Statistica Sinica **21** (2011), 1807-1829
doi:http://dx.doi.org/10.5705/ss.2009.090

SCORE BASED GOODNESS-OF-FIT TESTS FOR TIME SERIES

Shiqing Ling and Howell Tong

Hong Kong University of Science & Technology
and London School of Economics and Political Science

Abstract: This paper studies a class of tests useful for testing goodness of fit of a wide variety of time series models. These tests are based on a class of empirical processes marked by certain scores. Major advantages of these tests are that they are easy to implement, require only weak conditions that are usually satisfied in practical applications, the relevant critical values are readily available without bootstrap, and are more powerful than the Ljung-Box test, the Li-Mak test and the Koul-Stute test in all the cases we have tried. A comparison with the Fan-Zhang test is included. We also extend the class of tests to include score-like statistics.

Key words and phrases: Empirical process, goodness-of-fit test, nonlinear time series, score, time series models.

1. Introduction

Let $\{y_t : t = 0, \pm 1, \pm 2, \ldots\}$ be a strictly stationary sequence of real-valued random variables defined on the probability space $(\Omega, \mathcal{F}, P)$; let $\mathcal{F}_t$ be the $\sigma-$field generated by $\{y_t, y_{t-1}, \cdots\}$. We assume that the mean and the variance of y_t are both finite. Suppose we wish to summarize the information contained in a finite number of observations of the time series by building a parametric model. Now, goodness-of-fit tests constitute an essential stage in parametric modelling and there are numerous tests available in the literature, see, e.g., Li (2004) for a fairly comprehensive account. Formally, we postulate that the time series is generated by the model

$$y_t = \mu_t(\theta) + \eta_t \sqrt{h_t(\theta)}\,, \tag{1.1}$$

where $\theta \in \Theta, \Theta$ being a proper subset of the $p-$dimensional Euclidean space, $\mu_t(\theta)$ and $h_t(\theta)$ are, respectively, the conditional mean function and the conditional variance function of y_t given $\mathcal{F}_{t-1}$, $\{\eta_t, t = 0, \pm 1, \ldots\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with a common distribution F, zero mean and unit variance, and η_t is independent of $\mathcal{F}_{t-1}$. Note that sometimes θ is called the nuisance parameter with its true value denoted by θ_0.

Within such a framework, a goodness-of-fit test tests the (composite) null hypothesis that the given data follow model (1.1). The alternative hypothesis is merely that the null hypothesis does not hold; the test is sometimes called an omnibus test or a portmanteau test, accordingly. Given an alternative parametric model other tests would be more relevant, e.g. the likelihood ratio test. The likelihood ratio test can be generalised to cover the case of an alternative non-parametric model such as

$$y_t = \mu_t(\cdot) + \eta_t \sqrt{h_t(\cdot)}\,, \tag{1.2}$$

where $\mu_t(\cdot)$ and $h_t(\cdot)$ are respectively the conditional mean function and the conditional variance function of y_t given $\mathcal{F}_{t-1}$, $\{\eta_t, t = 0, \pm 1, \ldots\}$ is as defined above. With suitable constructions, χ^2-asymptotics can be retained, in which the degrees of freedom tend to infinity as the sample size increases to infinity. (Fan and Yao (2003)). Fan and Zhang (2004) introduced some important developments with this approach; for implementation, they resorted to bootstrapping for the critical values. For likelihood ratio tests, generalised or not, the case of a mis-specified alternative hypothesis remains challenging.

Todate, many of the goodness-of-fit tests in time series are *residual-based*. For example, the classic portmanteau test of Box and Pierce (1970) and its improvement by Ljung and Box (1978) are based on the sample autocorrelations of the residuals. In the context of goodness of fit of nonlinear time series models, the McLeod and Li test (1983) and the Li and Mak test (1994) are based on the sample autocorrelations of the squared residuals. Based on a generalized spectral approach of the residuals, Hong and Lee (2003) and Escanciano (2008) proposed some new diagnostic tests for model (1.1). The former focuses on the independence assumption of the 'noise process'; the latter requires us to approximate the critical values by bootstrap, but allows non-i.i.d. variables and checks for many lags in $\mathcal{F}_{t-1}$. More recently, perhaps influenced by the empirical distribution function approach in the goodness-of-fit test for independent observations, substantial developments for time series data have taken place in the form of tests based on empirical processes marked by certain *residuals*, see, e.g., Stute (1997), Koul and Stute (1999), Stute et al. (2006), and Escanciano (2007). Of course, the use of marked empirical processes in hypothesis testing in time series has a longer history; see, e.g., An and Cheng (1991). In all these developments, residuals play a pivotal role.

In the 1980s, unification of numerous classical goodness-of-fit tests, such as those developed by Quenouille (1947, 1949), Walker (1950, 1952), Bartlett and Diananda (1950), as well as some of the later ones such as the Box-Pierce test, was achieved by the observation that a Lagrangian multiplier (LM) test with an appropriately chosen alternative hypothesis results in a test that is the

large-sample equivalent of the above goodness-of-fit tests. For example, Newbold (1980) showed that the LM test for an $ARMA(p, q)$ model against the alternative of an $ARMA(p+m, q)$ model is asymptotically equivalent to a goodness-of-fit test based on the first m sample autocorrelations of the residuals. For more details of the unification, see, e.g., Hosking (1978, 1980) and Godfrey (1979). For extension of the idea to tests for linearity see, e.g., Tong (1990, Sec. 5.3). This unification is significant. It suggests that a *score* (statistic) may be a fundamentally more useful pivot in the construction of goodness-of-fit tests since an LM test uses the score evaluated under the null hypothesis. This paper develops goodness-of-fit tests for time series that are based on empirical processes marked by certain scores; we later generalise the approach to include their equivalents.

This paper is organized as follows. Section 2 gives the generic form of the test statistic, then gives explicit expressions for various commonly used models. Section 3 presents the null distribution, with critical values evaluated, and studies the local power of our tests. Section 4 reports simulation results, including some comparative studies. Section 5 illustrates our approach with the Hang Seng Index. Section 6 draws some conclusions.

2. Score-Based Empirical Process Approach to Goodness-of-Fit Tests

Given observations $\{y_1, \ldots, y_n\}$ from model (1.1) and initial values $\{y_s : s \leq 0\}$, let $\hat{\theta}_n$ denote the maximum likelihood estimator of θ_0 under H_0.

Assumption 1.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \Sigma^{-1} \sum_{t=1}^{n} \frac{D_t(\theta_0)}{\sqrt{n}} + o_p(1), \tag{2.1}$$

where $D_t(\theta_0)$ is the score of θ evaluated at θ_0 and $\Sigma = E[D_t(\theta_0)D_t'(\theta_0)]$, the information matrix.

This is a mild condition for maximum likelihood estimation of parameters in time series models and is generally satisfied under standard conditions. We return to this point later.

Let $I\{B\}$ denote the indicator function of the event B. Our test statistic is based on the score-marked empirical process

$$T_n(x, \theta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} D_t(\theta) I\{y_{t-1} \leq x\}. \tag{2.2}$$

Since θ_0 is usually not fully specified in practice we replace it by $\hat{\theta}_n$, noting Assumption 1, and study $T_n(x, \hat{\theta}_n)$. Let $\Sigma_x = E[D_t(\theta_0)D_t'(\theta_0)I\{y_{t-1} \leq x\}]$ and $A = \inf\{x : \Sigma = \Sigma_x\}$. When y_t has a support on R, then we have $A =$

∞, generally. Let $\hat{\Sigma}_{nx} = \sum_{t=1}^{n}[D_t(\hat{\theta}_n)D_t'(\hat{\theta}_n)I\{y_{t-1} \le x\}]/n$ and $\hat{\Sigma}_n$ be the estimators of Σ_x and Σ, respectively, where $\hat{\Sigma}_n = \hat{\Sigma}_{nA}$. We define our generic test statistic via a linear transformation of $T_n(x, \hat{\theta}_n)$ as

$$S_n^a = \max_{a \le x \le A} \frac{[\beta' \hat{\Sigma}_{nx}^{-1} T_n(x, \hat{\theta}_n)]^2}{\beta'(\hat{\Sigma}_{na}^{-1} - \hat{\Sigma}_n^{-1})\beta}, \tag{2.3}$$

where β is a nonzero $p \times 1$ constant vector. When $p = 1$, S_n^a is equivalent to the weighted LR-test for $H_0 : y_t = \mu_t(\theta) + \varepsilon_t$ with Gaussian white noise $\{\varepsilon_t\}$ against the alternative

$$y_t = \mu(\theta, y_{t-1}) + \mu(\theta_1, y_{t-1})I\{y_{t-1} \le x\} + \varepsilon_t, \tag{2.4}$$

with weight $(1-\Sigma\Sigma_x^{-1})/(1-\Sigma\Sigma_a^{-1})$, where $\theta_1 \in \Theta$ is another unknown parameter. This connection is similar to the situation pertaining to the classic goodness-of-fit tests in time series mentioned earlier. For general p, we can replace the threshold variable y_{t-1} in $I\{y_{t-1} \le x\}$ by y_{t-r} or by $\hat{\theta}_n'(y_{t-1}, \ldots, y_{t-p})'$ as in Stute et al. (2006). In fact, we can replace it by any function $\xi_{t-1} = g(y_{t-1}, y_{t-2}, \ldots)$ and our theory still holds as long as ξ_{t-1} has a positive conditional density given $\{y_{t-2}, y_{t-3}, \ldots\}$; see Ling and Tong (2006). This assumption is usually satisfied. We have chosen y_{t-1} to keep the procedure simple, in the absence of a general theory for an optimal choice. Our approach can be extended to multivariate time series models with y_{t-1} replaced by a suitable choice of ξ_{t-1}.

Typically, the quantity a is taken as an early quantile of the process values. It should, however, ensure that $\hat{\Sigma}_{na}^{-1}$ exists. Unlike Chan (1991), the limiting distribution of S_n^a does not depend on the choice of a since the weight function cancels out the related component. Note that $\max_\beta S_n^a$ does not have a limiting distribution as simple as that of S_n^a; the latter is described by Theorem 2 in Section 3. Note also that S_n^a is invariant with respect to $\|\beta\|$. If we denote the normalized score $\hat{\Sigma}_{nx}^{-1}T_n(x, \hat{\theta}_n)$ by $U_n(x) = (u_1(x), \ldots, u_p(x))'$, then $\beta' U_n(x) = \sum_{i=1}^{p} \beta_i u_i(x)$ can be interpreted as a weighted score, and each $u_i(x)$ is the marked-score along the direction of the $i-$th coordinator in θ. The optimal choice of β remains an open problem. A simple choice for β is $(1, \ldots, 1)'$, which means that we attach equal weight to each $u_i(x)$. The simulation in Section 4 suggests that this choice together with a around the $5p\%$ quantile of data produces good power. Another choice is $\beta = \hat{\theta}_n$, but the simulation results in Section 4 suggest that this is not as good.

When the alternative of model (1.1) is its threshold counterpart, $T_n(x, \hat{\theta}_n)$ is precisely the score function in the LR test. However, the LR test is a quadratic form of $T_n(x, \theta)$ and its limiting distribution is a functional of a Brownian bridge with a complicated covariance matrix. In this case, except for AR models in Chan

(1990, 1991) and Chan and Tong (1990), we have to use a simulation method to obtain its critical values case by case; see, e.g., Wong and Li (1997, 2000). We should mention that the initial values $\{y_s : s \le 0\}$ are typically not available and are usually replaced by some constants. However, for most stationary models such as ARMA or GARCH models, the initial values do not affect the asymptotic properties of the estimated parameters or our test $S_n^{(a)}$; see Assumption 4 of Hong and Lee (2003).

We can always construct the test S_n^a because we have $D_t(\theta)$ from the model fitting stage. Since S_n^a is generally just the maximum of n different numbers, it is as easy to implement as the Ljung-Box test and the McLeod-Li test.

Example 1. Consider the double AR (DAR) model

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \eta_t \sqrt{\omega + \sum_{i=1}^{p} \alpha_i y_{t-i}^2},$$

where $\omega, \alpha_i > 0$, $t \in \{-p, \ldots, 0, 1, 2, \ldots\}$, and $\{\eta_t\}$ is an independent random sequence with $\eta_t \sim N(0,1)$. Here, $\theta = (\theta_1', \theta_2')'$ with $\theta_1 = (\phi_1, \ldots, \phi_p)'$ and $\theta_2 = (\omega, \alpha_1, \ldots, \alpha_p)'$. We take $\hat{\theta}_n$ as the MLE of θ_0. Under conditions for strict stationarity, Ling (2004, 2007) shows that (2.1) holds and

$$D_t(\theta) = \Big\{ \frac{Y_{1t-1}' \varepsilon_t(\theta)}{\theta_2' Y_{2t-1}}, -\frac{Y_{2t-1}'}{2\theta_2' Y_{2t-1}} \Big[1 - \frac{\varepsilon_t^2(\theta)}{\theta_2' Y_{2t-1}} \Big] \Big\}',$$

where $\varepsilon_t(\theta) = y_t - \theta_1' Y_{1t-1}$, $Y_{1t} = (y_t, \ldots, y_{t-p+1})'$ and $Y_{2t} = (1, y_t^2, \ldots, y_{t-p+1}^2)'$. The expansion (2.1) holds since the information matrix is Σ in maximum likelihood estimation.

In some applications, instead of MLE, practitioners may prefer to use least squares estimation, or quasi-Gaussian MLE. For these, the test statistic S_n^a can still apply provided the score $D_t(\theta)$ is replaced by the derivative of a relevant loss function, the specific form of which is usually obvious as we show in the following examples. By an abuse of notation, we denote this derivative also by $D_t(\theta)$. Note that the information matrix is then usually of the form $E[D_t(\theta) D_t'(\theta)]/\gamma$, where γ is a positive constant, the exact value of which depends on the method of estimation, as also shown in the examples below. Thus, Assumption 1 remains essentially the same as stated, but with Σ replaced by Σ/γ.

Example 2. Consider the ARMA (p, q) model

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{i=1}^{q} \psi_i \varepsilon_{t-1} + \varepsilon_t,$$

where ε_t's are i.i.d. with mean zero and a finite variance σ^2. Here, $\theta = (\phi_1, \ldots, \phi_p, \psi_1, \ldots, \psi_q)'$. Under the usual stationarity and invertibility conditions (e.g., Weiss (1986), the conditional LSE, $\hat{\theta}_n$, of θ_0 satisfies the expansion (2.1) with $\gamma = \sigma^2$, and $D_t(\theta) = [\partial\varepsilon_t(\theta)/\partial\theta]\varepsilon_t(\theta)$, where $\varepsilon_t(\theta) = \psi^{-1}(B)\phi(B)y_t$. In particular, for the AR(2) model ($p = 2$ and $q = 0$), we have $D_t(\theta) = (y_{t-1}, y_{t-2})'(y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2})$.

Example 3. The TAR(p, q) model is

$$y_t = I\{y_{t-d} \le r\}(\phi_{10} + \sum_{i=1}^{p} \phi_{1i} y_{t-i}) + I\{y_{t-d} > r\}(\phi_{20} + \sum_{i=1}^{p} \phi_{2i} y_{t-i}) + \varepsilon_t,$$

where $\max_{i=1,2} \sum_{j=1}^{p} |\phi_{ij}| < 1$ and ε_t's are i.i.d. with $E\varepsilon_t^4 < \infty$. Assume that the AR function is discontinuous and $d > 0$ is a known integer. Here, $\theta = (\phi_{10}, \phi_{11}, \ldots, \phi_{1p}, \phi_{20}, \phi_{21}, \ldots, \phi_{2p})'$. Let $(\hat{\theta}_n, \hat{r}_n)$ be the LSE of (θ_0, r_0), where r_0 is the true value of r. From Chan (1993), we have $n(\hat{r}_n - r_0) = O_p(1)$ and (2.1) holds with $\gamma = \sigma^2$,

$$D_t(\theta) = \tilde{D}_t(\theta, r_0) \text{ and } \tilde{D}_t(\theta, r) = [Y'_{t-1} I\{y_{t-d} \le r\}, Y'_{t-1} I\{y_{t-d} > r\}]' \varepsilon_t(\theta, r),$$

where $\varepsilon_t(\theta, r) = y_t - I\{y_{t-d} \le r\}(\phi_{10} + \sum_{i=1}^{p} \phi_{1i} y_{t-i}) - I\{y_{t-d} > r\}(\phi_{20} + \sum_{i=1}^{p} \phi_{2i} y_{t-i})$ and $Y_{t-1} = (1, y_{t-1}, \ldots, y_{t-p})'$. We can show that $\sum_{t=1}^{n} \|\tilde{D}_t(\hat{\theta}_n, \hat{r}_n) - D_t(\hat{\theta}_n)\|/\sqrt{n} = o_p(1)$. Thus, S_n^a has asymptotically the same distribution when r_0 is replaced by $\hat{r}_n$.

Example 4. Consider the GARCH(r, s) model

$$y_t = \eta_t \sqrt{h_t} \text{ and } h_t = \alpha_0 + \sum_{i=1}^{r} \alpha_i y_{t-i}^2 + \sum_{i=1}^{s} \beta_i h_{t-i},$$

where the η_t are i.i.d. with $E\eta_t^2 = 1$ and $E\eta_t^4 < \infty$, $\alpha_0 > 0$, $\alpha_i's \ge 0$, $\alpha_r \ne 0$, $\beta_j's \ge 0$, and $\beta_s \ne 0$. Here, $\theta = (\alpha_0, \alpha_1, \ldots, \alpha_r, \beta_1, \cdots, \beta_s)'$. Let $\hat{\theta}_n$ be the quasi-maximum likelihood estimator of θ_0. Under the standard strict stationarity condition, Francq and Zakoïan (2004) show that (2.1) holds with $\gamma = E\eta_t^4 - 1$, and

$$D_t(\theta) = [2h_t(\theta)]^{-1}\Big[\frac{y_t^2}{h_t(\theta)} - 1\Big]\Big[\frac{\partial h_t(\theta)}{\partial\theta}\Big],$$

where $h_t(\theta) = \alpha_0 + \sum_{i=1}^{r} \alpha_i y_{t-i}^2 + \sum_{i=1}^{s} \beta_i h_{t-i}(\theta)$. For the important special case of the GARCH(1,1) model, $h_t(\theta) = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}(\theta)$ and $\partial h_t(\theta)/\partial\theta = [1, y_{t-1}^2, h_{t-1}(\theta)]' + \beta_1 \partial h_{t-1}(\theta)/\partial\theta$.

When extending the GARCH or ARMA model to the ARMA-GARCH model, we can use the MLE to estimate the associated parameters. For the general model

(1.1), we can use the quasi-Gaussian MLE to estimate the parameters. In this case, the score function is $D_t(\theta) = U_t(\theta)\psi(\eta_t(\theta))$, where

$$U_t(\theta) = \left[\frac{1}{\sqrt{h_t(\theta)}}\frac{\partial \mu_t(\theta)}{\partial \theta}, \frac{1}{2h_t(\theta)}\frac{\partial h_t(\theta)}{\partial \theta}\right] \text{ and } \psi(x) = \left[x, x^2 - 1\right]'.$$

We can see that $D_t(\theta_0)$ is a martingale difference and $\Sigma = E[D_t'(\theta_0)D_t(\theta_0)]$ if η_t is symmetric and $E\eta^4 = 3$. Thus, our test can be used.

3. Null Distribution and Local Power

To get the null distribution of S_n^a, we introduce assumptions as follows.

Assumption 2. *$D_t(\theta_0)$ is an $\mathcal{F}_t-$measurable, strictly stationary and ergodic martingale difference with $E(\|D_t(\theta_0)\|^{2(1+\iota)}) < \infty$ for some $\iota > 0$.*

Assumption 3. *$D_t(\theta)$ has the expansion $D_t(\theta) - D_t(\theta_0) = P_t(\theta^*)(\theta - \theta_0)'$ and $EP_t(\theta_0) = \Sigma/\gamma$, where θ^* lies between θ and θ_0, and for any fixed $C > 0$,*

$$\sup_{\sqrt{n}\|\theta-\theta_0\|\le C} \frac{1}{n}\sum_{t=1}^{n} \left\|P_t(\theta) - P_t(\theta_0)\right\| = o_p(1).$$

Here, $P_t(\theta)$ is an information-type matrix and is $[\partial D_t(\theta)/\partial\theta]$ if $D_t(\theta)$ is differentiable. The moment condition in Assumption 2 is minimal; Assumption 3 holds for most of the strictly stationary time series models met in practice. We first give a lemma whose proof is given in the Appendix.

Lemma 1. *Under Assumptions* 1, 2, *and* 3,

(a) $\sup_{x\in R\cup\{\infty\}} \|\hat{\Sigma}_{nx} - \Sigma_x\| = o_p(1)$,

(b) $\sup_{x\in R} \left\|T_n(x,\hat{\theta}_n) - T_n(x,\theta_0) - \frac{\Sigma_x\Sigma^{-1}}{\sqrt{n}}\sum_{t=1}^{n} D_t(\theta_0)\right\| = o_p(1)$.

The weak convergence of $\{T_n(x,\hat{\theta}_n) : x \in R\}$ is a corollary of Theorem 3 of Escanciano (2007), as follows.

First, let $R_\gamma = [\gamma_1, \gamma_2] \subset [x_0, A)$ for some $x_0 \in R$, where A is defined as in (2.3). Let $D[R_\gamma]$ denote the space of real-valued functions on R_γ which are right continuous and have left-hand limits, and let it be equipped with the Skorohod topology as in Billingsley (1968). The weak convergence on $D^p[x_0, A)$ is defined as on $D[R_\gamma] \times \cdots \times D[R_\gamma]$ (p factors) for *any* interval $[\gamma_1, \gamma_2]$, and is denoted by $\Longrightarrow$.

Theorem 1. *Suppose that Assumptions* 1, 2, *and* 3 *hold, and that η_t has a bounded density f in R. If Σ_{x_0} is positive definite for some $x_0 \in R$, then*

$$T_n(x,\hat{\theta}_n) \Longrightarrow G_p(x) \text{ in } D^p[x_0, A)$$

under H_0, where $\{G_p(x) : x \in [x_0, A)\}$ is a $p-$dimensional Gaussian process with mean zero and covariance kernel $K_{xy} = \Sigma_{x\wedge y} - \Sigma_x\Sigma^{-1}\Sigma_y$; almost all paths of $G_p(x)$ are continuous in x.

We first note that $\Sigma_x^{-1}T_n(x,\hat{\theta}_n) \Rightarrow G_{0p}(x)$ in $D^p[x_0, A)$ under H_0, where $\{G_{0p}(x)\}$ is a $p \times 1$ vector Gaussian process on $[x_0, A)$ with mean zero and covariance kernel $K_{xy} = \Sigma_{x\vee y}^{-1} - \Sigma^{-1}$. An important observation is that $\{G_{0p}(x)\}$ has independent increments with $E\{[G_{0p}(x)-G_{0p}(y)][G_{0p}(x)-G_{0p}(y)]'\} = \Sigma_y^{-1} - \Sigma_x^{-1}$ when $x > y$. For marked empirical processes, the covariance kernel usually has the form $\sigma_{x\wedge y} - u_x'\Sigma^{-1}u_y$. For Theorem 1, $\sigma_{x\wedge x} = u_x = \Sigma_x$. This is the key for the process $\{G_{0p}(x)\}$ to have independent increments. For marked processes (such as the residual-marked process) for which $\sigma_{x\wedge x} \neq u_x$, we cannot obtain a process with independent increments after normalization.

Since the components of $G_{0p}(x)$ are dependent, its covariance kernel does not admit a simple transformation and neither does a quadratic form or the maximum of all its components. However, for any constant β, $\beta'G_{0p}(x)$ has the rather simple covariance kernel $\sigma_x \wedge \sigma_y$, where $\sigma_x = \beta'(\Sigma_x^{-1} - \Sigma^{-1})\beta$. For any finite constant $a \in [x_0, A)$, σ_x/σ_a is a continuous and strictly decreasing function in terms of x and runs through $[0, 1]$ when x runs from A to a. Thus, $B(\tau) \equiv \beta'G_{0p}(x)/\sqrt{\sigma_a}$ is a standard Brownian motion on $\tau = \sigma_x/\sigma_a \in [0, 1]$. Let $b \in [a, A)$ be a constant and

$$S_n^a(b) = \max_{a\le x\le b} \frac{[\beta'\hat{\Sigma}_{nx}^{-1}T_n(x,\hat{\theta}_n)]^2}{\beta'(\hat{\Sigma}_{na}^{-1} - \hat{\Sigma}_n^{-1})\beta}.$$

Theorem 1 and the Continuous Mapping Theorem yield the main result.

Theorem 2. *If the assumptions of Theorem 1 hold, then, for any $p \times 1$ nonzero constant vector β, we have*

$$\lim_{b\to A}\lim_{n\to\infty} P\,[S_n^a(b) \le x] = P\left[\max_{\tau\in[0,1]} B^2(\tau) \le x\right]$$

for any $a \in [x_0, A)$ and any $x \in R$, where $B(\tau)$ is a standard Brownian motion on $C[0, 1]$.

From this theorem, the constant C_α such that $P[\max_{\tau\in[0,1]} B^2(\tau) \ge C_\alpha] = \alpha$ can be used as an approximate critical value of S_n^a for rejecting the null H_0 at the significance level α. From Shorack and Wellner (1986, p.34), we have

$$P\Big[\max_{\tau\in[0,1]} B^2(\tau) \ge x\Big] = 1 - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp\Big[-\frac{(2k+1)^2\pi^2}{8x}\Big],$$

for all $x > 0$, and $C_{0.1} = 3.83$, $C_{0.05} = 5.00$, and $C_{0.01} = 7.63$.

We next study the asymptotic local power of S_n^a. Let $r_{1t} = r_1(y_{t-1}, y_{t-2}, \ldots)$ and $r_{2t} = r_2(y_{t-1}, y_{t-2}, \ldots)$ be $\mathcal{F}_{t-1}-$measurable random variables for $t = 0, \pm 1, \ldots$. Consider the local alternative hypothesis

$$y_t = \mu_t(\theta) + \frac{r_{1t}}{\sqrt{n}} + \eta_t \sqrt{h_t(\theta) + \frac{r_{2t}}{\sqrt{n}}}.$$

Assume that η_t is normal and independent of y_s for $s \leq 0$, under both H_0 and H_{1n}. Let $m(x) = E[D_t(\theta_0)\zeta_t I\{y_{t-1} \leq x\}] - \Sigma_x \Sigma^{-1} E[D_t(\theta_0)\zeta_t]$, where $\zeta_t = \eta_t r_{1t}/\sqrt{h_t(\theta_0)} + (1 - \eta_t^2) r_{2t}/h_t(\theta_0)$.

Theorem 3. *If the assumptions of Theorem 2 hold and* $0 < Er_{1t}^2 + Er_{2t}^2 < \infty$ *under* H_0, *then under* H_{1n}, *it follows that*

(a) $T_n(x, \hat{\theta}_n) \Longrightarrow m(x) + G_p(x)$ *in* $D^p[R]$,

(b) $\lim_{b \to A} \lim_{n \to \infty} P\left[S_n^a(b) \leq z\right] = P\left[\max_{\tau \in [0,1]} [u(\tau) + B(\tau)]^2 \leq z\right]$,

for any $z \in R$, *where* $u(\tau) = \beta' \Sigma_x^{-1} m(x)/[\beta'(\Sigma_a^{-1} - \Sigma^{-1})\beta]^{1/2}$ *with* x *such that* $\sigma_x = \tau$, *and* $G_p(x)$ *and* $B(\tau)$ *are defined as in Theorems* 1–2.

This shows that S_n^a has good local power if $u(\tau) \neq 0$; otherwise it has no local power. It is unlikely that $u(\tau) = 0$, unless $\zeta_t = \beta' D_t(\theta_0)$. When n and b are large, we have $P(S_n^a > C_\alpha) \approx P\Big\{ \max_{\tau \in [0,1]} [u(\tau) + B(\tau)]^2 > C_\alpha \Big\} \to 1$ if $\max_{\tau \in [0,1]} |u(\tau)| \to \infty$.

4. Simulation Results

To conduct a theoretical study of the various goodness-of-fit tests in time series would be ideal. However, there are difficulties. First, the composite nature of the alternative hypothesis means that the power is likely to change from one alternative model to the next and there are (uncountably) infinitely many possible models. Second, the distribution of the test statistic (especially for finite samples) is typically unknown or unavailable under the alternative hypothesis, except for trivial cases and for large samples. Third, asymptotic power is usually not practically informative. For example, when testing the goodness of fit of an AR(2) model against the alternative of an AR(3) model, both the Ljung-Box test and our test have power 1 asymptotically.

Therefore, we compare the performances of tests on the basis of simulations. For a summary of similar simulation-based comparative studies of tests in this vein; see, e.g., Li (2004). This section examines the performance of the test statistic S_n^a in finite samples through Monte Carlo experiments. In all the experiments, we take a as the $5p$%-quantile of data $\{y_1, \ldots, y_n\}$ and use 1,000 independent replications.

Table 1. Sizes of S_n^a for Null Hypothesis H_0: ARMA(1,1) model at Significance Level α(1,000 replications).

		$n=100$			$n=200$			$n=400$		
α		0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
ϕ	ψ				$\beta=(1,1)'$					
-0.8	-0.5	0.019	0.041	0.066	0.009	0.029	0.069	0.008	0.052	0.097
-0.5	-0.5	0.008	0.036	0.066	0.010	0.033	0.079	0.013	0.037	0.090
0.0	-0.5	0.009	0.037	0.078	0.006	0.040	0.085	0.009	0.030	0.072
0.8	-0.5	0.015	0.051	0.105	0.009	0.041	0.080	0.012	0.036	0.079
-0.8	0.5	0.007	0.035	0.067	0.009	0.033	0.072	0.007	0.043	0.084
0.0	0.5	0.006	0.031	0.068	0.010	0.043	0.087	0.007	0.037	0.084
0.5	0.5	0.002	0.030	0.071	0.008	0.046	0.085	0.012	0.047	0.093
0.8	0.5	0.013	0.049	0.100	0.015	0.044	0.092	0.013	0.056	0.097
					$\beta=(\hat{\phi}_n,\hat{\psi}_n)'$					
-0.8	-0.5	0.014	0.053	0.093	0.015	0.051	0.096	0.006	0.040	0.090
-0.5	-0.5	0.008	0.029	0.065	0.007	0.041	0.086	0.009	0.043	0.071
0.0	-0.5	0.002	0.015	0.046	0.007	0.038	0.082	0.010	0.032	0.062
0.8	-0.5	0.007	0.029	0.072	0.009	0.031	0.060	0.007	0.028	0.068
-0.8	0.5	0.011	0.038	0.065	0.005	0.033	0.067	0.005	0.038	0.070
0.0	0.5	0.008	0.030	0.072	0.007	0.033	0.074	0.008	0.032	0.081
0.5	0.5	0.005	0.035	0.070	0.009	0.036	0.079	0.012	0.034	0.086
0.8	0.5	0.011	0.049	0.097	0.014	0.053	0.092	0.018	0.057	0.098

We first study the size and the power of S_n^a when the null hypothesis is the ARMA(1,1) model, $y_t = \phi y_{t-1} + \psi\varepsilon_{t-1} + \varepsilon_t$, where ε_t is i.i.d. $N(0,1)$. We take $\beta = (1,1)'$ and $(\hat{\phi}_n, \hat{\psi}_n)'$. For the size, the true parameters are taken to be $(\phi,\psi) = (-0.8,-0.5)$, $(-0.5,-0.5)$, $(0.0,-0.5)$, $(0.8,-0.5)$, $(-0.8,0.5)$, $(0.0,0.5)$, $(0.5,0.5)$ and $(0.8,0.5)$. The sample sizes are $n = 100$, 200 and 400. Table 1 summarizes the results when the significance level α is 0.01, 0.05 and 0.1. It shows that the sizes are fairly close to their nominal values although there is evidence of conservatism.

To study the power of S_n^a, we consider two alternatives:

$$\text{TARMA Model } y_t = 0.5y_{t-1} + 0.5\varepsilon_{t-1} - \theta(y_{t-1} + \varepsilon_{t-1})I\{y_{t-1} \le 0\} + \varepsilon_t,$$
$$\text{BL Model } y_t = 0.5y_{t-1} + 0.5\varepsilon_{t-1} - \theta y_{t-2}\varepsilon_{t-1} + \varepsilon_t.$$

The first is an example of the threshold ARMA models proposed by Tong (1978, 1990), while the second is an example of the bilinear models (or BL models, for short); see, e.g., Granger and Andersen (1978). We first compare our tests with two commonly used tests, namely the Ljung-Box $Q_n(m)$ test and the Li-Mak $Q_n^2(m)$ test. We take $\theta = 0.1$, 0.2, 0.3, 0.4 and 0.5 and compare the power of S_n^a with $Q_n(m)$ and $Q_n^2(m)$ at level 0.05 when $n = 100$, 200 and 400. The results are reported in Table 2 when $m = 6$ and $\beta = (1,1)'$, and in Table 3 when $m = 12$

Table 2. Powers of S_n^a, $Q_n(m)$ and $Q_n^2(m)$ for Null Hypothesis H_0: ARMA(1,1) Model at Significance Level 0.05 [$\beta = (1,1)'$ and 1,000 replications].

θ		0.0	0.1	0.2	0.3	0.4	0.5
				H_1: TARMA Model			
	S_n^a	0.026	0.081	0.146	0.419	0.681	0.884
$n = 100$	$Q_n(6)$	0.053	0.051	0.056	0.070	0.082	0.102
	$Q_n^2(6)$	0.027	0.028	0.027	0.024	0.029	0.038
	S_n^a	0.035	0.118	0.403	0.780	0.985	0.997
$n = 200$	$Q_n(6)$	0.066	0.061	0.077	0.091	0.118	0.159
	$Q_n^2(6)$	0.035	0.044	0.052	0.061	0.090	0.128
	S_n^a	0.035	0.180	0.704	0.980	1.000	1.000
$n = 400$	$Q_n(6)$	0.052	0.057	0.064	0.095	0.186	0.324
	$Q_n^2(6)$	0.034	0.033	0.036	0.067	0.120	0.189
				H_1: BL Model			
	S_n^a		0.052	0.114	0.203	0.263	0.289
$n = 100$	$Q_n(6)$		0.053	0.051	0.045	0.064	0.081
	$Q_n^2(6)$		0.030	0.051	0.104	0.163	0.229
	S_n^a		0.073	0.228	0.406	0.455	0.440
$n = 200$	$Q_n(6)$		0.072	0.067	0.069	0.081	0.098
	$Q_n^2(6)$		0.041	0.096	0.206	0.313	0.408
	S_n^a		0.165	0.495	0.738	0.721	0.634
$n = 400$	$Q_n(6)$		0.044	0.044	0.045	0.053	0.095
	$Q_n^2(6)$		0.044	0.161	0.393	0.593	0.626

and $\beta = (\hat{\phi}_n, \hat{\psi}_n)'$. When $m = 6$, their sizes (i.e. the case with $\theta = 0.0$) in Table 2 show that, like the S_n^a test, there is apparently some evidence of conservatism for the Li-Mak test. When $m = 12$, the sizes (i.e. the case with $\theta = 0.0$) in Table 3 show that Ljunge-Box test tends to over-reject when $n = 100$ and 200, which is because its distribution is not approximated well by the χ^2-distribution when $n - m$ is small. When the alternative is the threshold ARMA model, Tables 2-3 show that the power of S_n^a increases when the sample size n or θ increases, while both $Q_n(m)$ and $Q_n^2(m)$ have much less, and in some cases almost no, power. When the alternative is the bilinear ARMA model, we have a similar conclusion for $Q_n(m)$, but $Q_n^2(m)$ performs much better than $Q_n(m)$ although not nearly as well as S_n^a. The power of S_n^a is higher when $\beta = (1,1)'$ than when $\beta = (\hat{\phi}_n, \hat{\psi}_n)'$ for the TARMA model, but it shows little difference for the BL model. For the BL model, there is also evidence to suggest that the power of S_n^a is affected adversely as θ approaches the boundary of invertibility, which is approximately 0.6. It seems that our test is more powerful against the TARMA alternative than

Table 3. Powers of S_n^a, $Q_n(m)$ and $Q_n^2(m)$ for Null Hypothesis H_0: ARMA(1,1) Model at Significance Level 0.05 [$\beta = (\hat{\phi}_n, \hat{\psi}_n)'$ and 1,000 replications].

θ		0.0	0.1	0.2	0.3	0.4	0.5
			H_1: TARMA Model				
	S_n^a	0.030	0.061	0.170	0.304	0.416	0.438
$n = 100$	$Q_n(12)$	0.121	0.110	0.123	0.135	0.164	0.209
	$Q_n^2(12)$	0.030	0.032	0.029	0.029	0.028	0.030
	S_n^a	0.046	0.112	0.362	0.674	0.878	0.709
$n = 200$	$Q_n(12)$	0.090	0.094	0.098	0.139	0.205	0.314
	$Q_n^2(12)$	0.044	0.039	0.037	0.051	0.052	0.067
	S_n^a	0.047	0.211	0.674	0.952	0.970	0.930
$n = 400$	$Q_n(12)$	0.062	0.065	0.088	0.165	0.336	0.577
	$Q_n^2(12)$	0.057	0.052	0.051	0.055	0.071	0.108
			H_1: BL Model				
	S_n^a		0.035	0.098	0.210	0.206	0.234
$n = 100$	$Q_n(12)$		0.053	0.051	0.045	0.064	0.081
	$Q_n^2(12)$		0.039	0.058	0.091	0.128	0.175
	S_n^a		0.083	0.272	0.455	0.463	0.480
$n = 200$	$Q_n(12)$		0.086	0.091	0.097	0.097	0.101
	$Q_n^2(12)$		0.058	0.086	0.167	0.237	0.288
	S_n^a		0.189	0.582	0.747	0.826	0.721
$n = 400$	$Q_n(12)$		0.057	0.062	0.058	0.062	0.095
	$Q_n^2(12)$		0.069	0.121	0.246	0.359	0.419

against the BL alternative. This is consistent with the interpretation of the test given in §2 by reference to an L-M test.

We next study the size and the power of S_n^a when the null hypothesis is the GARCH(1,1) model, $y_t = \eta_t\sqrt{h_t}$ *and* $h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}$, where η_t is i.i.d. $N(0,1)$. For the size, the true parameters are taken to be $\alpha_0 = 0.1$ and $(\alpha_1, \beta_1) = (0.3, 0.4)$, (0.3,0.5), (0.3,0.6), (0.2,0.7), (0.1,0.8), (0.3,0.7), (0.2,0.8), and (0.1,0.9). $\beta = (1,1,1)'$ and $(\hat{\alpha}_{0n}, \hat{\alpha}_{1n}, \hat{\beta}_{1n})'$. The sample sizes are n =100, 200 and 400. Table 4 summarizes the results when the significance level α is 0.01, 0.05 and 0.1, respectively. It shows that the sizes of S_n^a are fairly close to their nominal values, although there is some evidence of over-rejection when $\alpha = 0.01$ and conservatism when $\alpha = 0.10$.

The power of of S_n^a is studied via two alternatives:

$$\text{TGARCH} \quad \sqrt{h_t} = 0.1 + 0.3|y_{t-1}| + 0.4\sqrt{h_{t-1}} + \theta|y_{t-1}|I\{y_{t-1} \le 0\},$$

$$\text{NAGARCH} \quad h_t^{3/4} = 0.1 + 0.3\,|(\theta - \text{sgn}(\eta_t))y_t|^{3/2} + 0.4h_{t-1}^{3/4}.$$

Table 4. Sizes of S_n^a for Null Hypothesis H_0: GARCH(1,1) model at Significance Level α (1,000 replications).

		$n = 100$			$n = 200$			$n = 400$		
α		0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
α_1	β_1				$\beta = (1, 1, 1)'$					
0.3	0.4	0.016	0.031	0.055	0.005	0.027	0.056	0.008	0.035	0.063
0.3	0.5	0.007	0.027	0.051	0.007	0.025	0.057	0.007	0.034	0.064
0.3	0.6	0.008	0.029	0.056	0.009	0.032	0.062	0.008	0.033	0.063
0.2	0.7	0.009	0.032	0.058	0.010	0.036	0.069	0.008	0.037	0.069
0.1	0.8	0.011	0.041	0.070	0.011	0.034	0.071	0.007	0.037	0.069
0.3	0.7	0.010	0.036	0.068	0.008	0.038	0.070	0.011	0.038	0.073
0.2	0.8	0.010	0.042	0.069	0.008	0.035	0.068	0.010	0.037	0.078
0.1	0.9	0.015	0.055	0.090	0.016	0.043	0.084	0.012	0.048	0.092
					$\beta = (\hat{\alpha}_{0n}, \hat{\alpha}_n, \hat{\beta}_n)'$					
0.3	0.4	0.004	0.027	0.049	0.005	0.027	0.056	0.005	0.027	0.053
0.3	0.5	0.007	0.027	0.051	0.007	0.025	0.057	0.005	0.029	0.052
0.3	0.6	0.008	0.029	0.056	0.009	0.032	0.062	0.008	0.033	0.063
0.2	0.7	0.009	0.032	0.058	0.010	0.036	0.069	0.008	0.037	0.069
0.1	0.8	0.011	0.041	0.070	0.011	0.034	0.071	0.007	0.037	0.069
0.3	0.7	0.010	0.036	0.068	0.008	0.038	0.070	0.011	0.038	0.073
0.2	0.8	0.008	0.035	0.068	0.008	0.035	0.068	0.010	0.037	0.078
0.1	0.9	0.015	0.055	0.090	0.016	0.043	0.084	0.012	0.048	0.092

The first model is a threshold GARCH that is a special case of models proposed by Taylor (1986) and Schwert (1989). The second is a nonlinear asymmetric GARCH model proposed by Engle and Ng (1993). We take $\theta = 0.4$, 0.6, 0.8, 1.0 and 1.2. The sample sizes are $n = 100$, 200 and 400. Again, we compare the power of S_n^a with those of $Q_n(m)$ and $Q_n^2(m)$. The sizes of $Q_n(6)$ and $Q_n^2(6)$ are very close to their corresponding nominal values; see Li and Mak (1994) and Wong and Ling (2005) for simulation evidence. The results reported in Table 5 are for the significance level 0.05 when $\beta = (1, 1)'$. In all cases, S_n^a is more powerful than $Q_n(6)$ and $Q_n^2(6)$. In particular, when the alternative is the NAGARCH model, S_n^a can reject GARCH with power reaching 50 percent, while both $Q_n(6)$ and $Q_n^2(6)$ have virtually no power. Again it seems that our test is more powerful against the TGARCH alternative than against the NAGARCH alternative. Similar conclusions hold when $\beta = (\hat{\alpha}_{0n}, \hat{\alpha}_{1n}, \hat{\beta}_{1n})'$ and $m = 12$. Details are available from the authors.

We also carried out some experiments when $\beta = (1, \delta)$ with δ =0, $\pm$0.2, $\pm$0.4, $\pm$0.6 and $\pm$0.8 for the null ARMA(1,1) model. The sizes are relatively stable. But, for the alternative TARMA(1,1) model, S_n^a is less powerful than when $\beta = (1, 1)$, and is more powerful when $|\delta| > 0.4$ and less powerful when $|\delta| \leq 0.4$ than when $\beta = (\hat{\phi}_n, \hat{\psi}_n)'$. We carried out some experiments by taking a

Table 5. Powers of S_n^a, $Q_n(m)$ and $Q_n^2(m)$ for Null Hypothesis H_0: GARCH(1,1) Model at Significance Level 0.05 [$\beta = (1,1)'$ and 1,000 replications].

θ		0.4	0.6	0.8	10.0	10.2
		H_1: TGARCH Model				
	S_n^a	0.310	0.543	0.614	0.618	0.670
$n = 100$	$Q_n(6)$	0.153	0.156	0.176	0.286	0.457
	$Q_n^2(6)$	0.072	0.083	0.069	0.108	0.177
	S_n^a	0.478	0.845	0.766	0.666	0.751
$n = 200$	$Q_n(6)$	0.138	0.167	0.196	0.273	0.502
	$Q_n^2(6)$	0.085	0.066	0.070	0.138	0.286
	S_n^a	0.680	0.978	0.896	0.737	0.834
$n = 400$	$Q_n(6)$	0.125	0.132	0.146	0.230	0.512
	$Q_n^2(6)$	0.145	0.117	0.080	0.157	0.479
		H_1: NAGARCH Model				
	S_n^a	0.099	0.143	0.217	0.322	0.457
$n = 100$	$Q_n(6)$	0.075	0.082	0.091	0.103	0.115
	$Q_n^2(6)$	0.030	0.040	0.046	0.050	0.052
	S_n^a	0.116	0.173	0.283	0.454	0.649
$n = 200$	$Q_n(6)$	0.074	0.079	0.087	0.100	0.109
	$Q_n^2(6)$	0.040	0.040	0.046	0.060	0.062
	S_n^a	0.127	0.214	0.393	0.630	0.863
$n = 400$	$Q_n(6)$	0.051	0.055	0.063	0.068	0.072
	$Q_n^2(6)$	0.034	0.044	0.053	0.065	0.073

as the 10%-quantile of data $\{y_1, \ldots, y_n\}$ when the null is the GARCH (1,1) model. Compared with those in Table 4, the sizes of S_n^a are closer to their nominal levels when the level is 0.01 and 0.05, and are more conservative when the level is 0.1. In each case, the power is higher than in Table 5. For example, when $n = 400$ and the alternative is the NAGARCH model, the powers of S_n^a with $\beta = (1,1,1)'$ are 0.121, 0.940, 0.334, 0.533 and 0.776, respectively, for θ at 0.4, 0.6, 0.8, 1.0, and 1.2.

We now compare our test with that of Koul and Stute (1999). Since the test in Koul and Stute (1999) has not been extended to cover the ARMA model or the GARCH model in the literature, we only consider the null AR(1), $y_t = \phi y_{t-1} + \varepsilon_t$, where ε_t is i.i.d. $N(0,1)$. In this case, their test statistic is $KS_n \equiv \max_{x \le x_0} |V_n(x)|/[\sigma_n G_n(x_0)]$, where $G_n(x_0) = \sum_{t=1}^n I\{y_{t-1} \le x_0\}/n$, $\sigma_n^2 = \sum_{t=1}^n$

Table 6. Sizes and Powers of S_n^a and KS_n for Null Hypothesis H_0: AR(1) Model at Significance Level 0.05 [$\beta = 1$ and 1,000 replications].

θ		0.0	0.1	0.2	0.3	0.4	0.5
				H_1: TAR Model			
	S_n^a	0.042	0.061	0.109	0.171	0.266	0.392
$n = 100$	KS_n	0.046	0.055	0.073	0.113	0.168	0.248
	S_n^a	0.044	0.092	0.184	0.350	0.546	0.747
$n = 200$	KS_n	0.054	0.088	0.158	0.281	0.472	0.630
	S_n^a	0.038	0.106	0.306	0.618	0.855	0.960
$n = 400$	KS_n	0.056	0.098	0.272	0.587	0.820	0.947
				H_1: BL Model			
	S_n^a		0.032	0.060	0.111	0.146	0.158
$n = 100$	KS_n		0.062	0.085	0.109	0.113	0.119
	S_n^a		0.068	0.179	0.317	0.423	0.473
$n = 200$	KS_n		0.078	0.114	0.134	0.149	0.155
	S_n^a		0.119	0.404	0.667	0.808	0.840
$n = 400$	KS_n		0.081	0.129	0.177	0.206	0.251

$\varepsilon_t^2(\hat{\phi}_n)/n$, and

$$V_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[I\{y_{i-1} \le x\} - \frac{1}{n} \sum_{j=1}^{n} \frac{y_{j-1} y_{i-1} I\{y_{j-1} \le y_{i-1} \wedge x\})}{n^{-1} \sum_{k=1}^{n} y_{k-1}^2 I\{y_{k-1} \ge y_{j-1}\}} \right] \varepsilon_t(\hat{\phi}_n),$$

where $\varepsilon_t(\hat{\phi}_n) = y_t - \hat{\phi}_n y_{t-1}$ and $\hat{\phi}_n$ is the LSE of ϕ. We take x_0 to be the 95%th quantile of data set $\{y_1, \ldots, y_n\}$. In the simulation, $\phi = 0.5$ and sample size $n = 100$, 200, and 400. Alternatives are TAR and bilinear (BL) models:

$$\text{TAR Model } y_t = 0.5 y_{t-1} - \theta y_{t-1} I\{y_{t-1} \le 0\} + \varepsilon_t,$$
$$\text{BL Model } y_t = 0.5 y_{t-1} - \theta y_{t-2} \varepsilon_{t-1} + \varepsilon_t.$$

Table 6 reports the sizes (i.e. case with $\theta = 0.0$) and powers. It can be seen that both tests have little power when (i) $\theta \le 0.2$ and $n = 100$; (ii) $\theta = 0.1$ and $n = 200$. Except for these cases, the S_n^a test is uniformly more powerful than the KS_n test. In addition, it only needs n iterations to compute the $T_n(x, \hat{\theta})$ in the S_n^a test, while it needs n^3 iterations to compute the $V_n(x)$ in the KS_n test. For each cell in Table 6 when $n = 200$, the S_n^a test takes 15 seconds, while the KS_n test takes 3 hours and 40 minutes (running on a Pentium IV at HKUST).

Finally, we compare our test with the generalized likelihood-based test with bias reduction proposed by Fan and Zhang (2004). We first use the null and

alternative models exactly as in Example 1 of Fan and Zhang (2004). Specifically, the null model is an AR(3) model given by model (4.1) with $\beta = 0$. The alternative model is

$$y_t = \{\theta_1(1-\beta) + \beta\nu(y_{t-3})\}y_{t-1} + \theta_2 y_{t-2} + \theta_3 y_{t-3} + \varepsilon_t, \tag{4.1}$$

where $\nu(x) = 0.95I\{-.5 \le x < 0\} - 1.8xI\{0 \le x \le 5\}$, $\{\varepsilon_t\}$ are independently and identically distributed N(0, 1) random variables, and β is a given parameter. The true values of the θ-parameters are $\theta_1 = 0.8$, $\theta_2 = -0.56$, and $\theta_3 = 0.6$. For each fixed β, we simulate a time series of length $n = 500$ and use 1,000 replications for different choices of β. Figure 1 shows the power functions of our test (LT) and the Fan-Zhang test (FZ) when the significance level is 0.05. When $\beta = 0$, the power becomes the size of the test.

From Figure 1, we can see that our test is more powerful than the Fan and Zhang test when $0 < \beta \le 0.8$ (roughly), but is less powerful when $0.8 < \beta \le 1$. Because it is not known whether model (4.1) is stationary or not and there is empirical evidence to suggest that it may not be stationary when $\beta > 0.8$, we repeated the simulation study with the following model, which we know is stationary. (Tong (1990, p.464)). It is a two-regime TAR(3) model:

$$y_t = [0.5I\{y_{t-1} \le 0\} + (0.5-\beta)I\{y_{t-1} > 0\}]y_{t-1} - 0.3y_{t-2} - 0.1y_{t-3} + \varepsilon_t, \tag{4.2}$$

where $\beta \in [0, 1]$. Figure 2 shows the power functions of LT and FZ when the significance level is 0.05. Again, the power becomes the size of the test when $\beta = 0$. We are puzzled by the very low power of the Fan-Zhang test in this case, but can offer no explanation.

5. The Hang Seng Index

We used the S_n^a tests to investigate the Hang Seng Index (HSI) for the Hong Kong stock market. Each period of two years from 01/06/1988-31/05/1996 was considered. The model we used to fit the data was the AR-GARCH model

$$y_t = \phi y_{t-1} + \varepsilon_t, \tag{5.1}$$

$$\varepsilon_t = \eta_t\sqrt{h_t} \text{ and } h_t = \alpha_0 + f\varepsilon_{t-i}^2 + gh_{t-i}. \tag{5.2}$$

The results are summarized in Table 7. In this table, the values in the parenthesis are the corresponding asymptotic standard deviations of the estimated parameters and LF is the value of log-likelihood function. As in Section 4, $Q_n(6)$ and $Q_n^2(6)$ are the Ljung-Box test and Li-Mak test, respectively. Both tests suggest that this model fits the data adequately. We used the statistic S_n^a with $\beta = (1, \ldots, 1)'$ and a being the $5p\%$-quantile of data $\{y_1, \ldots, y_n\}$ to test the null model (5.1)-(5.2). The S_n^a test rejects the null model for all the four periods at

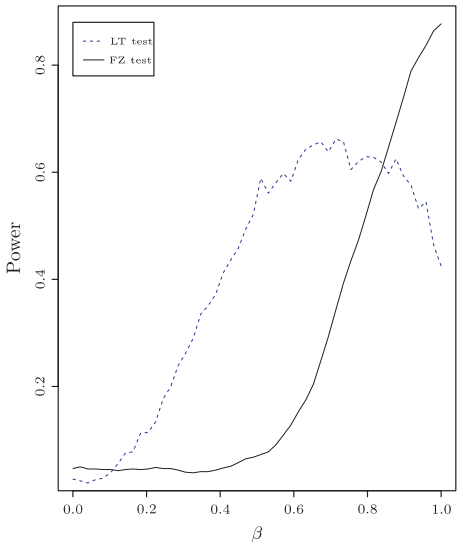

Figure 1. The power functions of our test (LT) and the Fan-Zhang test (FZ) at the 5% significance level, based on 1,000 simulations and for different choices of β for model (4.1).

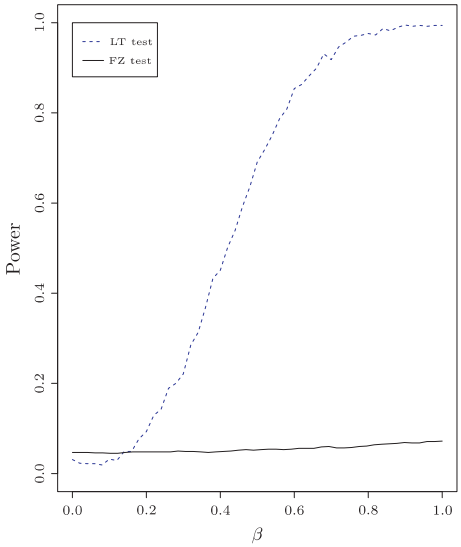

Figure 2. The power functions of our test (LT) and the Fan-Zhang test (FZ) at the 5% significance level, based on 1,000 simulations and for different choices of β for model (4.2).

the 0.05-significance level. We should mention that the same model and data set were used for testing the normality of η_t in Koul and Ling (2006). It is interesting

Table 7. Empirical Results for Hong Seng Index Fitted AR(1)−GARCH(1,1) Models

Periods	n	ϕ	α_0	f	g	LF	$Q_n(6)$	$Q_n^2(6)$	S_n^a
1/6/88−31/5/90	493	0.242	0.083	0.223	0.772	-1170.9	60.38	00.61	260.92
		(0.055)	(0.027)	(0.048)	(0.031)				
1/6/90−31/5/92	495	0.186	0.526	0.203	0.442	-1310.7	20.97	00.92	70.61
		(0.056)	(0.162)	(0.070)	(0.145)				
1/6/92−31/5/94	498	0.117	0.322	0.242	0.664	250.5	70.09	30.05	250.63
		(0.050)	(0.108)	(0.057)	(0.068)				
1/6/94−31/5/96	497	0.128	0.053	0.069	0.900	-950.6	70.30	00.65	360.13
		(0.048)	(0.029)	(0.025)	(0.034)				

to see if a DAR model, $y_t = \phi y_{t-1} + \eta_t \sqrt{\omega + \alpha y_{t-1}^2}$, is adequate for these periods. It turns out that the values of S_n^a were 0.54, 1.77, 10.55, and 6.33, respectively. Thus, our test suggests that the fitted DAR(1) is adequate at the 0.05 level for the first two periods but not for the last two periods.

6. Conclusions

This paper has developed a general approach to goodness-of-fit tests that are easy to construct for a wide variety of time series models, ranging from the linear model to the nonlinear model, and from the constant variance model to the ARCH-type model. The critical values of the test statistics are available without the need to bootstrap. To illustrate the versatility of our general approach, we have detailed the construction for four specific models. Simulation results suggest that our test works well compared with classical portmanteau tests such as the Ljung-Box test and the Li-Mak test, and the recent Koul-Stute test (when comparison is possible) and Fan-Zhang test. We have demonstrated the efficacy of our approach in an application to the Hang Seng Index.

Size distortion is a fairly common problem among goodness-of-fit tests in time series; ours is no exception. In the event of several competing tests each with size distortion, there seems to be some general agreement that in practice a conservative test is preferred to one that over-rejects. (See, e.g., Kheoh and McLeod (1992)). As suggested in Li (2004, p.11), this is particularly the case when their power is comparable. For our approach, it is worthwhile to investigate whether the distortion is due to a in (2.3) when the sample size is not large enough. Another challenging open problem is the optimal choice of β. Of course, it is not inconceivable that the problem might turn out to be as intractable as the number of lags used in, for example, the Ljung-Box test or the Li-Mak test.

Acknowledgements

We thank Professor Andy Wood for useful comments and suggestions, Professor Wenyang Zhang for his programs and discussion, and the Hong Kong Research Grants Commission (Grant #HKUST6016/07P and HKUST602609 to SQL and #HKU7049/03P to HT), the University of Hong Kong (Distinguished Visiting Professorship), the National University of Singapore (Saw Swee Hock Professorship) and the UK EPSRC grant, all to HT, for partial support.

Appendix: Proofs of Lemma 1 and Theorem 3

Proof. (a). Note that $E\|D_t(\theta_0)\|^2 < \infty$ and $E\|P_t(\theta_0)\| < \infty$. We have

$$\max_{1\le t\le n} \|D_t(\theta_0)\| = o_p(n^{1/2}) \; and \; \max_{1\le t\le n} \|P_t(\theta_0)\| = o_p(n),$$

see e.g., Chung (1968, p.93). By Assumption 3, we have

$$\begin{aligned}
&\Big\|D_t(\hat\theta_n)D_t'(\hat\theta_n) - D_t(\theta_0)D_t'(\theta_0)\Big\| \\
&\quad\le \Big\|D_t(\hat\theta_n) - D_t(\theta_0)\Big\|^2 + 2\|D_t'(\theta_0)\|\Big\|D_t(\hat\theta_n) - D_t(\theta_0)\Big\| \\
&\quad= O_p(\frac{1}{n})\Big\|P_t(\hat\theta_n^*)\Big\|^2 + O_p(\frac{1}{\sqrt{n}})\|D_t(\theta_0)\|\Big\|P_t(\hat\theta_n^*)\Big\| \\
&\quad= O_p(\frac{1}{n})\Big\|P_t(\hat\theta_n^*) - P_t(\theta_0)\Big\|^2 + o_p(1)\Big\|P_t(\hat\theta_n^*) - P_t(\theta_0)\Big\| + o_p(n), \qquad \text{(A.1)}
\end{aligned}$$

where $O_p(\cdot)$ and $o_p(\cdot)$ hold uniformly in $t = 1, \ldots, n$, and $\hat\theta_n^*$ lies between $\hat\theta_n$ and θ_0. Note that $\sqrt{n}(\hat\theta_n^* - \theta_0) = O_p(1)$. For any $\varepsilon > 0$, there exists a constant C such that

$$P(\|\sqrt{n}(\hat\theta_n^* - \theta_0)\| > C) \le \frac{\varepsilon}{2}.$$

By Assumption 3 again, we have

$$\begin{aligned}
&P\left(\frac{1}{n^2}\sum_{t=1}^n \Big\|P_t(\hat\theta_n^*) - P_t(\theta_0)\Big\|^2 \ge \varepsilon\right) \\
&\quad\le P\left(\left[\frac{1}{n}\sup_{\sqrt{n}\|\theta-\theta_0\|\le C}\sum_{t=1}^n \Big\|P_t(\hat\theta_n^*) - P_t(\theta_0)\Big\|\right]^2 > \varepsilon\right) + \frac{\varepsilon}{2} \le \varepsilon, \qquad \text{(A.2)}
\end{aligned}$$

as n is large enough. Similarly, we have

$$P\left(\frac{1}{n}\sum_{t=1}^n \Big\|P_t(\hat\theta_n^*) - P_t(\theta_0)\Big\| \ge \varepsilon\right) \le \varepsilon. \qquad \text{(A.3)}$$

By (A.1)−(A.3), we can show that

$$\frac{1}{n}\sum_{t=1}^{n}\left\|D_t(\hat{\theta}_n)D_t'(\hat{\theta}_n) - D_t(\theta_0)D_t'(\theta_0)\right\| = o_p(1).$$

Using this equality, we have

$$\sup_{x\in R\cup\{\infty\}}\|\hat{\Sigma}_{nx} - \Sigma_x\| \le \frac{1}{n}\sup_{x\in R\cup\{\infty\}}\left\|\Delta_n(x)\right\| + o_p(1), \tag{A.4}$$

where $\Delta_n(x) = \sum_{t=1}^{n} D_t(\theta_0)D_t'(\theta_0)I\{y_{t-1}\le x\} - \Sigma_x$. By the Ergodic Theorem, for each fixed x, $\Delta_n(x) = o(1)$ and $\Delta_n(\infty) = o(1)$, a.s.. Thus, for any $\varepsilon > 0$, there exists a constant M such that

$$\sup_{x\le -M}\|\Delta_n(x)\| \le \frac{1}{n}\sum_{t=1}^{n}\|D_t(\theta_0\|^2 I\{y_{t-1}\le -M\} + \|\Sigma_{-M}\| \le \frac{\varepsilon}{2}. \tag{A.5}$$

Furthermore, by the Ergodic Theorem, for a large M we have

$$\begin{aligned}\sup_{x\ge M}\|\Delta_n(x)\| &\le \sup_{x\ge M}\|\Delta_n(\infty) - \Delta_n(x)\| + o(1)\\ &\le \frac{1}{n}\sum_{t=1}^{n}\|D_t(\theta_0\|^2 I\{y_{t-1}\ge M\} + E[\|D_t(\theta_0\|^2 I\{y_{t-1}\ge M\}]\\ &\le o(1) + \frac{\varepsilon}{2},\end{aligned} \tag{A.6}$$

as $n\to\infty$. Using a standard piece-wised argument, we can show that $\sup_{|x|\le M}\|\Delta_n(x)\| = o_p(1)$. Furthermore, by (A.4)−(A.6), we can claim that (a) holds. (b) comes directly from Assumptions 1 and 3 and (a) of this lemma. This completes the proof.

Proof of Theorem 3. Let P_{0n} denote the joint distribution of $(y_1,\ldots,y_n)$ under H_0 and P_{1n} that under H_{1n}. Let the log-likelihood ratio of P_{1n} to P_{0n} be denoted by Λ_n. Then

$$\Lambda_n = -\frac{1}{2}\sum_{t=1}^{n}\left[\log h_{nt} - \log h_t(\theta_0) - \frac{\varepsilon_{nt}^2}{h_{nt}} + \frac{(y_t-\mu_t(\theta_0))^2}{h_t(\theta_0)}\right],$$

where $\varepsilon_{nt} = y_t - \mu_t(\theta_0) - r_{1t}/\sqrt{n}$ and $h_{nt} = h_t(\theta_0) + r_{2t}/\sqrt{n}$. Using Le Cam's third lemma in Van der Vaart and Wellner (1996) and either Theorem 2.1 in Ling and McAleer (2003) or by a direct method, we can show that (a) holds. Part (b) follows directly from (a). This completes the proof.

References

An, H. Z. and Cheng, B. (1991). A Kolmogorov-Smirnov type statistic with application to test for nonlinearty in time-series. *Internat. Statist. Rev.* **59**, 287-307.

Bartlett, M. S. and Diananda, P. H. (1950). Extension of Quenouille's test for autoregressive schemes. *J. Roy. Statist. Soc. Ser. B* **12**, 108-115.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Box, G. E. P. and Pierce, D. A. (1970). Distribution of the residual autocorrelations in autoregressive integrated moving average time series models. *J. Amer. Statist. Assoc.* **65**, 1509-1526.

Chan, K. S. (1990). Testing for threshold autoregression. *Ann. Statist.* **18**, 1886-1893.

Chan, K. S. (1991). Percentage points of likelihood ratio tests for threshold autoregression. *J. Roy. Statist. Soc. B* **53**, 691-696.

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Statist.* **21**, 520-533.

Chan, K. S. and Tong, H. (1990). On likelihood ratio tests for threshold autoregression. *J. Roy. Statist. Soc. Ser. B* **52**, 469-476.

Chung, K. L. (1968). *A Course in Probability Theory*. Academic Press, New York.

Engle, R. and Ng, V. (1993). Measuring and testing the impact of news on volatility. *J. Finance* **48**, 1749-1778.

Escanciano, J. C. (2007). Model checks using residual marked empirical processes. *Statist. Sinica* **17**, 115-138.

Escanciano, J. C. (2008). Joint and marginal specification tests for conditional mean and variance models. *J. Econometrics* **143**, 74-87.

Fan, J. and Yao, Q. W. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

Fan, J. and Zhang, W. (2004). Generalized likelihood ratio tests for spectral density. *Biometrika* **91**, 195-209.

Francq, C. and Zakoïan, J. M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* **10**, 605-637.

Godfrey, L. G. (1979). Testing the adequacy of a time series model. *Biometrika* **66**, 67-72.

Granger, C. W. J. and Andersen, A. P. (1978). *An Introduction to Bilinear Time Series Models*. Vandenhoeck & Ruprecht, Gttingen.

Hong, Y. and Lee, T. H. (2003). Diagnostic checking for the adequacy of nonlinear time series models. *Econometric Theory* **19**, 1065-1121.

Hosking, J. R. M. (1978). A unified derivation of the asymptotic distribution of goodness-of-fit statistics for autoregressive time-series models. *J. Roy. Statist. Soc. Ser. B* **40**, 341-349.

Hosking, J. R. M. (1980). Lagrange-multiplier tests of time series models. *J. Roy. Statist. Soc. Ser. B* **42**, 170-181.

Kheoh, T. S. and McLeod, A. I. (1992). Comparison of two modified portmanteau tests for model adequacy. *Comput. Statist. Data Anal.* **14**, 99-106.

Koul, H. L. and Stute, W. (1999). Nonparametric model checks for time series. *Ann. Statist.* **27**, 204-236.

Koul, H. L. and Ling, S. (2006). Fitting an error distribution in some heteroscedastic time series models. *Ann. Statist.* **34**, 994-1012.

Li, W. K. (2004). *Diagnostic Checks in Time Series*. Chapman&Hall/CRC, New York.

Li, W. K. and Mak, T. K. (1994). On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. *J. Time Series Anal.* **15**, 627-636.

Ling, S. (2004). Estimation and testing of stationarity for double autoregressive models. *J. Roy. Statist. Soc. Ser. B* **66**, 63-68.

Ling, S. (2007). A double AR(p) model: structure and estimation. *Statist. Sinica* **17**, 161-175..

Ling, S. and McAleer, M. (2003). Adaptive estimation in nonstationry ARMA models with GARCH noises. *Ann. Statist.* **31**, 642-674.

Ling, S. and Tong, H. (2006). Weak convergence of a gereral marked empirical process and goodnessof-fit tests for time series models. Unpublished Working Paper, HKUST, `http://www.math.ust.hk/~maling/paper/Tong-wg8-1.pdf`.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**, 297-303.

McLeod, A. I. and Li, W. K.(1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *J. Time Series Anal.* **4**, 269-73.

Newbold, P. (1980). The equivalence of two tests of time series model adequacy. *Biometrika* **10**, 57-69.

Quenouille, M. H. (1947). A large-sample test for goodness of fit of autoregressive scheme. *J. Roy. Statist. Soc. Ser. A* **110**, 123-129.

Quenouille, M. H. (1949). Approximate tests of correlation in time series. *J. Roy. Statist. Soc. B* **11**, 68-84.

Schwert, G. W. (1989). Why does stock market volatility change over time?. *J. Finance* **44**, 1115-1153.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. John Wiley, New York.

Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25**, 613-641.

Stute, W., Quindimil, M. P., Manteiga, W. G. and Koul, H. L. (2006). Model checks of higher order time series. *Statist. Probab. Lett.* **76**, 1385-1396.

Taylor, S. (1986). *Modelling Financial Time Series*. Wiley, New York.

Tong, H. (1978). On a threshold model. *Pattern Recognition and Signal Processing*. (Edited by C. H. Chen). Sijthoff and Noordhoff, Amsterdam.

Tong, H. (1990). *Nonlinear Time Series. A Dynamical System Approach*. Oxford University Press, Oxford.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer-Verlag, New York.

Walker, A. M. (1950). Note on a generalization of the large sample goodness of fit test for linear autoregressive schemes. *J. Roy. Statist. Soc. Ser. B* **12**, 102-107.

Walker, A. M. (1952). Some properties of the asymptotic power functions of goodness-of-fit tests for linear autoregressive schemes. *J. Roy. Statist. Soc. Ser. B* **14**, 117-134.

Weiss, A. A. (1986). Asymptotic theory for ARCH models: estimation and testing. *Econometric Theory* **2**, 107-131.

Wong, C. S. and Li, W. K. (1997). Testing for threshold autoregression with conditional heteroscedasticity. *Biometrika* **84**, 407-418.

Wong, C. S. and Li, W. K. (2000). Testing for double threshold autoregressive conditional heteroscedastic model. *Statist. Sinica* **10**, 173-189.

Wong, H. and Ling, S. (2005). Mixed portmanteau tests for time series. *J. Time Series Anal.* **26**, 569-579.

Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong.

E-mail: maling@ust.hk

Department of Statistics, London School of Economics & Political Science, Houghton Street, London WC2A 2AE, United Kingdom.

E-mail: howell.tong@gmail.com

(Received April 2009; accepted June 2010)