Multivariate Tests of Mean–Variance Efficiency With Possibly Non-Gaussian Errors: An Exact Simulation-Based Approach

Marie-Claude BEAULIEU
CIRANO and Department of Finance and Insurance, Laval University, Ste-Foy, Quebec, Canada G1K 7P4 (*Marie-Claude.Beaulieu@fas.ulaval.ca*)

Jean-Marie DUFOUR
CIRPEE, CIREQ, and Department of Economic Science, University of Montreal, Montreal, Quebec, Canada H3C 3J7 (*jean.marie.dufour@umontreal.ca*)

Lynda KHALAF
CIREQ, GREEN, and Department of Economics, Carleton University, Ottawa, Ontario, Canada K1S 5B6 (*lynda-khalaf@carleton.ca*)

We develop exact mean–variance efficiency tests of the market portfolio in the context of (conditional and unconditional) capital asset pricing models (CAPM), allowing for a wide class of possibly non-Gaussian error distributions. The proposed procedures are applicable in a general multivariate linear regression framework, and exactness is achieved through Monte Carlo test techniques. We also perform exact multivariate diagnostic checks. Empirical results show that the Gaussian assumption is rejected, temporal instabilities are apparent, and mean–variance efficiency is rejected over several subperiods, but finite-sample methods that allow for nonnormality and conditioning information substantially reduce the number of rejections.

KEY WORDS: Bootstrap; Capital asset pricing model; Generalized autoregressive conditional heteroscedasticity; Monte Carlo test; Multivariate linear regression; Nonnormality.

1. INTRODUCTION

The capital asset pricing model (CAPM) is one of the most commonly used models in theoretical and applied finance; (for reviews and references, see Campbell, Lo, and MacKinlay 1997; Shanken 1996; Cochrane 2001; DeRoon and Nijman 2001; Fama and French 2004). Since the work of Gibbons (1982), empirical tests on the CAPM are often conducted within a *multivariate linear regression* (MLR). In this context, standard asymptotic theory provides a poor approximation to the finite-sample distribution of test statistics, even with fairly large samples (see Shanken 1996, sec. 3.4.2; Campbell et al. 1997, chap. 5; Dufour and Khalaf 2002b). In particular, test size distortions grow quickly when the number of equations increases. As a result, the conclusions of MLR-based empirical studies on the CAPM can be strongly affected and can lead to spurious rejections.

Consequently, several exact and Bayesian methods have been proposed to assess mean–variance efficiency (see Jobson and Korkie 1982; MacKinlay 1987, 1995; Gibbons, Ross, and Shanken 1989 [henceforth GRS]; Stewart 1997; Kandel, McCulloch, and Stambaugh 1995). These methods typically require Gaussian distributional assumptions. However, it has long been recognized that financial returns exhibit nonnormalities (Fama 1965). Although the CAPM can be derived from expected utility maximization under various non-Gaussian assumptions on the return cross-sectional distribution, such as the multivariate t (see Ingersoll 1987; Berk 1997), finite-sample tests for mean–variance efficiency in non-Gaussian CAPMs are not yet available.

Indeed, mean–variance efficiency tests that relax normality include (a) large-sample GMM or bootstrap techniques (Affleck-Graves and McDonald 1989; MacKinlay and Richardson 1991; Fama and French 1993; Jagannathan and Wang 1996; Ferson and Harvey 1999; Groenwold and Fraser 2001); (b) semiparametric asymptotic procedures specific to elliptical distributions (Hodgson, Linton, and Vorkink 2002; Vorkink 2003; Hodgson and Vorkink 2003); (c) parametric procedures based on postulating a non-Gaussian distribution, such as the multivariate t (Fiorentini, Sentana, and Calzolari 2003; Zhou 1993); and (d) non-Gaussian Bayesian procedures (Tu and Zhou 2004). In all of these approaches, the distributional theory of test statistics is either approximate or does not formally take into account nuisance parameter uncertainty in a fitted parametric distribution. In particular, Hodgson et al. (2002) reported size problems on high-dimensional systems and restricted their analysis to systems with three or four portfolios, whereas Vorkink (2003) proceeded on a portfolio-by-portfolio basis, not on the whole system. In the parametric case, Zhou (1993) proposed simulation-based p values for the GRS statistic given a few elliptical distributions, while selecting their tail area parameters by trial and error.

In this article we propose finite-sample unconditional and conditional multivariate mean–variance efficiency tests in possibly non-Gaussian CAPMs. The conditional specifications allow model coefficients to vary as functions of a number

Journal of Business & Economic Statistics
October 2007, Vol. 25, No. 4
DOI 10.1198/073500106000000468

of instruments, as described by Shanken (1996, sec. 2.3.4), Cochrane (2001, chap. 8), and DeRoon and Nijman (2001). Conditional testing is important because portfolios that are conditionally efficient might be unconditionally inefficient (see Hansen and Richard 1987; Cochrane 2001, chap. 8).

We use finite-sample results from Dufour and Khalaf (2002b) on testing *uniform linear* (UL) restrictions in MLR models with a given, possibly non-Gaussian, disturbance distribution. For such hypotheses, the null distributions of standard test statistics are invariant to MLR coefficients and error variances and covariances. In this case, Monte Carlo (MC) test techniques (see Dufour 2006) can be applied to obtain exact p values. On observing that mean–variance efficiency restrictions take the UL form when the risk-free rate is observable, we show that efficiency can be tested exactly under general distributional assumptions that include the Gaussian and a wide spectrum of non-Gaussian distributions, both elliptically symmetric and nonelliptical. Single- and multi-beta models are covered by these results.

To control for the parameters that define the hypothesized non-Gaussian distribution, such as the degrees of freedom for the multivariate t (a problem not considered in Dufour and Khalaf 2002b), we use a two-stage procedure as follows: (1) We build an exact confidence set (with level $1-\alpha_1$) for the nuisance parameter, through "inversion" of a distributional goodness-of-fit (GF) test, and (2) maximize the p value for the mean–variance efficiency test (which depends on the nuisance parameter) over this confidence set. Referring the latter *maximized MC* (MMC) p value to an α_2 cutoff provides a test with exact level $\alpha_1+\alpha_2$ (see Dufour and Kiviet 1996; Dufour 2006). We stick here to the original notion of test level in the presence of nuisance parameters (Lehmann 1986, chap. 3): A test has *level* α if the probability of rejecting the null hypothesis is not greater than α for any data-generating process compatible with the null hypothesis.

Furthermore, we evaluate the specification of the model using GF tests on the error distribution and serial dependence tests. All procedures rely on properly standardized multivariate ordinary least squares (OLS) residuals, which provide statistics invariant to MLR coefficients and error variances and covariances and allows easy application of MC tests. The GF tests compare multivariate skewness and kurtosis criteria with a simulation-based estimate of their expected value under the hypothesized (normal or nonnormal) distribution, which can be viewed as extensions of Mardia's (1970) procedures. The diagnostic checks combine (as in Shanken 1990) standardized individual-equation versions of the generalized autoregressive conditional heteroscedasticity (GARCH) tests suggested by Engle (1982) and Lee and King (1993), and the variance-ratio tests of Lo and MacKinlay (1988); we also test for heteroscedasticity linked to conditioning on market returns (Vorkink 2003). Our exact combination method relies on simulation (as in Dufour and Khalaf 2002a and Dufour, Khalaf, Bernard, and Genest 2004) to avoid the Bonferroni bounds applied by Shanken (1990). Such bounds require one to divide the level of each individual test by the number of tests, leading to possibly large power losses if the MLR includes many equations (i.e., many portfolios). All tests are performed under normal and nonnormal error distributions.

The proposed tests are applied to an unconditional and a conditional CAPM with observable risk-free rates and both multivariate normal and multivariate t distributions. We consider monthly returns on New York Stock Exchange (NYSE) portfolios, constructed from the University of Chicago Center for Research in Security Prices (CRSP) database (1926–1995). Our results show the following: (a) Multivariate normality is rejected; (b) multivariate residual checks suggest temporal instabilities for both the unconditional and the conditional models; (c) although mean–variance efficiency is rejected over several subperiods, using finite-sample methods and allowing for nonnormal errors reduces the number of subperiods for which efficiency is rejected and the strength of the evidence against it; and (d) using conditioning information has nonnegligible effects on tests of mean–variance efficiency and substantially reduces the number of rejections.

The article is organized as follows. In Section 2 we set up the framework. In Section 3 we describe existing tests and propose extensions for nonnormal distributions. In Section 4 we discuss how to deal with nuisance parameters in the error distribution. We describe GF and diagnostic tests in Section 5 and report the empirical results in Section 6. We conclude in Section 7.

2. FRAMEWORK

Let R_{it}, $i=1,\dots,n$, be returns on n securities for period t, and let $\tilde{R}_{Mt}$ be the return on a benchmark portfolio ($t=1,\dots,T$). Following Gibbons et al. (1989), the (unconditional) CAPM, which assumes time-invariant betas, can be assessed by testing

$$\mathcal{H}_E: a_i = 0, \qquad i=1,\dots,n, \tag{1}$$

in the context of the MLR model,

$$r_{it} = a_i + \beta_i \tilde{r}_{Mt} + \varepsilon_{it}, \qquad t=1,\dots,T, i=1,\dots,n, \tag{2}$$

where $r_{it}=R_{it}-R_{ft}$, $\tilde{r}_{Mt}=\tilde{R}_{Mt}-R_{ft}$, R_{ft} is the riskless rate of return, and ε_{it} is a random disturbance.

In general, the CAPM also allows for the possibility of time-varying betas. As discussed by Shanken (1996, sec. 2.3.4), Cochrane (2001, chap. 8), and DeRoon and Nijman (2001), this can be accommodated using conditioning information, such as lagged variables (or instruments, known at time t). In particular, model parameters can be viewed as linear functions of q conditioning variables $z_{1t},\dots,z_{qt}$; depending on whether only the betas or both the intercepts and the betas are allowed to vary, this leads to alternative specifications:

$$\begin{aligned} r_{it} &= \bar{a}_i + \beta_{it}\tilde{r}_{Mt} + \varepsilon_{it}, \\ \beta_{it} &= \bar{\beta}_i + \sum_{j=1}^{q} d_{ji} z_{jt} \end{aligned} \tag{3}$$

and

$$\begin{aligned} r_{it} &= a_{it} + \beta_{it}\tilde{r}_{Mt} + \varepsilon_{it}, \\ a_{it} &= \bar{a}_i + \sum_{j=1}^{q} c_{ji} z_{jt}, \\ \beta_{it} &= \bar{\beta}_i + \sum_{j=1}^{q} d_{ji} z_{jt}, \end{aligned} \tag{4}$$

$t = 1, \ldots, T, i = 1, \ldots, n$. Model (3) entails the expanded regression,

$$r_{it} = \bar{a}_i + \bar{\beta}_i \tilde{r}_{Mt} + \sum_{j=1}^{q} d_{ji}(\tilde{r}_{Mt} z_{jt}) + \varepsilon_{it}, \qquad t = 1, \ldots, T, i = 1, \ldots, n. \quad (5)$$

Assuming that the regressor matrix has full column rank, efficiency can be assessed by testing

$$\bar{\mathcal{H}}_{E1} : \bar{a}_i = 0, \qquad i = 1, \ldots, n. \quad (6)$$

Similarly, (4) leads to the following equation:

$$r_{it} = \bar{a}_i + \bar{\beta}_i \tilde{r}_{Mt} + \sum_{j=1}^{q} c_{ji} z_{jt} + \sum_{j=1}^{q} d_{ji}(\tilde{r}_{Mt} z_{jt}) + \varepsilon_{it}, \qquad t = 1, \ldots, T, \ i = 1, \ldots, n. \quad (7)$$

Assuming again that the corresponding regressor matrix has full column rank, efficiency can be assessed by testing $a_{it} = 0$ for all i and t or, equivalently,

$$\bar{\mathcal{H}}_{E2} : \bar{a}_i = 0, \qquad c_{ji} = 0, \qquad i = 1, \ldots, n, j = 1, \ldots, q. \quad (8)$$

For further reference, we set $\mathcal{O}(l, m)$ to be the $l \times m$ zero matrix and

$$\begin{aligned} \iota_T &= (1, \ldots, 1)', \\ \tilde{\mathbf{r}}_M &= (\tilde{r}_{1M}, \ldots, \tilde{r}_{TM})', \\ \mathbf{r}_i &= (r_{1i}, \ldots, r_{Ti})' \end{aligned} \quad (9)$$

and

$$\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_q], \qquad \mathbf{z}_j = (z_{j1}, \ldots, r_{jT})', \qquad j = 1, \ldots, q. \quad (10)$$

We also use "$*$" to denote element-by-element rowwise matrix multiplication; for example, if $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_T]'$ is a $T \times l$ matrix and $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_T]'$ is an $T \times m$ matrix, then $\mathbf{A} * \mathbf{D}$ is the $T \times (lm)$ matrix with tth row equal to $\mathbf{A}_t' \otimes \mathbf{D}_t'$, that is, $\mathbf{A} * \mathbf{D} = [\mathbf{A}_1 \otimes \mathbf{D}_1, \ldots, \mathbf{A}_T \otimes \mathbf{D}_T]'$.

The foregoing models are special cases of the MLR model,

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}, \quad (11)$$

where $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_n]$ is a $T \times n$ matrix of dependent variables, $\mathbf{X}$ is a $T \times k$ full-column rank matrix of regressors, and $\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_n] = [\mathbf{V}_1, \ldots, \mathbf{V}_T]'$ is a $T \times n$ matrix of disturbances. Furthermore, the hypotheses $\mathcal{H}_E$, $\bar{\mathcal{H}}_{E1}$, and $\bar{\mathcal{H}}_{E2}$ belong to the UL class; that is, they have the form

$$\mathcal{H}_0 : \mathbf{H}\mathbf{B} = \mathcal{O}(h, n), \quad (12)$$

where $\mathbf{H}$ is a fixed $h \times k$ matrix of rank h. Indeed, (1)–(2), (5)–(6), and (7)–(8) constitute special cases of (11)–(12) obtained by taking one of the three following definitions:

$$\mathbf{Y} = [\mathbf{r}_1, \ldots, \mathbf{r}_n], \qquad \mathbf{X} = [\iota_T, \tilde{\mathbf{r}}_M], \qquad \mathbf{H} = (1, 0); \quad (13)$$

$$\begin{aligned} \mathbf{Y} &= [\mathbf{r}_1, \ldots, \mathbf{r}_n], \\ \mathbf{X} &= [\iota_T, \tilde{\mathbf{r}}_M, \tilde{\mathbf{r}}_M * \mathbf{z}], \\ \mathbf{H} &= [1, \mathcal{O}(1, q+1)]; \end{aligned} \quad (14)$$

$$\begin{aligned} \mathbf{Y} &= [\mathbf{r}_1, \ldots, \mathbf{r}_n], \\ \mathbf{X} &= [\iota_T, \mathbf{z}, \tilde{\mathbf{r}}_M, \tilde{\mathbf{r}}_M * \mathbf{z}], \\ \mathbf{H} &= [\mathbf{I}_{q+1}, \mathcal{O}(q+1, q+1)]. \end{aligned} \quad (15)$$

In this context we apply a formal statistical approach to obtain simple finite-sample tests under alternative error distributions (assuming that we can condition on $\mathbf{X}$, that is, we can take $\mathbf{X}$ as fixed for statistical analysis). More precisely, we consider the general case

$$\mathbf{V}_t \equiv (\varepsilon_{1t}, \ldots, \varepsilon_{nt})' = \mathbf{J}\mathbf{W}_t, \qquad t = 1, \ldots, T, \quad (16)$$

where $\mathbf{J}$ is an unknown nonsingular matrix and the distribution of the vector $\mathbf{w} = \text{vec}(\mathbf{W})$, $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_T]'$ is either known (hence free of nuisance parameters) or specified up to an unknown finite-dimensional nuisance-parameter (denoted by ν); we call $\mathbf{w}$ the vector of *normalized disturbances* and its distribution the *normalized disturbance distribution*. Let $\boldsymbol{\Sigma} = \mathbf{J}\mathbf{J}'$, so that $\det(\boldsymbol{\Sigma}) \neq 0$. For example, we assume that $\mathbf{W}_t \sim \mathcal{F}(\nu)$, $t = 1, \ldots, T$, where $\mathcal{F}(\cdot)$ represents a known distribution function. Later we consider both the case where the error distribution does not involve nuisance parameters,

$$\mathbf{W}_t \sim \mathcal{F}(\nu_0) \quad \text{where } \nu_0 \text{ is specified}, \quad (17)$$

and the case where it does,

$$\mathbf{W}_t \sim \mathcal{F}(\nu) \quad \text{where } \nu \text{ is unknown}. \quad (18)$$

This assumption includes as special cases the Gaussian distribution,

$$\mathbf{V}_1, \ldots, \mathbf{V}_T \overset{\text{iid}}{\sim} \text{N}[\mathbf{0}, \boldsymbol{\Sigma}]; \quad (19)$$

all elliptically symmetric distributions, such as the multivariate t; and cases where $\mathbf{W}_1, \ldots, \mathbf{W}_T$ are independent and identically distributed (iid) according to any given nonelliptical distribution. In this regard, conditioning on further instruments [as in models (5) and (7)]—rather than only on the market portfolio—can make the iid error hypothesis more plausible.

3. MEAN–VARIANCE EFFICIENCY TESTS WITH A KNOWN NORMALIZED DISTURBANCE DISTRIBUTION

In this section we consider testing $\mathcal{H}_E$, $\bar{\mathcal{H}}_{E1}$, and $\bar{\mathcal{H}}_{E2}$ [in (1), (6), or (8)] under the distributional assumption (16). The test statistics used are Gaussian likelihood ratios (LRs),

$$\begin{aligned} LR &= T \ln(\Lambda), \qquad \Lambda = |\hat{\boldsymbol{\Sigma}}_0| / |\hat{\boldsymbol{\Sigma}}|, \\ \hat{\boldsymbol{\Sigma}} &= \hat{\mathbf{U}}'\hat{\mathbf{U}}/T, \qquad \hat{\boldsymbol{\Sigma}}_0 = \hat{\mathbf{U}}_0'\hat{\mathbf{U}}_0/T; \end{aligned} \quad (20)$$

$$\begin{aligned} \hat{\mathbf{U}} &= \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}, \\ \hat{\mathbf{B}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \\ \hat{\mathbf{U}}_0 &= \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_0; \end{aligned} \quad (21)$$

and

$$\hat{\mathbf{B}}_0 = \hat{\mathbf{B}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'[\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}\hat{\mathbf{B}}, \quad (22)$$

where $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{H}$ are defined as in (13), (14), or (15), depending on the null hypothesis ($\mathcal{H}_E$, $\bar{\mathcal{H}}_{E1}$, or $\bar{\mathcal{H}}_{E2}$). On using the results of Dufour and Khalaf (2002b, sec. 3 and the app.), the null distribution of the LRs can be characterized as follows (under a possibly non-Gaussian error distribution).

Theorem 1 (Null distribution of Gaussian LR statistics for mean–variance efficiency). Under (16), the *LR* statistic defined in (20) for testing $\mathcal{H}_E$ against the (unrestricted) model (2) [resp. $\bar{\mathcal{H}}_{E1}$ against (5), or $\bar{\mathcal{H}}_{E2}$ against (7)] is distributed like

$$L(\mathbf{W}) \equiv T \ln\big(|\mathbf{W}'\mathbf{M}_0\mathbf{W}|/|\mathbf{W}'\mathbf{M}\mathbf{W}|\big)$$

under the null hypothesis, where $\mathbf{M}_0 = \mathbf{M} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}' \times [\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}']^{-1}\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and $\mathbf{H}$ is defined in (13) [resp. in (14) and (15)]. If, furthermore, the Gaussian assumption (19) holds and $T-k-n+1 \geq 1$, where k is the number of columns in $\mathbf{X}$, then $[(T-k-n+1)/n](\Lambda - 1) \sim F(n, T-k-n+1)$ under $\mathcal{H}_E$ and $\bar{\mathcal{H}}_{E1}$, and

$$\frac{\rho\tau - 2\lambda}{n(q+1)}\big(\Lambda^{1/\tau} - 1\big) \sim F[n(q+1),\ \rho\tau - 2\lambda] \qquad (23)$$

under $\bar{\mathcal{H}}_{E2}$ when $\min(n,\ q+1) \leq 2$, where $\rho = (T-k) - [(n-q)/2]$, $\lambda = [n(q+1) - 2]/4$, and

$$\tau = \begin{cases} \left\{\dfrac{n^2(q+1)^2 - 4}{n^2 + (q+1)^2 - 5}\right\}^{1/2} & \text{if } n^2 + (q+1)^2 - 5 > 0 \\ 1 & \text{otherwise.} \end{cases}$$

The foregoing analysis easily extends to multi-beta models of the form

$$r_{it} = a_i + \sum_{j=1}^{s} \beta_{ji}\tilde{r}_{jt} + \varepsilon_{it}, \qquad t = 1, \ldots, T, i = 1, \ldots, n, \quad (24)$$

where $\tilde{r}_{jt} = \tilde{R}_{jt} - R_{ft}$ and $\tilde{R}_{jt}, j = 1, \ldots, s$, are returns on s benchmark portfolios. In such models the hypothesis being tested entails a portfolio of the benchmark portfolios that is mean–variance efficient (see Gibbons et al. 1989; henceforth GRS). Unconditional efficiency tests follow from Theorem 1 with $\mathbf{X} = [\iota_T, \tilde{\mathbf{r}}]$, $\tilde{\mathbf{r}} = [\tilde{\mathbf{r}}_1, \ldots, \tilde{\mathbf{r}}_s]$, $\tilde{\mathbf{r}}_j = (\tilde{r}_{1j}, \ldots, \tilde{r}_{Tj})'$, $j = 1, \ldots, s$, and $\mathbf{H} = [1, \mathcal{O}(1, s)]$. Furthermore, conditional efficiency tests are covered by Theorem 1 in the context of an expanded MLR of the form (11), where $\mathbf{Y} = [\mathbf{r}_1, \ldots, \mathbf{r}_n]$, with $\mathbf{X} = [\iota_T, \tilde{\mathbf{r}}, \tilde{\mathbf{r}} * \mathbf{z}]$ and $\mathbf{H} = [1, \mathcal{O}(1, qs+s)]$ if the coefficients of $\tilde{\mathbf{r}}$ are assumed to be linear functions of the instruments and $\mathbf{X} = [\iota_T, \mathbf{z}, \tilde{\mathbf{r}}, \tilde{\mathbf{r}} * \mathbf{z}]$ and $\mathbf{H} = [\mathbf{I}_{q+1}, \mathcal{O}(q+1, qs+s)]$ if the intercepts also are linear functions of the instruments.

Theorem 1 entails that the distribution of the LR statistic in (20) does not depend on $\mathbf{B}$ and $\boldsymbol{\Sigma}$. This property holds under conditions much more general than elliptical symmetry (as emphasized in the earlier literature on the CAPM). Therefore, given draws from the distribution of the disturbance matrix $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_T]$, an exact p value may be obtained using the MC test technique as follows:

1. Using the distributional assumption (16) and a given value of ν [as in (17)], generate N iid replications of the disturbance matrix $\mathbf{W}$.
2. This yields N simulated values of the test statistic, applying the relevant pivotal transform $L(\mathbf{W})$ from Theorem 1.
3. Calculate the exact Monte Carlo p value from the rank of the observed LR relative to the simulated ones [see (A.2) in Sec. A.1].

Further details are supplied in Section A.1. By the general theory of MC tests, we observe that the size of a simulation-based test can be perfectly controlled even with a very small number of MC replications. For example, 19 replications are sufficient to obtain a test of size .05. For power considerations, there is in principle an advantage in using a larger number of replications, but the power gain from using more than 100 or 200 replications is typically very small. (For further discussion on this issue, see Dufour and Kiviet 1996; Dufour and Khalaf 2002a,b; Dufour et al. 2004; Dufour 2006.)

We denote by $\hat{p}_N(LR_0|\nu)$ the MC p value so obtained, where LR_0 is the observed value of the LR statistic and ν represents the distributional parameter used. We consider the case where ν is taken as unknown in Section 4. It is noteworthy that the latter MC test approach is useful even with Gaussian errors, as in $\bar{\mathcal{H}}_{E2}$, because an analytical distribution is not always available.

Two other results follow from Theorem 1. First, our Gaussian LR test is equivalent to the Hotelling T^2 test proposed by MacKinlay (1987) and Gibbons et al. (1989). In the context of (24), the latter apply tests based on the following distributional result:

$$\frac{T-s-n}{n(T-s-1)} Q \sim F(n, T-s-n),$$
$$\text{with } Q = T\hat{\mathbf{a}}'\left[\frac{T}{T-k}\hat{\boldsymbol{\Sigma}}\right]^{-1}\hat{\mathbf{a}}\big/\big[1 + \bar{\mathbf{r}}'\hat{\boldsymbol{\Delta}}^{-1}\bar{\mathbf{r}}\big], \quad (25)$$

where $\hat{\mathbf{a}}$ is the vector of intercept OLS estimates, $[T/(T-k)]\hat{\boldsymbol{\Sigma}}$ is the OLS-based unbiased estimator of $\boldsymbol{\Sigma}$, $\bar{\mathbf{r}}$, and $\hat{\boldsymbol{\Delta}}$ include the time series means and sample covariance matrix corresponding to the right-side returns. On observing that Q and Λ are related by the monotonic transformation $\Lambda - 1 = Q/(T-s-1)$, where $s = k-1$ (see Stewart 1997), we see that GRS's results follow from Theorem 1 under normal errors. Second, it is easy to see that our results extend beyond the mean–variance efficiency hypothesis and cover any hypothesis of the form (12) on a MLR of the form (2) describing returns. In this case, the null distribution of the LR statistic follows from Theorem 1 for the specific $\mathbf{H}$ matrix considered. For hypotheses where $\min(h, n) > 2$ (such as $\bar{\mathcal{H}}_{E2}$), the MC approach is necessary even with Gaussian errors, because a transformation of the LR statistic with a Fisher distribution (as for GRS statistic) does not seem to be available.

4. MEAN–VARIANCE EFFICIENCY TESTS WITH AN INCOMPLETELY SPECIFIED ERROR DISTRIBUTION

In this section we extend the foregoing results to the case of (16) where ν is unknown. To formally account for the problem of estimating ν, we apply the following MMC approach (see Dufour and Kiviet 1996), which involves two stages. First, we build an exact confidence set for ν with level $1 - \alpha_1$, which we denote by $\mathcal{C}(\mathbf{Y})$, with $\mathbf{Y}$ referring to the return data. Next, on applying Theorem 1 and the MC algorithm in Section A.1 (summarized earlier), we can obtain a MC p value $\hat{p}_N(LR_0|\nu_0)$ for each $\nu_0 \in \mathcal{C}(\mathbf{Y})$. Setting

$$Q_{\mathrm{U}}(LR_0) = \sup_{\nu_0 \in \mathcal{C}(\mathbf{Y})} \hat{p}_N(LR_0|\nu_0), \qquad (26)$$

the critical region

$$Q_{\mathrm{U}}(LR_0) \leq \alpha_2 \tag{27}$$

has level $\alpha_1 + \alpha_2$. In other words, if we construct the nuisance parameter confidence set with level α_1 and refer the sup p value to the cutoff level α_2, then the global level of the two-stage test is $\alpha = \alpha_1 + \alpha_2$. In the empirical application considered next, we use $\alpha_1 = \alpha_2 = \alpha/2$.

Because a procedure for deriving an exact confidence set for ν is not readily available (even with multivariate t errors), we provide one here. Given the recent literature documenting the dramatically poor performance of asymptotic Wald-type confidence intervals, we prefer to "invert" a test for the null hypothesis (16) where $\nu = \nu_0$ for known ν_0. Specifically, suppose that some test statistic [denoted by $\mathcal{T}(\mathbf{Y})$] is available for the latter hypothesis; we provide one in Section 5.1. Inverting $\mathcal{T}(\mathbf{Y})$ implies assembling the ν_0 values that are not rejected at a specific significance level. This may be carried out as follows: Using, for example, a grid search over the relevant values of ν_0, compute the statistic associated with $\nu = \nu_0$ from the observed sample [say $\mathcal{T}_0(\mathbf{Y})$] and its p value [say $\hat{p}_N(\mathcal{T}_0(\mathbf{Y})|\nu_0)$, obtained by MC test techniques], conforming with (16). The confidence set for ν (which is not necessarily a bounded confidence interval) at level α_1 corresponds to the values of ν_0 such that $\hat{p}_N(\mathcal{T}_0(\mathbf{Y})|\nu_0) > \alpha_1$ (see Dufour 1990; Dufour and Kiviet 1996).

5. EXACT DIAGNOSTIC CHECKS

In this section we present multivariate specification tests, including distributional GF tests—which we invert to estimate ν—and checks for departures from the hypothesis of iid errors.

5.1 Goodness-of-Fit Tests

The null hypotheses of concern here are (19) (normal errors), and (17) or (18) (e.g., multivariate t errors with known or unknown degrees of freedom). The test criteria considered use the multivariate skewness and kurtosis measures

$$SK = \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\hat{d}_{st}^{3} \qquad \text{and} \qquad KU = \frac{1}{T}\sum_{t=1}^{T}\hat{d}_{tt}^{2}, \tag{28}$$

where $\hat{d}_{st}$ are the elements of the matrix $\hat{\mathbf{D}} = \hat{\mathbf{U}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{U}} = T\hat{\mathbf{U}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'$. These statistics were introduced by Mardia (1970) in models in which the regressor reduces to a vector of 1's. Zhou (1993, p. 1935, footnote 5) proposed using these statistics to test elliptical distributions, but did not provide a finite-sample theory for their application to residuals from MLR models. In our context, these statistics are distributed like $\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} d_{st}^3$ and $\frac{1}{T}\sum_{t=1}^{T} d_{tt}^2$, where d_{st} is the (s,t)th element of the matrix

$$\mathbf{D} = T\mathbf{M}\mathbf{W}(\mathbf{W}'\mathbf{M}\mathbf{W})^{-1}\mathbf{W}'\mathbf{M}, \qquad \mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_T]' \tag{29}$$

(see Dufour, Khalaf, and Beaulieu 2003). This implies that SK and KU are pivotal (i.e., invariant to $\mathbf{B}$ and $\boldsymbol{\Sigma}$). Two further adjustments are applied: a simulation-based "centering" of the test statistics and a formal procedure for combining them into a single test.

Centering involves using both measures in excess of expected values consistent with the hypothesized error distribution. In view of (17), the resulting statistics are denoted by $\overline{SK}(\nu_0)$ and $\overline{KU}(\nu_0)$. In the Gaussian case (19), we use the simplified notations $\overline{SK}$ and $\overline{KU}$. In view of the absence of an analytical form for the expected values, the latter are evaluated by simulation, yielding the following simulation-based statistics:

$$\begin{aligned} ESK(\nu_0) &= |SK - \overline{SK}(\nu_0)| \qquad \text{and} \\ EKU(\nu_0) &= |KU - \overline{KU}(\nu_0)| \end{aligned} \tag{30}$$

in the general case; in the Gaussian case, the test statistics are denoted by $ESK = |SK - \overline{SK}|$ and $EKU = |KU - \overline{KU}|$. This modification preserves pivotality. The MC technique thus may be applied to derive exact p values (using N replications); the resulting simulation-based p values are denoted $\hat{p}_N(ESK(\nu_0)|\nu_0)$, $\hat{p}_N(EKU(\nu_0)|\nu_0)$ in the general case and by $\hat{p}_N(ESK_0)$, $\hat{p}_N(EKU_0)$ under the Gaussian hypothesis (see Sec. A.2.2 for more details). The observed and simulated statistics must be obtained conditional on the same average skewness and kurtosis measures to this ensure that they remain exchangeable (see Dufour 2006).

This procedure allows us to obtain exact individual p values for each statistic. To obtain a joint test, we propose rejecting the null hypothesis if at least one of the individual p values is significantly small. To avoid relying on Boole–Bonferroni rules in defining the cutoff level, we use the following combined statistic (see Dufour et al. 2003):

$$CSK(\nu_0) = 1 - \min\{\hat{p}_N(ESK(\nu_0)|\nu_0), \hat{p}_N(EKU(\nu_0)|\nu_0)\}, \tag{31}$$

or $CSK = 1 - \min\{\hat{p}_N(ESK), \hat{p}_N(EKU)\}$ in the Gaussian case. This combination method preserves invariance to $\mathbf{B}$ and $\boldsymbol{\Sigma}$. Therefore, under (19), a MC p value for CSK can be easily obtained. Under (17), pivotality allows us to obtain a MC p value given a known value $\nu = \nu_0$, which is denoted by $\hat{p}_N(CSK_0(\nu_0)|\nu_0)$, where CSK_0 refers to the observed value of the statistic. To account for an unknown ν, the values of ν_0 for which $\hat{p}_N(CSK_0(\nu_0)|\nu_0)$ exceeds the desired significance level (say α_1) are assembled in a set. This set defines the class of distributions of the form of (18) that are consistent with the data; if this set is empty, then (18) is rejected at level α_1. Details of the algorithm are given in Section A.2.3.

5.2 Multivariate Checks for Serial Dependence and Generalized Autoregressive Conditional Heteroscedasticity

We now present the tests we apply to assess departure from iid errors, specifically, tests against conditional heteroscedasticity and variance ratio tests (see Dufour et al. 2005). The null hypotheses of concern are (17), (18), and (19).

When pursuing a univariate approach, standard diagnostics may be applied to each equation in (2). For instance, the Engle generalized autoregressive conditional heteroscedasticity (GARCH) test statistic (Engle 1982) for equation i, denoted by E_i, is given by T multiplied by the coefficient of determination in the regression of the squared OLS residuals $\hat{\varepsilon}_{it}^2$ on a constant and $\hat{\varepsilon}_{i,t-j}^2$, $j = 1, \ldots, \bar{q}$. The Lee–King test (Lee and King

1993) exploits the one-sided nature of the problem and is based on statistics of the form

$$
\begin{aligned}
LK_i = \Bigg(\Bigg\{ (T-\bar{q}) \sum_{t=\bar{q}+1}^{T} [(\hat{\varepsilon}_{it}^2/\hat{\sigma}_i^2 - 1)] \sum_{j=1}^{\bar{q}} \hat{\varepsilon}_{i,t-j}^2 \Bigg\} \\
\Bigg/ \Bigg\{ \sum_{t=\bar{q}+1}^{T} (\hat{\varepsilon}_{it}^2/\hat{\sigma}_i^2 - 1)^2 \Bigg\}^{1/2} \Bigg) \\
\times \Bigg(\Bigg\{ (T-\bar{q}) \sum_{t=\bar{q}+1}^{T} \Bigg(\sum_{j=1}^{\bar{q}} \hat{\varepsilon}_{i,t-j}^2 \Bigg)^2 \\
- \Bigg(\sum_{t=\bar{q}+1}^{T} \Bigg(\sum_{j=1}^{\bar{q}} \hat{\varepsilon}_{i,t-j}^2 \Bigg) \Bigg)^2 \Bigg\}^{1/2} \Bigg)^{-1}, \qquad (32)
\end{aligned}
$$

where $\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_{it}^2$ and its asymptotic null distribution is standard normal. The variance ratio test statistic, VR_i (Lo and MacKinlay 1988), is

$$
VR_i = 1 + 2\sum_{j=1}^{K}\left(1 - \frac{j}{K}\right)\hat{\rho}_{ij},
$$

$$
\hat{\rho}_{ij} = \frac{\sum_{t=j+1}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{i,t-j}}{\sum_{t=1}^{T} \hat{\varepsilon}_{ti}^2}, \; j = 1, \ldots, K, \qquad (33)
$$

where $VR_i - 1 \overset{\text{asy}}{\sim} \mathrm{N}[0, 2(2K-1)(K-1)/(3K)]$ under the iid null hypothesis.

Such univariate tests may not be appropriate in multivariate regression. Indeed, the error covariance, which appears as a nuisance parameter, is typically not taken into consideration if a series of univariate tests is applied. Furthermore, the problem of combining test decisions over all equations is not straightforward, because the individual tests are not independent (see Shanken 1990). In view of this, we consider the following multivariate modification of these tests (see Dufour et al. 2005). Let $\tilde{W}_{it}$ denote the elements of the *standardized residuals* matrix,

$$
\tilde{\mathbf{W}} = \hat{\mathbf{U}}\mathbf{S}_{\hat{\mathbf{U}}}^{-1}, \qquad (34)
$$

where $\mathbf{S}_{\hat{\mathbf{U}}}$ is the Cholesky factor of $\hat{\mathbf{U}}'\hat{\mathbf{U}}$; that is, $\mathbf{S}_{\hat{\mathbf{U}}}$ is the (unique) upper-triangular matrix such that $\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{S}_{\hat{\mathbf{U}}}'\mathbf{S}_{\hat{\mathbf{U}}}$. We obtain standardized versions of E_i, LK_i, and VR_i (denoted by $\tilde{E}_i$, $\widetilde{LK}_i$, and $\widetilde{VR}_i$) after replacing $\hat{\varepsilon}_{it}$ by $\tilde{W}_{it}$ in the formulas for these statistics. As in Section 5.1, the test criteria from the different equations are then combined through joint statistics of the form

$$
\begin{aligned}
\tilde{E} &= 1 - \min_{1\le i\le n} [p(\tilde{E}_i)], \\
\widetilde{LK} &= 1 - \min_{1\le i\le n} [p(\widetilde{LK}_i)], \qquad (35) \\
\widetilde{VR} &= 1 - \min_{1\le i\le n} [p(\widetilde{VR}_i)],
\end{aligned}
$$

where $p(\tilde{E}_i)$, $p(\widetilde{LK}_i)$, and $p(\widetilde{VR}_i)$ refer to p values that may be obtained by applying a MC test method or using asymptotic null distributions to decrease execution time. In our context, $\tilde{\mathbf{W}}$ has a distribution that is completely determined by the distribution of $\mathbf{W}$ given $\mathbf{X}$, provided that $\mathbf{J}$ [in (16)] is lower triangular (see also Dufour et al. 2003). Consequently, the null distributions of the joint test statistics $\tilde{E}_i$, $\widetilde{LK}_i$, and $\widetilde{VR}_i$ do not depend on $\mathbf{B}$ and $\boldsymbol{\Sigma}$; thus, under (19), MC p values for $\tilde{E}$, $\widetilde{LK}$, and $\widetilde{VR}$ are easy to obtain. Otherwise, we can derive an exact MC p value given $\nu = \nu_0$ (known), which are denoted by $\hat{p}_N(\tilde{E}|\nu_0)$, $\hat{p}_N(\widetilde{LK}|\nu_0)$, and $\hat{p}_N(\widetilde{VR}|\nu_0)$. The unknown ν problem is solved by applying a MMC strategy, we compute

$$
\sup_{\nu_0\in\mathcal{C}(\mathbf{Y})} \hat{p}_N(\tilde{E}|\nu_0), \qquad \sup_{\nu_0\in\mathcal{C}(\mathbf{Y})} \hat{p}_N(\widetilde{LK}|\nu_0),
$$

and

$$
\sup_{\nu_0\in\mathcal{C}(\mathbf{Y})} \hat{p}_N(\widetilde{VR}|\nu_0),
$$

where $\mathcal{C}(\mathbf{Y})$ refers to the same α_1-level confidence set considered for the efficiency test, and refer these MMC p values to a cutoff α_2. This provides exact MMC tests with level $\alpha_1 + \alpha_2$.

5.3 Conditional Heteroscedasticity Under Elliptical Distributions

Finally, we also test for irregularities arising from modeling elliptical returns through distributional assumptions on error terms (as we have proceeded so far), because the latter statistical approach may lead to conditional heteroscedasticity of the following form: The variance of the vector $(r_{1t}, r_{2t}, \ldots, r_{nt})'$ is proportional to a quadratic function of $\tilde{r}_{Mt}$—specifically, the standardized square of the deviation of $\tilde{r}_{Mt}$ from its time series mean, which we denote by $\tilde{z}_{Mt}$ (see Hodgson et al. 2002; Vorkink 2003; Zhou 1993; Kan and Zhou 2003). In multi-beta contexts, $\tilde{z}_{Mt}$ is the tth element of the matrix $\bar{\mathbf{r}}'\hat{\boldsymbol{\Delta}}^{-1}\bar{\mathbf{r}}$, which appears in (25). For example, for the multivariate t with $\kappa \ge 2$, the variance proportionality factor is

$$
\delta_t = (\kappa - 2 + \tilde{z}_{Mt})/(\kappa - 1). \qquad (36)
$$

We proceed as for the GARCH test, using the univariate LM statistic for equation i, which is equal to T multiplied by the coefficient of determination from the regression of the squared OLS residuals on a constant and $\tilde{z}_{Mt}$. This statistic is obtained from the standardized residuals of each equation, leading to n statistics denoted by $\widetilde{BP}_i$, which are combined through the minimum approximate p value as

$$
\widetilde{BP} = 1 - \min_{1\le i\le n} [p(\widetilde{BP}_i)]. \qquad (37)
$$

If heteroscedasticity of the form (36) is accepted as the correct pattern, it is straightforward to correct the efficiency tests described in Section 3 by simply weighting (i.e., dividing) each observation (dependent variables and regressors) with the corresponding value of $\delta_t^{1/2}$. In other words, the model is reestimated by using the corresponding generalized least squares estimator [leading to weighted maximum likelihood (ML)-type or quasi-ML estimators (QMLEs)], which can provide statistical efficiency gains through the use of conditioning information. Some of the results presented in Section 6 use this correction.

6. EMPIRICAL ANALYSIS

Our empirical analysis focuses on unconditional and conditional mean–variance efficiency tests of the market portfolio [formally, tests of (1) in the context of (2), tests of (6) in the context of (5), and tests of (8) in the context of (7)], where the errors follow multivariate normal and Student-t distributions. For the Student distributions, we assume that (16) holds with

$$\mathbf{W}_t = \frac{\mathcal{Z}_{1t}}{(\mathcal{Z}_{2t}/\kappa)^{1/2}}, \tag{38}$$

where $\mathcal{Z}_{1t}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_n$ and $\mathcal{Z}_{2t} \sim \chi^2(\kappa)$ and is independent of $\mathbf{Z}_{1t}$.

We use nominal monthly returns for January 1926–December 1995, obtained from the University of Chicago's CRSP. We formed 12 portfolios of NYSE firms grouped by standard two-digit industrial classification (SIC), as done by Breeden, Gibbons, and Litzenberger (1989). As done by Breeden et al. (1989), we excluded firms with SIC Code 39 (Miscellaneous manufacturing industries) from the dataset for portfolio formation. For each month, the industry portfolios comprise those firms for which the return, the price per common share, and the number of shares outstanding are recorded by the CRSP. Furthermore, portfolios are value-weighted each month. To assess the testable implications of the asset pricing models, we measure the market return by the value-weighted NYSE returns, also available from the CRSP. We measure the risk-free rate by the 1-month Treasury Bill rate, also from the CRSP. The instruments used for our conditional analysis are most prominent in the conditional asset pricing literature (see, e.g., Ferson and Harvey 1999) and include the lagged value of a 1-month Treasury Bill yield, the dividend yield of the Standard & Poor 500 index, the spread between Moody's Baa and Aaa corporate bond yields, the spread between a 10-year and a 1-year Treasury Bond yield, and the difference between the 1-month lagged returns of a 3-month and a 1-month Treasury Bill. Because the instruments are not available before the mid-1960s, we restrict our conditional analysis to the post-1965 period. Our results on efficiency tests are summarized in Tables 1 and 2. All MC tests were applied with 999 replications. The returns for October 1987 and January of every year are excluded from the dataset; the same analysis including these observations yields qualitatively similar results.

Table 1 reports tests of the unconditional CAPM over 5-year subperiods. We also ran the analysis with 10-year subperiods; the results are not significantly affected by such modifications. A notable feature emerges from Table 1: Test decisions (concerning MLR errors and the zero-intercept restriction) vary consistently over time. Such effects are documented in empirical work on the CAPM (see Black 1993; Fama and French 2004). Indeed, temporal instabilities have motivated subperiod model analysis and spurred further research aimed at capturing time-varying betas. Our results, which allow for short time spans, reveal temporal instabilities even when accounting for non-Gaussian errors. Our analysis of the conditional model (discussed later) points out to similar problems in the latter context.

Table 1 reports (in columns 1–3) the p values of the exact multinormality tests based on *ESK*, *EKU*, and *CSK* (see Sec. 5.1). These tests allow us to evaluate whether observed residuals exhibit non-Gaussian behavior through excess skewness and kurtosis. For most subperiods, normality is rejected. These results are interesting, because, although it is well accepted in the finance literature that continuously compounded returns are skewed and leptokurtic, empirical evidence of nonnormality is weaker for monthly data. For instance, Affleck-Graves and McDonald (1989) rejected normality in about 50% of the stocks that they studied. Our results, which are exact (i.e., cannot reject spuriously), indicate much stronger evidence

Table 1. Normality and unconditional efficiency tests

	Normality tests			Efficiency tests						
Sample	*SK*	*KU*	*CSK*	*LR*	p_∞	$p_{\mathcal{N}}$	Q_{U}	$\mathcal{C}(\mathbf{Y})$	$Q_{\mathrm{U}}^{\mathrm{GLS}}$	$\mathcal{C}_{\mathrm{GLS}}(\mathbf{Y})$
1927–1930	**.001**	**.001**	**.001**	16.104	.1866	.364	.357	3–12	.396	3–15
1931–1935	**.001**	**.001**	**.001**	16.257	.1798	.313	.322	3–8	.268	3–9
1936–1940	**.001**	**.001**	**.001**	16.018	.1904	.319	.333	4–26	.483	3–26
1941–1945	**.004**	**.002**	**.004**	25.869	**.0112**	**.045**	.049	≥5	.049	≥4
1946–1950	**.001**	**.001**	**.001**	37.196	**.0002**	**.003**	**.004**	4–26	**.004**	2–24
1951–1955	**.001**	**.002**	**.001**	36.510	**.0003**	**.004**	**.005**	5–31	**.007**	2–33
1956–1960	**.024**	**.003**	**.003**	43.841	**.0000**	**.002**	**.002**	≥5	**.002**	≥2
1961–1965	.594	.479	.631	39.098	**.0001**	**.002**	**.002**	≥7	**.002**	≥4
1966–1970	**.011**	**.002**	**.004**	36.794	**.0002**	**.003**	**.003**	≥5	**.003**	≥3
1971–1975	**.001**	**.002**	**.001**	21.094	**.0490**	.120	.129	4–24	.112	4–30
1976–1980	**.001**	**.001**	**.001**	28.373	**.0049**	**.023**	.026	4–17	**.014**	2–18
1981–1985	**.001**	**.001**	**.001**	27.189	**.0073**	**.033**	.035	5–34	.033	2–30
1986–1990	**.028**	**.020**	**.030**	35.747	**.0007**	**.003**	**.005**	≥5	**.006**	≥2
1991–1995	.177	.311	.239	16.752	.1592	.299	.305	≥15	.293	≥6

NOTE: Numbers in bold indicate test results that are significant at the .05 level. Columns 1–3 report p values for multinormality tests. Columns 1 and 2 pertain to the null hypotheses of no excess skewness and no excess kurtosis in the residuals of each subperiod. The p values in column 3 correspond to the combined statistic CSK designed to jointly test for the presence of skewness and kurtosis; individual and joint tests are obtained by applying (30) and (31) under the assumption of multivariate normal errors in the context of (2). Column 4 presents the quasi-LR statistic defined in (20) to test $\mathcal{H}_E$ defined by (1) in the context of (2); columns 5, 6, and 7 are the associated p values using, respectively, the asymptotic chi-squared distribution, the corresponding (pivotal) MC test obtained under the assumption of multivariate normal errors, and a MMC test assuming a multivariate $t(\kappa)$ error distribution where the p value is maximized over a confidence set for κ with level $1-\alpha_1 = .975$. In the latter case, the maximized p value for the corresponding efficiency test is significant at level .05 if it is not larger than $\alpha_2 = .025$. The confidence set for κ is reported in column 8; see Section 4 for details on its construction. Columns 9 and 10 are the GLS (weighted QMLE) counterparts of 7–8, using the variance weights (36) to correct for heteroscedasticity.

Table 2. Normality and conditional efficiency tests

	Normality tests			Conditional efficiency tests				
Sample	*SK*	*KU*	*CSK*	*LR*	p_∞	$p_\mathcal{N}$	Q_U	$\mathcal{C}(\mathbf{Y})$
Model (7)								
1966–1970	.085	**.017**	**.033**	122.545	**.0002**	.111	.125	≥4
1971–1975	.778	.986	.908	130.384	**.0000**	.057	.067	≥6
1976–1980	.095	.118	.137	147.084	**.0000**	**.012**	**.021**	≥4
1981–1985	.707	.095	.141	155.475	**.0000**	**.004**	**.005**	≥4
1986–1990	.114	**.032**	**.046**	109.736	**.0028**	.300	.344	≥3
1991–1995	.611	.501	.645	113.462	**.0013**	.207	.225	≥6
1966–1995	**.001**	**.001**	**.001**	162.050	**.0000**	**.001**	**.001**	3–16
Model (5)								
1966–1970	.275	.014	.025	34.344	**.0006**	**.011**	**.015**	≥4
1971–1975	.093	.139	.130	26.166	**.0102**	.072	.087	≥5
1976–1980	**.013**	**.002**	**.001**	31.903	**.0014**	**.021**	**.023**	≥4
1981–1985	**.019**	**.024**	**.028**	32.655	**.0011**	**.019**	.026	≥4
1986–1990	**.019**	**.015**	**.028**	31.932	**.0014**	**.020**	**.024**	≥4
1991–1995	.160	.381	.200	17.976	.1164	.338	.347	≥11
1966–1995	**.001**	**.001**	**.001**	39.790	**.0001**	**.001**	**.001**	4–15

NOTE: Numbers in bold indicate test results that are significant at the .05 level. Columns 1–3 report p values for multinormality tests. Columns 1 and 2 pertain to the null hypothesis of no excess skewness and no excess kurtosis in the residuals of each subperiod. The p values in column 3 correspond to the combined statistic CSK designed to jointly test for the presence of skewness and kurtosis; individual and joint tests are obtained by applying (30) and (31) under the assumption of multivariate normal errors, in the context of (5) and (7). Column 4 presents the quasi-LR statistic defined in (20) to test $\overline{\mathcal{H}}_{E1}$ defined by (6) [for model (5)], and $\overline{\mathcal{H}}_{E2}$ defined by (8) [for model (7)]; columns 5, 6, and 7 are the associated p values using, respectively, the asymptotic chi-squared distribution, the corresponding (pivotal) MC test obtained under the assumption of multivariate normal errors, and a MMC test assuming a multivariate $t(\kappa)$ error distribution where the p value is maximized over a confidence set for κ with level $1-\alpha_1 = .975$. In the latter case, the maximized p value for the corresponding efficiency test is significant at level .05 if it is not larger than $\alpha_2 = .025$. The confidence set for κ is reported in column 8; see Section 4 for details on its construction.

against normality. This also confirms the results of Richardson and Smith (1993), who provided evidence against multivariate normality based on asymptotic tests (see also Fiorentini et al. 2003). Of course, this evidence provides further motivation for using our approach to test mean–variance efficiency under non-Gaussian errors.

Columns 4–7 of Table 1 present the LR statistics for unconditional mean–variance efficiency, the corresponding asymptotic p values obtained from the asymptotic $\chi^2(n)$ distribution (p_∞), the exact Gaussian-based MC p values ($p_\mathcal{N}$), and the maximized MC p values based on the Student-t error model (Q_U). Column 8 gives the confidence set $\mathcal{C}(\mathbf{Y})$ for the number of degrees of freedom, κ. These results show that asymptotic p values are quite often spuriously significant (e.g., for 1941–1955), and that the maximal p values exceed the Gaussian-based p value. It is "easier" to reject the testable implications under normality. For instance, at the 5% level of confidence, we find 10 rejections (out of the 14 subperiods) of the null hypothesis for the asymptotic $\chi^2(12)$ test, 9 for the MC p values under normality, and 6 under the Student-t distribution. Under the Student distribution, the tests *jointly* assess the mean–variance efficiency hypothesis and the unknown degrees of freedom parameters in the error distribution. Because the confidence level for the nuisance parameter is .975 ($\alpha_1 = .025$), p values for the efficiency tests should be compared with $\alpha_2 = .025$ to ensure that the overall level of the test is $\alpha = \alpha_1 + \alpha_2 = .05$; see Section 4.

These findings differ from those of Zhou (1993), who found no change in the rejection rates of mean–variance efficiency using elliptical distributions other than the normal. This may be due to the fact that we explicitly take into account nuisance parameter uncertainty (e.g., the fact that the degrees-of-freedom parameter is unknown). Interestingly, whenever the results obtained under non-Gaussian distributions differ from those obtained under the Gaussian distribution, the Gaussian distributional assumption is strongly rejected. Our results clearly indicate that GRS-type tests are sensitive to the hypothesized error distribution. Of course, this observation is relevant when the hypothesized distributions are empirically consistent with the data. Focusing on the t distributions with parameters not rejected by exact GF tests, we see that the decision of the MMC mean–variance efficiency test can change relative to the F-based test.

It is usual to aggregate the efficiency test results over all subperiods, in some manner. For instance, Gibbons and Shanken (1987) proposed two aggregate statistics, which, in terms of our notation, may be expressed as

$$GS_1 = -2\sum_{j=1}^{14} \ln(p_\mathcal{N}[j]) \qquad \text{and}$$
$$GS_2 = \sum_{j=1}^{14} \Psi^{-1}(p_\mathcal{N}[j]), \tag{39}$$

where $[j]$ refers to the subperiods and $\Psi^{-1}(\cdot)$ provides the standard normal deviate corresponding to $p_\mathcal{N}[j]$. If the mean–variance efficiency hypothesis holds across all subperiods, then $GS_1 \sim \chi^2(2 \times 14)$, whereas $GS_2 \sim \mathrm{N}(0, 14)$. It is noteworthy that the same aggregation methods can be applied to our test problem even under (16) by replacing, in (39), $p_\mathcal{N}[j]$ with $Q_{\mathrm{U}[j]}$, the MMC p values obtained imposing (16). Indeed, as is observed by Gibbons and Shanken (1987), the F-distribution is not needed to obtain the null distribution of these combined statistics. All that is needed is a continuous null distribution (a hypothesis satisfied by normal and Student-t errors) and, of course, independence across subperiods. Our results, under normal and Student-t errors, are $GS_1 = 102.264$ and 101.658 and $GS_2 = 28.476$ and 28.397; the associated p values are extremely small. If independence is upheld, as done by Gibbons and Shanken (1987), then this implies that mean–variance efficiency is jointly rejected by our data. If one questions independence and prefers to combine using Bonferroni-based criteria, then the smallest p value is .002, which, when referred to $.025/14 \simeq .002$, comes close to a rejection. In the context of a MC with 999 replications, the smallest possible p values are .001, .002, and so forth. To allow for a fair Bonferroni test, it is preferable to consider the level $.028/14 = .002$. This means that in every period the pretest confidence set should be applied with $\alpha_1 = .022$ to allow .028 to the mean–variance efficiency test. The results reported in the foregoing tables are robust to this change in level.

Finally, Table 3 presents the results of our multivariate exact diagnostic checks for departures from the iid assumption—namely, our proposed multivariate versions of the Engle, Lee–King, and variance ratio tests; we use 12-month lags. The results show very few rejections of the null hypothesis at both the 1% and 5% levels of significance. This implies that in our statistical framework and for the time spans analyzed, iid errors provide an acceptable working assumption. Our heteroscedasticity tests also show that analyzing mean–variance efficiency through elliptical distributional assumptions on the errors is statistically valid in our sample.

An advantage of our methodology is that weighted QMLE-based tests (i.e., tests based on weighted QMLE) may easily be conducted following the methodology that we have described herein, in the context of an MLR weighted by the necessary variance correction term, by, for example, using the variance weights (36) in the case of the multivariate-t (see also Vorkink 2003, footnote 4) as described at the end of Section 5. For illustrative purposes, we report the corrected p values for multivariate t-type tests, in column 9 of Table 1. Our results show that the decision of our tests is not notably affected when we correct for time-varying volatility. It is noteworthy that the latter GLS-based correction does use (in some form) conditioning information.

We now turn to Table 2, which reports our conditional test results for the two models (5) and (7) over 5-year intervals and over the whole sample. We retain the same layout as in Table 1, except of course that the GLS approach is no longer justified and thus is not applied in this context. The companion diagnostic tests are given in Table 4. Although at first glance, the subperiod analysis may appear unnecessary, given that the conditional model is supposed to account for time-varying betas, care must be exercised when interpreting the full-sample test results. From Table 2, we see that for both models (5) and (7): (a) the efficiency hypotheses when assessed using the whole sample are soundly rejected using asymptotic or MC p values, (b) the confidence sets on the degrees-of-freedom parameter appear dramatically tighter, and (c) normality is definitely rejected. Unfortunately, our diagnostic tests (see Table 4) reveal significant departures from the statistical foundations underlying the latter tests (even when allowing for nonnormal errors); thus temporal instabilities cast doubt on the full-sample analysis. The tests in Table 4 are applied in the context of the conditional model (7); because the latter nests model (5) and the unconditional model as well, the results of Table 4 indicate temporal instabilities for all three models.

When we move to subperiod analysis, which appears to be appropriate in the present context, we see that the test results do not differ considerably from the unconditional case. First, asymptotic p values are quite often spuriously significant, particularly in the case of model (7); indeed, as may be seen in Table 2, there is a large difference between the asymptotic and the MC

Table 3. Multivariate diagnostics, unconditional CAPM

	Normal errors			Student-t errors			
Sample	$\tilde{E}$	$\widetilde{LK}$	$\widetilde{VR}$	$\tilde{E}$	$\widetilde{LK}$	$\widetilde{VR}$	$\widetilde{BP}$
1927–1930	**.001**	.356	**.004**	**.013**	.301	**.004**	.285
1931–1935	**.022**	.748	.069	.082	.659	.066	**.016**
1936–1940	.075	.612	.855	.124	.587	.867	.087
1941–1945	.824	.979	.163	.843	.982	.177	.034
1946–1950	**.003**	.804	.063	**.017**	.784	.068	.880
1951–1955	.139	.353	.111	.168	.321	.120	.591
1956–1960	.987	.628	.093	.994	.628	.095	.347
1961–1965	.339	.207	.577	.375	.195	.584	.771
1966–1970	**.027**	.274	.821	.043	.278	.847	.961
1971–1975	.280	.224	.218	.316	.212	.224	**.013**
1976–1980	**.004**	**.011**	.165	**.016**	**.013**	.183	.406
1981–1985	**.027**	.103	.208	.050	.103	.217	.583
1986–1990	**.033**	.453	.346	.077	.442	.366	.279
1991–1995	.803	.236	.088	.821	.252	.092	.585

NOTE: Numbers shown are p values associated with the combined tests $\tilde{E}$, $\widetilde{LK}$, and $\widetilde{VR}$ defined by (35), in the context of model (2). $\tilde{E}$ and $\widetilde{LK}$ are multivariate versions of Engle's and Lee–King's GARCH tests, and $\widetilde{VR}$ is a multivariate version of Lo and MacKinlay's variance ratio tests; see Section 5.2. $\widetilde{BP}$ [defined in (37)] is the conditional heteroscedasticity test as function of the benchmark returns, which is relevant for elliptical nonnormal errors; see Section 5.3. The MC p values in columns 1–3 are based on pivotal statistics, while those in columns 4–7 are MMC p values obtained by maximizing over confidence sets (with level .975) of distributional nuisance parameters. The confidence sets used are those reported in Table 1 (column 8). Numbers in bold indicate test results significant at level .05.

Table 4. Multivariate diagnostics, conditional CAPM

	Normal errors			Student-t errors		
Sample	$\tilde{E}$	$\widetilde{LK}$	$\widetilde{VR}$	$\tilde{E}$	$\widetilde{LK}$	$\widetilde{VR}$
1966–1970	.297	.235	.166	.333	.239	.178
1971–1975	.131	.095	.924	.188	.108	.929
1976–1980	**.012**	.740	.669	**.020**	.744	.683
1981–1985	.137	.108	.628	.172	.110	.629
1986–1990	.264	.766	.932	.338	.767	.933
1991–1995	.878	.178	.473	.878	.184	.495
1966–1975	.331	.083	.417	.348	.087	.425
1976–1985	.290	**.005**	.690	.348	**.008**	.706
1986–1995	**.015**	.647	.190	.038	.639	.207
1966–1995	**.001**	**.001**	.392	**.021**	**.001**	.414

NOTE: Numbers shown are p values associated with the combined tests $\tilde{E}$, $\widetilde{LK}$, and $\widetilde{VR}$, defined by (35), in the context of model (5). $\tilde{E}$ and $\widetilde{LK}$ are multivariate versions of Engle's and Lee–King's GARCH tests, and $\widetilde{VR}$ is a multivariate version of Lo and MacKinlay's variance ratio tests; see Section 5.2. $\widetilde{BP}$ [defined in (37)] is the conditional heteroscedasticity test as function of the benchmark returns, which is relevant for elliptical nonnormal errors. The MC p values in columns 1–3 are based on pivotal statistics, whereas those in columns 4–11 are MMC p values obtained by maximizing over confidence sets (with level .975) of distributional nuisance parameters. The confidence sets used are those reported in Table 2 (column 8). Numbers in bold indicate test results which are significant at the .05 level.

(Gaussian and non-Gaussian) p values. Of course, the number of restrictions tested in this case is 6 per equation (globally, 72 constraints), whereas the problem of testing intercepts involves 12 constraints. Also note that the expanded regression includes 12 regressors for 12 equations, so the number of "effective observations" available for the test is quite small. This observation may suggest that power considerations underlie our observed nonrejections for the shorter subsample, although the simulation studies reported by Dufour and Khalaf (2002b) indicate very good power properties for sample sizes as small as 25 observations even in high-dimensional MLR models. Recall that an F-test (of the GRS type) is unavailable for model (7), so our MC exact test approach is quite useful even given Gaussian errors. Similar considerations hold for the diagnostic tests: simulation results reveal good power for samples of sizes comparable to those used in this article, especially when the system involves a large number of equations (see Dufour et al. 2005).

Second, as in the unconditional case, the Student-t maximal p values exceed the Gaussian p value. For instance, for model (5), at the 5% significance level, we find five rejections (out of the six subperiods) of the null hypothesis for the asymptotic test, four for the MC test under normality, and three under the Student-t distribution. For model (7), at the 5% significance level, we find six rejections (out of the six subperiods) of the null hypothesis for the asymptotic test and two for the MC test under normality and the Student-t distribution. Not surprisingly, in the subperiods in which the conditional models are rejected, the unconditional model is also rejected. In general, model (7) is rejected in fewer subperiods relative to model (5) and the unconditional model (over the 1966–1995 subsample, where the data allow estimation of the conditional models).

In the case of (7), it might be useful to assess the significance of the intercepts only or, alternatively, to assess the contribution of the instruments in explaining excess returns. Interestingly, the MC p values for our test on the intercepts for the six subperiods are .759, .933, .075, .318, .617, and .485 under normality and .771, .946, .080, .339, .645, and .519 given t-errors. We thus see that at the 5% level, our rejections of efficiency are driven by the significance of instruments.

In view of time instabilities, the conditional efficiency test applied to the full sample is unreliable. Thus, to aggregate our subperiod analysis, once again we resort to the combined statistics used by Gibbons and Shanken (1987) as in the unconditional case. Our results, under normal and Student-t errors are [p values are reported in brackets] $GS_1 = 35.572$ [.00038] and 33.006 [.00097] and $GS_2 = 9.052$ [.00011] and 8.415 [.00029] for model (7), and $GS_1 = 39.572$ [.00001] and 37.703 [.00017] and $GS_2 = 10.331$ [.00000] and 9.839 [.00000] for model (5). The latter p values imply that mean–variance efficiency is jointly rejected with our data. Once again, if we question independence and prefer to combine using Bonferroni-based criteria, then the smallest p value for (7) is .004 under normality and .005 with t-errors; the latter, when compared with $.025/6 \simeq 0.004$, comes close to a rejection. Efficiency on the aggregate in model (5) fails to be rejected by the Bonferroni rule. Viewed collectively, our subperiod and aggregate tests indicate that the method used to incorporate conditioning information has nonnegligible implications on mean–variance efficiency.

These results motivate the use of alternative models that capture conditioning information in more parsimonious approaches (i.e., with fewer degrees-of-freedom losses). Inevitably, such approaches, as well as nonlinear stochastic discount factor–based models, will lead to instrumental variable contexts (see the foregoing references on GMM-based tests of the CAPM), for which the literature on exact testing is still scarce.

7. CONCLUSION

In this article we have proposed exact mean–variance efficiency tests in the context of unconditional and conditional CAPMs with Gaussian or non-Gaussian disturbances. We have also shown how to deal with—in finite samples—Student-t errors which may involve unknown parameters. Our empirical results clearly show that the normality assumption does not fit CAPM error returns, even for monthly data. In contrast, Student-t distributions appear to be consistent with the data. Exact unconditional mean–variance efficiency tests, which formally account for nonnormality, fail to reject efficiency for three out of nine subperiods for which Gaussian-based tests are significant. The conditional models analyzed herein provide a better fit, but the efficiency restrictions are rejected for at least half of the six subperiods considered. The conditional results are notably sensitive to the method used to incorporate conditioning information. Overall, although mean–variance efficiency is rejected for several subperiods, using finite-sample methods and allowing for nonnormal errors reduces the number of subperiods for which efficiency is rejected and the strength of the evidence against it.

Although we focused here on mean–variance efficiency tests, it is noteworthy that the proposed methodology applies to several interesting asset pricing tests, including problems in which Hotelling's test (exploited by GRS and MacKinlay 1987) and Rao's F test (see Stewart 1997; Dufour and Khalaf 2002b, the app.) have been used. In view of its fundamental importance, mean–variance efficiency is one of the few MLR-based problems that have been approached from an exact perspective in econometrics, but some authors have recognized that hypotheses dealing with the joint significance of the coefficients of two regression coefficients across equations can also be tested by applying Rao's F test. Examples include intertemporal asset pricing tests by Shanken (1990, footnote 18). Furthermore, as discussed by Shanken (1996), econometric tests of spanning also fall within this class. Indeed, spanning tests (see the survey in DeRoon and Nijman 2001) may be written in terms of a model of GRS form. However, the hypothesis is more restrictive in the sense that, in addition to the restrictions on the intercepts, the betas of each regression must sum to 1. These hypotheses fit into our UL framework. Alternatively, assessing the significance of squared market returns in the context of a three-moment asset pricing model (see, e.g., Barone-Adesi, Gagliardini, and Urga 2004) can be carried out using our framework. The results in this article extend available exact tests of these important financial problems beyond the Gaussian context.

The fact remains that the results presented in this article are specific to UL hypotheses. Not all linear hypotheses may be cast in this form. In earlier work (Beaulieu, Dufour and Khalaf 2005), we studied extensions to nonlinear problems, including tests of mean–variance efficiency in the context of Black's version of the CAPM. Finally, we note that an apparent shortcoming of our exact tests comes from the fact that the right-side benchmark may be observed with errors. The development of exact tests that correct for errors-in-variables problems also appears to be an important issue, and we are currently pursuing research on it.

ACKNOWLEDGMENTS

The authors thank Christian Gouriéroux, Raymond Kan, Blake LeBaron, Mathilda Yared, Guofu Zhou, two anonymous referees, an associate editor, and the editor, as well as seminar participants at the 2000 EC^2 meetings, CREST (Paris), the University of British Columbia, the University of Toronto, the 2001 Canadian Economic Association meetings, CIRANO, and the Deutsche Bundesbank for several useful comments. This work was supported by the Canada Research Chair Program (Chair in Econometrics, Université de Montréal, and Chair in Environmental and Financial Econometric Analysis, Université Laval), the 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 Canada Council for the Arts (Killam Fellowship), 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). This article was also partly written at the Centre de recherche en Économie et Statistique (INSEE, Paris) and the Institut für Wirtschaftsforschung Halle (Germany).

APPENDIX: TESTS

This appendix summarizes the MC test method (given a right-tailed test) as it applies to the test statistics considered in this article. (For proofs and references, see Dufour 2006.)

A.1 Monte Carlo Tests

Let $S(y, \mathbf{X})$ be a test statistic that can be rewritten in the form

$$S(y, \mathbf{X}) = \bar{S}(\mathbf{W}, \mathbf{X}) \tag{A.1}$$

under the null hypothesis, where $\mathbf{W}$ is defined by (16) and the distribution of $\mathbf{W}$ is known. For example, $S(y, \mathbf{X})$ could be the LR statistic considered in Theorem 1. Then the conditional distribution of $S(y, \mathbf{X})$, given $\mathbf{X}$, is completely determined by the matrix $\mathbf{X}$ and the conditional distribution of $\mathbf{W}$ given $\mathbf{X}$; that is, $S(y, \mathbf{X})$ is pivotal. We can then proceed as follows to obtain an exact critical region:

1. Let $S^{(0)}$ be the observed test statistic (based on data).
2. By Monte Carlo methods, draw N iid replications of $\mathbf{W}: \mathbf{W}_{(j)} = [\mathbf{W}_1^{(j)}, \ldots, \mathbf{W}_n^{(j)}], j = 1, \ldots, N$.
3. From each simulated error matrix $\mathbf{W}_{(j)}$, compute the statistics $S^{(j)} = \bar{S}(\mathbf{W}_{(j)}, \mathbf{X}), j = 1, \ldots, N$. For instance, in the case of the QLR statistic underlying Theorem 1, calculate $L(\mathbf{W}_{(j)}) = T \ln(|\mathbf{W}'_{(j)}\mathbf{M}_0\mathbf{W}_{(j)}|/|\mathbf{W}'_{(j)}\mathbf{M}\mathbf{W}_{(j)}|), j = 1, \ldots, N$.
4. Compute the MC p value $\hat{p}_N[S] \equiv p_N(S^{(0)}, S)$, where
$$p_N(x, S) \equiv \frac{NG_N(x, S) + 1}{N + 1} \tag{A.2}$$
and
$$G_N(x, S) \equiv \frac{1}{N}\sum_{j=1}^{N} I_{[0,\infty)}(S^{(j)} - x),$$
$$I_{[0,\infty)}(x) = \begin{cases} 1 & \text{if } x \in [0, \infty) \\ 0 & \text{if } x \notin [0, \infty). \end{cases} \tag{A.3}$$
In other words, $p_N(S^{(0)}; S) = [NG_N(S^{(0)}; S) + 1]/(N + 1)$, where $NG_N(S^{(0)}; S)$ is the number of simulated values that are greater than or equal to $S^{(0)}$. When $S^{(0)}, S^{(1)}, \ldots, S^{(N)}$ are all distinct [an event with probability 1 when the vector $(S^{(0)}, S^{(1)}, \ldots, S^{(N)})'$ has an absolutely continuous distribution], $\hat{R}_N(S^{(0)}) = N + 1 - NG_N(S^{(0)}; S)$ is the rank of $S^{(0)}$ in the series $S^{(0)}, S^{(1)}, \ldots, S^{(N)}$.
5. The MC critical region is $\hat{p}_N[S] \le \alpha$, $0 < \alpha < 1$. If $\alpha(N + 1)$ is an integer and the distribution of S is continuous under the null hypothesis $\mathcal{H}_E$, then, under $\mathcal{H}_E$,
$$\mathsf{P}\big[\hat{p}_N[S] \le \alpha\big] = \alpha. \tag{A.4}$$

The foregoing algorithm is valid for any fully specified distribution of W. Now consider the case in which the distribution of W involves a nuisance parameter as in (16). In this case, given ν, (A.2) yields an MC p value that we denote by $\hat{p}_N[S|\nu]$, where the conditioning on ν is emphasized for further reference. The test defined by $\hat{p}_N[S \mid \nu] \le \alpha$ has size α [in the sense of (A.4)] for known ν. Treating ν as a nuisance parameter, the test based on

$$\sup_{\nu \in \Phi_0} \hat{p}_N[S|\nu] \le \alpha, \tag{A.5}$$

where Φ_0 is a nuisance parameter set consistent with $\mathcal{H}_E$, is *exact at level* α (see Dufour 2006). Note that no asymptotic argument on the number N of MC replications is required to obtain the latter result; this is the fundamental difference between the latter procedure and the (closely related) parametric bootstrap method, which in this context would correspond to a test based on $\hat{p}_N[S|\hat{\nu}_0]$, where $\hat{\nu}_0$ is any point estimate of ν. In earlier work (Dufour and Khalaf 2002b), we call the test based on simulations using a point nuisance parameter estimate a local MC (LMC) test. The term "local" reflects the fact that the underlying MC p value is based on a specific choice for the nuisance parameter. Furthermore, we show that LMC nonrejections are exactly conclusive in the following sense: If $\hat{p}_N[S|\hat{\nu}_0] > \alpha$, then the exact MMC test is clearly not significant at level α.

A.2 MC Skewness and Kurtosis Tests

The algorithm for implementing the MC skewness and kurtosis tests can be decomposed in three wide steps. A more detailed discussion has been given by Dufour et al. (2003).

A.2.1 Estimating Expected Skewness and Kurtosis.

A1. Draw N_0 iid replications, $\bar{\mathbf{W}}_{(i)} = [\bar{\mathbf{W}}_1^{(i)}, \dots, \bar{\mathbf{W}}_n^{(i)}]$, $i = 1, \dots, N_0$, according to the hypothesized distribution with $\nu = \nu_0$.

A2. From each simulated error matrix $\bar{\mathbf{W}}_{(i)}$, compute [see (29)]

$$\mathbf{D}_{(i)} = T\mathbf{M}\bar{\mathbf{W}}_{(i)}\big[\bar{\mathbf{W}}'_{(i)}\mathbf{M}\bar{\mathbf{W}}_{(i)}\big]^{-1}\bar{\mathbf{W}}'_{(i)}\mathbf{M}, \quad i = 1, \dots, N_0, \quad \text{(A.6)}$$

and compute the corresponding statistics SK and KU, applying (28). This provides N_0 replications of the latter statistics, $\overline{SK}_{(i)}$ and $\overline{KU}_{(i)}$, $i = 1, \dots, N_0$.

A3. Compute the average values,

$$\overline{SK}(\nu_0) = \sum_{i=1}^{N_0} \overline{SK}_{(i)}/N_0 \quad \text{and} \quad \overline{KU}(\nu_0) = \sum_{i=1}^{N_0} \overline{KU}_{(i)}/N_0. \quad \text{(A.7)}$$

We call $\overline{SK}(\nu_0)$ and $\overline{KU}(\nu_0)$ the *reference simulated moments*. Two questions arise at this stage: (a) how to obtain exact cutoff points for $ESK(\nu_0)$ and $EKU(\nu_0)$ in (30), and (b) how to obtain a size-correct simultaneous test that combines these two tests. We first address the individual p values issue, which may be run as in Section A.1.

A.2.2 Individual Excess Skewness and Kurtosis Tests.

B1. Using the values $\overline{SK}(\nu_0)$ and $\overline{KU}(\nu_0)$ obtained at steps A1–A3, compute the test statistics based on the observed data, $E^{(0)} = [ESK_{\mathrm{M}}^{(0)}(\nu_0), EKU_{\mathrm{M}}^{(0)}(\nu_0)]'$.

B2. Independent of the data and the draws of steps A1–A3, generate N iid realizations of W according to the hypothesized distribution with $\nu = \nu_0$.

B3. Using the same values $\overline{SK}(\nu_0)$ and $\overline{KU}(\nu_0)$ as for the observed data, compute the statistics $ESK_{\mathrm{M}}(\nu_0)$ and $EKU_{\mathrm{M}}(\nu_0)$ associated with each one of these MC samples: $E^{(j)} = [ESK_{\mathrm{M}}^{(j)}(\nu_0), EKU_{\mathrm{M}}^{(j)}(\nu_0)]'$, $j = 1, \dots, N$. It is easy to see that the $N+1$ vectors $E^{(j)}$, $j = 0, 1, \dots, N$, are exchangeable under the null hypothesis.

B4. Compute a simulated p value for any one of the test statistics in $E^{(0)}$: $\hat{p}_N[ESK_{\mathrm{M}}(\nu_0)]$, $\hat{p}_N[EKU_{\mathrm{M}}(\nu_0)]$, where $\hat{p}_N[\cdot]$ is defined in Section A.1 for each statistic in E [see (A.2)]. The null hypothesis is rejected at level α by the test $ESK_{\mathrm{M}}(\nu_0)$ if $\hat{p}_N[ESK_{\mathrm{M}}(\nu_0)] \le \alpha$, and similarly for $EKU_{\mathrm{M}}(\nu_0)$. By the exchangeability of $E^{(j)}$, $j = 0, 1, \dots, N$, and provided that E follows a continuous distribution, this procedure satisfies the size constraint, that is,

$$\mathsf{P}\big[\hat{p}_N[ESK_{\mathrm{M}}(\nu_0)] \le \alpha\big] = \mathsf{P}\big[\hat{p}_N[EKU_{\mathrm{M}}(\nu_0)] \le \alpha\big] = \alpha, \quad \text{(A.8)}$$

under the null hypothesis.

A.2.3 Combined Excess Skewness and Kurtosis Test.

C1. Generate a set of reference simulated moments (according to A1–A3), the observed value of $E^{(0)}$ (according to B1), and the N corresponding simulated statistics.

C2. For each test statistic considered, obtain the p value functions determined by simulated statistics (generated at step C1), $p_N(S^{(0)}; S)$, for $S = ESK_{\mathrm{M}}(\nu_0)$, $EKU_{\mathrm{M}}(\nu_0)$, where the function $p_N(S^{(0)}; S)$ is defined in Section A.1.

C3. Independent of the previous simulations and the data, generate N_1 additional iid realizations of W according to the hypothesized distribution with $\nu = \nu_0$. Choose N_1 so that $\alpha(N_1 + 1)$ is an integer.

C4. Using the reference simulated values and the N_1 draws generated at steps C1 and C3, compute the corresponding simulated statistics $EE^{(l)} = [ESK_{\mathrm{M}}^{(l)}(\nu_0), EKU_{\mathrm{M}}^{(l)}(\nu_0)]'$, $l = 1, \dots, N_1$.

C5. Using the p value functions $p_N(\cdot; \cdot)$ obtained at step C2, evaluate the simulated p values for the observed and the N_1 additional simulated statistics, $\hat{p}_N^{(l)}[S] = p_N(S^{(l)}; S)$, $l = 0, 1, \dots, N_1$, for $S = ESK_{\mathrm{M}}(\nu_0)$, $EKU_{\mathrm{M}}(\nu_0)$.

C6. From the latter, compute the corresponding values of the combined test statistics

$$CSK_{\mathrm{M}}^{(l)}(\nu_0) = 1 - \min\big\{\hat{p}_N\big[ESK_{\mathrm{M}}^{(l)}(\nu_0)\big], \hat{p}_N\big[EKU_{\mathrm{M}}^{(l)}(\nu_0)\big]\big\}, \quad l = 0, 1, \dots, N_1. \quad \text{(A.9)}$$

Again, it is easy to see that the vectors $CSK_{\mathrm{M}}^{(l)}(\nu_0)$, $l = 0, 1, \dots, N_1$, are exchangeable.

C7. The combined test $CSK_{\mathrm{M}}(\nu_0)$ rejects the null hypothesis at level α if $\hat{p}_{N_1}[CSK_{\mathrm{M}}(\nu_0)] \equiv p_{N_1}(CSK_{\mathrm{M}}^{(0)}; CSK_{\mathrm{M}}(\nu_0)) \le \alpha$, where the p value function $p_{N_1}(\cdot|\cdot)$ is based on the simulated variables $CSK_{\mathrm{M}}^{(l)}(\nu_0)$, $l = 0, 1, \dots, N_1$.

This test has level α because the variables $CSK_{(l)}$, $l = 0, 1, \dots, N$, are exchangeable under the null hypothesis.

[Received July 2004. Revised March 2006.]

REFERENCES

Affleck-Graves, J., and McDonald, B. (1989), "Nonnormalities and Tests of Asset Pricing Theories," *Journal of Finance*, 44, 889–908.

Barone-Adesi, G., Gagliardini, P., and Urga, G. (2004), "Testing Asset Pricing Models With Coskewness," *Journal of Business & Economic Statistics*, 22, 474–485.

Beaulieu, M.-C., Dufour, J.-M., and Khalaf, L. (2005), "Testing Black's CAPM With Possibly Non-Gaussian Errors: An Exact Identification-Robust Simulation-Based Approach," technical report, CIRANO and CIREQ, University of Montréal.

Berk, J. B. (1997), "Necessary Conditions for the CAPM," *Journal of Economic Theory*, 73, 245–257.

Black, F. (1993), "Beta and Return," *Journal of Portfolio Management*, 20, 8–17.

Breeden, D. T., Gibbons, M., and Litzenberger, R. H. (1989), "Empirical Tests of the Consumption Based CAPM," *Journal of Finance*, 44, 231–262.

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

Cochrane, J. H. (2001), *Asset Pricing*, Princeton, NJ: Princeton University Press.

DeRoon, F. A., and Nijman, T. E. (2001), "Testing for Mean–Variance Spanning: A Survey," *Journal of Empirical Finance*, 8, 111–155.

Dufour, J.-M. (1990), "Exact Tests and Confidence Sets in Linear Regressions With Autocorrelated Errors," *Econometrica*, 58, 475–494.

——— (2006), "Monte Carlo Tests With Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics," *Journal of Econometrics*, 133, 443–477.

Dufour, J.-M., and Khalaf, L. (2002a), "Exact Tests for Contemporaneous Correlation of Disturbances in Seemingly Unrelated Regressions," *Journal of Econometrics*, 106, 143–170.

——— (2002b), "Simulation Based Finite- and Large-Sample Tests in Multivariate Regressions," *Journal of Econometrics*, 111, 303–322.

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.

——— (2005), "Multivariate Residual-Based Finite-Sample Tests for Serial Dependence and GARCH With Applications to Asset Pricing Models," technical report, CIRANO and CIREQ, University of Montréal.

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, 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.

Engle, R. F. (1982), "Autoregressive Conditional Heteroscedasticity With Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987–1008.

Fama, E. F. (1965), "The Behaviour of Stock Prices," *Journal of Business*, 60, 401–424.

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.

——— (2004), "The Capital Asset Pricing Model: Theory and Evidence," *The Journal of Economic Perspectives*, 18, 25–46.

Ferson, W. E., and Harvey, C. R. (1999), "Conditioning Variables and the Cross-Section of Stock Returns," *The Journal of Finance*, 54, 1325–1360.

Fiorentini, G., Sentana, E., and Calzolari, G. (2003), "Maximum Likelihood Estimation and Inference in Multivariate Conditionally Heteroskedastic Dynamic Models With Student *t* Innovations," *Journal of Business & Economic Statistics*, 21, 532–546.

Gibbons, M. R. (1982), "Multivariate Tests of Financial Models: A New Approach," *Journal of Financial Economics*, 10, 3–27.

Gibbons, M. R., Ross, S. A., and Shanken, J. (1989), "A Test of the Efficiency of a Given Portfolio," *Econometrica*, 57, 1121–1152.

Gibbons, M. R., and Shanken, J. (1987), "Subperiod Aggregation and the Power of Multivariate Tests of Portfolio Efficiency," *Journal of Financial Economics*, 19, 389–394.

Groenwold, N., and Fraser, P. (2001), "Tests of Asset-Pricing Models: How Important Is the iid-Normal Assumption," *Journal of Empirical Finance*, 8, 427–449.

Hansen, L. P., and Richard, S. F. (1987), "The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models," *Econometrica*, 55, 587–613.

Hodgson, D. J., Linton, O., and Vorkink, K. (2002), "Testing the Capital Asset Pricing Model Efficiently Under Elliptical Symmetry: A Semiparametric Approach," *Journal of Applied Econometrics*, 17, 617–639.

Hodgson, D. J., and Vorkink, K. (2003), "Efficient Estimation of Conditional Asset Pricing Models," *Journal of Business & Economic Statistics*, 21, 269–283.

Ingersoll, J. (1987), *Theory of Financial Decision Making*, Totowa, NJ: Rowman & Littlefield.

Jagannathan, R., and Wang, Z. (1996), "The Conditional CAPM and the Cross-Section of Expected Returns," *The Journal of Finance*, 51, 3–53.

Jobson, J., and Korkie, B. (1982), "Potential Performance and Tests of Portfolio Efficiency," *Journal of Financial Economics*, 10, 433–466.

Kan, R., and Zhou, G. (2003), "Modeling Non-Normality Using Multivariate *t*: Implications for Asset Pricing," technical report, Rotman School of Management, University of Toronto.

Kandel, S., McCulloch, R., and Stambaugh, R. F. (1995), "Bayesian Inference and Portfolio Efficiency," *Review of Financial Studies*, 8, 1–53.

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 & Economic Statistics*, 11, 17–27; corr. 12 (1994), 139.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Wiley.

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.

MacKinlay, A. C. (1987), "On Multivariate Tests of the Capital Asset Pricing Model," *Journal of Financial Economics*, 18, 341–372.

——— (1995), "Multifactor Models Do Not Explain Deviations From the Capital Asset Pricing Model," *Journal of Financial Economics*, 38, 3–28.

MacKinlay, A. C., and Richardson, M. P. (1991), "Using Generalized Method of Moments to Test Mean–Variance Efficiency," *Journal of Finance*, 46, 511–527.

Mardia, K. V. (1970), "Measures of Multivariate Skewness and Kurtosis With Applications," *Biometrika*, 57, 519–530.

Richardson, M., and Smith, T. (1993), "A Test for Multivariate Normality in Stock Returns," *Journal of Business*, 66, 295–321.

Shanken, J. (1990), "Intertemporal Asset Pricing: An Empirical Investigation," *Journal of Econometrics*, 45, 99–120.

——— (1996), "Statistical Methods in Tests of Portfolio Efficiency: A Synthesis," in *Handbook of Statistics 14: Statistical Methods in Finance*, eds. G. S. Maddala and C. R. Rao, Amsterdam: North-Holland, pp. 693–711.

Stewart, K. G. (1997), "Exact Testing in Multivariate Regression," *Econometric Reviews*, 16, 321–352.

Tu, J., and Zhou, G. (2004), "Data-Generating Process Uncertainty: What Difference Does It Make in Portfolio Decisions," *Journal of Financial Economics*, 72, 385–421.

Vorkink, K. (2003), "Return Distributions and Improved Tests of Asset Pricing Models," *Review of Financial Studies*, 16, 845–874.

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