Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: an exact simulation-based approach*

Marie-Claude Beaulieu † Université Laval

Jean-Marie Dufour ‡ Université de Montréal

Lynda Khalaf § Carleton University

First version: March 2002
Revised: June 2004, July 2005, February 2006, May 2006
This version: February 19, 2008
Compiled: February 19, 2008, 11:52pm

This paper has been published in the *Journal of Business and Economic Statistics*, 25 (2007), 4, 398-410.

*The authors thank Christian Gouriéroux, Raymond Kan, Blake LeBaron, Mathilda Yared, Guofu Zhou, two anonymous referees, an Associate Editor, the Editor Torben Andersen, 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 paper was also partly written at the Centre de recherche en Économie et Statistique (INSEE, Paris) and the Institut für Wirtschaftsforschung Halle (Germany).

† CIRANO and Département de finance et assurance, Université Laval. Mailing address: Département de finance et assurance, Pavillon Palasis-Prince, Université Laval, Ste-Foy, Québec, Canada G1K 7P4. TEL: 1 (418) 656-2926, FAX: 1 (418) 656-2624; e-mail: Marie-Claude.Beaulieu@fas.ulaval.ca

‡ 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

§ Canada Research Chair in Environmental and Financial Econometric Analysis (Université Laval), 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

ABSTRACT

In this paper, we propose exact likelihood-based mean-variance efficiency tests of the market portfolio in the context of the capital asset pricing model (CAPM), allowing for a wide class of error distributions which include normality and multivariate t as special cases. Both unconditional and conditional versions of the CAPM are considered. These tests are developed in the framework of multivariate linear regressions (MLR). It is well known that, despite their simple statistical structure, standard asymptotically justified MLR-based tests are unreliable in finite samples. In financial econometrics, exact tests have been proposed only for a few specific hypotheses [Jobson and Korkie (Journal of Financial Economics, 1982), MacKinlay (Journal of Financial Economics, 1987), Gibbons, Ross and Shanken (Econometrica, 1989)], most of which depend on normality. For the Gaussian unconditional model, our tests correspond to Gibbons, Ross and Shanken's mean-variance efficiency tests. Our framework casts more evidence on whether the normality assumption is too restrictive when testing the CAPM. We also apply exact multivariate goodness-of-fit tests and diagnostic checks, and obtain a set estimate for the intervening nuisance parameters. Our results show the following: (i) multivariate normality is rejected; (ii) multivariate residual checks suggest temporal instabilities, for both the unconditional and the conditional models; (iii) although mean-variance efficiency is rejected over several subperiods, using finite-sample methods and allowing for non-normal errors reduces the number of subperiods for which efficiency is rejected and the strength of the evidence against it; (iv) the use of conditioning information has non-negligible effects on tests of mean-variance efficiency and substantially reduces the number of rejections.

Key words: capital asset pricing model; CAPM; conditional CAPM; mean-variance efficiency; non-normality; multivariate linear regression; uniform linear hypothesis; exact test; Monte Carlo test; bootstrap; nuisance parameters; specification test; diagnostics; GARCH; variance ratio test.

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

RÉSUMÉ

Dans cet article nous proposons des tests exacts, basés sur la vraisemblance, de l'efficience du portefeuille de marché dans l'espace moyenne-variance. Ces tests, utilisés ici dans le contexte du modèle du CAPM (Capital Asset Pricing Model), permettent de considérer diverses classes de distributions incluant la loi normale. Les tests sont développés dans le cadre de modèles de régression linéaires multivariés (RLM). Il est, par ailleurs, bien établi que, malgré leur structure simple, les écart-types et tests usuels asymptotiques de ces modèles ne sont pas fiables. En économétrie financière, des tests en échantillons finis ont été proposés seulement pour quelques hypothèses spécifiques, lesquels dépendent pour la plupart de l'hypothèse de normalité [Jobson et Korkie (Journal of Financial Economics, 1982), MacKinlay (Journal of Financial Economics, 1987), Gibbons, Ross et Shanken (Econometrica, 1989)]. Dans le contexte gaussien, nos tests d'efficience correspondent à ceux de Gibbons, Ross et Shanken. Dans un contexte non-gaussien, nous reconsidérons l'efficience moyenne-variance du portefeuille de marché en permettant des distributions multivariées de Student et des « mélanges de lois normales ». Notre démarche nous permet d'évaluer si l'hypothèse de normalité est trop restrictive lorsque l'on teste le CAPM. Nous appliquons aussi des tests diagnostiques multivariés (incluant des tests pour les effets d'hétéroscédasticité conditionnelle et une généralisation multivariée des tests de ratio de variance), des tests de spécification, et nous obtenons aussi un estimateur ensembliste pour les paramètres de nuisance pertinents. Nos résultats montrent que: (i) l'hypothèse de normalité multivariée est rejetée sur la plupart des sous-périodes; (ii) les tests diagnostiques suggèrent la présence de changement structurel à la fois pour le modèle inconditionnel et le modèle conditionel; (iii) bien que l'hypothèse d'efficience du portefeuille de marché soit rejetée sur plusieurs sous-périodes, l'utilisation de méthodes exactes et la prise en compte de la non-normalité des erreurs réduit le nombre de périodes pour lesquelles l'efficience est rejetée ainsi que la force de l'évidence contre celle-ci; (iv) l'utilisation de variables conditionnantes a des effets notables sur les tests d'efficience et réduit substantiellement le nombre des rejets.

Mots-clefs: modèle d'évaluation d'actifs financiers; CAPM; efficience de portefeuille; non-normalité; modèle de régression multivarié; hypothèse linéaire uniforme; test exact; test de Monte Carlo; bootstrap; paramètres de nuisance; test de spécification; tests diagnostiques; GARCH; test de ratio des variances.

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

Contents

List of Propositions and Theorems

List of Tables

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), and 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, Section 3.4.2), Campbell et al. (1997, Chapter 5), and Dufour and Khalaf (2002*b*). 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 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), Gibbons, Ross and Shanken (1989, henceforth GRS), Stewart (1997), MacKinlay (1995) and Kandel, McCulloch and Stambaugh (1995). These methods typically require Gaussian distributional assumptions. However, it has long been recognized that financial returns exhibit non-normalities [Fama (1965)]. Though 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) or Berk (1997)], finite-sample tests for mean-variance efficiency in non-Gaussian CAPM's are unavailable as yet.

Indeed, mean-variance efficiency tests which relax normality include: (i) 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)]; (ii) semiparametric asymptotic procedures specific to elliptical distributions [Hodgson, Linton and Vorkink (2002), Vorkink (2003), Hodgson and Vorkink (2003)]; (iii) parametric procedures based on postulating a non-Gaussian distribution, such as the multivariate t [Fiorentini, Sentana and Calzolari (2003), Zhou (1993)]; (iv) non-Gaussian Bayesian procedures [Tu and Zhou (2004)]. In all 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) report size problems on high-dimensional systems and restrict their analysis to systems with 3 or 4 portfolios, while Vorkink (2003) proceeds on a portfolio-by-portfolio basis, not the whole system. In the parametric case, Zhou (1993) proposes simulation-based p-values for the GRS statistic given a few elliptical distributions, while selecting their tail area parameter by trial and error.

In this paper, we propose finite-sample unconditional and conditional multivariate mean-variance efficiency tests in possibly non-Gaussian CAPM's. The conditional specifications allow model coefficients to vary as functions of a number of instruments, as described by Shanken (1996, section 2.3.4), Cochrane (2001, Chapter 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) and Cochrane (2001, Chapter 8)].

We use finite-sample results from Dufour and Khalaf (2002*b*) 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 which include the Gaussian and a wide spectrum of non-Gaussian distributions, both elliptically symmetric and non-elliptical. 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 by Dufour and Khalaf (2002*b*)], 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; (2) we 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 cut-off provides a test with exact level $\alpha_1 + \alpha_2$ [see Dufour and Kiviet (1996) and Dufour (2006)]. We stick here to the original notion of test level in the presence of nuisance parameters [Lehmann (1986, Chapter 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: (1) GF tests on the error distribution, and (2) serial dependence tests. All procedures rely on properly standardized multivariate ordinary least squares (OLS) residuals, which provides statistics *invariant* to MLR coefficients and error variances and covariances; this allows an 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 non-normal) 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 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 heteroskedasticity linked to conditioning on market returns [Vorkink (2003)]. Our exact combination method relies on simulation [as in Dufour and Khalaf (2002*a*) 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 non-normal error distributions.

The tests proposed are applied to an unconditional and a conditional CAPM with observable risk-free rates, and multivariate normal as well as 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) data base (1926-1995). Our results show the following: (i) multivariate normality is rejected; (ii) multivariate residual checks suggest temporal instabilities, for both the unconditional and the conditional models; (iii) although mean-variance efficiency is rejected over several subperiods, using finite-sample methods and allowing for non-normal errors reduces the number of subperiods for which efficiency is rejected and the strength of the evidence against it; (iv) the use of conditioning information has non-negligible effects on tests of mean-variance efficiency and substantially reduces the number of rejections.

The paper is organized as follows. Section 2 sets the framework. In Section 3, we describe existing tests and propose extensions for non-normal distributions. In Section 4, we discuss how to deal with nuisance parameters in the error distribution. GF and diagnostic tests are described in Section 5. In Section 6, we report the empirical results. We conclude in Section 7.

2. Framework

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

$$\mathcal{H}_E : a_i = 0\,, \quad i = 1, \ldots, n, \tag{2.1}$$

in the context of the MLR model

$$r_{it} = a_i + \beta_i \tilde{r}_{\mathrm{M}t} + \varepsilon_{it}\,, \quad t = 1, \ldots, T,\ \ i = 1, \ldots, n, \tag{2.2}$$

where $r_{it} = R_{it} - R_{ft},\ \ \tilde{r}_{\mathrm{M}t} = \tilde{R}_{\mathrm{M}t} - R_{ft}$, 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 in Shanken (1996, section 2.3.4), Cochrane (2001, Chapter 8) and DeRoon and Nijman (2001), this can be accommodated by 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 $z_{1t}, \ldots, z_{qt}$: depending on whether only the *betas* or both the intercepts and the betas are allowed to vary, this leads to alternative specifications:

$$r_{it} = \bar{a}_i + \beta_{it} \tilde{r}_{\mathrm{M}t} + \varepsilon_{it}\,, \quad \beta_{it} = \bar{\beta}_i + \sum_{j=1}^{q} d_{ji} z_{jt}\,, \tag{2.3}$$

$$r_{it} = a_{it} + \beta_{it} \tilde{r}_{\mathrm{M}t} + \varepsilon_{it}\,, \quad a_{it} = \bar{a}_i + \sum_{j=1}^{q} c_{ji} z_{jt}\,, \quad \beta_{it} = \bar{\beta}_i + \sum_{j=1}^{q} d_{ji} z_{jt}\,, \tag{2.4}$$

$t = 1, \ldots, T,\ \ i = 1, \ldots, n$. Model (2.3) entails the expanded regression

$$r_{it} = \bar{a}_i + \bar{\beta}_i \tilde{r}_{\mathrm{M}t} + \sum_{j=1}^{q} d_{ji} \left(\tilde{r}_{\mathrm{M}t} z_{jt}\right) + \varepsilon_{it}\,, \quad t = 1, \ldots, T,\ \ i = 1, \ldots, n. \tag{2.5}$$

On assuming that the regressor matrix has full rank for each i, efficiency can be assessed by testing

$$\bar{\mathcal{H}}_{E1} : \bar{a}_i = 0\,, \quad i = 1, \ldots, n. \tag{2.6}$$

Similarly, (2.4) leads to the following equation:

$$r_{it} = \bar{a}_i + \bar{\beta}_i \tilde{r}_{\mathrm{M}t} + \sum_{j=1}^{q} c_{ji} z_{jt} + \sum_{j=1}^{q} d_{ji} \left(\tilde{r}_{\mathrm{M}t} z_{jt}\right) + \varepsilon_{it}\,, \quad t = 1, \ldots, T,\ \ i = 1, \ldots, n. \tag{2.7}$$

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

$$\bar{\mathcal{H}}_{E2} : \bar{a}_i = 0 \,, \quad c_{ji} = 0, \quad i = 1, \ldots , n, \quad j = 1, \ldots , q. \tag{2.8}$$

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

$$\iota_T = (1, \ldots , 1)' \,, \quad \tilde{r}_{\mathrm{M}} = (\tilde{r}_{1\mathrm{M}}, \ldots , \tilde{r}_{T\mathrm{M}})', \quad r_i = (r_{1i}, \ldots , r_{Ti})' \,, \tag{2.9}$$

$$z = [z_1, \ldots , z_q] \,, \quad z_j = (z_{j1}, \ldots , r_{jT})' \,, \quad j = 1, \ldots , q. \tag{2.10}$$

We also use the $*$ symbol to denote element by element row-wise matrix multiplication; for example, if $A = [A_1, \ldots , A_T]'$ is a $T \times l$ matrix and $D = [D_1, \ldots , D_T]'$ is an $T \times m$ matrix, then $A * D$ is the $T \times (lm)$ matrix with t-th row equal to $A_t' \otimes D_t'$, *i.e.* $A * D = [A_1 \otimes D_1, \ldots , A_T \otimes D_T]'$.

The foregoing models are special cases of the MLR model:

$$Y = XB + U \tag{2.11}$$

where $Y = [Y_1, \ldots , Y_n]$ is a $T \times n$ matrix of dependent variables, X is a $T \times k$ full-column rank matrix of regressors, and $U = [U_1, \ldots , U_n] = [V_1, \ldots , V_T]'$ 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, *i.e.* they have the form:

$$\mathcal{H}_0 : HB = \mathcal{O}(h,\, n) \tag{2.12}$$

where H is a fixed $h \times k$ matrix of rank h. Indeed, (2.1) - (2.2), (2.5) - (2.6) and (2.7) - (2.8) each constitute a special case of (2.11) - (2.12) obtained by taking respectively one of the following definitions:

$$Y = [r_1, \ldots , r_n], \quad X = [\iota_T, \tilde{r}_{\mathrm{M}}] \,, \quad H = (1,\, 0) \,, \tag{2.13}$$

$$Y = [r_1, \ldots , r_n] \,, \quad X = [\iota_T, \tilde{r}_{\mathrm{M}}, \tilde{r}_{\mathrm{M}} * z] \,, \quad H = [1,\ \mathcal{O}(1, q+1)] \,, \tag{2.14}$$

$$Y = [r_1, \ldots , r_n] \,, \quad X = [\iota_T,\ z,\ \tilde{r}_{\mathrm{M}},\ \tilde{r}_{\mathrm{M}} * z] \,, \quad H = [I_{q+1},\ \mathcal{O}(q+1, q+1)]. \tag{2.15}$$

In this context, we apply a formal statistical approach to obtain simple finite-sample tests under alternative error distributions (assuming we can condition on X, *i.e.* we can take X as fixed for statistical analysis). More precisely, we consider the general case:

$$V_t \equiv (\varepsilon_{1t}, \ldots , \varepsilon_{nt})' = JW_t \;,\;\; t = 1, \ldots ,\; T \,, \tag{2.16}$$

where J is an unknown, non-singular matrix and the distribution of the vector $w = vec(W)$, $W = [W_1, \ldots , W_T]'$ is either: (i) known (hence, free of nuisance parameters), or (ii) specified up to an unknown finite dimensional nuisance-parameter (denoted ν); we call w the vector of *normalized disturbances* and its distribution the *normalized disturbance distribution*. Let $\Sigma = JJ'$ so $\det(\Sigma) \neq 0$. For example, we assume that $W_t \sim \mathcal{F}(\nu) \,,\;\; t = 1, \ldots ,\; T$, where $\mathcal{F}(.)$ represents a known distribution function. Below, we consider both the case where the error distribution does not involve nuisance parameters,

$$W_t \sim \mathcal{F}(\nu_0) \text{ where } \nu_0 \text{ is specified,} \tag{2.17}$$

and the one where it does,

$$W_t \sim \mathcal{F}(\nu) \text{ where } \nu \text{ is unknown.} \tag{2.18}$$

This assumption includes as special cases the Gaussian distribution,

$$V_1, \ldots, V_T \overset{i.i.d.}{\sim} N[0, \Sigma], \tag{2.19}$$

all elliptically symmetric distributions, such as the multivariate t, and cases where $W_1, \ldots, W_T$ are independent identically distributed (*i.i.d.*) according to any given non-elliptical distribution. In this regard, conditioning on further instruments [as in models (2.5) and (2.7)] – rather than only the market portfolio – can make the *i.i.d.* 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 (2.1), (2.6) or (2.8)] under the distributional assumption (2.16). The test statistics used are Gaussian likelihood ratios:

$$LR = T\,\ln(\Lambda),\ \ \Lambda = |\hat{\Sigma}_0|/|\hat{\Sigma}|, \quad \hat{\Sigma} = \hat{U}'\hat{U}/T, \quad \hat{\Sigma}_0 = \hat{U}_0'\hat{U}_0/T\,, \tag{3.20}$$

$$\hat{U} = Y - X\hat{B}, \quad \hat{B} = (X'X)^{-1}X'Y, \quad \hat{U}_0 = Y - X\hat{B}_0\,, \tag{3.21}$$

$$\hat{B}_0 = \hat{B} - \left(X'X\right)^{-1} H'\big[H\left(X'X\right)^{-1} H'\big]^{-1} H\hat{B}\,, \tag{3.22}$$

where Y, X and H are defined as in (2.13), (2.14) or (2.15), depending on the null hypothesis ($\mathcal{H}_E$, $\bar{\mathcal{H}}_{E1}$ or $\bar{\mathcal{H}}_{E2}$). On using the results in Dufour and Khalaf (2002*b*, section 3 and Appendix), the null distribution of LR can be characterized as follows (under a possibly non-Gaussian error distribution).

Theorem 3.1 NULL DISTRIBUTION OF GAUSSIAN LR STATISTICS FOR MEAN-VARIANCE EFFICIENCY. *Under* (2.16)*, the LR statistic defined in* (3.20) *for testing $\mathcal{H}_E$ against the* (*unrestricted*) *model* (2.2) [resp., $\bar{\mathcal{H}}_{E1}$ *against* (2.5)*, or and* $\bar{\mathcal{H}}_{E2}$ *against* (2.7)] *is distributed like*

$$L(W) \equiv T\,\ln\left(|W'M_0W|/|W'MW|\right)$$

under the null hypothesis, where $M_0 = M + X(X'X)^{-1}H'[H(X'X)^{-1}H']^{-1}H(X'X)^{-1}X'$, $M = I - X(X'X)^{-1}X'$, *and H is defined in* (2.13) [resp., *in* (2.14) *and* (2.15)]. *If furthermore the Gaussian assumption* (2.19) *holds and* $T - k - n + 1 \geq 1$, *where k is the number of columns in 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*

$$\frac{\rho\tau - 2\lambda}{n(q+1)}(\Lambda^{1/\tau} - 1) \sim F[n(q+1),\, \rho\tau - 2\lambda] \tag{3.23}$$

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*

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

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

$$r_{it} = a_i + \sum_{j=1}^{s} \beta_{ji}\tilde{r}_{jt} + \varepsilon_{it}\ , \quad t = 1,\ \ldots\ ,\ T,\ \ i = 1,\ \ldots\ ,\ n, \tag{3.24}$$

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 under test entails that there is a portfolio of the benchmark portfolios that is mean-variance efficient [see Gibbons et al. (1989, henceforth GRS)]. Unconditional efficiency tests follow from Theorem **3.1** with $X = [\iota_T,\ \tilde{r}]$, $\tilde{r} = [\ \tilde{r}_1,\ \ldots\ ,\ \tilde{r}_s]$, $\tilde{r}_j = (\tilde{r}_{1j},\ \ldots\ ,\ \tilde{r}_{Tj})'$, $j = 1,\ \ldots\ ,\ s$, and $H = [1,\ \mathcal{O}(1,\ s)]$. Furthermore, conditional efficiency tests are covered by Theorem **3.1** in the context of an expanded MLR of the form (2.11), where $Y = [r_1,\ \ldots\ ,\ r_n]$, with: $X = [\iota_T,\ \tilde{r},\ \tilde{r} * z]$ and $H = [1,\ \mathcal{O}(1,\ qs+s)]$ if the coefficients of $\tilde{r}$ are assumed to be linear functions of the instruments, $X = [\iota_T,\ z,\ \tilde{r},\ \tilde{r} * z]$ and $H = [I_{q+1},\ \mathcal{O}(q+1,\ qs+s)]$ if the intercepts also are linear functions of the instruments.

Theorem **3.1** entails that the distribution of the LR statistic in (3.20) does not depend on B and Σ. This property holds under conditions much more general than elliptical symmetry (which was emphasized in the earlier econometric literature on the CAPM). So, given draws from the distribution of the disturbance matrix $W = [W_1,\ \ldots\ ,\ W_T]$, an exact p-value may be obtained using the MC test technique as follows:

1. using the distributional assumption (2.16) and a given value of ν [as in (2.17)], generate N *i.i.d.* replications of the disturbance matrix W;

2. this yields N simulated values of the test statistic, applying the relevant pivotal transform $L(W)$ from Theorem **3.1**;

3. the exact Monte Carlo p-value is then calculated from the rank of the observed LR relative to the simulated ones [see (A.2) in Appendix A].

Further details are supplied in Appendix A. By the general theory of Monte Carlo tests, we observe that the size of a simulation-based test can be *perfectly controlled* even with a very small number of Monte Carlo 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 a number of replications larger than 100 or 200 is typically very small; for further discussion and evidence on this issue, see Dufour and Kiviet (1996), Dufour and Khalaf (2002*a*, 2002*b*), Dufour et al. (2004), and Dufour (2006).

We shall 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 **3.1**. *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 (3.24), the latter apply tests based on the following distributional result:

$$\frac{T-s-n}{n(T-s-1)}Q \sim F(n,\, T-s-n)\,, \text{ with } Q = T\,\hat{a}'\left[\frac{T}{T-k}\hat{\Sigma}\right]^{-1}\hat{a}\,/\left[1+\bar{r}'\hat{\Delta}^{-1}\bar{r}\right], \quad (3.25)$$

where $\hat{a}$ is the vector of intercept OLS estimates, $[T/(T-k)]\hat{\Sigma}$ is the OLS-based unbiased estimator of Σ, $\bar{r}$ and $\hat{\Delta}$ include respectively the time-series-means and sample covariance matrix corresponding to the right-hand-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 **3.1** 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 (2.12) on a MLR of the form (2.2) describing returns. In this case, the null distribution of the LR statistic follows from Theorem **3.1** for the specific 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, since 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 above results to the case of (2.16) 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 will be denoted by $\mathcal{C}(Y)$ where Y refers to the return data. Next, on applying Theorem **3.1** and the MC algorithm in Appendix A (summarized above) we can obtain for each $\nu_0 \in \mathcal{C}(Y)$ a MC p-value $\hat{p}_N(LR_0|\nu)$. Setting

$$Q_U(LR_0) = \sup_{\nu_0 \in \mathcal{C}(Y)} \hat{p}_N(LR_0|\nu_0)\,, \quad (4.1)$$

the critical region

$$Q_U(LR_0) \leq \alpha_2 \quad (4.2)$$

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 cut-off 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 $\alpha_1 = \alpha_2 = \alpha/2$.

Since a procedure to derive 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 (2.16) where $\nu = \nu_0$ for known ν_0. Specifically, suppose some test statistic [denoted $\mathcal{T}(Y)$] is available for the latter hypothesis (we provide one in Section 5.1 below). Inverting $\mathcal{T}(Y)$ 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 [denote it $\mathcal{T}_0(Y)$] and its p-value [obtain it e.g. by MC test techniques and denote it $\hat{p}_N(\mathcal{T}_0(Y)|\nu_0)$] conforming with (2.16). 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 $\hat{p}_N(\mathcal{T}_0(Y)|\nu_0) > \alpha_1$; see Dufour (1990) and 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 *i.i.d.* errors.

5.1. Goodness-of-fit tests

The null hypotheses of concern here are (2.19) [normal errors], and (2.17) or (2.18) [e.g. multivariate t errors with known or unknown degrees-of-freedom]. The test criteria considered use the multivariate skewness and kurtosis measures:

$$SK = \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\hat{d}_{st}^3\,, \qquad KU = \frac{1}{T}\sum_{t=1}^{T}\hat{d}_{tt}^2\,, \tag{5.1}$$

where $\hat{d}_{st}$ are the elements of the matrix $\hat{D} = \hat{U}'\hat{\Sigma}^{-1}\hat{U} = T\,\hat{U}(\hat{U}'\hat{U})^{-1}\hat{U}'$. These statistics were introduced by Mardia (1970) in models where the regressor reduces to a vector of ones. Zhou (1993, p. 1935, footnote 5) proposed to use them to test elliptical distributions, without however providing a finite-sample theory for their application to residuals from MLR models. In our context, these statistics are distributed, respectively, like $\frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}d_{st}^3$ and $\frac{1}{T}\sum_{t=1}^{T}d_{tt}^2$, where d_{st} is the $(s,\,t)$-th element of the matrix

$$D = T\,MW\left(W'MW\right)^{-1}W'M\,, \quad W = [W_1,\,\ldots\,,\,W_T]'\,; \tag{5.2}$$

see Dufour, Khalaf and Beaulieu (2003). This implies that SK and KU are pivotal (invariant to B and Σ). Two further adjustments are applied: (1) a simulation-based "centering" of the test statistics; (2) 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 (2.17), the resulting statistics are denoted $\overline{SK}(\nu_0)$ and $\overline{KU}(\nu_0)$. In the Gaussian case (2.19), 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:

$$ESK(\nu_0) = \left|SK - \overline{SK}(\nu_0)\right|\,, \quad EKU(\nu_0) = \left|KU - \overline{KU}(\nu_0)\right|\,, \tag{5.3}$$

in the general case; in the Gaussian case, the test statistics are denoted $ESK = \left|SK - \overline{SK}\right|$ and $EKU = \left|KU - \overline{KU}\right|$. This modification preserves pivotality. The MC technique may thus be applied to derive exact p-values (using N replications): the resulting simulation-based p-values are denoted $\hat{p}_N\big(ESK(\nu_0)|\nu_0\big)$, $\hat{p}_N\big(EKU(\nu_0)|\nu_0\big)$ in the general case, and $\hat{p}_N(ESK_0)$, $\hat{p}_N(EKU_0)$ under the Gaussian hypothesis; see Appendix B.2 for more details. The observed and simulated statistics have to be obtained conditional on the same average skewness and kurtosis measures; this ensures that they remain exchangeable; see Dufour (2006).

This procedure allows to obtain exact individual p-values for each statistic. To obtain a joint test, we propose to reject 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 cut-off level), we use the following combined statistic [see Dufour et al. (2003)]:

$$CSK(\nu_0) = 1 - \min\left\{\hat{p}_N\big(ESK(\nu_0)|\nu_0\big),\ \hat{p}_N\big(EKU(\nu_0)|\nu_0\big)\right\} \tag{5.4}$$

or $CSK = 1 - \min\left\{\hat{p}_N\big(ESK\big),\ \hat{p}_N\big(EKU\big)\right\}$ in the Gaussian case. This combination method preserves invariance to B and Σ. So under (2.19), a MC p-value for CSK can be easily obtained. Under (2.17), pivotality allows to obtain a MC p-value given a known value $\nu = \nu_0$, which is denoted $\hat{p}_N(CSK_0(\nu_0)|\nu_0)$ 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 (2.18) which are consistent with the data; if this set is empty, then (2.18) is rejected at level α_1; details of the algorithm are given in Appendix B.3.

5.2. Multivariate checks for serial dependence and GARCH

We now present the tests we apply to assess departure from *i.i.d.* errors, specifically, tests against conditional heteroskedasticity and variance ratio tests; see Dufour, Khalaf and Beaulieu (2005). The null hypotheses of concern are (2.17), (2.18) or (2.19).

If one pursues a univariate approach, standard diagnostics may be applied to each equation in (2.2). For instance, the Engle GARCH test statistic [Engle (1982)] for equation i, denoted 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:

$$LK_i = \frac{\left\{(T-\bar{q})\sum\limits_{t=\bar{q}+1}^{T}\left[(\hat{\varepsilon}_{it}^2/\hat{\sigma}_i^2 - 1)\right]\sum\limits_{j=1}^{\bar{q}}\hat{\varepsilon}_{i,t-j}^2\right\} \Big/ \left\{\sum\limits_{t=\bar{q}+1}^{T}(\hat{\varepsilon}_{it}^2/\hat{\sigma}_i^2 - 1)^2\right\}^{1/2}}{\left\{(T-\bar{q})\sum\limits_{t=\bar{q}+1}^{T}\left(\sum\limits_{j=1}^{\bar{q}}\hat{\varepsilon}_{i,t-j}^2\right)^2 - \left(\sum\limits_{t=\bar{q}+1}^{T}\left(\sum\limits_{j=1}^{\bar{q}}\hat{\varepsilon}_{i,t-j}^2\right)\right)^2\right\}^{1/2}} \tag{5.5}$$

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:

$$VR_i = 1 + 2\sum_{j=1}^{K}\Big(1 - \frac{j}{K}\Big)\hat{\rho}_{ij}, \quad \hat{\rho}_{ij} = \frac{\sum_{t=j+1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{i,t-j}}{\sum_{t=1}^{T}\hat{\varepsilon}_{ti}^2}, \quad j = 1, \ldots, K, \tag{5.6}$$

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

Such univariate tests may not be appropriate in multivariate regressions. Indeed, the error covariance, which appears as a nuisance parameter, is typically not taken into consideration if a series of univariate tests are applied. Furthermore, the problem of combining test decisions over all equations is not straightforward, since 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

$$\tilde{W} = \hat{U}\, S_{\hat{U}}^{-1} \tag{5.7}$$

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

$$\tilde{E} = 1 - \min_{1\le i\le n}\big[p(\tilde{E}_i)\big]\,, \quad \widetilde{LK} = 1 - \min_{1\le i\le n}\big[p(\widetilde{LK}_i)\big]\,, \quad \widetilde{VR} = 1 - \min_{1\le i\le n}\big[p(\widetilde{VR}_i)\big]\,, \tag{5.8}$$

where $p(\tilde{E}_i)$, $p(\widetilde{LK}_i)$ and $p(\widetilde{VR}_i)$ refer to p-values (which may be obtained by applying a MC test method or by using asymptotic null distributions to cut execution time). In our context, $\tilde{W}$ has a distribution which is completely determined by the distribution of W given X, provided J [in (2.16)] 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 B and Σ, so under (2.19), 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 $\hat{p}_N(\tilde{E}|\nu_0)$, $\hat{p}_N(\widetilde{LK}|\nu_0)$ and $\hat{p}_N(\widetilde{VR}|\nu_0)$. The unknown ν problem is solved by applying a MMC strategy: we compute

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

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

5.3. Conditional heteroskedasticity under elliptical distributions

Finally, we also test for irregularities that arise from modeling elliptical returns via distributional assumptions on error terms (as we have proceeded so far), because the latter statistical approach may lead to conditional heteroskedasticity of the following form: the variance of the vector $(r_{1t},\, r_{2t},\, \ldots\, ,\, r_{nt})'$ is proportional to a quadratic function of $\tilde{r}_{\mathrm{M}t}$, specifically, the standardized square of the deviation of $\tilde{r}_{\mathrm{M}t}$ from its time series mean which we denote $\tilde{z}_{\mathrm{M}t}$; see Hodgson et al. (2002), Vorkink (2003), Zhou (1993) and Kan and Zhou (2003). In multi-beta contexts, $\tilde{z}_{\mathrm{M}t}$ is the t-th element of the matrix $\overline{r}'\hat{\Delta}^{-1}\overline{r}$ which appears in (3.25). For example, for the multivariate t with

$\kappa \geq 2,$ the variance proportionality factor is

$$\delta_t = (\kappa - 2 + \tilde{z}_{\mathrm{M}t})/(\kappa - 1)\,. \tag{5.9}$$

We proceed as for the GARCH test, using the univariate Lagrange multiplier (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 $\tilde{z}_{\mathrm{M}t}$. This statistic is obtained from the standardized residuals of each equation, leading to n statistics denoted $\widetilde{BP}_i$, which are combined through the minimum approximate p-value as follows:

$$\widetilde{BP} = 1 - \min_{1\leq i\leq n}\left[p(\widetilde{BP}_i)\right]. \tag{5.10}$$

If heteroskedasticity of the form (5.9) 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 re-estimated by using the corresponding generalized least squares estimator [leading to weighted ML-type or quasi ML estimators (QMLE)], 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 (2.1) in the context of (2.2), tests of (2.6) in the context of (2.5) and tests of (2.8) in the context of (2.7)], where the errors follow multivariate normal and Student t distributions. For Student distributions, we assume that (2.16) holds with

$$W_t = \mathcal{Z}_{1t}/(\mathcal{Z}_{2t}/\kappa)^{1/2}\;, \tag{6.1}$$

where $\mathcal{Z}_{1t}$ is multivariate normal with mean zero and covariance matrix I_n, and $\mathcal{Z}_{2t} \sim \chi^2(\kappa)$ and is independent of Z_{1t}.

We use nominal monthly returns from January 1926 to December 1995, obtained from the University of Chicago's Center for Research in Security Prices (CRSP). We form 12 portfolios of New York Stock Exchange (NYSE) firms grouped by standard two-digit industrial classification (SIC), as in Breeden, Gibbons and Litzenberger (1989). As in Breeden et al. (1989), firms with SIC code 39 (Miscellaneous manufacturing industries) are excluded from the data set 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 CRSP. Furthermore, portfolios are value-weighted in each month. In order to assess the testable implications of the asset pricing models, we measure the market return by the value-weighted NYSE returns, also available from CRSP. The risk-free rate is measured by the one-month Treasury bill rate, also from 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 one-month

Treasury bill yield, the dividend yield of the Standard and Poor's 500 index, the spread between Moody's Baa and Aaa corporate bond yield, the spread between a ten year and a one year Treasury bond yield, and the difference between the one-month lagged returns of a three-month and a one-month Treasury bill. Since the instruments are not available before the mid-sixties, 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 data set; the same analysis including these observations yields qualitatively similar results.

Table 1 reports tests of the unconditional CAPM over subperiods of 5 years. 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) and 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 below) points out to similar problems in the latter context.

In Table 1, we report in columns (1) - (3) the p-values of the exact multi-normality tests based on ESK, EKU and CSK (see Section 5.1). These tests allow one to evaluate whether observed residuals exhibit non-Gaussian behavior through excess skewness and kurtosis. For most subperiods, normality is rejected. These results are interesting since, although it is well accepted in the finance literature that continuously compounded returns are skewed and leptokurtic, empirical evidence of non-normality is weaker for monthly data. For instance, Affleck-Graves and McDonald (1989) reject normality in about 50% of the stocks they study. Our results, which are exact (*i.e.*, cannot reject spuriously), indicate much stronger evidence against normality. This also confirms the results of Richardson and Smith (1993) who provide 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.

In columns (4)-(7) of Table 1, we 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). The confidence set $\mathcal{C}(Y)$ for the number of degrees of freedom κ appears in column (8). These results show that asymptotic p-values are quite often spuriously significant (*e.g.*, for 1941-55). Furthermore, 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 ten rejections [out of the fourteen subperiods] of the null hypothesis for the asymptotic $\chi^2(12)$ test, nine for the MC p-values under normality and six 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. Since the confidence level for the nuisance parameter is 0.975 $(\alpha_1 = 0.025)$, p-values for the efficiency tests should be compared with $\alpha_2 = 0.025$ to ensure the overall level of the test is $\alpha = \alpha_1 + \alpha_2 = 0.05$; see Section 4.

Table 1. Normality and unconditional efficiency tests

	Normality tests			Efficiency Tests						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Sample	SK	KU	CSK	LR	p_∞	$p_{\mathcal{N}}$	Q_U	$\mathcal{C}(Y)$	Q_U^{GLS}	$\mathcal{C}_{GLS}(Y)$
1927-30	**.001**	**.001**	**.001**	16.104	.1866	.364	.357	$3-12$	.396	$3-15$
1931-35	**.001**	**.001**	**.001**	16.257	.1798	.313	.322	$3-8$	.268	$3-9$
1936-40	**.001**	**.001**	**.001**	16.018	.1904	.319	.333	$4-26$	.483	$3-26$
1941-45	**.004**	**.002**	**.004**	25.869	**.0112**	**.045**	.049	≥ 5	.049	≥ 4
1946-50	**.001**	**.001**	**.001**	37.196	**.0002**	**.003**	**.004**	$4-26$	**.004**	$2-24$
1951-55	**.001**	**.002**	**.001**	36.510	**.0003**	**.004**	**.005**	$5-31$	**.007**	$2-33$
1956-60	**.024**	**.003**	**.003**	43.841	**.0000**	**.002**	**.002**	≥ 5	**.002**	≥ 2
1961-65	594	.479	.631	39.098	**.0001**	**.002**	**.002**	≥ 7	**.002**	≥ 4
1966-70	**.011**	**.002**	**.004**	36.794	**.0002**	**.003**	**.003**	≥ 5	**.003**	≥ 3
1971-75	**.001**	**.002**	**.001**	21.094	**.0490**	.120	.129	$4-24$	.112	$4-30$
1976-80	**.001**	**.001**	**.001**	28.373	**.0049**	**.023**	.026	$4-17$	**.014**	$2-18$
1981-85	**.001**	**.001**	**.001**	27.189	**.0073**	**.033**	.035	$5-34$	.033	$2-30$
1986-90	**.028**	**.020**	**.030**	35.747	**.0007**	**.003**	**.005**	≥ 5	**.006**	≥ 2
1991-95	.177	.311	.239	16.752	.1592	.299	.305	≥ 15	.293	≥ 6

Note – Numbers in bold indicate test results which are significant at level 0.05. Columns (1)-(3) report p-values for multinormality tests: columns (1)-(2) pertain respectively 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 (5.3) and (5.4) under the assumption of multivariate normal errors in the context of (2.2). Column (4) presents the quasi-LR statistic defined in (3.20) to test $\mathcal{H}_E$ defined by (2.1) in the context of (2.2); columns (5), (6) and (7) are the associated p-values using, respectively, the asymptotic chi-square 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 $1-\alpha_1 = 0.975$. In the latter case, the maximized p-value for the corresponding efficiency test is significant at level 0.05 if it is not larger than $\alpha_2 = 0.025$. 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 (5.9) to correct for heteroskedasticity.

Table 2. Normality and conditional efficiency tests

	Normality tests			Conditional Efficiency Tests				
(A) Model (2.7)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	SK	KU	CSK	LR	p_∞	$p_\mathcal{N}$	Q_U	$\mathcal{C}(Y)$
1966-70	.085	**.017**	**.033**	122.545	**.0002**	.111	.125	≥ 4
1971-75	.778	.986	.908	130.384	**.0000**	.057	.067	≥ 6
1976-80	.095	.118	.137	147.084	**.0000**	**.012**	**.021**	≥ 4
1981-85	.707	.095	.141	155.475	**.0000**	**.004**	**.005**	≥ 4
1986-90	.114	**.032**	**.046**	109.736	**.0028**	.300	.344	≥ 3
1991-95	.611	.501	.645	113.462	**.0013**	.207	.225	≥ 6
1966-95	**.001**	**.001**	**.001**	162.050	**.0000**	**.001**	**.001**	$3-16$
(B) Model (2.5)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	SK	KU	CSK	LR	p_∞	$p_\mathcal{N}$	Q_U	$\mathcal{C}(Y)$
1966-70	.275	.014	.025	34.344	**.0006**	**.011**	**.015**	≥ 4
1971-75	.093	.139	.130	26.166	**.0102**	.072	.087	≥ 5
1976-80	**.013**	**.002**	**.001**	31.903	**.0014**	**.021**	**.023**	≥ 4
1981-85	**.019**	**.024**	**.028**	32.655	**.0011**	**.019**	.026	≥ 4
1986-90	**.019**	**.015**	**.028**	31.932	**.0014**	**.020**	**.024**	≥ 4
1991-95	.160	.381	.200	17.976	.1164	.338	.347	≥ 11
1966-95	**.001**	**.001**	**.001**	39.790	**.0001**	**.001**	**.001**	$4-15$

Note – Numbers in bold indicate test results which are significant at level 0.05. Columns (1)-(3) report p-values for multinormality tests: columns (1)-(2) pertain respectively 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 (5.3) and (5.4) under the assumption of multivariate normal errors, in the context of (2.5) [in Panel B] and (2.7) [in Panel A]. Column (4) presents the quasi-LR statistic defined in (3.20) to test $\overline{\mathcal{H}}_{E1}$ defined by (2.6) [in Panel B], and $\overline{\mathcal{H}}_{E2}$ defined by (2.8) [in Panel A]; columns (5), (6) and (7) are the associated p-values using, respectively, the asymptotic chi-square 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 $1-\alpha_1 = 0.975$. In the latter case, the maximized p-value for the corresponding efficiency test is significant at level 0.05 if it is not larger than $\alpha_2 = 0.025$. The confidence set for κ is reported in column (8); see Section 4 for details on its construction.

These findings differ from those of Zhou (1993), who found no change in 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 follows:

$$GS_1 = -2\sum_{j=1}^{14} \ln(p_{\mathcal{N}}[j]) \ , \quad GS_2 = \sum_{j=1}^{14} \Psi^{-1}(p_{\mathcal{N}}[j]) \ , \tag{6.2}$$

where $[j]$ refers to the subperiods, and $\Psi^{-1}(.)$ 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 $GS_2 \sim N(0,\, 14)$. It is worth noting that the same aggregation methods can be applied to our test problem even under (2.16) by replacing, in (6.2), $p_{\mathcal{N}}[j]$ with $Q_{U[j]}$, the MMC p-values obtained imposing (2.16). Indeed, as is observed by Gibbons and Shanken (1987), the F-distribution is not necessary 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 respectively, 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 in Gibbons and Shanken (1987), this implies that mean-variance efficiency is jointly rejected by our data. If one questions independence and prefers to combine using Bonferroni-based criteria, the smallest p-value is .002 which when referred to $.025/14 \simeq 0.002$ 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 $.028/14 = 0.002$. This means that in every period, the pre-test confidence set should be applied with $\alpha_1 = 0.022$ to allow 0.028 to the mean-variance efficiency test. The results reported in the above tables are robust to this change in level.

Finally, Table 3 presents the results of our multivariate exact diagnostic checks for departures from the *i.i.d.* 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 both at the 1% and 5% level of significance. This implies that, in our statistical framework and for the time spans analyzed, *i.i.d.* errors provide an acceptable working assumption. Our heteroskedasticity 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

Table 3. Multivariate diagnostics, unconditional CAPM

	Normal errors			Student t errors			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sample	$\tilde{E}$	$\widetilde{LK}$	$\widetilde{VR}$	$\tilde{E}$	$\widetilde{LK}$	$\widetilde{VR}$	$\widetilde{BP}$
1927-30	**.001**	.356	**.004**	**.013**	.301	**.004**	.285
1931-35	**.022**	.748	.069	.082	.659	.066	**.016**
1936-40	.075	.612	.855	.124	.587	.867	.087
1941-45	.824	.979	.163	.843	.982	.177	.034
1946-50	**.003**	.804	.063	**.017**	.784	.068	.880
1951-55	.139	.353	.111	.168	.321	.120	.591
1956-60	.987	.628	.093	.994	.628	.095	.347
1961-65	.339	.207	.577	.375	.195	.584	.771
1966-70	**.027**	.274	.821	.043	.278	.847	.961
1971-75	.280	.224	.218	.316	.212	.224	**.013**
1976-80	**.004**	**.011**	.165	**.016**	**.013**	.183	.406
1981-85	**.027**	.103	.208	.050	.103	.217	.583
1986-90	**.033**	.453	.346	.077	.442	.366	.279
1991-95	.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 (5.8), in the context of model (2.2). $\tilde{E}$ and $\widetilde{LK}$ are multivariate versions of Engle's and Lee-King's GARCH tests, while $\widetilde{VR}$ is a multivariate version of Lo and MacKinlay's variance ratio tests; see Section 5.2. $\widetilde{BP}$ [defined in (5.10)] is the conditional heteroskedasticity test as function of the benchmark returns, which is relevant for elliptical non-normal 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 0.975) of distributional nuisance parameters. The confidence sets used are those reported in Table 1 [column (8)]. Numbers in bold indicate test results which are significant at level 0.05.

weighted QMLE) may easily be conducted following the methodology we have described here, in the context of an MLR weighted by the necessary variance correction term, for example by using the variance weights (5.9) 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 columns (9) of Table 1. Results show that the decision of our tests is not notably affected when we correct for time varying volatility. It is worth noting 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 (2.5) [Panel B] and (2.7) [Panel A] over intervals of 5 years 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 is thus not applied in this context. The companion diagnostic tests are shown in Table 4. While the subperiod analysis may at first sight appear unnecessary, given that the conditional model is supposed to account for time-varying *betas*, care must be exercised in interpreting the full-sample test results. From Table 2, we see that for both models (2.5) and (2.7): (i) the efficiency hypotheses when assessed using

Table 4. Multivariate diagnostics, conditional CAPM

	Normal errors			Student t errors		
	(1)	(2)	(3)	(4)	(5)	(6)
Sample	$\tilde{E}$	$\widetilde{LK}$	$\widetilde{VR}$	$\tilde{E}$	$\widetilde{LK}$	$\widetilde{VR}$
1966-70	.297	.235	.166	.333	.239	.178
1971-75	.131	.095	.924	.188	.108	.929
1976-80	**.012**	.740	.669	**.020**	.744	.683
1981-85	.137	.108	.628	.172	.110	.629
1986-90	.264	.766	.932	.338	.767	.933
1991-95	.878	.178	.473	.878	.184	.495
1966-75	.331	.083	.417	.348	.087	.425
1976-85	.290	**.005**	.690	.348	**.008**	.706
1986-95	**.015**	.647	.190	.038	.639	.207
1966-95	**.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 (5.8), in the context of model (2.5). $\tilde{E}$ and $\widetilde{LK}$ are multivariate versions of Engle's and Lee-King's GARCH tests, while $\widetilde{VR}$ is a multivariate version of Lo and MacKinlay's variance ratio tests; see Section 5.2. $\widetilde{BP}$ [defined in (5.10)] is the conditional heteroskedasticity test as function of the benchmark returns, which is relevant for elliptical non-normal errors. The MC p-values in columns (1)-(3) are based on pivotal statistics, while those in columns (4)-(11) are MMC p-values obtained by maximizing over confidence sets (with level 0.975) 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 level 0.05.

the whole sample, are soundly rejected, using asymptotic or MC p-values, (ii) the confidence sets on the degrees-of-freedom parameter appear dramatically tighter, and (iii) normality is definitely rejected. Unfortunately, our diagnostic tests [refer to Table 4] reveal significant departures from the statistical foundations underlying the latter tests (even when allowing for non-normal errors); temporal instabilities thus cast doubt on the full sample analysis. The tests in Table 4 are applied in the context of the conditional model (2.7); since the latter nests model (2.5) 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 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 (2.7); indeed, as may be seen from Panel A of Table 2, there is a large difference between the asymptotic and the MC (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 non-rejections for the shorter sub-sample, although the simulation studies reported in Dufour and Khalaf (2002*b*) 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 (2.7), 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 the ones used in this paper, 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 (2.5), 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 (2.7), at the 5% significance level, we find six rejections [out of the six subperiods] of the null hypothesis for the asymptotic test, two for the MC test under normality and the Student t distribution. Not surprisingly, in the subperiods where the conditional models are rejected, the unconditional model is also rejected. In general, model (2.7) is rejected in fewer subperiods relative to model (2.5) and the unconditional model (over the 1966-95 sub-sample, where the data allows to estimate the conditional models).

In the case of (2.7), 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. So to aggregate our subperiod analysis, we resort once again to the combined statistics used by Gibbons and Shanken (1987) as in the unconditional case. Our results, under normal and Student t errors respectively, 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 (2.7), and $GS_1 = 39.572$ [.00001] and 37.703 [.00017] and $GS_2 = 10.331$ [.00000] and 9.839 [.00000], for model (2.5). The latter p-values imply that mean-variance efficiency is jointly rejected with our data. Once again, if one questions independence and prefers to combine using Bonferroni-based criteria, the smallest p-value for (2.7) is .004 under normality and .005 with t-errors; the latter when compared with $.025/6 \simeq 0.004$ come close to a rejection. Efficiency on the aggregate in model (2.5) fails to be rejected by the Bonferroni rule. Viewed collectively, our subperiod and aggregate tests indicate that the method one uses to incorporate conditioning information has non negligible implications on mean-variance efficiency.

These results motivate the use of alternative models which capture conditioning information in more parsimonious approaches (i.e. with fewer degrees-of-freedom losses). Inevitably, such approaches as well as non-linear stochastic discount factor based models, will lead to Instrumental Variable contexts (see the above cited references on GMM-based tests of the CAPM), where the literature on exact testing is still scarce.

7. Conclusion

In this paper, we have proposed exact mean-variance efficiency tests in the context of unconditional and conditional CAPM's, with Gaussian or non-Gaussian disturbances. Further, we have shown how to deal – in finite samples – which may involve Student t errors with possibly unknown parameters.

Our empirical results show clearly that the normality assumption does not fit CAPM error returns, even for monthly data. By contrast, Student t distributions appear to be consistent with the data. Exact unconditional mean-variance efficiency tests, which formally account for non-normality, fail to reject efficiency for 3 out of 9 subperiods for which Gaussian-based tests are significant. The conditional models analyzed provide a better fit, but the efficiency restrictions are rejected on at least half of the 6 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 non-normal 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 worth noting that the proposed methodology applies to several interesting asset pricing tests including problems where the Hotelling test [exploited by GRS and MacKinlay (1987)] and Rao's F test [see Stewart (1997) and Dufour and Khalaf (2002*b*, Appendix)] have been used. In view of its fundamental importance, mean-variance efficiency is one of the few MLR-based problems which 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 exactly applying Rao's F test. Examples include inter-temporal asset pricing tests in Shanken (1990, footnote 18). Furthermore, as discussed in Shanken (1996), econometric tests of spanning fall within this class. Indeed, spanning tests [see the survey of DeRoon and Nijman (2001)] may be written in terms of a model of the GRS form. The hypothesis is however more restrictive, in the sense that, in addition to the restrictions on the intercepts, the betas of each regression must sum to one. 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 paper extend available exact tests of these important financial problems beyond the Gaussian context.

The fact remains that the results presented in this paper are specific to UL hypotheses. Not all linear hypotheses may be cast in this form. In Beaulieu, Dufour and Khalaf (2005), we study extensions to non-linear 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-hand-side benchmark may be observed with errors. The development of exact tests which correct for error-in-variable problems also appears to be an important issue, and we are pursuing research on it.

Appendix

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

A. Monte Carlo tests

Let $S(y,\, X)$ be a test statistic which can be rewritten in the form

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

under the null hypothesis, where W is defined by (2.16) and the distribution of W is known. For example, $S(y,\, X)$ could be the LR statistic considered in Theorem **3.1**. Then the conditional distribution of $S(y,\, X),$ given $X,$ is completely determined by the matrix X and the conditional distribution of W given X, *i.e.* $S(y,\, X)$ 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 *i.i.d.* replications of W : $W_{(j)} = [W_1^{(j)},\, \ldots\, ,\, W_n^{(j)}]\,,$ $j = 1,\, \ldots\, ,\, N\,.$
3. From each simulated error matrix $W_{(j)}$, compute the statistics $S^{(j)} = \bar{S}(W_{(j)},\, X)\,, \quad j = 1\,,\, \ldots\, ,\, N\,.$ For instance, in the case of the QLR statistic underlying Theorem **3.1**, calculate $L(W_{(j)}) = T\, \ln\big(|W'_{(j)} M_0 W_{(j)}|/|W'_{(j)} M W_{(j)}|\big)\,,\ j = 1\,,\, \ldots\, ,\, N\,.$
4. Compute the MC p-value $\hat{p}_N[S] \equiv p_N(S^{(0)}\,;\, S),$ where

   $$p_N(x\,;\, S) \equiv \frac{NG_N(x\,;\, S) + 1}{N+1}\,, \tag{A.2}$$

   $$G_N(x\,;\, S) \equiv \frac{1}{N}\sum_{j=1}^{N} I_{[0,\,\infty)}(S^{(j)} - x)\,,\ \ I_{[0,\,\infty)}(x) = \begin{cases} 1, & if\ x \in [0,\, \infty)\,, \\ 0, & if\ x \notin [0,\, \infty)\,. \end{cases} \tag{A.3}$$

   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 which are greater than or equal to $S^{(0)}$. When $S^{(0)},\, S^{(1)},\, \ldots\, ,\, S^{(N)}$ are all distinct [an event with probability one when the vector $\big(S^{(0)},\, S^{(1)},\, \ldots\, ,\, S^{(N)}\big)'$ 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: $\hat{p}_N[S] \le \alpha\,, \quad 0 < \alpha < 1\,.$ 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 above algorithm is valid for any fully specified distribution of W. Consider now the case where the distribution of W involves a nuisance parameter as in (2.16). In this case, given ν, (A.2) yields a MC p-value which we will denote $\hat{p}_N[S \mid \nu]$ where the conditioning on ν is emphasized for further reference. The test defined by $\hat{p}_N[S \mid \nu] \leq \alpha$ has size α [in the sense of (A.4)] for known ν. Treating ν as a nuisance parameter, the test based on

$$\sup_{\nu \in \Phi_0} \hat{p}_N[S \mid \nu] \leq \alpha \tag{A.5}$$

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 $\hat{p}_N[S \mid \hat{\nu}_0]$, where $\hat{\nu}_0$ is any *point* estimate of ν. In Dufour and Khalaf (2002*b*), 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 non-rejections are *exactly* conclusive in the following sense: if $\hat{p}_N[S \mid \hat{\nu}_0] > \alpha$, then the exact MMC test is clearly not significant at level α.

B. 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 is available in Dufour et al. (2003).

B.1. Estimating expected skewness and kurtosis

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

A2. From each simulated error matrix $\bar{W}_{(i)}$, compute [see (5.2)]

$$D_{(i)} = TM\bar{W}_{(i)}\left[\bar{W}'_{(i)}M\bar{W}_{(i)}\right]^{-1}\bar{W}'_{(i)}M, \quad i = 1, \ldots , N_0 , \tag{B.1}$$

and compute the corresponding statistics SK and KU, applying (5.1). This provides N_0 replications of the latter statistics: $\overline{SK}_{(i)}$ and $\overline{KU}_{(i)}$, $i = 1, \ldots , N_0$.

A3. Compute the average values:

$$\overline{SK}(\nu_0) = \sum_{i=1}^{N_0} \overline{SK}_{(i)}/N_0 , \quad \overline{KU}(\nu_0) = \sum_{i=1}^{N_0} \overline{KU}_{(i)}/N_0 . \tag{B.2}$$

We call $\overline{SK}(\nu_0)$ and $\overline{KU}(\nu_0)$ the *reference simulated moments*. Two questions arise at this stage: (i) how to obtain exact cut-off points for $ESK(\nu_0)$ and $EKU(\nu_0)$ in (5.3), and (ii) how to obtain a size-correct simultaneous test which combines these two tests. Let us first address the individual p-values issue, which may be run as in Appendix A above.

B.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. Independently of the data and the draws of steps A1-A3, generate N *i.i.d.* 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: $E^{(j)} = [ESK_{\mathrm{M}}^{(j)}(\nu_0),\, EKU_{\mathrm{M}}^{(j)}(\nu_0)]'$, $j = 1, \ldots, N$. It is easy to see that the $N + 1$ vectors $E^{(j)}$, $j = 0, 1, \ldots, N$ 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 Appendix A 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 $\hat{p}_N[ESK_{\mathrm{M}}(\nu_0)] \leq \alpha$, and similarly for $EKU_{\mathrm{M}}(\nu_0)$. By the exchangeability of $E^{(j)}$, $j = 0, 1, \ldots, N$, and provided E follows a continuous distribution, this procedure satisfies the size constraint, *i.e.*

$$\mathsf{P}\big[\hat{p}_N[ESK_{\mathrm{M}}(\nu_0)] \leq \alpha\big] = \mathsf{P}\big[\hat{p}_N[EKU_{\mathrm{M}}(\nu_0)] \leq \alpha\big] = \alpha \tag{B.3}$$

under the null hypothesis.

B.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 Appendix A.

C3. Independently of the previous simulations and the data, generate N_1 additional *i.i.d.* realizations of W according to the hypothesized distribution with $\nu = \nu_0$. N_1 is chosen 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: $EE^{(l)} = [ESK_{\mathrm{M}}^{(l)}(\nu_0),\, EKU_{\mathrm{M}}^{(l)}(\nu_0)]'$, $l = 1, \ldots, N_1$.

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: $\hat{p}_N^{(l)}[S] = p_N(S^{(l)}\,;\,S)$, $l = 0, 1, \ldots, N_1$, 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:

$$CSK_{\mathrm{M}}^{(l)}(\nu_0) = 1 - \min\left\{\hat{p}_N[ESK_{\mathrm{M}}^{(l)}(\nu_0)],\, \hat{p}_N[EKU_{\mathrm{M}}^{(l)}(\nu_0)]\right\}, \quad l = 0,\, 1,\, \ldots\,,\, N_1\,. \tag{B.4}$$

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

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

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

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 and Economic Statistics* **22**(4), 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, Centre interuniversitaire de recherche en analyse des organisations (CIRANO) and Centre interuniversitaire de recherche en économie quantitative (CIREQ), Université de 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**(1), 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 University Press, New Jersey.

Cochrane, J. H. (2001), *Asset Pricing*, Princeton University Press, Princeton, New Jersey.

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.

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**(2), 443–477.

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., 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. and Beaulieu, M.-C. (2005), Multivariate residual-based finite-sample tests for serial dependence and GARCH with applications to asset pricing models, Technical report, CIRANO and CIREQ, Université de 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**(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.

Engle, R. F. (1982), ‘Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation’, *Econometrica* **50**(4), 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.

Fama, E. F. and French, K. R. (2004), ‘The Capital Asset Pricing Model: Theory and evidence’, *The Journal of Economic Perspectives* **18**(3), 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 and Economic Statistics* **21**(4), 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**(3), 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 and Economic Statistics* **21**, 269–283.

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

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, Toronto, Canada.

Kandel, S., McCulloch, R. and Stambaugh, R. F. (1995), 'Bayesian inference and portfolio efficiency', *The 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 and Economic Statistics* **11**, 17–27. Correction 12 (1994), 139.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, 2nd edn, John Wiley & Sons, New York.

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.

MacKinlay, A. C. (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', *The 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.

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.

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', *The Review of Financial Studies* **16**(3), 845–874.

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