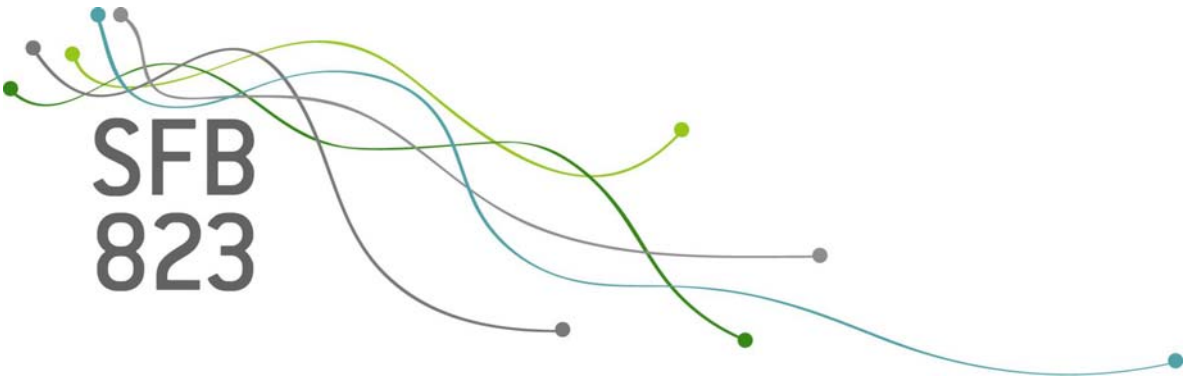
SFB 823

Discriminating between GARCH and stochastic volatility via nonnested hypotheses testing

Philip Messow

Nr. 27/2013

Discussion Paper



Discriminating between GARCH and Stochastic Volatility via nonnested hypotheses testing[☆]

Philip Messow[a,b]

[a] *Fakultät Statistik, Universität Dortmund, D-44221 Dortmund, Germany*
[b] *Ruhr Graduate School in Economics, D-45128 Essen, Germany*

Abstract

GARCH- and Stochastic Volatility (SV)-models are the main workhorses for describing unobserved volatility in asset returns. Because economic theory behind these models is not the same and estimating SV-models is much more difficult, discriminating between these two rival models is of interest. This paper suggests a nonnested testing procedure going back to Davidson and MacKinnon (1981) that does not implicitly assume that one of the models is the correct one. We illustrate the proposed test by applying it to ten daily stock index return series and five exchange rate return series.

Keywords:

Nonnested Testing, Stochastic Volatility, Model Selection

JEL: C10, C22, C52

[☆]Research supported by Deutsche Forschungsgemeinschaft (SFB 823). I thank the Ruhr Graduate School in Economics for the financial support.

Email address: messow@statistik.tu-dortmund.de (Philip Messow)

July 3, 2013

1. Introduction

Modelling conditional volatility is among the most important tasks of financial econometrics. Two competing model classes, with a different economic interpretation, are the main workhorses in this field, the GARCH-models, where the conditional volatility is described by past observations and the class of SV-models, where additional uncertainty enters via some extra error term. These competing models look quite similar in continuous time, but dissimilar in discrete time (Fleming and Kirby, 2003). While GARCH-models are much easier to estimate, SV-models need fewer restrictions on conditional moments than GARCH-models. From a practitioners point of view it would be good to know if the estimation of a much more difficult model is worth the effort. Furthermore, GARCH- and SV-models yield different economic interpretations. Due to the second innovation within the framework of the SV-model, the conditional variance process is a function of latent variables, which can be interpreted as the random and uneven flow of information (e.g. information about other assets and markets, volume of transactions or the order book). The GARCH-model in lieu thereof assumes that the conditional variance is perfectly explained by past observations. This economic aspect as well as the practical handling raises interest in discriminating between these both classes.

Tests to decide whether a GARCH- or a SV-model is appropriate go back to Kim et al. (1998) and normally rely on nested hypothesis testing. Popular examples are Kobayashi and Shi (2005) and Franses et al. (2008). One ma-

jor disadvantage of this type of model selection technique is that these tests implicity assume that one of the models is the true data generating process (DGP). But models are just approximations to the true DGP. The goal of a model selection technique should be to find a good approximation of the true DGP. That would include that neither the specific (nested) GARCH- nor the specific SV-model is a good approximation to the true DGP. In this paper we circumvent this problem by applying the popular C-test of Davidson and MacKinnon (1981) to the problem of discriminating between GARCH- and SV-models. Using this method it is possible that both, none or just one of the models are rejected. Because this kind of test normally suffers from size distortion in the form of overrejection for finite samples, we use a bootstrapped version of the test and compare the performance of the normal and the bootstrapped test.

2. The models

Bollerslev (2008) lists more than 100 different GARCH-type models in his glossary. This raises interest into the question of picking an appropriate model out of the infinite universe of GARCH-models. Hansen and Lunde (2005) compare 330 ARCH-type models and find no evidence that more sophisticated ARCH-models outperform the GARCH(1,1)-model, even though the GARCH(1,1) cannot capture the asymmetric response to shocks.

The GARCH(1,1)-model includes one lag of the conditional variance within the standard ARCH(1)-framework

$$y_t = \varepsilon_t \sigma_t \tag{1}$$

$$\sigma_t^2 = \phi + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{2}$$

ε_t is an IID process with zero mean and variance of unity. In most applications ε_t is assumed to be NID(0,1). To ensure the existence of the conditional variance and for avoiding the degeneration of the process $\phi > 0$ and $\alpha, \beta \geq 0$ must hold (Carnero et al., 2004). Moreover $\alpha + \beta < 1$ must hold for (weakly) covariance stationarity of y_t. The model can be estimated by a standard Maximum-Likelihood (ML)-procedure.

For the class of stochastic volatility models we follow Harvey et al. (1994) and define a (simple) SV-model as

$$y_t = \varepsilon_t \sigma_t \tag{3}$$

$$h_t = ln\ \sigma_t^2 = \gamma + \pi h_{t-1} + \xi_t, \tag{4}$$

where $\epsilon \overset{\text{N}}{\sim} (0,1)$ and $\xi_t \overset{\text{N}}{\sim} (0, \sigma_\xi^2)$. Formula 4 can be seen as the discrete-time approximation to the continuous-time Orstein-Uhlenbeck process used in financial econometrics mostly for modeling short term interest rates. Because y_t is a product of two processes, both of these processes must be stationary to ensure the stationary of y_t, that is $|\pi| < 1$ for ensuring the stationarity of h_t. This simple model behaves like the GARCH(1,1). It has excess kur-

tosis $exp(\sigma_\xi^2)$. Estimation is a little bit more advanced than the estimation of the GARCH-model due to the additional nuisance parameter. By using a state space representation of (3)-(4) and approximate $log(\varepsilon_t^2)$ by a mixture of two normally distributed random variables, one centered at zero, a Quasi-Newton-Raphson-method can be used to maximize the resulting ML-function.

3. Testing nonnested hypotheses

This chapter focus on hypotheses testing when the considered hypotheses are nonnested. In the following we will introduce the C-test proposed by Davidson and MacKinnon (1981) for discriminating between two rival (nonlinear) models and we will make use of this test for selecting a GARCH- or SV-model from section 2. Suppose a researcher wants to find out if economic theory behind these models is supported by empirical data. Using (1)-(2) and (3)-(4) one may want to test if one of the following hypotheses holds

$$H_0 : y_t = f_t(\theta_1) + \eta_{1t} \tag{5}$$

$$H_1 : y_t = g_t(\theta_2) + \eta_{2t}, \tag{6}$$

where θ_1 and θ_2 describe the parameter vector of the proposed models. By forming the (possibly) nonlinear regression

$$y_t = f_t(\hat{\theta}_1) + \alpha g_t(\hat{\theta}_2) + \eta_t \tag{7}$$

with both $\hat{\theta}_1$ and $\hat{\theta}_2$ the estimated parameter vectors, one can test H_0. α is estimated conditional on these estimates using a standard least squares procedure and the test statistic then reads $\hat{C} = \frac{\hat{\alpha}}{sd(\hat{\alpha})}$. It would also be possible to estimate θ_2 and α jointly, but the proposed procedure is preferred for nonlinear models (Davidson and MacKinnon, 1981).

Under H_1, $\hat{\alpha} \xrightarrow{p} 1$. But to test H_1 one needs to carry out a second regression, substituting H_0 and H_1. This is needed, because the test for H_0 is not valid for testing H_1 (Davidson and MacKinnon, 1981). Because of this sequential testing, it is possible that both models are rejected, neither is rejected or that one but not the other is rejected.

This accounts for the possible outcome that neither the proposed GARCH- nor the SV-model is a good approximation to the true data generating process, or that the true DGP is sufficiently close to both models.

3.1. Bootstrapped based testing

The test often overrejects in finite samples and the extent of this overrejection depends on the level of significance (Davidson and MacKinnon, 2002). One way to deal with this problem is to using a bootstrapped test statistic. By doing so, the finite sample performance of the tests can be enhanced dramatically (Fan and Li, 1995; Davidson and MacKinnon, 2002; Godfrey, 1998). To deal with autocorrelation we use the moving block bootstrap with a block length of $T^{\frac{1}{4}}$ for the simulation based testing (Hall et al., 1995). For the empirical application we combine the ideas of the wild- and blockbootstrap to account for dependent and heteroskedastic data. An al-

ternative for an appropriate bootstrap procedure robust to underlying heteroskedasticity would be the pairs bootstrap, but Flachaire (2003) compares different heteroskedasticity-robust bootstrap procedures and finds that the wild bootstrap of Davidson and Flachaire (2008) outperforms other wild and pairs bootstrap methods. The bootstrap procedure accounting for both heteroskedastic and autocorrelated observations looks like this:

1.) Estimate both models and calculate the test statistic $\hat{C}$.

2.) Estimation of the model under H_0 yields unbiased parameter estimates and thus provides the bootstrap data-generating process (DGP).

$$y_t^* = f_t(\theta_1) + \eta_t^* \; \xi_t \; a_t, \tag{8}$$

where $a_t = \sqrt{\frac{n}{n-k}}$ and $\xi_t = \begin{cases} 1, & \text{with probability 0.5} \\ -1, & \text{with probability 0.5} \end{cases}$. After the rescaling is done, the residuals are blocked using the moving block procedure mentioned above with a block length of $T^{1/4}$.

3.) B bootstrap samples are drawn from 8. B needs to be chosen such that the level of significance times $(B + 1)$ is an integer.

4.) For each B, the bootstrapped test statistic C^* is computed similar to the original test statistic.

5.) The bootstrap p-value is computed by

$$p^*(\hat{C}) = \frac{1}{B}\sum_{j=1}^{B} \mathbf{1}_{(C_j^* \geq \hat{C})}, \tag{9}$$

where $\mathbf{1}_{(.)}$ is an indicator function.

The bootstrap p-value converges faster to the true p-value than the asymptotic p-value does, given that the bootstrap test statistic's distribution converges to the true distribution as sample size is increasing and thus the bootstrap test statistic is asymptotic pivotal (Beran, 1988). As shown by Davidson and MacKinnon (2002), the test statistic for the standard linear regression model is asymptotically pivotal except one special case ($\theta_1 = 0$). Therefore we assume for the time being that this property holds for this (more complicated) model, too.

3.2. Finite sample properties

This section compares the performance of the test with its bootstrapped counterpart. We use both models as data generating processes with the following parameterizations that are typical for returns of stock indices:

GARCH: $\phi = 0.0001,\ \alpha = 0.09,\ \beta = 0.9$

SV: $\gamma = -0.005,\ \phi = 0.98,\ \sigma_\xi = 0.01.$

Table 1 and 2 report the results of a Monte Carlo Simulation with 1000 replications for the empirical size. The corresponding null hypothesis for table 1 is $H_0 : GARCH$ and $H_0 : SV$ for table 2. As mentioned above, it is often assumed that the test statistic follows a t_{n-k-1} distribution even though it is

well known that the distribution can be quite different. Because the sample size is really big (T=1000 up to 5000) the corresponding t-distribution is (almost) similar to the N(0,1)-distribution and we assume that $\hat{C} \sim N(0,1)$. The sample sizes were chosen to reflect typical sample sizes of empirical studies, because the proposed models are normally calibrated to daily data of at least three years. The purpose of the simulation is to test whether the the assumed distribution of the test statistic is viable and if by using a bootstrapped based test statistic the empirical size bias can be reduced.

Table 1 shows that the test almost always keeps its theoretical level of significance for all sample sizes. The bias seems to diminish as sample size increases. The bootstrapped version of the test enhances the performance to some extent given that the performance was already good. Especially the overrejection of the small sample sizes for a level of significance of 0.1 is reduced within the bootstrap framework (see table 3). Things change if we exchange the model under H_0 from GARCH to SV. If the DGP is the SV-model, the test overrejects for all levels of significance and all sample sizes. Using the bootstrapped version of the test the performance is enhanced dramatically. The empirical size meets the theoretical level of significance and thus the bootstrapped version of the test is able to discriminate between the proposed models.

Table 1: Empirical size of the C-test (DGP=GARCH-model)

	T				
Level of significance	1000	2000	3000	4000	5000
0.01	0.009	0.006	0.009	0.008	0.012
0.05	0.050	0.053	0.045	0.048	0.049
0.10	0.107	0.115	0.094	0.094	0.098

Table 2: Empirical size of the C-test (DGP=SV-model)

	T				
Level of significance	1000	2000	3000	4000	5000
0.01	0.025	0.023	0.018	0.020	0.022
0.05	0.090	0.085	0.097	0.084	0.088
0.10	0.168	0.169	0.154	0.155	0.149

Table 5 and 6 reports the empirical power of the bootstrapped version of the test. The power of the test is evaluated for different values of α and the difference from 0 of the true parameter value is displayed by $\Delta\ \alpha$. The power results are very encouraging especially for the empirically most crucial sample sizes. If the sample size is increased, the power increases too in a rapid fashion. By increasing $\Delta\ \alpha$, the test is able to detect the false null hypothesis much faster.

4. Empirical application

This section uses the proposed test to discriminate between GARCH and SV-models for modeling return series of economic quantities. We apply the test to stock index return series and to exchange rate return series. From a

Table 3: Empirical size of the bootstrapped C-test (DGP=GARCH-model)

	T				
Level of significance	1000	2000	3000	4000	5000
0.01	0.010	0.008	0.013	0.012	0.008
0.05	0.045	0.049	0.039	0.045	0.046
0.10	0.097	0.098	0.089	0.106	0.095

Table 4: Empirical size of the bootstrapped C-test (DGP=SV-model)

	T				
Level of significance	1000	2000	3000	4000	5000
0.01	0.010	0.008	0.012	0.013	0.007
0.05	0.049	0.044	0.055	0.051	0.050
0.10	0.104	0.098	0.095	0.093	0.106

Table 5: Empirical power of the bootstrapped C-test (DGP=SV-model)

	T				
$\Delta\ \alpha$	1000	2000	3000	4000	5000
0.01	0.27	0.44	0.48	0.58	0.63
0.02	0.50	0.73	0.82	0.86	0.95
0.03	0.72	0.91	0.94	0.96	1.00
0.04	0.84	0.99	0.99	0.99	1.00
0.05	0.92	0.99	0.99	1.00	1.00

Table 6: Empirical power of the bootstrapped C-test (DGP=GARCH-model)

	T				
$\Delta\ \alpha$	1000	2000	3000	4000	5000
0.01	0.45	0.76	0.92	0.98	0.99
0.02	0.52	0.86	0.97	0.99	1.00
0.03	0.66	0.95	0.99	1.00	1.00
0.04	0.72	0.96	0.99	1.00	1.00
0.05	0.77	0.98	1.00	1.00	1.00

theoretical point of view, one could argue that for stock index return series the additional nuisance parameter in the SV-model can be used to reproduce the more pronounced uncertainty in emerging markets compared to the G8-countries. Hence we want to shed light on the question whether our proposed test confirm these theoretical considerations. We use ten years of daily data ranging from 11/27/2002 to 11/27/2012 for the following countries: USA, Germany, France, Great Britian, Japan, Russia, Brasil, China, Taiwan and South Korea. The first five countries are considered to be one of the most developed countries in the world, the latter five have the highest weighting within the MSCI Emerging Markets index. Figure 1 shows four selected stock index return series. For all return series the typical volatility clusters are observable, with the most pronounced clustering for the Russian stock index. Furthermore the volatility for the emerging countries is more pronounced than for the developed countries. Because the sample size for all ten time series is roughly 2500, we use a blocklength of 7 for the bootstrap. For each index we run our proposed test two times substituting the null hypothesis

$H_{01} : GARCH$ to $H_{02} : SV$ for the second run.

Figure 2 shows exemplarily the distribution of the bootstrapped test statistic for the HANGSENG return series. The left hand side corresponds to $H_{01} : SV$ and the right hand side corresponds to $H_{02} : GARCH$. The added lines reflect appropriate density functions of a normal distribution for both null hypotheses.

Table 7 summarizes the results for both null hypotheses. It turns out that for $H_{01} : GARCH$, the null is rejected for all ten stock index return series, indicating that the GARCH(1,1)-model seems not to be a good model for describing the returns of the last ten years. Four out of ten times the SV-model is also rejected. The level of industrialization seems not to matter as both models are rejected for two more developed countries and also for two emergent countries. But for three among the four asian countries both models are rejected, indicating that one needs special care for modeling these return series.

On the one hand the results are an indication that the pretty simple model specifications we used here are not able to mimic the behavior of the return series observed in the real world and more sophisticated model specifications should be used. On the other hand one could interpret the results as the need for an additional error term during turbulent times at the financial markets as the sample includes the financial crisis from 2007 up to today.

Another field of application of the proposed models are exchange rate returns. We apply the test to five different exchange rate return series: US-Dollar to Euro, British Pound to Euro, Yen to Euro, British Pound to US-

Table 7: Test statistics for selected stock index returns

Stock Index	$H_0 : GARCH$	$H_0 : SV$
$DOWJONES$	6.30^{***}	-0.60
DAX	4.36^{***}	1.80
CAC	4.06^{***}	-6.18^{***}
$FTSE$	3.26^{***}	1.18
$NIKKEI$	2.73^{**}	4.43^{***}
$BOVESPA$	4.43^{***}	0.17
$HANGSENG$	6.36^{***}	-2.24^{**}
$KOSPI$	5.19^{***}	-0.02
RTS	4.93^{***}	1.48
$TAIEX$	5.40^{***}	4.30^{***}

Notes. Level of significance: *:10%; **:5%; ***:1%

Dollar and Swiss Franc to Euro. Figure 3 shows the corresponding time series. For all time series the typical volatility clustering is observable. Worth noting is the peak of the Swiss Franc to Euro series at 09/06/2011. On this day, the Swiss central bank introduced a minimum level for the exchange rate of Swiss Franc to Euro of 1.20 and the exchange rate on 09/05/2011 was 1.1122. Due to the announcement the exchange rate climbed up to the minimum level and resulted in an artificially high one-day return. Table 8 shows the results for the incorporated exchange rates. As for the stock index returns it stands out that the GARCH-model is always rejected in presence of the SV-model. For the Japanese Yen to Euro and Swiss Franc to Euro time series both models are rejected. This results are in line with the previ-

Table 8: Test statistics for selected exchange rate returns

Exchange Rate	$H_0 : GARCH$	$H_0 : SV$
US Dollar to Euro	-38.10^{***}	-1.42
British Pound to Euro	-36.63^{***}	-1.16
Japanese Yen to Euro	-31.62^{***}	-2.88^{***}
British Pound to US Dollar	-37.62^{***}	-6.27^{***}
Swiss Franc to Euro	-32.15^{***}	0.11

Notes. Level of significance: *:10%; **:5%; ***:1%

ous results as both models were rejected for the stock index return series of selected Asian countries (TAIEX, HANGSENG, NIKKEI), indicating that the pretty simple models used for the empirical application are not capable of describing the dynamics of Asian financial markets. In lieu thereof the SV-model adequately describes the dynamics of three out of five exchange rate returns.

It is possible that the turbolent last years increase the need for more sophisticated models also for exchange rate returns.

As for the stock index return application, figure 4 shows exemplarily the distribution of the bootstrapped test statistic for the Swiss Franc to Euro series. As before, the left hand side corresponds to $H_{01} : SV$ and the right hand side corresponds to $H_{02} : GARCH$ and the shape of the bootstrapped distribution is close to that of the normal distribution.

5. Possible extensions

Using empirical data it is not clear which null hypothesis is the natural one. From a practitioners point of view there is a continuum of competing models that need to be tested to pick an appropriate one. One possible extension for testing M different models at once that are all capable of explaining some (economic) variable y, $y = f_m(\theta_m) + u_m \; \forall \; m \in \mathbb{M} := \{1, ..., M\}$, and is robust to the sequential testing problem, is the MJ-test introduced by Hagemann (2012). The general procedure works like this:

1.) For each model, run regression

$$y = \left(1 - \sum_{l \in \mathbb{M} \setminus \{m\}} a_{l,m}\right) \; f_m(\theta_m) + \sum_{l \in \mathbb{M} \setminus \{m\}} a_{l,m} \; f_l(\theta_l) + \mu$$

and compute the test statistic $C_{n,m}$. Let $\Xi_n := \{C_{n,m} \; \forall \; m \in \mathbb{M}\}$ and $MC_n := min \; \Xi_n$.

2.) Test $H_0 : m^* \in \mathbb{M}$ against $H_1 : m^* \notin \mathbb{M}$ and reject the hypothesis, if $MC_n > \chi^2_{M-1,1-\alpha}$, where m^* stands for the correct model.

This type of test is an intersection-union test of Berger (1982). It tries to find out if $m^* \in \mathbb{M}$ and if this hypothesis is not rejected, $\bar{m} = argmin \; \Xi_n$ is the natural candidate due to the fact that only the model with the smallest test statistic can possibly be the correct model.

Using this idea, we can compare in a fairly simple way more than just two different competing SV/GARCH-models at the same time.

6. Conclusion

This paper has proposed a simple test for discriminating between nonnested GARCH- and SV-models. Within this framework it is possible to reject or accept both model types and thus the test does not implicitly assume that one of the models has to be the correct one. This respects the fact that all models are just approximations to the unknown true data generating process. Applying the test to exchange rate and stock index returns, the SV-model is preferred to the GARCH-model. But for some time series both models are rejected, indicating that these rather simple models may not be adequate for describing the turbulent last years reasonably well.

Extending the proposed test to compare more than just two models out of the infinite universe of GARCH- and SV-models is a topic for further research.

References

Beran, R., 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements. Journal of the American Statistical Association 83, 687–697.

Berger, R., 1982. Multiparameter hypothesis testing and acceptance sampling. Technometrics 24, 295–300.

Bollerslev, T., 2008. Glossary to ARCH(GARCH). CREATES Research Paper 2008-49.

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

Davidson, R., Flachaire, E., 2008. The wild bootstrap, tamed at last. Journal of Econometrics 146, 162–169.

Davidson, R., MacKinnon, J.G., 1981. Several tests for model specification in the presence of alternative hypotheses. Econometrica 49, 781–793.

Davidson, R., MacKinnon, J.G., 2002. Bootstrap J tests of nonnested linear regression models. Journal of Econometrics 109, 167–193.

Fan, Y., Li, Q., 1995. Bootstrapping J-type tests for non-nested regression models. Economics Letters 48, 107–112.

Flachaire, E., 2003. Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. EUREQUA, Universite Paris 1 Pantheon-Sorbonne.

Fleming, J., Kirby, C., 2003. A closer look at the relation between GARCH and stochastic autoregressive volatility. Journal of Financial Econometrics 1, 365–419.

Franses, P.H., van der Leij, M., Paap, R., 2008. A simple test for GARCH against a stochastic volatility model. Journal of Financial Econometrics 6, 291–306.

Godfrey, L., 1998. Tests of non-nested regression models: small sample adjustments and monte carlo evidence. Journal of Econometrics 84, 59–74.

Hagemann, A., 2012. A simple test for regression specification with non-nested alternatives. Journal of Econometrics 166, 247–254.

Hall, P., Horowitz, J.L., Jing, B.Y., 1995. On blocking rules for the bootstrap with dependent data. Biometrika 82, 561–574.

Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Journal of Applied Econometrics 20, 873–889.

Harvey, A., Ruiz, E., Shepard, N., 1994. Multivariate stochastic variance models. Review of Economic Studies 61, 247–264.

Kim, S., Shepard, N., Chib, S., 1998. Stochastic volatility: Likelihood inference and comparison with ARCH models. Review of Economic Studies 65, 361–393.

Kobayashi, M., Shi, X., 2005. Testing for EGARCH against stochastic volatility models. Journal of Time Series Analysis 26, 135–150.

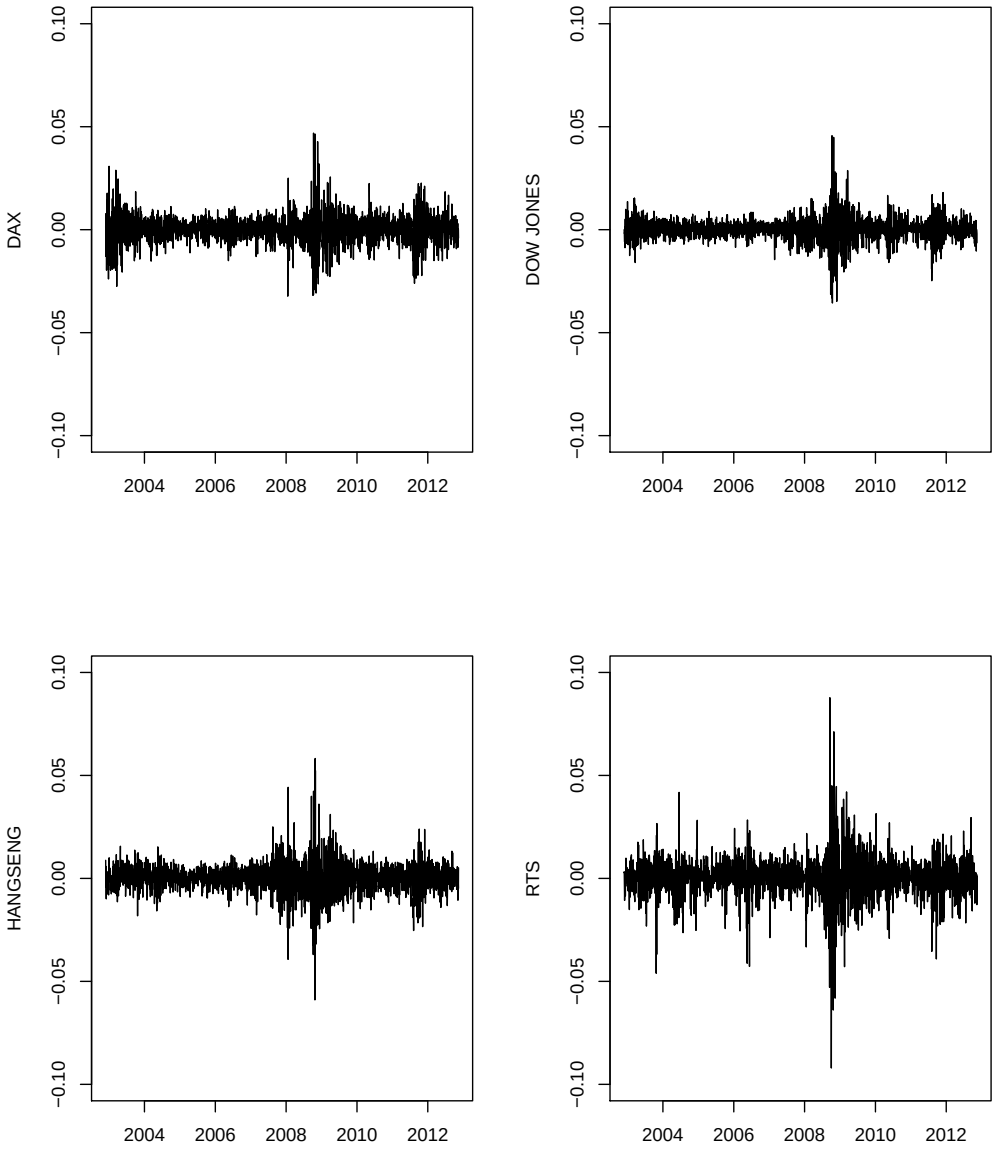


Figure 1: Returns of selected stock indices from 11/2002 - 11/2012

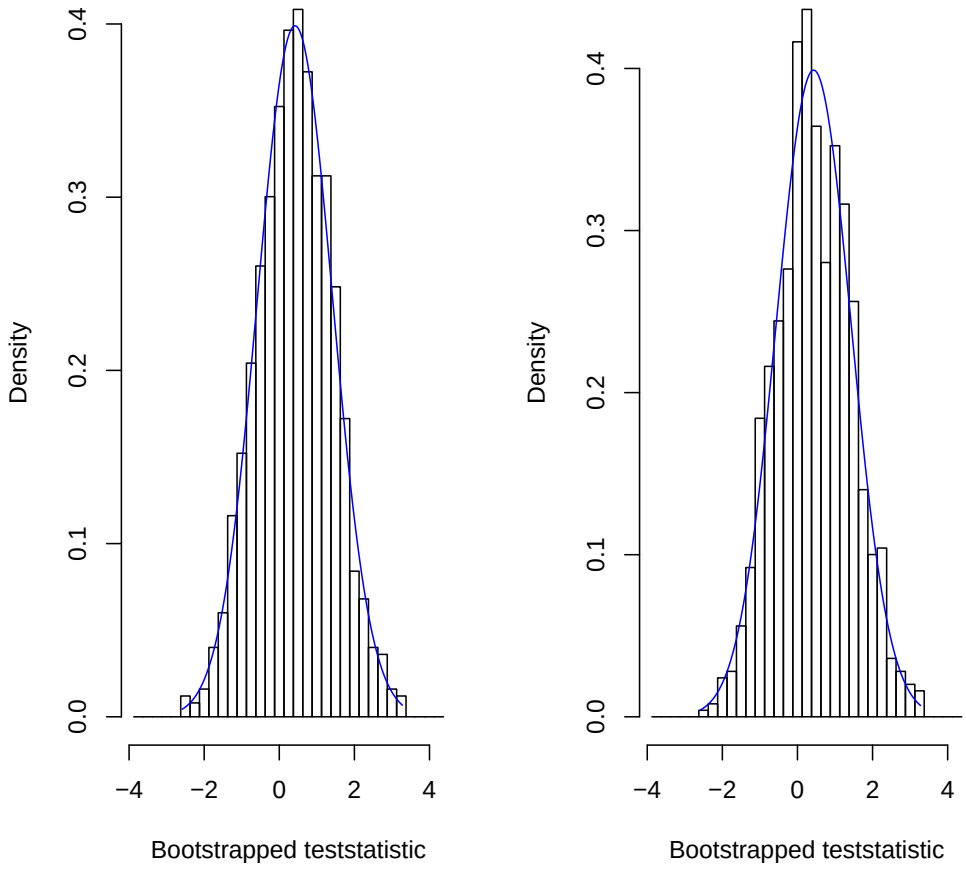


Figure 2: Distribution of the bootstrapped test statistic for the HANGSENG return series

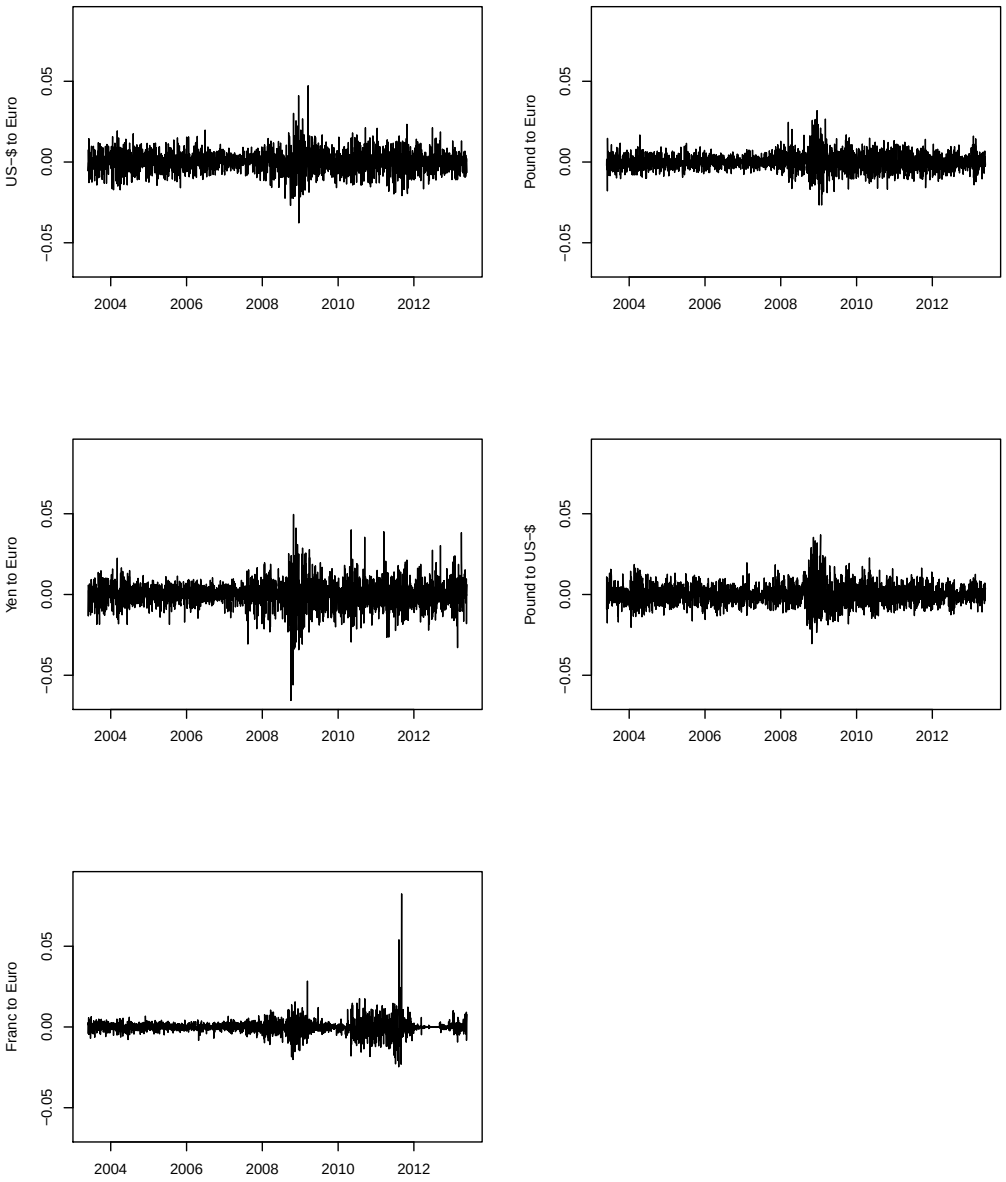


Figure 3: Returns of selected exchange rates from 05/2003 - 05/2013

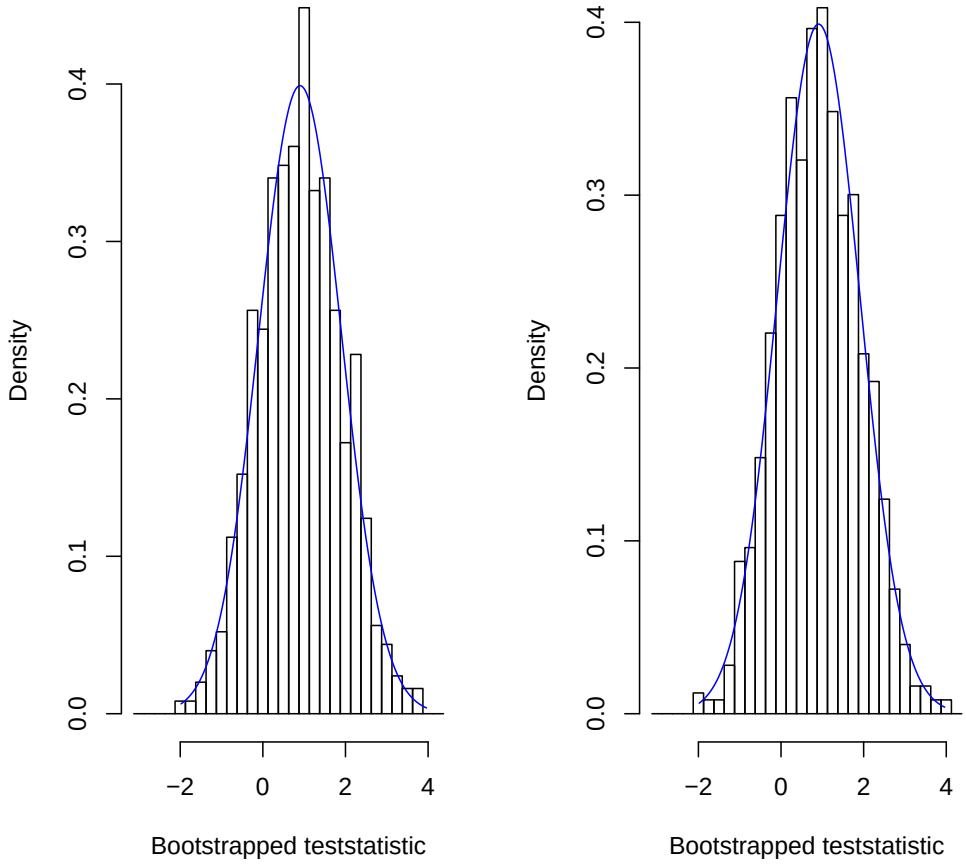


Figure 4: Distribution of the bootstrapped test statistic for the Swiss Franc to Euro return series