_{t-1}, \ldots)$. The stationarity and ergodicity properties of $\{y_t\}_{t\in\mathrm{N}}$ are preserved by any continuous transformation, so $\{x_t(\lambda_0, \psi_0)\}_{t\in\mathrm{N}}$ and $\{y_t^2\}_{t\in\mathrm{N}}$ are also ergodic and strictly stationary processes (see White (2001), Theorem 3.35). Hence, we can apply Corollary AIII.3 of Sin and White (1992) to establish the LIL for this process. This implies that $|\hat{\psi}_n - \psi_0| = O_{a.s.}(a_n n^{-0.5})$ and $\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t(\lambda_0, \psi_0)\right| = O_{a.s.}(a_n)$, where $|\cdot|$ denotes the Euclidean norm. By a mean value expansion

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t(\lambda_0, \hat{\psi}_n) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t(\lambda_0, \psi_0) + (\hat{\psi}_n - \psi_0)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\frac{\partial x_t}{\partial \psi}(\lambda_0, \psi_0) + o_{a.s.}(1)\right]$$

where

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t(\cdot,\psi_0) + o_{a.s.}(a_n) = O_{a.s.}(a_n)$$

and hence,

$$\sqrt{n}(\hat{Q}_{2n} - \hat{Q}_{1n}) = B_0\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t(\lambda_0,\psi_0) + o_{a.s.}(a_n) = O_{a.s.}(a_n) \tag{27}$$

and $SC_n = O_{a.s.}(a_n) - c_n = -c_n(1 + o_{a.s.}(1))$. Thus, SC_n is negative almost surely and the desired result follows.

3. From (27) we have

$$SC_n = a_n\left[B_0\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t(\lambda_0,\psi_0)/a_n - C_0 + o_{a.s.}(1)\right] - a_n\left[c_n/a_n - \hat{C}_n + (\hat{C}_n - C_0)\right] \tag{28}$$

where $C_0 = \sqrt{2B_0'V_0B_0}$ and $\hat{C}_n = \sqrt{2\hat{B}_n'V_n\hat{B}_n}$. By the LIL, the first term on the RHS is non-positive, so if $\hat{C}_n - C_0 \to_{a.s.} 0$ and $c_n \geq a_n\hat{C}_n$ then $SC_n \leq 0$ a.s. The strong consistency of $(\hat{\lambda}_n, \hat{\theta}_n^1, \hat{\theta}_n^2)$ and the smoothness of $A(\cdot)$ and $F(\cdot)$ imply that $\hat{B}_n - B_0 \to_{a.s.} 0$. Thus, it suffices to show that $\sum_{(u,s)}(\hat{V}_n^{us} - V_0^{us}) = o_{a.s.}(1)$ where $\hat{V}_n^{us}$ and V_0^{us} are the (u,s) elements of $\hat{V}_n$ and V_0, respectively. Using the LIL we can show similarly to 2.2 that each element of $(\hat{\lambda}_n - \lambda_0)$ is $O_{a.s.}(a_n n^{-0.5})$. This result combined with Assumptions 2.1, 2.2(iii), 2.3 and the c_r inequality imply that $\hat{V}_n - V_0 \to_{a.s.} 0$ element-wise by Theorem 3 of De Jong (2000), and the desired result follows.

Proof of Theorem 3

1. The proof follows Theorem 1 in De Jong (2000) with some changes to accommodate the absence of the kernel weighting function. We need to show that the second term on the RHS of (23) is $o_{a.s.}(1)$. As in Theorem 2.1 and 2.2, we can show that $\hat{\theta}_n^1$, $\hat{\theta}_n^2$ and $\hat{\lambda}_n$ are strongly consistent estimates, and it suffices to establish that $\sum_{j=1}^{q_n}\hat{v}_n(j) - v_0(j) \to_{a.s.} 0$.

Let $z_t = (y_t^2 - \psi_0)$ and $\tilde{v}_n(j) = \frac{1}{n}\sum_{t=1}^{n} z_t z_{t-j}$ and let $\|X\|_\delta = (E|X|^\delta)^{1/\delta}$. We begin by establishing that $\sum_{j=1}^{q_n}\tilde{v}_n(j) - v_0(j) \to_{a.s.} 0$. Consider the adapted mixingale $\{z_t z_{t-j}\}_{t\in\mathrm{N}}$ with respect to the subfields $\mathcal{F}_t = \sigma(y_t, y_{t-1}, \ldots)$. Similar to Hansen (1992), we can show that for $j \geq 0$, and $\varphi > \delta \geq 1$

$$||E(z_t z_{t-j}|\mathcal{F}_{t-m}) - E(z_t z_{t-j})||_\delta \leq \{4\alpha(m-i)^{1/\delta - 1/\varphi}\mathrm{I}(m > j) + 2\mathrm{I}(m \leq j)\}||y_t^2||_{2\varphi}^2$$

where $\mathrm{I}(\cdot)$ is the indicator function. Therefore, using Lemma 2 of Hansen (1991),

$$\left\|\max_{2^m \leq n \leq 2^{m+1}}\left|\sum_{t=1}^{n} z_t z_{t-j} - E(z_t z_{t-j})\right|\right\|_{r/2} \leq C_2(C_1 + 2j)2^{(m+1)2/r} \tag{29}$$

for some $C_1, C_2 > 0$. Hence,

$$\left\| \max_{2^m \le n \le 2^{m+1}} \left| \sum_{j=1}^{q_n - 2} \tilde{v}_n(j) - v_0(j) \right| \right\|_{r/2}^{r/2} \le \left\| \max_{2^m \le n \le 2^{m+1}} \sum_{j=1}^{q_n} \frac{1}{n} \left| \sum_{t=1}^{n} z_t z_{t-j} - E(z_t z_{t-j}) \right| \right\|_{r/2}^{r/2}$$

$$\le \left(\sum_{j=1}^{q_{2^{m+1}}} 2^{-m} \left\| \max_{2^m \le n \le 2^{m+1}} \left| \sum_{t=1}^{n} z_t z_{t-j} - E(z_t z_{t-j}) \right| \right\|_{r/2} \right)^{r/2}$$

$$\le \left(\sum_{j=1}^{q_{2^{m+1}}} 2^{-m} C_2 (C_1 + 2j) 2^{(m+1)2/r} \right)^{r/2} = O(2^{m(1-r/2)} q_{2^{m+1}}^r)$$

The last inequality uses (29) and since $q_n = O(n^{1/2-1/r}(\log n)^{-1/r-\eta})$ for some $\eta > 0$, $\sum_{m=0}^{\infty} 2^{m(1-r/2)} q_{2^{m+1}}^r < \infty$. Hence we can apply Lemma 1 of De Jong (2000) to show that $\left| \sum_{j=1}^{q_n} \tilde{v}_n(j) - v_0(j) \right| = o_{a.s.}(1)$. After some algebra

$$\begin{aligned} \sum_{j=1}^{q_n} \hat{v}_n(j) - \tilde{v}_n(j) &= \sum_{j=1}^{q_n} \frac{1}{n} \sum_{t=1}^{n} (y_t^2 + y_{t-j}^2 + \psi_0 + o_{a.s.}(1))(\hat{\psi}_n - \psi_0) \\ &= O_{a.s.}(q_n a_n n^{-1/2}) = o_{a.s.}(1) \end{aligned}$$

which implies that $SC_n = \sqrt{n}(\Delta_n + o_{a.s.}(1))$ and because $\limsup_{n\to\infty} \Delta_n < 0$, the desired result follows.

2. By using similar arguments as in the proof of Theorem 2.1 and 2.2 we can apply the LIL to x_t element–wise, so that $\sqrt{n} SC_n = \sqrt{n}\Delta_n + O_{a.s.}(q_n a_n) - c_n = -c_n(1 + o_{a.s.}(1))$, and the desired result follows.

3. By applying the LIL element-wise, we can bound $\Delta\hat{Q}_n/a_n$ by $\tilde{C}_0 = \sum_{i=1}^{q_n} V_0^i |B_0^i|$ and

$$\sqrt{n} SC_n = a_n q_n \left[\sqrt{n}\Delta_n / a_n q_n + o_{a.s.}(1) - c_n / a_n q_n + \hat{C}_n / q_n - (\tilde{C}_n - \tilde{C}_0)/q_n \right]$$

and since $\frac{1}{q_n} \sum_{i=1}^{q_n} (\hat{V}_n^i - V_0^i) = o_{a.s.}(1)$ (note that $\tilde{C}_0 = O(q_n)$ and $\Delta_n = o(a_n q_n n^{-0.5})$), the desired result follows.

References

Altissimo F. and Corradi V. (1997). A LIL for m-estimators and applications to hypothesis testing with nuisance parameters. Manuscript, University of Pennsylvania.

Altissimo F. and Corradi V. (2002). Bounds for inference with nuisance parameters present only under the alternatives. *Econometrics Journal*, 5, 494-518.

Altisimo F. and Corradi V. (2003). Strong rules for detection the number of breaks in a time series. *Journal of Econometrics*, 117, 207-244.

Andersen T.G. and Sorensen B.E.(1996). GMM estimation of a stochastic volatility model: a Monte Carlo study. *Journal of Business and Economic Statistics*, 14, 328-352.

Bai J. (1999). Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, 91, 299-323.

Bai X., Russell J. R. and Tiao G. C. (2003). Kurtosis of GARCH and stochastic volatility models with non-normal innovations. *Journal of Econometrics*, 114, 349-360.

Baillie R.T. and Chung H. (2001). Estimation of GARCH models from the autocorrelations of the squares of a process. *Journal of Time Series Analysis*, 22, 631-650.

Bartlett M.S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, 37, 1-16.

Bauwens, L. and Rombouts, J. (2004) Econometrics. In: *Handbook of Computational Statistics*, Volume I, edited by J. Gentle, W. Haerdle, and Y. Mori. Springer Verlag, Berlin.

Bauwens, L. and Veredas, D. (2004). The stochastic conditional duration model: a latent factor model for the analysis of financial durations. *Journal of Econometrics*, 119/2, 381-412.

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

Bollerslev T., Engle R.F. and D.B. Nelson (1994). ARCH Models. In: R.F. Engle and L. McFadden (eds.), *Handbook of Econometrics*, 4, 2959–3038, Elsevier Science, Amsterdam.

Brooks C., Burke S.P. and Persand G.(2001). Benchmarks and the accuracy of GARCH model estimation. *International Journal of Forecasting*, 17, 45-56.

Broto C. and Ruiz E. (2004). Estimation methods for stochastic volatility models: a survey. *Journal of Economic Surveys*, 18, 613-649.

Calzolari G., Fiorentini G. and Sentana E.(2001). Indirect inference estimation of conditionally heteroskedastic factor models. CEMFI Unpublished Paper.

Carnero, M.A., Pena D. and Ruiz E. (2004). Persistence and kurtosis in GARCH and stochastic volatility models. *Journal of Financial Econometrics*, 2, 319-342.

Carrasco, M. and Chen X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, 18, 17-39.

Chesney M. and Scott L. O. (1989). Pricing European options: a comparison of the modified Black-Scholes model and a random variance model. *Journal of Financial and Qualitative Analysis*, 24, 267-284.

Comte F. and Lieberman O.(2003). Asymptotic theory for multivariate GARCH processes. *Journal of Multivariate Analysis*, 84, 61-84.

Danielsson, J. (1994). Stochastic volatility in asset prices: estimation with simulated maximum likelihood. *Journal of Econometrics*, 64, 375-400.

Danielsson, J. and J.F. Richard (1993). Accelerated Gaussian importance sampler with application to dynamic latent variable models. *Journal of Applied Econometrics*, 8, S153-S173.

Davidson J. (1994). *Stochastic limit theory.* New York, Oxford University Press.

De Jong R. M. (2000). A strong consistency proof for heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometric Theory*, 16, 262-268.

Duffie D. and Singleton K.J.(1993). Simulated moments estimation of Markov models of asset prices. *Econometrica*, 61, 929-952

Dufour J-M. and P. Valéry (2006). On a simple two-stage closed-form estimator for a stochastic volatility in a general linear regression. *Advances in Econometrics*, 20, 259–288.

Engle, R.F. and Lee G. (1996). Estimating diffusions models of stochastic volatility. In P.E. Rossi (ed.) Modelling stock market volatility: bridging the gap to continuous time, San Diego: Academic Press, pp. 333-355.

Engle, R.F. and Russell J. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66, 1127–1162.

Gallant A.R., Hsieh D. and Tauchen G. (1997). Estimation of stochastic volatility model with diagnostics. *Journal of Econometrics*, 81, 159-192.

Gallant A.R. and White H. (1988). *A unified theory of estimation and inference for nonlinear dynamic models*. Basil Blackwell, New York.

Gerlach, R. and Tuyl, F. (2006). MCMC methods for comparing stochastic volatility and GARCH models. *International Journal of Forecasting*, 22, 91–107.

Ghysels, E., Harvey A.C. and Renault E.(1996). *Stochastic Volatility*. In G.S. Maddala and C.R. Rao (eds.), *Statistical Methods in Finance*, North Holland, Amsterdam, 119-191.

Granger C. W. J., King, M. L. and White H. (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometric*s, 67, 173-187.

Hafner C.M. and Herwartz H.(2000). Testing for linear autoregressive dynamics under heteroskedasticity. *Econometrics Journal*, 3, 177-197.

Han C. and P.C.B. Phillips (2006). GMM with many moment conditions. *Econometrica*, 74, 147-192.

Hansen B.E. (1991). Strong laws for dependent heterogenous processes. *Econometric Theory*, 7, 213-221.

Hansen B.E. (1992). Consistent covariance matrix estimation for dependent heterogeneous processes. *Econometrica*, 60, 967-972.

Harvey A.C., Ruiz E. and Shephard N.G.(1994). Multivariate stochastic variance models. *Review of Economic Studies*, 61, 247-264.

Jacquier E., Polson N.G. and Rossi P.E. (1994). Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business and Economic Statistics*, 12, 371-417.

Jiang G.J. and van der Sluis P.J. (2000). Index option pricing models with stochastic volatility and stochastic interest rates. *European Finance Review*, 3, 273-310.

Kim S., Shephard N. and Chib S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies,* 65, 361-393.

Kristensen, D. and Linton O. (2006). A closed form estimator for the GARCH (1,1) model. *Econometric Theory*, 22, 323-337.

Liesenfeld R. and Jung R.C. (2000). Stochastic volatility models: conditional normality versus heavy-tailed distributions. *Journal of Applied Econometrics*, 15, 137-160.

Liesenfeld, R. and Richard, J.-F. (2003). Univariate and Multivariate Stochastic Volatility Models: Estimation and Diagnostics. *Journal of Empirical Finance*, 10, 505-531.

McLeish D.L. (1975). A maximal inequality and dependent strong laws. *The Annals of Probability*, 5, 829-839.

Melino A. and Turnbull S.M.(1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, 45, 239-265.

Newey W.K. and West K.D. (1987). A simple positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703-708.

Preminger A. and Wettstein D. (2005). Using the penalized likelihood method for model selection with nuisance parameters present only under the alternative: An application to switching regression models. *Journal of Time Series Analysis*, 26, 715-741.

Rivers D. and Vuong Q. H. (2002). Model selection tests for nonlinear dynamic models. *The Econometrics Journal*, 5, 1-39.

Ruiz E. (1994). Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of Econometrics*, 63, 289-306.

Sandmann G. and Koopman S.J. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *Journal of Econometrics*, 87, 271-301.

Scott D.J. (1973). Central limit theorems for martingales and for processes with stationary increments using a Skorokhod representations approach. *Advances in Applied Probability*, 5, 119-137.

Shephard N. (1996). Statistical aspects of ARCH and stochastic volatility. In O.E. Barndorff-Nielsen, D.R. Cox and D.V. Hinkley (eds.), Statistical Models in Econometric, Finance and other Fields, London: Chapman and Hall, pp. 1-67.

Sin C.Y. and White H. (1992). Information criteria for selecting possibly misspecified parametric models. Department of Economics Discussion Paper 92-47, San Diego: University of California.

Sin C.Y. and White H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics* 71, 207-225.

Stinchcombe M. B. and White H. (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory*, 14, 295-325.

Stock J. and M. Yogo (2005). Asymptotic distributions of instrumental variables statistics with many instruments. In D.W.K. Andrews and J.H. Stock, eds. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 109-120.

Storti G. (2006). Minimum distance estimation of GARCH(1,1) models. *Computational Statistics and Data Analysis, forthcoming.*

Taylor S. (1986). *Modeling Financial Time Series.* John Wiley and Sons, New York.

Teräsvirta T. (1996). Two stylized facts and the GARCH(1,1) model. WP series in Economics and Finance 96, Stockholm School of Economics.

White H. (1994). *Estimation, inference and specification analysis.* New York: Cambridge University Press.

Wong C.K.J. (2002). Estimating stochastic volatility: new method and comparison. Ph.D. thesis, Oxford University, U.K.