Hypothesis testing based on a vector of statistics

Maxwell L. King, Xibin Zhang[1]

Department of Econometrics and Business Statistics, Monash University, Australia

Muhammad Akram

Mary MacKillop Institute for Health Research, Australian Catholic University

August 31, 2017

Revised: July 1, 2018

Abstract: This paper presents a new approach to hypothesis testing based on a vector of statistics. It involves simulating the statistics under the null hypothesis and then estimating the joint density of the statistics. This allows the p-value of the smallest acceptance region test to be estimated. We prove this p-value is a consistent estimate under some regularity conditions. The small-sample properties of the proposed procedure are investigated in the context of testing for autocorrelation, testing for normality, and testing for model misspecification through the information matrix. We find that our testing procedure has appropriate sizes and good powers.

JEL classification: C01, C12, C14.

Keywords: Cross-market prediction, Information matrix test, Markov chain Monte Carlo, Multivariate kernel density, p-value

[1]Address: 900 Dandenong Road, Caulfield East, Victoria 3145, Australia. Telephone: +61 3 99032130. Fax: +61 3 99032007. Email: xibin.zhang@monash.edu.

1 Introduction

Statistical hypothesis testing is an extremely important technique in the practice of econometrics, particularly with respect to diagnostic checking of model specification. This is how econometricians are best able to combat the severe problem of uncertainty in model specification. Such testing procedures need to be as accurate as possible due to constraints on data availability. Fortunately, advances in computer power and simulation based methods have allowed greater scope in the design of high quality tests. The purpose of any test is to accurately control the probability of wrongly rejecting the null hypothesis (known as the size of the test), while at the same time ensuring a high probability of correctly rejecting the null hypothesis (known as the power of the test).

There are a number of popular tests whose construction involves a desire to test whether a $d \times 1$ vector of statistics denoted by $\boldsymbol{t}(\boldsymbol{y})$, where $\boldsymbol{y}$ is a $T \times 1$ vector of observations, is significantly different from zero. Examples include the Lagrange multiplier test (see Godfrey, 1991) where $\boldsymbol{t}(\boldsymbol{y})$ is a vector of Lagrange multipliers, the Wald test (see Engle, 1984) where $\boldsymbol{t}(\boldsymbol{y})$ is a vector of parameter estimates minus their null hypothesis values, the Durbin-Wu-Hausman test (see Davidson and MacKinnon, 1993, Section 7.9) in which $\boldsymbol{t}(\boldsymbol{y})$ is a vector of differences between two types of parameter estimates, the information matrix test (White, 1982) where $\boldsymbol{t}(\boldsymbol{y})$ is vector of differences between the corresponding elements of the negative Hessian matrix and the outer product of the score vector, and conditional moment tests (see Pagan and Vella, 1989) where $\boldsymbol{t}(\boldsymbol{y})$ is a set of sample moments. The general method of test construction then involves determining the asymptotic covariance matrix V of $\boldsymbol{t}(\boldsymbol{y})$ and using a consistent estimate of V, denoted as $\widehat{V}$, to obtain the quadratic form

$$\boldsymbol{t}(\boldsymbol{y})'\widehat{V}^{-1}\boldsymbol{t}(\boldsymbol{y}). \tag{1}$$

Under suitable regularity conditions, the latter can be shown to have a χ^2 asymptotic distribution when the null hypothesis holds, which can be used to test whether $\boldsymbol{t}(\boldsymbol{y})$ is significantly different from zero.

Unfortunately, these tests can have poor size and power properties, particularly in small samples, because of $\widehat{V}$ being a poor estimate of the true covariance matrix of $\boldsymbol{t}(\boldsymbol{y})$. Research that have found such outcomes include papers by Gregory and Veall (1985, 1986, 1987), Lafontaine and White (1986), Mantel (1987), Taylor (1987), Orme (1990), Chesher and Spady (1991), Davidson and MacKinnon (1992) and Dhaene and Hoorelbeke (2004). In the case of testing non-linear restrictions in the parameters of a linear model using the Wald test, Breusch and Schmidt (1988) were able to demonstrate the possibility of obtaining any value for the Wald statistic by rewriting the restrictions. In fact, the involvement of $\widehat{V}$ in (1) is purely for standardization and the convenience of reducing a vector of statistics, $\boldsymbol{t}(\boldsymbol{y})$, to a scalar test statistic. This does raise the question of whether this can be done in a better and less harmful way.

There are similarities with this particular test construction problem and multiple testing of a common null hypothesis using the same data also known as induced testing (or combined testing). A very large and growing literature on diagnostic testing of all kinds of econometric models (see, for example, Breusch and Pagan, 1980; Engle, 1984; Tauchen, 1985; Wooldridge, 2001; Godfrey, 2009) allows researchers to check the adequacy of a chosen model by applying a range of diagnostic tests, each of which is designed to detect a particular form of model inadequacy. A major problem is how best to control the overall probability of rejecting the model when it is true. For example, five statistically independent tests applied at the 5% level will result in a 22.6% chance of at least one rejection when the null hypothesis model is true. Of course, it is unlikely that five diagnostic tests applied to the same model will be mutually independent, so in actual fact, this probability could be higher or lower than 22.6%. The core issue is how we should conduct these tests in order to control the overall probability of rejecting the model when it is true with an eye to having good power to reject the model when it is not true. For this type of application, $\boldsymbol{t}(\boldsymbol{y})$ denotes the vector of test statistics.

An early solution to this problem involves the use of the Bonferroni inequality to set an upper bound on the overall probability of rejecting the model. Unfortunately, this approach is known to be conservative and therefore to have reduced power, particularly when some of the individual tests are highly correlated. There have been a number of attempts to improve this approach (see, for example, Holm, 1979; Simes, 1986), but it is still regarded as a conservative approach to the problem. We have also seen the development of joint LM tests which test for multiple forms of misspecification (see, for example, Bera and Jarque, 1982; Baltagi and Li, 1995; Baltagi, Bresson, and Pirotte, 2006; Baltagi, Jung, and Song, 2010).

Dufour and his coauthors[2] developed a framework to address this problem based on Monte Carlo testing introduced by Dwass (1957) and Barnard (1963). It requires being able to calculate a p-value (or approximate p-value) for each individual test statistic $t_i(\boldsymbol{y})$ and then combining the m p-values either by calculating the minimum p-value, the product of p-values or a weighted product of p-values (or some other variation of these approaches) to produce a single value that can be used as test statistic in a Monte Carlo test. The latter involves a Monte Carlo simulation to calculate an empirical p-value of this new test statistic. When it can be shown that all the individual tests that give rise to the m p-values are similar tests (i.e., their distributions under the null hypothesis do not depend on nuisance parameters), then this approach leads to an exact α-level test provided $\alpha(N+1)$ is an integer where α is the desired significance level and N is the number of Monte Carlo replications.

In the case where the null distribution of $\boldsymbol{t}(\boldsymbol{y})$ depends on nuisance parameters, Dufour (2006) proposed the maximized Monte Carlo (MMC) test which involves maximizing the simulated p-value with respect to the nuisance parameters. This results in an exact test in the sense that the test's size, which is a function of the nuisance parameters, is always less than or equal to α. Unfortunately, because the size of this test is likely to be less than α at the true (unknown) value of the nuisance parameters, it is likely to be a conservative test with some loss of power. In view of this, Dufour also suggested the Consistent Set Estimate MMC test, which involves restricting the maximization search to a confidence set for the nuisance parameters under the null hypothesis. A further suggestion involves obtaining a consistent estimate of the nuisance

[2] Dufour, Farhat, Gardiol, and Khalaf (1998); Dufour and Khalaf (2001, 2002); Dufour, Khalaf, and Beaulieu (2003, 2010); Dufour, Khalaf, Bernard, and Genest (2004); Dufour, Farhat, and Khalaf (2004); Dufour (2006); Beaulieu, Dufour, and Khalaf (2007, 2009); Bernard, Idoudi, Khalaf, and Yélou (2007); Dufour, Khalaf, and Voia (2015).

parameters and using these estimates in the Monte Carlo simulations to apply the Monte Carlo test. Both alternatives have asymptotic justifications.

A bootstrap solution has been outlined by Westfall and Young (1993), Godfrey (2005) and Godfrey (2009, Section 4.3). Godfrey's procedure involves first calculating p-values for each of the individual tests, preferably via a series of first-stage bootstrap sampling. These p-values are then combined via the minimum p-values approach and a second-stage of bootstrap sampling is used to calculate an approximate overall p-value that is asymptotically justified.

This paper proposes an alternative approach to hypothesis testing based on a vector of statistics, $\boldsymbol{t}(\boldsymbol{y})$: $R^T \longrightarrow R^d$ $(T > d)$. It assumes that each of the elements of $\boldsymbol{t}(\boldsymbol{y})$ has been chosen because individually they have good power to detect a particular deviation from the null hypothesis. It is further assumed that collectively $\boldsymbol{t}(\boldsymbol{y})$ provides a good summary of the evidence contained in $\boldsymbol{y}$ that might point to the null hypothesis being false. The approach involves asking the question based on the observed value of $\boldsymbol{t}(\boldsymbol{y})$, do we think the null hypothesis is true? If we know the joint density function for $\boldsymbol{t}(\boldsymbol{y})$ under the null hypothesis denoted as $f(\boldsymbol{t})$, then following Hyndman (1996) we can calculate the p-value for the observed value of $\boldsymbol{t}(\boldsymbol{y})$. Typically we do not know this joint density function, and even if we did, it is possible that $f(\boldsymbol{t})$ will depend on unknown nuisance parameter values. Our approach is to simulate independent values of $\boldsymbol{t}(\boldsymbol{y})$ under the null hypothesis using consistent estimates of any nuisance parameters that $f(\boldsymbol{t})$ depends on, and then use a multivariate kernel density estimator to estimate the density.

The contribution we make in this paper can be viewed as a way to use simulation methods to (approximately) control the probability of falsely rejecting the null hypothesis based on a vector of statistics. This is done by approximating the smallest acceptance region test. We prove that under some regularity conditions, the estimated p-value of our proposed test procedure is a consistent estimate of the true p-value, thus proving our test is an asymptotic smallest acceptance region test. We show that one version of our test can be regarded as a Monte Carlo test as outlined by Dufour (2006) and therefore can be an exact test when $f(\boldsymbol{t})$ does not depend on unknown nuisance parameters. When $f(\boldsymbol{t})$ does depend on nuisance parameters, Dufour's

(2006) MMC procedure can be applied to a version of our test to yield an exact, but potentially conservative test. We also investigate the small-sample size and power properties of the new test through simulation studies. The most revealing study involves applying the procedure to the information matrix test which is of the form of (1) where $\boldsymbol{t}(\boldsymbol{y})$ is a vector of differences between corresponding elements of the negative Hessian matrix and the outer product of the score vector. The new procedure's estimated sizes are not significantly different from their nominal sizes, and its powers are almost always higher than, and in some cases more than double that of the best of two existing tests.

The rest of this paper is organized as follows. Section 2 presents the new testing procedure when $f(\boldsymbol{t})$ is independent of nuisance parameters and discusses its properties. In Section 3, we examine the performance of the new testing procedure through Monte Carlo simulations by comparing its size and power with that of a well known existing test and two-sided versions of Dufour's (2006) minimum p-value and product of p-value tests. In Section 4, we present the testing procedure when $f(\boldsymbol{t})$ depends on nuisance parameters. Section 5 briefly describes the information matrix test and its limitations. We present a simulation study of the new testing procedure applied to the information matrix test in Section 6. An empirical application of the proposed test is conducted through the information matrix test in Section 7. Section 8 concludes the paper.

2 The test procedure for statistics independent of nuisance parameters

2.1 Test procedure

We shall begin by first describing the main ideas behind our new testing procedure. Assume that we wish to test the null hypothesis that the $T \times 1$ vector of observations $\boldsymbol{y}$ has a particular parametric data generating process (DGP) using d statistics denoted as t_i, for $i = 1, 2, \cdots, d$. Let $\boldsymbol{t} = (t_1, t_2, \cdots, t_d)'$ represent the $d \times 1$ vector of statistics called the test vector hereafter. We also assume that each of the component tests is a two-sided test[3] based on accepting the null hypothesis if

$$c_{1i} < t_i < c_{2i},$$

where c_{1i} and c_{2i} are critical values, for $i = 1, 2, \cdots, d$. Let the joint density function of $\boldsymbol{t}$ under the null hypothesis be denoted by $f(\boldsymbol{t})$. For the moment, we assume that $f(\boldsymbol{t})$ does not depend on nuisance parameters. Let $\widehat{\boldsymbol{t}}$ denote the calculated value of the test vector $\boldsymbol{t}$ using the available data.

Essentially, we wish to ask the question, is the observed value of $\widehat{\boldsymbol{t}}$ consistent with the null hypothesis being true? The p-value is a useful tool for answering this question. It is defined as the probability under the null hypothesis of finding a value of the test vector as extreme as or more extreme than the value we have found from the data, namely $\widehat{\boldsymbol{t}}$. Thus, if we have the joint density of $\boldsymbol{t}$ denoted by $f(\boldsymbol{t})$, under the null hypothesis, the p-value of the test vector is the probability of obtaining a value of $\boldsymbol{t}$ such that $f(\boldsymbol{t}) < f(\widehat{\boldsymbol{t}})$ holds. Once calculated, the p-value can be used to conduct the test at any level of significance. For example, at the 5% significance level, if the p-value is less than 0.05 then the null hypothesis is rejected. Otherwise, it cannot be rejected. The resultant acceptance region is optimal in the sense that by its construction, it is the smallest $(1-\alpha)$ acceptance region in the d-dimensional sample space of $\boldsymbol{t}$. If we believe $\boldsymbol{t}(\boldsymbol{y})$ provides a good summary of the evidence contained in $\boldsymbol{y}$ that might point to the null hypothesis being false then this is a desirable property to have. We call a test with is property a smallest acceptance region (SAR) test.

Typically the d-dimensional density $f(\boldsymbol{t})$ is unknown. We can estimate it by applying a multivariate kernel density estimator to a sample of independent drawings from $f(\boldsymbol{t})$ which can be obtained by repeatedly simulating the DGP under the null hypothesis and then calculating $\boldsymbol{t}$ for each simulated data set. Because $f(\boldsymbol{t})$ does not depend on unknown parameters, we can set unknown parameters in the DGP to any value we wish in order to conduct these simulations. Let $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m\}$ denote such a sample. The general form of the kernel density estimator of $\boldsymbol{t}$ is

[3] Some preliminary results for one-sided component tests are reported in King et al. (2009).

given by

$$\widehat{f}_{m,H}(\boldsymbol{t}) = \frac{1}{m}\sum_{i=1}^{m}|H|^{-1/2}K(H^{-1/2}(\boldsymbol{t}-\boldsymbol{t}_i)), \tag{2}$$

where $K(\cdot)$ is a kernel function, and H is a positive definite matrix of bandwidths known as the bandwidth matrix (see, for example, Wand and Jones, 1995; Scott, 2015).

There are two ways in which the new testing procedure can be implemented in practice. The first, which we call the double simulation (DS) method, involves two separate rounds of simulation as follows:

(i) Based on the data under test, calculate $\widehat{\boldsymbol{t}}$.

(ii) Using any convenient form of the data generating process under the null hypothesis, simulate the model m times and calculate m independent values of $\boldsymbol{t}$ denoted as $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m$.

(iii) Use the sample generated in (ii) to estimate the joint density $f(\boldsymbol{t})$ by $\widehat{f}_{m,H}(\boldsymbol{t})$ via (2).

(iv) Repeat (ii) to generate a second sample of n values of $\boldsymbol{t}$ denoted as $\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, \cdots, \boldsymbol{t}^{(n)}$, which are independent of those originally calculated in step (ii).

(v) Use this second sample to calculate $\widehat{f}_{m,H}(\boldsymbol{t}^{(i)})$, for $i = 1, 2, \cdots, n$. The p-value of the joint test is estimated by the relative frequency for which $\widehat{f}_{m,H}(\boldsymbol{t}^{(i)}) < \widehat{f}_{m,H}(\widehat{\boldsymbol{t}})$ holds.

The second test procedure involves only one round of simulation and is therefore called the single simulation (SS) method. After completing steps (i) and (ii) above, the remaining steps are as follows:

(iii′) Use the sample generated in (ii) to estimate the joint density $f(\boldsymbol{t})$ at $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m$ by the leave-one-out kernel density estimator

$$\widehat{g}_{m,H}(\boldsymbol{t}_i) = \frac{1}{m-1}\sum_{j=1;j\neq i}^{m}|H|^{-1/2}K(H^{-1/2}(\boldsymbol{t}_i-\boldsymbol{t}_j)). \tag{3}$$

(iv′) The p-value of the joint test is estimated by the relative frequency for which $\widehat{g}_{m,H}(\boldsymbol{t}_i) < \widehat{f}_{m,H}(\widehat{\boldsymbol{t}})$ holds.

If $f(\boldsymbol{t})$ were known, the p-value of the proposed SAR testing procedure would be

$$p_0 = \Pr\left\{\boldsymbol{t} : f(\boldsymbol{t}) < f(\widehat{\boldsymbol{t}}) | f(\boldsymbol{t})\right\}. \tag{4}$$

In our testing procedure, the p-value denoted as $\widehat{p}_{T,m}$, is defined via the kernel density estimator under the probability measure implied by the true density of $\boldsymbol{t}$. This means that[4]

$$\widehat{p}_{T,m} = \Pr\left\{\boldsymbol{t} : \widehat{f}_{m,H}(\boldsymbol{t}) < \widehat{f}_{m,H}(\widehat{\boldsymbol{t}}) | f(\boldsymbol{t})\right\}. \tag{5}$$

In our proposed testing procedure, $\widehat{p}_{T,m}$ is approximated by $\widehat{p}_{T,m,n}$ that is the relative frequency of observing $\widehat{f}_{m,H}(\boldsymbol{t}) < \widehat{f}_{m,H}(\widehat{\boldsymbol{t}})$ in the second-round simulation of Step (v). Because we need to estimate $f(\boldsymbol{t})$ and p_0, the resultant test is no longer a SAR test, but we can consider it to be an approximate SAR test.

The DS method can be regarded as a Monte Carlo test as it fulfills the requirements outlined by Dufour (2006). To confirm this, we first need to view $\widehat{f}_{m,H}(\boldsymbol{t})$ as our test statistic and observe that

$$\widehat{f}_{m,H}\left(\boldsymbol{t}^{(i)}\right), \text{ for } i = 1, 2, \cdots, n \text{ and } \widehat{f}_{m,H}(\widehat{\boldsymbol{t}})$$

all share the same distribution and are exchangeable by construction. Because neither the function $\widehat{f}_{m,H}(\cdot)$ nor $\boldsymbol{t}^{(i)}$, for $i = 1, 2, \cdots, n$, depend on unknown parameter values, this shared distribution is independent of unknown parameters. In the case where there is a zero probability that any two $\widehat{f}_{m,H}\left(\boldsymbol{t}^{(i)}\right)$ values, for $i = 1, 2, \cdots, n$, will be equal, our test procedure applied at the α level of significance is an exact test provided $\alpha(n+1)$ is an integer (Dufour, 2006, Proposition 2.2). This does not require n to be particularly large. If there is a non-zero probability of two $\widehat{f}_{m,H}\left(\boldsymbol{t}^{(i)}\right)$ values being equal, then our procedure can be amended as outline in Dufour's (2006) Proposition 2.3 to make it exact.

[4]Equation (4) defines a low density region which is the complement of a high density region. Samworth and Wand (2010) investigated the asymptotic properties of the plug-in type bandwidth selector for kernel density estimation that aims to define a high density region. Recent studies on high density regions include Tsybakov (1997), Cadre (2006), Mason and Polonik (2009) and Cadre, Pelletier, and Pudlo (2013) among others.

2.2 Bandwidth selection

The multivariate kernel density estimator depends on the choice of a bandwidth matrix and the choice of a kernel function. It is generally accepted in the statistical literature that the performance of the kernel density estimator is mainly determined by the bandwidth matrix, and only in a minor way by the choice of a kernel function. The bandwidth matrix can be either a full matrix or a diagonal matrix. A full bandwidth matrix is able to incorporate any possible correlation between any pair of the d dimensions. However, the number of nonzero bandwidths to be estimated in a full bandwidth matrix grows dramatically as d increases. Consequently, a full bandwidth matrix encounters more computing complexity in selecting a bandwidth matrix that is optimal with respect to a chosen criterion than a diagonal bandwidth matrix does. As discussed by Wand and Jones (1993) in the situation of bivariate kernel density estimation, a diagonal bandwidth matrix allows for the flexibility of choosing a different bandwidth in each dimension and is often appropriate. Therefore, we use a diagonal bandwidth matrix in this new testing procedure, where the bandwidth matrix is denoted as $H = \text{diag}\{h_1^2, h_2^2, \cdots, h_d^2\}$.

According to Scott (2015) and Bowman and Azzalini (1997), when data are observed from the multivariate normal density and the diagonal bandwidth matrix is used, the optimal bandwidth matrix that minimizes the mean integrated squared error (MISE) between the true density and its estimator can be approximated by

$$h_i = \sigma_i \left\{ \frac{4}{(d+2)m} \right\}^{1/(d+4)},$$

for $i = 1, 2, \cdots, d$, where σ_i is the standard deviation of the ith variate and can be replaced by its sample estimator in practice. We call this the normal reference rule (NRR) which is also known as the rule-of-thumb method in the literature. This bandwidth selection method is often used in many applications of multivariate kernel density estimation in the absence of any other practical bandwidth selection methods, despite the fact that the data might not be Gaussian.

Zhang, King, and Hyndman (2006) presented a Markov chain Monte Carlo (MCMC) sampling algorithm to estimate the bandwidth matrix in multivariate kernel density estimation. The bandwidth matrix chosen through this sampling algorithm tends to produce a more accurate density estimator than that chosen through the NRR. However, the MCMC bandwidth selector is far more computationally costly than the NRR. A general guideline for selecting one of the two bandwidth selectors is as follows. When the required computing time is not of serious concern in the testing procedure, one may use the MCMC bandwidth selector. Otherwise, one may use the NRR to choose bandwidths.

As the performance of the kernel density estimator is only slightly affected by the choice of a kernel function, we will not investigate the issue of the choice of kernel. Throughout this paper, we use the product of d univariate Gaussian kernels in the kernel density estimator of $f(\boldsymbol{t})$.

3 Monte Carlo experiments for invariant test statistics

We conducted two separate Monte Carlo experiments in order to study the small sample size and power performance of the new test procedure, where the test vector's null distribution does not depend on nuisance parameters. The testing problems involved are (i) testing for serial correlation in a stationary time series; and (ii) testing for normality in a simple random sample. In each case we compared the performance of four versions of our test procedure, the double and single simulation methods, each using the MCMC bandwidth selector and the NRR bandwidth vector with three existing tests.

As simulations of simulations can be very time consuming, we used the following approach to estimate the size and power of the test procedure in the case of the DS method.

(a) Choose convenient values of the parameters in the DGP under the null hypothesis and repeatedly simulate the DGP to obtain two independent simple random samples of m values of $\boldsymbol{t}$, denoted as $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m$, and $\boldsymbol{t}_1^*, \boldsymbol{t}_2^*, \cdots, \boldsymbol{t}_m^*$.

(b) For the particular choice of bandwidth matrix (NRR or MCMC), compute the value of kernel density given by (2), for $\boldsymbol{t}_i^*$, namely $\widehat{f}_{m,H}(\boldsymbol{t}_i^*)$, for $i = 1, 2, \cdots, m$.

(c) Simulate the DGP under which size or power is to be estimated and calculate a third simple random sample of n values of $\boldsymbol{t}$ denoted as $\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, \cdots, \boldsymbol{t}^{(n)}$. For each $\boldsymbol{t}^{(j)}$, count the number of $\widehat{f}_{m,H}(\boldsymbol{t}_i^*)$ values, for which $\widehat{f}_{m,H}(\boldsymbol{t}_i^*) < \widehat{f}_{m,H}(\boldsymbol{t}^{(j)})$. The estimated probability of rejection of the null hypothesis is the relative frequency that

$$(\text{count}+1)/(n+1) < \alpha,$$

holds for $j = 1, 2, \cdots, n$.

For the SS method, only one simple random sample at Step (a) is calculated, denoted as $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m$. In place of $\widehat{f}_{m,H}(\boldsymbol{t}_i^*)$ in Steps (b) and (c), values of the leave-one-out kernel density $\widehat{g}_{m,H}(\boldsymbol{t}_i)$, for $i = 1, 2, \cdots, m$, are used.

3.1 Testing for serial correlation of unknown order and form

3.1.1 Experiment design

The first Monte Carlo experiment involves the classical problem of testing the null hypothesis that an observed time series is white noise against the alternative that it contains serial correlation of unknown order and form (see King, 1987). The null hypothesis is of the form

$$y_t = \mu + \varepsilon_t, \qquad \text{for } t = 1, 2, \cdots, T, \tag{6}$$

where μ is an unknown parameter and ε_t are independent and identically distributed (iid) as $N(0, \sigma^2)$. The alternative is that there is some serial correlation in ε_t and it is assumed that it can best be detected by examining r_j, the jth order autocorrelation coefficient, for $j = 1, 2, \cdots, d$, where

$$r_j = \frac{\sum_{t=j+1}^{T} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-j}}{\sum_{t=1}^{T} \widehat{\varepsilon}_t^2},$$

in which $\widehat{\varepsilon}_t$, for $t = 1, 2, \cdots, T$, are the ordinary least squares (OLS) residuals from fitting (6) to the observed time series. In other words,

$$t_j = r_j, \qquad \text{for } j = 1, 2, \cdots, d,$$

in this problem. Note that r_j is invariant to the values of μ and σ^2 under (6) and so $f(\boldsymbol{t})$ does not depend on nuisance parameters.

A standard testing procedure for this problem is to use the Portmanteau test proposed by Box and Pierce (1970) and extended by Ljung and Box (1978). It involves rejecting the null hypothesis for large values of the Portmanteau test statistic given by

$$Q_d = T(T+2)\sum_{j=1}^{d}\frac{r_j^2}{T-j}. \tag{7}$$

The Monte Carlo experiment involved comparing sizes and powers of the Portmanteau test based on (7) applied using simulated critical values with four forms of the new procedure, the first two being the SS and DS methods using NRR bandwidth parameters and the second two being the SS and DS methods using MCMC bandwidth parameters for $d = 4$ and $d = 6$, respectively.

Also included in the comparison are two-sided versions of the minimum p-value and product of p-value tests from Dufour, Khalaf, and Voia (2015). Because both the Portmanteau test and our new procedure are two-sided in nature, we applied two-sided versions of the minimum p-value and product of p-value tests as follows:

(i) Assume under H_0, $\sqrt{T}r_j \sim N(0,1)$ asymptotically (equation (39) in Dufour, Khalaf, and Voia, 2015). Let $s_j = \sqrt{T}r_j$. If $r_j > 0$, then calculate

$$\delta_j = \Pr\left(z > s_j | z \sim N(0,1)\right),$$

otherwise ($r_j \leq 0$), calculate

$$\delta_j = \Pr\left(z < s_j | z \sim N(0,1)\right),$$

for $j = 1, \cdots, d$.

(ii) The required individual p-values are $p_j = 2\delta_j$, and the test statistic for the minimum p-value test is

$$p = \min_{j=1,\cdots,d} p_j,$$

while that of the product of p-values test is

$$q = \prod_{j=1}^{d} p_j.$$

(iii) The tests are applied by first calculating p and q (denoted as $\widehat{p}$ and $\widehat{q}$) using the actual data and then simulating the calculation of p and q n times under the null hypothesis. The number of simulated values of p (and q) smaller than $\widehat{p}$ (and $\widehat{q}$) is counted (denoted as count), allowing the empirical p-value to calculated as

$$\widehat{p}_n = (\text{count} + 1)/(n+1).$$

Provided n is chosen so that $\alpha(n+1)$ is an integer, then these two tests are exact tests at the α level of significance. We set $n = 19,999$.

Sizes were calculated by simulating the DGP using (6) with $\varepsilon_t \sim IN(0,1)$ and $\mu = 1$. Powers were calculated for four different DGPs for ε_t in (6), these being ε_t generated by

(i) the stationary first-order autoregressive (AR(1)) process given by

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

or equivalently $(1-\rho L)\varepsilon_t = u_t$, where $\rho = 0.25$, $u_t \sim IN(0,1)$ and L is the lag operator;

(ii) the stationary second-order autoregressive (AR(2)) process given by

$$(1-\rho_1 L)(1-\rho_2 L)\varepsilon_t = u_t$$

with $(\rho_1, \rho_2) = (0.05, 0.1)$ and $(0.05, 0.2)$ and $u_t \sim IN(0,1)$;

(iii) the stationary third-order autoregressive (AR(3)) process given by

$$(1-\rho_1 L)(1-\rho_2 L)(1-\rho_3 L)\varepsilon_t = u_t$$

with $(\rho_1, \rho_2, \rho_3) = (0.05, 0.1, 0.15)$ and $(0.05, 0.1, 0.2)$ and $u_t \sim IN(0,1)$; and

(iv) the stationary fourth-order autoregressive (AR(4)) process given by

$$(1-\rho_1 L)(1-\rho_2 L)(1-\rho_3 L)(1-\rho_4 L)\varepsilon_t = u_t$$

with $(\rho_1, \rho_2, \rho_3, \rho_4) = (0.05, 0.1, 0.15, 0.15)$, $(0.05, 0.1, 0.05, 0.05)$ and $(0.05, 0.05, 0.05, 0.05)$ and $u_t \sim IN(0,1)$.

All DGPs were run for $T = 50, 100, 200$ and 500; all tests were applied at the 10%, 5% and 1% significance levels and, for the new test procedure, m and n were set to 19,999.

3.1.2 Simulation results

The size and power results are presented in Tables 1–3. With respect to sizes, all seven tests have appropriate sizes for $T = 50$, 100 and 200. There is a tendency for all tests to be slightly under size for $T = 500$. Given the exact tests also show this tendency, we believe it is caused by randomness. While there are minor differences in sizes between the SS and DS methods, none are statistically significant. In fact there are no discernible differences in sizes between any pair of tests. All appear to be doing an excellent job of controlling size.

With respect to power, the new test does very well for $T = 50$ with all versions almost always being more powerful than the best of the existing tests. For larger sample sizes, they perform relatively better for larger significance levels. There is also a tendency for their relative power performance to be better for $d = 4$ compared to $d = 6$. The MCMC based versions typically are slightly more powerful than their respective NRR based test. Overall there is a pattern of the new tests being more powerful than the existing tests. This is best illustrated by the following summary statistics for the MCMC based DS (SS) versions. Out of the 64 combinations of d, T and AR process, the new test is more powerful or at least as powerful as the most powerful of the existing tests on 51 (52) occasions when $\alpha = 0.1$, 46 (46) times when $\alpha = 0.05$, and 35 (31) occasions when $\alpha = 0.01$.

The minimum p-value test is almost always the most powerful test for $T \geq 100$ against the AR(1) process with its relative performance being better for $d = 6$ compared to $d = 4$. It also has the best power for $T = 500$ against near AR(1) processes. On the other hand, for all AR(4) processes, as well as the AR(3) process with equal ρ_j values, it has the worst power of all seven tests. The product of p-values test is the best only occasionally when $d = 6$ and more often for $d = 4$ when it favors lower values of α, larger samples and more complicated AR processes.

3.2 Testing for normality

3.2.1 Experiment design

In many statistical situations, random observations are often assumed to be normally distributed for the purpose of statistical inferences. Therefore, it is important to be able to test for normality (see, for example, Shapiro and Wilk, 1965; D'Agostino, 1971, 1972; Bowman and Shenton, 1975; Pearson, D'Agostino, and Bowman, 1977; Jarque and Bera, 1980, 1987; Spiegelhalter, 1980; Thode, 2002; Dufour, Khalaf, and Beaulieu, 2003; Dufour, Farhat, and Khalaf, 2004).

The second Monte Carlo experiment involved the problem of testing the null hypothesis that a simple random sample is independently and identically normally distributed with unknown mean (μ) and unknown variance (σ^2) against the alternative that it is non-normally distributed. In other words, (6) is the model for the null hypothesis. Evidence of non-normality is often obtained from sample measures of skewness and kurtosis denoted as $\sqrt{b_1}$ and b_2, respectively, where

$$b_1 = \widehat{\mu}_3^2 / \widehat{\mu}_2^3, \qquad b_2 = \widehat{\mu}_4 / \widehat{\mu}_2^2,$$

and $\widehat{\mu}_i = \sum_{t=1}^{T} (y_t - \widehat{\mu})^i / T$, for $i = 2, 3, 4$, with $\widehat{\mu} = \sum_{t=1}^{T} y_t / T$. Jarque and Bera (1980, 1987), D'Agostino and Stephens (1986), Urzúa (1996) and Thode (2002) have discussed omnibus tests for normality that combine information from $\sqrt{b_1}$ and b_2, and Monte Carlo tests have been recommended by Dufour, Farhat, Gardiol, and Khalaf (1998), Dufour and Khalaf (2001), Dufour, Khalaf, and Beaulieu (2010) and Dufour, Farhat, and Khalaf (2004). As Dufour and Khalaf (2001) observe, the joint distribution of $\sqrt{b_1}$ and b_2 under the null hypothesis of normality, is independent of nuisance parameters

This experiment involves comparing the small-sample properties of four versions of our test procedure based on the test vector $(\sqrt{b_1}, b_2)'$ in which $\sqrt{b_1}$ takes the sign of $\widehat{\mu}_3$, with the Jarque-Bera test (Jarque and Bera, 1980, 1987) whose test statistics is

$$\mathrm{JB} = T\left[\frac{(\sqrt{b_1}\,)^2}{6} + \frac{(b_2 - 3)^2}{24}\right]. \tag{8}$$

The versions of the new test procedure used are the four combinations of the DS and SS methods with the NRR or MCMC bandwidth selector. Note that the DS method yields an exact test in this case because n was set to 19,999.

A small number of simulation studies have revealed that the size of the JB test is incorrect for small- and moderate-sized samples particularly in the context of the linear regression model. A more straightforward solution is to use Monte Carlo simulations to obtain correct critical values, which is the approach we used in this study (see, for example, Dufour and Khalaf, 2001; Poitras, 2006).

Also included in the comparison were two-sided versions of the minimum p-value and product of p-value tests from Dufour, Farhat, and Khalaf (2004) and Dufour (2006). We used

$$\sqrt{T/6}\sqrt{b_1} \sim N(0,1) \text{ and } \sqrt{T/24}(b_2-3) \sim N(0,1), \text{ asymptotically,}$$

in order to calculate the individual p-values required for these tests.

Sizes were calculated by simulating (6) with $\varepsilon_t \sim IN(0,1)$. Powers were calculated for four alternative distributions for ε_t, these being the Student's t distribution with 5 degrees of freedom denoted t_5, the χ^2 distribution with 3 degrees of freedom denoted χ^2_3, the Gamma distribution with shape and scale parameters being 2 and 1, and the log normal distribution with mean zero and standard deviation 0.5. These tests were compared for sample sizes of $T = 30, 50, 75,$ and 100 with the values of m and n both being 19,999.

3.2.2 Simulation results

The size and power results are given in Table 4. All seven tests have excellent sizes, and there is no discernible difference between the estimated sizes of the seven tests, including between the SS and DS versions of our test.

With regards to power, a version of the new test procedure is more often than not the most powerful (or equal most powerful) of the seven tests. Across the 48 combinations of α, T and non-normality, an existing test is more powerful than every version of the new test on only 14 occasions. Unlike for the previous case of testing for autocorrelation, no particular version of the new test dominates. Collectively, the relative power of the new tests are best against t_5 and

χ^2_3 distributions and weakest against the log-normal distribution, particularly for $T = 30$ and 50. Against all distributions, the relative power of the new test is poorest for $T = 30$ and for $\alpha = 0.10$ when the true distribution is non-symmetric, With a few exceptions, the product of p-values test is typically the least powerful of the seven tests for $\alpha = 0.01$ and 0.05, while the minimum p-value test is generally the least powerful for $\alpha = 0.10$ and $T \geq 50$, except against the log-normal distribution. The JB test has best power against non-symmetric distributions for larger T values and $\alpha = 0.10$,

Out of the three existing tests, the minimum p-value test has the highest power almost always for $\alpha = 0.01$ and against non-symmetric distributions for $\alpha = 0.05$. The JB test is typically the most powerful of the three when $\alpha = 0.10$. The product of p-values test also performs relatively well for $\alpha = 0.10$ and has good relative power for $\alpha = 0.05$ against the t_5 distribution.

4 The testing procedure for non-invariant test statistics

4.1 Testing procedure

So far we have concentrated on statistics whose distribution under the null hypothesis does not depend on nuisance parameters. In this case, there is no issue of how to simulate $\boldsymbol{t}$ under the null hypothesis in order to estimate its density because we can choose any null hypothesis DGP for this purpose. When the distribution of $\boldsymbol{t}$ under the null hypothesis depends on the value of one or more nuisance parameters, which we denote by γ, then we have a more complex testing problem. In particular, the unknown density $f(\boldsymbol{t})$ now becomes a function of γ and will be denoted as $f(\boldsymbol{t}, \gamma)$. We are assuming that the DGP under the null hypothesis is fully parameterized in the sense that if γ is known then $f(\boldsymbol{t}, \gamma)$ is a single distribution for all DGPs under the null hypothesis with this value of γ.

Conditional on a value for γ, denoted as γ^*, we can apply either of the procedures outlined in Section 2.1, by simulating the DGP under the null hypothesis, with γ being set to γ^* and any remaining parameters being set to convenient values because $f(\boldsymbol{t}, \gamma)$ does not depend on their values. The kernel density $\widehat{f}_{m,H}(\boldsymbol{t})$ given by (2) and the leave-one-out kernel density estimator

(3) now become functions of γ^*, denoted as $\widehat{f}_{m,H}(\boldsymbol{t},\gamma^*)$ and $\widehat{g}_{m,H}(\boldsymbol{t},\gamma^*)$, respectively. In both cases, the resultant p-value is dependent on our choice of γ^* and does not result in a test that effectively controls size, either exactly or approximately, unless γ^* is chosen with controlling size in mind.

In the case of the DS procedure, we can view $\widehat{f}_{m,H}(\boldsymbol{t},\gamma^*)$ as our statistic and note that there is a value of γ^* (namely γ_0, the true value of γ under the null hypothesis) for which, conditional on γ^*,

$$\widehat{f}_{m,H}\left(\boldsymbol{t}^{(i)},\gamma^*\right),\ \text{for } i=1,2,\cdots,n \text{ and } \widehat{f}_{m,H}(\hat{\boldsymbol{t}},\gamma^*)$$

all share the same distribution and are exchangeable. Assuming there is a zero probability of any two $\widehat{f}_{m,H}(\boldsymbol{t}^{(i)},\gamma^*)$ values being exactly equal, this implies that our DS test procedure fulfils the conditions of Dufour (2006) Proposition 4.1 so that the MMC test is an exact α-level test provided n is chosen so that $\alpha(n+1)$ is an integer.

If we denote the p-value from the DS procedure with $\gamma=\gamma^*$ as $\widehat{p}_{T,m,n}(\gamma^*)$, the MMC test in this case involves calculating

$$\sup_{\gamma^*} \widehat{p}_{T,m,n}(\gamma^*) \tag{9}$$

and using it as the test's p-value. As noted in the Introduction, the MMC test is very likely to be a conservative test with the probability of a Type I error being well below the desired level, α, when $\gamma=\gamma_0$. Also, our experience with the simulated annealing algorithm which is used to solve (9), is that it can be a very computationally demanding algorithm. Dufour's (2006) Consistent Set Estimate MMC test mitigates these concerns to some degree although this test is only exact asymptotically.

The ideal solution is to set γ^* to the true value of γ, namely γ_0, under the null hypothesis. Unfortunately, it is very unlikely we will know γ_0. Our preferred approach involves what might be considered the next best option and that is to set γ^* equal to a consistent estimate of γ under the null. This involves the following variation to Step (ii) in the procedure given in Section 2.1:

(ii′) Estimate γ using the available T observations and assuming the null hypothesis is true and denote this estimate as $\widehat{\gamma}_T$. Using $\gamma = \widehat{\gamma}_T$ and any convenient values of the remaining parameters in the model under the null hypothesis, simulate the model m times and calculate m independent values of $\boldsymbol{t}$ denoted as $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m$.

The remainder of the procedure is as outlined in Section 2.1. There can be situations in which it is difficult to determine which parameters $\boldsymbol{t}$ depends on under the null hypothesis. In these cases, we can use estimated values of all the model's parameters to be safe when simulating $\boldsymbol{t}$, without any loss.

4.2 Consistency of the estimated p-value

Let $\widehat{f}_m(\boldsymbol{t}, \widehat{\gamma}_T)$ denote the kernel density estimator of $f(\boldsymbol{t}, \gamma)$ obtained during the first-round simulation involving m replications, where the bandwidth symbol H is omitted for simplicity. Let p_0 denote the p-value defined through the true density of $\boldsymbol{t}$ as follows.

$$p_0 = \Pr\left\{\boldsymbol{t} : f(\boldsymbol{t}, \gamma_0) < f(\widehat{\boldsymbol{t}}, \gamma_0) \middle| f(\boldsymbol{t}, \gamma_0)\right\}, \tag{10}$$

which in our testing procedure, is estimated by $\widehat{p}_{T,m}$, the p-value defined via the kernel density estimator under the probability measure implied by the true density of $\boldsymbol{t}$. This means that

$$\widehat{p}_{T,m} = \Pr\left\{\boldsymbol{t} : \widehat{f}_m\big(\boldsymbol{t}, \widehat{\gamma}_T\big) < \widehat{f}_m\big(\widehat{\boldsymbol{t}}, \widehat{\gamma}_T\big) \middle| f\big(\boldsymbol{t}, \widehat{\gamma}_T\big)\right\}. \tag{11}$$

In our testing procedure, $\widehat{p}_{T,m}$ is approximated by $\widehat{p}_{T,m,n}$, the relative frequency of observing $\widehat{f}_m(\boldsymbol{t}^{(i)}, \widehat{\gamma}_T) < \widehat{f}_m(\widehat{\boldsymbol{t}}, \widehat{\gamma}_T)$, for $i = 1, 2, \cdots, n$, during the second-round simulation of Step (v).

Assumption 1: $\widehat{\gamma}_T$ is a strongly consistent estimate of γ_0 under H_0.

Assumption 2: $f(\boldsymbol{t}, \gamma)$ is continuous in γ.

Assumption 3: $f(\boldsymbol{t}, \gamma)$ as a density function of $\boldsymbol{t}$, meets the smoothness conditions given in Masry (1996).

Assumption 4: The kernel function $K(\cdot)$ and the bandwidth matrix H are chosen to ensure the kernel density estimator is uniformly consistent on a bounded set in which $f(\boldsymbol{t}, \gamma) > 0$.

Theorem 1: Under the assumptions 1 to 4, as $T \to \infty$, $m \to \infty$ and $n \to \infty$, the estimated p-value denoted as $\widehat{p}_{T,m,n}$ is a consistent estimate of the true p-value of our testing procedure.

Proof: The proof of Theorem 1 is given in the Appendix.

Assumptions 1–4 are different to those imposed by Dufour (2006, Section 6). In our set up, his condition relate to $\widehat{f}_{m,H}\left(\boldsymbol{t}^{(i)},\gamma^*\right)$ and $\widehat{f}_{m,H}\left(\widehat{\boldsymbol{t}},\gamma^*\right)$. Our assumption 1 is satisfied by many standard econometric estimators, although each comes with its own regularity conditions. Assumption 2 imposes some restrictions on the type of testing problem we can consider. Assumption 3 is rather mild and requires the absolute difference between the joint density of any two components of $\boldsymbol{t}$ and the product of their marginal densities to be bounded. Our choices of kernel function and bandwidth selectors satisfy Assumption 4.

We have therefore shown that the DS procedure based on using a strongly consistent estimate of the nuisance parameter γ in order to simulate $\boldsymbol{t}$ when required, results in an asymptotically justified test. It is also an asymptotically SAR (ASAR) test given that p_0 is the p-value for a SAR test.

5 Information matrix test

It is often important to test whether a model is correctly specified. White (1982) showed that when a model is correctly specified and estimated by maximizing the likelihood function, the information matrix should be asymptotically equal to the negative Hessian matrix. The information matrix test introduced by White (1982), aims to test the significance of the discrepancy between the negative Hessian and the outer product of the score vector, where the lower triangular components of the matrix of such differences are organized into a vector which we call the test vector in this paper. Chesher (1984) showed that the IM test can be viewed as a Lagrange multiplier (LM) test for specification error against the alternative of parameter heterogeneity. Chesher (1983) and Lancaster (1984) presented an TR^2 version of the IM test, where T is the sample size and R^2 is the goodness of fit obtained through the OLS regression of a column of ones on a matrix whose elements are functions of the first and second derivatives of the log-likelihood function. For the normal fixed regressor linear model, Hall (1987) showed that the LM version of the IM test can be asymptotically decomposed into the sum of three components, where one is the general test for heteroscedasticity proposed by White (1982), and the other two components aim to test certain forms of normality.

The use of the IM test in applied econometrics is limited because its actual size obtained according to the asymptotic critical value often differs greatly from its nominal size. This phenomenon has been detected in Monte Carlo experiments conducted by Taylor (1987), Orme (1990), Chesher and Spady (1991) and Davidson and MacKinnon (1992). Davidson and MacKinnon (1992) proposed to deal with this problem by using the double-length artificial regressions to compute a variant of the IM test statistic, but models for discrete, censored or truncated data cannot be dealt with via this method. Chesher and Spady (1991) suggested obtaining the critical value for the IM test from the Edgeworth expansion through order $O(T^{-1})$ of the finite-sample distribution of the test statistic. Their Monte Carlo investigation indicates that such a critical value provides a good approximation to the true critical value obtained through the exact distribution of the IM test, and such an approximation was found to be superior to the usual χ^2 approximation in some cases. In the examples considered by Chesher and Spady (1991), the Edgeworth expansions are independent of the parameters of the models being tested, and therefore, the IM test statistic is pivotal (Horowitz, 1994). However, this is not a general case, and it is often very difficult to decide whether the IM test statistic is pivotal.

Horowitz (1994) proposed a bootstrapping procedure to obtain critical values for the IM test and demonstrated the capability of bootstrapping to overcome the incorrect-size problem in finite samples. He showed that one can easily obtain good finite-sample critical values for the IM test through bootstrapping rather than through Edgeworth expansions or other algebraically complicated manipulations. Moreover, he discussed the power performance of three versions of the IM test through Monte Carlo simulation. His results showed that all three versions of the IM test considered have much lower powers computed using size-corrected critical values than those computed according to asymptotic critical values. Therefore, it seems that getting the size right and achieving higher power are different tasks.

Most existing versions of the IM test rely on an estimate of the asymptotic covariance matrix of the test vector. The analytical form of the asymptotic covariance matrix is complicated and involves the third derivative of the log-likelihood function. Lancaster (1984) showed that the covariance matrix of White's (1982) IM test can be estimated without calculating the third derivative of the log-likelihood. Dhaene and Hoorelbeke (2004) indicated that the incorrect-size problem results from the inaccurate estimate of the covariance matrix. They proposed to estimate the covariance matrix of the test vector through parametric bootstrapping.

Let $f(y|\theta)$ denote the density for a postulated model where θ is a $r \times 1$ vector of parameters. Let $\boldsymbol{y} = (y_1, y_2, \cdots, y_T)'$ be the vector of observations, and $\ell(\boldsymbol{y}|\theta) = \log f(\boldsymbol{y}|\theta)$ the logarithmic density. We introduce the following notation.

$$A(\theta) = E\left[\frac{\partial^2 \ell(\boldsymbol{y}|\theta)}{\partial\theta\partial\theta'}\right], \qquad A_T(\boldsymbol{y},\theta) = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \ell(y_t|\theta)}{\partial\theta\partial\theta'},$$
$$B(\theta) = E\left[\frac{\partial \ell(\boldsymbol{y}|\theta)}{\partial\theta}\frac{\partial \ell(\boldsymbol{y}|\theta)}{\partial\theta'}\right], \qquad B_T(\boldsymbol{y},\theta) = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \ell(y_t|\theta)}{\partial\theta}\frac{\partial \ell(y_t|\theta)}{\partial\theta'},$$

where expectations are taken with respect to the true density. When the model is correctly specified, the true density is $f(\boldsymbol{y}|\theta)$. Let θ_0 be the true value of θ.

The information matrix test is based on the information-matrix equality, which states that $A(\theta_0) + B(\theta_0) = 0$ when the model is correctly specified. Given the vector of T independent observations, $\boldsymbol{y}$, the information-matrix test investigates the statistical significance of $A_T(\boldsymbol{y},\widehat{\theta}) + B_T(\boldsymbol{y},\widehat{\theta})$, where $\widehat{\theta}$ is the maximum likelihood estimator of θ.

Let $\boldsymbol{t}$ denote the vector of indicators (test vector) whose elements are D_{ij}, for $i = 1, 2, \cdots, r$, and $j = 1, \cdots, i$, where

$$D_{ij} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial \ell(y_t|\theta)}{\partial\theta_i}\frac{\partial \ell(y_t|\theta)}{\partial\theta_j} + \frac{\partial^2 \ell(y_t|\theta)}{\partial\theta_i\partial\theta_j}\right]\Big|_{\theta=\widehat{\theta}}. \tag{12}$$

White (1982) shows that under regularity conditions, the IM test statistic is of the form

$$\xi = T\boldsymbol{t}'\widehat{V}^{-1}\boldsymbol{t}, \tag{13}$$

where $\widehat{V}$ is the consistent estimator of the covariance matrix of $\boldsymbol{t}$ under H_0. Under the null hypothesis, ξ is distributed asymptotically χ_q^2 with $q = r(r+1)/2$, and this is based on the asymptotic null distribution of $\boldsymbol{t} \sim N(0, V(\theta))$. The IM test statistic depends on the estimate of the covariance matrix which in turn depends on the estimate of θ.

Our proposed IM testing procedure aims to estimate the joint density of the test vector, $\boldsymbol{t}$, and does not depend on $\widehat{V}$. We believe basing the test directly on the vector $\boldsymbol{t}$ and not needing an estimate of the covariance matrix of $\boldsymbol{t}$ under H_0 will give our proposed test a clear advantage. However, the null distribution of $\boldsymbol{t}$ is highly dependent on θ, the unknown model parameters.

Since $A_T(\boldsymbol{y}, \widehat{\theta}) + B_T(\boldsymbol{y}, \widehat{\theta})$ is a symmetric matrix, a test of the complete IM identity can be based on the lower triangular elements of $A_T(\boldsymbol{y}, \widehat{\theta}) + B_T(\boldsymbol{y}, \widehat{\theta})$ (or D_{ij}). However, according to White (1982), in many situations it is inappropriate to base the test on all q indicators because some indicators may be identically zero, furthermore, some indicators may be linear combinations of other indicators. In either case, it is appropriate to ignore such indicators. In the remainder of this paper, the IM tests are based on the maximum number of linearly independent non-zero indicators.

6 Simulation study of the new test applied to the IM test

This section reports a Monte Carlo simulation study which aims to compare the finite-sample size and power performance of our proposed SS method based on $\boldsymbol{t}$ with the Lancaster (1984) form of the IM test denoted by $\mathrm{IM_L}$ and the Dhaene and Hoorelbeke (2004) form of the IM test denoted by $\mathrm{IM_{DH}}$.

6.1 Null and alternative hypotheses

The study covers two different settings. The null hypothesis in the first setting is the normal linear regression model given by

$$y_t = x_t'\beta + u_t, \tag{14}$$

for $t = 1, 2, \cdots, T$, where $u_t \sim IN(0, \sigma^2)$, x_t is a $k \times 1$ vector of regressors, and β is a $k \times 1$ vector of parameters. Following Dhaene and Hoorelbeke (2004), we examine the power of the IM test under the heteroscedastic alternative of

$$y_t = x_t'\beta + u_t, \qquad u_t \sim N\left(0, |x_t'\beta|\right), \qquad \text{for } t = 1, 2, \cdots, T. \tag{15}$$

In the second setting, the null model is the Tobit model given by

$$y_t = \begin{cases} x_t'\beta + u_t & \text{if } x_t'\beta + u_t > 0 \\ 0 & \text{if } x_t'\beta + u_t \leq 0 \end{cases}, \tag{16}$$

for $t = 1, 2, \cdots, T$, $u_t \sim N(0, \sigma^2)$, x_t is a $k \times 1$ vector of regressors, and β is a $k \times 1$ vector of parameters. It will be convenient to re-parameterize the model as

$$hy_t = \begin{cases} x_t'b + v_t & \text{if } x_t'b + v_t > 0 \\ 0 & \text{if } x_t'b + v_t \leq 0 \end{cases}, \tag{17}$$

where $h = 1/\sigma$, $b = \beta/\sigma$, and $v_t \sim N(0, 1)$, for $t = 1, 2, \cdots, T$.

Following Horowitz (1994), we examine the power of the IM tests under the models given by

$$y_t = \max(0, x_t'\beta + u_t), \qquad u_t \sim N\left(0, \exp(0.5\, x_t'\beta)\right), \tag{18}$$

and

$$y_t = \max\left(0, x_t'\beta + 0.75 x_{t,2} x_{t,3} + u_t\right), \qquad u_t \sim N(0, 1), \tag{19}$$

for $t = 1, 2, \cdots, T$, where $x_{t,2}$ and $x_{t,3}$ are the two non-intercept components of x_t. Note that model (18) involves a heteroscedastic alternative while model (19) has an incorrect mean function.

The experiments consist of applying both forms of IM tests along with the proposed method of testing to the linear regression and Tobit models. In both models, x_t consists of an intercept component and either one or two additional variables. The values of x_t are fixed in repeated samples. The values of the β parameters are $(0.75, 1)'$ and $(0.75, 1, 1)'$ when x_t is 2×1 and 3×1, respectively. The non-intercept components of x_t are sampled independently from the standard normal distribution. The value of σ^2 is set to 1 in all of the experiments. The sample sizes are 50, 100, 200 and 300, and $n = m = 1000$. The $\boldsymbol{t}$ vector is 5×1 in the one-regressor case and 9×1 in the two-regressor case. Size-corrected critical values, which were obtained via simulation under the null hypothesis with known true parameters, were used for computing the powers of the IM_L and IM_{DH} tests. For the IM_{DH} test statistic, we used 50 parametric bootstrap samples to estimate the covariance matrix, $\widehat{V}$, following Dhaene and Hoorelbeke (2004). It should be noted that when the IM test statistic is not pivotal (i.e. for the Tobit model), these size-corrected critical values for the IM_L and IM_{DH} tests cannot be calculated in a practical application because the true parameter values under the null hypothesis are unknown. We have used these critical values in the simulation so that the powers of the respective tests can be compared fairly.

In a separate simulation using recommended χ^2 critical values, we found estimated sizes for the IM_L test and $\alpha = 0.10$, 0.05 and 0.01 of 0.712, 0.611 and 0.417 respectively, when $T = 50$ and 0.553, 0.462 and 0.298 when $T = 100$. For the IM_{DH} test we did see an improvement in these estimated sizes to, respectively, 0.164, 0.114 and 0.064 when $T = 50$ and for $T = 100$, to 0.151, 0.104 and 0.048. All these estimated sizes are highly significantly larger than their nominal sizes. Clearly the use of χ^2 critical values for these two tests is not appropriate, particularly if we wish to compare powers.

6.2 Simulation results from the linear model

The results for the linear model are presented in Tables 5 and 6 for sizes and powers, respectively. From Table 5, we see that the sizes derived through the proposed test and both versions of IM test are very close to their corresponding nominal sizes for both one-regressor and two-regressors models. In fact none of the estimated sizes are significantly different from their nominal values. This is particularly encouraging for the new procedure given it involves estimation of a nine-dimensional joint density in the two-regressor case.

From Table 6, we see that the power of the proposed method is, with one exception when $T = 50$ and $\alpha = 0.10$ in the two-regressor case, always more powerful than both the IM_L and IM_{DH} tests, typically by a large margin. We see settings in which the new test provides four or five-fold increases in power, although this advantage declines as the powers approach one. The fact that the IM_{DH} test is always more powerful than the IM_L test demonstrates that there is a power advantage in using bootstrap methods to estimate the covariance matrix. Our proposed procedure, which by-passes this estimation problem, provides an even bigger power advantage.

6.3 Simulation results from the Tobit model

The size and power results for the Tobit model are presented in Tables 7–9. In Table 7, we see that the estimated sizes for the proposed testing procedure are not significantly different to their nominal sizes for both the one-regressor and two-regressor models. On the other hand, the sizes for the IM_{L} and IM_{DH} tests have mixed behavior. At the 1% level, the sizes seem to be over rejecting the null hypothesis particularly for smaller sample sizes. At the 5% and 10%, levels, the sizes are close to their nominal levels with only one occurrence of significant over rejection.

Table 8 presents the estimated powers of the tests, when model (18) is the true alternative hypothesis. The new test is always more powerful than both the IM_{L} and IM_{DH} tests with the one exception of when the powers reach their maximum of one. The biggest improvements in power occur for smaller sample sizes and smaller significant levels when the new test often has more than double the power of the best existing test. The test is typically more powerful than the IM_{L} test. Also, the powers of all three tests are generally higher for the two-regressor model than for the one-regressor model. This is likely to be because the two-regressor model has a higher degree of heteroscedasticity.

Table 9 presents the estimated powers of the IM_{L}, IM_{DH} and the new test when model (19) is the true model. With the exception of a couple of cases when $\alpha = 0.01$ and the sample size is small, the new test is always the most powerful of the three tests with IM_{DH} test typically being the second most powerful.

Overall, the results of our simulation study suggest that at least in the context of IM testing, the new test has excellent finite sample size and power properties. There is no evidence to suggest the sizes are different from their nominal sizes and in some cases the powers can be more than double those of the existing tests. The requirement in the two-regressor model case of needing to estimate a nine-dimensional joint density function does not seem to hinder the performance of the proposed test.

7 Test for misspecification of cross-market prediction models

Stock market analysts have found that the Australian stock market often follows the overnight US stock market, and sometimes major European stock markets as well. Therefore, it is possible to forecast the Australian stock market daily return based on the overnight stock market returns on the US stock index and a major European stock index. Such a cross-market predictive relationship is one type of cross-market relationship, which is a broad area that has been extensively studied in the finance literature (see, for example, Forbes and Rigobon, 2002; Longstaff, 2010; Rapach, Strauss, and Zhou, 2013; Bekaert, Ehrmann, Fratzscher, and Mehl, 2014). In order to model such a predictive relationship, one possible model is

$$y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 z_{t-1} + \varepsilon_t, \tag{20}$$

for $t = 1, 2, \cdots, T$, where y_t is the daily return on the Australian S&P/ASX 200 index, and x_{t-1} and z_{t-1} are the lagged US S&P 500 and German DAX daily returns standardized by the corresponding daily VIX measures, respectively.[5]

In order to capture the stylized facts of y_t, one reasonable specification is to assume that ε_t, for $t = 1, 2, \cdots, T$, are conditional heteroscedastic. Another way to capture the stylized facts of y_t is to standardize y_t by its daily VIX and assume that ε_t, for $t = 1, 2, \cdots, T$, are iid $N(0, \sigma^2)$. The latter specification might be inappropriate when there exists a remaining GARCH effect even after the standardization of y_t. Nonetheless, the latter specification can be tested through each of the three versions of the IM test discussed in the previous section.

We collected a sample of these three stock daily returns from the 5th January 2010 to the 31st July 2014 with the sample size being 1110. All three daily return series were standardized by their respective daily VIX measures and the errors, $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_T$, were assumed to be iid $N(0, \sigma^2)$. The model was estimated using OLS, and the parameter vector $(\beta_0, \beta_1, \beta_2, \sigma)$ estimate

[5] If the regressors are not standardized, the cross-market linear relationship is very likely to be twisted by high volatilities, which usually occur when stock returns are extremely negative or positive values.

was $(0.0054, 0.7879, 0.1295, 0.6331)$. Applying the Lagrange multiplier test for GARCH effects to the residuals[6], we find that the p-values are 0.0000 for lags up to five, 0.0002 for lags up to ten, and 0.0058 for lags up to twenty. Thus, the Lagrange multiplier test rejects the null hypothesis of no GARCH effect in the residual series of (20) at the 1% significance level.

According to this finding, the model given by (20) with iid $N(0,\sigma^2)$ errors is likely to be misspecified. The IM test based on our testing procedure, as well as the other two versions of the IM test, were carried out to test for such misspecification. Our testing procedure produces a p-value of 0.0094, while the p-values of $\mathrm{IM_{DH}}$ and $\mathrm{IM_L}$ tests are 0.0073 and 0.3573, respectively. Thus, our test and the $\mathrm{IM_{DH}}$ test both reject this specification at the 1% significance level. However, the Lancaster version of the IM test does not reject this specification.

Investors sometimes may be interested in how positive daily returns in the Australian stock market are affected by the overnight daily returns in the US market and a major European market. Such a cross-market relationship might be explained by a Tobit model:

$$y_t = \begin{cases} \beta_0 + \beta_1 x_{t-1} + \beta_2 z_{t-1} + u_t & \text{if } \beta_0 + \beta_1 x_{t-1} + \beta_2 z_{t-1} + u_t > 0 \\ 0 & \text{if } \beta_0 + \beta_1 x_{t-1} + \beta_2 z_{t-1} + u_t \le 0 \end{cases}, \tag{21}$$

where u_t, for $t = 1, 2, \cdots, T$, are iid $N(0, 1/h^2)$. Note that there is no evidence supporting the Tobit specification of this model. Using the maximum likelihood estimation method, we fitted the model to a shorter sample of daily returns from the 3rd January 2013 to the 31st July 2014, a sample of 382 observations. The parameter vector $(\beta_0, \beta_1, \beta_2, h)$ was estimated as $(0.0075, 0.8752, 0.0879, 1.4937)$.

To test whether the Tobit specification is misspecified, we carried out the IM test based on our testing procedure together with the other two versions of the IM test. The p-value of our test is 0.0001, and the null hypothesis of no misspecification is rejected. However, the p-values of the $\mathrm{IM_{DH}}$ and $\mathrm{IM_L}$ tests are respectively, 0.1312 and 0.3989, which do not reject the null hypothesis at significance levels of 10% and less.

[6]See, for example, Engle (1982) and Lee and King (1993) for discussion of the Lagrange multiplier test.

8 Conclusion

This paper presents a new procedure for hypothesis testing based on a vector of statistics. It involves simulating the statistics under the null hypothesis and then estimating their joint density using a multivariate kernel density estimator. This allows the p-value of the vector of statistics to be estimated giving rise to an ASAR test. Simulation is straight forward when the null distribution of the vector of statistics is independent of nuisance parameters, and it is possible to conduct the test in such a way that it is an exact size test. In the case where the null distribution of the test vector is dependent on nuisance parameters, these parameters are first estimated assuming the null hypothesis is true and then these parameter estimates are used to simulate the vector of statistics under the null hypothesis. We prove that the resultant p-value is a consistent estimate of the true p-value under some regularity conditions. The small-sample properties of the proposed procedure were investigated via simulation in the context of testing for autocorrelation, testing for normality, and testing for model misspecification through the information matrix. We find that our testing procedure has appropriate sizes and that powers that are typically better than, or in some cases as good as those of existing tests.

It appears that for relatively simple testing problems with few nuisance parameters, such as testing for autocorrelation and non-normality in a random sample, the new procedure typically has a slight advantage in terms of power. We see clear evidence of that advantage increasing as we turn to the more difficult problem of testing for misspecification via the information matrix in the linear regression model and the Tobit model. The standard approach in these more complicated settings is to derive the asymptotic distribution of the vector of statistics, estimate the asymptotic covariance matrix and calculate the usual quadratic form that has an asymptotic χ^2 distribution under the null hypothesis and assuming regularity conditions hold. Each of the steps of test construction involves an approximation that has the potential to affect the power of the resultant test. Our approach focuses directly on estimating the small-sample null distribution of the vector of statistics in order to estimate the overall p-value. When that distribution is known, our test is optimal in the sense that by its construction, it has the smallest acceptance region of tests of that size in the d-dimensional sample space of $\boldsymbol{t}$.

An important step in our procedure is the selection of bandwidth values for kernel density estimation. We found that the MCMC based approach is slightly better than the NRR particularly for large values of d, although there is a very big difference in the computational time required. So if this is an issue then the use of the NRR can provide very acceptable results.

We applied three versions of the information test for misspecification to some cross-market prediction models of daily returns on the Australian stock market. We find the three tests can have quite different estimated p-values which can result in different test outcomes. The results of tests for GARCH effects in the model's residuals tend to support the reliability of our proposed test procedure over the other two tests.

There are two possible weaknesses of the procedure that deserve comment. The first is the curse of dimensionality when estimating the joint density function. It is well documented (see Epanechnikov, 1969; Scott, 2015) that as the dimension, d, of the joint density increases, the difficulty of estimating the density increases. To achieve the same level of accuracy larger samples are needed as d increases. Fortunately we have control over the number of simulations of $\boldsymbol{t}$ used to estimate the density. For larger values of d, we therefore recommend using as many simulations as possible. Also, as might be expected, the level of accuracy does depend on the smoothness/regularity of the true density. This suggests that the procedure might work better for larger sample sizes of the observed data, T, when d is large. We don't yet have an answer to the obvious question of how large can d be. Our simulations for the IM test showed that the procedure worked extremely well for $d = 9$ using only 1000 simulations of $\boldsymbol{t}$. We suspect it will work well for bigger values of d, particularly if a much higher number of simulations is used. Further research is needed to determine guidelines for working with large d values.

The second issue concerns the distributional assumption made about the disturbances in the model being simulated. For example, with respect to testing for serial correlation, we assumed that the null hypothesis was given by (6) with ε_t being iid $N(0,\sigma^2)$. This normality assumption[7] has been questioned by many in literature (see, for example, Godfrey, 2009). An obvious question is how might our procedure be applied when we wish to make a much less specific assumption about ε_t. Our suggestion is to assume the density of ε_t is a member of a family of distributions such as the Pearson family (see Bera and Jarque, 1982, for an application) or the Student's t family if one is willing to assume a symmetric heavy-tailed distribution for ε_t (see Zhang and King, 2008). Assuming membership of a family of distributions does then require estimation of unknown parameters (such as c_0, c_1 and c_2 for the Pearson family and the degrees of freedom, τ, for the Student's t family) to determine which member of the family we need to simulate from. Again, further research is needed to determine how well this approach works in practice.

Acknowledgements

We thank Graham Elliott, Jiti Gao, Rob Hyndman, Oliver Linton, Zen Lu, Esfandiar Maasoumi, Hashem Pesaran, Aman Ullah and Halbert White for their useful comments and suggestions. We are particular grateful for some very constructive suggestions from two reviewers which helped improve the paper. This research was supported under Australian Research Council's *Discovery Projects* funding scheme. An early version of this paper was presented at the Econometric Society World Congress held in Shanghai in August 2010 and was entitled "A new procedure for multiple testing of econometric models".

Appendix: Proof of Theorem 1

For multivariate kernel density estimation, Li and Racine (2007) showed that under the smoothness conditions on the true density given in Masry (1996), the kernel density estimator is uniformly consistent on a bounded set, in which the true density is greater than zero. This implies

[7] We could have made the more general assumption that the vector $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$ follows a spherically symmetric distribution with a joint density of the form $\phi(\varepsilon'\varepsilon)$ which includes $N(0,\sigma^2 I_n)$ with I_n denoting the $n \times n$ identity matrix as a special case. As King (1981) notes, all statistics that are invariant to the scale of ε have the same distribution for all spherically symmetric distributions including normality.

that for any γ value of interest,

$$\sup|\widehat{f}_m(\boldsymbol{t},\gamma)-f(\boldsymbol{t},\gamma)|\longrightarrow 0, \text{almost surely}, \tag{22}$$

as $m\longrightarrow\infty$.

Let

$$I_1=\left\{\boldsymbol{t}: f(\boldsymbol{t},\gamma_0)<f(\widehat{\boldsymbol{t}},\gamma_0)\right\}, \quad \text{and} \quad I_2=\left\{\boldsymbol{t}: \widehat{f}_m(\boldsymbol{t},\widehat{\gamma}_T)<\widehat{f}_m(\widehat{\boldsymbol{t}},\widehat{\gamma}_T)\right\}.$$

Then

$$p_0=\int_{I_1} f(\boldsymbol{t},\gamma_0)d\boldsymbol{t}, \quad \text{and} \quad \widehat{p}_{T,m,n}=\sum_{\boldsymbol{t}^{(i)}\in I_2} 1/n,$$

where $\boldsymbol{t}^{(i)}$, $i=1,2,\cdots,n$, are n simulated values of $\boldsymbol{t}$ from $f(\boldsymbol{t},\widehat{\gamma}_T)$. Note that as $n\longrightarrow\infty$,

$$\widehat{p}_{T,m,n}\to\widehat{p}_{T,m},$$

where

$$\widehat{p}_{T,m}=\int_{I_2} f(\boldsymbol{t},\widehat{\gamma}_T)dt.$$

Hence we need to show that $\widehat{p}_{T,m}\to p_0$ as $T\to\infty$ and $m\to\infty$.

Note that I_2 can be rewritten as

$$I_2=\left\{\boldsymbol{t}: \widehat{f}_m(\widehat{\boldsymbol{t}},\widehat{\gamma}_T)-f(\widehat{\boldsymbol{t}},\widehat{\gamma}_T)+f(\widehat{\boldsymbol{t}},\widehat{\gamma}_T)-f(\boldsymbol{t},\widehat{\gamma}_T)+f(\boldsymbol{t},\widehat{\gamma}_T)-\widehat{f}_m(\boldsymbol{t},\widehat{\gamma}_T)>0\right\}.$$

As $m\to\infty$,

$$I_2\to I_3=\left\{\boldsymbol{t}: f(\boldsymbol{t},\widehat{\gamma}_T)<f(\widehat{\boldsymbol{t}},\widehat{\gamma}_T)\right\}, \text{almost surely},$$

because from (22),

$$\widehat{f}_m(\widehat{\boldsymbol{t}},\widehat{\gamma}_T)-f(\widehat{\boldsymbol{t}},\widehat{\gamma}_T)\to 0, \text{almost surely},$$

and

$$f(\boldsymbol{t},\widehat{\gamma}_T)-\widehat{f}_m(\boldsymbol{t},\widehat{\gamma}_T)\to 0, \text{almost surely}.$$

Therefore, we now need to show that $\widehat{p}_T\to p_0$ as $T\to\infty$, where

$$\widehat{p}_T=\int_{I_3} f(\boldsymbol{t},\widehat{\gamma}_T)dt.$$

We have

$$p_0 - \widehat{p}_T = \int_{I_1 \cap I_3} \big(f(\boldsymbol{t}, \gamma_0) - f(\boldsymbol{t}, \widehat{\gamma}_T)\big) d\boldsymbol{t} + \int_{I_1 \cap \overline{I}_3} f(\boldsymbol{t}, \gamma_0) d\boldsymbol{t} - \int_{\overline{I}_1 \cap I_3} f(\boldsymbol{t}, \widehat{\gamma}_T) d\boldsymbol{t} \qquad (23)$$
$$\stackrel{\Delta}{=} A_{1,T} + A_{2,T} - A_{3,T}.$$

Let $I_s = \bigcup_T (I_1 \cap I_3)$ where $\bigcup_T (A)$ denotes the union for all values of T of the set A. It follows that

$$\begin{aligned} \left|A_{1,T}\right| &= \left|\int_{I_1 \cap I_3} (f(\boldsymbol{t}, \gamma_0) - f(\boldsymbol{t}, \widehat{\gamma}_T)) d\boldsymbol{t}\right| \\ &\le \int_{I_1 \cap I_3} \left|f(\boldsymbol{t}, \gamma_0) - f(\boldsymbol{t}, \widehat{\gamma}_T)\right| d\boldsymbol{t} \\ &\le \int_{I_s} \left|f(\boldsymbol{t}, \gamma_0) - f(\boldsymbol{t}, \widehat{\gamma}_T)\right| d\boldsymbol{t} \\ &\to 0, \text{ almost surely,} \end{aligned}$$

as $T \to \infty$ because of strong consistency of $\widehat{\gamma}_T$ and because $f(\boldsymbol{t}, \gamma)$ is continuous in γ.

Note that $\boldsymbol{t} \in I_1 \cap \overline{I}_3$ implies

$$f(\boldsymbol{t}, \gamma_0) < f(\widehat{\boldsymbol{t}}, \gamma_0), \text{ and } f(\boldsymbol{t}, \widehat{\gamma}_T) \ge f(\widehat{\boldsymbol{t}}, \widehat{\gamma}_T).$$

Given that $\widehat{\gamma}_T$ is strongly consistent, $\widehat{\gamma}_T$ almost surely converges to γ_0. It suggests that $I_1 \cap \overline{I}_3$ has measure zero as $T \to \infty$. Therefore, $A_{2,T} \to 0$ as $T \to \infty$.

Similarly, we can show that $\overline{I}_1 \cap I_3$ has measure zero as $T \to \infty$, which implies $A_{3,T} \to 0$ as $T \to \infty$. This completes the proof.

References

Baltagi, B. H., Bresson, G., Pirotte, A., 2006. Joint LM test for homoskedasticity in a one-way error component model. Journal of Econometrics 134 (2), 401–417.

Baltagi, B. H., Jung, B. C., Song, S. H., 2010. Testing for heteroskedasticity and serial correlation in a random effects panel data model. Journal of Econometrics 154 (2), 122–124.

Baltagi, B. H., Li, Q., 1995. Testing AR(1) against MA(1) disturbances in an error component model. Journal of Econometrics 68 (1), 133–151.

Barnard, G. A., 1963. Comment on ‘The spectral analysis of point processes’ by M. S. Bartlett. Journal of the Royal Statistical Society, Series B 25, 294.

Beaulieu, M.-C., Dufour, J.-M., Khalaf, L., 2007. Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: An exact simulation-based approach. Journal of Business & Economic Statistics 25 (4), 398–410.

Beaulieu, M.-C., Dufour, J.-M., Khalaf, L., 2009. Finite sample multivariate tests of asset pricing models with coskewness. Computational Statistics & Data Analysis 53 (6), 2008–2021.

Bekaert, G., Ehrmann, M., Fratzscher, M., Mehl, A. J., 2014. Global crises and equity market contagion. The Journal of Finance 69, 2597–2649.

Bera, A. K., Jarque, C. M., 1982. Model specification tests: A simultaneous approach. Journal of Econometrics 20 (1), 59–82.

Bernard, J.-T., Idoudi, N., Khalaf, L., Yélou, C., 2007. Finite sample multivariate structural change tests with application to energy demand models. Journal of Econometrics 141 (2), 1219–1244.

Bowman, A. W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis. Oxford University Press, London.

Bowman, K. O., Shenton, L. R., 1975. Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. Biometrika 62 (2), 243–250.

Box, G. E. P., Pierce, D. A., 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association 65 (332), 1509–1526.

Breusch, T. S., Pagan, A. R., 1980. The Lagrange multiplier test and its applications to model specification in econometrics. The Review of Economic Studies 47 (1), 239–253.

Breusch, T. S., Schmidt, P., 1988. Alternative forms of the Wald test: How long is a piece of string? Communications in Statistics — Theory and Methods 17 (8), 2789–2795.

Cadre, B., 2006. Kernel estimation of density level sets. Journal of Multivariate Analysis 97 (4), 999–1023.

Cadre, B., Pelletier, B., Pudlo, P., 2013. Estimation of density level sets with a given probability content. Journal of Nonparametric Statistics 25 (1), 261–272.

Chesher, A., 1983. The information matrix test: Simplified calculation via a score test interpretation. Economics Letters 13 (1), 45–48.

Chesher, A., 1984. Testing for neglected heterogeneity. Econometrica 52 (4), 865–872.

Chesher, A., Spady, R., 1991. Asymptotic expansions of the information matrix test statistic. Econometrica 59 (3), 787–815.

D'Agostino, R. B., 1971. An omnibus test of normality for moderate and large size samples. Biometrika 58 (2), 341–348.

D'Agostino, R. B., 1972. Small sample probability points for the D test of normality. Biometrika 59 (1), 219–221.

D'Agostino, R. B., Stephens, M. A., 1986. Goodness-of-Fit Techniques. Marcel Dekker, New York.

Davidson, R., MacKinnon, J. G., 1992. A new form of the information matrix test. Econometrica 60 (1), 145–157.

Davidson, R., MacKinnon, J. G., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York.

Dhaene, G., Hoorelbeke, D., 2004. The information matrix test with bootstrap-based covariance matrix estimation. Economics Letters 82 (3), 341–347.

Dufour, J.-M., 2006. Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. Journal of Econometrics 133 (2), 443–477.

Dufour, J.-M., Farhat, A., Gardiol, L., Khalaf, L., 1998. Simulation-based finite sample normality tests in linear regressions. The Econometrics Journal 1 (1), 154–173.

Dufour, J.-M., Farhat, A., Khalaf, L., 2004. Tests multiples simulés et tests de normalité basés sur plusieurs moments dans les modèles de régression. L'Actualité économique 80 (2-3), 501–522.

Dufour, J.-M., Khalaf, L., 2001. Monte Carlo test methods in econometrics. In: Baltagi, B. (Ed.), A Companion to Theoretical Econometrics. Blackwell, Oxford, pp. 494–519.

Dufour, J.-M., Khalaf, L., 2002. Exact tests for contemporaneous correlation of disturbances in seemingly unrelated regressions. Journal of Econometrics 106 (1), 143–170.

Dufour, J.-M., Khalaf, L., Beaulieu, M.-C., 2003. Exact skewness-kurtosis tests for multivariate normality and goodness-of-fit in multivariate regressions with application to asset pricing models. Oxford Bulletin of Economics and Statistics 65 (s1), 891–906.

Dufour, J.-M., Khalaf, L., Beaulieu, M.-C., 2010. Multivariate residual-based finite-sample tests for serial dependence and ARCH effects with applications to asset pricing models. Journal of Applied Econometrics 25 (2), 263–285.

Dufour, J.-M., Khalaf, L., Bernard, J.-T., Genest, I., 2004. Simulation-based finite-sample tests for heteroskedasticity and ARCH effects. Journal of Econometrics 122 (2), 317–347.

Dufour, J.-M., Khalaf, L., Voia, M., 2015. Finite-sample resampling-based combined hypothesis tests, with applications to serial correlation and predictability. Communications in Statistics-Simulation and Computation 44 (9), 2329–2347.

Dwass, M., 1957. Modified randomization tests for nonparametric hypotheses. The Annals of Mathematical Statistics 28 (1), 181–187.

Engle, R. F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50 (4), 987–1007.

Engle, R. F., 1984. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In: Griliches, Z., Intrilligtor, M. D. (Eds.), Handbook of Econometrics. Vol. 2. North-Holland, Amsterdam, pp. 775–826.

Epanechnikov, V. A., 1969. Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications 14 (1), 153–158.

Forbes, K. J., Rigobon, R., 2002. No contagion, only interdependence: Measuring stock market comovements. The Journal of Finance 57 (5), 2223–2261.

Godfrey, L. G., 1991. Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches. No. 16. Cambridge University Press, Cambridge.

Godfrey, L. G., 2005. Controlling the overall significance level of a battery of least squares diagnostic tests. Oxford Bulletin of Economics & Statistics 67 (2), 263–279.

Godfrey, L. G., 2009. Bootstrap Tests for Regression Models. Palgrave Macmillan, Basingstoke.

Gregory, A. W., Veall, M. R., 1985. Formulating Wald tests of nonlinear restrictions. Econometrica, 1465–1468.

Gregory, A. W., Veall, M. R., 1986. Wald tests of common factor restrictions. Economics Letters 22 (2), 203–208.

Gregory, A. W., Veall, M. R., 1987. Formulating Wald tests of the restrictions implied by the rational expectations hypothesis. Journal of Applied Econometrics 2 (1), 61–68.

Hall, A., 1987. The information matrix test for the linear model. The Review of Economic Studies 54 (2), 257–263.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6 (2), 65–70.

Horowitz, J. L., 1994. Bootstrap-based critical values for the information matrix test. Journal of Econometrics 61 (2), 395–411.

Hyndman, R., 1996. Computing and graphing highest density regions. The American Statistician 50 (2), 120–126.

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

Jarque, C. M., Bera, A. K., 1987. A test for normality of observations and regression residuals. International Statistical Review 55 (2), 163–172.

King, M. L., 1981. The alternative Durbin-Watson test: An assessment of Durbin and Watson's choice of test statistic. Journal of Econometrics 17 (1), 51–66.

King, M. L., 1987. Testing for autocorrelation in linear regression models: A survey. In: King, M. L., Giles, D. E. A. (Eds.), Specification Analysis in the Linear Model. Routledge and Kegan Paul, London, pp. 19–73.

King, M. L., Akram, M., Zhang, X., 2009. A new procedure for multiple one-sided hypothesis testing. In: Econometric Society Australasian Meeting. Canberra, July 2009.

Lafontaine, F., White, K. J., 1986. Obtaining any Wald statistic you want. Economics Letters 21 (1), 35–40.

Lancaster, T., 1984. The covariance matrix of the information matrix test. Econometrica 52 (4), 1051–1053.

Lee, J. H. H., King, M. L., 1993. A locally most mean powerful based score test for ARCH and GARCH regression disturbances. Journal of Business & Economic Statistics 11 (1), 17–27.

Li, Q., Racine, J. S., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press, New Jersey.

Ljung, G. M., Box, G. E. P., 1978. On a measure of lack of fit in time series models. Biometrika 65 (2), 297–303.

Longstaff, F. A., 2010. The subprime credit crisis and contagion in financial markets. Journal of Financial Economics 97 (3), 436–450.

Mantel, N., 1987. Understanding Wald's test for exponential families. The American Statistician 41 (2), 147–148.

Mason, D. M., Polonik, W., 2009. Asymptotic normality of plug-in level set estimates. The Annals of Applied Probability 19 (3), 1108–1142.

Masry, E., 1996. Multivariate regression estimation local polynomial fitting for time series. Stochastic Processes and Their Applications 65 (1), 81–101.

Orme, C., 1990. The small-sample performance of the information-matrix test. Journal of Econometrics 46 (3), 309–331.

Pagan, A., Vella, F., 1989. Diagnostic tests for models based on individual data: A survey. Journal of Applied Econometrics 4, S29–S59.

Pearson, E. S., D'Agostino, R. B., Bowman, K. O., 1977. Tests for departure from normality: Comparison of powers. Biometrika 64 (2), 231–246.

Poitras, G., 2006. More on the correct use of omnibus tests for normality. Economics Letters 90 (3), 304–309.

Rapach, D. E., Strauss, J. K., Zhou, G., 2013. International stock return predictability: What is the role of the United States? The Journal of Finance 68 (4), 1633–1662.

Samworth, R. J., Wand, M. P., 2010. Asymptotics and optimal bandwidth selection for highest density region estimation. The Annals of Statistics 38 (3), 1767–1792.

Scott, D. W., 2015. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, New York.

Shapiro, S. S., Wilk, M. B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3-4), 591–611.

Simes, R. J., 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73 (3), 751–754.

Spiegelhalter, D. J., 1980. An omnibus test for normality for small samples. Biometrika 67 (2), 493–196.

Tauchen, G., 1985. Diagnostic testing and evaluation of maximum likelihood models. Journal of Econometrics 30 (1-2), 415–443.

Taylor, L. W., 1987. The size bias of White's information matrix test. Economics Letters 24 (1), 63–67.

Thode, H. C., 2002. Testing for Normality. Marcel Dekker, New York.

Tsybakov, A. B., 1997. On nonparametric estimation of density level sets. The Annals of Statistics 25 (3), 948–969.

Urzúa, C. M., 1996. On the correct use of omnibus tests for normality. Economics Letters 53 (3), 247–251.

Wand, M. P., Jones, M. C., 1993. Comparison of smoothing parameterizations in bivariate kernel density estimation. Journal of the American Statistical Association 88 (422), 520–528.

Wand, M. P., Jones, M. C., 1995. Kernel Smoothing. Chapman and Hall, New York.

Westfall, P. H., Young, S. S., 1993. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. John Wiley & Sons, New York.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50 (1), 1–25.

Wooldridge, J. M., 2001. Diagnostic testing. In: Baltagi, B. (Ed.), A Companion to Theoretical Econometrics. Blackwell, Oxford, pp. 180–200.

Zhang, X., King, M. L., 2008. Box-Cox stochastic volatility models with heavy-tails and correlated errors. Journal of Empirical Finance 15 (3), 549–566.

Zhang, X., King, M. L., Hyndman, R. J., 2006. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. Computational Statistics & Data Analysis 50 (11), 3009–3031.

Table 1: The estimated sizes of the new test procedures and the Portmanteau test when $d = 4$ and $d = 6$.

Dimension	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
d=4	50	0.10	0.100	0.099	0.101	0.101	0.098	0.100	0.100
		0.05	0.051	0.051	0.048	0.048	0.050	0.051	0.052
		0.01	0.010	0.010	0.010	0.010	0.009	0.010	0.010
	100	0.10	0.098	0.101	0.101	0.101	0.103	0.101	0.102
		0.05	0.048	0.049	0.047	0.049	0.051	0.051	0.051
		0.01	0.009	0.009	0.009	0.009	0.010	0.010	0.011
	200	0.10	0.101	0.102	0.103	0.105	0.103	0.104	0.102
		0.05	0.050	0.050	0.050	0.051	0.053	0.051	0.052
		0.01	0.010	0.011	0.011	0.011	0.012	0.010	0.010
	500	0.10	0.097	0.095	0.093	0.096	0.097	0.095	0.095
		0.05	0.045	0.046	0.045	0.045	0.046	0.045	0.045
		0.01	0.009	0.009	0.009	0.009	0.010	0.009	0.009
d=6	50	0.10	0.104	0.103	0.102	0.102	0.100	0.104	0.104
		0.05	0.054	0.054	0.053	0.052	0.052	0.052	0.053
		0.01	0.010	0.010	0.010	0.010	0.010	0.011	0.010
	100	0.10	0.099	0.100	0.100	0.101	0.102	0.103	0.103
		0.05	0.049	0.050	0.050	0.050	0.051	0.049	0.050
		0.01	0.010	0.010	0.010	0.011	0.011	0.011	0.011
	200	0.10	0.101	0.100	0.102	0.102	0.102	0.102	0.103
		0.05	0.052	0.052	0.053	0.053	0.053	0.052	0.051
		0.01	0.012	0.011	0.010	0.011	0.012	0.009	0.010
	500	0.10	0.093	0.092	0.092	0.090	0.091	0.095	0.094
		0.05	0.044	0.045	0.045	0.046	0.046	0.044	0.046
		0.01	0.008	0.008	0.009	0.009	0.011	0.008	0.009

Table 2: Estimated powers of the new tests and three competing tests when $d = 4$.

DGP	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
AR(1): ρ_1=0.25	50	0.10	0.411	0.417	0.418	0.423	0.387	0.365	0.370
		0.05	0.294	0.299	0.298	0.309	0.278	0.259	0.263
		0.01	0.128	0.134	0.124	0.131	0.115	0.106	0.103
	100	0.10	0.645	0.651	0.656	0.659	0.658	0.610	0.627
		0.05	0.520	0.529	0.519	0.531	0.549	0.482	0.504
		0.01	0.276	0.288	0.268	0.282	0.326	0.259	0.277
	200	0.10	0.898	0.900	0.903	0.906	0.918	0.881	0.897
		0.05	0.835	0.838	0.838	0.847	0.868	0.805	0.833
		0.01	0.633	0.653	0.616	0.646	0.728	0.586	0.633
	500	0.10	0.999	0.999	0.999	0.999	1.000	0.999	0.999
		0.05	0.997	0.998	0.997	0.989	0.999	0.996	0.998
		0.01	0.985	0.988	0.988	0.989	0.994	0.983	0.989
AR(2): ρ_1=0.05	50	0.10	0.208	0.210	0.212	0.214	0.154	0.172	0.170
ρ_2=0.10		0.05	0.125	0.126	0.125	0.126	0.088	0.105	0.104
		0.01	0.041	0.042	0.036	0.038	0.022	0.031	0.030
	100	0.10	0.256	0.261	0.263	0.266	0.214	0.235	0.234
		0.05	0.163	0.166	0.167	0.169	0.132	0.149	0.148
		0.01	0.055	0.057	0.056	0.058	0.037	0.053	0.052
	200	0.10	0.363	0.369	0.382	0.388	0.324	0.348	0.352
		0.05	0.249	0.255	0.261	0.265	0.221	0.243	0.246
		0.01	0.097	0.099	0.099	0.101	0.084	0.093	0.097
	500	0.10	0.638	0.645	0.647	0.651	0.616	0.628	0.639
		0.05	0.521	0.531	0.523	0.531	0.489	0.507	0.519
		0.01	0.284	0.298	0.275	0.286	0.273	0.293	0.302
AR(2): ρ_1=0.05	50	0.10	0.370	0.374	0.381	0.382	0.294	0.315	0.313
ρ_2=0.20		0.05	0.261	0.263	0.266	0.273	0.199	0.221	0.222
		0.01	0.117	0.119	0.107	0.113	0.070	0.089	0.087
	100	0.10	0.542	0.551	0.553	0.557	0.490	0.499	0.507
		0.05	0.419	0.424	0.418	0.429	0.373	0.378	0.387
		0.01	0.214	0.222	0.196	0.210	0.176	0.189	0.197
	200	0.10	0.789	0.797	0.800	0.804	0.777	0.758	0.774
		0.05	0.692	0.701	0.698	0.704	0.683	0.650	0.672
		0.01	0.445	0.464	0.458	0.469	0.466	0.413	0.443
	500	0.10	0.990	0.990	0.990	0.990	0.991	0.985	0.988
		0.05	0.978	0.979	0.978	0.979	0.981	0.969	0.976
		0.01	0.911	0.923	0.923	0.929	0.938	0.900	0.919

Table 2: Estimated powers of the new tests and three competing tests when $d = 4$ (continued).

DGP	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
AR(3): ρ_1=0.10	50	0.10	0.347	0.352	0.356	0.360	0.253	0.298	0.292
ρ_2=0.10		0.05	0.241	0.249	0.251	0.254	0.164	0.217	0.212
ρ_3=0.10		0.01	0.110	0.121	0.106	0.113	0.057	0.100	0.094
	100	0.10	0.488	0.484	0.494	0.496	0.405	0.470	0.467
		0.05	0.371	0.374	0.372	0.380	0.291	0.369	0.365
		0.01	0.192	0.184	0.190	0.202	0.121	0.207	0.203
	200	0.10	0.707	0.717	0.708	0.720	0.635	0.714	0.712
		0.05	0.599	0.604	0.599	0.611	0.513	0.620	0.614
		0.01	0.367	0.361	0.377	0.394	0.287	0.425	0.421
	500	0.10	0.964	0.963	0.966	0.967	0.934	0.968	0.967
		0.05	0.933	0.931	0.934	0.939	0.881	0.946	0.943
		0.01	0.807	0.830	0.803	0.828	0.712	0.870	0.865
AR(3): ρ_1=0.05	50	0.10	0.379	0.380	0.376	0.380	0.289	0.309	0.312
ρ_2=0.05		0.05	0.271	0.271	0.269	0.270	0.190	0.214	0.219
ρ_3=0.20		0.01	0.122	0.121	0.117	0.120	0.064	0.086	0.083
	100	0.10	0.566	0.570	0.559	0.562	0.508	0.506	0.519
		0.05	0.449	0.451	0.438	0.442	0.385	0.388	0.401
		0.01	0.234	0.242	0.218	0.223	0.181	0.195	0.205
	200	0.10	0.816	0.822	0.811	0.816	0.793	0.780	0.797
		0.05	0.720	0.727	0.706	0.719	0.702	0.683	0.703
		0.01	0.480	0.500	0.490	0.501	0.488	0.449	0.481
	500	0.10	0.992	0.993	0.992	0.993	0.993	0.989	0.992
		0.05	0.983	0.984	0.982	0.983	0.985	0.978	0.982
		0.01	0.940	0.945	0.929	0.942	0.950	0.932	0.943
AR(4): ρ_1=0.05	50	0.10	0.533	0.533	0.533	0.533	0.390	0.454	0.452
ρ_2=0.10		0.05	0.422	0.424	0.424	0.429	0.283	0.367	0.367
ρ_3=0.15		0.01	0.255	0.258	0.247	0.252	0.122	0.217	0.208
ρ_3=0.15	100	0.10	0.745	0.748	0.734	0.737	0.625	0.699	0.698
		0.05	0.649	0.656	0.648	0.650	0.510	0.610	0.607
		0.01	0.444	0.457	0.463	0.474	0.280	0.436	0.434
	200	0.10	0.934	0.937	0.935	0.936	0.881	0.922	0.922
		0.05	0.890	0.895	0.890	0.894	0.806	0.880	0.878
		0.01	0.745	0.765	0.761	0.769	0.605	0.756	0.757
	500	0.10	1.000	1.000	1.000	1.000	0.998	1.000	0.999
		0.05	0.998	0.999	0.999	0.999	0.995	0.998	0.998
		0.01	0.990	0.993	0.993	0.994	0.973	0.994	0.994

Table 2: Estimated powers of the new tests and three competing tests when $d = 4$ (continued).

DGP	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
AR(4): ρ_1=0.05	50	0.10	0.334	0.337	0.339	0.342	0.233	0.278	0.275
ρ_2=0.10		0.05	0.236	0.237	0.238	0.242	0.149	0.198	0.198
ρ_3=0.05		0.01	0.106	0.108	0.104	0.105	0.047	0.087	0.082
ρ_3=0.10	100	0.10	0.464	0.469	0.454	0.459	0.355	0.421	0.421
		0.05	0.343	0.350	0.343	0.349	0.252	0.319	0.317
		0.01	0.164	0.172	0.165	0.167	0.096	0.171	0.169
	200	0.10	0.657	0.664	0.657	0.663	0.564	0.641	0.636
		0.05	0.542	0.551	0.548	0.556	0.438	0.539	0.535
		0.01	0.304	0.324	0.325	0.343	0.222	0.345	0.344
	500	0.10	0.936	0.939	0.936	0.939	0.890	0.935	0.935
		0.05	0.889	0.896	0.890	0.896	0.811	0.895	0.893
		0.01	0.731	0.750	0.747	0.756	0.603	0.783	0.776
AR(4): ρ_1=0.05	50	0.10	0.218	0.219	0.219	0.222	0.149	0.173	0.170
ρ_2=0.05		0.05	0.137	0.138	0.138	0.139	0.085	0.110	0.108
ρ_3=0.05		0.01	0.048	0.048	0.043	0.045	0.021	0.036	0.034
ρ_3=0.05	100	0.10	0.265	0.269	0.265	0.270	0.204	0.235	0.232
		0.05	0.172	0.175	0.168	0.171	0.124	0.155	0.152
		0.01	0.061	0.064	0.060	0.063	0.033	0.062	0.060
	200	0.10	0.356	0.362	0.363	0.368	0.290	0.344	0.339
		0.05	0.249	0.253	0.246	0.252	0.189	0.247	0.243
		0.01	0.100	0.106	0.102	0.106	0.066	0.109	0.106
	500	0.10	0.611	0.613	0.602	0.611	0.513	0.612	0.606
		0.05	0.483	0.496	0.485	0.496	0.377	0.504	0.496
		0.01	0.250	0.264	0.273	0.298	0.170	0.314	0.302

Table 3: Estimated powers of the new tests and three competing tests when $d = 6$.

DGP	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
AR(1): ρ_1=0.25	50	0.10	0.366	0.374	0.358	0.371	0.350	0.326	0.327
		0.05	0.256	0.269	0.249	0.265	0.247	0.225	0.228
		0.01	0.102	0.110	0.099	0.109	0.101	0.085	0.083
	100	0.10	0.566	0.582	0.568	0.591	0.616	0.539	0.560
		0.05	0.437	0.453	0.434	0.455	0.500	0.402	0.430
		0.01	0.217	0.226	0.213	0.221	0.293	0.203	0.218
	200	0.10	0.850	0.864	0.859	0.872	0.895	0.827	0.855
		0.05	0.765	0.790	0.778	0.798	0.841	0.737	0.775
		0.01	0.564	0.587	0.570	0.597	0.691	0.492	0.546
	500	0.10	0.997	0.998	0.997	0.998	0.999	0.996	0.998
		0.05	0.993	0.995	0.992	0.995	0.999	0.992	0.995
		0.01	0.974	0.977	0.968	0.973	0.994	0.961	0.977
AR(2): ρ_1=0.05	50	0.10	0.196	0.197	0.196	0.199	0.143	0.159	0.157
ρ_2=0.10		0.05	0.117	0.121	0.121	0.124	0.082	0.093	0.091
		0.01	0.037	0.037	0.038	0.039	0.019	0.027	0.026
	100	0.10	0.223	0.233	0.237	0.244	0.195	0.210	0.209
		0.05	0.139	0.144	0.137	0.141	0.116	0.124	0.126
		0.01	0.045	0.046	0.039	0.040	0.033	0.042	0.042
	200	0.10	0.319	0.328	0.329	0.340	0.289	0.310	0.314
		0.05	0.214	0.223	0.219	0.227	0.191	0.209	0.211
		0.01	0.080	0.081	0.079	0.085	0.069	0.072	0.076
	500	0.10	0.575	0.588	0.576	0.599	0.558	0.564	0.578
		0.05	0.440	0.458	0.435	0.459	0.441	0.435	0.455
		0.01	0.225	0.234	0.217	0.232	0.240	0.222	0.241
AR(2): ρ_1=0.05	50	0.10	0.336	0.342	0.333	0.341	0.266	0.285	0.284
ρ_2=0.20		0.05	0.233	0.240	0.227	0.238	0.177	0.193	0.193
		0.01	0.091	0.095	0.090	0.095	0.062	0.072	0.069
	100	0.10	0.472	0.490	0.475	0.486	0.447	0.442	0.449
		0.05	0.351	0.365	0.352	0.362	0.333	0.314	0.325
		0.01	0.164	0.168	0.162	0.165	0.153	0.150	0.156
	200	0.10	0.719	0.733	0.720	0.737	0.734	0.689	0.714
		0.05	0.604	0.620	0.606	0.625	0.637	0.575	0.603
		0.01	0.382	0.382	0.381	0.392	0.422	0.339	0.368
	500	0.10	0.978	0.982	0.979	0.983	0.987	0.971	0.979
		0.05	0.952	0.961	0.955	0.964	0.973	0.943	0.960
		0.01	0.859	0.870	0.864	0.876	0.925	0.835	0.874

Table 3: Estimated powers of the new tests and three competing tests when $d = 6$ (continued).

DGP	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
AR(3): ρ_1=0.10	50	0.10	0.320	0.327	0.321	0.333	0.229	0.272	0.267
ρ_2=0.10		0.05	0.222	0.227	0.226	0.234	0.148	0.191	0.187
ρ_3=0.10		0.01	0.097	0.100	0.092	0.096	0.050	0.083	0.077
	100	0.10	0.432	0.446	0.427	0.443	0.367	0.423	0.419
		0.05	0.323	0.335	0.305	0.320	0.257	0.316	0.315
		0.01	0.160	0.167	0.144	0.148	0.106	0.172	0.167
	200	0.10	0.643	0.660	0.641	0.662	0.582	0.660	0.660
		0.05	0.529	0.553	0.523	0.544	0.463	0.562	0.560
		0.01	0.320	0.329	0.321	0.337	0.249	0.363	0.364
	500	0.10	0.935	0.947	0.937	0.948	0.911	0.951	0.951
		0.05	0.886	0.903	0.892	0.906	0.849	0.918	0.920
		0.01	0.749	0.765	0.764	0.778	0.675	0.811	0.816
AR(3): ρ_1=0.05	50	0.10	0.343	0.347	0.353	0.359	0.258	0.282	0.281
ρ_2=0.05		0.05	0.242	0.249	0.249	0.257	0.171	0.192	0.195
ρ_3=0.20		0.01	0.102	0.105	0.106	0.108	0.057	0.075	0.072
	100	0.10	0.504	0.514	0.504	0.515	0.460	0.452	0.465
		0.05	0.388	0.400	0.385	0.393	0.341	0.333	0.347
		0.01	0.199	0.203	0.190	0.195	0.159	0.167	0.172
	200	0.10	0.749	0.765	0.730	0.748	0.751	0.724	0.746
		0.05	0.640	0.664	0.620	0.642	0.656	0.617	0.642
		0.01	0.423	0.433	0.389	0.408	0.445	0.385	0.413
	500	0.10	0.983	0.986	0.983	0.986	0.990	0.981	0.985
		0.05	0.965	0.971	0.963	0.969	0.979	0.962	0.972
		0.01	0.901	0.913	0.887	0.898	0.936	0.883	0.911
AR(4): ρ_1=0.05	50	0.10	0.492	0.498	0.494	0.504	0.356	0.422	0.419
ρ_2=0.10		0.05	0.390	0.398	0.387	0.401	0.254	0.333	0.331
ρ_3=0.15		0.01	0.223	0.227	0.217	0.219	0.110	0.194	0.185
ρ_3=0.15	100	0.10	0.685	0.700	0.677	0.692	0.582	0.655	0.655
		0.05	0.582	0.600	0.572	0.585	0.463	0.555	0.557
		0.01	0.377	0.389	0.382	0.396	0.250	0.386	0.382
	200	0.10	0.902	0.912	0.899	0.908	0.848	0.895	0.895
		0.05	0.844	0.858	0.840	0.851	0.767	0.845	0.845
		0.01	0.689	0.699	0.697	0.708	0.561	0.705	0.707
	500	0.10	0.998	0.999	0.998	0.999	0.997	0.999	0.999
		0.05	0.995	0.997	0.996	0.997	0.993	0.996	0.997
		0.01	0.982	0.995	0.987	0.988	0.965	0.989	0.990

Table 3: Estimated powers of the new tests and three competing tests when $d = 6$ (continued).

DGP	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
AR(4): ρ_1=0.05	50	0.10	0.314	0.319	0.315	0.323	0.214	0.253	0.252
ρ_2=0.10		0.05	0.219	0.224	0.215	0.223	0.134	0.175	0.173
ρ_3=0.05		0.01	0.091	0.094	0.087	0.091	0.042	0.073	0.069
ρ_3=0.10	100	0.10	0.403	0.418	0.418	0.430	0.323	0.377	0.375
		0.05	0.296	0.307	0.302	0.313	0.220	0.276	0.276
		0.01	0.138	0.142	0.139	0.146	0.082	0.143	0.141
	200	0.10	0.595	0.611	0.597	0.609	0.509	0.588	0.587
		0.05	0.473	0.497	0.488	0.505	0.392	0.487	0.485
		0.01	0.263	0.271	0.286	0.300	0.190	0.293	0.293
	500	0.10	0.899	0.912	0.898	0.908	0.854	0.909	0.909
		0.05	0.834	0.854	0.838	0.851	0.768	0.856	0.858
		0.01	0.666	0.685	0.683	0.694	0.561	0.714	0.718
AR(4): ρ_1=0.05	50	0.10	0.211	0.212	0.207	0.210	0.140	0.159	0.157
ρ_2=0.05		0.05	0.131	0.135	0.126	0.129	0.080	0.098	0.095
ρ_3=0.05		0.01	0.044	0.043	0.039	0.040	0.019	0.031	0.028
ρ_3=0.05	100	0.10	0.234	0.240	0.234	0.242	0.185	0.209	0.209
		0.05	0.152	0.158	0.154	0.158	0.110	0.132	0.132
		0.01	0.051	0.052	0.050	0.051	0.029	0.051	0.050
	200	0.10	0.320	0.329	0.327	0.339	0.257	0.312	0.309
		0.05	0.217	0.230	0.222	0.233	0.164	0.221	0.217
		0.01	0.085	0.088	0.089	0.093	0.055	0.089	0.087
	500	0.10	0.541	0.561	0.550	0.565	0.452	0.559	0.556
		0.05	0.413	0.429	0.434	0.443	0.330	0.442	0.440
		0.01	0.213	0.221	0.240	0.245	0.144	0.254	0.251

Table 4: Estimated sizes and powers of the new test procedures, Dufour's (2006) exact tests and JB test.

Dimension	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
Size	30	0.10	0.099	0.099	0.100	0.100	0.101	0.099	0.099
		0.05	0.051	0.051	0.050	0.050	0.050	0.050	0.051
		0.01	0.010	0.010	0.010	0.010	0.011	0.011	0.011
	50	0.10	0.099	0.098	0.098	0.097	0.099	0.099	0.101
		0.05	0.046	0.046	0.047	0.048	0.048	0.048	0.048
		0.01	0.011	0.011	0.011	0.010	0.011	0.011	0.011
	75	0.10	0.104	0.104	0.103	0.103	0.102	0.103	0.104
		0.05	0.051	0.051	0.051	0.052	0.051	0.051	0.052
		0.01	0.010	0.010	0.009	0.009	0.010	0.010	0.011
	100	0.10	0.103	0.103	0.103	0.102	0.101	0.100	0.100
		0.05	0.053	0.053	0.053	0.053	0.054	0.055	0.054
		0.01	0.011	0.011	0.011	0.011	0.011	0.011	0.011
Student t_5	30	0.10	0.416	0.414	0.417	0.416	0.385	0.397	0.396
		0.05	0.324	0.322	0.320	0.319	0.299	0.310	0.309
		0.01	0.177	0.178	0.179	0.177	0.179	0.179	0.180
	50	0.10	0.542	0.544	0.540	0.539	0.502	0.519	0.519
		0.05	0.442	0.445	0.444	0.445	0.421	0.433	0.432
		0.01	0.289	0.290	0.284	0.283	0.287	0.278	0.280
	75	0.10	0.665	0.666	0.664	0.664	0.620	0.637	0.637
		0.05	0.568	0.569	0.569	0.569	0.547	0.558	0.556
		0.01	0.379	0.381	0.383	0.382	0.389	0.374	0.379
	100	0.10	0.743	0.743	0.738	0.739	0.704	0.717	0.718
		0.05	0.659	0.659	0.663	0.663	0.641	0.652	0.652
		0.01	0.480	0.480	0.478	0.480	0.489	0.469	0.475
χ^2_3	30	0.10	0.829	0.834	0.832	0.835	0.848	0.845	0.844
		0.05	0.730	0.740	0.738	0.744	0.745	0.647	0.683
		0.01	0.489	0.503	0.478	0.481	0.478	0.395	0.399
	50	0.10	0.976	0.976	0.977	0.977	0.970	0.977	0.978
		0.05	0.948	0.945	0.947	0.947	0.930	0.880	0.900
		0.01	0.818	0.809	0.804	0.805	0.759	0.635	0.647
	75	0.10	0.999	0.666	0.664	0.664	0.620	0.637	0.637
		0.05	0.997	0.997	0.996	0.996	0.992	0.990	0.991
		0.01	0.973	0.971	0.976	0.976	0.922	0.823	0.841
	100	0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.05	1.000	1.000	1.000	1.000	0.999	1.000	1.000
		0.01	0.998	0.998	0.998	0.998	0.986	0.941	0.956

Table 4: Estimated sizes and powers of the new test procedures, Dufour's (2006) exact tests and JB test (continued).

Dimension	T	α	Single simulation		Double simulation		Minimum	Product	Q
			NRR	MCMC	NRR	MCMC	p	p	
Gamma	30	0.10	0.720	0.723	0.724	0.726	0.764	0.745	0.749
		0.05	0.601	0.608	0.612	0.616	0.639	0.545	0.576
		0.01	0.359	0.370	0.349	0.350	0.366	0.302	0.305
	50	0.10	0.933	0.933	0.935	0.935	0.930	0.936	0.938
		0.05	0.874	0.869	0.876	0.877	0.864	0.786	0.815
		0.01	0.670	0.660	0.654	0.656	0.632	0.514	0.523
	75	0.10	0.993	0.993	0.994	0.994	0.990	0.994	0.994
		0.05	0.981	0.981	0.981	0.981	0.972	0.954	0.962
		0.01	0.905	0.902	0.909	0.910	0.829	0.702	0.720
	100	0.10	1.000	1.000	1.000	1.000	0.999	1.000	1.000
		0.05	0.998	0.998	0.998	0.998	0.995	0.994	0.995
		0.01	0.980	0.979	0.980	0.979	0.949	0.859	0.881
log normal	30	0.10	0.718	0.720	0.722	0.723	0.773	0.746	0.754
		0.05	0.616	0.619	0.625	0.628	0.670	0.598	0.622
		0.01	0.420	0.430	0.407	0.407	0.436	0.385	0.388
	50	0.10	0.920	0.921	0.923	0.923	0.934	0.927	0.932
		0.05	0.869	0.866	0.873	0.874	0.880	0.822	0.843
		0.01	0.704	0.699	0.696	0.697	0.694	0.609	0.614
	75	0.10	0.988	0.988	0.990	0.990	0.989	0.990	0.990
		0.05	0.975	0.974	0.975	0.975	0.975	0.956	0.963
		0.01	0.905	0.903	0.907	0.907	0.869	0.787	0.798
	100	0.10	0.999	0.999	0.998	0.998	0.998	0.999	0.999
		0.05	0.996	0.996	0.996	0.996	0.995	0.993	0.994
		0.01	0.977	0.977	0.976	0.975	0.961	0.909	0.921

Table 5: Estimated sizes of the IM tests with samples generated from the linear regression model.

Test	Sample size	One-regressor			Two-regressor		
		0.01	0.05	0.10	0.01	0.05	0.10
New test	50	0.010	0.051	0.100	0.011	0.048	0.098
	100	0.009	0.046	0.095	0.010	0.050	0.099
	200	0.009	0.051	0.097	0.008	0.049	0.099
	300	0.010	0.050	0.102	0.009	0.048	0.097
IM_L	50	0.011	0.052	0.103	0.010	0.051	0.102
	100	0.011	0.050	0.100	0.011	0.051	0.100
	200	0.008	0.051	0.103	0.011	0.052	0.101
	300	0.011	0.050	0.103	0.009	0.048	0.103
IM_{DH}	50	0.010	0.053	0.102	0.011	0.051	0.102
	100	0.011	0.049	0.097	0.010	0.049	0.098
	200	0.010	0.051	0.101	0.009	0.050	0.101
	300	0.010	0.054	0.107	0.010	0.048	0.096

Table 6: Estimated powers of the IM tests with samples generated from the linear model.

Tests	Sample size	One-regressor			Two-regressor		
		0.01	0.05	0.10	0.01	0.05	0.10
New test	50	0.106	0.363	0.533	0.030	0.121	0.216
	100	0.491	0.801	0.913	0.243	0.572	0.743
	200	0.975	0.998	1.000	0.738	0.944	0.976
	300	0.999	1.000	1.000	0.927	0.991	0.998
IM_L	50	0.017	0.074	0.154	0.014	0.055	0.107
	100	0.034	0.268	0.475	0.022	0.130	0.263
	200	0.422	0.803	0.914	0.202	0.500	0.680
	300	0.751	0.962	0.993	0.396	0.781	0.911
IM_{DH}	50	0.025	0.133	0.281	0.020	0.104	0.219
	100	0.078	0.473	0.713	0.051	0.287	0.500
	200	0.580	0.948	0.986	0.196	0.649	0.832
	300	0.910	0.993	0.999	0.485	0.873	0.952

Table 7: Estimated sizes of the IM tests with samples generated from the Tobit model.

Tests	Sample size	One-regressor 0.01	One-regressor 0.05	One-regressor 0.10	Two-regressor 0.01	Two-regressor 0.05	Two-regressor 0.10
New test	50	0.011	0.055	0.109	0.009	0.049	0.104
	100	0.009	0.048	0.097	0.009	0.055	0.112
	200	0.011	0.057	0.106	0.012	0.046	0.098
	300	0.013	0.047	0.087	0.011	0.054	0.107
IM_L	50	0.018	0.058	0.105	0.012	0.047	0.089
	100	0.016	0.056	0.108	0.015	0.051	0.092
	200	0.016	0.052	0.100	0.013	0.049	0.098
	300	0.014	0.049	0.100	0.017	0.065	0.124
IM_{DH}	50	0.020	0.052	0.102	0.016	0.056	0.100
	100	0.020	0.060	0.100	0.019	0.053	0.107
	200	0.016	0.055	0.112	0.010	0.045	0.116
	300	0.015	0.057	0.111	0.015	0.054	0.107

Table 8: Estimated powers of the IM tests with samples generated from (18).

Test	Sample size	One-regressor 0.01	One-regressor 0.05	One-regressor 0.10	Two-regressor 0.01	Two-regressor 0.05	Two-regressor 0.10
New test	50	0.156	0.362	0.503	0.120	0.330	0.471
	100	0.247	0.501	0.629	0.259	0.598	0.726
	200	0.454	0.782	0.866	0.647	0.926	0.951
	300	0.699	0.917	0.948	0.818	0.992	1.000
IM_L	50	0.054	0.128	0.197	0.052	0.103	0.187
	100	0.062	0.198	0.320	0.120	0.234	0.351
	200	0.196	0.425	0.572	0.311	0.652	0.766
	300	0.396	0.674	0.782	0.706	0.841	0.952
IM_{DH}	50	0.043	0.167	0.307	0.029	0.110	0.223
	100	0.112	0.349	0.536	0.097	0.315	0.506
	200	0.365	0.712	0.858	0.475	0.795	0.930
	300	0.622	0.891	0.943	0.762	0.976	1.000

Table 9: Estimated powers of the IM tests with samples generated from (19).

Test	Sample size	Significance level		
		0.01	0.05	0.10
New test	50	0.025	0.115	0.201
	100	0.041	0.184	0.320
	200	0.222	0.580	0.742
	300	0.476	0.844	0.930
IM_L	50	0.033	0.077	0.144
	100	0.056	0.117	0.182
	200	0.139	0.291	0.395
	300	0.258	0.440	0.578
IM_{DH}	50	0.022	0.074	0.135
	100	0.029	0.120	0.238
	200	0.142	0.422	0.589
	300	0.428	0.752	0.872