Finite-sample resampling-based combined hypothesis tests, with applications to serial correlation and predictability *

Jean-Marie Dufour †
McGill University

Lynda Khalaf ‡
Carleton University

Marcel Voia§
Carleton University

October 2013

*This work was supported by the William Dow Chair in Political Economy (McGill University), the Bank of Canada (Research Fellowship), the Toulouse School of Economics (Pierre-de-Fermat Chair of excellence), the Universitad Carlos III de Madrid (Banco Santander de Madrid Chair of excellence), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Fonds de recherche sur la société et la culture and NATECH (Government of Québec).

† William Dow Professor of Economics, Department of Economics, McGill University, and Centre interuniversitaire de recherche en économie quantitative, Centre interuniversitaire de recherche en analyse des organisations. Mailing address: Leacock Building, Rm 519, 855 Sherbrooke Street West, Montréal, Québec, Canada H3A 2T7. TEL: (514) 398 8879; FAX: (514) 398 4938; e-mail: jean-marie.dufour@mcgill.ca.

‡ Economics Department Carleton University, and Centre for Monetary and Financial Economics [Carleton University], Centre de Recherche de l'Environnement, de l'Agroalimentaire, des Transports et de l'Énergie [Université Laval], Centre interuniversitaire de recherche en économie quantitative. Mailing address: Loeb Building 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada. Tel (613) 520-2600-8697; FAX: (613)-520-3906. email: Lynda_Khalaf@carleton.ca.

§ Economics Department, Carleton University and Centre for Monetary and Financial Economics [Carleton University]. Mailing address: Loeb Building 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada. Tel (613) 520-2600-3546; FAX: (613)-520-3906. email: Marcel_Voia@carleton.ca.

Abstract

This paper suggests Monte Carlo multiple test procedures which are provably valid in finite samples. These include combination methods originally proposed for independent statistics and further improvements which formalize statistical practice. We also adapt the Monte Carlo test method to non-continuous combined statistics. The methods suggested are applied to test serial dependence and predictability. In particular, we introduce and analyze new procedures that account for endogenous lag selection. A simulation study illustrates the properties of the proposed methods. Results show that concrete and non-spurious power gains (over standard combination methods) can be achieved through the combined Monte Carlo test approach, and confirm arguments in favour of variance-ratio type criteria.

Keywords: Monte Carlo test; induced test; test combination; simultaneous inference; Variance ratio.

1 Introduction

Combining multiple non-independent tests is a common problem in statistics and econometrics. Indeed, econometric models often suggest to test several hypotheses or the same hypothesis using several tests, all from the same data. The first problem is typically associated with the construction of simultaneous confidence regions (simultaneous inference), while the second one involves combining different tests which are valid under the same hypothesis, but with power properties that vary depending on the alternative hypothesis. Tests obtained by combining several separate tests are called induced tests (or combined tests). Although different, these two problems raise related difficulties and require similar techniques. For general discussions of these issues, see Miller (1981), Folks (1984), Savin (1984), Dufour (1989), and Dufour and Torrès (1998); for econometric applications, see Dufour and Khalaf (2002), Dufour, Khalaf, and Beaulieu (2003), Dufour, Khalaf, Bernard, and Genest (2004), Bernard, Idoudi, Khalaf, and Yelou (2007), Dufour, Khalaf, and Beaulieu (2010), Bennett (2012b), and Bennett and Thompson (2012).

In the case of induced tests, the problem consists in controlling the global level of the procedure in a situation where the distribution of each test statistic is usually known or relatively easy to compute, but the joint distribution is unknown or difficult to establish. Relying on critical points that control the level of each test individually leads to a global Type I error which can exceed by far the level of the individual tests. For example, if the level of each one of 10 tests is equal to 0.05, the probability that at least one of these tests is significant is typically much higher (up to 0.50). It is therefore imperative to account for the relationship between the different statistics. In the case of simultaneous tests for different hypotheses, the problem is to control the probability of rejecting at least one true hypothesis in a set which may be large (possibly infinite). This problem is a generalization of the former where several distinct hypotheses are examined rather than one, so again we must take into account the joint distribution of the statistics. In addition, it is well known that rejection using a joint procedure can be difficult to interpret as its constituents can react differently to different alternatives. An important further question is thus to determine what aspect of a joint hypothesis results in its rejection, for example, for model specification purposes.

Specification testing is one of the basic problems which motivate multiple tests. For example, autocorrelation and goodness-of-fit tests typically suggest one to consider several *moments* via *portmanteau* methods which combine transformations of asymptotically uncorrelated individual criteria, for example: (i) normality tests which combine skewness and kurtosis [(Jarque and Bera, 1980, JB), Kiefer and Salmon (1983), Dufour, Farhat, Gardiol, and Khalaf (1998), Dufour, Farhat, and Khalaf (2004)]; (ii) portmanteau serial correlation tests [Box and Pierce (1970)] or their variance-ratio counterparts, proposed by Cochrane (1988) and Lo and MacKinlay (1988) to test market efficiency (also called *predictability* tests); see also Lo and MacKinlay (1989), Chou and Denning (1993), Fong, Koh, and Ouliaris (1997), Whang and Kim (2003), Wright (2000), Yilmaz (2003), Kim (2006) and Charles and Darne (2009). Such tests are justified asymptotically, but it is well known that their finite-sample performance can be very unsatisfactory. Tests on multivariate models are another typical example. Dimensionality difficulties arise in this case. For example, we can transform an m-dimensional test of normality or heteroskedasticity into a series of m univariate tests. In this case as well, finite-sample methods are scarce, while the available asymptotic methods often behave poorly in finite samples; see Bewley and Theil (1987), Deschamps (1996), Zhou (1993), Kilian and Demiroglu (2000), Dufour and Khalaf (2002), Dufour, Khalaf, and Beaulieu (2003), Bernard, Idoudi, Khalaf, and Yelou (2007), and Dufour, Khalaf, and Beaulieu (2010).

Overall, the distributional issues raised by using several statistics are difficult. A common method in such contexts consists in using bounds (*e.g.,* bounds based on Boole-Bonferroni inequalities), but these are typically conservative and can have a negative effect on power. Reflecting advances in computer technology, simulation-based resampling techniques have recently been used when the derivation of exact (for a given sample size) or asymptotic distributions is complex. These include the bootstrap [see, *e.g.*, Hall (1992), Efron and Tibshirani (1993), Davison and Hinkley (1997), Horowitz (1997), Chernick (2008)] and the Monte Carlo (MC) test method [Dwass (1957), Barnard (1963), Dufour and Kiviet (1996), Dufour and Khalaf (2001), Dufour (2006)]. For testing hypotheses, these methods typically improve level control in finite samples. In addition, for some non-standard problems [*e.g.*, when certain parameters are not identified under the null hypothesis], simulation-based approaches are not only more reliable, but may be easier to implement than available asymptotic counterparts; see Dufour and Khalaf (2001), Dufour, Khalaf, Bernard, and Genest (2004), and Bernard, Idoudi, Khalaf, and Yelou (2007).

Resampling techniques can be exploited to solve multiple testing problems. Specifically, the bootstrap can improve the quality of asymptotic approximations; see Westfall and Young (1993), White (2000), Dudoit and van der Laan (2008), and Bennett (2012a, 2012b). But asymptotic improvements or refinements may not be sufficient to solve finite-sample size distortions. In this paper, we argue for MC multiple test procedures which are provably valid in finite samples. More specifically, this paper makes two main contributions.

First, we propose a unified framework under which the MC test method solves the combination problem in finite samples and thus avoids reliance on Bonferroni or other bounds, with focus on induced tests. These include parametric possibly non-Gaussian hypotheses and even non-parametric problems. This framework allows us to reinterpret previously proposed procedures and to consider new applications. The latter include: (i) combination methods originally proposed for independent statistics, specifically the procedures suggested by Tippett (1931), Fisher (1932) and Pearson (1933); (ii) further refinements which reflect statistical practice, and a number of alternative combination methods previously not considered. We show *analytically* how the MC test technique solves combination problems for any sample size. An attractive relationship between the Fisher-Pearson and portmanteau tests also emerges. Further, we note that some of the combined statistics which arise naturally in this framework are not continuous. To deal with this issue, we adapt the MC test method conformably using the randomized tie-breaking procedure from Dufour (2006).

Second, using this framework, we revisit some examples of induced tests which routinely appear in time series analysis: serial dependence and predictability tests. We consider autocorrelation Box-Pierce-type and variance-ratio statistics, and we study several exact procedures based on such criteria. To do this, we propose new tests which: (i) formalize the practice of analyzing correlograms, and (ii) allow data-based lag selection. We further show that the MC test technique allows one to use asymptotic p-values in the construction of an exact serial dependence or predictability test, even though these p-values could lead to inaccurate inference when used in the conventional way. Formally, our joint test procedure involves converting all individual tests to an approximate p-value form, in order to combine them. When the overall procedure is simulated, the fact that individual p-values are approximated does not prevent controlling the global test level.

We present a simulation study to assess the usefulness of the proposed procedures. Results can be summarized as follows. (1) Tests based on asymptotic distributions can be either over-sized or under-sized. Their MC counterparts always have the correct size when the form of the underlying distribution is correctly specified. For under-sized tests, this can translate into notable power gains, *e.g.* for tests based on variance-ratio criteria. (2) Whether sup- or Tippett-type, the proposed combined tests perform better than the MC versions of available tests (such as portmanteau tests for serial correlation). (3) There is little difference in the relative powers of different MC tests (power rankings), under normal or fat-tailed distributions. Effective power improvements, due to the size correction, are stronger with t-errors. (4) Variance-ratio tests exhibit better power than Box-Pierce-type tests. This confirms existing arguments in favour of such criteria, once the size control problem is solved by the MC test approach.

The plan of the paper is as follows. In Section 2, we present our unified test framework. Section 3 discusses the serial dependence application. The simulation study is reported in Section 4. We conclude in Section 5.

2 Framework and joint test methods

Consider m statistics S_i, $i = 1, \ldots, m$, which may not be independent, each designed to test the null hypothesis H_{0i} (where some of the hypotheses H_{0i} may be identical). To simplify the exposition (and without loss of generality), we assume the hypothesis H_{0i} is rejected at level α when S_i is large, *i.e.* $S_i \geq c_i$ where c_i is a critical point such that $P[S_i \geq c_i] \leq \alpha$ when H_{0i} is true. Equivalently, the test $S_i \geq c_i$ can be considered significant at level α when $p_i \leq \alpha$ where p_i is the marginal significance level of the test (p-value), *i.e.* $p_i = G_i(S_i)$ where $G_i(x) = \mathsf{P}[S_i \geq x]$ is the survival function of S_i under H_{0i}. We further assume each statistic S_i follows a continuous distribution under H_{0i}. In this case, we easily see that p_i has a uniform distribution on the interval $(0, 1)$ under the null hypothesis:

$$p_i \sim U(0,\, 1) \text{ under } H_{0i}\,. \tag{1}$$

The problem of interest can be formulated as follows: how can we combine these tests to assess the joint hypothesis

$$H_0 : \text{the hypotheses } H_{01}, \ldots,\, H_{0m} \text{ are all true} \tag{2}$$

in a way that controls the probability of rejecting the joint hypothesis H_0?

To do this, we propose to apply the technique of MC tests, which can be summarized as follows. First, we obtain a combined statistic, denoted $\bar{S}$. Again, without loss of generality, we assume the test based on the statistic $\bar{S}$ rejects H_0 when $\bar{S}$ is large. Several combination rules are considered:

1. tests based on the minimum p-value [Tippett (1931)]:

$$p_{\min} = \min_{i=1,\ldots,m} \{p_i\}\ , \quad S_{\min} = 1 - \min_{i=1,\ldots,m} \{p_i\}\ ; \tag{3}$$

   H_0 is rejected when $p_{\min}$ is small, or, equivalently, when $S_{\min}$ is large;

2. tests based on the product of the p-values [Fisher (1932), Pearson (1933)]

$$p_\times = \prod_{i=1}^{m} p_i \tag{4}$$

or one of the following transformations of this product

$$S_\times = 1 - \prod_{i=1}^{m} p_i\,, \quad S_{\ln} = -2\sum_{i=1}^{m} \ln(p_i)\,; \tag{5}$$

H_0 is rejected when $p_\times$ is small, or equivalently, when $S_\times$ (or $S_{\ln}$) is large;

3. tests based on a weighted product of p-values (or a weighted sum of the logarithms of the p-values):

$$S^*_\times = 1 - \prod_{i=1}^{m} p_i^{w_i}\,, \quad S^*_{\ln} = -2\sum_{i=1}^{m} w_i \ln(p_i)\ , \tag{6}$$

where the weights may reflect prior beliefs [Good (1955)] or depend on the p-values [Wilkinson (1951)].

We focus here on two variants of the weighted product procedures. In the first one, we assign zero weight to non-significant individual p-values, which corresponds to (6) with

$$\begin{aligned} w_i &= 1, \text{ if } p_i \le \alpha^*, \quad j = 1, \ldots, m, \\ &= 0, \text{ otherwise,} \end{aligned} \tag{7}$$

where α^* is set as desired and may even be equal to the targeted overall significance level α. In the second one, only the $\tilde{m}$ smallest p-values are included in the test statistic, where $\tilde{m} < m$ is preset. Formally, if $p_{(1)} \le \ \ldots\ \le p_{(i)} \le \ \ldots\ \le p_{(m)}$ are the ordered individual p-values, this corresponds to (6) with

$$\begin{aligned} w_i &= 1, \text{ if } p_i \le p_{(\tilde{m})}, \quad j = 1, \ldots, m, \\ &= 0, \text{ otherwise,} \end{aligned} \tag{8}$$

where $\tilde{m}$ is set as desired. A basic advantage of our approach is that Bonferroni-type bounds are no longer necessary to control the level of the combined test. As long as the weighting index does not depend on nuisance parameters under H_0, our method remains applicable.

If the above proposed statistics are independent with continuous distributions, it is easy to calculate their joint distribution under the null hypothesis. In this case, the individual p-values are independent and identically distributed (i.i.d.) according to $U(0,\, 1)$ distributions, so we have:

$$\begin{aligned} \mathsf{P}[p_{\min} \le \alpha_0] &= 1 - \mathsf{P}[p_1 > \alpha_0,\ \ldots\ ,\, p_m > \alpha_0] = 1 - \prod_{i=1}^{m} \mathsf{P}[p_i > \alpha_0] \\ &= 1 - (1-\alpha_0)^m\,. \end{aligned} \tag{9}$$

We can then choose $\alpha_0 = 1 - (1-\alpha)^{1/m}$ to ensure that the critical region $p_{\min} \le \alpha_0$ has level α. Similarly, in this case,

$$p_\times \sim \prod_{i=1}^{m} U_i \quad \text{where} \quad U_1,\, \dots\,,\, U_m \overset{i.i.d.}{\sim} U(0,\, 1)\,, \tag{10}$$

a distribution which is easy to evaluate (and simulate). Note

$$S_{\ln} = -2\ln(p_\times) \sim \chi^2(2m) \quad \text{under } H_0\,, \tag{11}$$

so critical values can be obtained from the $\chi^2(2m)$ distribution.

When the S_i statistics are not independent, these results are no longer valid, and deriving relevant distributions may be difficult. However, in many situations, this distribution is easy to simulate under H_0, which suggests the following bootstrap-type strategy. Denote by $\bar{S}_0$ the statistic calculated from the observed sample where any choice within the above defined criteria [(3), (5), or (6)] can be considered. For a given number of replications N, let $\bar{S}_1\,,\ \dots\,,\ \bar{S}_N$ denote simulated counterparts of $\bar{S}$ (for example, MC or bootstrap replications) which have the same distribution as $\bar{S}$ under H_0. Further details will be provided on how these may be obtained in the next section, for a specific case.

We can then calculate an empirical p-value from the rank of $\bar{S}_0$ [denoted $\hat{R}_N(\bar{S}_0)$] in the series $\bar{S}_0\,,\ \bar{S}_1\,,\ \dots\,,\ \bar{S}_N$, which leads to the critical region:

$$\hat{p}_N(\bar{S}_0) = \frac{N\hat{G}_N(\bar{S}_0;\, \bar{S}_1\,,\ \dots\,,\ \bar{S}_N) + 1}{N+1} \le \alpha \tag{12}$$

where

$$\hat{p}_N(x) = \frac{N\hat{G}_N(x) + 1}{N+1}\,, \tag{13}$$

$$\hat{G}_N(x) = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{[0,\infty)}(\bar{S}_j - x),\ \mathbf{1}_A(x) = \begin{cases} 1, & if\ x \in A \\ 0, & if\ x \notin A\,. \end{cases} \tag{14}$$

In (12), $N\hat{G}_N(\bar{S}_0; \bar{S}_1\,,\ \dots\,,\ \bar{S}_N)$ is the number of simulated statistics greater than or equal to $\bar{S}_0$. In the following theorem, we establish the following property: if the distribution of the statistics under the null hypothesis can be simulated and does not depend on any unknown parameter, a critical region of the form (12) has level α, provided $\alpha(N+1)$ is an integer.

Theorem 1 *Consider m (not necessarily distinct) hypotheses H_{0i}, $i = 1, \dots, m$, and for each hypothesis H_{0i} a test statistic S_i, where $S_1, \dots, S_m$ may not be independent. Let $\bar{S} = g(S_1, \dots, S_m)$ be a test statistic of the form (3), (5) or (6) for the joint hypothesis*

$$H_0 : \textit{the hypotheses } H_{01}, \dots, H_{0m} \textit{ are all true,}$$

$\bar{S}_0$ the observed value of $\bar{S}$, and $\bar{S}_1, \dots, \bar{S}_N$ additional real random variables. If, under H_0, the joint distribution $(S_1, \dots, S_m)$ is unique (free of nuisance parameters), and $\bar{S}_0, \bar{S}_1, \dots, \bar{S}_N$ are exchangeable with $\mathsf{P}[\bar{S}_i = \bar{S}_j] = 0$ for $i \ne j$, then, for $0 < \alpha < 1$,

$$\mathsf{P}\left[\hat{p}_N(\bar{S}_0) \le \alpha\right] \le \alpha \tag{15}$$

and when N is chosen so that $\alpha(N+1)$ is an integer,

$$\mathsf{P}\left[\hat{p}_N(\bar{S}_0) \le \alpha\right] = \alpha\,, \tag{16}$$

where $\hat{p}_N(x)$ is defined by (13).

Proof. Let $R_N(\bar{S}_j)$ be the rank of $\bar{S}_j$ when $\bar{S}_0, \bar{S}_1, \ldots, \bar{S}_N$ are ranked in increasing order. Since the random variables $\bar{S}_0, \bar{S}_1, \ldots, \bar{S}_m$ are exchangeable and ties have zero probability $\big(\mathsf{P}[\bar{S}_i = \bar{S}_j] = 0$ for $i \ne j\big)$, all rankings are equally probable, and the vector $[R_N(\bar{S}_0), R_N(\bar{S}_1), \ldots, R_N(\bar{S}_N)]'$ is random permutation of the vector $[1, 2, \ldots, N+1]'$. Consequently, for each $j = 0, 1, \ldots, N$, we have

$$\mathsf{P}\big[R_N(\bar{S}_j) = k\big] = \frac{1}{N+1}\,,\ k = 1, 2, \ldots, N+1. \tag{17}$$

and, with probably one,

$$R_N(\bar{S}_0) = N + 1 - N\hat{G}_N(\bar{S}_0)\,,\ \hat{p}_N(\bar{S}_0) = \frac{N+2-R_N(\bar{S}_0)}{N+1} \tag{18}$$

Thus,

$$\mathsf{P}\big[R_N(\bar{S}_0) \le k\big] = \frac{k}{N+1}\,,\ k = 1, 2, \ldots, N+1, \tag{19}$$

$$\begin{aligned}\mathsf{P}\big[R_N(\bar{S}_0) \ge k\big] &= \mathsf{P}\big[R_N(\bar{S}_0) = k\big] + \big[R_N(\bar{S}_0) > k\big] = \frac{1}{N+1} + 1 - \frac{k}{N+1} \\ &= \frac{N+2-k}{N+1}\,,\ k = 1, 2, \ldots, N+1,\end{aligned} \tag{20}$$

hence

$$\mathsf{P}\left[\hat{p}_N(\bar{S}_0) \le \frac{k}{N+1}\right] = \mathsf{P}\left[R_N(\bar{S}_0) \ge N+2-k\right] = \frac{k}{N+1}\,,\ k = 1, 2, \ldots, N+1. \tag{21}$$

If $0 < \alpha < 1$, this entails (15), and for $\alpha(N+1)$ an integer (16). ■

It is easy to see that the above theorem applies when the statistics $\bar{S}_0, \bar{S}_1, \ldots, \bar{S}_m$ are i.i.d. with continuous distribution under H_0. However, for the combined statistic (7), ties have non-zero probability. In the examples considered below, we also propose other non-continuous statistics. Nevertheless, the technique of MC tests can be adapted to discrete distributions using the following randomized tie-breaking procedure; for proofs and further references, see Dufour (2006).

Draw $N+1$ uniformly distributed variates $\tilde{Z}_0, \tilde{Z}_1, \ldots, \tilde{Z}_N$, independently of $(\bar{S}_0, \bar{S}_1, \ldots, \bar{S}_N)$, and arrange the pairs $(\bar{S}_j, \tilde{Z}_j)$ following the lexicographic order:

$$(\bar{S}_i, \tilde{Z}_i) \ge (\bar{S}_j, \tilde{Z}_j) \Leftrightarrow \left[\bar{S}_i > \bar{S}_j \quad \textit{or} \quad (\bar{S}_i = \bar{S}_j \ \ \textit{and}\ \tilde{Z}_i \ge \tilde{Z}_j)\right]. \tag{22}$$

This leads to the MC p-value $\tilde{p}_N(\bar{S}_0)$, where

$$\tilde{p}_N(x) = \frac{N\tilde{G}_N(x) + 1}{N+1}\,, \tag{23}$$

$$\tilde{G}_N(x) = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[0,\infty)}(x - \bar{S}_i) + \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[0]}(\bar{S}_i - x)\ \mathbf{1}_{[0,\infty)}(\tilde{Z}_i - \tilde{Z}_0).$$

The resulting critical region $\tilde{p}_N(\bar{S}_0) \leq \alpha$ has the correct level provided $\alpha(N+1)$ is an integer, *i.e.*

$$\mathsf{P}\left[\hat{p}_N(\bar{S}_0) \leq \alpha\right] \leq \mathsf{P}\left[\tilde{p}_N(\bar{S}_0) \leq \alpha\right] = \frac{I\left[\alpha(N+1)\right]}{N+1}\ , \quad for \ \ 0 \leq \alpha \leq 1\ . \tag{24}$$

The proposed joint test procedure can be summarized as follows. All individual tests are converted to an approximate p-value form, and then combined into a joint criterion whose distribution under H_0 is free of nuisance parameters and can be simulated. When the combined criterion is simulated, the fact that underlying p-values are approximate does not prevent controlling the global test level, so we can get exact combined tests even if the individual p-values are not themselves exact. In other words, provided the statistics are nuisance-parameter-free under H_0, (16) and (24) hold whether the individual p-values p_i, $i = 1, ..., m$ [as in (3), (5), or (6)] are exact, approximate or asymptotic.

It is worth noting that N may be as small as 19 to get a level of 0.05. Power may improve with more replications, but controlling test size does not depend on increasing the number of replications, as in a standard bootstrap. For theoretical insights explaining this feature for MC test methods in general, see Dufour (2006). Theorem 1 underscores this property, for exactness obtains for given N. The simulation study in section 4 shows good power with just 99 replications.

3 Joint serial correlation and predictability tests

To illustrate the usefulness of the above general procedure, this section focuses on serial correlation and predictability tests [1] in the linear model:

$$y_t = x_t'\beta + u_t, \quad u_t = \sigma\varepsilon_t,\ t = 1,\ \ldots\ ,\ T, \tag{25}$$

where $x_t = (1,\ x_{t2},\ \ldots\ ,\ x_{tk})'$, β is a $k \times 1$ vector of unknown coefficients, σ is a scale parameter (which may be random), $\varepsilon = (\varepsilon_1,\ \ldots\ ,\ \varepsilon_T)'$ is a random error term with mean zero, and the distribution of ε, conditional on X, is completely specified:

$$\varepsilon_1, \ldots, \varepsilon_T \text{ are i.i.d. following } F_0\ , \tag{26}$$

where F_0 is a given distribution. For example, we could consider the Gaussian distribution

$$\varepsilon_1, \ldots, \varepsilon_T \overset{i.i.d.}{\sim} N[0,\ 1]\,. \tag{27}$$

Let $y \equiv (y_1,\ \ldots\ ,\ y_T)'$, $X \equiv (x_1,\ \ldots\ ,\ x_T)'$ and $u \equiv (u_1,\ \ldots\ ,\ u_T)'$. The problem of interest consists in assessing

$$\rho_j = 0, \quad for j = 1, 2, ..., \tag{28}$$

where

$$\mathsf{E}(\varepsilon_t\varepsilon_{t-j}) = \rho_j, \quad t = j+1, ..., T. \tag{29}$$

The following assumptions will also be tested and/or maintained.

[1]Here the predicatbility tests refer to varaiance ratio tests. We mentain this terminology as it is widely used in the finance literature. In particular, to evaluate return predictability, variance ratio test are employed, hence the name of predictatbility tests was used.

Assumption 1 *The distribution of the random vector ε is continuous and completely specified; the hypothesis of i.i.d. normal errors is of course a special case.*

Assumption 2 *The regressor matrix X is fixed or independent of the error term u.*

To derive finite-sample tests for the above problem, we consider the OLS residuals:

$$\hat{u} = \left(\hat{u}_1, \ldots, \hat{u}_T\right)', \quad \hat{u}_t = y_t - x_t'\hat{\beta}, \ t = 1, \ \ldots \ \ T, \quad \hat{\beta} = (X'X)^{-1}X'y. \tag{30}$$

The test statistics we shall use are functions of the standardized residual vector $\hat{u}/\hat{\sigma}$, where

$$\hat{\sigma}^2 = \sum_{t=1}^{T} \hat{u}_t^2/T = \hat{u}'\hat{u}/T. \tag{31}$$

Theorem 2 *In the context of the linear regression* (*25*) *along with Assumptions 1 and 2, the conditional distribution of the scaled residual vector $\hat{u}/\hat{\sigma}$, given X, only depends on the distribution of $(\varepsilon_1, \ \ldots \ , \ \varepsilon_T)'$.*

Proof. On observing that $\hat{\sigma} = (\hat{u}'\hat{u}/T)^{1/2}$ and

$$\hat{u} = M_X u, \quad M_X = I_n - X\left(X'X\right)^{-1}X', \tag{32}$$

it is easy to see that

$$\frac{\hat{u}}{\hat{\sigma}} = T^{1/2}\frac{M_X u}{(u'M_X u)^{1/2}} = T^{1/2}\frac{M_X\,(u/\sigma)}{((u/\sigma)'\,M_X\,(u/\sigma))^{1/2}} = T^{1/2}\frac{M_X\varepsilon}{(\varepsilon' M_X \varepsilon)^{1/2}} \tag{33}$$

which establishes the desired result (when X is fixed). This means that $\hat{u}/\hat{\sigma}$ has a known distribution under all hypotheses which completely specify the distribution of the random vector ε. ■

Most commonly used serial correlation tests are based on residual empirical autocorrelations:

$$\hat{\rho}_j = \frac{\sum_{t=j+1}^{T}\hat{u}_t\hat{u}_{t-j}}{\sum_{t=1}^{T}\hat{u}_t^2}, \ j = 1, \ldots, m, \tag{34}$$

where m is usually pre-set (given the size of the sample). Indeed, the well-known Ljung-Box statistic [Ljung and Box (1978)] is

$$LB(J) = T(T+2)\sum_{j=1}^{J}\frac{\hat{\rho}_j^2}{T-j}\,. \tag{35}$$

In location-scale models, the asymptotic null distribution of $LB(J)$ is $\chi^2(J)$. In practical applications, this limiting distribution is also informally used with regression residuals; see Dezhbakhsh (1990).

Another choice of test involves the variance-ratio statistic, proposed by Cochrane (1988) and Lo and MacKinlay (1988) to test market efficiency. Heteroskedastic-robust versions of this test are not of particular interest here, so we focus on the statistic

$$VR(J) = 1 + 2\sum_{j=1}^{J-1}(1 - \frac{j}{J})\hat{\rho}_j \tag{36}$$

which can be viewed as an estimate of the ratio

$$\mathcal{VR}(J) = \frac{V(\hat{u}_t - \hat{u}_{t-J})}{JV(\hat{u}_t)}$$

where $V(\hat{u}_t - \hat{u}_{t-J})$ is the variance of the lag differences $\hat{u}_t - \hat{u}_{t-J}$, and $V(\hat{u}_t)$ is the residual variance. Under the null hypothesis, $V(\hat{u}_t - \hat{u}_{t-J})$ is J times $V(\hat{u}_t)$, for all J, hence deviations from a ratio of one can be viewed as evidence against the null hypothesis. The asymptotic null distribution of $VR(J)$ is given by

$$VR(J) \overset{asy}{\sim} N[1, 2(2J-1)(J-1)/(3J)]. \tag{37}$$

Attempts to improve the latter approximation include the bootstrap-based algorithms of Malliaropulos (1996), Politis, Romano, and Wolf (1997), and Kim (2006), and a subsampling-based modification by Whang and Kim (2003). See Wright (2000) for an alternative statistic based on signs and ranks, and Charles and Darne (2009) for a general overview. Chou and Denning (1993), Fong, Koh, and Ouliaris (1997) and Yilmaz (2003) emphasize the importance of the joint interpretation of the variance ratios for all relevant J.

Let us first observe that the empirical autocorrelations are indeed a function of the standardized residual vector. To see this, let

$$\hat{u}_{[1:T-j]} = \left(\hat{u}_1, \ldots, \hat{u}_{T-j}\right)' = \underline{A}_{[j]}\hat{u}, \quad \hat{u}_{[j+1:T]} = \left(\hat{u}_{j+1}, \ldots, \hat{u}_T\right)' = \overline{A}_{[j]}\hat{u}$$

where

$$\underline{A}_{[j]} = [I_{T-J}, \text{zeros}(T-j, j)], \quad \overline{A}_{[j]} = [\text{zeros}(T-j, j), I_{T-J}]$$

are selection matrices with dimension $(T-j) \times T$. Then, for all lags j, we have:

$$\hat{\rho}_j = \frac{\hat{u}'_{[1:T-j]}\hat{u}_{[j+1:T]}}{\hat{u}'\hat{u}} = \frac{\hat{u}'\underline{A}'_{[j]}\overline{A}_{[j]}\hat{u}}{T\hat{\sigma}^2} = T^{-1}\left(\hat{u}/\hat{\sigma}\right)'\underline{A}'_{[j]}\overline{A}_{[j]}\left(\hat{u}/\hat{\sigma}\right). \tag{38}$$

On using Theorem 2, it follows that the joint distribution of the autocorrelations $\hat{\rho}_j$ only depends on the distribution of the vector ε. Under the null hypothesis (28) and the Assumptions 1 and 2, (38) implies that the autocorrelations in question are jointly pivotal. This property is shared with any statistic which depends on the data only through these autocorrelations.

Among many statistics which may be used, we have considered the following ones.

1. The MC versions of the tests based on the Ljung-Box and variance-ratio statistics [in (35) and (36)] with $J = m$.

2. The minimum p-value test, denoted $AC_{\min}$, based on the individual autocorrelation [see (3)]: here S_i corresponds to $\hat{\rho}_i^2$, and p_i is obtained using

$$\sqrt{T}\,\hat{\rho}_i \overset{asy}{\sim} N[0, 1], \ i = 1, \ldots, m. \tag{39}$$

3. The minimum p-value statistic, denoted $VR_{\min}$, based on a sequence of variance ratios [see (3)]: S_i corresponds to $VR(i)$ as defined by (36), $i = 1, \ldots, m$, and p_i is obtained using (37).

4. The Ljung-Box statistic (35) with $J = l^{\mathsf{e}}$ and l^{e} is the lag which corresponds to the largest significant [at the 5% level] autocorrelation (until a maximal lag length as permitted by the data). To assess significance, we use $\hat{\rho}_i^2$, and the approximate distribution (39). We denote this statistic LB_{e}. If none of the autocorrelations is significant, LB_{e} is set to zero.

5. The variance-ratio in (36) where $J = l^{\mathsf{e}}$, where l^{e} is the lag corresponding to the largest significant [at level 5%] variance ratio. To assess significance we use $|VR(J)|$ and the approximate distribution (37). We denote this statistic VR_{e}. If none of the variance-ratio statistics is significant, VR_{e} is set to zero.

6. The combined criterion [see (5)], denoted $AC_{\times}$, based on the product of the p-values p_i associated with $\hat{\rho}_i^2$, $i = 1, \ldots, m$, each obtained using the approximate distribution (39).

7. The combined criterion [see (6)], denoted $AC_{\times}^{*}$, based on the product of the significant p-values [at level 5%] associated with $\hat{\rho}_i^2$, $i = 1, \ldots, m$; the individual p-values are computed from the approximate null distribution (39).

8. The combined criterion [see (5)], denoted $VR_{\times}$, based on the product of the p-values p_i associated with $|VR(i)|$ as defined by (36), $i = 1, \ldots, m$, using (37).

9. The combined criterion (6) based on the product of the significant [at the 5% level] p-values associated with $|VR(i)|$ as defined by (36) with $i = 1, \ldots, m$, and obtained using (37). We denote this statistic $VR_{\times}^{*}$.

Since the conditions of Theorem 1 hold for all these statistics, it follows that the MC p-values provided by (16) or (23) would have the correct size for any sample size. Observe we can set the individual significance levels underlying $VR_{\times}^{*}$, VR_{e}, $AC_{\times}^{*}$ and LB_{e}, at 5% and still obtain a test with global level 5%. Size control is achieved even if approximate distributions are used to calculate the individual p-values. The MC test method achieves size control as long as joint pivotality holds.

Three properties further explain why computational expense is not an issue for our proposed combination methods. (1) The proposed joint tests are exact even if individual p-values are themselves not exact. Inexpensive standard asymptotic approximations [for example normal or χ^2] can be used for individual tests with no effect on the finite-sample properties of the joint test. (2) Joint test criteria are the minimum or the (possibly weighted) product of the individual p-values so obtained. These operations are also inexpensive. (3) While we must replicate (1) and (2) N times, both operations are inexpensive, and N need not be very large as in a standard bootstrap. Our simulation study was conducted with $N = 99$ to underscore this feature.

4 Simulation study

To illustrate the performance of the serial correlation tests, we consider the following experiment. The base model is (25). The regressors are generated as *i.i.d.* standard normal (kept fixed over the simulation). Sample sizes of $T = 32,\ 60$, are used and k (the number of regressors) is set as the largest integer less than or equal to $\sqrt{T}$. Under the null hypothesis, the error terms u_t, $t = 1, \ldots, T$, are drawn as *i.i.d.*, from the following distribution: $N(0,1)$, $\chi^2(2)$, $U[-.5,.5]$, $t(5)$ and Cauchy. For the power study, we assume the AR(2) error process

$$u_t = \rho_1 u_{t-1} + \rho_1 u_{t-2} + \eta_t,\ t = 1, \ldots, T,$$

where the (fixed) initial values are zero and the error terms η_t, $t = 1, \ldots, T$, are *i.i.d.* $N(0,1)$ and $t(5)$. We consider: $(\rho_1, \rho_2) = (.5, .2), (.7, -.2), (1, -.2), (1.3, -.5)$. All statistics defined in the previous section are studied. MC tests are applied with $N = 99$; randomized ranks are used for non-continuous statistics. Each study relies on 1000 replications.[2] The results are summarized in Tables 1– 3. The main features of these can be summarized as follows.

1. The performance of the asymptotic tests is unsatisfactory. In the presence of normal, χ^2 and uniform errors, the Ljung-Box test is oversized; the problem gets worse when more lags are considered. This issue is important since practitioners tend to consider as many lags as possible with these tests. When errors are Cauchy or t-distributed, the Ljung-Box test seems undersized. Turning to the asymptotic variance ratio, it is evident that the test is severely undersized, in all cases. In particular, no rejections at all are observed under the null for $T = 32$ with Cauchy or t-distributed disturbances. The sizes of the MC version of both the Ljung-Box and variance-ratio tests are controlled in all cases. This clearly affects the power of the latter test, which improves sometimes dramatically. For instance, for $T = 32$ and $t(5)$ errors, with $\rho_1 = .7$ and $\rho_2 = -.2$, empirical rejections increase from $\simeq$19% (for the asymptotic variance ratio) to $\simeq$78% (for its MC version); see Table 3.

2. The Tippett-type autocorrelation tests tend to outperform the standard Ljung-Box test, as more lags are used. The best test in this category is the one based on the significant autocorrelations. Observe however that the $AC_{\min}$ statistic performs equally well and sometimes marginally better in this example. The same holds for the variance-ratio criteria, except for two observations: (i) the power advantage of the Tippett-type tests is in general stronger; (ii) the $VR_{\min}$ statistic performs as well as but not better than its Tippet counterpart. The proposed combined criteria perform better than the MC version of available test statistics. For example (see Table 1), with $T = 32$ and normal errors, $\rho_1 = .7$ and $\rho_2 = -.2$, empirical rejections increase from 18% (for the MC variance ratio) to $\simeq$66% (for its min-p or Tippet MC version); for the same values of $\rho_1$and ρ_2, with $T = 60$ and $t(5)$ errors, the power jumps from $\simeq$40 to $\simeq$ 97%.

3. There is no apparent difference in the MC tests power ranking, with normal or t-errors. As outlined above, effective power improvements, which result from size-correction, are more visible with t-errors, since the available asymptotic tests perform worse in this case.

[2]As an example of execution time: one run, using 32 bit GAUSS with a 1.66 GHz CPU, for a sample size of 62 along with $N = 99$, for all the considered statistics executed in one algorithm, ends in less than 17 seconds.

4. The endogenous-lag criteria do not provide improvements over the tests based on the min-p and p-value product statistics.

5. The variance-ratio tests appear preferable to the Ljung-Box-type tests. Though both statistics are functions of the sample autocorrelations, the variance ratio exploits further features of white-noise behavior, including variance linearity (over the sampling interval); see Lo and MacKinlay (1988). This may confer a power advantage to these tests, which is revealed in our results once test size is controlled by the MC test method.

5 Conclusion

This paper suggests MC multiple test procedures which are provably valid in finite samples. These include combination methods originally proposed for independent statistics and further improvements which formalize statistical or econometric practice. We also adapt the MC method for non-continuous combined statistics. These methods are applied to serial dependence and predictability tests. We propose new tests which allow, among others, endogenous lag selection. We conduct a simulation study to illustrate the usefulness of the proposed procedures. In general, our results show that concrete and non-spurious power gains (over standard combination methods) can be achieved through our multiple Monte Carlo test approach, and confirms arguments in favour of variance-ratio type criteria.

To conclude, it is worth revisiting the above discussed fat-tailed case with Student-t errors, and allow for the possibility of an unknown degrees-of-freedom parameter, denoted ν. To deal with the latter as a nuisance-parameter, various procedures have been suggested and applied [for different though related test problems] in Dufour, Khalaf, and Beaulieu (2003), Dufour, Khalaf, Bernard, and Genest (2004), Bernard, Idoudi, Khalaf, and Yelou (2007), Beaulieu, Dufour, and Khalaf (2007), Dufour, Khalaf, and Beaulieu (2010), and Beaulieu, Dufour, and Khalaf (2012). These consist of maximizing the MC p-value for the tested hypothesis (which depends on the nuisance parameter) over the relevant nuisance parameter space. For the problem at hand, the joint distributions of the combined criteria depend on ν. Any relevant [*i.e.* conforming with the null hypothesis] value for ν can lead to an empirical p-value as outlined in section (2), given the value of ν in question. This leads to a p-value "function", denoted $\hat{p}_N(\,\cdot\,|\,\nu)$. The *maximized MC* method introduced by Dufour (2006) and applied in the above-cited works involves (numerically) maximizing the p-value function $\hat{p}_N(\,\cdot\,|\,\nu)$ over all relevant values of ν. The test critical region corresponds to referring the supremum $\sup_\nu\,[\hat{p}_N(\,\cdot\,|\,\nu)]$ to a given level α.

An alternative method originally proposed by Dufour and Kiviet (1996) and denoted the consistent set maximized MC [CSMMC] method involves two stages: (1) an exact confidence set is built for ν, and (2) the MC p-value $\hat{p}_N(.|\nu)$ is maximized over all values of ν in the latter confidence set. So far, the latter method was applied [in the above cited works] using Bonferroni-type bounds over each stage. While extending our analysis to this case is beyond the scope of this paper, it is intuitively appealing to treat the CSMMC test, in turn, as a combined test, and re-sample the whole procedure. Results available so far suggest that simulation-based combination methods to treat distributional nuisance parameters is a promising avenue for further research.

Table 1: Empirical rejections: Size and power of serial correlation tests; normal errors

			AR(2) parameters ρ_1, ρ_2				
T	Lags (m)	Statistic	$0, 0$	$.5, .2$	$.7, -.2$	$1, -.2$	$1.3, -.5$
32	5	LB_∞	6.6	51.2	52.2	89.3	96.6
		LB	4.3	45.5	45.2	85.0	95.1
		$AC_\times$	4.4	45.3	44.8	84.7	95.2
		$AC_{\min}$	3.9	45.6	55.1	89.9	97.2
		VR_∞	1.5	66.3	45.3	90.6	95.2
		VR	3.6	71.0	51.5	92.9	96.8
		$VR_\times$	3.5	70.0	64.1	95.4	98.8
		$VR_{\min}$	4.2	67.7	66.9	95.2	98.9
32	10	LB_∞	8.3	49.6	49.4	86.4	95.0
		LB	4.5	38.5	35.4	76.5	90.2
		$AC_\times$	4.9	39.0	35.9	76.8	89.9
		$AC_{\min}$	3.5	42.9	50.4	87.7	96.3
		VR_∞	0.6	44.6	17.1	63.8	60.6
		VR	3.7	57.9	29.4	74.2	73.6
		$VR_\times$	3.5	70.6	51.5	92.7	97.0
		$VR_{\min}$	4.1	69.4	66.6	95.3	98.9
32	15	LB_∞	9.8	48.3	48.7	83.4	93.9
		LB	4.6	35.6	32.6	70.4	84.8
		$AC_\times$	4.7	37.1	35.4	72.2	85.9
		$AC_{\min}$	3.5	42.6	50.1	87.2	96.3
		VR_∞	0.3	27.1	7.0	41.0	35.2
		VR	3.5	42.7	18.2	57.7	51.9
		$VR_\times$	3.6	66.8	44.9	89.5	95.5
		$VR_{\min}$	4.1	69.4	66.6	95.3	98.9
32	≤ 15	$AC^*_\times$	3.9	41.5	40.5	80.3	91.3
		LB_{e}	4.0	27.4	25.3	59.5	75.4
		$VR^*_\times$	3.6	72.3	66.2	95.8	98.7
		VR_{e}	7.4	31.1	27.8	23.0	16.8

Note – Frequencies are given in percentages (%). LB and KS are the MC Ljung-Box and variance-ratio tests [see (35) - (36)] with $J = m$, where m is reported in column 2. LB_∞ and KS_∞ are their asymptotic counterparts, using the $\chi^2(m)$ for the former, and (37) for the latter. $AC_{\min}$ is (3) where S_i corresponds to $\hat{\rho}_i^2$, $i = 1, \ldots, m$, and p_i is obtained using (39); $AC_\times$ is its product counterpart. $VR_{\min}$ is (3) where S_i corresponds to $|VR(i)|$ with $i = 1, \ldots, m$, and p_i is obtained using (37); $VR_\times$ is its product counterpart. LB_{e} is the Ljung-Box statistic (35) with $J = l^{\mathsf{e}}$ and l^{e} is the lag which corresponds to the largest significant [at the 5% level] autocorrelation. VR_{e} is the variance-ratio (36) where $J = l^{\mathsf{e}}$ and l^{e} is the lag which corresponds to the largest significant [at the 5% level] variance ratio. $AC^*_\times$ is (6) based on the product of the significant p-values [at the 5% level] associated with $\hat{\rho}_i^2$, $i = 1, \ldots, m$. $VR^*_\times$ is (6) based on the product of the significant [at the 5% level] p-values associated with $|VR(i)|$ for $i = 1, \ldots, m$.

Table 1 (continued)

			AR(2) parameters ρ_1, ρ_2				
T	Lags (m)	Statistic	$0, 0$	$.5, .2$	$.7, -.2$	$1, -.2$	$1.3, -.5$
60	5	LB_∞	7.1	92.4	94.6	99.9	100
		LB	4.4	91.0	91.2	99.9	100
		$AC_\times$	4.5	90.1	88.6	99.9	100
		$AC_{\min}$	4.6	89.8	96.7	100	100
		VR_∞	2.7	95.4	82.1	99.8	100
		VR	3.7	95.8	82.3	99.8	100
		$VR_\times$	3.9	96.7	95.9	100	100
		$VR_{\min}$	3.6	96.5	98.6	100	100
60	10	LB_∞	8.2	89.0	87.8	99.8	100
		LB	5.2	84.3	80.3	99.2	100
		$AC_\times$	5.1	83.8	76.8	99.0	100
		$AC_{\min}$	4.2	88.5	95.1	100	100
		VR_∞	1.7	86.9	45.2	93.9	91.4
		VR	4.0	88.4	52.4	94.9	93.9
		$VR_\times$	4.0	95.7	86.6	99.9	100
		$VR_{\min}$	4.0	96.4	98.4	100	100
60	15	LB_∞	9.5	87.6	83.9	99.5	99.9
		LB	4.8	82.0	71.8	98.6	99.9
		$AC_\times$	4.3	81.0	70.4	98.1	99.8
		$AC_{\min}$	4.4	87.7	94.5	100	100
		VR_∞	1.9	73.1	29.2	84.1	73.0
		VR	5.4	80.0	38.0	89.0	80.5
		$VR_\times$	3.9	94.4	79.8	98.1	98.3
		$VR_{\min}$	4.1	96.4	98.3	100	100
60	≤ 15	$AC_\times^*$	4.6	85.0	75.6	99.6	99.9
		LB_{e}	4.3	70.3	54.0	96.9	99.2
		$VR_\times^*$	4.0	96.0	91.4	100	100
		VR_{e}	7.1	35.7	27.5	28.4	15.7

Table 2: Empirical size of serial correlation tests with non-normal errors

			Error distribution				
T	Lags (m)	Statistic	$N(0,1)$	$\chi^2(2)$	$U[-.5,.5]$	$t(5)$	Cauchy
32	5	LB_∞	6.6	4.3	7.5	2.2	1.6
		LB	4.3	4.1	4.8	3.8	4.9
		$AC_\times$	4.4	4.6	4.9	3.2	4.6
		$AC_{\min}$	3.9	4.4	4.8	3.6	4.5
		VR_∞	1.5	2.0	1.3	0.8	0.7
		VR	3.6	4.4	4.2	4.0	4.2
		$VR_\times$	3.5	4.7	7.7	3.9	4.8
		$VR_{\min}$	4.2	5.2	4.8	3.1	5.2
32	10	LB_∞	8.3	5.8	8.1	2.7	2.6
		LB	4.5	4.7	4.9	3.8	4.7
		$AC_\times$	4.9	4.7	4.8	3.8	4.8
		$AC_{\min}$	3.5	4.1	4.9	4.0	4.5
		VR_∞	0.6	0.3	0.8	0.2	0.3
		VR	3.7	3.8	4.0	4.1	3.2
		$VR_\times$	3.5	4.8	4.3	3.4	4.4
		$VR_{\min}$	4.1	5.8	4.8	3.3	5.0
32	15	LB_∞	9.8	6.9	10.3	3.3	2.4
		LB	4.6	3.7	4.0	3.6	4.9
		$AC_\times$	4.7	4.5	4.5	4.2	4.8
		$AC_{\min}$	3.5	3.5	4.7	3.5	4.2
		VR_∞	0.3	0.1	0.3	0.0	0.0
		VR	3.5	4.3	4.2	4.1	3.5
		$VR_\times$	3.6	4.3	4.5	3.5	3.5
		$VR_{\min}$	4.1	4.7	4.8	3.3	5.0
32	≤ 15	$AC^*_\times$	3.9	4.4	4.0	4.0	4.4
		LB_{e}	4.0	5.3	4.7	3.5	4.2
		$VR^*_\times$	3.6	5.3	4.7	6.0	6.3
		VR_{e}	7.4	5.3	4.7	6.0	6.1

Note – For definitions, see Table 1.

Table 2 (continued)

			Error distribution				
T	Lags (m)	Statistic	$N(0,1)$	$\chi^2(2)$	$U[-.5,.5]$	$t(5)$	$Cauchy$
60	5	LB_∞	7.1	4.4	6.4	2.9	1.8
		LB	4.4	3.7	4.6	4.3	3.2
		$AC_\times$	4.5	3.8	4.4	4.7	3.3
		$AC_{\min}$	4.6	3.6	4.7	3.9	3.6
		VR_∞	2.7	2.5	3.4	1.7	1.0
		VR	3.7	4.2	4.3	4.4	4.3
		$VR_\times$	3.9	4.1	3.8	3.5	3.7
		$VR_{\min}$	3.6	3.7	4.1	4.3	3.5
60	10	LB_∞	8.2	6.2	7.5	2.9	2.0
		LB	5.2	4.0	4.8	3.8	4.2
		$AC_\times$	5.1	4.0	5.2	3.6	4.2
		$AC_{\min}$	4.2	3.9	4.0	4.3	3.8
		VR_∞	1.7	1.1	1.8	0.8	0.3
		VR	4.0	3.7	3.6	4.5	3.7
		$VR_\times$	4.0	3.6	4.4	3.5	3.3
		$VR_{\min}$	4.0	3.2	3.9	3.9	3.2
60	15	LB_∞	9.5	5.0	8.6	3.1	2.1
		LB	4.8	3.7	4.2	3.9	4.4
		$AC_\times$	4.3	3.4	4.7	4.3	4.4
		$AC_{\min}$	4.4	4.4	3.8	4.2	4.4
		VR_∞	1.9	0.8	1.4	0.5	0.3
		VR	5.4	4.6	4.3	4.4	3.1
		$VR_\times$	3.9	3.5	4.4	3.9	3.3
		$VR_{\min}$	4.1	3.4	4.0	4.1	3.2
60	≤ 15	$AC^*_\times$	4.6	3.6	4.4	4.3	4.7
		LB_{e}	4.3	4.4	3.8	4.2	4.4
		$VR^*_\times$	4.0	4.0	4.0	3.9	3.8
		VR_{e}	7.1	3.4	4.0	4.0	4.1

Table 3: Power of serial correlation tests; $t(5)$ errors

			AR(2) parameters ρ_1, ρ_2				
T	Lags (m)	Statistic	$0, 0$	$.5, .2$	$.7, -.2$	$1, -.2$	$1.3, -.5$
32	15	LB_∞	3.3	50.0	40.0	81.4	90.2
		LB	3.6	52.4	42.7	82.5	90.4
		$AC_\times$	4.2	55.0	43.4	83.9	90.9
		$AC_{\min}$	3.5	52.7	55.0	93.0	97.1
		VR_∞	0.0	35.8	19.1	51.4	36.4
		VR	4.1	56.0	77.7	67.7	55.5
		$VR_\times$	3.5	77.3	52.7	95.1	97.5
		$VR_{\min}$	3.3	80.5	79.7	98.0	99.4
32	≤ 15	$AC^*_\times$	4.0	52.7	55.7	93.0	97.1
		LB_e	3.5	56.3	55.3	90.8	96.2
		$VR^*_\times$	6.0	79.7	76.9	97.8	99.1
		VR_e	6.0	79.2	76.0	97.8	99.2
60	15	LB_∞	3.1	86.1	73.5	99.8	99.7
		LB	3.9	86.4	76.8	99.8	99.7
		$AC_\times$	4.3	86.4	74.3	99.5	99.5
		$AC_{\min}$	4.2	86.6	92.4	99.8	100
		VR_∞	0.5	78.5	25.5	87.2	76.5
		VR	4.4	85.8	39.9	93.0	85.9
		$VR_\times$	3.9	97.8	87.3	99.7	99.7
		$VR_{\min}$	4.2	98.7	98.9	100	100
60	≤ 15	$AC^*_\times$	4.3	86.6	92.4	99.8	100
		LB_e	4.2	89.2	86.1	99.8	100
		$VR^*_\times$	3.9	98.7	98.8	100	100
		VR_e	4.0	99.2	97.3	100	100

Note – For definitions, see Table 1.

References

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., J.-M. Dufour, and L. Khalaf (2007): "Multivariate Tests of Mean-Variance Efficiency with Possibly Non-Gaussian Errors: An Exact Simulation-Based Approach," *Journal of Business and Economic Statistics*, 25, 398–410.

——— (2012): "Identification-Robust Estimation and Testing of the Zero-Beta CAPM," *The Review of Economics Studies*, forthcoming.

Bennett, C. (2012a): "On Bootstrap Minimum P-value Tests," Discussion paper, Vanderbilt University.

——— (2012b): "Resampling-based Multiple Testing: A Synthesis of New and Existing Results," Discussion paper, Vanderbilt University.

Bennett, C., and B. Thompson (2012): "Graphical Procedures for Multiple Comparisons under General Dependence," Discussion paper, Vanderbilt University.

Bernard, J.-T., N. Idoudi, L. Khalaf, and C. Yelou (2007): "Finite Sample Multivariate Structural Change Tests with Application to Energy Demand Models," *Journal of Econometrics*, 141, 1219–1244.

Bewley, R., and H. Theil (1987): "Monte Carlo Testing for Heteroscedasticity in Equation Systems," *Advances in Econometrics*, 6, 1–15.

Box, G. E. P., and D. A. Pierce (1970): "Distribution of Residual Autocorrelations In Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association*, 65, 1509–1526.

Charles, A., and O. Darne (2009): "Variance-Ratio Tests of Random Walk; an Overview," *Journal of Economic Surveys*, 23, 503–527.

Chernick, M. R. (2008): *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons, Inc., second edn.

Chou, P.-H., and K. Denning (1993): "A Simple Multiple Variance Ratio Test," *Journal of Econometrics*, 58, 385–401.

Cochrane, J. H. (1988): "How Big is the Random Walk in GNP," *Journal of Political Economy*, 96, 893–920.

Davison, A., and D. Hinkley (1997): *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge (UK).

Deschamps, P. (1996): "Monte Carlo Methodology for LM and LR Autocorrelation Tests in Multivariate Regressions," *Annales d'Économie et de Statistique*, 43, 150–169.

Dezhbakhsh, H. (1990): "The Inapropriate Use of Serial Correlation Tests in Dynamic Linear Models," *The Review of Economics and Statistics*, 72, 126–132.

Dudoit, S., and M. J. van der Laan (2008): *Multiple Testing Procedures with Applications to Genomics*. Springer, New York.

Dufour, J.-M. (1989): "Nonlinear Hypotheses, Inequality Restrictions, and Non-Nested Hypotheses: Exact Simultaneous Tests in Linear Regressions," *Econometrica*, 57, 335–355.

——— (2006): "Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics," *Journal of Econometrics*, 133, 443–478.

Dufour, J.-M., A. Farhat, L. Gardiol, and L. Khalaf (1998): "Simulation-Based Finite Sample Normality Tests in Linear Regressions," *The Econometrics Journal*, 1, 154–173.

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

Dufour, J.-M., and L. Khalaf (2001): "Monte Carlo Test Methods in Econometrics," in *Companion to Theoretical Econometrics*, ed. by B. Baltagi, Blackwell Companions to Contemporary Economics, chap. 23, pp. 494–519. Basil Blackwell, Oxford, U.K.

——— (2002): "Exact Tests for Contemporaneous Correlation of Disturbances in Seemingly Unrelated Regressions," *Journal of Econometrics*, 106, 143–170.

Dufour, J.-M., L. Khalaf, and M.-C. Beaulieu (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, 891–906.

——— (2010): "Multivariate Residual-Based Finite-Sample Tests for Serial Dependence and GARCH with Applications to Asset Pricing Models," *Journal of Applied Econometrics*, 25, 263–285.

Dufour, J.-M., L. Khalaf, J.-T. Bernard, and I. Genest (2004): "Simulation-Based Finite-Sample Tests for Heteroskedasticity and ARCH Effects," *Journal of Econometrics*, 122(2), 317–347.

Dufour, J.-M., and J. F. Kiviet (1996): "Exact Tests for Structural Change in First-Order Dynamic Models," *Journal of Econometrics*, 70, 39–68.

Dufour, J.-M., and O. Torrès (1998): "Union-Intersection and Sample-Split Methods in Econometrics with Applications to SURE and MA Models," in *Handbook of Applied Economic Statistics*, ed. by D. E. A. Giles, and A. Ullah, pp. 465–505. Marcel Dekker, New York.

Dwass, M. (1957): "Modified Randomization Tests for Nonparametric Hypotheses," *Annals of Mathematical Statistics*, 28, 181–187.

Efron, B., and R. J. Tibshirani (1993): *An Introduction to the Bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York.

Fisher, R. A. (1932): *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Folks, J. L. (1984): "Combination of Independent Tests," in *Handbook of Statistics 4: Nonparametric Methods*, ed. by P. R. Krishnaiah, and P. K. Sen, pp. 113–121. North-Holland, Amsterdam.

Fong, M. F., S. K. Koh, and S. Ouliaris (1997): "Joint Variance-Ratio Tests of the Martingale Hypothesis for Exchange Rates," *Journal of Business and Economic Statistics*, 15, 51–59.

Good, I. J. (1955): "On the Weighted Combination of Significance Tests," *Journal of the Royal Statistical Society, Series B*, 17, 264–265.

Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Horowitz, J. L. (1997): "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in *Advances in Economics and Econometrics*, ed. by D. Kreps, and K. W. Wallis, vol. 3, pp. 188–222. Cambridge University Press, Cambridge, U.K.

Jarque, C. M., and A. K. Bera (1980): "Efficient Tests for Normality, Heteroscedasticity and Serial Independence of Regression Residuals," *Economics Letters*, 6, 255–259.

Kiefer, N. M., and M. Salmon (1983): "Testing Normality in Econometric Models," *Economic Letters*, 11, 123–127.

Kilian, L., and U. Demiroglu (2000): "Residual-Based Tests for Normality in Autoregressions: Asymptotic Theory and Simulation Evidence," *Journal of Business and Economic Statistics*, 18, 40–50.

Kim, J. H. (2006): "Wild bootstrapping variance ratio tests," *Economics Letters*, 92, 38–43.

Ljung, G. M., and G. E. P. Box (1978): "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 65, 297–303.

Lo, A., and C. MacKinlay (1988): "Stock Prices do not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, 1, 41–66.

——— (1989): "The Size and Power of the Variance Ratio Test in Finite Samples: A Monte Carlo Investigation," *Journal of Econometrics*, 40, 203–238.

Malliaropulos, D. (1996): "Are Long-Horizon Stock Returns Predictable? A Bootstrap Analysis," *Journal of Business Finance and Accounting*, 23, 93–106.

Miller, Jr., R. G. (1981): *Simultaneous Statistical Inference*. Springer-Verlag, New York, second edn.

Pearson, K. (1933): "On a Method of Determining Whether a Sample of Size n Supposed to Have Been Drawn from a Parent Population," *Biometrika*, 25, 379–410.

Politis, D. N., J. P. Romano, and M. Wolf (1997): "Subsampling for Heteroskedastic Time Series," *Journal of Econometrics*, 81, 281–317.

Savin, N. E. (1984): "Multiple Hypothesis Testing," in *Handbook of Econometrics, Volume 2*, ed. by Z. Griliches, and M. D. Intrilligator, chap. 14, pp. 827–879. North-Holland, Amsterdam.

Tippett, L. H. (1931): *The Methods of Statistics*. Williams and Norgate, London.

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

Whang, Y, J., and J. Kim (2003): "A Multiple Variance Ratio Test Using Subsampling," *Economics Letters*, 79, 225–230.

White, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.

Wilkinson, B. (1951): "A Statistical Consideration in Psychological Research," *Psychology Bulletin*, 48, 156–158.

Wright, J. H. (2000): "Alternative Variance-Ratio Tests Using Ranks and Signs," *Journal of Business and Economic Statistics*, 18, 1–9.

Yilmaz, K. (2003): "Martingale Property of Exchange Rates and Central Bank Intervention," *Journal of Business and Economic Statistics*, 21, 383–395.

Zhou, G. (1993): "Asset-Pricing Tests under Alternative Distributions," *The Journal of Finance*, 48, 1927–1942.