# Comparison of Misspecification Tests Designed for Nonlinear Time Series Models

Leena Kalliovirta

Department of Political and Economic Studies

University of Helsinki

November 29, 2010

**Abstract**

We use GARCH and regime switching models to compare the reliability of recently proposed misspecification tests. Our simulations indicate that simple moment based LM type tests are more reliable than other moment based tests or tests that employ the empirical distribution function or non-parametric methods.

*JEL classification:* C22, C52

*Keywords:* Quantile residual, LM type test, Khmaladze's martingale transformation, Empirical process, Non-parametric test

# 1 Introduction

The misspecification tests of Bai (2003), Duan (2003), Hong and Li (2005) and Kalliovirta (2006) are applicable to a wide class of models including linear and non-linear time series models, and these tests properly take into account the uncertainty caused by parameter estimation. Because the tests employ different methodologies, they may perform differently in finite samples. We study their performance by simulating GARCH and regime-switching models.

# 2 Compared tests

All the tests considered here use quantile residuals. These residuals, unlike traditional residuals, are reliable in non-linear models also (Kalliovirta 2006). They exist for any fully specified parametric model with continuous CDF $F(\boldsymbol{\theta}_0, \mathbf{Y})$. Here $\boldsymbol{\theta}_0$ is the true parameter value that generates the observed $T \times 1$ vector $\mathbf{Y}$. The definition of uniformly distributed quantile residual for each $t$ is $U_t = F_{t-1}(\boldsymbol{\theta}_0, Y_t)$, where $F_{t-1}$ is the conditional CDF implied by the model. Similarly, the normally distributed quantile residual is $R_t = \Phi^{-1}(U_t)$, where $\Phi^{-1}(\cdot)$ is the inversed CDF of the standard normal distribution. If the estimated model is correctly specified and $\widehat{\boldsymbol{\theta}}_T$ is a consistent estimator for $\boldsymbol{\theta}_0$, then vector of $\hat{U}_t = F_{t-1}(\widehat{\boldsymbol{\theta}}_T, Y_t)$s (or equivalently, $\hat{R}_t$s) are asymptotically i.i.d. This implies that the hypothesis of a correct specification and properties of quantile residuals are conveniently connected, which makes quantile residuals useful in model evaluation.

Bai (2003) generalizes the Kolmogorov-Smirnov test by applying the Khmaladze's martingale transformation to remove the effect of parameter estimation in the empirical process based on $\hat{U}_t$s. Duan (2003) considers four different tests based on moments of modified $\hat{R}_t$s and removes the effect of parameter estimation by including first order approximations in the test statistics. A Cramér-von Mises type test statistics of Hong and Li (2005) use nonparametric estimates of the density function of $\hat{U}_t$s. Kalliovirta (2006) uses first order approximations to correct the effect of parameter estimation in three LM type tests based on moments of $\hat{R}_t$s.

Henceforth, we denote the test of Bai (2003) with $KS$, the tests of Duan (2003) with $D_i$, $i = 1, 2, 3, 4$, the tests of Hong and Li (2005) with $HL_i$, $i = 1, ..., 20$, and the tests of Kalliovirta (2006) with $N$, $A$, and $H$[1]. In addition, the results of Bai (2003) allow us to introduce alternative test statistics based on the transformed empirical process $\hat{W}_T(r)$,[2] $r \in [0, 1]$, of $\hat{U}_t$s. Thus, we consider also the following statistics: Anderson-Darling type test $AD_{c,d} = \sup_{c \le r \le d} \left| \hat{W}_T(r) \right|$ for some $c < d \in (0, 1)$; Cramér-von Mises type test $CM_1 = \int_0^1 \left| \hat{W}_T(r) \right| dr$ and $CM_2 = \int_0^1 \left| \hat{W}_T(r) \right|^2 dr$; and Pearson's goodness-of-fit type test $S_l = \sum_{i=1}^{l} \left[ \hat{W}_T(r_i) - \hat{W}_T(r_{i-1}) \right]^2 / (r_i - r_{i-1})$ with $0 < r_1 < \cdots < r_{l-1} < 1$.[3] Furthermore, we compute the standard, (non-transformed) empirical process based, Pearson's goodness-of-fit test $P_l$.

[1] The abbreviations indicate the type of possible misspecifications: $N$ for non-normality, $A$ for autocorrelation, and $H$ for heteroscedasticity.

[2] For details on $\hat{W}_T(r)$, see Bai (2003), page 533.

[3] Bai (2003) shows that $\hat{W}_T(r)$ converges weakly to a standard Brownian motion. Thus, $S \xrightarrow{d} \chi_l^2$. The critical values of the other statistics need to be simulated. For example, 5% level critical values, computed using $10^5$ replications with sample of $10^5$ observations, are 2.24 for $KS$, 2.00 for $AD_{0.2,0.8}$, 1.14 for $CM_1$, and 1.67 for $CM_2$ tests.

# 3 Models

Assume that $\varepsilon_t \sim n.i.d.(0,1)$ are independent of $\eta_t \sim n.i.d.(0,1)$, and $\mathsf{I}_A$ is the indicator function of a set $A$. We generate data using models:

1) N(0,1), $Y_t = \varepsilon_t$;

2) GARCH(1,1), $Y_t = 0.52 + \sigma_t \varepsilon_t$ with $\sigma_t = 0.04 + 0.05 y_{t-1}^2 + 0.82 \sigma_{t-1}^2$;

3) MAR(3,1,0),

$Y_t = (0.50 + 0.30 Y_{t-1}) \mathsf{I}_{\eta_t < 0} + (1.75 + 0.60 Y_{t-1}) \mathsf{I}_{0 < \eta_t \leq 1} + (3.0 + 0.85 Y_{t-1}) \mathsf{I}_{\eta_t > 1} + \varepsilon_t$;

and

4) MAR(3,1,0)-GARCH(1,1),

$Y_t = 0.24 \mathsf{I}_{\eta_t < -0.5} + 1.57 \mathsf{I}_{-0.5 < \eta_t \leq 0.75} + 3.14 \mathsf{I}_{\eta_t > 0.75} + 0.83 Y_{t-1} + \sigma_t \varepsilon_t$

with $\sigma_t = 0.06 + 0.16 y_{t-1}^2 + 0.82 \sigma_{t-1}^2$.

Distribution of $Y_t$ generated by a MAR model can be positively or negatively skewed, peaked or flat, and multi-modal. For more details, see Kalliovirta (2006).

In power comparisons, we also estimate submodels of MAR(3,1,0):

a) MAR(3,0,0), $Y_t = \mu_1 \mathsf{I}_{\eta_t < c_1} + \mu_2 \mathsf{I}_{c_1 < \eta_t \leq c_2} + \mu_3 \mathsf{I}_{\eta_t > c_2} + \varepsilon_t$;

b) MAR(2,1,0), $Y_t = (\mu_1 + \phi_1 Y_{t-1}) \mathsf{I}_{\eta_t < c_1} + (\mu_2 + \phi_2 Y_{t-1}) \mathsf{I}_{c_1 \leq \eta_t} + \varepsilon_t$;

and

c) MAR(2,0,0), $Y_t = \mu_1 \mathsf{I}_{\eta_t < c_1} + \mu_2 \mathsf{I}_{c_1 \leq \eta_t} + \varepsilon_t$.

# 4 Simulations

Tables 1, 2, and 3 report size and power of the tests at 5% *nominal* level. Thus, we make no size corrections in misspecified models. The sample sizes vary from 100 to 3000, and results base on 2000 replications. We obtained the MLEs of the

**Table 1:** Size and power at 5% level.

| | Size | | | Size | | Power | | | |
|---|---|---|---|---|---|---|---|---|---|
| | N(1,0) | | | GARCH(1,1) | | GARCH(1,1) data | | | |
| | | | | | | N($\mu$,$\sigma^2$) estimated | | | |
| *Sample* | 500 | 1000 | 2000 | 500 | 1000 | 500 | 1000 | 2000 | 3000 |
| $KS^{(a)}$ | **16.0** | 9.8 | 5.1 | **15.1** | 9.7 | 17.9 | 13.8 | 10.8 | 8.7 |
| $AD^{(a)}_{0.2,0.8}$ | 0.6 | 0.5 | 0.1 | 0.1 | 0.4 | 0.6 | 0.3 | 0.2 | 0.2 |
| $CM^{(a)}_1$ | 0.4 | 0.4 | 0.1 | 0.7 | 0.3 | 0.5 | 0.2 | 0.2 | 0 |
| $CM^{(a)}_2$ | 0.8 | 0.6 | 0.2 | 1.3 | 0.5 | 0.1 | 0.7 | 0.4 | 0.1 |
| $S^{(a),(b)}_l$ | **26.2** | **22.4** | **17.9** | **25.6** | **19.4** | 30.0 | 26.9 | 24.7 | 20.8 |
| $P^{(b)}_l$ | 1.7 | 2.4 | 2.1 | 2.6 | 2.3 | 2.0 | 1.9 | 2.7 | 2.1 |
| $D_1$ | 7.5 | 7.1 | 6.4 | 0 | 0 | 0 | 0 | 1.4 | 4.3 |
| $D_2$ | 5.4 | 6.1 | 6.0 | 2.7 | 2.4 | 9.7 | 9.2 | 11.1 | 13.3 |
| $D_3$ | 5.7 | 6.0 | 5.2 | 5.3 | 4.9 | 5.3 | 4.8 | 6.5 | 6.4 |
| $D_4$ | 5.3 | 5.2 | 4.8 | 3.6 | 3.3 | 6.7 | 10.3 | 21.8 | 36.4 |
| $HL_1$ | 4.8 | 5.4 | 4.9 | 3.7 | 3.2 | 5.4 | 6.1 | 9.6 | 12.8 |
| $HL_2$ | 4.4 | 4.8 | 5.6 | 4.0 | 4.4 | 5.5 | 6.7 | 8.9 | 10.1 |
| $N$ | 5.4 | 5.4 | 5.4 | 6.0 | 4.8 | 8.1 | 9.6 | 11.2 | 14.4 |
| $A^{(c)}_1$ | 4.3 | 4.9 | 4.6 | 4.8 | 4.5 | 7.2 | 6.5 | 6.2 | 7.3 |
| $H^{(c)}_3$ | 5.8 | 5.6 | 6.0 | 5.7 | 5.9 | 37.4 | 61.7 | 86.3 | **94.1** |

NOTES: (a) We use function $\dot{g}(r,\widehat{\boldsymbol{\theta}}_T) = (1, -\Phi^{-1}(r,\widehat{\boldsymbol{\theta}}_T), 1 - \Phi^{-2}(r,\widehat{\boldsymbol{\theta}}_T))'$, given in Bai(2003), to compute the Khmaladze's martingale transformation.

(b) Bandwidth $3.49\widehat{std}(\hat{U}_t)/T^{1/3}$ decides the number of classes $l$ (Scott 1979).

(c) The subscript signifies the number of lags employed in the test statistics.

**Table 2:** MAR(3,1,0) data, size and power at 5% level.

| | Size | | Power | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAR(3,1,0) | | MAR(2,1,0) | | MAR(3,0,0) | | MAR(2,0,0) | | N($\mu$,$\sigma^2$) | |
| *Sample* | 500 | 1000 | 500 | 1000 | 250 | 500 | 100 | 250 | 100 | 250 |
| $KS^{(d)}$ | **66.0** | **54.5** | 51.6 | 40.2 | 30.9 | 39.2 | 46.2 | 69.4 | 66.1 | 75.2 |
| $AD^{(d)}_{0.2,0.8}$ | **17.6** | **10.7** | 7.5 | 4.7 | 17.9 | 30.4 | 26.0 | 56.3 | 2.1 | 2.5 |
| $S_l^{(b),(d)}$ | **76.9** | **67.0** | 77.8 | 82.5 | 30.4 | 33.6 | 49.6 | 66.6 | 84.1 | 97.0 |
| $P_l^{(b)}$ | 0.4 | 0.3 | 4.9 | 16.5 | 0.7 | 0.7 | 1.0 | 3.1 | 11.2 | 50.9 |
| $D_1$ | 0 | 0 | 0 | 0 | 0.1 | 0 | 0.4 | 0.5 | 2.2 | 1.3 |
| $D_2$ | 2.0 | 2.6 | 3.1 | 3.8 | 4.6 | 5.9 | 7.5 | 20.7 | 26.4 | 41.9 |
| $D_3$ | 5.0 | 5.1 | 4.8 | 4.6 | 16.4 | 23.7 | 16.5 | 32.9 | 28.8 | 81.0 |
| $D_4$ | 5.1 | 4.5 | 6.0 | 4.6 | 4.4 | 6.0 | 4.5 | 6.7 | 3.9 | 9.2 |
| $HL_1$ | 0.3 | 0.5 | 12.3 | 40.3 | **99.1** | **100** | 63.3 | **99.4** | 86.6 | **100** |
| $HL_2$ | 0.9 | 0.9 | 10.3 | 28.1 | 23.6 | 59.4 | 10.4 | 37.8 | 45.9 | **91.4** |
| $N$ | 4.2 | 3.8 | 61.4 | **94.6** | 5.0 | 12.3 | 6.6 | 32.7 | 64.9 | **99.0** |
| $A_1^{(c)}$ | 5.3 | 5.4 | 15.6 | 20.7 | **100** | **100** | **99.6** | **100** | **99.4** | **100** |
| $H_1^{(c)}$ | 4.5 | 5.8 | 3.7 | 4.3 | 80.8 | **98.2** | 34.6 | 83.0 | 58.5 | **91.8** |

NOTES: (b) and (c) See Table 1.

(d) We use an estimated function $\dot{\bar{g}}_T(r, \hat{\boldsymbol{\theta}}_T)$ in the Khmaladze's martingale transformation.

parameters using cml package in GAUSS.

The behaviour of the tests classifies them roughly into three groups. One group is formed by the tests $KS$, $AD$, $CM_1$, $CM_2$, $S_l$, and $P_l$. These tests are unreliable in size and exhibit no (against GARCH in Table 1) or occational power (Tables 2

**Table 3:** MAR(3,1,0)-GARCH(1,1) data, size and power at 5% level.

| | Size of | | | Power | | | Power | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAR(3,1,0)-GARCH(1,1) | | | MAR(3,1,0) | | | GARCH(1,1) | | |
| *Sample* | 500 | 1000 | 2000 | 500 | 1000 | 2000 | 100 | 250 | 500 |
| $KS^{(d)}$ | **11.0** | 6.3 | 3.3 | 43.0 | 29.3 | 20.9 | 33.5 | 25.1 | 26.6 |
| $AD^{(d)}_{0.2,0.8}$ | 0.7 | 0.5 | 0.1 | 7.7 | 4.2 | 4.7 | 15.7 | 12.7 | 17.9 |
| $S_l^{(b),(d)}$ | **17.9** | **13.5** | 7.8 | 58.5 | 55.1 | 53.8 | 44.1 | 50.4 | 72.1 |
| $P_l^{(b)}$ | 1.0 | 1.4 | 1.2 | 2.7 | 5.7 | 14.9 | 20.2 | 40.7 | 72.5 |
| $D_1$ | 0 | 0 | 0 | 0 | 0.1 | 0.5 | 3.8 | 0.5 | 0.1 |
| $D_2$ | 1.0 | 1.8 | 1.2 | 9.0 | 15.6 | 19.2 | 10.5 | 22.2 | 43.6 |
| $D_3$ | 4.9 | 4.6 | 4.5 | 5.1 | 6.9 | 5.7 | 76.7 | **95.9** | **98.2** |
| $D_4$ | 4.0 | 4.7 | 4.5 | 28.8 | 50.1 | 77.8 | 10.0 | 23.8 | 46.8 |
| $HL_1$ | 1.6 | 2.7 | 3.9 | 34.1 | 75.8 | **97.6** | **99.9** | **100** | **100** |
| $HL_2$ | 2.4 | 2.8 | 2.6 | 30.8 | 71.7 | **96.5** | **91.4** | **100** | **100** |
| $N$ | 3.2 | 2.9 | 3.6 | 13.6 | 31.5 | 57.9 | 10.2 | 27.5 | 42.4 |
| $A_1^{(c)}$ | 4.9 | 5.0 | 5.1 | 11.5 | 11.2 | 13.1 | **100** | **100** | **100** |
| $H_1^{(c)}$ | 4.1 | 5.5 | 5.6 | **94.3** | **100** | **100** | 3.8 | 11.7 | 22.6 |

NOTES: See Tables 1 and 2.

and 3)[4]. This power is, however, exaggerated in $KS$ and $S_l$ by their oversizeness. Overall, a comparison to $P_l$ shows that the Khmaladze's martingale transformation provides no improvement.

The tests of Duan (2003) comprise another group that has adequate size in a linear model (Table 1), but are undersized, especially $D_1$, in nonlinear models. Further, these tests are rather powerless, especially $D_1$ and $D_2$. In the largest

[4] Behavior of $AD$, $CM_1$, and $CM_2$ is identical. Thus, we only report $AD$ in Tables 2 and 3.

samples $D_3$ and $D_4$ detect the misspecifications in dependence structure of the mean and variance. In contrast, the misspecification in the number of regimes (Table 2) remains undetected.

Finally, the tests $HL_1$, $HL_2$, $N$, $A$, and $H$ comprise the last group, in which both size and power properties are adequate. Although sizes are unadjusted, we can conclude that $H_3$ has superior power to $HL_1$[5] in detecting GARCH (Table 1). In practice one can only use nominal levels. Therefore, one should prefer tests that are reliable in size and exhibit best power at nominal levels. In this sense, we may conclude that the tests of Kalliovirta (2006) outperform. In addition, these tests are able to give hints where misspecification lies. For example, $H_3$ indicates heteroscedasticity in Tables 1 and 3, $N$ wrong distribution in Table 2, and $A$ autocorrelation in Tables 2 and 3. In comparison, the test $HL_1$ detects misspecification in autocorrelation structure, but less so if misspecification is in conditional heteroscedasticity or in distribution. Furhermore, $HL_1$ implies no hints why a model is rejected.

# 5 Conclusions

In our simulations, the tests based on Khmaladze's martingale transformation, including the test of Bai (2003), are unreliable. The moment based tests of Duan (2003) exhibit undersizeness and insufficient power. The non-parametric tests of Hong and Li (2005) are sometimes undersized and lack power against GARCH, for example. The LM type tests of Kalliovirta (2006) have accurate size and best power at nominal levels against the misspecifications considered.

[5] We computed all $HL_i$, $i = 1, ..., 20$.They were identical in size and $HL_1$ always had the best power among them.

**Acknowledgements** *The author thanks Pentti Saikkonen, Markku Lanne, Jukka Nyblom, and Bent Nielsen for helpful comments that improved this work. The author is responsible for any remaining errors. The author is grateful to the Academy of Finland, the Yrjö Jahnsson foundation, the Okobank Group Research Foundation, and the Finnish Foundation for Advancement of Securities Markets for financial support.*

## References

Bai, J.: 2003, Testing parametric conditional distributions of dynamic models, *The Review of Economics and Statistics* **85**(3), 531–549.

Duan, J.-C.: 2003, A specification test for time series models by a normality transformation. working paper, University of Toronto.

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

Kalliovirta, L.: 2006, Misspecification tests based on quantile residuals. HECER Discussion Paper, No 124.

Scott, D. W.: 1979, On optimal and data-based histograms, *Biometrika* **66**(3), 605–610.