Multivariate Tests of Mean-Variance Efficiency and Spanning With a Large Number of Assets and Time-Varying Covariances

Sermin Gungor

Funds Management and Banking Department, Bank of Canada, Ottawa, Ontario K1A 0G9, Canada (sgungor@bankofcanada.ca)

Richard Luger

Department of Finance, Insurance and Real Estate, Laval University, Quebec City, Quebec G1V 0A6, Canada (richard.luger@fsa.ulaval.ca)

November 16, 2016

ABSTRACT: We develop a finite-sample procedure to test the mean-variance efficiency and spanning hypotheses, without imposing any parametric assumptions on the distribution of model disturbances. In so doing, we provide an exact distribution-free method to test uniform linear restrictions in multivariate linear regression models. The framework allows for unknown forms of non-normalities as well as time-varying conditional variances and covariances among the model disturbances. We derive exact bounds on the null distribution of joint F statistics in order to deal with the presence of nuisance parameters, and we show how to implement the resulting generalized non-parametric bounds tests with Monte Carlo resampling techniques. In sharp contrast to the usual tests which are not even computable when the number of test assets is too large, the power of the proposed test procedure potentially increases along both the time and cross-sectional dimensions.

KEY WORDS: Exact distribution-free inference; Monte Carlo bounds test; Multi-beta asset pricing model; Multivariate linear regression; Multivariate GARCH

1 Introduction

A benchmark portfolio of assets is said to be mean-variance efficient with respect to a given set of test assets if it is not possible to combine it with the test assets to obtain another portfolio with the same expected return as the benchmark portfolio, but a lower variance. With multiple benchmark portfolios, the question becomes whether some combination of them is efficient. The mean-variance efficiency hypothesis is a testable implication of the validity of linear factor asset pricing models, such as the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965), or more generally of the Arbitrage Pricing Theory (APT) of Ross (1976); see Sentana (2009) for a recent survey of the econometrics of mean-variance efficiency tests. A more stringent hypothesis is that of mean-variance spanning, which states that the minimum-variance frontier of the benchmark portfolios plus the test assets coincides with the frontier of the benchmark portfolios only; see DeRoon and Nijman (2001) for a survey. When spanning holds, the addition of the new assets does not improve the efficiency frontier for a mean-variance optimizing investor. This means that the extra assets are not worth holding, either long or short (Cheung et al., 2009).

The most prominent tests of these hypotheses are those by Gibbons et al. (1989) (GRS) in the case of mean-variance efficiency and by Huberman and Kandel (1987) (HK) for the spanning hypothesis. These tests take the form of either likelihood ratio (LR) tests or system-wide F tests conducted within a multivariate linear regression (MLR) model, where the number of equations in the system equals the number of test assets. The CAPM and APT are single-period models, so in order to test their implications it is necessary to make an assumption concerning the time-series behavior of returns. The exact, finite-sample distributional theory for the GRS and HK tests rests on the assumption that the MLR model disturbances are independent and identically distributed (i.i.d.) each period according to a multivariate normal distribution. This assumption can be questionable when dealing with financial asset returns, since there has long been ample evidence that financial returns exhibit non-normalities; see e.g. Fama (1965), Blattberg and Gonedes (1974), Affleck-Graves and McDonald (1989), and Zhou (1993). Beaulieu et al. (2007, 2010) (BDK) extend the GRS and HK approaches for testing mean-variance efficiency and spanning. Their simulation-based procedure does not necessarily assume normality but it does nevertheless require that the disturbance distribution be parametrically specified, at least up to a finite number of unknown nuisance parameters, e.g. Student-t with unknown degrees of freedom.

In this paper, we extend the ideas of Gungor and Luger (2009, 2013) to obtain a finite-sample procedure to test mean-variance efficiency and spanning that relaxes four restrictions of the GRS and HK tests: (i) the assumption of independent disturbances, (ii) the assumption of identically distributed disturbances, (iii) the assumption of normally distributed disturbances, and (iv) the restriction on the number of test assets. Indeed, any procedure (e.g. GRS, HK, BDK) based on standard estimates of the disturbance covariance matrix requires that the size of the cross-section, N, be less than that of the time series, T, in order to avoid singularities and hence be computable. In sharp contrast, our approach is based on F statistics computed in turn for each equation of the MLR model and thus remains applicable no matter the number N of included equations. This idea of using equation-by-equation statistics that leave aside the effects of disturbance covariances follows Affleck-Graves and McDonald (1990) and Hwang and Satchell (2012). We propose the use of vector norms to combine the resulting N statistics, and we then derive exact bounds around the unknown null distribution of the aggregate F statistic in order to deal with the presence of nuisance parameters that arise in our statistical framework. In so doing, we provide a new method to test uniform (within equation) linear restrictions in MLR models, of which the efficiency and spanning hypotheses are special cases. The resulting generalized bounds tests bear resemblance to the well-known test of Durbin and Watson (1950, 1951) for autocorrelated disturbances in regression models.

The developed procedure rests on a multivariate conditional symmetry assumption for the MLR model disturbances, which includes the multivariate normal distribution assumed by GRS and HK. In fact, the maintained symmetry condition encompasses the entire class of elliptically symmetric distributions which play a very important role in mean-variance analysis because they guarantee full compatibility with expected utility maximization regardless of investor preferences; see Chamberlain (1983), Owen and Rabinovitch (1983), and Berk (1997). Unlike Gungor and Luger (2009, 2013), this framework also leaves open the possibility of unknown forms of time-varying conditional non-normalities and other distribution heterogeneities, such as time-varying conditional covariance structures. Many popular models, e.g. multivariate GARCH and stochastic volatility models with symmetrically distributed innovations, are compatible with our statistical framework. The null distribution of the equation-by-equation F statistics is characterized by a sign-permutation principle which preserves the cross-sectional covariance structure among the model disturbances. We rely on the Monte Carlo resampling techniques of Dwass (1957), Barnard (1963), and Birnbaum (1974) to obtain computationally inexpensive and yet exact p-values, no matter the sample size; see Dufour and Khalaf (2001) for a survey of Monte Carlo test in econometrics. In sharp contrast to the GRS and HK tests that are not computable when $N > T$, the power of the proposed test procedure potentially increases with both T and N.

Pesaran and Yamagata (2012) (PY) also develop (asymptotic) tests of the mean-variance efficiency hypothesis that can be applied when $N > T$. Similar to our approach, the PY tests use an aggregation of t statistics computed equation by equation. In order to deal with the presence of a non-trivial cross-sectional correlation structure, the PY test statistic is scaled by a threshold estimator of the average squares of pairwise disturbance correlations. The theory underlying the use of this threshold estimator nevertheless places certain restrictions on the allowable disturbance correlations. Specifically, it assumes weakly and sparsely correlated disturbances. So not surprisingly, our simulation experiments show that the asymptotically standard normal PY test has better power than ours when the model disturbances are uncorrelated in the cross-section. But as the degree of cross-sectional disturbance correlation increases (and whether the correlation structure is time-varying or not), the proposed test procedure does better than the PY test. Moreover, the PY approach based on t statistics is specifically tailored to the mean-variance efficiency hypothesis; it does not yield a general testing procedure for any MLR restriction. This leaves the new tests as the only ones available to test the mean-variance spanning hypothesis or any other uniform linear restrictions in MLR models when $N > T$.

It is important to note that large N, small T situations are quite common in empirical finance applications. Indeed, it is a usual practice to test asset pricing models over relatively short subperiods owing to concerns about parameter stability; see Campbell et al. (1997, Ch. 5), Gungor and Luger (2009, 2013), Ray et al. (2009), and Pesaran and Yamagata (2012) for examples. If $N > T$, one may ask: "Why not form portfolios to decrease the number of test assets?" Since Roll (1977), it has long been recognized that portfolio groupings can result in a loss of information about the cross-sectional behavior of individual stocks. Specifically, individual asset deviations from the pricing model can cancel out in the formation of portfolios, thereby destroying test power. As Lo and MacKinlay (1990) explain, the selection of assets to be included in a given portfolio is almost never at random, but is often based on some of the stock's empirical characteristics such as the market value of the companies' equity. This way of sorting stocks into groups based on variables that are correlated with returns is a questionable practice since it favors a rejection of the asset pricing model under consideration. Liang (2000) argues that even when the sort is based on a variable estimated using prior data, measurement error in this variable can also lead to a spurious rejection. If anything then, it seems more natural to try to increase the number of test assets in order to boost the probability of rejecting the null hypothesis when it is false. Indeed, an expansion of the investment universe should help detect violations of the null hypothesis, provided of course that more informative test assets get included in the MLR model.

The paper is organized as follows. In Section 2 we formally introduce the mean-variance efficiency and spanning hypotheses along with the exact GRS and HK tests. In Section 3 we develop our test procedure in the general MLR context. Section 4 presents the results of our simulation study comparing the performance of the new procedure with the GRS and PY tests of mean-variance efficiency, and to the HK test of mean-variance spanning. Section 5 presents an illustrative empirical application with a large number of individual stocks as test assets, and finally Section 6 concludes.

2 Hypotheses and exact tests

Consider an investment universe comprising a risk-free asset, K portfolios of risky assets, and an additional set of N risky assets. We are interested in the relation between the minimum-variance frontier spanned by the K benchmark portfolios and the frontier of the $N + K$ assets. At time t, the risk-free return is denoted by r_{ft}, the returns on the K benchmark portfolios are denoted by $\mathbf{r}_{Kt}$, and the returns on the other N test assets are denoted by $\mathbf{r}_t$. Correspondingly, the time-t excess returns are denoted by $\mathbf{z}_t = \mathbf{r}_t - r_{ft}$ and $\mathbf{z}_{Kt} = \mathbf{r}_{Kt} - r_{ft}$.

2.1 Mean-variance efficiency

Suppose the excess returns $\mathbf{z}_t$ are described by the following model:

$$\mathbf{z}_t = \mathbf{a} + \boldsymbol{\beta}\mathbf{z}_{Kt} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\mathbf{a}$ is an N-vector of intercepts (or *alphas*), $\boldsymbol{\beta}$ is an $N \times K$ matrix of linear regression coefficients (or *betas*), and $\boldsymbol{\varepsilon}_t$ is an N-vector of model disturbances such that $E[\boldsymbol{\varepsilon}_t \,|\, \mathbf{z}_{Kt}] = \mathbf{0}$ and $E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}$. If a portfolio of the K benchmark portfolios is mean-variance efficient (i.e. it minimizes variance for a given level of expected return), then $E[\mathbf{z}_t] = \boldsymbol{\beta} E[\mathbf{z}_{Kt}]$. These N conditions of the usual expected return-beta representation can be assessed by testing the null hypothesis:

$$H_E : \mathbf{a} = \mathbf{0}, \tag{2}$$

in the context of model (1). Observe that forming P portfolios of the test assets with weights $\boldsymbol{\omega}_p$ to deal with large N amounts to testing $H_0^p : \boldsymbol{\omega}_p'\mathbf{a} = \mathbf{0}$, for $p = 1, ..., P$, as opposed to H_E in (2). Gungor and Luger (2013) use a split-sample technique to formalize this approach without introducing any of the data-snooping size distortions (i.e. the appearance of statistical significance when the null hypothesis is true) discussed in Lo and MacKinlay (1990). It is clear, however, that $\mathbf{a} = \mathbf{0}$ implies $\boldsymbol{\omega}_p'\mathbf{a} = \mathbf{0}$, but not *vice versa*. Indeed, H_0^p may hold even if H_E is false.

GRS propose a multivariate F test of H_E that all the pricing errors comprising the vector $\mathbf{a}$ are jointly equal to zero. Their test assumes that the vectors of disturbance terms $\boldsymbol{\varepsilon}_t$, $t = 1, ..., T$, in (1) are independent and normally distributed around zero with a cross-sectional covariance matrix that is time-invariant, conditional on the $T \times K$ collection of factors $\mathbf{Z}_K = [\mathbf{z}_{K1}, ..., \mathbf{z}_{KT}]'$; i.e., $\boldsymbol{\varepsilon}_t \,|\, \mathbf{Z}_K \sim$ i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma})$. Under normality, the methods of maximum likelihood and ordinary least squares (OLS) yield the same unconstrained estimates of $\mathbf{a}$ and $\boldsymbol{\beta}$:

$$\begin{aligned}
\hat{\mathbf{a}} &= \bar{\mathbf{z}} - \hat{\boldsymbol{\beta}}\bar{\mathbf{z}}_{Kt}, \\
\hat{\boldsymbol{\beta}} &= \left[\sum_{t=1}^{T}(\mathbf{z}_t - \bar{\mathbf{z}})(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'\right]\left[\sum_{t=1}^{T}(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'\right]^{-1},
\end{aligned}$$

where $\bar{\mathbf{z}} = T^{-1}\sum_{t=1}^{T}\mathbf{z}_t$ and $\bar{\mathbf{z}}_K = T^{-1}\sum_{t=1}^{T}\mathbf{z}_{Kt}$. With $\hat{\mathbf{a}}$ and $\hat{\boldsymbol{\beta}}$ in hand, the unconstrained estimate of the disturbance covariance matrix is found as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{z}_t - \hat{\mathbf{a}} - \hat{\boldsymbol{\beta}}\mathbf{z}_{Kt}\right)\left(\mathbf{z}_t - \hat{\mathbf{a}} - \hat{\boldsymbol{\beta}}\mathbf{z}_{Kt}\right)'. \tag{3}$$

For the constrained model, which sets the vector $\mathbf{a}$ in (1) equal to zero, the estimates are

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_0 &= \left[\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_{Kt}'\right]\left[\sum_{t=1}^{T}\mathbf{z}_{Kt}\mathbf{z}_{Kt}'\right]^{-1}, \\
\hat{\boldsymbol{\Sigma}}_0 &= \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{z}_t - \hat{\boldsymbol{\beta}}_0\mathbf{z}_{Kt}\right)\left(\mathbf{z}_t - \hat{\boldsymbol{\beta}}_0\mathbf{z}_{Kt}\right)'.
\end{aligned} \tag{4}$$

The GRS test statistic for H_E is

$$J_{E,1} = \frac{(T-N-K)}{N}\left[1 + \bar{\mathbf{z}}_K'\hat{\boldsymbol{\Omega}}^{-1}\bar{\mathbf{z}}_K\right]^{-1}\hat{\mathbf{a}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{a}}, \tag{5}$$

where $\hat{\boldsymbol{\Omega}} = T^{-1}\sum_{t=1}^{T}(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)(\mathbf{z}_{Kt} - \bar{\mathbf{z}}_K)'$. Equivalently, the GRS test statistic can be written as

$$J_{E,1} = \frac{(T-N-K)}{N}\left[\frac{|\hat{\boldsymbol{\Sigma}}_0|}{|\hat{\boldsymbol{\Sigma}}|} - 1\right], \tag{6}$$

which shows that $J_{E,1}$ can be interpreted as an LR test (Campbell et al., 1997, Ch. 5). Under the null hypothesis H_E, the statistic $J_{E,1}$ follows a central F distribution with N degrees of freedom in the numerator and $(T-N-K)$ degrees of freedom in the denominator.

2.2 Mean-variance spanning

Mean-variance spanning occurs when the minimum-variance frontier of $\mathbf{r}_{Kt}$ (with $K \geq 2$) is the same as the minimum-variance frontier of $\mathbf{r}_{Kt}$ and $\mathbf{r}_t$. To formulate the spanning hypothesis, consider the statistical model:

$$\mathbf{r}_t = \mathbf{a} + \boldsymbol{\beta}\mathbf{r}_{Kt} + \boldsymbol{\varepsilon}_t, \tag{7}$$

where the disturbance vector $\boldsymbol{\varepsilon}_t$ now satisfies $E[\boldsymbol{\varepsilon}_t \,|\, \mathbf{r}_{Kt}] = \mathbf{0}$ and $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}$. Note that this model is specified in terms of returns, not excess returns. HK show that mean-variance spanning imposes on model (7) the $2N$ restrictions:

$$H_S : \mathbf{a} = \mathbf{0},\ \boldsymbol{\delta} = \mathbf{0}, \tag{8}$$

where $\boldsymbol{\delta} = \boldsymbol{\iota}_N - \boldsymbol{\beta}\boldsymbol{\iota}_K$ and $\boldsymbol{\iota}_i$ is an i-vector of ones. When H_S holds, then for every test asset, we can find a portfolio of the K benchmark assets that has the same mean (since $\mathbf{a} = \mathbf{0}$ and $\boldsymbol{\beta}\boldsymbol{\iota}_K = \boldsymbol{\iota}_N$) but a lower variance than the test asset (since $\mathrm{Cov}(\mathbf{r}_{Kt}, \boldsymbol{\varepsilon}_t') = \mathbf{0}$ and $\boldsymbol{\Sigma}$ is positive definite). In this case, the N test assets do not improve the mean-variance frontier already spanned by the K benchmark assets; see Kan and Zhou (2012) for more details.

Just like the GRS test, the one proposed by HK to assess the spanning hypothesis H_S assumes that the disturbances in (7) are normally distributed. Specifically, if we let the $T \times K$ collection of benchmark returns be collected in $\mathbf{R}_K = [\mathbf{r}_{K1}, ..., \mathbf{r}_{KT}]'$, then the exactness of the HK test rests on the assumption that $\boldsymbol{\varepsilon}_t \,|\, \mathbf{R}_K \sim$ i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma})$.

For the unconstrained model, the OLS parameter estimates resemble those for the GRS efficiency test. In the case of model (7), they are given by

$$\begin{aligned} \hat{\mathbf{a}} &= \bar{\mathbf{r}} - \hat{\boldsymbol{\beta}}\bar{\mathbf{r}}_{Kt}, \\ \hat{\boldsymbol{\beta}} &= \left[\sum_{t=1}^{T}(\mathbf{r}_t - \bar{\mathbf{r}})(\mathbf{r}_{Kt} - \bar{\mathbf{r}}_K)'\right]\left[\sum_{t=1}^{T}(\mathbf{r}_{Kt} - \bar{\mathbf{r}}_K)(\mathbf{r}_{Kt} - \bar{\mathbf{r}}_K)'\right]^{-1}, \end{aligned}$$

where $\bar{\mathbf{r}} = T^{-1}\sum_{t=1}^{T}\mathbf{r}_t$ and $\bar{\mathbf{r}}_K = T^{-1}\sum_{t=1}^{T}\mathbf{r}_{Kt}$. The unconstrained estimate of the disturbance covariance matrix is then

$$\hat{\mathbf{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{r}_t - \hat{\mathbf{a}} - \hat{\boldsymbol{\beta}}\mathbf{r}_{Kt}\right)\left(\mathbf{r}_t - \hat{\mathbf{a}} - \hat{\boldsymbol{\beta}}\mathbf{r}_{Kt}\right)'. \tag{9}$$

Following Campbell et al. (1997, Ch. 6), the restrictions in (8) can be imposed by partitioning the matrix $\boldsymbol{\beta}$ into $[\mathbf{b}_1, \mathbf{C}]$, where the $N \times 1$ vector $\mathbf{b}_1$ is the first column of $\boldsymbol{\beta}$ and $\mathbf{C}$ is the remainder $N \times (K-1)$ matrix. Conformably, we partition the vector $\mathbf{r}_{Kt}$ into its first row $\mathbf{r_{1t}}$ and its last $K-1$ rows $\mathbf{r}_{(K-1)t}$. With these partitions, the model in (7) can be written as

$$\mathbf{r}_t = \mathbf{a} + \mathbf{b}_1\mathbf{r}_{1t} + \mathbf{C}\mathbf{r}_{(K-1)t} + \boldsymbol{\varepsilon}_t,$$

and the constraint $\boldsymbol{\beta}\boldsymbol{\iota}_K = \boldsymbol{\iota}_N$ becomes $\mathbf{b}_1 + \mathbf{C}\boldsymbol{\iota}_{K-1} = \boldsymbol{\iota}_N$. Upon substitution of the restrictions $\mathbf{a} = \mathbf{0}$ and $\mathbf{b}_1 = \boldsymbol{\iota}_N - \mathbf{C}\boldsymbol{\iota}_{K-1}$, we obtain the constrained version:

$$\mathbf{r}_t - \boldsymbol{\iota}_N\mathbf{r}_{1t} = \mathbf{C}(\mathbf{r}_{(K-1)t} - \boldsymbol{\iota}_{K-1}\mathbf{r}_{1t}) + \boldsymbol{\varepsilon}_t. \tag{10}$$

The constrained estimates are then given by

$$\begin{aligned}
\hat{\mathbf{C}}_0 &= \left[\sum_{t=1}^{T}(\mathbf{r}_t - \boldsymbol{\iota}_N\mathbf{r}_{1t})(\mathbf{r}_{(K-1)t} - \boldsymbol{\iota}_{K-1}\mathbf{r}_{1t})'\right] \\
&\quad \times \left[\sum_{t=1}^{T}(\mathbf{r}_{(K-1)t} - \boldsymbol{\iota}_{K-1}\mathbf{r}_{1t})(\mathbf{r}_{(K-1)t} - \boldsymbol{\iota}_{K-1}\mathbf{r}_{1t})'\right]^{-1}, \\
\hat{\mathbf{b}}_{1,0} &= \boldsymbol{\iota}_N - \hat{\mathbf{C}}_0\boldsymbol{\iota}_{K-1}, \\
\hat{\mathbf{\Sigma}}_0 &= \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{r}_t - \hat{\boldsymbol{\beta}}_0\mathbf{r}_{Kt}\right)\left(\mathbf{r}_t - \hat{\boldsymbol{\beta}}_0\mathbf{r}_{Kt}\right)',
\end{aligned} \tag{11}$$

where $\hat{\boldsymbol{\beta}}_0 = [\hat{\mathbf{b}}_{1,0}, \hat{\mathbf{C}}_0]$.

The HK test statistic takes the following LR form:

$$J_S = \frac{(T-N-K)}{N}\left[\sqrt{\frac{|\hat{\mathbf{\Sigma}}_0|}{|\hat{\mathbf{\Sigma}}|}} - 1\right], \tag{12}$$

and, under the null hypothesis H_S, the statistic J_S follows a central F distribution with $2N$ degrees of freedom in the numerator and $2(T-N-K)$ degrees of freedom in the denominator. As Kan and Zhou (2012) and Peñaranda and Sentana (2012) point out, the original expression given in Huberman and Kandel (1987) contains a typo, whereby the square root is missing from the ratio of determinants. The correct expression shown in (12) is also found in Jobson and Korkie (1989).

3 Exact non-parametric tests

In this section we develop non-parametric bounds tests of efficiency and spanning that relax four assumptions of the exact $J_{E,1}$ and J_S tests discussed previously: (i) the assumption of independent disturbances, (ii) the assumption of identically distributed disturbances, (iii) the assumption of normally distributed disturbances, and (iv) the restriction that $N \leq T - K - 1$.

3.1 MLR framework

The specifications in (1) and (7) are special cases of the MLR model:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{\varepsilon}, \tag{13}$$

where $\mathbf{Y}$ is a $T \times N$ matrix of dependent variables, $\mathbf{X}$ is a $T \times (K+1)$ matrix of regressors, and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_T]'$ is the $T \times N$ matrix of model disturbances. The parameters are collected in $\mathbf{B} = [\mathbf{a}, \boldsymbol{\beta}]'$, a $(K+1) \times N$ matrix. In the case of model (1) we define $\mathbf{Y} = [\mathbf{z}_1, ..., \mathbf{z}_T]'$ and $\mathbf{X} = [\boldsymbol{\iota}_T, \mathbf{Z}_K]$, and for model (7) we take $\mathbf{Y} = [\mathbf{r}_1, ..., \mathbf{r}_T]'$ and $\mathbf{X} = [\boldsymbol{\iota}_T, \mathbf{R}_K]$. From here on we shall make explicit when necessary the dependence on $\mathbf{Y}$ to distinguish some statistics computed with the original sample of dependent variables from those computed with "bootstrap" samples, which later will be denoted by $\tilde{\mathbf{Y}}$.

In the terminology of Berndt and Savin (1977), the mean-variance efficiency and spanning hypotheses are prominent examples of so-called uniform linear (within equation) restrictions on the parameters of (13) which can be written as

$$H_0 : \mathbf{H}\mathbf{B} = \mathbf{D}, \tag{14}$$

where $\mathbf{H}$ is an $h \times (K+1)$ matrix of constants of rank h, and $\mathbf{D}$ is an $h \times N$ matrix of constants. Indeed, the efficiency hypothesis in (2) obtains upon setting $\mathbf{H} = [1, 0, ..., 0]$ and $\mathbf{D} = [0, ..., 0]$. For the spanning hypothesis in (8), we set

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 1 \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} 0 & ... & 0 \\ 1 & ... & 1 \end{bmatrix}.$$

The distinguishing feature of (14) is that the same hypothesis is tested on all the equations comprising the MLR system in (13); see Stewart (1997) for further discussion and examples of such restrictions.

With the MLR model in (13), the unrestricted OLS estimates and residuals are given as usual by

$$\begin{aligned} \hat{\mathbf{B}}(\mathbf{Y}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \\ \hat{\varepsilon}(\mathbf{Y}) &= \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}(\mathbf{Y}) = \mathbf{M}\mathbf{Y} = \mathbf{M}\boldsymbol{\varepsilon}, \end{aligned} \tag{15}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Here the i^{th} column of $\hat{\mathbf{B}}(\mathbf{Y}) = [\hat{\mathbf{B}}_1(\mathbf{Y}), ..., \hat{\mathbf{B}}_N(\mathbf{Y})]$ minimizes the i^{th} diagonal element of the sum-of-squares and cross-products matrix $\boldsymbol{\mathcal{E}} = (\mathbf{Y} - \mathbf{X}\mathbf{B})'(\mathbf{Y} - \mathbf{X}\mathbf{B})$. The estimated version of this matrix is

$$\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y}) = \hat{\varepsilon}'(\mathbf{Y})\hat{\varepsilon}(\mathbf{Y}). \tag{16}$$

Minimizing the diagonal sum-of-squares in $\boldsymbol{\mathcal{E}}$ subject to the restrictions in (14) yields the following constrained estimates and residuals:

$$\begin{aligned} \hat{\mathbf{B}}_0(\mathbf{Y}) &= \hat{\mathbf{B}}(\mathbf{Y}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'\left[\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'\right]^{-1}\left[\mathbf{H}\hat{\mathbf{B}}(\mathbf{Y}) - \mathbf{D}\right], \\ \hat{\varepsilon}_0(\mathbf{Y}) &= \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_0(\mathbf{Y}) = \mathbf{M}_0\mathbf{Y} = \mathbf{M}_0\boldsymbol{\varepsilon}, \end{aligned} \tag{17}$$

with $\mathbf{M}_0 = \mathbf{M} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'\left[\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'\right]^{-1}\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and where $\hat{\mathbf{B}}(\mathbf{Y})$ and $\mathbf{M}$ already appear in (15). The corresponding restricted residual sum-of-squares and cross-products matrix is

$$\hat{\boldsymbol{\mathcal{E}}}_0(\mathbf{Y}) = \hat{\varepsilon}_0'(\mathbf{Y})\hat{\varepsilon}_0(\mathbf{Y}). \tag{18}$$

The GRS and HK test statistics in (5) and (12) are constructed specifically for the mean-variance efficiency and spanning hypotheses in (2) and (8), respectively, which are special cases of H_0 in (14). More generally, some commonly used criteria for H_0 are: (i) the LR criterion (Bartlett, 1947; Wilks, 1932), (ii) the Lawley-Hotelling trace criterion (Bartlett, 1939; Hotelling, 1947, 1951; Lawley, 1938), (iii) the Bartlett-Nanda-Pillai trace criterion (Bartlett, 1939; Nanda, 1950; Pillai, 1955), and (iv) the maximum root criterion (Roy, 1953). All these test criteria are functions of the roots $m_1, ..., m_N$ of the determinantal equation:

$$\left|\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y}) - m\hat{\boldsymbol{\mathcal{E}}}_0(\mathbf{Y})\right| = 0,$$

where the matrices $\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y})$ and $\hat{\boldsymbol{\mathcal{E}}}_0(\mathbf{Y})$ are defined in (16) and (18), respectively. Under H_0 and when certain other conditions hold, Dufour and Khalaf (2002, Theorem 3.1) show that the joint distribution of $m_1, ..., m_N$ does not depend on nuisance parameters so that test criteria obtained as functions of these roots are pivotal (i.e. free of nuisance parameters). For this result to be operational, however, one needs to proceed like GRS, HK, and BDK by assuming a parametric distribution for the disturbances of the MLR model, e.g. $\boldsymbol{\varepsilon}_t \,|\, \mathbf{X} \sim$ i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma})$. This is precisely what we are trying to avoid here. Indeed, we wish to leave free the distribution of disturbances and to allow for time-varying conditional covariance structures of unknown form. Moreover the matrices $\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y})$ and $\hat{\boldsymbol{\mathcal{E}}}_0(\mathbf{Y})$ become singular when $N > T$, meaning that none of the usual statistics can be computed.

The test procedure we propose is also derived from (16) and (18), but does not require the determinants $|\hat{\boldsymbol{\Sigma}}|$ and $|\hat{\boldsymbol{\Sigma}}_0|$ seen in (6) and (12) for the GRS and HK tests, thereby avoiding the singularity problem. The distributional theory underlying our approach rests on a multivariate symmetry assumption, which includes the normal distribution assumed by GRS and HK as a special case. In the following, the symbol "$\overset{d}{=}$" stands for the equality in distribution.

Assumption 1 (Reflective symmetry). *The cross-sectional disturbance vectors* $\boldsymbol{\varepsilon}_t$, $t = 1, ..., T$, *which constitute the rows of* $\boldsymbol{\varepsilon}$ *in (13), are jointly continuous and reflectively symmetric so that*

$$(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, ..., \boldsymbol{\varepsilon}_T \,|\, \mathbf{X}) \overset{d}{=} (\pm\boldsymbol{\varepsilon}_1, \pm\boldsymbol{\varepsilon}_2, ..., \pm\boldsymbol{\varepsilon}_T \,|\, \mathbf{X}),$$

where $\pm\boldsymbol{\varepsilon}_t$ *means that the entire vector* $\boldsymbol{\varepsilon}_t$ *is assigned either a positive or negative sign with probability* $1/2$.

This assumption is satisfied whenever the vectors $\boldsymbol{\varepsilon}_t$, for $t = 1, ..., T$, are continuous and reflectively symmetric in the sense that $\boldsymbol{\varepsilon}_t \overset{d}{=} -\boldsymbol{\varepsilon}_t$, conditional on $\mathbf{X}$ and $\boldsymbol{\varepsilon}_\tau$, $\tau \neq t$. This reflective symmetry condition can be equivalently expressed in terms of the conditional density function as $f_t(\boldsymbol{\varepsilon}_t) = f_t(-\boldsymbol{\varepsilon}_t)$. Recall that a random variable x is symmetric around zero if and only if $x \overset{d}{=} -x$, so the symmetry assumption made here represents the most direct non-parametric extension of univariate symmetry; see Serfling (2006) for more concepts of multivariate symmetry. The class of distributions encompassed by Assumption 1 is very large and includes elliptically symmetric distributions, which play a very important role in mean-variance analysis because they guarantee full compatibility with expected utility maximization regardless of investor preferences (Berk, 1997; Chamberlain, 1983; Owen and Rabinovitch, 1983).

Several popular models of time-varying covariances, such as (possibly high-dimensional) multivariate GARCH or stochastic volatility models, satisfy the symmetry condition in Assumption 1. For example, suppose the conditional cross-sectional covariance matrix of model disturbances at time t is $\mathbf{\Sigma}_t$ and that the disturbances themselves are governed by

$$\boldsymbol{\varepsilon}_t = \mathbf{\Sigma}_t^{1/2} \boldsymbol{\eta}_t,$$

where $\{\boldsymbol{\eta}_t\}$ is an i.i.d. sequence of random vectors drawn from a symmetric distribution (e.g. multivariate normal or Student-t) and $\mathbf{\Sigma}_t^{1/2}$ is an $N \times N$ "square root" matrix such that $\mathbf{\Sigma}_t^{1/2} \mathbf{\Sigma}_t^{1/2} = \mathbf{\Sigma}_t$. If $\mathbf{\Sigma}_t^{1/2}$ and $\boldsymbol{\eta}_t$ are conditionally independent given $\mathbf{X}$ and $\boldsymbol{\varepsilon}_\tau$, $\tau \neq t$, then Assumption 1 is satisfied.

3.2 Test procedure

The proposed test procedure is based on equation-by-equation F statistics which can be computed from the unrestricted and restricted OLS estimates in (15) and (17). Consider the $N \times 1$ vector of F statistics:

$$\mathbf{F}(\mathbf{Y}) = \frac{\Big(\mathrm{diag}\big\{\hat{\boldsymbol{\mathcal{E}}}_0(\mathbf{Y})\big\} - \mathrm{diag}\big\{\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y})\big\}\Big)/h}{\mathrm{diag}\big\{\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y})\big\}/(T-K-1)}, \tag{19}$$

where $\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y})$ and $\hat{\boldsymbol{\mathcal{E}}}_0(\mathbf{Y})$ are the unrestricted and restricted $N \times N$ residual sum-of-squares and cross-products matrices in (16) and (18), respectively; $\mathrm{diag}\{\cdot\}$ returns the diagonal elements of a square matrix. Here h equals the number of rows of $\mathbf{H}$ in (14), and the division between the vectors appearing in the numerator and denominator is performed element-wise. The i^{th} element of the N-vector $\mathbf{F}(\mathbf{Y}) = [F_1(\mathbf{Y}), ..., F_N(\mathbf{Y})]'$ is the usual single-equation F statistic:

$$F_i(\mathbf{Y}) = \frac{\left(RSS_{0,i}(\mathbf{Y}) - RSS_i(\mathbf{Y})\right)/h}{RSS_i(\mathbf{Y})/(T-K-1)},$$

where the residual sum-of-squares terms $RSS_i(\mathbf{Y})$ and $RSS_{0,i}(\mathbf{Y})$ correspond to elements $[i,i]$ of $T\hat{\mathbf{\Sigma}}$ and $T\hat{\mathbf{\Sigma}}_0$, respectively; recall that $\hat{\mathbf{\Sigma}}$ is an unrestricted covariance matrix estimate as in (3) and (9), and $\hat{\mathbf{\Sigma}}_0$ is the restricted counterpart as in (4) and (11). Note that the degrees-of-freedom term $(T-K-1)/h$ could be omitted from (19) since it is just a constant under the proposed permutation approach.

The $F_i(\mathbf{Y})$ statistics comprising $\mathbf{F}(\mathbf{Y})$ could also be calculated from the restricted and unrestricted sum of squared residuals of the following models:

$$\mathbf{y}_i = \boldsymbol{\iota}_T a_i + \boldsymbol{x}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \tag{20}$$

for $i = 1, ..., N$, where $\mathbf{y}_i$ corresponds to column i of $\mathbf{Y}$ and $\boldsymbol{x}$ represents columns 2 through $K+1$ of $\mathbf{X}$. Here the scalar a_i is the i^{th} element of $\mathbf{a}$ and the K-vector $\boldsymbol{\beta}_i$ corresponds to the i^{th} column of $\boldsymbol{\beta}'$. When testing the efficiency hypothesis in (2) for instance, the $F_i(\mathbf{Y})$ statistics are related to the usual t statistic for $a_i = 0$. Indeed, let $\hat{a}_i$, $\hat{\boldsymbol{\beta}}_i$ denote the OLS estimates of a_i, $\boldsymbol{\beta}_i$ in (20) and consider the following squared t statistic:

$$t_i^2 = \frac{\hat{a}_i^2(\boldsymbol{\iota}_T'\mathbf{M}_{\boldsymbol{x}}\boldsymbol{\iota}_T)}{T\hat{\sigma}_i^2/(T-K-1)}, \tag{21}$$

where $\mathbf{M}_{\boldsymbol{x}} = \mathbf{I} - \boldsymbol{x}(\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'$ is the matrix that projects onto the orthogonal complement to the span of $\boldsymbol{x}$, and $\hat{\sigma}_i^2 = \hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_i/T$ with $\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \boldsymbol{\iota}_T\hat{a}_i - \boldsymbol{x}\hat{\boldsymbol{\beta}}_i$. In this case with $h = 1$, it is well known that $F_i(\mathbf{Y}) = t_i^2$ (Davidson and MacKinnon, 2004, p. 144).

The elements of $\mathbf{F}(\mathbf{Y})$ can be combined in different ways to obtain a joint test. A seemingly natural choice is simply to use the average F statistic $\bar{\mathbf{F}}(\mathbf{Y}) = \sum_{i=1}^N N^{-1}F_i(\mathbf{Y})$, which was proposed by Hwang and Satchell (2012) to test the mean-variance efficiency hypothesis; see also Affleck-Graves and McDonald (1990) for a similar idea. The average F statistic can be interpreted geometrically in relation to the GRS statistic. To see how, consider the problem of testing the Sharpe-Lintner version of the CAPM whereby the excess returns of the market portfolio constitute the only factor on the right-hand side of (1). Given (21), the average F statistic can be written as $\bar{\mathbf{F}}(\mathbf{Y}) = cN^{-1}\sum_{i=1}^N \hat{a}_i^2/\hat{\sigma}_i^2$ with $c = \boldsymbol{\iota}_T'\mathbf{M}_{\boldsymbol{x}}\boldsymbol{\iota}_T(T-K-1)/T$ in this case. Moreover, the results of GRS for the market model show that

$$\hat{\mathbf{a}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{a}} = \frac{\hat{\mu}_p^2}{\hat{\sigma}_p^2} - \frac{\hat{\mu}_b^2}{\hat{\sigma}_b^2},$$

where p refers to the *ex post* tangency portfolio constructed from the N test assets plus the benchmark market portfolio, b. As Hwang and Satchell (2012) explain, this last expression can be further decomposed as

$$\frac{\hat{\mu}_p^2}{\hat{\sigma}_p^2} - \frac{\hat{\mu}_b^2}{\hat{\sigma}_b^2} = \sum_{i=1}^N \frac{\hat{a}_i^2}{\hat{\sigma}_i^2} + \hat{\Delta},$$

where the discrepancy term $\hat{\Delta}$ is a function of the off-diagonal elements of the disturbance variance-covariance matrix. The left-hand side of this expression measures the *ex post* maximum pricing error, whereas $\sum_{i=1}^N \hat{a}_i^2/\hat{\sigma}_i^2$ on the right-hand side corresponds to the mean pricing error scaled by N. The discrepancy between these two measures depends on the sample estimates of the disturbance covariances. Hwang and Satchell (2012) argue that the average pricing error can be more informative than the maximum pricing error.

The constant term N^{-1} in the definition of the average F statistic plays no role under our proposed resampling scheme. So an equivalent test is obtained from $\mathbf{F}_1(\mathbf{Y}) = \sum_{i=1}^N F_i(\mathbf{Y})$, which corresponds to the 1-norm of the vector $\mathbf{F}(\mathbf{Y})$ since each component $F_i(\mathbf{Y}) \geq 0$. More generally, our approach applies to any p-norm defined as

$$\mathbf{F}_p(\mathbf{Y}) = \left(\sum_{i=1}^{N} F_i(\mathbf{Y})^p \right)^{1/p}, \tag{22}$$

for an integer $p \geq 1$. When $p \to \infty$, we obtain the maximum norm:

$$\mathbf{F}_{max}(\mathbf{Y}) = \max \Big\{ F_1(\mathbf{Y}), ..., F_N(\mathbf{Y}) \Big\}, \tag{23}$$

which picks out the individual F statistic suggesting the greatest violation of the null hypothesis. In Section 4 we compare the performance of our procedure based on (22) for $p = 1, ..., 4$ versus (23), and we find that power is generally increasing in p, meaning that $\mathbf{F}_{max}(\mathbf{Y})$ is the preferred statistic in the present context.

In our statistical framework built upon the reflective symmetry condition in Assumption 1, the distribution of $\mathbf{F}_p(\mathbf{Y})$ and $\mathbf{F}_{max}(\mathbf{Y})$ under H_0 depends on the values of $\mathbf{B}$ left unspecified by the null hypothesis. We deal with the presence of these nuisance parameters by establishing exact bounds to the H_0-distribution of the test statistics. Before doing so, it is worth emphasizing again that (22) and (23) can be calculated even if $N > T$, since the constituent $F_i(\mathbf{Y})$ statistics can be calculated one equation at a time. Observe also that $\mathbf{F}_p(\mathbf{Y})$ and $\mathbf{F}_{max}(\mathbf{Y})$ potentially have power increasing with both T and N. To see this, consider the efficiency hypothesis (2) and statistic (21). As the time series lengthens, the precision with which the a_is are estimated should improve, thereby increasing power. Furthermore, it will become more likely that non-zero a_is will be detected as more informative test assets are included in the MLR model, i.e. ones for which the "signal-to-noise" ratio in (21) is relatively large. The simulation study in Section 4 illustrates this point.

3.2.1 Building blocks

The bounds we establish to deal with the nuisance parameters that arise in our context (i.e. the elements of $\mathbf{B}$ not restricted by H_0) are based on a point null hypothesis of the form:

$$H_0^* : H_0 \text{ and } \mathbf{B} = \mathbf{B}^*, \tag{24}$$

where $\mathbf{B}^*$ are specified values that ensure compatibility with the null hypothesis, i.e. so that $H_0^* \subseteq H_0$. Define $\boldsymbol{\varepsilon}^* = \mathbf{Y} - \mathbf{X}\mathbf{B}^*$ and note that under H_0^* these residuals correspond to $\boldsymbol{\varepsilon}$, the true model disturbances. Observe also that H_0^* in (24) depends by construction on the choice of $\mathbf{B}^*$ and so do the H_0^*-restricted residuals, $\boldsymbol{\varepsilon}^*$.

Let $\tilde{\mathbf{s}} = [\tilde{s}_1, ..., \tilde{s}_T]'$ denote a T-vector comprising independent Bernoulli random variables such that $\Pr[\tilde{s}_t = 1] = \Pr[\tilde{s}_t = -1] = 1/2$, for all t, and define a bootstrap sample of dependent variables as

$$\tilde{\mathbf{Y}} = \mathbf{X}\mathbf{B}^* + \tilde{\mathbf{s}} \odot \boldsymbol{\varepsilon}^*, \tag{25}$$

where the notation $\tilde{\mathbf{s}} \odot \boldsymbol{\varepsilon}^*$ means that, for $t = 1, ..., T$, the scalar $\tilde{s}_t$ multiplies every element in row t of $\boldsymbol{\varepsilon}^*$. Doing so preserves the contemporaneous covariance structure among the row elements of $\boldsymbol{\varepsilon}^*$. Then, under H_0^* in (24) and conditional on $\mathbf{X}$, we have that $\mathbf{Y} \overset{d}{=} \tilde{\mathbf{Y}}$, for each of the 2^T possible realizations of $\tilde{\mathbf{Y}}$. From Theorem 1.3.7 in Randles and Wolfe (1979), we know that if $\mathbf{Y} \overset{d}{=} \tilde{\mathbf{Y}}$ and $\mathcal{F}(\cdot)$ is a measurable function (possibly vector-valued) defined on the common support of $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$, then $\mathcal{F}(\mathbf{Y}) \overset{d}{=} \mathcal{F}(\tilde{\mathbf{Y}})$. For our purposes, $\mathcal{F}(\mathbf{Y})$ will denote either $\mathbf{F}_p(\mathbf{Y})$ in (22) or $\mathbf{F}_{max}(\mathbf{Y})$ in (23).

Proposition 1 (Equally likely property). *Suppose the MLR model in (13) with Assumption 1 holds. Let $\tilde{\mathbf{Y}}$ be a bootstrap sample generated according to (25) for a given realization of $\tilde{\mathbf{s}}$ and consider the statistic $\mathcal{F}(\tilde{\mathbf{Y}})$ computed using the bootstrap sample. Then, under H_0^* in (24) and given $\mathbf{X}$, the 2^T values of $\mathcal{F}(\tilde{\mathbf{Y}})$ that can be obtained from all the possible realizations of $\tilde{\mathbf{s}}$ are equally likely values for $\mathcal{F}(\mathbf{Y})$, the original test statistic.*

This result is a straightforward extension to the multivariate case of the general methods described in Randles and Wolfe (1979, §11.1) for constructing distribution-free procedures. Proposition 1 shows that $\mathcal{F}(\mathbf{Y})$ is pivotal under H_0^*, meaning that its bootstrap distribution does not depend on any nuisance parameters. In principle, critical values could be found from the conditional distribution of $\mathcal{F}(\mathbf{Y})$ derived from the 2^T equally likely possibilities represented by $\mathcal{F}(\tilde{\mathbf{Y}})$. Determination of this distribution from a complete enumeration of all possible realizations of $\tilde{\mathbf{s}}$ is obviously impractical. To circumvent this problem and still obtain exact p-values, we use the Monte Carlo (MC) test technique (Barnard, 1963; Birnbaum, 1974; Dwass, 1957).

The MC test proceeds by generating $M-1$ random samples $\tilde{\mathbf{Y}}_1, ..., \tilde{\mathbf{Y}}_{M-1}$, each one according to (25). With each such sample, the statistic $\mathcal{F}(\cdot)$ is computed to yield $\mathcal{F}(\tilde{\mathbf{Y}}_m)$ for $m = 1, ..., M-1$. Proposition 1 implies that the M statistics $\mathcal{F}(\tilde{\mathbf{Y}}_1), ..., \mathcal{F}(\tilde{\mathbf{Y}}_{M-1}), \mathcal{F}(\mathbf{Y})$ are exchangeable under H_0^*. Note that the bootstrap distribution of the $\mathcal{F}(\cdot)$ statistic is discrete, meaning that ties among the resampled values can occur, at least theoretically. A test with size α can be obtained by applying the following tie-breaking rule (Dufour, 2006). Draw M i.i.d. variates U_m, $m = 1, ..., M$, from a continuous uniform distribution on $[0, 1]$, independently of the $\mathcal{F}(\cdot)$ statistics, randomly pair the U and $\mathcal{F}(\cdot)$ statistics, and compute the lexicographic rank of $\big(\mathcal{F}(\mathbf{Y}), U_M\big)$ according to

$$\tilde{R}_M\big[\mathcal{F}(\mathbf{Y})\big] = 1 + \sum_{m=1}^{M-1} \mathbb{I}\Big[\mathcal{F}(\mathbf{Y}) > \mathcal{F}(\tilde{\mathbf{Y}}_m)\Big] + \sum_{m=1}^{M-1} \mathbb{I}\Big[\mathcal{F}(\mathbf{Y}) = \mathcal{F}(\tilde{\mathbf{Y}}_m)\Big] \times \mathbb{I}\big[U_M > U_m\big], \qquad (26)$$

where $\mathbb{I}[A]$ is the indicator function of event A.

Upon recognizing that the pairs $\big(\mathcal{F}(\tilde{\mathbf{Y}}_1), U_1\big), ..., \big(\mathcal{F}(\tilde{\mathbf{Y}}_M), U_{M-1}\big), \big(\mathcal{F}(\mathbf{Y}), U_M\big)$ are exchangeable under H_0^*, we then know from Lemma 2.3 in Dufour (2006) that the lexicographic ranks are uniformly distributed over the integers $1, ..., M$; i.e., $\Pr\big[\tilde{R}_M[\mathcal{F}(\mathbf{Y})] = m\big] = 1/M$, for $m = 1, ..., M$. So the MC p-value can be defined as

$$\tilde{p}_M\big[\mathcal{F}(\mathbf{Y})\big] = \frac{M - \tilde{R}_M\big[\mathcal{F}(\mathbf{Y})\big] + 1}{M}, \qquad (27)$$

where $\tilde{R}_M\big[\mathcal{F}(\mathbf{Y})\big]$ is the rank of $\big(\mathcal{F}(\mathbf{Y}), U_M\big)$, given by (26). If αM is an integer, then the critical region $\tilde{p}_M\big[\mathcal{F}(\mathbf{Y})\big] \leq \alpha$ has exactly size α in the sense that

$$\Pr\Big[\,\tilde{p}_M\big[\mathcal{F}(\mathbf{Y})\big] \leq \alpha \,\big|\, \mathbf{X}\Big] = \alpha,$$

under the point null hypothesis H_0^* in (24).

The MC test of H_0^* paves the way for our proposed bounds tests of H_0, the hypothesis of interest. The basic idea is to obtain both a liberal test and a conservative test, each with nominal level α. The null hypothesis H_0 will be accepted when it is not rejected by the liberal test, and it will be rejected when the conservative test is significant.

3.2.2 Bounds MC tests

The liberal and conservative tests are based on the point null hypothesis in (24) specified with $\mathbf{B}^* = \hat{\mathbf{B}}_0$, the OLS estimate of $\mathbf{B}$ obtained under H_0. By construction, we have $\mathbf{H}\hat{\mathbf{B}}_0 = \mathbf{D}$ so that H_0^* is compatible with H_0. The H_0^*-residuals now correspond to those obtained under H_0 so that $\boldsymbol{\varepsilon}^* = \hat{\boldsymbol{\varepsilon}}_0$, where we have dropped the dependence on $\mathbf{Y}$ seen in (17).

Denote by $p_M^L\big[\mathcal{F}(\mathbf{Y})\big]$ the associated MC p-value computed according to (27), where the superscript indicates that this is a *liberal* p-value in the sense that $\Pr\big[\tilde{p}_M^L\big(\mathcal{F}(\mathbf{Y})\big) > \alpha \,|\, \mathbf{X}\big] \leq 1-\alpha$, under H_0. The logic of the decision rule which consists of accepting H_0 when $\tilde{p}_M^L\big(\mathcal{F}(\mathbf{Y})\big) > \alpha$ follows from the fact that $H_0^* \subseteq H_0$; i.e., if H_0^* is not rejected, then neither is H_0. Dufour (2006) refers to such a test as a *local* MC test.

The conservative test also focuses on $H_0^*: H_0$ and $\mathbf{B} = \hat{\mathbf{B}}_0$, but introduces a test statistic specifically for this point null hypothesis. Let the residual sum-of-squares and cross-products matrix at H_0^* be written as $\boldsymbol{\mathcal{E}}^* = \boldsymbol{\varepsilon}^{*\prime}\boldsymbol{\varepsilon}^*$, which corresponds to (18), and consider the $N \times 1$ vector of test statistics:

$$\mathbf{F}^C(\mathbf{Y}) = \frac{\Big(\text{diag}\big\{\boldsymbol{\mathcal{E}}^*\big\} - \text{diag}\big\{\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y})\big\}\Big)/h}{\text{diag}\big\{\hat{\boldsymbol{\mathcal{E}}}(\mathbf{Y})\big\}/(T-K-1)},$$

whose superscript stands for *conservative*. When computed with the original sample $\mathbf{Y}$, we have $\mathbf{F}^C(\mathbf{Y}) = \mathbf{F}(\mathbf{Y})$ since we set $\mathbf{B}^* = \hat{\mathbf{B}}_0$. Observe also that $\text{diag}\big\{\boldsymbol{\varepsilon}^{*\prime}\boldsymbol{\varepsilon}^*\big\} = \text{diag}\big\{(\tilde{\mathbf{s}} \odot \boldsymbol{\varepsilon}^*)'(\tilde{\mathbf{s}} \odot \boldsymbol{\varepsilon}^*)\big\}$, for any possible realization of $\tilde{\mathbf{s}}$. So with any bootstrap sample $\tilde{\mathbf{Y}}$ generated according to (25), the following inequalities hold:

$$\text{diag}\big\{\boldsymbol{\mathcal{E}}^*\big\} \geq \text{diag}\big\{\hat{\boldsymbol{\mathcal{E}}}_0(\tilde{\mathbf{Y}})\big\} \geq \text{diag}\big\{\hat{\boldsymbol{\mathcal{E}}}(\tilde{\mathbf{Y}})\big\}, \tag{28}$$

where the comparisons are element-wise. This follows from the fact that an OLS restricted residual sum of squares cannot be smaller than a less restricted one (Davidson and MacKinnon, 2004, §3.8). The inequalities in (28) imply that $\mathbf{F}(\tilde{\mathbf{Y}}) \leq \mathbf{F}^C(\tilde{\mathbf{Y}})$.

As we did before in (22) or (23), the conservative statistics comprising the $\mathbf{F}^C(\cdot)$ vector can be aggregated using any p-norm. In obvious notation, let $\mathcal{F}^C(\cdot)$ denote either $\mathbf{F}_p^C(\cdot)$ or $\mathbf{F}_{max}^C(\cdot)$. The foregoing discussion shows that $\mathcal{F}(\cdot) \leq \mathcal{F}^C(\cdot)$ and hence

$$\Pr[\mathcal{F}(\cdot) > \zeta] \leq \Pr[\mathcal{F}^C(\cdot) > \zeta], \tag{29}$$

for any $\zeta \in \mathbb{R}$. To see how this result will be exploited, let ζ_α be a critical value such that $\Pr[\mathcal{F}(\mathbf{Y}) > \zeta_\alpha \,|\, \mathbf{X}] = \alpha$ when H_0 holds; similarly define ζ_α^C via $\Pr[\mathcal{F}^C(\mathbf{Y}) > \zeta_\alpha^C \,|\, \mathbf{X}] = \alpha$ under H_0^*. It follows from (29) that $\zeta_\alpha \leq \zeta_\alpha^C$, meaning that $\Pr[\mathcal{F}(\mathbf{Y}) > \zeta_\alpha^C \,|\, \mathbf{X}] \leq \alpha$ when $\mathcal{F}(\mathbf{Y})$ follows its H_0-distribution. The consequence is that $\mathcal{F}(\mathbf{Y}) > \zeta_\alpha^C \Rightarrow \mathcal{F}(\mathbf{Y}) > \zeta_\alpha$. In words, if the joint F bounds test based on ζ_α^C is significant, then for sure the exact joint F test based on ζ_α is also significant at level α. In order to operationalize the bounds test, we use the MC test technique.

Proposition 2 (Bounds MC p-values). *Suppose the MLR model in (13) with Assumption 1 holds. Further, consider a statistic $\mathcal{F}(\mathbf{Y})$ for testing H_0 and the corresponding conservative test statistic $\mathcal{F}^C(\mathbf{Y})$. Define liberal and conservative MC p-values as*

$$\tilde{p}_M^L\big[\mathcal{F}(\mathbf{Y})\big] = \frac{M - \tilde{R}_M\big[\mathcal{F}(\mathbf{Y})\big] + 1}{M} \quad \text{and} \quad \tilde{p}_M^C\big[\mathcal{F}(\mathbf{Y})\big] = \frac{M - \tilde{R}_M^C\big[\mathcal{F}(\mathbf{Y})\big] + 1}{M},$$

where $\tilde{R}_M\big[\mathcal{F}(\mathbf{Y})\big]$ *and* $\tilde{R}_M^C\big[\mathcal{F}(\mathbf{Y})\big]$ *are the lexicographic ranks of* $\mathcal{F}(\mathbf{Y})$ *among* $\mathcal{F}(\tilde{\mathbf{Y}}_m)$ *and* $\mathcal{F}^C(\tilde{\mathbf{Y}}_m)$*,* $m = 1, ..., M-1$*, respectively. Here the* $\tilde{\mathbf{Y}}_m$*s are bootstrap samples generated according to (25) and the lexicographic ranks are computed as*

$$\begin{aligned}\tilde{R}_M\big[\mathcal{F}(\mathbf{Y})\big] &= 1 + \sum_{m=1}^{M-1} \mathbb{I}\left[\mathcal{F}(\mathbf{Y}) > \mathcal{F}(\tilde{\mathbf{Y}}_m)\right] + \sum_{m=1}^{M-1} \mathbb{I}\left[\mathcal{F}(\mathbf{Y}) = \mathcal{F}(\tilde{\mathbf{Y}}_m)\right] \times \mathbb{I}\big[U_M > U_m\big], \\ \tilde{R}_M^C\big[\mathcal{F}(\mathbf{Y})\big] &= 1 + \sum_{m=1}^{M-1} \mathbb{I}\left[\mathcal{F}(\mathbf{Y}) > \mathcal{F}^C(\tilde{\mathbf{Y}}_m)\right] + \sum_{m=1}^{M-1} \mathbb{I}\left[\mathcal{F}(\mathbf{Y}) = \mathcal{F}^C(\tilde{\mathbf{Y}}_m)\right] \times \mathbb{I}\big[U_M > U_m\big],\end{aligned}$$

where U_m*,* $m = 1, ..., M$*, are i.i.d. uniform variates on* $[0, 1]$*, independently of the F statistics. If* αM *is an integer, then* $\Pr\left[\tilde{p}_M^L\big(\mathcal{F}(\mathbf{Y})\big) > \alpha \,|\, \mathbf{X}\right] \leq 1 - \alpha$ *and* $\Pr\left[\tilde{p}_M^C\big(\mathcal{F}(\mathbf{Y})\big) \leq \alpha \,|\, \mathbf{X}\right] \leq \alpha$*, under* H_0 *in (14).*

This result follows from Proposition 2.4 in Dufour (2006) on the validity of MC tests for general statistics. An important remark about Proposition 2 above is that a given bootstrap sample $\tilde{\mathbf{Y}}_m$ serves to compute both $\mathcal{F}(\tilde{\mathbf{Y}}_m)$ and $\mathcal{F}^C(\tilde{\mathbf{Y}}_m)$. Furthermore, the same collection of uniform draws $U_1, ..., U_M$ should be used to compute both $\tilde{R}_M\big[\mathcal{F}(\mathbf{Y})\big]$ and $\tilde{R}_M^C\big[\mathcal{F}(\mathbf{Y})\big]$. These requirements ensure that the liberal and conservative MC p-values do not yield conflicting answers.

The result in Proposition 2 suggests the following MC bounds test of $H_0 : \mathbf{HB} = \mathbf{D}$ at level α:

$$\begin{cases} \text{Reject } H_0 \text{ when } \tilde{p}_M^C\big(\mathcal{F}(\mathbf{Y})\big) \leq \alpha; \\ \text{Accept } H_0 \text{ when } \tilde{p}_M^L\big(\mathcal{F}(\mathbf{Y})\big) > \alpha; \\ \text{Consider the test inconclusive, otherwise.} \end{cases} \tag{30}$$

The logic of this decision rule is the same as with the well-known bounds test of Durbin and Watson (1950, 1951) for autocorrelated disturbances in regression models. For further discussion and examples of such bounds procedures, see Dufour (1989, 1990), Dufour and Kiviet (1996), Stewart (1997), and Dufour and Khalaf (2002).

4 Simulation study

This section presents the results of simulation experiments to examine the performance of the proposed procedure for testing the mean-variance efficiency and spanning hypotheses. Here we simply use $\mathbf{F}_p$ and $\mathbf{F}_{max}$ to refer to the MC test procedure based on the statistics in (22) and (23).

The MC tests are performed at the nominal $\alpha = 0.05$ significance level with $M - 1 = 99$ random samples.

We consider the MLR model in (13) given for convenience again here as

$$\mathbf{y}_t = \mathbf{a} + \mathbf{B}\mathbf{x}_{Kt} + \boldsymbol{\varepsilon}_t, \tag{31}$$

for $t = 1, ..., T$, where $\mathbf{y}_t$ and $\mathbf{x}_{Kt}$ are interpreted as vectors of excess returns when we examine the efficiency hypothesis, and simply as returns in the case of mean-variance spanning. The benchmark portfolio returns are generated as standard normal variables, which is a rather innocuous choice since the proposed tests are conditional on the realized values of $\mathbf{x}_{Kt}$. Here we let $K = 1, 3$ and the elements of $\mathbf{B}$ are uniformly distributed over $[0.5, 1.5]$. The model disturbances in (31) have the following factor structure:

$$\boldsymbol{\varepsilon}_t = \boldsymbol{\varphi} f_t + \lambda \boldsymbol{e}_t, \tag{32}$$

where $\boldsymbol{e}_t \sim N(\mathbf{0}, \mathbf{I})$. The common factor f_t evolves according to a stochastic volatility process of the form:

$$f_t = \exp(h_t/2)\eta_t, \text{ with } h_t = \phi h_{t-1} + \xi_t, \tag{33}$$

where η_t follows a Student-t distribution with 3 degrees of freedom (standardized to have unit variance); and ξ_t follows a normal distribution with mean zero and variance 0.1. The specification in (32) and (33) implies that $\text{Var}(\varepsilon_{it} \,|\, \mathfrak{F}_{t-1}) = \varphi_i^2 \text{Var}(f_t \,|\, \mathfrak{F}_{t-1}) + \lambda^2$ and $\text{Cov}(\varepsilon_{it}, \varepsilon_{jt} \,|\, \mathfrak{F}_{t-1}) = \varphi_i \varphi_j \text{Var}(f_t \,|\, \mathfrak{F}_{t-1})$, where $\mathfrak{F}_t$ is the time-t information set. So the autoregressive parameter ϕ determines the persistence over time of shocks to the cross-sectional covariance structure. We examine two polar cases by setting the autoregressive parameter in (33) as either $\phi = 0$ (no persistence) or $\phi = 0.99$ (nearly integrated), and the recursion is started with $h_1 = \xi_1$. The power of the efficiency and spanning tests depends on the disturbance variance through the values of $\boldsymbol{\varphi}$ and λ in (32). We draw the elements of $\boldsymbol{\varphi}$ as $\varphi_i \sim U[0, \varphi_{\max}]$ and we consider the following pairs of values for $(\varphi_{\max}, \lambda)$: $(0, 0.8)$ and $(1, 0.2)$. With this design, the normality assumption made by GRS and HK is satisfied when $\varphi_{\max} = 0$. When examining the power of the efficiency tests, the elements of $\mathbf{a}$ are generated as $a_i \sim U[-0.1, 0.1]$. Recall that the spanning hypothesis places restrictions on the elements of both $\mathbf{a}$ and $\boldsymbol{\delta}$. So we investigate the power of the spanning tests under two scenarios: (i) $a_i \sim U[-0.1, 0.1]$, $\delta_i = 0$, and (ii) $a_i = 0$, $\delta_i \sim U[-0.2, 0.2]$. Finally, we let the sample size vary as $T = 60, 100$ and the number of test assets as $N = 50, 100, 200, 400$.

Even though we are mainly concerned with testing mean-variance efficiency and spanning when $N > T$, we nevertheless include some cases in which the GRS $J_{E,1}$ and the HK J_S tests are computable. As we mentioned in the introduction, Pesaran and Yamagata (2012) also develop tests of the efficiency hypothesis (2) in large N situations. Of the two tests they propose, the one that allows for the presence of cross-sectional correlations is computed as

$$J_{E,2} = \frac{N^{-1/2}\sum_{i=1}^{N}\left(t_i^2 - \frac{v}{v-2}\right)}{\left(\frac{v}{v-2}\right)\sqrt{\frac{2(v-1)}{v-4}\left[1+(N-1)\hat{\rho}^2\right]}},$$

where t_i^2 is the squared t statistic defined in (21), $v = T - K - 1$ is a degrees-of-freedom term, and $\hat{\rho}^2$ is a threshold estimator of the average squares of pairwise disturbance correlations given by

$$\hat{\rho}^2 = \frac{2}{N(N-1)}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\hat{\rho}_{ij}^2\mathbb{I}[v\hat{\rho}_{ij}^2 \geq \theta_N],$$

with $\hat{\rho}_{ij} = \hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_j/\sqrt{(\hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_i)(\hat{\boldsymbol{\varepsilon}}_j'\hat{\boldsymbol{\varepsilon}}_j)}$; recall that $\hat{\boldsymbol{\varepsilon}}_i$ are the OLS residuals from (20). PY suggest selecting the threshold value via $\sqrt{\theta_N} = \Phi^{-1}\big(1 - p_N/2\big)$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function and $p_N = \alpha/(N-1)$. Assuming, as in GRS, that $\boldsymbol{\varepsilon}_t \,|\, \mathbf{X} \sim$ i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma})$, as well as some other regularity conditions, PY show that $J_{E,2}$ is asymptotically $N(0,1)$ when mean-variance efficiency holds. PY argue that this asymptotic result continues to hold even for non-Gaussian disturbances, assuming that N grows at a sufficiently slower rate than T.

The empirical size and power (in percentage) of $J_{E,1}$, $J_{E,2}$, and the proposed $\mathbf{F}_p$, for $p = 1, ..., 4$, and $\mathbf{F}_{max}$ tests are reported in Tables 1 and 2 for $K = 1$ and 3, respectively. Table 3 compares the new tests for the spanning hypothesis with the HK test, J_S. In each table, the symbol "-" indicates cases when the GRS test or the HK test is not computable and the entries set in bold show the most powerful tests. From Panel A of each table, we see that all the tests respect the nominal level constraint. Indeed, the empirical size of the conservative MC tests is always strictly less than 5%, as expected from the developed theory, while that of $J_{E,1}$, $J_{E,2}$, and J_S stays relatively close to 5%.

The simulation results clearly show the power of $\mathbf{F}_p$ increasing in p with the best power achieved by $\mathbf{F}_{max}$. Tables 1 and 2 further reveal that the power of $J_{E,2}$ is far better than that of $J_{E,1}$ and the proposed tests when the model disturbances are i.i.d. both over time and in the cross-section ($\phi = 0$, $\varphi_{\max} = 0$). Note that increasing N from 60 to 100 with i.i.d. disturbances yields relatively little additional power for the new tests, if any at all. When the disturbances are cross-sectionally correlated, however, $J_{E,2}$ is dominated by one of the other tests. The pattern is that $J_{E,1}$ is the better test when it is computable. But as soon as $N > T$, the power ranking has $\mathbf{F}_{max}$ in first place quite far ahead of the $J_{E,2}$ test. For instance, when $\phi = 0, \varphi_{\max} = 1, \lambda = 0.2$ and $T = 60, N = 400$, the power of $\mathbf{F}_{max}$ is about 96% while that of $J_{E,2}$ is about 31%. The reason is that the theory underlying the use of the threshold estimator in $J_{E,2}$ assumes that the correlation matrix is sparse, i.e. with only a finite number of non-zero correlations that vanish as N grows. On the contrary, Assumption 1 allows for any correlation structure. The power results in Tables 1 and 2 are all the more remarkable considering the distribution-free nature of the new tests. A comparison of Tables 1 and 2 reveals that all the tests tend to have relatively lower power when K increases. The reason why the bounds tests become more conservative is that increasing K from 1 to 3 triples the number of nuisance parameters in the testing problem, thereby increasing the inequalities in (28).

Table 3 tells a similar story when examining the mean-variance spanning hypothesis. Here we see that the J_S test is preferred when $N < T$, but a larger number of test assets leaves the new tests as the only ones available to assess the spanning hypothesis. Table 3 again shows that the $\mathbf{F}_p$ and $\mathbf{F}_{max}$ tests have low power in the i.i.d. case. As before, however, we see that the presence of cross-sectional correlation among the model disturbances restores the power of the new tests. Our general conclusion is that $\mathbf{F}_{max}$ has the better power when these correlations become stronger. We proceed next to an empirical illustration of the $\mathbf{F}_{max}$ test.

5 Empirical application

Our empirical illustration uses monthly returns on 452 individual stocks traded on the NYSE, AMEX, and NASDAQ markets for the 39-year period from January 1973 to December 2011 (468 months). These are all the stocks for which data is available in the Centre for Research in Securities Prices (CRSP) monthly files for this sample period. We use the one-month U.S. Treasury bill as the risk-free asset when forming excess returns. It is also quite common in the empirical finance literature to test asset pricing models over subperiods owing to concerns about parameter stability. So here we also divide the 39 years into seven 5-year, one 4-year, three 10-year, and one 9-year subperiods. This breakdown follows Campbell et al. (1997, Ch. 5), Gungor and Luger (2009, 2013), and Ray et al. (2009). As in Pesaran and Yamagata (2012), we complement the subperiod analysis by performing the tests using the returns observed over 60-month rolling windows.

5.1 Efficiency assessment

We assess the efficiency hypothesis first in the context of the Sharpe-Lintner version of the CAPM using the excess returns of a value-weighted stock market index of all stocks listed on the NYSE, AMEX, and NASDAQ as proxy for the market risk factor. Second, we test the more general Fama and French (1993) three-factor model, which adds two risk factors to the CAPM specification: (i) the average returns on three small capitalization portfolios minus the average return on three big market capitalization portfolios, and (ii) the average return on two value portfolios minus the average return on two growth portfolios.

Table 4 reports the p-values of the mean-variance efficiency tests, where columns 2–4 pertain to the CAPM and columns 5–7 are for the Fama-French model. The new test procedure is applied here with the $\mathbf{F}_{max}$ statistic using $M = 500$, so the smallest possible MC p-value is 0.2%. Based on the decision rule in (30) with $\alpha = 5\%$, we report only the conservative MC p-value if $\tilde{p}_M^C\big(\mathbf{F}_{max}(\mathbf{Y})\big) \leq \alpha$, whereas the liberal MC p-value is reported when $\tilde{p}_M^L\big(\mathbf{F}_{max}(\mathbf{Y})\big) > \alpha$. Recall that the MC tests may yield an inconclusive outcome when $\tilde{p}_M^C\big(\mathbf{F}_{max}(\mathbf{Y})\big) > \alpha$ and $\tilde{p}_M^L\big(\mathbf{F}_{max}(\mathbf{Y})\big) \leq \alpha$. In these inconclusive cases, we report both the conservative and liberal MC p-values. We set in bold the table entries that correspond to a rejection of the null hypothesis at the 5% significance level.

Looking at the full sample results, we see that the GRS $J_{E,1}$ and the MC test do not reject efficiency in the CAPM, but the $J_{E,2}$ test indicates a decisive rejection of that null hypothesis. In the subperiods, the $J_{E,2}$ and $\mathbf{F}_{max}$ test outcomes agree in most cases, except in the 5-year subperiod 1/03–12/07 and two of the three 10-year subperiods (1/73–12/82, 1/83–12/92). Overall, the CAPM finds strong support from the MC test. This is further corroborated by the 60-month rolling-window p-values shown in Panel (a) of Figure 1. We clearly see the p-values staying above the cutoff line, indicating non-rejections of the CAPM.

Turning now to the Fama-French model, we see from Table 4 that mean-variance efficiency finds broad support across tests and time periods. This can also be gleaned from Panel (b) of Figure 1 where the rolling-window MC p-values, while again fluctuating a lot from month to month, never indicate a rejection of the efficiency hypothesis. Given that the CAPM is never rejected in the subperiods by the MC test, it is then entirely coherent to find that the Fama-French portfolios are generally efficient as well since the latter three-factor model nests the single-factor model. On the contrary, in the 5-year subperiod 1/88–12/92 and the 9-year subperiod 1/03–12/11, the PY test indicates a non-rejection of the CAPM at the 5% significance level but then surprisingly rejects the Fama-French model. The message to take away from Figure 1 is that even though it never quite dips below the 5% cutoff line, the new non-parametric test displays non-trivial power with empirical p-values showing a great deal of variation and often times moving towards a rejection of the mean-variance efficiency hypothesis.

5.2 Spanning assessment

The mean-variance spanning hypothesis is evaluated in the context of the Fama-French model using the HK test in (12) and the non-parametric $\mathbf{F}_{max}$ test. From the results in Table 5, it is immediately clear that mean-variance spanning is strongly rejected, suggesting that the individual stocks can improve the efficiency frontier spanned by the three Fama-French portfolios. Over the seven 5-year and the one 4-year subperiods, the MC test shows rejections more than half the time. On the other hand, the spanning hypothesis is decisively rejected in the full 39-year period, in the two of the three 10-year subperiods (1/73–12/82, 1/83–12/92), and in the 9-year subperiod (1/03–12/11), which suggests that mean-variance spanning is less likely to hold when assessed over longer periods.

Focusing on the 60-month rolling-window results, Figure 2 shows that $\mathbf{F}_{max}$ doesn't reject the spanning hypothesis in the 1980s and from the late-1990s until about 2005. The rather sustained rejections occurring towards the end of the sample period are particularly noteworthy, since the correlations among asset returns generally tend to increase during high volatility periods. For instance, the results suggest that an investor holding the Fama-French portfolios would have benefited from diversification during the recent financial crisis.

6 Conclusion

The starting point for the econometric analysis of linear factor asset pricing models, such as the CAPM or APT models, is an assumption about the time-series behavior of returns. For example, the well-known GRS and HK exact tests of mean-variance efficiency and spanning, respectively, assume that returns, conditional on the factor portfolio realizations, are i.i.d. through time and jointly multivariate normal. This assumption is at odds with a huge body of empirical evidence as it precludes not only non-normalities, but also multivariate GARCH-type effects. Another shortcoming of these tests is that they can no longer be computed when the number of test assets (i.e. the number of equations in the MLR) is too large relative to the available time series. This is rather unfortunate since it is natural to try to use as many test assets as possible in order to boost test power. Indeed, as the test asset universe expands, it should become more likely that violations of the null hypothesis will be detected.

In this paper we have proposed an exact test procedure that overcomes these problems, without imposing any parametric assumptions on the MLR disturbance distribution. Our statistical framework leaves open the possibility of unknown forms of time-varying non-normalities and many other distribution heterogeneities, such as time-varying conditional variances and covariances. We derived liberal and conservative bounds on the null distribution of joint F statistics in order to deal with the presence of nuisance parameters, and we have shown how to implement the exact test procedure with Monte Carlo resampling techniques. The null distribution of the proposed bounds tests is obtained conditional on the absolute values of the model residuals, since only their signs are randomized according to (25). The Lehmann and Stein (1949) impossibility theorem shows that such sign tests are the only ones which yield valid inference when one wishes to remain completely agnostic about disturbance distribution heterogeneities; see also Dufour (2003) for more on this point. It is important to bear in mind that even though we found the GRS, PY, and HK tests to be fairly robust to deviations from their underlying assumption of i.i.d. $N(\mathbf{0}, \mathbf{\Sigma})$ model disturbance vectors, there is no theoretical guarantee that this would always be the case.

A very appealing feature of our approach is that it remains applicable no matter the number of equations in the MLR. And in fact the results of our simulation study show that the power of the proposed tests potentially increases along both the time and cross-sectional dimensions. This makes the new test procedure a very useful way of assessing mean-variance efficiency and spanning, especially when the MLR includes a large number of correlated disturbances. Finally, note that our approach applies not only to these hypotheses, but to any uniform linear restriction in the MLR model. Investigating the performance of our test procedure for other MLR restrictions is the subject of ongoing research.

Acknowledgements

We thank Antonio Diez de los Rios, Jean-Marie Dufour, Eleonora Granziera, two anonymous referees, and the joint editor Rong Chen for helpful comments. Earlier versions of this article were presented at the Bank of Canada 2013 Fellowship Learning Exchange Workshop, the 2013 Meetings of the Canadian Economics Association, the 2013 Meetings of the Society for Computational Economics, the 2013 Meetings of the Midwest Econometrics Group, and Syracuse University. The views in this article are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Bank of Canada.

References

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

Affleck-Graves, J., McDonald, B., 1990. Multivariate tests of asset pricing: the comparative power of alternative statistics. Journal of Financial and Quantitative Analysis 25, 163–185.

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

Bartlett, M., 1939. A note on tests of significance in multivariate analysis. Mathematical Proceedings of the Cambridge Philosophical Society 35, 180–185.

Bartlett, M., 1947. Multivariate analysis. Journal of the Royal Statistical Society (Supplement) 9, 176–197.

Beaulieu, M.-C., Dufour, J.-M., Khalaf, L., 2007. Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors. Journal of Business and Economic Statistics 25, 398–410.

Beaulieu, M.-C., Dufour, J.-M., Khalaf, L., 2010. Asset-pricing anomalies and spanning: multivariate and multifactor tests with heavy-tailed distributions. Journal of Empirical Finance 17, 763–782.

Berk, J., 1997. Necessary conditions for the CAPM. Journal of Economic Theory 73, 245–257.

Berndt, E., Savin, E., 1977. Conflict among criteria for testing hypotheses in the multivariate linear regression model. Econometrica 45, 1263–1277.

Birnbaum, Z., 1974. Computers and unconventional test statistics. In: Proschan, F., Serfling, R. (Eds.), Reliability and Biometry. SIAM, Philadelphia, pp. 441–458.

Blattberg, R., Gonedes, N., 1974. A comparison of the stable and Student distributions as statistical models for stock prices. Journal of Business 47, 244–280.

Campbell, J., Lo, A., MacKinlay, A., 1997. The Econometrics of Financial Markets. Princeton University Press.

Chamberlain, G., 1983. A characterization of the distributions that imply mean-variance utility functions. Journal of Economic Theory 29, 185–201.

Cheung, C., Kwan, C., Mountain, D., 2009. On the nature of mean-variance spanning. Finance Research Letters 6, 106–113.

Davidson, R., MacKinnon, J., 2004. Econometric Theory and Methods. Oxford University Press.

DeRoon, F., Nijman, T., 2001. Testing for mean-variance spanning: a survey. Journal of Empirical Finance 8, 111–155.

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

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

Dufour, J.-M., 2003. Identification, weak instruments, and statistical inference in econometrics. Canadian Journal of Economics 36, 767–808.

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

Dufour, J.-M., Khalaf, L., 2001. Monte Carlo test methods in econometrics. In: Baltagi, B. (Ed.), A Companion to Theoretical Econometrics. Basil Blackwell, Oxford, UK, pp. 494–510.

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

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

Durbin, J., Watson, G., 1950. Testing for serial correlation in least squares regression I. Biometrika 37, 409–420.

Durbin, J., Watson, G., 1951. Testing for serial correlation in least squares regression II. Biometrika 38, 159–178.

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

Fama, E., 1965. The behavior of stock-market prices. Journal of Business 38, 34–105.

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

Gibbons, M., Ross, S., Shanken, J., 1989. A test of the efficiency of a given portfolio. Econometrica 57, 1121–1152.

Gungor, S., Luger, R., 2009. Exact distribution-free tests of mean-variance efficiency. Journal of Empirical Finance 16, 816–829.

Gungor, S., Luger, R., 2013. Testing linear factor pricing models with large cross-sections: a distribution-free approach. Journal of Business and Economic Statistics 31, 66–77.

Hotelling, H., 1947. Multivariate quality control. In: Eisenhart, C., Hastay, M., Wallis, W. (Eds.), Techniques of Statistical Analysis. McGraw-Hill, New York.

Hotelling, H., 1951. A generalized t test and measure of multivariate dispersion. In: Neyman, J. (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. The University of California Press, Berkeley, California, pp. 23–41.

Huberman, G., Kandel, S., 1987. Mean-variance spanning. Journal of Finance 42, 873–888.

Hwang, S., Satchell, S., 2012. Testing linear factor models on individual stocks using the average F-test. European Journal of Finance, 1–36.

Jobson, J., Korkie, B., 1989. A performance interpretation of multivariate tests of asset set intersection, spanning, and mean-variance efficiency. Journal of Financial and Quantitative Analysis 24, 185–204.

Kan, R., Zhou, G., 2012. Tests of mean-variance spanning. Annals of Economics and Finance 13, 145–193.

Lawley, D., 1938. A generalization of Fisher's z test. Biometrika 30, 180–187.

Lehmann, E., Stein, C., 1949. On the theory of some non-parametric hypotheses. Annals of Mathematical Statistics 20, 28–45.

Liang, B., 2000. Portfolio formation, measurement errors, and beta shifts: a random sampling approach. Journal of Financial Research 23, 261–284.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics 47, 13–37.

Lo, A., MacKinlay, A., 1990. Data-snooping biases in tests of financial asset pricing models. Review of Financial Studies 3, 431–467.

Nanda, D., 1950. Distribution of the sum of roots of a determinantal equation under a certain condition. Annals of Mathematical Statistics 21, 432–439.

Owen, J., Rabinovitch, R., 1983. On the class of elliptical distributions and their applications to the theory of portfolio choice. Journal of Finance 38, 745–752.

Peñaranda, F., Sentana, E., 2012. Spanning tests in return and stochastic discount factor mean-variance frontiers: a unifying approach. Journal of Econometrics 170, 303–324.

Pesaran, H., Yamagata, T., 2012. Testing CAPM with a large number of assets. SSRN Working Paper and IZA Discussion Paper No. 6469.

Pillai, K., 1955. Some new test criteria in multivariate analysis. Annals of Mathematical Statistics 26, 117–121.

Randles, R., Wolfe, D., 1979. Introduction to the Theory of Nonparametric Statistics. Wiley, New York.

Ray, S., Savin, N., Tiwari, A., 2009. Testing the CAPM revisited. Journal of Empirical Finance 16, 721–733.

Roll, R., 1977. A critique of the asset pricing theory's tests; Part I: On past and potential testability of the theory. Journal of Financial Economics 4, 129–176.

Ross, S., 1976. The arbitrage theory of capital asset pricing. Journal of Economic Theory 13, 341–360.

Roy, S., 1953. On a heuristic method of test construction and its use in multivariate analysis. Annals of Mathematical Statistics 24, 220–238.

Sentana, E., 2009. The econometrics of mean-variance efficiency: a survey. Econometrics Journal 12, 65–101.

Serfling, R., 2006. Multivariate symmetry and asymmetry. In: Kotz, S., Balakrishnan, N., Read, C., Vidakovic, B. (Eds.), Encyclopedia of Statistical Sciences, 2nd Edition. Wiley, pp. 5338–5345.

Sharpe, W., 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. Journal of Finance 19, 425–442.

Stewart, K., 1997. Exact testing in multivariate regression. Econometric Reviews 16, 321–352.

Wilks, S., 1932. Certain generalizations in the analysis of variance. Biometrika 24, 471–494.

Zhou, G., 1993. Asset-pricing tests under alternative distributions. Journal of Finance 48, 1927–1942.

Table 1. Comparison of empirical size and power of mean-variance efficiency tests: 1 benchmark portfolio

		$\phi=0,\ \varphi_{\max}=0,\ \lambda=0.8$				$\phi=0,\ \varphi_{\max}=1,\ \lambda=0.2$				$\phi=0.99,\ \varphi_{\max}=1,\ \lambda=0.2$			
T	$N=$	50	100	200	400	50	100	200	400	50	100	200	400
Panel A: Size													
60	$J_{E,1}$	5.2	-	-	-	5.4	-	-	-	5.2	-	-	-
	$J_{E,2}$	6.8	4.2	6.0	5.7	5.8	6.8	5.9	6.3	6.1	6.7	6.1	5.6
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	1.8	1.2	1.4	2.1	1.0	1.3	1.4	1.3
	$\mathbf{F}_2$	0.2	0.1	0.0	0.0	1.8	1.3	1.4	2.2	1.3	1.4	1.4	1.7
	$\mathbf{F}_3$	0.8	0.3	0.0	0.2	1.7	1.3	1.5	2.3	1.3	1.3	1.5	1.7
	$\mathbf{F}_4$	1.3	0.7	0.4	0.3	1.4	1.3	1.6	2.1	1.5	1.3	1.5	1.5
	$\mathbf{F}_{max}$	1.2	1.1	1.1	0.9	1.4	1.1	1.3	1.2	1.6	1.5	1.6	0.8
100	$J_{E,1}$	5.0	-	-	-	4.5	-	-	-	4.8	-	-	-
	$J_{E,2}$	6.0	5.3	5.0	5.8	6.5	6.2	5.6	6.5	6.2	6.5	5.9	6.0
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	1.2	0.8	0.9	1.2	1.8	1.2	2.0	1.5
	$\mathbf{F}_2$	0.1	0.0	0.0	0.0	1.4	0.9	0.9	1.3	1.7	1.2	2.1	1.6
	$\mathbf{F}_3$	0.6	0.1	0.0	0.1	1.3	0.9	0.9	1.3	1.7	1.1	2.2	1.6
	$\mathbf{F}_4$	1.0	0.4	0.1	0.4	1.3	1.2	0.8	1.3	1.7	1.2	2.2	1.6
	$\mathbf{F}_{max}$	1.1	1.3	0.5	1.2	1.3	1.3	0.5	1.3	2.1	1.0	1.3	1.8
Panel B: Power with $a_i \sim U[-0.1, 0.1]$													
60	$J_{E,1}$	10.4	-	-	-	**95.8**	-	-	-	**96.1**	-	-	-
	$J_{E,2}$	**40.6**	**63.1**	**85.4**	**98.4**	30.5	33.5	29.7	31.3	35.6	37.1	36.9	38.6
	$\mathbf{F}_1$	0.1	0.0	0.0	0.0	7.0	6.7	5.6	6.5	11.9	10.4	10.1	11.3
	$\mathbf{F}_2$	1.0	0.6	0.2	0.1	26.9	25.8	22.8	22.8	29.9	28.9	29.1	30.6
	$\mathbf{F}_3$	3.1	1.8	1.9	0.9	50.6	54.4	59.1	64.5	46.1	51.1	56.5	61.2
	$\mathbf{F}_4$	3.9	3.3	3.7	3.2	61.6	72.2	79.0	85.4	55.6	64.2	73.2	79.6
	$\mathbf{F}_{max}$	4.2	4.8	5.3	5.5	71.0	**82.0**	**91.3**	**96.6**	68.0	**78.4**	**89.0**	**94.7**
100	$J_{E,1}$	42.6	-	-	-	**100.0**	-	-	-	**100.0**	-	-	-
	$J_{E,2}$	**72.8**	**92.2**	**99.3**	**100.0**	69.5	71.0	73.4	75.3	53.3	54.2	54.9	58.0
	$\mathbf{F}_1$	0.4	0.0	0.0	0.0	18.4	14.3	12.9	14.5	23.5	23.7	24.5	25.4
	$\mathbf{F}_2$	4.5	3.5	1.5	0.8	61.9	64.8	70.4	75.2	52.0	54.1	56.7	58.1
	$\mathbf{F}_3$	9.1	8.4	7.0	9.5	82.5	89.9	94.2	96.5	70.5	76.8	82.9	85.7
	$\mathbf{F}_4$	10.9	10.0	9.3	12.3	89.1	95.1	98.4	99.2	79.2	86.9	93.4	95.7
	$\mathbf{F}_{max}$	11.1	9.8	9.7	11.4	93.7	**98.3**	**99.7**	**100.0**	88.6	**94.8**	**99.0**	**99.8**

Notes: This table reports the empirical size in Panel A and power in Panel B of the GRS $J_{E,1}$ test, the PY $J_{E,2}$ test, and the proposed MC bounds tests with $M=100$ based on the $\mathbf{F}_p$, $p=1,\ldots,4$, and $\mathbf{F}_{max}$ statistics. The MLR model disturbances are i.i.d. both over time and in the cross-section when $\phi=0$ and $\varphi_{\max}=0$; a higher value of $\varphi_{\max}$ implies stronger cross-sectional covariances; a non-zero value of ϕ makes the covariance structure time-dependent. Entries are percentage rates, the nominal level is 5%, and the results are based on 1000 replications. The symbol "-" is used whenever the GRS test is not computable and the entires set in bold indicate the most powerful test.

Table 2. Comparison of empirical size and power of mean-variance efficiency tests: 3 benchmark portfolio

T	$N=$	$\phi=0,\ \varphi_{\max}=0,\ \lambda=0.8$				$\phi=0,\ \varphi_{\max}=1,\ \lambda=0.2$				$\phi=0.99,\ \varphi_{\max}=1,\ \lambda=0.2$			
		50	100	200	400	50	100	200	400	50	100	200	400
Panel A: Size													
60	$J_{E,1}$	5.0	-	-	-	5.0	-	-	-	3.9	-	-	-
	$J_{E,2}$	4.8	6.9	6.2	4.7	6.7	6.0	6.0	6.1	7.5	4.8	5.8	6.3
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.1	0.2	0.2	0.1	0.0
	$\mathbf{F}_2$	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.1	0.2	0.3	0.2	0.0
	$\mathbf{F}_3$	0.0	0.1	0.0	0.0	0.0	0.4	0.3	0.1	0.2	0.3	0.2	0.1
	$\mathbf{F}_4$	0.0	0.2	0.0	0.0	0.1	0.3	0.3	0.1	0.1	0.3	0.2	0.1
	$\mathbf{F}_{max}$	0.1	0.5	0.1	0.2	0.2	0.3	0.3	0.1	0.1	0.1	0.5	0.0
100	$J_{E,1}$	4.7	-	-	-	5.0	-	-	-	5.4	-	-	-
	$J_{E,2}$	5.6	6.1	5.6	6.1	6.2	6.7	7.1	8.0	5.0	6.2	6.1	7.7
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.4	0.2	0.4
	$\mathbf{F}_2$	0.0	0.0	0.0	0.0	0.2	0.1	0.1	0.1	0.1	0.4	0.2	0.5
	$\mathbf{F}_3$	0.0	0.0	0.0	0.0	0.2	0.1	0.1	0.1	0.1	0.4	0.2	0.5
	$\mathbf{F}_4$	0.0	0.0	0.0	0.0	0.2	0.1	0.1	0.1	0.1	0.4	0.2	0.5
	$\mathbf{F}_{max}$	0.1	0.0	0.0	0.1	0.2	0.2	0.2	0.0	0.3	0.1	0.0	0.1
Panel B: Power with $a_i \sim U[-0.1, 0.1]$													
60	$J_{E,1}$	8.8	-	-	-	**89.0**	-	-	-	**88.9**	-	-	-
	$J_{E,2}$	**39.9**	**62.2**	**83.4**	**97.6**	33.8	29.3	31.2	29.1	37.3	34.2	36.7	35.0
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	0.4	0.6	0.8	0.6	1.3	0.9	0.7	0.7
	$\mathbf{F}_2$	0.0	0.0	0.0	0.0	2.7	1.6	2.0	1.8	5.1	3.9	3.1	4.3
	$\mathbf{F}_3$	0.0	0.1	0.0	0.0	10.4	7.7	9.7	8.0	13.6	13.2	10.9	14.8
	$\mathbf{F}_4$	0.3	0.3	0.0	0.0	20.6	21.3	24.9	27.0	22.2	22.8	26.4	32.0
	$\mathbf{F}_{max}$	0.9	0.6	0.7	0.6	35.1	**47.5**	**54.3**	**71.8**	35.5	**44.7**	**58.9**	**68.5**
100	$J_{E,1}$	42.7	-	-	-	**100.0**	-	-	-	**100.0**	-	-	-
	$J_{E,2}$	**67.8**	**90.3**	**99.7**	**100.0**	63.1	66.9	70.4	73.2	50.4	51.8	53.9	56.2
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	1.5	1.9	1.1	1.0	4.8	3.8	3.8	4.7
	$\mathbf{F}_2$	0.0	0.0	0.0	0.0	12.7	11.6	8.8	7.9	17.4	17.5	15.3	18.9
	$\mathbf{F}_3$	0.2	0.0	0.0	0.0	37.0	44.2	44.6	46.4	35.0	41.3	40.7	43.7
	$\mathbf{F}_4$	0.5	0.4	0.1	0.2	55.2	67.8	73.2	79.3	48.1	59.1	64.3	68.4
	$\mathbf{F}_{max}$	0.9	1.6	1.3	1.8	75.0	**88.0**	**95.5**	**98.6**	70.6	**83.1**	**92.5**	**95.7**

Notes: This table mimics Table 1, except that here the returns are generated according to the MLR model with $K = 3$.

Table 3. Comparison of empirical size and power of mean-variance spanning tests: 3 benchmark portfolios

		$\phi=0,\ \varphi_{\max}=0,\ \lambda=0.8$				$\phi=0,\ \varphi_{\max}=1,\ \lambda=0.2$				$\phi=0.99,\ \varphi_{\max}=1,\ \lambda=0.2$			
T	$N=$	50	100	200	400	50	100	200	400	50	100	200	400
Panel A: Size													
60	J_S	4.9	-	-	-	5.7	-	-	-	4.2	-	-	-
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	0.6	0.4	0.9	1.0	0.5	1.1	0.7	0.5
	$\mathbf{F}_2$	0.0	0.0	0.0	0.0	0.7	0.4	0.9	1.1	0.5	1.2	0.7	0.6
	$\mathbf{F}_3$	0.1	0.0	0.0	0.0	0.7	0.4	1.0	1.1	0.5	1.2	0.7	0.7
	$\mathbf{F}_4$	0.3	0.2	0.0	0.0	0.7	0.4	0.9	1.2	0.4	1.0	0.6	0.6
	$\mathbf{F}_{max}$	0.4	0.4	0.4	0.2	1.0	0.7	0.8	0.8	1.1	0.6	0.8	0.6
100	J_S	6.3	-	-	-	5.4	-	-	-	5.2	-	-	-
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	0.6	0.8	0.4	0.5	0.7	0.6	0.9	0.6
	$\mathbf{F}_2$	0.0	0.0	0.0	0.0	0.5	1.1	0.5	0.5	0.9	0.7	0.9	0.6
	$\mathbf{F}_3$	0.0	0.0	0.0	0.0	0.6	1.2	0.7	0.5	0.9	0.7	1.1	0.6
	$\mathbf{F}_4$	0.4	0.0	0.2	0.0	0.8	1.0	0.7	0.5	0.9	0.8	1.1	0.5
	$\mathbf{F}_{max}$	0.9	0.7	0.8	0.9	0.8	0.6	0.6	0.7	0.8	1.0	1.0	0.3
Panel B: Power with $a_i \sim U[-0.1, 0.1]$, $\delta_i = 0$													
60	J_S	**7.7**	-	-	-	**64.5**	-	-	-	**65.9**	-	-	-
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	2.5	1.5	2.6	3.0	3.0	3.7	3.1	2.3
	$\mathbf{F}_2$	0.1	0.0	0.0	0.0	6.1	4.0	4.8	5.7	9.4	9.3	7.6	7.4
	$\mathbf{F}_3$	0.4	0.3	0.0	0.0	16.9	14.2	15.1	15.3	17.4	19.8	20.5	18.8
	$\mathbf{F}_4$	1.8	0.6	0.2	0.1	27.4	30.7	35.4	36.7	25.3	32.6	36.7	38.0
	$\mathbf{F}_{max}$	3.2	1.4	1.5	2.0	45.1	**54.9**	**68.7**	**78.2**	39.5	**56.4**	**66.3**	**74.2**
100	J_S	**26.2**	-	-	-	**100.0**	-	-	-	**100.0**	-	-	-
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	5.6	4.7	5.0	3.7	10.9	11.2	11.1	10.6
	$\mathbf{F}_2$	0.1	0.0	0.0	0.0	22.5	19.8	20.0	17.6	24.4	25.6	26.5	24.4
	$\mathbf{F}_3$	1.3	0.4	0.2	0.0	49.0	55.6	60.9	60.2	42.4	48.9	52.0	52.3
	$\mathbf{F}_4$	1.8	1.5	1.9	0.5	64.2	74.4	82.4	87.2	54.6	65.2	71.2	76.8
	$\mathbf{F}_{max}$	2.5	4.0	4.0	3.5	81.9	**92.6**	**97.5**	**99.2**	75.7	**86.9**	**93.4**	**98.1**
Panel C: Power with $a_i = 0$, $\delta_i \sim U[-0.2, 0.2]$													
60	J_S	**10.2**	-	-	-	**73.7**	-	-	-	**73.5**	-	-	-
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	4.0	2.4	3.7	4.5	5.8	6.7	5.7	5.7
	$\mathbf{F}_2$	0.0	0.0	0.0	0.0	13.0	10.0	10.9	12.5	16.7	16.9	16.6	13.9
	$\mathbf{F}_3$	0.5	0.1	0.0	0.0	30.0	31.7	35.2	38.8	30.1	34.9	33.5	37.1
	$\mathbf{F}_4$	1.3	1.0	0.6	0.1	43.1	49.7	60.3	65.4	42.5	51.2	53.2	57.7
	$\mathbf{F}_{max}$	2.3	2.3	2.9	1.6	58.0	**73.0**	**82.7**	**90.7**	58.2	**70.1**	**78.1**	**89.9**
100	J_S	**38.8**	-	-	-	**100.0**	-	-	-	**100.0**	-	-	-
	$\mathbf{F}_1$	0.0	0.0	0.0	0.0	11.7	9.1	9.1	7.7	17.2	17.9	18.2	17.6
	$\mathbf{F}_2$	0.1	0.2	0.0	0.0	45.7	41.6	46.7	46.5	39.3	40.7	39.9	40.0
	$\mathbf{F}_3$	1.6	0.8	0.5	0.2	73.0	78.0	83.4	85.7	60.0	64.4	68.4	71.8
	$\mathbf{F}_4$	2.5	2.3	2.4	1.5	83.5	90.5	93.3	96.6	71.5	79.1	84.3	89.2
	$\mathbf{F}_{max}$	4.3	5.8	5.8	6.4	91.2	**96.8**	**99.9**	**99.9**	85.2	**93.7**	**97.2**	**99.3**

Notes: This table reports the empirical size in Panel A and power in Panels B and C of the HK J_S test and the proposed MC bounds tests with $M = 100$ based on the $\mathbf{F}_p$, $p = 1, \ldots, 4$, and $\mathbf{F}_{max}$ statistics. The MLR disturbances are i.i.d. both over time and in the cross-section when $\phi = 0$ and $\varphi_{\max} = 0$; a higher value of $\varphi_{\max}$ implies stronger cross-sectional covariances; a non-zero value of ϕ makes the covariance structure time-dependent. Entries are percentage rates, the nominal level is 5%, and the results are based on 1000 replications. The symbol "-" is used whenever the HK test is not computable and the entires set in bold indicate the most powerful test.

Table 4. Mean-variance efficiency tests: CAPM and Fama-French model

	CAPM			Fama-French Model		
Time period	$J_{E,1}$	$J_{E,2}$	$\mathbf{F}_{max}$	$J_{E,1}$	$J_{E,2}$	$\mathbf{F}_{max}$
39-year period						
1/73–12/11	0.999	**0.000**	0.302	0.956	0.055	0.260
5-year subperiods and a 4-year subperiod						
1/73–12/77	-	0.487	0.544	-	0.638	0.312
1/78–12/82	-	0.178	0.158	-	0.989	0.768
1/83–12/87	-	0.258	0.910	-	0.482	0.916
1/88–12/92	-	0.419	0.674	-	**0.038**	0.154
1/93–12/97	-	0.972	0.986	-	0.929	0.640
1/98–12/02	-	0.999	0.852	-	0.999	0.994
1/03–12/07	-	**0.000**	0.274	-	**0.000**	0.360
1/08–12/11	-	0.973	0.918	-	0.986	0.778
10-year subperiods and a 9-year subperiod						
1/73–12/82	-	**0.003**	0.196	-	0.637	0.770
1/83–12/92	-	**0.000**	0.662	-	**0.000**	0.028, 0.900
1/93–12/02	-	0.802	0.948	-	0.988	0.756
1/03–12/11	-	0.076	0.196	-	**0.045**	0.112

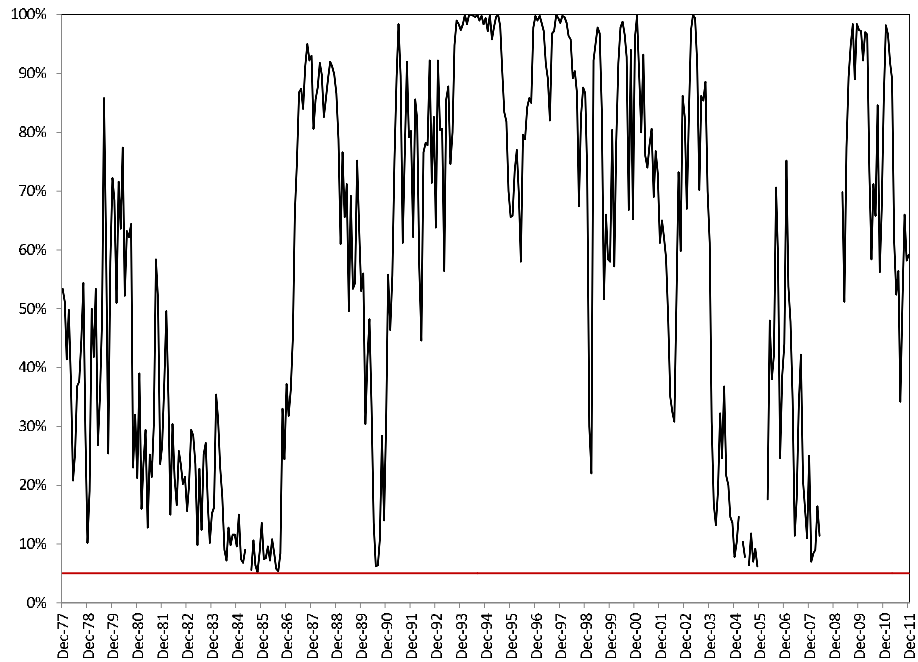
Notes: The results are based on 452 individual stock returns. The entries are p-values and those set in bold represent significant cases at the 0.05 level. The conservative MC p-value is reported when $\tilde{p}_M^C\big(\mathbf{F}_{max}(\mathbf{Y})\big) \leq 0.05$, whereas the liberal MC p-value is reported when $\tilde{p}_M^L\big(\mathbf{F}_{max}(\mathbf{Y})\big) > 0.05$, and both are reported when the outcome is inconclusive. The symbol "-" is used whenever the GRS test is not computable.

Table 5. Mean-variance spanning tests: Fama-French model

Time period	J_S	$\mathbf{F}_{max}$
39-year period		
1/73–12/11	**0.029**	**0.002**
5-year subperiods and a 4-year subperiod		
1/73–12/77	-	**0.014**
1/78–12/82	-	0.066
1/83–12/87	-	0.170
1/88–12/92	-	**0.022**
1/93–12/97	-	**0.010**
1/98–12/02	-	0.012, 0.070
1/03–12/07	-	**0.038**
1/08–12/11	-	**0.016**
10-year subperiods and a 9-year subperiod		
1/73–12/82	-	**0.004**
1/83–12/92	-	**0.004**
1/93–12/02	-	0.020, 0.152
1/03–12/11	-	**0.010**

Notes: The results are based on 452 individual stock returns. The entries are p-values and those set in bold represent significant cases at the 0.05 level. The conservative MC p-value is reported when $\tilde{p}_M^C\big(\mathbf{F}_{max}(\mathbf{Y})\big) \leq 0.05$, whereas the liberal MC p-value is reported when $\tilde{p}_M^L\big(\mathbf{F}_{max}(\mathbf{Y})\big) > 0.05$, and both are reported when the outcome is inconclusive. The symbol "-" is used whenever the HK test is not computable.

(a) Mean-variance efficiency tests: CAPM



(b) Mean-variance efficiency tests: Fama-French model

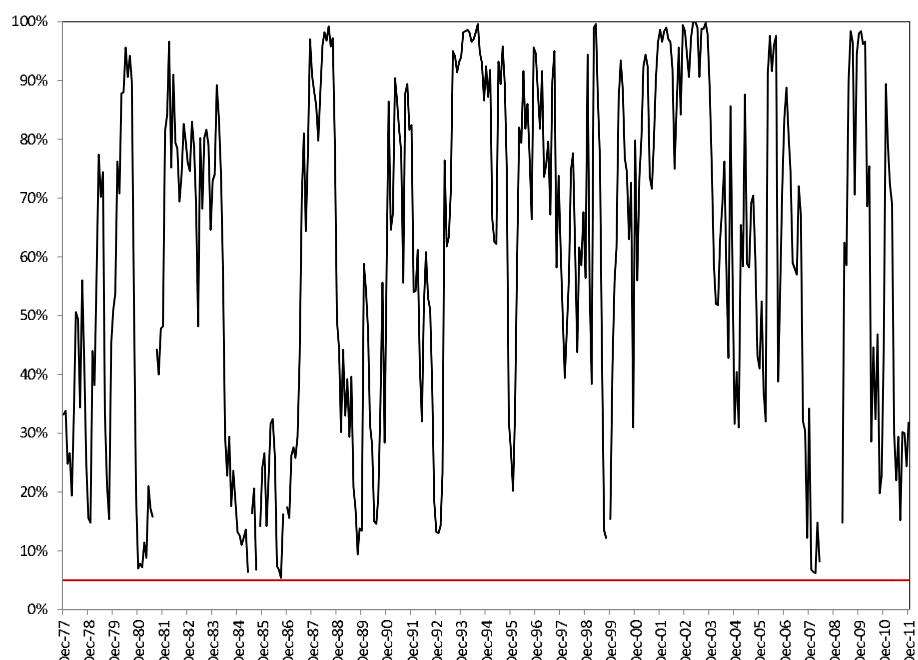


Figure 1. Time variation in p-values (as percentage rates) of the $\mathbf{F}_{max}$ test of mean-variance efficiency based on the CAPM (panel a) and the 3-factor Fama-French model (panel b) using a 60-month rolling window. The conservative MC p-value is plotted when $\tilde{p}_M^C(\mathbf{F}_{max}(\mathbf{Y})) \leq 5\%$, whereas the liberal MC p-value is plotted when $\tilde{p}_M^L(\mathbf{F}_{max}(\mathbf{Y})) > 5\%$. The discontinuities in the series indicate periods of inconclusive test outcomes.

Mean-variance spanning tests: Fama-French model

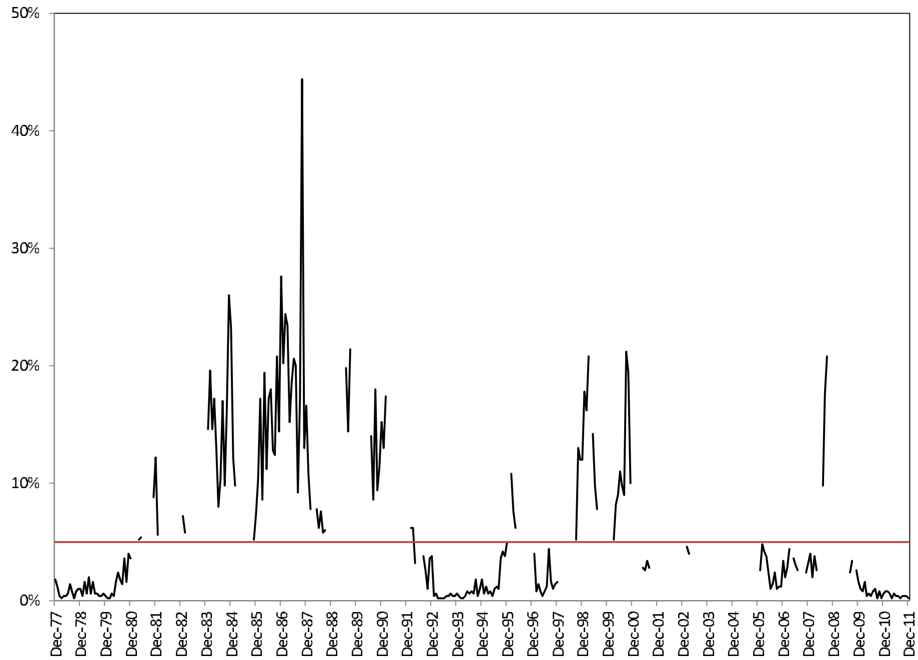


Figure 2. Time variation in p-values (as percentage rates) of the $\mathbf{F}_{max}$ test of mean-variance spanning based on the 3-factor Fama-French model using a 60-month rolling window. The conservative MC p-value is plotted when $\tilde{p}_M^C(\mathbf{F}_{max}(\mathbf{Y})) \leq 5\%$, whereas the liberal MC p-value is plotted when $\tilde{p}_M^L(\mathbf{F}_{max}(\mathbf{Y})) > 5\%$. The discontinuities in the series indicate periods of inconclusive test outcomes.