Multivariate residual-based finite-sample tests for serial dependence and ARCH effects with applications to asset pricing models*

Jean-Marie Dufour † McGill University

Lynda Khalaf ‡ Carleton University

Marie-Claude Beaulieu § Université Laval

First version: February 2003
Revised: September 2003, October 2004, July 2006, September 2007
This version: February 2008
Compiled: August 21, 2008, 1:14pm

*The authors thank the Editor Tim Bollerslev, and three anonymous referees for several helpful comments. This work was supported by the William Dow Chair in Political Economy (McGill University), the Canada Research Chair Program (Chair in Econometrics, Université de Montréal, and Chair in Environmental and Financial Econometric Analysis, Université Laval), the RBC Chair in Financial Innovations (Université Laval), the Bank of Canada (Research Fellowship), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), the Institut de finance mathématique de Montréal (IFM2), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], 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 (Québec), and the Fonds de recherche sur la nature et les technologies (Québec), and a Killam Fellowship (Canada Council for the Arts). This paper was also partly written at the Centre de recherche en économie et statistique (INSÉE, Paris), the Institut für Wirtschaftsforschung Halle (Germany), and the Finance Division at the University of British Columbia.

† William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 519, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 8879; FAX: (1) 514 398 4938; e-mail: jean-marie.dufour@mcgill.ca . Web page: http://www.jeanmariedufour.com

‡ Economics Department, Carleton University, CIREQ, and Groupe de recherche en économie de l'énergie, de l'environnement et des ressources naturelles (GREEN), Université Laval. Mailing address: Economics Department, Carleton University, Loeb Building 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada. TEL: 1 (613) 520 2600 ext. 8697; FAX: 1 (613) 520 3906; e-mail: Lynda_Khalaf@carleton.ca

§ RBC Chair in Financial Innovations and Département de finance et assurance, Université Laval, CIRANO, and Centre Interuniversitaire sur le risque, les politiques economiques et l'emploi (CIRPEE). Mailing address: Département de finance et assurance, Pavillon Palasis-Prince, Université Laval, Québec, Québec G1K 7P4, Canada. TEL: 1 (418) 656-2926, FAX: 1 (418) 656-2624; e-mail: Marie-Claude.Beaulieu@fas.ulaval.ca

ABSTRACT

In this paper, we propose several finite-sample specification tests for multivariate linear regressions (MLR). We focus on tests for serial dependence and ARCH effects with possibly non-Gaussian errors. The tests are based on properly standardized multivariate residuals to ensure invariance to error covariances. The procedures proposed provide: (i) exact variants of standard multivariate portmanteau tests for serial correlation [Hosking (1980)] as well as ARCH effects [Duchesne and Lalancette (2003)], and (ii) exact versions of the diagnostics presented by Shanken (1990) which are based on combining univariate specification tests. Specifically, we combine tests across equations using a Monte Carlo (MC) test method so Bonferroni-type bounds can be avoided. The procedures considered are evaluated in a simulation experiment: the latter shows that standard asymptotic procedures suffer from serious size problems, while the MC tests suggested display excellent size and power properties, even when the sample size is small relative to the number of equations, with normal or Student-t errors. The tests proposed are applied to the Fama-French three-factor model. Our findings suggest that the *i.i.d.* error assumption provides an acceptable working framework once we allow for non-Gaussian errors within 5 year subperiods, whereas temporal instabilities clearly plague the full-sample data set.

Key words : capital asset pricing model; CAPM; non-normality; multivariate linear regression; exact test; Monte Carlo test; bootstrap; nuisance parameters; multivariate specification test; multivariate diagnostics; ARCH effects; multivariate variance ratio test.

Journal of Economic Literature classification: C3; C12; C33; C15; G1; G12; G14.

Contents

List of Tables

List of Propositions and Theorems

1. Introduction

The multivariate linear regression (MLR) model is one of the most widely used models in statistics, econometrics and finance; see Stewart (1997), Dufour and Khalaf (2002*b*) and the references therein. Well-known financial applications include: (i) tests of market portfolio efficiency in the context of the capital asset pricing model (CAPM) [see Shanken (1986), MacKinlay (1987), Gibbons, Ross and Shanken (1989), Affleck-Graves and McDonald (1989), Zhou (1993), Fama and French (1993, 1995), Shanken (1996), and Beaulieu, Dufour and Khalaf (2007)]; (ii) spanning tests [see De Roon and Nijman (2001) and the references therein]; and (iii) event studies [Binder (1985), Schipper and Thompson (1985)]. This paper focuses on diagnostic procedures designed to check the statistical assumptions underlying such tests. Indeed, a common feature of MLR models consists in assuming that the disturbances in different equations are correlated across equations, but otherwise constitute independent identically distributed (*i.i.d.*) random vectors. Violation of the latter condition can affect the inferences based on the model (such as mean-variance efficiency or spanning tests). This underscores the importance of performing diagnostics *before* implementing the tests.

As emphasized by Kroner and Ng (1998), the existing literature on multivariate diagnostics is sparse compared to the univariate case. Perhaps because of this, diagnostics in empirical MLR-based financial studies – such as Engle's ARCH test (Engle (1982)), the Ljung-Box [Ljung and Box (1978)] and variance ratio [Lo and MacKinlay (1988, 1989)] tests – are often conducted on an equation by equation basis. Although univariate tests can provide some guidance, contemporaneous correlation of disturbances entails that statistics from individual equations are not independent, so combining test decisions over all equations raises joint testing problems; for insights and empirical evidence on the importance of multivariate diagnostic testing in finance, see *e.g.* Richardson and Smith (1993) and Shanken (1990).

In this context, joint diagnostics are typically based either on asymptotic approximations or on Bonferroni-type bounds. The procedures suggested following the first approach involve test statistics which formally incorporate cross-sectional dependence, yet are asymptotically free of nuisance parameters; see Godfrey (1988), Richardson and Smith (1993), and the recent literature on multivariate GARCH which may be traced back to Bollerslev, Engle and Wooldridge (1988) [see Engle and Kroner (1995), Kroner and Ng (1998) and the survey in Bauwens, Laurent and Rombouts (2006)]. Although this may lead to convenient test procedures, including the well known portmanteau serial-correlation test [Hosking (1980)] and its ARCH extensions [Duchesne and Lalancette (2003), Ling and Li (1997)], the fact remains that cross-equation correlations can still affect the null distributions of the test statistics in finite samples. In systems with many equations (*e.g.*, many portfolios), the number of correlations can be quite large relative to the sample size, leading to serious degrees-of-freedom losses and size distortions. As a result, asymptotic approximations perform poorly in finite samples; see Shanken (1996), Campbell, Lo and MacKinlay (1997, Chapter 5), and Dufour and Khalaf (2002*a*, 2002*b*, 2003). Alternatively, Bonferroni-based bound joint tests require one to divide the significance level of each individual test by the number of tests [see Dufour (1990), Shanken (1990), Dufour and Torrès (1998), Dufour and Khalaf (2002*a*)]. While this guards against spurious rejections, it can also lead to severe power losses if the number of equations is large. Despite the above problems, very few finite sample exact specification tests have been proposed for MLR models.[1]

In this paper, we consider the problem of testing the specification of MLR models. We focus on: (1) detecting the presence of ARCH-type heteroskedasticity, and (2) detecting (linear) serial dependence. We propose procedures based on least squares residuals, hence computationally simple. In order to avoid the nuisance parameter problem raised by the unknown error covariance matrix, we apply a multivariate rescaling transformation which eliminates the unknown covariance matrix from the residual distribution. In this way, we get multivariate standardized residuals which are location-scale invariant, hence do not depend on the (unknown) regression coefficients or the error covariance matrix.

The tests against ARCH effects include multivariate extensions of the univariate procedures proposed by Engle (1982) and Lee and King (1993), as well as exact variants of the multivariate procedures studied by Duchesne and Lalancette (2003). The tests for linear serial dependence are multivariate versions of the univariate portmanteau Ljung-Box [Ljung and Box (1978)] and variance ratio [Lo and MacKinlay (1988, 1989)] tests, and exact variants of the multivariate diagnostics proposed by Hosking (1980). All these tests are applied to properly standardized residuals. None of the exact procedures is based on a Bonferroni bound [*i.e.* they do not require one to divide the significance level by the number of equations], with obvious consequences on test power. To overcome multiple-test difficulties as well as the fact that the test statistics have distributions which are difficult to evaluate analytically, we obtain exact test p-values via the Monte Carlo (MC) test technique [Dwass (1957), Barnard (1963), Dufour and Kiviet (1996, 1998), Dufour (2006)]; *i.e.*, the level condition is satisfied for any given sample size, using a finite (possibly small) number of MC replications.

The proposed multivariate procedures also constitute an interesting contribution to the theory of simulation-based testing. We show that the MC technique allows one to use asymptotic p-values in the construction of an exact test, even though these p-values could lead to highly inaccurate inference if used in the conventional way. Indeed, our joint test procedure involves converting all individual tests to an approximate p-value form, in order to combine them (*e.g.*, through their minimum). When the overall procedure is simulated, the fact that approximate or asymptotic distributions are used to obtain the individual p-values does not preclude *exactly* controlling the level of the test.

Our methodology also deals, from a finite-sample perspective, with non-normal errors. Formally, this allows one to test for time varying variances with fat-tailed error distributions, such as the Student-t with possibly unknown degrees of freedom. The latter parameter typically affects the null distribution of the diagnostic test statistic. To control the significance level given such difficulties, we apply a "maximized MC" (MMC) test, where the MC p-value for the tested hypothesis (which depends on the nuisance parameter) is maximized over the relevant nuisance parameter set [Dufour (2006)].

The procedures considered are evaluated in a simulation experiment. Our results reveal that standard multivariate procedures including Bonferroni-based ones suffer from serious size prob-

[1]One exception includes work on testing the independence between the disturbances in different equations [see Dufour and Khalaf (2002*a*)]. But this problem is relatively simple, for the null hypothesis sets the error covariances to zero.

lems. In contrast, our MC and MMC tests display excellent size and power properties, even when the sample size is small relative to the number of equations.

The tests proposed are applied to the Fama-French three-factor model, using monthly data for the period 1965-2000. We analyze the model over the full sample, as well as over 5-year subperiods. Our results reveal temporal instabilities for the full-sample data set. In general however, significant departures from the *i.i.d.* hypothesis are less evident over the subperiods, once we allow for non-Gaussian errors. These results, in view of our simulation study [which illustrates the power of our tests for sample designs compatible with our subperiod analysis], suggest that the *i.i.d.* error assumption provides an acceptable working framework for the Fama-French model, within 5 year subperiods, but not over a longer time span.

The paper is organized as follows. In Section 2, we describe the statistical framework studied and derive the relevant invariance results which underlie our finite-sample testing approach. In Section 3, we present the test criteria considered and the associated testing strategy. The simulation study is reported in Section 4. Section 5 presents our empirical analysis. We conclude in Section 6.

2. Framework and distributional theory

Many asset pricing models take the multivariate regression form

$$Y = XB + U \tag{2.1}$$

where $Y = [y_1, \ldots, y_n]$ is a $T \times n$ matrix of observations on n dependent variables, X is an $T \times k$ full-column rank matrix, B is a $k \times n$ matrix of unknown coefficients, $U = [u_1, \ldots, u_n] = [U_1, \ldots, U_T]'$ is a $T \times n$ matrix of random errors with $u_i = (u_{i1}, \ldots, u_{iT})'$, $i = 1, \ldots, n$. For instance, an s-factor asset pricing model can be written as

$$r_{it} = a_i + \sum_{j=1}^{s} b_{ij}\tilde{r}_{jt} + u_{it},\ t = 1, \ldots, T,\ i = 1, \ldots, n, \tag{2.2}$$

where $r_{it} = R_{it} - R_t^F$, $\tilde{r}_{jt} = \tilde{R}_{jt} - R_t^F$, R_{it}, $i = 1, \ldots, n$, are returns on n portfolios (over period t), R_t^F is the riskless rate of return, $\tilde{R}_{jt}$, $j = 1, \ldots, s$ are returns on s benchmark factors, and u_{it} is a random disturbance. Clearly, this model is a special case of (2.1) where

$$Y = [r_1, \ldots, r_n],\ r_i = (r_{i1}, \ldots, r_{iT})',\ i = 1, \ldots, n, \tag{2.3}$$

$$X = [\iota_T, \tilde{r}_1, \ldots, \tilde{r}_s],\ r_j = (\tilde{r}_{j1}, \ldots, \tilde{r}_{jT})',\ j = 1, \ldots, s, \tag{2.4}$$

$k = s + 1$, ι_T is a vector of ones, and U is the $T \times n$ matrix which includes the errors u_{it}.

We assume we can condition on X, *i.e.* we can take X as fixed for statistical analysis. Furthermore, we restrict the error distribution as follows:

$$U_t = JW_t,\ t = 1, \ldots, T, \tag{2.5}$$

where J is an unknown non-singular lower triangular matrix, and the vector $vec(W_1, \ldots, W_T)$

has a distribution which is either: (i) fully specified, or (ii) specified up to an unknown nuisance parameter ν. Let $W = [w_1, \ldots, w_n] = [W_1, \ldots, W_T]'$, $\quad w_i = (w_{i1}, \ldots, w_{iT})'$ so (2.5) entails that

$$W = U(J^{-1})'. \tag{2.6}$$

This restriction aims to sort-out the following two characteristics of the error distribution: (i) the random term W_t so the joint distribution of $vec\,(W_1, \ldots, W_T)$ gives the fundamental data generating process [DGP]; and (ii) the matrix J which sets the "scale", defined as

$$\Sigma = JJ',$$

i.e. J sets both variance parameters and coefficients representing cross-equation correlations. Special cases of (2.5) [considered in Section 5] include the *i.i.d.* Gaussian assumption:

$$W_1, \ldots, W_T \overset{i.i.d.}{\sim} N[0, I_n], \tag{2.7}$$

and the case where $W_1, \ldots, W_T$ are *i.i.d.* Student,

$$W_1, \ldots, W_T \overset{i.i.d.}{\sim} t(\kappa), \tag{2.8}$$

where the degree-of-freedom parameter κ is either: (i) known [*hence, the fundamental DGP is free of nuisance parameters*], or (ii) unknown and needs to be estimated from the data [κ *is a nuisance parameter*].

The least squares estimate of B is $\hat{B} = (X'X)^{-1}X'Y$ with corresponding residuals

$$\hat{U} = \left[\hat{u}_1, \ldots, \hat{u}_n\right] = \left[\hat{U}_1, \ldots, \hat{U}_T\right], \quad \hat{u}_i = (\hat{u}_{i1}, \ldots, \hat{u}_{iT})'. \tag{2.9}$$

Note that the Gaussian-based quasi maximum likelihood estimators for this model are $\hat{B}$ and

$$\hat{\Sigma} = \frac{1}{T}\hat{U}'\hat{U}. \tag{2.10}$$

The statistics we consider are based on the multivariate standardized residual matrix

$$\tilde{W} = \hat{U}\, S_{\hat{U}}^{-1} \tag{2.11}$$

where $S_{\hat{U}}$ is the Cholesky factor of $T^{-1}\hat{U}'\hat{U}$, *i.e.* $S_{\hat{U}}$ is the (unique) upper triangular matrix such that

$$\hat{\Sigma} = S_{\hat{U}}'S_{\hat{U}}, \quad \hat{\Sigma}^{-1} = \left(\hat{U}'\hat{U}/T\right)^{-1} = S_{\hat{U}}^{-1}\left(S_{\hat{U}}^{-1}\right)'. \tag{2.12}$$

For presentation clarity, we use the following notation: $\tilde{W} = \left[\tilde{w}_1, \ldots, \tilde{w}_n\right] = \left[\tilde{W}_1, \ldots, \tilde{W}_T\right]'$, $\tilde{w}_i = (\tilde{w}_{i1}, \ldots, \tilde{w}_{iT})'$, so (2.11) implies

$$\tilde{W}_t = \left(S_{\hat{U}}^{-1}\right)'\hat{U}_t. \tag{2.13}$$

The validity of our proposed diagnostics relies on the following representation of $\tilde{W}$.

Theorem 2.1 INVARIANCE OF CHOLESKY-STANDARDIZED MULTIVARIATE RESIDUALS. *Under* (2.1), *and for all error distributions compatible with* (2.5), *the standardized residual matrix defined in* (2.11) *satisfies the identity*

$$\tilde{W} = \hat{U}\, S_{\hat{U}}^{-1} = \hat{W}\, S_{\hat{W}}^{-1} \tag{2.14}$$

where $\hat{W} = MW$, $M = I - X(X'X)^{-1}\, X'$ *and* $S_{\hat{W}}$ *is the Cholesky factor of* $T^{-1}\hat{W}'\hat{W}$, *and thus follows a distribution which does not depend on* B *and* J.

The proofs of the theorems appear in Appendix A. Equation (2.14) reexpresses the standardized residual $\tilde{W}$ as a function of $\hat{W} = MW$, which implies that the distribution of $\tilde{W}$ is completely determined by the distribution of W given X. Under assumption (2.5), the distribution of W does not depend on B nor J. For example, under (2.7), $W_1,\ \ldots\ ,\ W_T$ are *i.i.d.* $N[0,\ I_n]$, while under (2.8) the distribution of $W_1,\ \ldots\ ,\ W_T$ is defined by the degrees-of-freedom parameter κ. This entails that B and J (and thus Σ) are simply evacuated from the distribution of $\tilde{W}$. This invariance result holds for all statistics which depend on the data only through $\tilde{W}$, when the MLR is estimated in the (2.1) form.[2]

Theorem **2.1** has crucial implications for diagnostic tests associated with model (2.1)-(2.5) [which becomes, in this case, the null hypothesis]. Indeed, the recent theory of MC test methods [Dufour (2006)] allows to make use of such invariance properties to derive valid test p-values. The MC method is an exact simulation-based procedure which yields an empirical p-value [denoted $\hat{p}_N(\cdot)$] for the considered test statistic, based on the rank of the observed statistic relative to a set of N simulated ones. The latter are drawn imposing the null hypothesis. The MC procedure thus relates to the parametric bootstrap, in the sense that it entails simulating the null distribution of the test statistic.

When the latter simulated distribution does not involve unknown parameters, the MC test method perfectly controls the size of the test for given T and N. For the problem under consideration, this occurs when $vec(W_1,\ \ldots\ ,\ W_T)$ has a fully specified distribution. In this case, exact p-values can be obtained as long as the statistic considered, say $S = S(\hat{U})$, can be rewritten as a function of W and X :

$$S = S(\hat{U}) = \bar{S}\,(W,\ X)\,. \tag{2.15}$$

As in Theorem **2.1**, the latter notation implies that the function $\bar{S}\,(W,\ X)$ evacuates B and J (and thus Σ) out. In view of Theorem **2.1**, this leads to consider statistics – presented below, in Section 3 – which depend on the data only through $\tilde{W}$.

When $vec(W_1,\ \ldots\ ,\ W_T)$ has a distribution which depends on an unknown parameter ν, any relevant [*i.e.* conforming with the null hypothesis] value for ν can lead to an empirical p-value [based on the rank of the observed statistic relative to a set of N simulated ones, drawn given the

[2]This invariance result may not hold if $S_{\hat{U}}'S_{\hat{U}}$ is replaced by another "plausible" factorization of $\hat{\Sigma}$; for example, the appropriate invariance does not occur if the Cholesky factor $S_{\hat{U}}$ is replaced by the usual square root $\hat{\Sigma}^{1/2}$ [for the definition of the square root of a matrix, see Harville (1997, section 21.9)]. This is easy to check numerically.

value of ν in question]; this leads to a p-value "function", denoted $\hat{p}_N(\cdot\,|\,\nu)$. In this case, standard bootstrap methods rely on a consistent point estimate $\hat{\nu}$ of ν which imposes the null hypothesis; the associated approximate p-value [*i.e.* $\hat{p}_N(\cdot\,|\,\hat{\nu})$] would lead, under standard regularity conditions, to an asymptotically [for infinite T and N] valid test. We rather rely on a sup-type MC procedure [introduced by Dufour (2006) and denoted maximized MC (MMC)] which controls the significance level by construction, for finite T and N. The associated critical region corresponds to referring the supremum $\sup_\nu\,[\hat{p}_N(\cdot\,|\,\nu)]$ to a given level α.

We also consider a modified version of the MMC technique [see Dufour (2006) and Dufour and Kiviet (1996)] denoted confidence-set-based MMC (CSMMC) which 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. For an overall α-level, the confidence set and the CSMMC test should be applied with levels $1-\alpha_1$ and $\alpha-\alpha_1$, respectively. Detailed algorithms for all the statistics we consider is provided in Section 3.

There are no theoretical arguments which favor either MMC or CSMMC methods. While the MMC method may appear relatively conservative (since the MC p-value is considered over all values of ν irrespective of the sample information on this parameter), recall that the CSMMC p-value (which, in contrast, uses estimated values of ν) needs to be referred to $\alpha-\alpha_1$. Nevertheless, it is intuitively appealing to consider a CSMMC procedure where the underlying confidence set incorporates information on the goodness-of-fit (GF) of the hypothesized error distribution. In this way, we formally deal with the joint characteristic of the null hypothesis which imposes distributional constraints, in addition to the properties under test (here: no serial correlation, no ARCH effects).

Note finally that the invariance result of Theorem **2.1** holds for multivariate linear models and does not necessarily apply to nonlinear models. Extensions to such models may be feasible – for example, through an exploitation of the MMC method – but this goes beyond the scope of this paper.

3. Multivariate specification tests

In this section, we use the above results to derive multivariate specification tests. The proposed tests are formally valid for any parametric null hypothesis of type (2.5). In Section (5), we focus on assumptions (2.7) and (2.8) with unknown κ.

3.1. Combined equation-by-equation tests

Standard diagnostics may be applied to the residuals of each equation in (2.1). We focus on serial dependence tests based on the popular Ljung-Box [Ljung and Box (1978)] statistic (applied to the i-th equation)

$$LB_i = T(T+2)\sum_{g=1}^{G}\frac{\hat{\rho}_{ig}^2}{T-g}\,, \quad \hat{\rho}_{ig} = \frac{\sum_{t=g+1}^{T}\hat{u}_{it}\hat{u}_{i,t-g}}{\sum_{t=1}^{T}\hat{u}_{it}^2} \tag{3.1}$$

and the variance ratio [Lo and MacKinlay (1988, 1989)] statistic

$$VR_i = 1 + 2\sum_{g=1}^{G}\Big(1 - \frac{g}{G}\Big)\hat{\rho}_{ig}\,, \tag{3.2}$$

where G refers to the maximum number of lags used. We also consider tests for ARCH effects based on Engle-type procedures [Engle (1982), Lee and King (1993)]. The Engle statistic for equation i (denoted E_i), is given by $T\,\times$ (the coefficient of determination in the regression of the equation's squared OLS residuals $\hat{u}_{it}^2$ on a constant and $\hat{u}_{i,t-g}^2$, $g = 1, \ldots, G$). Lee-King's (one-sided) statistic [for equation i] where $\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}^2$ is:

$$LK_i = \frac{\Big\{(T-G)\sum\limits_{t=G+1}^{T}\big[(\hat{u}_{it}^2/\hat{\sigma}_i^2) - 1\big]\sum\limits_{g=1}^{G}\hat{u}_{i,t-g}^2\Big\}\Big/\Big\{\sum\limits_{t=G+1}^{T}[(\hat{u}_{it}^2/\hat{\sigma}_i^2) - 1]^2\Big\}^{1/2}}{\Big\{(T-G)\sum\limits_{t=G+1}^{T}\Big(\sum\limits_{g=1}^{G}\hat{u}_{i,t-g}^2\Big)^2 - \Big(\sum\limits_{t=G+1}^{T}\sum\limits_{g=1}^{G}\hat{u}_{i,t-g}^2\Big)^2\Big\}^{1/2}}. \tag{3.3}$$

In view of Theorem **2.1**, we obtain standardized versions of these test statistics, denoted respectively $\tilde{LB}_i$, $\tilde{VR}_i$, $\tilde{E}_i$ and $\tilde{LK}_i$, replacing $\hat{u}_{it}$ by $\tilde{w}_{it}$ [the elements of the matrix $\tilde{W}$ from (2.11)] in the formula for these statistics. $\tilde{LB}_i$, $\tilde{VR}_i$, $\tilde{E}_i$ and $\tilde{LK}_i$, $i = 1, \ldots, n$, satisfy the conditions of Theorem **2.1** by construction. Hence, under (2.5), their *joint* distribution does not depend on the regression coefficient B nor the scale parameter Σ. We next construct the combined statistics:

$$\tilde{LB} = 1 - \min_{1\le i\le n}\big[p\big(\tilde{LB}_i\big)\big]\,, \qquad \tilde{VR} = 1 - \min_{1\le i\le n}\big[p\big(\tilde{VR}_i\big)\big] \tag{3.4}$$

$$\tilde{E} = 1 - \min_{1\le i\le n}\big[p\big(\tilde{E}_i\big)\big], \quad \tilde{LK} = 1 - \min_{1\le i\le n}\big[p\big(\tilde{LK}_i\big)\big], \tag{3.5}$$

where $p\big(\tilde{VR}_i\big)$, $p\big(\tilde{LB}_i\big)$, $p\big(\tilde{E}_i\big)$ and $p\big(\tilde{LK}_i\big)$ are individual p-values associated with $\tilde{LB}_i$, $\tilde{VR}_i$, $\tilde{E}_i$ and $\tilde{LK}_i$; these may be derived via the MC method, or using approximate null distributions. In Section 5, we use (respectively) the asymptotic distributions: $(VR_i - 1) \overset{asy}{\sim} N[0, 2(2G-1)(G-1)/3G]$, $LB_i \overset{asy}{\sim} \chi^2(G)$, $E_i \overset{asy}{\sim} \chi^2(G)$ and $LK_i \overset{asy}{\sim} N[0,1]$.

While several alternative combination procedures are available,[3] we focus on the form $\big(1 - \min_{1\le i\le n}\big[p(\,\cdot\,)\big]\big)$ which extends Tippett's procedure [Tippett (1931)] to the non-independent tests context. This procedure is intuitively appealing for the following reasons: the combined test rejects the null hypothesis if at least one of the individual (standardized) tests is significant. This is closely related to a Bonferroni-type procedure (as considered *e.g.* by Shanken (1990) in the context of an asset pricing problem similar to the one we study in Section 5), with the following fundamental difference: by the Boole-Bonferroni bound, the joint test is significant at level α, if at least one individual p-value is less than or equal to α/n (α divided by the number of tests). In contrast, we obtain a joint p-value, using the MC test method, for each of the combined statistics, so that such a

[3]See *e.g.* Dufour and Khalaf (2002*a*), Dufour, Khalaf, Bernard and Genest (2004), Dufour and Torrès (1998), Westfall and Young (1993), Savin (1984), Folks (1984) and Dufour, Farhat and Khalaf (2004).

level adjustment is no longer required; this yields obvious power advantages.

The following algorithm summarizes the MC procedure we use. For presentation clarity, we focus on the combined Engle test $\tilde{E}$; of course, the same procedure is applied to all criteria presented so far. Under the *i.i.d.* normal hypothesis (2.7), we proceed as follows.

1. From the observed data, compute the value of $\tilde{E}$ [using (3.5)] and denote it $\tilde{E}^{(0)}$.

2. Obtain N draws from the distribution of W [here (2.7)]; denote the drawn variates $W^{(j)}$, $j = 1, \ldots, N$.

3. For each draw, calculate $\hat{W}^{(j)} = MW^{(j)}$, $S_{\hat{W}}^{(j)}$, the Cholesky factor of $T^{-1}\hat{W}^{(j)\prime}\hat{W}^{(j)}$, and

$$\tilde{W}^{(j)} = \hat{W}^{(j)} \left(S_{\hat{W}}^{(j)}\right)^{-1} = \left[\tilde{w}_1^{(j)}, \ldots, \tilde{w}_n^{(j)}\right] \tag{3.6}$$

where $\tilde{w}_i^{(j)} = \left(\tilde{w}_{i1}^{(j)}, \ldots, \tilde{w}_{iT}^{(j)}\right)'$, $i = 1, \ldots, n$.

4. The simulated Engle criterion for equation i and the MC draw j, which will be denoted $E_i^{(j)}$, obtains as $T \times$ (the coefficient of determination in the regression of the squared $\tilde{w}_{it}^{(j)}$ on their G lags). Compute $\tilde{E}^{(j)} = 1 - \min_{1 \le i \le n} \left[p\left(\tilde{E}_i^{(j)}\right)\right]$, using the same distribution for approximating $p\left(\tilde{E}_i^{(j)}\right)$ – such as the $\chi^2(G)$ – as in step 1.

5. Given $\tilde{E}^{(j)}$, $j = 1, \ldots, N$, compute the number of simulated values greater than or equal to $\tilde{E}^{(0)}$ [denoted $N\hat{G}_N(\tilde{E}^{(0)})$]. The MC p-value is

$$\hat{p}_N\left(\tilde{E}\right) = [N\hat{G}_N\left(\tilde{E}^{(0)}\right) + 1]/(N + 1). \tag{3.7}$$

The null hypothesis is rejected at level α when $\hat{p}_N\left(\tilde{E}\right) \le \alpha$.

Provided $\alpha(N + 1)$ is an integer, the above test procedure has size α (for finite T and N), because $\tilde{E}^{(0)}$, $\tilde{E}^{(1)}$, $\ldots$, $\tilde{E}^{(N)}$ are exchangeable under the null hypothesis; see Dufour (2006). The similarities and differences between our test as described and a naive bootstrap can be explicitly seen from the latter algorithm. Indeed, under the *i.i.d.* normal hypothesis (2.7), a naive parametric bootstrap could be implemented replacing step 3 by the following.

3*. For each draw, $W^{(j)}$, $j = 1, \ldots, N$, and conditional on the observed regressor matrix, the MLR form (2.1), the Cholesky factor $[S_{\hat{U}}]$ of the observed (calculated from the observed data) matrix and the observed OLS estimator $\hat{B}$, reconstruct

$$Y^{(j)} = X\hat{B} + W^{(j)}S_{\hat{U}}, \quad j = 1, \ldots, N.$$

For each j, regress $Y^{(j)}$ on X and obtain the associated residual matrix $\hat{U}^{(j)}$, $\hat{\Sigma}^{(j)} = T^{-1}\hat{U}^{(j)\prime}\hat{U}^{(j)}$ and its associated Cholesky factor $S_{\hat{U}}^{(j)}$, which leads to a series of N simulated standardized residuals $\tilde{W}^{(j)} = \hat{U}^{(j)}\left(S_{\hat{U}}^{(j)}\right)^{-1}$, $\quad j = 1, \ldots, N$.

Now in view of Theorem **2.1**, we see that the latter can be drawn equivalently as described in step 3 (this is also easy to check numerically). So when $vec(W_1, \ldots, W_T)$ has a fully specified distribution [*i.e.*, no unknown parameter needs to be specified to obtain the $W^{(j)}$, $j = 1, \ldots, N$ draws], the MC test method is closely related to the naive parametric bootstrap. Exactness (for finite N and T) under our assumptions requires a p-value function as defined in (3.7) [notice the division by $N+1$] and a choice of N such that $N+1$ is an integer. In contrast, when $vec(W_1, \ldots, W_T)$ has a distribution which depends on an unknown parameter ν, our method differs markedly from the naive bootstrap because we do not use a point estimate of ν to obtain the draws $W^{(j)}$, $j = 1, \ldots, N$. Specifically, for the case of (2.8) where κ is an unknown nuisance parameter, we proceed as follows.

For each acceptable value of κ – we consider integer values ranging from 2 to $T - 2 - n$ (our effective sample size) – applying steps 2-4 above with $W^{(j)}$ according to the Student-*t* distribution [as in (2.8)], leads to a series of empirical p-values we denote $\hat{p}_N\left(\tilde{E}|\kappa\right)$. Clearly, our notation implies that $\hat{p}_N\left(\tilde{E}|\kappa\right)$ defines a MC p-value function [an empirical p-value, as a function of κ]. The MMC procedure involves relying on the maximal p-value, so the MMC critical region for a test with level α is

$$\sup_{\kappa} [\hat{p}_N(\tilde{E}|\kappa)] \leq \alpha.$$

We also consider the CSMMC modification to the latter technique, which involves two stages: (1) an exact confidence set denoted $CS(\kappa)$, with level α_1, is built for κ [the procedure, introduced in Dufour, Khalaf and Beaulieu (2003), which we apply for this purpose is summarized in Appendix B], and (2) the MC p-value $\hat{p}_N(\tilde{E}|\kappa)$ is maximized over all values of κ in the latter confidence set. Because of the pre-estimation stage, if an overall α-level test is desired, then a CSMMC critical region obtains as

$$\sup_{\kappa \in CS(\kappa)} [\hat{p}_N(\tilde{E}|\kappa)] \leq \alpha - \alpha_1.$$

In Section 5, we consider $\alpha_1 = 0.025$. The MMC [or CSMMC] p-values so obtained will be referred to α [or to $\alpha - \alpha_1$] and not to α/n [or to $(\alpha - \alpha_1)/n$]. It is evident that when n is large (we consider $n = 25$ portfolios in Section 5), this leads to sizeable power improvements relative to Bonferroni procedures.[4]

3.2. Multi-equation portmanteau criteria

In addition to the above combined criteria, we propose exact MC tests based on (standardized, when necessary) multi-equation portmanteau statistics. To define these statistics we use the following notation: for a given $T \times n$ matrix $Z = [Z_1, \ldots, Z_T]'$, let

$$C_Z(g) = T^{-1} \sum_{t=g+1}^{T} Z_t Z_{t-g}', \quad g = 0, 1, \ldots, G.$$

[4]The combined procedure presented here remains exact even if approximate individual *p*-values are used in the formulae for the combined statistics. Indeed, our joint test procedure starts by converting all individual test statistics to some *p*-value form, in order to combine them through their minimum; even if the latter "conversion" is performed using asymptotic distributions, the test's global level is controlled *exactly* when the overall procedure is simulated, *i.e.* when the MC technique is applied to the combined statistic.

We consider the serial-dependence statistic of Hosking (1980),

$$HM = T^2 \sum_{g=1}^{G} (T-g)^{-1} \mathrm{tr}\{C_{\hat{U}}(0)^{-1} C_{\hat{U}}(g) C_{\hat{U}}(0)^{-1} C_{\hat{U}}(g)'\} \tag{3.8}$$

and the extension aimed at detecting ARCH effects proposed by Duchesne and Lalancette (2003),

$$HM_2 = T^2 \sum_{g=1}^{G} (T-g)^{-1}\, \mathrm{tr}\{C_{\hat{U}^2}(0)^{-1} C_{\hat{U}^2}(g) C_{\hat{U}^2}(0)^{-1} C_{\hat{U}^2}(g)'\} \tag{3.9}$$

where $\hat{U}^2$ is the matrix of squared residuals.[5] We first observe that HM depends on the data via $\tilde{W}$ only and is thus location-scale invariant.

Theorem 3.1 INVARIANCE OF HOSKING'S STATISTIC. *Under* (2.1), *and for all error distributions compatible with* (2.5), *the Hosking statistic defined in* (3.8) *satisfies the identity*

$$HM = T^2 \sum_{g=1}^{G} (T-g)^{-1} \mathrm{tr}\{C_{\tilde{W}}(0)^{-1} C_{\tilde{W}}(g) C_{\tilde{W}}(0)^{-1} C_{\tilde{W}}(g)'\} \tag{3.10}$$

where $\tilde{W}$ *is the standardized residual matrix defined in* (2.11).

The latter invariance result is not satisfied by the HM_2 statistic. In contrast, the statistic proposed by Ling and Li (1997) and studied by Duchesne and Lalancette (2003) is location-scale invariant.

Theorem 3.2 INVARIANCE OF LING-LI'S STATISTIC. *Under* (2.1), *and for all error distributions compatible with* (2.5), *the Ling-Li ARCH test statistic*

$$\tilde{R} = \sum_{t=G+1}^{T} \big(\hat{U}_t' \hat{\Sigma}^{-1} \hat{U}_t - n\big)\big(\hat{U}_{t-G}' \hat{\Sigma}^{-1} \hat{U}_{t-G} - n\big) \Big/ \sum_{t=1}^{T} \big(\hat{U}_t' \hat{\Sigma}^{-1} \hat{U}_t - n\big)^2$$

satisfies the identity

$$\tilde{R} = \sum_{t=G+1}^{T} \big(\tilde{W}_t' \tilde{W}_t - n\big)\big(\tilde{W}_{t-G}' \tilde{W}_{t-G} - n\big) \Big/ \sum_{t=1}^{T} \big(\tilde{W}_t' \tilde{W}_t - n\big)^2 \tag{3.11}$$

where $\tilde{W} = [W_1, \ldots, W_T]'$ *is the standardized residual matrix defined in* (2.11).

The above invariance results obtain because residuals are standardized before they are squared; in contrast, to obtain HM_2, residuals are first squared then standardized. We thus propose the

[5]Note that Duchesne and Lalancette (2003) proposed to consider squares and cross-products of residuals. In view of our small sample [relative to the number of equations], the latter test is not always feasible, so we focus on squares only. Our exact approach can be extended (allowing for a large enough sample) to account for squares and cross-products of residuals.

following modification to the HM_2 statistic which consists in standardizing residuals before taking their squares:

$$\tilde{HS}_2 = T^2 \sum_{g=1}^{G} (T-g)^{-1} \mathrm{tr}\left\{ C_{\tilde{W}^2}^{-1}(0) C_{\tilde{W}^2}(g) C_{\tilde{W}^2}^{-1}(0) C'_{\tilde{W}^2}(g) \right\} \tag{3.12}$$

where $\tilde{W}^2$ is the matrix of squared residuals. The following standard asymptotic null distributions hold: $HM \overset{asy}{\sim} \chi^2(n^2 G)$, $HM_2 \overset{asy}{\sim} \chi^2(n^2 G)$ and $\tilde{R} \overset{asy}{\sim} \chi^2(G)$. We obtain exact MC versions of the latter tests, by applying the algorithm presented in Section 3.1, using the pivotal representations (3.10), (3.11) and (3.12) for the observed (step 1) and simulated (step 4) statistics.

4. Simulation study

We now present a small-scale simulation experiment to assess the performance of the proposed tests. The model considered is (2.1) with $T = 60$, $n = 12$, 20 or 40 equations, where the regressor matrix includes a constant and a standard normal variate (drawn only once). The sample size was fixed to match our empirical application reported in the next section. The tests are implemented with 2 and 12 lags. In all designs, $N = 999$ replications are used to implement MC tests, and the number of simulations in each experiment is 1000. Because of location-scale invariance, all tests are applied to the residuals generated as $\hat{U} = MW$, hence there is no need to specify values for the regression coefficients and error covariances. We study normal and t-errors with unknown degrees-of-freedom, so the rows of W are generated respectively as in (2.7) and (2.8). We set $\kappa = 5$ to draw the "observed" samples, but the tests were applied ignoring this information: formally, κ is considered unknown and the MMC test method is applied over the space $2 \leq \kappa \leq 10$.[6] To study the power of the tests considered, we introduce, in turn, ARCH(1), GARCH(1,1) and AR(1) and AR(2) effects in the first $m = n/3$, $n/2$ and $3n/4$ equations. This is done as follows; first, the W matrix is drawn, conforming with either (2.7) or (2.8); following our notational framework, if we denote by w_{it} the elements of the latter matrix, then

$$u_{it} = w_{it} h_{it}^{\frac{1}{2}},\ h_{it} = 1 + \left(\delta_1 w_{i,t-1}^2 + \delta_2\right) h_{i,t-1}\,,\ i = 1, \ldots, m,\ t = 1, \ldots, T, \tag{4.13}$$

give the errors of the m equations with ARCH. We consider: (1) $\delta_1 = \delta_2 = 0$ [the null hypothesis]; (2) $\delta_1 = 0.4$, $\delta_2 = 0$; (3) $\delta_1 = 0.9$, $\delta_2 = 0$; (5) $\delta_1 = 0.4$, $\delta_2 = 0.5$; (6) $\delta_1 = 0.25$, $\delta_2 = 0.65$. Following the same notation,

$$u_{it} = \rho_1 u_{i,t-1} + \rho_2 u_{i,t-2} + w_{it},\ i = 1, \ldots, m,\ t = 1, \ldots, T$$

give the errors of the m equations with serial correlation (with $u_{i,-1} = u_{i,0} = 0$). We consider: (a) $\rho_1 = \rho_2 = 0$ [the null hypothesis]; (b) $\rho_1 = 0.5$, $\rho_2 = 0$; (c) $\rho_1 = 0.9$, $\rho_2 = 0$; (d) $\rho_1 = 0.5$, $\rho_2 = 0.2$; and (e) $\rho_1 = 0.1$, $\rho_2 = -0.2$. For all configurations, the tests are applied with $G = 2$ or 12

[6] A wider range was allowed in our empirical application; in the case of the MC study, this restriction was adopted to keep execution time within manageable ease: the MMC test has to be applied 1000 times, for all chosen designs.

Table 1. Size of diagnostic tests

n	G	$\tilde{E}$	$\tilde{E}_B$	$\tilde{LK}$	$\tilde{LK}_B$	$\tilde{LB}$	$\tilde{LB}_B$	$\tilde{VR}$	$\tilde{VR}_B$	HM	$\tilde{HM}$	HM_2	$\tilde{HS}_2$
		Normal errors											
12	2	.034	.024	.048	.037	.052	.049	.057	.042	.032	.043	.054	.052
20		.050	.038	.055	.039	.040	.041	.044	.028	.025	.048	.042	.041
40		.048	.044	.051	.042	.056	.055	.054	.031	.003	.050	.002	.056
12	12	.050	.000	.036	.041	.056	.159	.040	.026	.117	.047	.153	.044
20		.052	.003	.051	.057	.052	.184	.045	.043	.179	.056	.243	.050
40		.048	.004	.064	.074	.052	.259	.049	.065	.551	.050	.485	.053
		Student-t errors											
12	2	.035	.057	.041	.037	.044	.049	.040	.032	.039	.013	.126	.014
20		.022	.048	.038	.034	.032	.042	.042	.033	.034	.012	.106	.009
40		.035	.062	.040	.038	.034	.041	.041	.025	.007	.019	.044	.025
12	12	.014	.017	.038	.034	.035	.132	.035	.035	.130	.017	.153	.025
20		.014	.013	.034	.036	.036	.172	.039	.043	.238	.020	.272	.027
40		.017	.009	.039	.044	.037	.210	.046	.067	.499	.019	.418	.021

Note – Numbers shown are empirical rejections for 5% nominal significance test levels when errors are *i.i.d.*, n is the number of equations in the system, T is the sample size and G is the number of lags used for each test. MC tests with $t(\kappa)$ errors are MMC tests, maximized over $2 \leq \kappa \leq 10$. $\tilde{E}$ and $\tilde{LK}$ refer to our generalized Engle and Lee-King joint tests defined in (3.5); $\tilde{E}_B$ and $\tilde{LK}_B$ are (respectively) their Bonferroni counterparts based on referring the minimum p-value to a $(5/n)\%$ level. $\tilde{LB}$ and $\tilde{VR}$ refer to our generalized Ljung-Box and Variance ratio joint tests defined in (3.4); $\tilde{LK}_B$ and $\tilde{VR}_B$ are their Bonferroni counterparts. HM denotes Hosking's multivariate asymptotic portmanteau test defined in (3.8), and $\tilde{HM}$ is its MC counterpart. HM_2 is the Hosking-type multivariate asymptotic ARCH test criterion defined in (3.9), and $\tilde{HS}_2$ is its standardized MC counterpart defined in (3.12).

lags. Results are reported in tables 1-3. We report empirical rejections (over the 1000 replications) for a nominal level of 5%.[7]

Results on test sizes (reported in Table 1) can be summarized as follows. Bonferroni-type tests can over-reject; this occurred in particular with the Ljung-Box combined test using 12 lags, even with normal errors. On recalling that we have relied on asymptotic individual equation p-values to derive the Bonferroni p-values, this result is driven by the poor performance of the individual equation tests. In other words, despite the important level correction required here (a division by n, where n = 12, 20 and 40), the Bonferroni procedure is not exact and remains unreliable.

The size of the asymptotic Hosking type tests can deviate arbitrarily from the nominal one; size distortions increase with the number of equations (n) and the number of lags used in the tests (G). Over-rejections can be very severe: empirical sizes nearing 50% (for a nominal level of 5%) were observed for large n and G. This observation is worth noting, given that available simulation studies

[7]For space considerations, since we find (as also observed by Duchesne and Lalancette (2003)) that the Ling-Lee statistic is dominated by the Hosking-type variant, we report results on HM_2 and its MC exact counterpart $\tilde{HS}_2$.

Table 2. Power of ARCH tests

Normal			$n=20$, $G=2$			$n=20$, $G=12$			$n=40$, $G=2$			$n=40$, $G=12$		
m	δ_1	δ_2	$\tilde{E}$	$\tilde{LK}$	$\tilde{HS}_2$	$\tilde{E}$	$\tilde{LK}$	$\tilde{HS}_2$	$\tilde{E}$	$\tilde{LK}$	$\tilde{HS}_2$	$\tilde{E}$	$\tilde{LK}$	$\tilde{HS}_2$
5	.40	0	.528	.503	.084	.173	.089	.066	.625	.569	.093	.216	.093	.062
	.90	0	.906	.886	.196	.510	.157	.144	.961	.954	.105	.671	.229	.096
	.25	.65	.422	.549	.083	.239	.334	.083	.479	.637	.072	.287	.374	.063
	.40	.50	.638	.761	.114	.366	.341	.108	.747	.867	.080	.442	.386	.680
10	.40	0	.711	.690	.167	.276	.126	.124	.776	.704	.109	.258	.118	.090
	.90	0	.986	.986	.472	.734	.262	.428	.996	.992	.260	.799	.313	.260
	.25	.65	.601	.748	.140	.353	.530	.197	.608	.766	.084	.368	.510	.085
	.40	.50	.841	.913	.231	.548	.521	.272	.879	.949	.110	.599	.531	.108
15	.40	0	.808	.791	.239	.316	.154	.255	.793	.765	.132	.295	.142	.105
	.90	0	.992	.994	.706	.813	.331	.735	.996	.998	.430	.860	.410	.373
	.25	.65	.677	.826	.267	.412	.599	.398	.659	.818	.094	.427	.546	.145
	.40	.50	.900	.973	.412	.607	.601	.529	.904	.972	.146	.645	.601	.189
Student			$n=20$, $G=2$			$n=20$, $G=12$			$n=40$, $G=2$			$n=40$, $G=12$		
5	.40	0	.552	.543	.076	.126	.086	.057	.689	.661	.059	.163	.099	.040
	.90	0	.853	.866	.137	.341	.175	.092	.949	.943	.103	.517	.231	.059
	.25	.65	.432	.617	.040	.264	.423	.051	.536	.730	.035	.348	.502	.034
	.40	.50	.623	.808	.061	.341	.421	.063	.775	.900	.047	.466	.502	.039
10	.40	0	.769	.782	.181	.181	.130	.097	.842	.828	.128	.212	.140	.070
	.90	0	.971	.977	.414	.541	.283	.279	.995	.997	.271	.666	.354	.121
	.25	.65	.615	.820	.133	.343	.653	.156	.710	.872	.064	.415	.655	.045
	.40	.50	.831	.948	.201	.482	.626	.220	.921	.976	.103	.596	.665	.062
15	.40	0	.874	.868	.367	.189	.156	.190	.793	.765	.193	.213	.175	.082
	.90	0	.993	.999	.760	.608	.379	.493	.996	.998	.551	.697	.418	.300
	.25	.65	.732	.902	.361	.404	.756	.386	.734	.899	.197	.452	.739	.156
	.40	.50	.907	.981	.505	.573	.736	.477	.916	.985	.282	.630	.746	.194

Note – Numbers shown are empirical rejections for 5% nominal significance test levels when errors are not *i.i.d.*, n is the number of equations in the system, T is the sample size and G is the number of lags used for each test. MC tests with Student $t(\kappa)$ errors are MMC tests, maximized over $2 \leq \kappa \leq 10$. ARCH effects are introduced in the first $m = n/3$, $n/2$ and $3n/4$ equations. δ_1 and δ_2 are the parameters of the ARCH process postulated for all equations. $\tilde{E}$ and $\tilde{LK}$ refer to the joint tests defined in (3.5). $\tilde{HS}_2$ is the standardized multivariate criterion defined in (3.12).

Table 3. Power of serial-correlation tests

Normal			$n=20$, $G=2$			$n=20$, $G=12$			$n=40$, $G=2$			$n=40$, $G=12$		
m	ρ_1	ρ_2	$\tilde{LB}$	$\tilde{VR}$	$\tilde{HM}$	$\tilde{LB}$	$\tilde{VR}$	$\tilde{HM}$	$\tilde{LB}$	$\tilde{VR}$	$\tilde{HM}$	$\tilde{LB}$	$\tilde{VR}$	$\tilde{HM}$
5	.5	0	.989	.995	.347	.730	.789	.241	.995	.998	.183	.736	.821	.112
	.9	0	1.00	1.00	.999	1.00	1.00	.973	1.00	1.00	.736	1.00	1.00	.475
	.5	.2	1.00	1.00	.692	.986	.996	.579	1.00	1.00	.292	.994	.999	.165
	.1	-.2	.121	.035	.076	.091	.041	.074	.145	.060	.074	.075	.053	.066
10	.5	0	1.00	.999	.823	.826	.907	.859	.998	.998	.557	.752	.883	.557
	.9	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.997	1.00	1.00	.985
	.5	.2	1.00	1.00	.984	.997	.999	.991	1.00	1.00	.718	.993	1.00	.588
	.1	-.2	.204	.053	.140	.115	.035	.152	.201	.077	.113	.120	.040	.091
15	.5	0	.999	1.00	.985	.847	.935	1.00	.999	1.00	.913	.779	.881	.867
	.9	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	.5	.2	1.00	1.00	1.00	.999	1.00	1.00	1.00	1.00	.942	.994	1.00	.915
	.1	-.2	.204	.047	.209	.144	.030	.338	.203	.078	.166	.124	.046	.180
Student			$n=20$, $G=2$			$n=20$, $G=2$			$n=40$, $G=2$			$n=40$, $G=12$		
5	.5	0	.983	.992	.154	.691	.753	.070	.994	.999	.095	.674	.798	.036
	.9	0	1.00	1.00	.976	1.00	1.00	.819	1.00	1.00	.516	1.00	1.00	.117
	.5	.2	1.00	.999	.427	.980	.995	.227	1.00	1.00	.172	.989	.998	.050
	.1	-.2	.119	.044	.024	.095	.039	.020	.120	.055	.033	.074	.038	.017
10	.5	0	.998	1.00	.590	.768	.903	.461	.997	.999	.321	.719	.842	.104
	.9	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.983	1.00	1.00	.644
	.5	.2	1.00	1.00	.939	.994	1.00	.259	1.00	1.00	.502	.988	1.00	.164
	.1	-.2	.172	.036	.038	.101	.029	.042	.131	.074	.051	.104	.042	.022
15	.5	0	.998	1.00	.925	.832	.929	.955	.995	.999	.735	.832	.929	.327
	.9	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.975
	.5	.2	1.00	1.00	.998	.998	1.00	1.00	1.00	1.00	.798	.988	.999	.372
	.1	-.2	.186	.043	.071	.106	.022	.100	.159	.072	.081	.101	.032	.037

Note – Numbers shown are empirical rejections for 5% nominal significance test levels, n is the number of equations in the system, T is the sample size and G is the number of lags used for each test. MC tests with Student $t(\kappa)$ errors are MMC tests, maximized over $2 \leq \kappa \leq 10$. AR effects are introduced in the first $m = n/3$, $n/2$ and $3n/4$ equations. ρ_1 and ρ_2 are the parameters of the AR process postulated for all equations. $\tilde{LB}$ and $\tilde{VR}$ refer to the joint tests defined in (3.4). $\tilde{HM}$ refers to the MC test based on the multivariate criterion defined in (3.8).

from the statistics literature typically consider a small number of equations (relative to the sample size). In the finance literature, one often relies on many portfolios with monthly data over five to ten years, which leads to a large number of equations (relative to the sample size).

The MC test procedure achieves level control. Size is controlled perfectly with normal errors (as expected, because the tests are nuisance parameter free in this case). In interpreting the empirical size in the case of Student$-t$ errors, recall that the maximized MC procedure satisfies the level condition, but its size may be lower than its level. We note some conservative performance, yet of course this question must be evaluated under the alternative hypothesis, where one may check whether under-rejections under the null hypothesis effectively translate into power problems.

We next turn to discussing the results of our power study (reported in tables 2 and 3). Our discussion only focuses on the level correct procedures. Results reveal the following.

Consider the ARCH experiment (Table 2). Over all configurations considered, the procedures which combine via the MC approach individual-equation criteria outperform by far their Hosking-type multivariate counterpart. As with all simulation studies, results may relate to the alternative considered. Here we have not considered time varying cross-correlations, yet the Hosking-type procedure are applied conformably (see Section 3.2). In the literature on combined tests [see *e.g.* Miller (1981), Folks (1984), Savin (1984), Dufour (1989), Dufour and Torrès (1998) and Dufour, Farhat and Khalaf (2004)], min-p-value-type (sup-type) procedures are often proposed as promising alternatives to their portmanteau (sum-type) counterparts, however available results on their relative merits are typically non-conclusive. Our MC procedures control the size (at least in the case of the normal distribution) of both procedures, or more explicitly, correct for their conservative character which stems from test combination, so we are able to compare their effective power reliably.[8] Our results provide an interesting case where one procedure improves on the other.

Lee-King type statistics outperform their Engle-type counterpart in the presence of GARCH, whereas Engle-type criteria appear superior under the ARCH alternative. For criteria based on either Engle or Lee-King statistics, using more lags (12 relative to 2) reduces power for both ARCH and GARCH directions. Power losses go, to some extent, against expectations in the latter case; note however that using more lags reduces the effective sample size, which seems to translate into important power losses. In a few cases however, particularly for large m (the number of equations with ARCH effects), Hosking-type procedures perform somewhat better with more lags, yet these tests (irrespective of the lags used) remain largely dominated by our combined univariate ones. In general, as expected, power improves with m.

Results of the AR experiment, concur, in many aspects with the ARCH case, except that the Hosking statistic performs somewhat better than its ARCH counterpart. Overall, the variance ratio based criterion displays the best performance, except for the AR(2) alternative with $\rho_1 = .1$ and $\rho_2 = -.2$, where either the Ljung-Box based or the Hosking criterion performs better. Lags typically cost power in this experiment as well, except in a few cases with large m (as with the ARCH case).

For both the AR and (G)ARCH experiments, results with Student-t errors show excellent power relative to the normal case. In some cases, the tests even seem to perform better with Student-t

[8] In the case of the Student-t distribution which may remain conservative, size issues are not due to combinations; rather, they relate to the degrees-of-freedom nuisance parameter.

errors. Our results also demonstrates the usefulness of the MMC procedure.

5. Application to Fama-French three-factor model

Our empirical analysis focuses on the asset pricing model (2.2) with $n = 25$, $s = 3$ and different distributional assumptions for stock market returns. The factors considered include: the return on the market portfolio, the average return on three small portfolios minus the average return on three large portfolios (SMB), and the average return on two value portfolios minus the average returns on two growth portfolios (HML).

We use Fama and French's data base, on monthly returns of 25 value-weighted portfolios from 1961-2000. The portfolios which are constructed at the end of June, are the intersections of five portfolios formed on size (market equity) and five portfolios formed on the ratio of book equity to market equity. The size breakpoints for year τ are the New York Stock Exchange (NYSE) market equity quintiles at the end of June of year τ. The ratio of book equity to market equity for June of year τ is the book equity for the last fiscal year end in $\tau - 1$ divided by market equity for December of year $\tau - 1$. The ratio of book equity to market equity are NYSE quintiles. The portfolios for July of year τ to June of year $\tau + 1$ include all NYSE, AMEX, NASDAQ stocks for which market equity data is available for December of year $\tau - 1$ and June of year τ, and (positive) book equity data for $\tau - 1$. The benchmark factors are defined as follows: (1) the excess return on the market is the value-weighted return on all NYSE, AMEX, and NASDAQ stocks [from the University of Chicago's Center for Research in Security Prices (CRSP)] minus the one-month Treasury bill rate [from Ibbotson Associates]; (2) SMB is the average return on three small portfolios minus the average return on three big portfolios, and (3) HML is the average return on two value portfolios minus the average return on two growth portfolios. Fama and French benchmark factors, SMB and HML, are constructed from six size/book-to-market benchmark portfolios that do not include ranges and do not incur transaction costs. The portfolios for these factors are rebalanced quarterly using two independent sorts, on size (market equity, ME) and book-to-market (the ratio of book equity to market equity, BE/ME). The size breakpoint (which determines the buy range for the small and big portfolios) is the median NYSE market equity. The BE/ME breakpoints (which determine the buy range for the growth, neutral, and value portfolios) are the 30th and 70th NYSE percentiles.

Results are reported in tables 4–5. We report, in addition to the asymptotic p-values (denoted p_∞) when available, the Bonferroni p-values (denoted $\hat{p}_B$) for each test combined, obtained as the minimum p-value over all equations, and three MC p-values: (i) the Gaussian based MC p-value (denoted $\hat{p}_g$), (ii) the Student-t based MMC p-value (maximized over all relevant degrees of freedom $\kappa \geq 2$) (denoted $\hat{p}_a$), and (ii) the Student-t based CSMMC (denoted $\hat{p}_i$) (where the maximization is restricted to the degrees of freedom not rejected by a prior goodness-of-fit test; see Appendix B); the associated confidence sets are reported in the last column of each Table. The tests use $G = 12$ lags. All MC procedures are implemented with $N = 999$ replications and the confidence set underlying the CSMMC procedure is applied with $\alpha_1 = 2.5\%$. Note that in the context of a MC test with 999 replications, the smallest possible p-values are .001, .002 and so on and so forth. A discussion of our results follows.

Recall that 2.5% must be added to the CSMMC p-values (denoted $\hat{p}_i$) since 2.5% was used to

Table 4. Univariate and multivariate ARCH tests

	Engle				Lee-King				Hosking-ARCH				κ
	$\hat{p}_B$	$\hat{p}_g$	$\hat{p}_i$	$\hat{p}_a$	$\hat{p}_B$	$\hat{p}_g$	$\hat{p}_i$	$\hat{p}_a$	p_∞	$\hat{p}_g$	$\hat{p}_i$	$\hat{p}_a$	
61-65	.110	.576	.590	.591	.079	.888	.879	.879	.879	1.00	1.00	1.00	8-34
66-70	.179	.832	.848	.848	.017	.399	.385	.385	.125	.655	.638	.638	13-34
71-75	.056	.239	.280	.298	.018	.390	.380	.380	.580	.910	.904	.904	10-34
76-80	.009	.012	.022	.101	.001	.026	.032	.032	1.00	.999	1.00	1.00	13-34
81-85	.031	.098	.132	.197	.035	.650	.646	.586	.832	.758	.733	.733	8-34
86-90	.191	.853	.868	.879	.009	.298	.315	.315	.392	.402	.403	.403	23-34
91-95	.056	.242	.261	.339	.001	.080	.084	.084	.429	.537	.544	.544	16-34
96-00	.003	.002	.014	.021	.002	.061	.070	.071	.000	.666	.634	.666	4-15
ALL	.000	.001	.001	.012	.000	.001	.001	.001	.000	.001	.001	.005	6-10

Note – Numbers shown are p-values, except for the last column which reports the confidence set for κ, where κ denotes the degrees-of-freedom parameter of the hypothesized Student-t distribution (the method for constructing this confidence set is presented in Appendix B). p_∞ refers to the test asymptotic p-value and $\hat{p}_g$ is the Gaussian based MC p-value. $\hat{p}_i$ is CSMMC p-value imposing $t(\kappa)$ errors and $\hat{p}_a$ is the MMC p-value over all degrees of freedom. $\hat{p}_B$ is the minimum p-value for each individual equation test over all equations. The individual equation test statistics are: Engle's TR^2 and Lee-King's statistic (3.3) defined in Section 3.1; their joint counterparts are defined in (3.5). Hosking-ARCH is the multivariate criterion defined in (3.9) [to obtain p_∞] and (3.12) [to obtain the MC p-values]. To obtain an α-level test, $\hat{p}_B$ and $\hat{p}_i$ as defined need to be referred (respectively) to the cut-off levels of $\alpha/25$ and $\alpha - .025$. The tests use $G = 12$ lags.

Table 5. Univariate and multivariate serial-correlation tests

	Ljung-Box				Variance ratio				Hosking				κ
	$\hat{p}_B$	$\hat{p}_g$	$\hat{p}_i$	$\hat{p}_a$	$\hat{p}_B$	$\hat{p}_g$	$\hat{p}_i$	$\hat{p}_a$	p_∞	$\hat{p}_g$	$\hat{p}_i$	$\hat{p}_a$	
61-65	.003	.291	.312	.312	.384	.963	.967	.974	.711	.973	.957	.957	8-34
66-70	.000	.106	.119	.127	.347	.866	.881	.881	.354	.818	.786	.786	13-34
71-75	.022	.619	.632	.632	.177	.285	.292	.292	.795	.956	.941	.941	10-34
76-80	.014	.546	.565	.570	.132	.193	.206	.206	.978	1.00	1.00	1.00	13-34
81-85	.010	.454	.471	.477	.099	.159	.167	.167	.205	.497	.493	.493	8-34
86-90	.001	.135	.145	.151	.182	.240	.241	.241	.619	.978	.964	.964	23-34
91-95	.000	.040	.049	.049	.070	.114	.122	.124	.432	.796	.774	.774	16-34
96-00	.013	.529	.541	.542	.093	.142	.156	.156	.018	.088	.183	.185	4-15
ALL	.000	.001	.003	.004	.182	.001	.001	.001	.000	.001	.001	.001	6-10

Note – Numbers shown are p-values, except for the last column which reports the confidence set for κ , where κ denotes the degrees-of-freedom parameter of the hypothesized Student-t distribution (the method for constructing the underlying confidence set for the degrees of freedom κ is presented in Appendix B). p_∞ refers to the test asymptotic p-value and $\hat{p}_g$ is the Gaussian based MC p-value. $\hat{p}_i$ is CSMMC p-value imposing $t(\kappa)$ errors. $\hat{p}_a$ is the MMC p-value over all degrees of freedom. $\hat{p}_B$ is the minimum p-value for each individual equation test over all equations. The individual equation test statistics are: the variance ratio (3.2) and the Ljung-Box criteria (3.1); the joint tests are defined in (3.4). Hosking is the multivariate criterion defined in (3.8). To obtain an α -level test, $\hat{p}_B$ and $\hat{p}_i$ as defined need to be referred (respectively) to the cut-off levels of $\alpha/25$ and $\alpha - .025$. The tests use $G = 12$ lags.

construct the underlying confidence set; in other words, if an overall level of 5% is desired, the cut-off level for $\hat{p}_i$ is $5.0 - 2.5 = 2.5\%$. Furthermore, for a joint Bonferroni type procedure with a level of $\alpha = 5\%$, $\hat{p}_B$ needs to be referred to a level of $\alpha/25 = 0.2\%$ [the system includes 25 equations here]. The joint p-values $\hat{p}_g$, $\hat{p}_i$ or $\hat{p}_a$ do not require any level adjustment. To illustrate the implications of level corrections, consider for example the case of tests based on Engle's statistic in the subperiod 1996-00: from a Bonferroni perspective (for a joint level of 5%), the joint test is not significant (the p-value is $.003 > .05/25$); however, all MC p-values are less than their relevant cutoffs which suggests significant ARCH effects. The same observation holds for: (i) the full sample tests based on the variance ratio statistic, and (ii) the tests based on Engle's statistic in the subperiod 1976-80; in this case however, the MMC test is not significant at 5% yet the CSMMC p-value is $.022 < .025$ so the CSMMC test is significant.

In view of our simulation results (Section 4), outcomes of the Bonferroni and the Hosking-type asymptotic procedures must be qualified because rejections may be spurious. Indeed, on comparing the Bonferroni to the MC p-values, we observe that rejection decisions are reversed in several instances, for example: (i) the tests based on Lee-King's statistic, in the subperiods 1991-95 and 1996-00; and (ii) the tests based on Ljung-Box statistics in the subperiods 1966-70, 1986-90 and 1991-95. In all these cases, the Bonferroni p-value is significant at 5% whereas even the Gaussian MC test is not significant at this level. A further decision reversal also deserves notice, namely the joint variance ratio case over the whole sample. In this case, although all MC tests are highly significant (the p-values range are .001), their Bonferroni counterpart is not significant at conventional levels (the p-values is .182). In the case of Hosking-type tests, on comparing the asymptotic and MC p-values, we also find that rejection decisions are reversed in several instances. Consider for example: (i) the Hosking-ARCH test in the 1996-00 subperiods, and in the latter subperiod; and (ii) the serial-correlation Hosking test in the 1996-00 subperiod. In all these cases the asymptotic tests are significant at 5% whereas the MC p-values exceed their relevant cutoffs even with normal errors.

The MMC and CSMMC test approaches lead, in a few cases, to conflicting decisions. For example, refer to the joint Engle-type test in 1976-80, where $\hat{p}_i = .022$ while $\hat{p}_a = .101$. In this case, if degrees of freedom which are not compatible with the data are allowed, ARCH effects may end up undetected. In contrast, consider: (i) the joint Lee-King-type test in 1976-80, where $\hat{p}_i = \hat{p}_a = .032$, and (ii) the joint Ljung-Box-type test in 1991-95, where $\hat{p}_i = \hat{p}_a = .049$. Here, on recalling that a level adjustment is required in the case of $\hat{p}_i$ [the cut-off level for an overall level of α is $\alpha - 2.5$], we see that the pre-estimation step may lead to power costs.[9]

Over the whole sample period, all tests are significant at 5%, except the Bonferroni Variance-Ratio test. In contrast, the various subperiod tests applied yield conflicting decisions. Indeed, the Hosking-type MC tests are all not significant at the 5% level, whereas a few rejections are noted at this level using our combined tests. Examples include: (i) the ARCH test based on Engle's statistic (allowing for normal or t-errors) in the 1976-80 and 1996-00 subperiods; (ii) the ARCH test based on Lee-King's statistic (allowing for normal or t-errors) in the 1976-80 subperiod; (iii)

[9]An exact multi-stage MC joint test which integrates the estimation and testing steps in conceptually feasible. The test would however involve several nested simulations; whether computational burdens translate into consequential power advantages is an open question, which is beyond the scope of this paper. On multi-stage MC tests, see Dufour et al. (2003).

the serial-correlation test based on Ljung-Box's statistic (allowing for normal or t-errors) in the 1991-95 subperiod. Nevertheless, in all cases where Hosking-type MC tests are significant at 5% [which occurs only with tests applied over the full sample], our combined tests are also significant at the same level. These findings are in line with our simulation results (reported in Section 4).

Departures from the *i.i.d.* hypothesis are less evident with non-Gaussian errors. Indeed, the Gaussian MC p-values are typically lower than the Student-t based ones. Overall, while the full sample MC tests are all significant at usual levels, the subperiod diagnostics do not detect serious deviations from the *i.i.d.* assumption, particularly if Student-t error distributions are formally accounted for. These results may suggest that serial dependence is negligible in the short run and important over the long run so that temporal dependence factors are slow moving, which is empirically intriguing. Such an interpretation may thus cast doubt on the tests' usefulness in modeling the short run dynamics of conditional return distributions.

Skepticism about test power with sub-period data must be weighed against our simulation results which reveal that all tests perform well with samples of 60 observations on as many as 40 equations. Admittedly, as with all simulation studies, results may relate to the experimental design considered and power issues are not necessarily ruled out. Nevertheless, our tests, as with most diagnostics conducted on regression error distributions, presume stable regression coefficients and, for that matter, a constant degrees-of-freedom parameter, over the test period. This hypothesis may not hold over the long term and calls for caution in confronting our sub-period to the full-sample test outcomes. On balance, our results suggest the following strategies for empirical asset pricing practice. While a regression of the form (2.2) with Fama-French factors and given (2.8) seems acceptable as a working framework within subperiods, the underlying risk-return relationship may be unstable over long time spans. Controlling for the long run dynamics of conditional distributions matters importantly, yet asset pricing tests with long spans of monthly returns require searching for stable factor structures.

6. Conclusion

Previous research typically assess MLR-based asset pricing statistical models using tests based on individual equations. Due to error cross-correlations, statistics from individual equations are not independent, which raises simultaneous test problems. In this paper, we consider a diagnostic test procedure that accounts for cross-equation correlations exactly, in possibly non-normal contexts. We consider tests for serial correlation and tests for ARCH effects. The procedures proposed provide exact variants of the standard multivariate portmanteau tests as well as exact diagnostics which consist in combining univariate specification tests. Our tests are invariant to MLR coefficients and error covariances; since in typical financial models, the covariance matrix is high dimensional, invariance to these nuisance parameters is a very useful property; with non-Gaussian errors, dependence on further unknown parameters is circumvented by applying MMC test techniques. From a theoretical perspective, our multivariate procedures illustrate the usefulness of the MC test procedure in combining non-independent tests exactly. Interestingly, we show that even if individual p-values are obtained using asymptotic arguments, they may be combined in a way which yields a joint exact test.

The procedures considered are evaluated via a simulation experiment, with sample sizes matching our empirical analysis. Our results reveal that available procedures including Bonferroni-based ones suffer from serious size problems. In contrast, our MC and MMC tests display excellent size and power properties. We find that combining individual equation criteria (after standardizing residuals) outperforms portmanteau approaches.

The tests proposed are applied to the Fama-French three-factor model, using monthly data. We analyze the model over the full sample [1965-2000] and over 5-year subperiods. Our results indicate significant instabilities for the full-sample case, although significant departures from the *i.i.d.* hypothesis are less evident over the subperiods, once we allow for non-Gaussian errors. Our simulation study does not reveal any power problems. Viewed collectively, our findings suggest that a multivariate regression with Student-*t* errors and Fama-French factors seems acceptable as a working framework over the short term, yet the underlying factor structure may be unstable over long time spans.

A. Appendix: Proofs

PROOF OF THEOREM **2.1** Using (2.6), (2.9) and (2.14), we have:

$$\tilde{W} = \hat{U}(J^{-1})'\big(J'S_{\hat{U}}^{-1}\big) = MU(J^{-1})'\big(J'S_{\hat{U}}^{-1}\big) = MW\big(J'S_{\hat{U}}^{-1}\big)\,, \tag{A.1}$$

$$\begin{aligned} S_{\hat{W}}'\,S_{\hat{W}} &= T^{-1}(W'MW) = (J^{-1})T^{-1}U'MU(J^{-1})' = (J^{-1})(\hat{U}'\hat{U}/T)(J^{-1})' \\ &= \big[J'(\hat{U}'\hat{U}/T)^{-1}J\big]^{-1} = \big[J'S_{\hat{U}}^{-1}\big(S_{\hat{U}}^{-1}\big)'J\big]^{-1} = \big[\big(J'S_{\hat{U}}^{-1}\big)\big(J'S_{\hat{U}}^{-1}\big)'\big]^{-1} \\ &= \big[\big(J'S_{\hat{U}}^{-1}\big)^{-1}\big]'\big(J'S_{\hat{U}}^{-1}\big)^{-1}. \end{aligned} \tag{A.2}$$

On observing that $\big(J'S_{\hat{U}}^{-1}\big)^{-1}$ is upper triangular, this means that $\big(J'S_{\hat{U}}^{-1}\big)^{-1}$ is the (unique) Cholesky factor of $T^{-1}(W'MW)$, hence $J'S_{\hat{U}}^{-1} = S_{\hat{W}}^{-1}$; on the unicity of the Cholesky factor, see Harville (1997, Section 14.5c, p. 229). Substituting $S_{\hat{W}}^{-1}$ into (A.1), we see that $\tilde{W} = MW\big(J'S_{\hat{U}}^{-1}\big) = \hat{W}S_{\hat{W}}^{-1}$. □

PROOF OF THEOREM **3.1** Using (2.11) and (2.13), we see that, for $g = 0, 1, \ldots, G$:

$$\begin{aligned} C_{\tilde{W}}(g) &= T^{-1}\sum_{t=g+1}^{T}\tilde{W}_t\tilde{W}_{t-g}' = T^{-1}\sum_{t=g+1}^{T}\big(S_{\hat{U}}^{-1}\big)'\hat{U}_t\hat{U}_{t-g}'S_{\hat{U}}^{-1} \\ &= \big(S_{\hat{U}}^{-1}\big)'\big[T^{-1}\sum_{t=g+1}^{T}\hat{U}_t\hat{U}_{t-g}'\big]S_{\hat{U}}^{-1} = \big(S_{\hat{U}}^{-1}\big)'C_{\hat{U}}(g)S_{\hat{U}}^{-1} \end{aligned}$$

hence

$$\begin{gathered} C_{\tilde{W}}(0)^{-1}C_{\tilde{W}}(g) = S_{\hat{U}}C_{\hat{U}}(0)^{-1}S_{\hat{U}}'\big(S_{\hat{U}}^{-1}\big)'C_{\hat{U}}(g)S_{\hat{U}}^{-1} = S_{\hat{U}}C_{\hat{U}}(0)^{-1}C_{\hat{U}}(g)S_{\hat{U}}^{-1}\,, \\ C_{\tilde{W}}(0)^{-1}C_{\tilde{W}}(g)' = S_{\hat{U}}C_{\hat{U}}(0)^{-1}S_{\hat{U}}'\big(S_{\hat{U}}^{-1}\big)'C_{\hat{U}}(g)'S_{\hat{U}}^{-1} = S_{\hat{U}}C_{\hat{U}}(0)^{-1}C_{\hat{U}}(g)'S_{\hat{U}}^{-1}\,, \\ C_{\tilde{W}}(0)^{-1}C_{\tilde{W}}(g)C_{\tilde{W}}(0)^{-1}C_{\tilde{W}}(g)' = S_{\hat{U}}C_{\hat{U}}(0)^{-1}C_{\hat{U}}(g)C_{\hat{U}}(0)^{-1}C_{\hat{U}}(g)'S_{\hat{U}}^{-1}. \end{gathered}$$

Since matrix multiplication commutes under the trace operator, we have:

$$\begin{aligned} \mathrm{tr}\big\{C_{\tilde{W}}(0)^{-1}C_{\tilde{W}}(g)C_{\tilde{W}}(0)^{-1}C_{\tilde{W}}(g)'\big\} &= \mathrm{tr}\big\{S_{\hat{U}}C_{\hat{U}}(0)^{-1}C_{\hat{U}}(g)C_{\hat{U}}(0)^{-1}C_{\hat{U}}(g)'S_{\hat{U}}^{-1}\big\} \\ &= \mathrm{tr}\big\{C_{\hat{U}}(0)^{-1}C_{\hat{U}}(g)C_{\hat{U}}(0)^{-1}C_{\hat{U}}(g)'\big\} \end{aligned}$$

which establishes that Hosking's criterion obtained from the residuals $\hat{U}$ is equal to the Hosking's criterion based on the standardized residuals $\tilde{W}$. □

PROOF OF THEOREM **3.2** Using (2.11), (2.12) and (2.13), we see that

$$\tilde{W}_t'\tilde{W}_t = \hat{U}_t'\big(S_{\hat{U}}^{-1}\big)\big(S_{\hat{U}}^{-1}\big)'\hat{U}_t = \hat{U}_t'\big(\hat{U}'\hat{U}/T\big)^{-1}\hat{U}_t = \hat{U}_t'\hat{\Sigma}^{-1}\hat{U}_t$$

hence

$$\tilde{R} = \sum_{t=G+1}^{T}\big(\tilde{W}_t'\tilde{W}_t - n\big)\big(\tilde{W}_{t-G}'\tilde{W}_{t-G} - n\big)\Big/\sum_{t=1}^{T}\big(\tilde{W}_t'\tilde{W}_t - n\big)^2.$$

□

B. Appendix: Monte Carlo goodness-of-fit tests

This appendix presents the set estimation method we use to obtain a confidence set for the nuisance parameter ν associated with assumption (2.5). The set is obtained by "inverting" a goodness-of-fit [GF] test, of level α_1 (we considered $\alpha_1 = 2.5\%$ here) for the error distribution underlying (2.5). In other words, the confidence set corresponds to the set of ν_0 values that are not rejected (at the α_1 level) by a GF test which assesses (2.5) imposing $\nu = \nu_0$. We use the multivariate skewness and kurtosis criteria introduced in Dufour et al. (2003) (see also Mardia (1970) and Zhou (1993)):

$$\mathrm{ESK}(\nu_0) = \left|\mathrm{SK} - \overline{\mathrm{SK}}(\nu_0)\right|, \quad \mathrm{EKU}(\nu_0) = \left|\mathrm{KU} - \overline{\mathrm{KU}}(\nu_0)\right|, \tag{B.1}$$

$$\mathrm{SK} = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{i=1}^{T}\hat{d}_{it}^3, \quad \mathrm{KU} = \frac{1}{T}\sum_{t=1}^{T}\hat{d}_{tt}^2, \tag{B.2}$$

where $\hat{d}_{it}$ are the elements of the matrix $\hat{U}(\hat{U}'\hat{U})^{-1}\hat{U}'$ and $\overline{\mathrm{SK}}(\nu_0)$ and $\overline{\mathrm{KU}}(\nu_0)$ are simulation-based estimates of the expected SK and KU given (2.5). Conditional on $\overline{\mathrm{SK}}(\nu_0)$ and $\overline{\mathrm{KU}}(\nu_0)$, these tests satisfy the conditions of Theorem **2.1**. Thus the MC test technique may be applied to obtain their corresponding exact p-values, $\hat{p}(\mathrm{ESK}_0\,|\nu_0)$, $\hat{p}(\mathrm{EKU}_0\,|\nu_0)$. To obtain a joint test we consider:

$$\mathrm{CSK} = 1 - \min\left\{\hat{p}(\mathrm{ESK}_0\,|\nu_0),\ \hat{p}(\mathrm{EKU}_0\,|\nu_0)\right\}. \tag{B.3}$$

The MC technique is applied to the CSK statistic.

References

Affleck-Graves, J. and McDonald, B. (1989), 'Nonnormalities and tests of asset pricing theories', *Journal of Finance* **44**, 889–908.

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.

Bauwens, L., Laurent, S. and Rombouts, J. V. K. (2006), 'Multivariate GARCH models: a survey', *Journal of Applied Econometrics* **21**, 79–109.

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

Binder, J. J. (1985), 'Measuring the effects of regulation with stock price data', *Rand Journal of Economics* **16**, 167–183.

Bollerslev, T., Engle, R. F. and Wooldridge, J. M. (1988), 'A Capital Asset Pricing Model with time-varying covariances', *Journal of Political Economy* **96**, 116–131.

Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*, Princeton University Press, New Jersey.

De Roon, F. A. and Nijman, T. E. (2001), 'Testing for mean-variance spanning: a survey', *Journal of Empirical Finance* **8**, 111–155.

Duchesne, P. and Lalancette, S. (2003), 'On testing for multivariate arch effects in vector time series models', *The Canadian Journal of Statistics* **31**, 275–292.

Dufour, J.-M. (1989), 'Nonlinear hypotheses, inequality restrictions, and non-nested hypotheses: Exact simultaneous tests in linear regressions', *Econometrica* **57**, 335–355.

Dufour, J.-M. (1990), 'Exact tests and confidence sets in linear regressions with autocorrelated errors', *Econometrica* **58**, 475–494.

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

Dufour, J.-M. and Khalaf, L. (2002*b*), 'Simulation based finite and large sample tests in multivariate regressions', *Journal of Econometrics* **111**(2), 303–322.

Dufour, J.-M. and Khalaf, L. (2003), Finite sample tests in seemingly unrelated regressions, *in* D. E. A. Giles, ed., 'Computer-Aided Econometrics', Marcel Dekker, New York, chapter 2, pp. 11–35.

Dufour, J.-M., Khalaf, L. and 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**, 891–906.

Dufour, J.-M., Khalaf, L., Bernard, J.-T. and Genest, I. (2004), 'Simulation-based finite-sample tests for heteroskedasticity and ARCH effects', *Journal of Econometrics* **122**(2), 317–347.

Dufour, J.-M. and Kiviet, J. F. (1996), 'Exact tests for structural change in first-order dynamic models', *Journal of Econometrics* **70**, 39–68.

Dufour, J.-M. and Kiviet, J. F. (1998), 'Exact inference methods for first-order autoregressive distributed lag models', *Econometrica* **66**, 79–104.

Dufour, J.-M. and Torrès, O. (1998), Union-intersection and sample-split methods in econometrics with applications to SURE and MA models, *in* D. E. A. Giles and A. Ullah, eds, 'Handbook of Applied Economic Statistics', Marcel Dekker, New York, pp. 465–505.

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

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation', *Econometrica* **50**(4), 987–1008.

Engle, R. F. and Kroner, K. F. (1995), 'Multivariate generalized ARCH', *Econometric Theory* **11**, 122–150.

Fama, E. F. and French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* **33**, 3–56.

Fama, E. F. and French, K. R. (1995), 'Size and book-to-market factors in earnings and returns', *The Journal of Finance* **50**, 131–155.

Folks, J. L. (1984), Combination of independent tests, *in* P. R. Krishnaiah and P. K. Sen, eds, 'Handbook of Statistics 4: Nonparametric Methods', North-Holland, Amsterdam, pp. 113–121.

Gibbons, M. R., Ross, S. A. and Shanken, J. (1989), 'A test of the efficiency of a given portfolio', *Econometrica* **57**, 1121–1152.

Godfrey, L. G. (1988), *Misspecification Tests in Econometrics : The Lagrange Multiplier Principle and Other Approaches*, Cambridge University Press, Cambridge, UK.

Harville, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, Springer-Verlag, New York.

Hosking, J. (1980), 'The multivariate portmanteau statistic', *Journal of the American Statistical Association* **75**, 602–608.

Kroner, K. F. and Ng, V. P. (1998), 'Modeling Asymmetric Comovements of Asset Returns', *The Review of Financial Studies* **11**, 817–844.

Lee, J. H. and King, M. L. (1993), 'A locally most mean powerful based score test for ARCH and GARCH regression disturbances', *Journal of Business and Economic Statistics* **11**, 17–27. Correction 12 (1994), 139.

Ling, S. and Li, W. K. (1997), 'Diagnostic checking on nonlinear multivariate time series with multivariate arch errors', *Journal of Time Series Analysis* **18**, 447–464.

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

Lo, A. and MacKinlay, C. (1988), 'Stock prices do not follow random walks: Evidence from a simple specification test', *Review of Financial Studies* **1**, 41–66.

Lo, A. and MacKinlay, C. (1989), 'The size and power of the variance ratio test in finite samples: A Monte Carlo investigation', *Journal of Econometrics* **40**, 203–238.

MacKinlay, A. C. (1987), 'On multivariate tests of the Capital Asset Pricing Model', *Journal of Financial Economics* **18**, 341–372.

Mardia, K. V. (1970), 'Measures of multivariate skewness and kurtosis with applications', *Biometrika* **57**, 519–530.

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

Richardson, M. and Smith, T. (1993), 'A test for multivariate normality in stock returns', *Journal of Business* **66**, 295–321.

Savin, N. E. (1984), Multiple hypothesis testing, *in* Z. Griliches and M. D. Intrilligator, eds, 'Handbook of Econometrics, Volume 2', North-Holland, Amsterdam, chapter 14, pp. 827–879.

Schipper, K. and Thompson, R. (1985), 'The impact of merger-related regulations using exact distributions of test statistics', *Journal of Accounting Research* **23**, 408–415.

Shanken, J. (1986), 'Testing portfolio efficiency when the zero-beta rate is unknown: A note', *Journal of Finance* **41**, 269–276.

Shanken, J. (1990), 'Intertemporal asset pricing: An empirical investigation', *Journal of Econometrics* **45**, 99–120.

Shanken, J. (1996), Statistical methods in tests of portfolio efficiency: A synthesis, *in* G. S. Maddala and C. R. Rao, eds, 'Handbook of Statistics 14: Statistical Methods in Finance', North-Holland, Amsterdam, pp. 693–711.

Stewart, K. G. (1997), 'Exact testing in multivariate regression', *Econometric Reviews* **16**, 321–352.

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

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

Zhou, G. (1993), 'Asset-pricing tests under alternative distributions', *The Journal of Finance* **48**, 1927–1942.