Testing Linear Factor Pricing Models with Large Cross-Sections: A Distribution-Free Approach

Sermin Gungor

Financial Markets Department, Bank of Canada, Ottawa, ON K1A 0G9, Canada

Richard Luger

Department of Risk Management and Insurance, Georgia State University, Atlanta, GA 30303 (rluger@gsu.edu)

Abstract

We develop a finite-sample distribution-free procedure to test the beta-pricing representation of linear factor pricing models. In sharp contrast to extant finite-sample tests, our framework allows for unknown forms of non-normalities, heteroskedasticity, and time-varying covariances. The power of the proposed test procedure increases as the time series lengthens and/or the cross-section becomes larger. So the criticism sometimes heard that non-parametric tests lack power does not apply here, since the number of test assets is chosen by the user. This also stands in contrast to the usual tests that lose power or may not even be computable if the number of test assets is too large.

KEY WORDS: Factor model; Beta pricing; CAPM; Mean-variance efficiency; Robust inference.

1 INTRODUCTION

Many asset pricing models predict that expected returns depend linearly on "beta" coefficients relative to one or more portfolios or factors. The beta is the regression coefficient of the asset return on the factor. In the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965), the single beta measures the systematic risk or co-movement with the returns on the market portfolio. Accordingly, assets with higher betas should offer in equilibrium higher expected returns. The Arbitrage Pricing Theory (APT) of Ross (1976), developed on the basis of arbitrage arguments, can be more general than the CAPM in that it relates expected returns with multiple beta coefficients. The intertemporal CAPM of Merton (1973), based on investor optimization and equilibrium arguments, also leads to multi-beta pricing.

Empirical tests of the validity of beta pricing relationships are often conducted within the context of multivariate linear factor models. When the factors are traded portfolios and a risk-free asset is available, exact factor pricing implies that the vector of asset return intercepts will be zero. These tests are interpreted as tests of the mean-variance efficiency of a benchmark portfolio in the single-beta model, or that some combination of the factor portfolios is mean-variance efficient in multi-beta models. (A portfolio is mean-variance efficient if it maximizes the expected return for a given level of variance.) In this context, standard asymptotic theory provides a poor approximation to the finite-sample distribution of the usual Wald and likelihood ratio (LR) test statistics, even with fairly large samples. Shanken (1996), Campbell, Lo, and MacKinlay (1997), and Dufour and Khalaf (2002) document severe size distortions for those tests, with overrejections growing quickly as the number of equations in the multivariate model increases. The simulation evidence in Ferson and Foerster (1994) and Gungor and Luger (2009) shows that tests based on the Generalized Method of Moments (GMM) à la MacKinlay and Richardson (1991) suffer from the same problem. As a result, commonly used empirical tests of beta-pricing representations can be severely affected and lead to erroneous rejections of their validity.

The assumptions underlying standard asymptotic arguments can be questionable when dealing with financial asset returns data. In the context of the consumption CAPM for example, Kocherlakota (1997) shows that the model disturbances are so heavy-tailed that they do not satisfy the Central Limit Theorem. In such an environment, standard methods of inference can lead to spurious rejections even asymptotically and Kocherlakota instead relies on jackknifing to devise a method of testing the consumption CAPM. Similarly, Affleck-Graves and McDonald (1989) and Chou and Zhou (2006) suggest the use of bootstrap techniques to provide more robust and reliable asset pricing tests.

There are very few methods that yield truly exact finite-sample tests. The most prominent one is probably the F test of Gibbons, Ross, and Shanken (1989) (GRS). The exact distribution theory for that test rests on the assumption that the vectors of model disturbances are independent and identically distributed (i.i.d.) each period according to a multivariate normal distribution. Yet there has long been ample evidence that financial returns exhibit non-normalities; see Fama (1965), Blattberg and Gonedes (1974), Hsu (1982), Affleck-Graves and McDonald (1989), and Zhou (1993). Beaulieu, Dufour, and Khalaf (2007) generalize the GRS approach for testing mean-variance efficiency. Their simulation-based approach 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. Gungor and Luger (2009) propose exact tests of the mean-variance efficiency of a single reference portfolio whose exactness does not depend on any parametric assumptions.

In this paper we extend the idea of Gungor and Luger (2009) to obtain tests of multi-beta pricing representations that relax three assumptions of the GRS test: (i) the assumption of identically distributed disturbances, (ii) the assumption of normally distributed disturbances, and (iii) the restriction on the number of test assets. The proposed test procedure is based on finite-sample pivots that are valid without any parametric assumptions about the specific distribution of the disturbances in the multi-factor model. We propose an adaptive approach based on a split-sample technique to obtain a single portfolio representation judiciously formed to avoid power losses that can occur in naive portfolio groupings. For other examples of split-sample techniques, see Jouneau-Sion and Torrès (2006), and Dufour and Taamouti (2010).

A very attractive feature of our approach is that it is applicable even if the number of test assets is greater than the length of the time series. This stands in sharp contrast to the GRS test or any other approach based on usual estimates of the disturbance covariance matrix. In order to avoid singularities and be computable, those approaches require the size of the cross-section to be less than that of the time series. In fact, great care must be taken when applying the GRS test since its power does not increase monotonically with the number of test assets and all the power may be lost if too many are included. This problem is related to the fact that the number of covariances that need to be estimated grows rapidly with the number of included test assets. As a result, the precision with which this increasing number of parameters can be estimated deteriorates given a fixed time-series length.

Our proposed test procedure exploits results from Coudin and Dufour (2009) on median regressions to construct sign-based statistics, one of which is a sign-based GMM statistic and the other is the sign analogue of the usual F test. The motivation for using signs comes from an impossibility result due to Lehmann and Stein (1949) that shows that the only tests which yield reliable inference under sufficiently general distributional assumptions, allowing non-normal, possibly heteroskedastic, independent observations are based on sign statistics. This means that all other methods, including the standard heteroskedasticity and autocorrelation-corrected (HAC) methods developed by White (1980) and Newey and West (1987) among others, which are not based on signs, cannot be proved to be valid and reliable for any sample size.

The paper is organized as follows. Section 2 presents the linear factor model used to describe the asset returns, the exact pricing implication, and the benchmark GRS test. We provide an illustration of the effects of increasing the number of test assets on the power of the GRS test. In Section 3 we develop the new test procedure. We begin that section by presenting the statistical framework and then proceed to describe each step of the procedure. Section 4 contains the results of simulation experiments designed to compare the performance of the proposed test procedure with several of the standard tests. In Section 5 we apply the procedure to test the Sharpe-Lintner version of the CAPM and the well-known Fama-French three-factor model. Section 6 concludes.

2 FACTOR MODEL

Suppose there exists a riskless asset for each period of time and define $\mathbf{r}_t$ as an $N \times 1$ vector of time-t returns on N assets in excess of the riskless rate of return. Suppose further that those excess returns are described by the linear K-factor model

$$\mathbf{r}_t = \mathbf{a} + \mathbf{B}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\mathbf{f}_t$ is a $K \times 1$ vector of common factor portfolio excess returns, $\mathbf{B}$ is the $N \times K$ matrix of betas (or factor loadings), and $\mathbf{a}$ and $\boldsymbol{\varepsilon}_t$ are $N \times 1$ vectors of factor model intercepts and disturbances, respectively. As usual, the disturbance vector $\boldsymbol{\varepsilon}_t$ is assumed to have well-defined first and second moments satisfying $E[\boldsymbol{\varepsilon}_t|\mathbf{f}_t] = \mathbf{0}$ and $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'|\mathbf{f}_t] = \boldsymbol{\Sigma}_t$, a finite $N \times N$ matrix.

Exact factor pricing implies that expected returns depend linearly on the betas associated to the factor portfolio returns via the condition

$$E[\mathbf{r}_t] = \mathbf{B}E[\mathbf{f}_t], \tag{2}$$

where the vector of expected excess returns on $\mathbf{f}_t$ represents market-wide risk premiums since they are common across all traded securities. The beta-pricing representation in (2) is a generalization of the CAPM of Sharpe (1964) and Lintner (1965), which asserts that the expected excess return on an asset is linearly related to its single beta. This beta measures the asset's systematic risk or co-movement with the excess return on the market portfolio—the portfolio of all invested wealth. Equivalently, the CAPM says that the market portfolio is mean-variance efficient in the investment universe comprising all possible assets. The pricing relationship in (2) is more general since it says that a combination (portfolio) of the factor portfolios is mean-variance efficient; see Jobson (1982), Jobson and Korkie (1982, 1985), Grinblatt and Titman (1987), Shanken (1987), and Huberman, Kandel, and Stambaugh (1987) for more on the relation between factor models and mean-variance efficiency.

The beta-pricing representation in (2) is a restriction on expected returns which can be assessed by testing the hypothesis

$$H_0 : \mathbf{a} = \mathbf{0}, \tag{3}$$

under the maintained factor structure specification in (1). If the pricing errors in $\mathbf{a}$ are in fact different from zero, then (2) does not hold, meaning that there is no way to combine the factor portfolios to obtain one that is mean-variance efficient.

GRS propose a multivariate F test of H_0 in (3) that all the pricing errors 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 $\mathbf{F} = [\boldsymbol{f}_1, ..., \boldsymbol{f}_T]'$; i.e., $\boldsymbol{\varepsilon}_t \,|\, \mathbf{F} \sim$ 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 $\mathbf{B}$:

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

where $\bar{\mathbf{r}} = T^{-1}\sum_{t=1}^{T}\mathbf{r}_t$ and $\bar{\mathbf{f}} = T^{-1}\sum_{t=1}^{T}\mathbf{f}_t$, and the estimate of the disturbance covariance matrix is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_t - \hat{\mathbf{a}} - \hat{\mathbf{B}}\mathbf{f}_t)(\mathbf{r}_t - \hat{\mathbf{a}} - \hat{\mathbf{B}}\mathbf{f}_t)'. \tag{4}$$

The GRS test statistic is

$$J_1 = \frac{T - N - K}{N}\left[1 + \bar{\mathbf{f}}'\hat{\boldsymbol{\Omega}}^{-1}\bar{\mathbf{f}}\right]^{-1}\hat{\mathbf{a}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{a}}, \tag{5}$$

where $\hat{\boldsymbol{\Omega}}$ is given by

$$\hat{\boldsymbol{\Omega}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{f}_t - \bar{\mathbf{f}})(\mathbf{f}_t - \bar{\mathbf{f}})'.$$

Under the null hypothesis H_0, the statistic J_1 follows a central F distribution with N degrees of freedom in the numerator and $(T - N - K)$ degrees of freedom in the denominator.

In practical applications of the GRS test, one needs to decide the appropriate number N of test assets to include. It might seem natural to try to use as many as possible in order to increase the probability of rejecting H_0 when it is false. As the test asset universe expands it becomes more likely that non-zero pricing errors will be detected, if indeed there are any. However, the choice of

N is restricted by T in order to keep the estimate of the disturbance covariance matrix in (4) from becoming singular, and the choice of T itself is often restricted owing to concerns about parameter stability. For instance, it is quite common to see studies where $T = 60$ monthly returns and N is between 10 and 30. The effects of increasing the number of test assets on test power is discussed in GRS, Campbell, Lo, and MacKinlay (1997, p. 206) and Sentana (2009). When N increases, three effects come into play: (i) the increase in the value of J_1's non-centrality parameter, which increases power, (ii) the increase in the number of degrees of freedom of the numerator, which decreases power, and (iii) the decrease in the number of degrees of freedom of the denominator due to the additional parameters that need to be estimated, which also decreases power.

To illustrate the net effect of increasing N on the power of the GRS test, we simulated model (1) with $K = 1$, where the returns on the single factor are random draws from the standard normal distribution. The elements of the independent disturbance vector were also drawn from the standard normal distribution thereby ensuring the exactness of the GRS test. We set $T = 60$ and considered $a_i = 0.05$, 0.10, and 0.15, for $i = 1, ..., N$, and we let the number of test assets N range from 1 to 58. The chosen values for a_i are well within the range of what we find with monthly stock returns. Figure 1 shows the power of the GRS test as a function of N, where for any given N the higher power is associated with greater pricing errors. In line with the discussion in GRS, this figure clearly shows the power of the test given this specification rising as N increases up to about one half of T, and then decreasing beyond that. The results in Table 5.2 of Campbell, Lo, and MacKinlay (1997) show several other alternatives against which the power of the GRS test declines as N increases. Furthermore, there are no general results about how to devise an optimal multivariate test. So great care must somehow be taken when choosing the number of test assets since power does not increase monotonically with N, and if the cross-section is too large, then the GRS test may lose all its power or may not even be computable. In fact, any procedure that relies on standard unrestricted estimates of the covariance matrix of regression disturbances will have this singularity problem when N exceeds T.

3 TEST PROCEDURE

In this section we develop a procedure to test exact factor pricing in the context of (1) that relaxes three assumptions of the GRS test: (i) the assumption of identically distributed disturbances, (ii) the assumption of normally distributed disturbances, and (iii) the restriction that $N \leq T - K - 1$.

Our approach is motivated by a classical theorem in non-parametric statistics due to Lehmann and Stein (1949) which states that the *only* tests which yield valid inference under sufficiently general distributional assumptions, allowing non-normal, possibly heteroskedastic, independent observations are ones that are conditional on the absolute values of the observations; i.e., they must be based on sign statistics. Conversely, if a test procedure does not satisfy this condition for all levels $0 < \alpha < 1$, then its true size is 1 irrespective of its nominal size (Dufour, 2003).

3.1 Statistical Framework

As in the GRS framework, we assume that the disturbance vectors $\boldsymbol{\varepsilon}_t$ in (1) are independently distributed over time, conditional on $\mathbf{F}$. We do not require the disturbance vectors to be identically distributed, but we do assume that they satisfy a multivariate symmetry condition each period. In what follows, the symbol $\overset{d}{=}$ stands for the equality in distribution.

Assumption. *The cross-sectional disturbance vectors $\boldsymbol{\varepsilon}_t$, for $t = 1, ..., T$, are mutually independent, continuous, and reflectively symmetric so that $\boldsymbol{\varepsilon}_t \overset{d}{=} -\boldsymbol{\varepsilon}_t$, conditional on $\mathbf{F}$.* (6)

The distributions encompassed by this assumption include elliptically symmetric ones, such as the well-known multivariate normal (assumed by GRS) and Student-t distributions. The reflective symmetry condition in (6) is less stringent than elliptical symmetry. For instance, a mixture (finite or not) of distributions each one elliptically symmetric around the origin is not necessarily elliptically symmetric but it is reflectively symmetric.

Assumption (6) does not require the vectors $\boldsymbol{\varepsilon}_t$ to be identically distributed nor does it restrict their degree of heterogeneity. This is a very attractive feature since it is well known that financial returns often depart quite dramatically from Gaussian conditions. In particular, the distribution of asset returns appears to have much heavier tails and is more peaked than a normal distribution. The present framework leaves open not only the possibility of unknown forms of non-normality, but also heteroskedasticity and time-varying covariances among the $\boldsymbol{\varepsilon}_t$s. For example, when $(\mathbf{r}_t, \boldsymbol{f}_t)$ are elliptically distributed but non-normal, the conditional covariance matrix of $\boldsymbol{\varepsilon}_t$ depends on the contemporaneous $\boldsymbol{f}_t$; see MacKinlay and Richardson (1991) and Zhou (1993). Here the disturbance covariance structure could be any function of the common factors (contemporaneous or not). The simulation study in Section 4 examines a contemporaneous heteroskedasticity specification.

3.2 Portfolio Formation

A usual practice in the application of the GRS test is to base it on portfolio groupings in order to have N much less than T. As Shanken (1996) notes, this has the potential effect of reducing the residual variances and increasing the precision with which $\mathbf{a} = (a_1, ..., a_N)'$ is estimated. On the other hand, as Roll (1979) emphasizes, individual stock expected return deviations under the alternative can cancel out in portfolios, which would reduce the power of the GRS test unless the portfolios are combined in proportion to their weighting in the tangency portfolio. So ideally, all the pricing errors that make up the vector $\mathbf{a}$ in (1) would be of the same sign to avoid power losses when forming an equally-weighted portfolio of the test assets.

Our approach here is an adaptive one based on a split-sample technique where the first subsample is used to obtain an estimate of $\mathbf{a}$. That estimate is then used to form a single portfolio that judiciously avoids power losses. Finally, a conditional test of exact factor pricing is performed using only the returns on that portfolio observed over the second subsample. This approach formalizes the usual practice of forming portfolios to deal with large N. It is important to note that our procedure does not introduce 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), since the estimation results are conditionally (on the factors) independent of the second subsample test outcomes.

Let $T = T_1 + T_2$. In matrix form, the first T_1 returns on asset i can be represented by

$$\mathbf{r}_i^{(1)} = a_i \boldsymbol{\iota} + \mathbf{F}^{(1)} \mathbf{b}_i + \boldsymbol{\varepsilon}_i^{(1)},$$

where $\mathbf{r}_i^{(1)} = [r_{i1}, ..., r_{iT_1}]'$ and $\mathbf{F}^{(1)} = [\boldsymbol{f}_1, ..., \boldsymbol{f}_{T_1}]'$ collect the time series of T_1 returns on asset i and the factors, respectively, $\boldsymbol{\iota}$ is a conformable vector of ones, $\mathbf{b}_i'$ is the i^{th} row of $\mathbf{B}$ in (1), and $\boldsymbol{\varepsilon}_i^{(1)} = [\varepsilon_{i1}, ..., \varepsilon_{iT_1}]'$.

Restriction. *Only the first T_1 observations on $\mathbf{r}_t$ and $\mathbf{f}_t$ are used to compute $\hat{a}_1, ..., \hat{a}_N$.* (7)

This restriction does not limit the choice of estimation method, so the estimates $\hat{a}_1, ..., \hat{a}_N$ could be obtained by OLS or any other method. Of course, T_1 must at least be enough to obtain the subsample estimates $\hat{a}_1, ..., \hat{a}_N$. A well-known problem with OLS though is that it is very sensitive to the presence of large disturbances and outliers; see Section 5 for evidence. An alternative estimation method is to minimize the sum of the absolute deviations in computing the regression lines (Bassett and Koenker, 1978). The resulting least absolute deviations (LAD) estimator may be more efficient than OLS in heavy-tailed samples where extreme observations are more likely to occur. The results reported below in the simulation study and the empirical application are based on LAD.

With the estimates $\hat{\boldsymbol{a}} = (\hat{a}_1, ..., \hat{a}_N)'$ in hand, a vector of statistically motivated "portfolio" weights $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \hat{\omega}_2, ..., \hat{\omega}_N)'$ is computed according to

$$\hat{\omega}_i = \text{sign}(\hat{a}_i) \frac{1}{N}, \tag{8}$$

for $i = 1, ..., N$, and these weights are then used to find the T_2 returns on a portfolio computed as $y_t = \sum_{i=1}^N \hat{\omega}_i r_{it}$, $t = T_1 + 1, ..., T$. We shall first provide a distributional result for y_t that holds under H_0 and then explain in what sense the weights in (8) will maximize the power of the proposed distribution-free tests. In the following, $\delta = \sum_{i=1}^N \hat{\omega}_i a_i$ is the sum of the weighted a_is over the second subsample.

Proposition 1. *Under H_0 in (3) and when (6) and (7) hold, y_t is represented by the single equation*

$$y_t = \delta + \mathbf{f}_t' \boldsymbol{\beta} + u_t, \ \textit{for } t = T_1 + 1, ..., T, \tag{9}$$

where $\delta = 0$ *and* $(u_{T_1+1}, ..., u_T) \overset{d}{=} (\pm|u_{T_1+1}|, ..., \pm|u_T|)$*, conditional on* $\mathbf{F}$.

Proof. See online Appendix.

The construction of a test based on a single portfolio grouping is reminiscent of a mean-variance efficiency test proposed in Bossaerts and Hillion (1995) based on $\boldsymbol{\iota}'\hat{\boldsymbol{a}}$ and another one proposed in Gungor and Luger (2009) that implicitly exploits $\boldsymbol{\iota}'\boldsymbol{a}$. Those approaches can suffer power losses depending on whether the a_is tend to cancel out when summed. Splitting the sample and applying the weights in (8) when forming the portfolio offsets that problem. Of course if one believes *a priori* that the a_is don't tend to cancel out, then there is no need to split the sample and the test can proceed simply with $\omega_i = 1/N$.

The portfolio weights in (8) are in fact optimal in a certain sense. To see how, recall that $\boldsymbol{\varepsilon}_t$ is assumed to have well-defined first and second moments satisfying $E[\boldsymbol{\varepsilon}_t|\mathbf{f}_t] = \mathbf{0}$ and $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'|\mathbf{f}_t] = \boldsymbol{\Sigma}_t$. Assumption (6) then implies that the mean and median (point of symmetry) coincide at zero for any component of $\boldsymbol{\varepsilon}_t$. The power of our test procedure depends on $E[\boldsymbol{\omega}'\mathbf{r}_t - \boldsymbol{\omega}'\mathbf{B}\mathbf{f}_t|\mathbf{f}_t] = \boldsymbol{\omega}'\mathbf{a}$. (The next section shows how we deal with the presence of the nuisance parameters comprising $\boldsymbol{\omega}'\mathbf{B}$.) As in the usual mean-variance portfolio selection problem, choosing $\boldsymbol{\omega}$ to increase $\boldsymbol{\omega}'\mathbf{a}$ also entails an increase in the portfolio's variance $\boldsymbol{\omega}'\boldsymbol{\Sigma}_t\boldsymbol{\omega}$, which decreases test power. So the problem we face is to maximize $\boldsymbol{\omega}'\mathbf{a}$ subject to a target value for the variance. Here we set the target to $\boldsymbol{\iota}'\boldsymbol{\Sigma}_t\boldsymbol{\iota}/N^2$, the variance of the naive, equally-weighted portfolio which simply allocates equally across the N assets. It is easy to see that $\boldsymbol{\omega} = \text{sign}(\boldsymbol{a})/N$ will maximize power while keeping the resulting portfolio variance as close as possible to the target. Of course $\text{sign}(\boldsymbol{a})$ is unknown, so (8) uses $\text{sign}(\hat{\boldsymbol{a}})$. The possible discrepancy between the achieved and target variance values is given by $\frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\big(\text{sign}(\hat{a}_i)\text{sign}(\hat{a}_j) - 1\big)\sigma_{ij,t}$, which depends on the off-diagonal (covariance) terms of $\boldsymbol{\Sigma}_t$ but not on any of its diagonal (variance) terms. Note that in an approximate APT factor model, those off-diagonal terms tend to zero as $N \to \infty$.

The weights in (8) are quite intuitive and represent the optimal choice in our distribution-free context where possible forms of distribution heterogeneity (e.g. time-varying variances and covariances) are left completely unspecified. Note that optimal weights in a strict mean-variance sense cannot be used here since finding those requires an estimate of the (possibly time-varying) covariance structure and that is precisely what we are trying to avoid.

3.3 Test Statistics

The model in (9) can be represented in matrix form as $\mathbf{y} = \delta\boldsymbol{\iota} + \mathbf{F}^{(2)}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{F}^{(2)} = [\boldsymbol{f}_{T_1+1}, ..., \boldsymbol{f}_T]'$ and the elements of $\mathbf{u}$ follow what Coudin and Dufour (2009) call a "mediangale" which is similar to the usual martingale difference concept except that the median takes the place of the expectation. The following result is an immediate consequence of the strict conditional mediangale property in Proposition 2.1 of Coudin and Dufour. Here we define $s[x] = 1$, if $x \geq 0$, and $s[x] = -1$, if $x < 0$.

Proposition 2. *When (6) and (7) hold, the T_2 disturbance sign vector*

$$s(\mathbf{y} - \delta\boldsymbol{\iota} - \mathbf{F}^{(2)}\boldsymbol{\beta}) = \left(s[y_{T_1+1} - \delta - \boldsymbol{f}'_{T_1+1}\boldsymbol{\beta}], ..., s[y_T - \delta - \boldsymbol{f}'_T\boldsymbol{\beta}]\right)'$$

follows a distribution free of nuisance parameters, conditional on $\mathbf{F}^{(2)}$. Its exact distribution can be simulated to any degree of accuracy simply by repeatedly drawing $\tilde{S}_{T_2} = (\tilde{s}_1, ..., \tilde{s}_{T_2})'$ whose elements are independent Bernoulli variables such that $\Pr[\tilde{s}_t = 1] = \Pr[\tilde{s}_t = -1] = 1/2$.

A corollary of this proposition is that any function of the disturbance sign vector and the factors, say $\Psi = \Psi(s(\mathbf{y} - \delta\boldsymbol{\iota} - \mathbf{F}^{(2)}\boldsymbol{\beta}); \mathbf{F}^{(2)})$, is also pivotal, conditional on $\mathbf{F}^{(2)}$. To see the usefulness of this result, consider the problem of testing

$$H_0(\delta_0, \boldsymbol{\beta}_0) : \delta = \delta_0,\ \boldsymbol{\beta} = \boldsymbol{\beta}_0, \tag{10}$$

where δ_0 and $\boldsymbol{\beta}_0$ are specified values. Following Coudin and Dufour, we consider two test statistics for (10) given by the quadratic forms

$$SX(\delta_0, \boldsymbol{\beta}_0) = s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{F}^{(2)}\boldsymbol{\beta}_0)'\mathbf{X}\mathbf{X}'s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{F}^{(2)}\boldsymbol{\beta}_0), \tag{11}$$

$$SP(\delta_0, \boldsymbol{\beta}_0) = s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{F}^{(2)}\boldsymbol{\beta}_0)'\mathbf{P}(\mathbf{X})s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{F}^{(2)}\boldsymbol{\beta}_0), \tag{12}$$

where $\mathbf{X} = [\boldsymbol{\iota}, \mathbf{F}^{(2)}]$ and $\mathbf{P}(\mathbf{X}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ projects orthogonally onto the subspace spanned by the columns of $\mathbf{X}$. Boldin, Simonova, and Tyurin (1997) show that these statistics can be associated with locally most powerful tests in the case of i.i.d. disturbances under some regularity conditions and Coudin and Dufour extend that proof to more general disturbances that are not necessarily i.i.d., but only satisfy the mediangale property. The statistic in (11) can be interpreted as a sign-based GMM statistic which exploits the property that each element of $s(y_t - \delta_0 - \boldsymbol{f}_t'\boldsymbol{\beta}_0)[1, \boldsymbol{f}_t']'$ is a conditional mediangale under $H_0(\delta_0, \boldsymbol{\beta}_0)$. Note also that (12) can be interpreted as a sign analogue of the usual F test for testing the hypothesis that all the coefficients in a regression of $s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{F}^{(2)}\boldsymbol{\beta}_0)$ on $\mathbf{X}$ are zero.

Under $H_0(\delta_0, \boldsymbol{\beta}_0)$ and conditional on $\mathbf{F}^{(2)}$, the statistics in (11) and (12) are distributed like $\widetilde{SX} = \tilde{S}_{T_2}'\mathbf{X}\mathbf{X}'\tilde{S}_{T_2}$ and $\widetilde{SP} = \tilde{S}_{T_2}'\mathbf{P}(\mathbf{X})\tilde{S}_{T_2}$, respectively. This means that appropriate critical values from the conditional distributions may be found to obtain finite-sample tests of $H_0(\delta_0, \boldsymbol{\beta}_0)$. For instance, consider the statistic in (12). The decision rule is then to reject $H_0(\delta_0, \boldsymbol{\beta}_0)$ at level α if $SP(\delta_0, \boldsymbol{\beta}_0)$ is greater than the $(1-\alpha)$-quantile of the distribution obtained by simulating $\widetilde{SP}$, say c_α^{SP}. (A critical value c_α^{SX} can be found in similar fashion by simulating values $\widetilde{SX}$.) When (11) and (12) are evaluated at the true parameter values δ and $\boldsymbol{\beta}$, Proposition 2 implies that $\Pr[SX(\delta, \boldsymbol{\beta}) > c_\alpha^{SX}] = \alpha$, and $\Pr[SP(\delta, \boldsymbol{\beta}) > c_\alpha^{SP}] = \alpha$ as well. So for all $0 < \alpha < 1$, the critical regions $\{SX(\delta_0, \boldsymbol{\beta}_0) > c_\alpha^{SX}\}$ and $\{SP(\delta_0, \boldsymbol{\beta}_0) > c_\alpha^{SP}\}$ each have size α. Note also that the critical values c_α^{SX} and c_α^{SP} only need to be computed once, since they do not depend on δ_0 and $\boldsymbol{\beta}_0$ in (10).

Here the value of interest is $\delta_0 = 0$ which means that we are dealing with point null hypotheses of the form

$$H_0(\boldsymbol{\beta}_0) : \delta = 0, \, \boldsymbol{\beta} = \boldsymbol{\beta}_0, \tag{13}$$

where $\beta_0 \in \mathcal{B}$, a set of admissible parameter values for $\boldsymbol{\beta}$. The null hypothesis implied by (3) that we wish to test is

$$H_0^p : \bigcup_{\beta_0 \in \mathcal{B}} H_0(\boldsymbol{\beta}_0), \tag{14}$$

the union of (13) taken over $\mathcal{B}$. In order to test such a hypothesis, we appeal to a *minimax* argument which may be stated as "reject the null whenever for all admissible values of the nuisance parameters under the null, the corresponding point null hypothesis is rejected;" see Savin (1984). In general, this rule consists of maximizing the p-value of the sample test statistic over the set of nuisance parameters. Here it amounts to minimizing the values of the SX and SP statistics over $\mathcal{B}$. To see why, define

$$SX_L = \inf_{\beta_0 \in \mathcal{B}} SX(0, \boldsymbol{\beta}_0) \text{ and } SP_L = \inf_{\beta_0 \in \mathcal{B}} SP(0, \boldsymbol{\beta}_0), \tag{15}$$

and observe that under H_0^p in (14) we have

$$0 \leq SP_L \leq SP(0, \boldsymbol{\beta}),$$

which shows that SP_L is boundedly pivotal. This property further implies under H_0^p that

$$\Pr[SP_L > c_\alpha^{SP}] \leq \Pr[SP(0, \boldsymbol{\beta}) > c_\alpha^{SP}] = \alpha.$$

In words, the test that rejects the null hypothesis H_0^p whenever $SP_L > c_\alpha^{SP}$ has level α. The same argument applies to (11) to get the critical region $SX_L > c_\alpha^{SX}$.

We compute SX_L and SP_L in (15) by searching over a grid $\mathcal{B}(\hat{\boldsymbol{\beta}})$ specified around LAD point estimates $\hat{\boldsymbol{\beta}}$, which are computed in the restricted (i.e. no intercept) median regression model $\mathbf{y} = \mathbf{F}^{(2)}\boldsymbol{\beta} + \mathbf{u}$. Of course, more sophisticated optimization methods such as simulated annealing could be used to find SX_L and SP_L. The advantage of the naive grid search is that it is completely reliable and feasible when the dimension of $\boldsymbol{\beta}$ is not too large. An important remark is that the search for SX_L and SP_L can be stopped and the null hypothesis can no longer be rejected at level α as soon as a grid point yields a non-rejection. For instance, if $SP(0, \hat{\boldsymbol{\beta}}) \leq c_\alpha^{SP}$ then SP_L does not reject either and H_0^p in (14) is not significant at level α.

3.4 Summary of Test Procedure

Suppose one wishes to use the SP_L statistic in (15). In a preliminary step, the reference distribution for that statistic is simulated to the desired degree of accuracy by generating a large number, say M, of simulated i.i.d. values $\widetilde{SP}_1, ..., \widetilde{SP}_M$ and the α-level critical value c_α^{SP} is determined from the simulated distribution. The rest of the test procedure then proceeds according to the following steps:

1. Compute the estimates $\hat{a}_i$ of a_i, for $i = 1, ..., N$, using the first T_1 observations on $\mathbf{r}_t$ and $\mathbf{f}_t$.

2. Compute the weights $\hat{\omega}_i$, $i = 1, ..., N$, according to:

$$\hat{\omega}_i = \text{sign}(\hat{a}_i)\frac{1}{N},$$

   and then compute T_2 portfolio returns as $y_t = \sum_{i=1}^{N} \hat{\omega}_i r_{it}$, for $t = T_1 + 1, ..., T$.

3. Find $SP_L = \inf_{\beta_0 \in \mathcal{B}} SP(0, \boldsymbol{\beta}_0)$.

4. Reject the null hypothesis H_0^p at level α if $SP_L > c_\alpha^{SP}$, or equivalently in terms of the p-value if $\hat{p}(SP_L) \leq \alpha$. Otherwise, accept H_0^p. Here the p-value can be computed as

$$\hat{p}(SP_L) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{I}[\widetilde{SP}_j > SP_L],$$

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

This procedure yields a distribution-free test in the sense that it remains exact over the class of all disturbance distributions satisfying (6). Note that the procedure is conditional on the factors, meaning that the critical value c_α^{SP} can only be obtained after the data have been observed.

4 SIMULATION EVIDENCE

We present the results of some simulation experiments to compare the performance of the proposed test procedure with several standard tests. The first of the benchmarks for comparison purposes is the GRS J_1 test in (5). The other benchmarks are the usual LR test (J_2), an adjusted LR test (J_3) suggested by Jobson and Korkie (1982), and a test based on GMM (J_4) proposed by MacKinlay and Richardson (1991). The latter is a particularly important benchmark here, since in principle it is robust to non-normality and heteroskedasticity of returns. We also include in our comparisons two distribution-free tests (SD and WD) developed by Gungor and Luger (2009) that are applicable even if N is large, but only for the single-factor case. Details about all these tests are given in the online Appendix.

The model specification we examine is given by

$$r_{it} = a_i + b_{i1}f_{1t} + b_{i2}f_{2t} + b_{i3}f_{3t} + \varepsilon_{it}, \text{ for } t = 1, ..., T, \;\; i = 1, ..., N, \tag{16}$$

with common factor returns following independent stochastic volatility processes of the form

$$f_{jt} = \exp(h_{jt}/2)\epsilon_{jt}, \text{ with } h_{jt} = \lambda_j h_{j,t-1} + \xi_{jt},$$

where the independent terms ϵ_{jt} and ξ_{jt} are both i.i.d. according to a standard normal distribution and the persistence parameters λ_j are set to 0.5. The disturbances in (16) are subject to contemporaneous heteroskedasticity of the form $\varepsilon_{it} = \exp(\lambda_i f_t^*/2)\eta_{it}$, where the innovations η_{it} are standard normal and the λ_is are randomly drawn from a uniform distribution between 1.5 and 2.5. We set $f_t^* = (f_{1t} + f_{2t} + f_{3t})/3$ so that all three factors contribute equally to the variance heterogeneity; in the single-factor version ($b_{i2} = b_{i3} = 0$) we set $f_t^* = f_{1t}$. It should be noted that such a contemporaneous heteroskedastic specification finds empirical support in Duffee (1995, 2001) and it is easy to see that it generates ε_{it}s with time-varying excess kurtosis—a well-known feature of asset returns. The b_{ij}s in (16) are randomly drawn from a uniform distribution between 0.5 and 1.5. All the tests are conducted at the nominal 5% level and critical values for SX_L and SP_L are determined using $M = 10,000$. In the experiments, we choose mispricing values a and set half the intercept values as $a_i = a$ and the other half as $a_i = -a$. We denote this in the tables as $|a_i| = a$. The estimates of a_i, $i = 1, ..., N$, in Step 1 of the procedure are found via LAD. Finally, each experiment comprises 1000 replications.

In the application of the test procedure, a choice needs to be made about where to split the sample. While this choice has no effect on the level of the tests, it obviously matters for their power. We do not have analytical results on how to split the sample, so we resort to simulations. Overall, the results presented in the online Appendix suggest that no less that 30% and no more than 50% of the time-series observations should be used as the first subsample in order to maximize power. Accordingly, we pursue the testing strategy represented by $T_1 = 0.4T$.

Tables 1 and 2 show the empirical size (Panel A) and power (Panel B) of the considered tests when $|a_i| = 0.15$ and $T = 60$, 120 and N takes on values between 10 and 500. The chosen value for $|a_i|$ is well within the range found in our empirical application, where the intercepts estimated with monthly stock returns range in values from -0.5 to 1.5. When they don't respect the nominal level constraint, the power results for the J tests are based on size-corrected critical values. It is important to emphasize that size-corrected tests are not feasible in practice, especially under the very general symmetry condition in (6). They are merely used here as theoretical benchmarks for the truly distribution-free tests. In particular, we wish to see how the power of the new tests compares to these benchmarks as T and N vary. Recall that the parametric tests are not computable when N exceeds T; those cases are indicated with "-" in the tables.

Panel A of Tables 1 and 2 reveals that all the parametric J tests have massive size distortions, and these over-rejections worsen as N increases for a given T. The sensitivity of the GRS test to contemporaneous heteroskedasticity is also documented in MacKinlay and Richardson (1991), Zhou (1993), and Gungor and Luger (2009). When $T = 120$ and $N = 10$, the J tests all have empirical sizes around 20%. The probability of a Type I error for all those tests exceeds 65% when N is increased to 50. In sharp contrast, the four distribution-free tests satisfy the nominal 5% level constraint, no matter T and N. In Panel B, we see the same phenomenon as in Figure 1: for a fixed T, the power of the GRS J_1 test rises and then eventually drops as N keeps on increasing. Note that J_1, J_2, and J_3 have identical size-corrected powers, since they are all related via monotonic transformations (Campbell, Lo, and MacKinlay 1997, Ch. 5). In contrast, the power of the SD and WD tests, as well as that of the new SX_L and SP_L tests, increases monotonically with N.

At this point, one may wonder what is the advantage of the new SX_L and SP_L tests over the SD and WD tests of Gungor and Luger (2009) since the latter display better power in Panel B of Table 1. Those tests achieve higher power because they eliminate the b_{i1}s from the inference problem through the use of long differences, whereas the new tests proceed through a minimization of the test statistics over the intervening nuisance parameter space. A limitation of the SD and WD tests, however, is that they are valid only under the assumption that the single-factor model disturbances are cross-sectionally independent. The online Appendix reports additional simulation results showing that the SD and WD tests are fairly robust to mild cross-sectional correlation, but start over-rejecting as the cross-sectional dependence increases and this problem is further exacerbated as N increases. Empirical sizes are also reported when the model disturbances are asymmetric, and the SX_L and SP_L tests are found to be quite robust to departures from symmetry.

Notice also that Table 2 has no results for the SD and WD tests, since they are not computable in the presence of multiple factors. The overall pattern in Table 2 echoes the previous findings for the J tests about their size distortions and diminishing power as N increases. What's new in Table 2 is that the SX_L and SP_L tests appear generally more conservative, so N needs to be increased further in order to attain the power levels seen in Table 1. In the empirical illustration that follows, we apply the new tests with $N = 10$, 100, and 503 test assets.

5 EMPIRICAL ILLUSTRATION

In this section we illustrate the new tests with two empirical applications. First, we examine the Sharpe-Lintner version of the CAPM. This single-factor model uses the excess returns of a value-weighted stock market index of all stocks listed on the NYSE, AMEX, and NASDAQ markets. Second, we test the more general three-factor model of Fama and French (1993), which adds two factors to the CAPM specification: (i) the average return on three small market 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. Note that the CAPM is nested within the Fama-French model. This means that if there was no sampling uncertainty, finding that the market portfolio is mean-variance efficient would trivially imply the validity of the three-factor model. Conversely, if the three-factor model does not hold, then the CAPM is also rejected.

We test both specifications with three sets of test assets comprising the stocks traded on the NYSE, AMEX, and NASDAQ markets for the 38-year period from January 1973 to December 2010 (456 months). The first two data sets are the monthly returns on 10 portfolios formed on size, and 100 portfolios formed on both size and the book-to-market ratio. Those two data sets are available in Ken French's online data library. The third data set comprises the returns on 503 individual stocks traded on the markets mentioned above. These represent all the stocks for which there is data in the Center for Research in Security Prices (CRSP) monthly stock files for the entire 38-year sample period. Finally, we use the one-month U.S. Treasury bill as the risk-free asset.

The full sample period contains several extreme observations. For instance, the returns during the stock market crash of October 1987 and the financial crisis of 2008 are obviously not representative of normal market activity; we discuss the effects of extreme observations in subsection 5.3. It is also quite common in the empirical finance literature to perform asset pricing tests over subperiods out of concerns about parameter stability. So in addition to the entire 38-year period, we also examine six 5-year, one 8-year, and three 10-year subperiods. For other examples of this practice, see Campbell, Lo, and MacKinlay (1997), Gungor and Luger (2009), and Ray, Savin, and Tiwari (2009). Here we present the results based on the 10 size portfolios and 503 individual stocks, while those based on the 100 size and book-to-market portfolios are reported in the online Appendix.

5.1 10 Size Portfolios

Table 3 reports the CAPM test results based on the ten size portfolios. The numbers reported in parenthesis are p-values and the entries in bold represent cases of significance at the 5% level. We see that the parametric J tests reject the null hypothesis over the entire sample period with p-values no more than 5%. The non-parametric SD and WD tests also indicate strong rejections. In contrast, the SX_L and SP_L tests clearly do not reject the mean-variance efficiency of the market index.

Looking at the subperiods, we see that the only rejection by the new tests occurs with SP_L in the 10-year subperiod 1/73–12/82. In the 5-year subperiod 1/98–12/02, the J_2 and J_4 tests reject the CAPM specification. The results for the J tests during the last 10-year subperiod from 1/93 to 12/02 agree with the rejection findings for the entire sample period. Besides the obvious differences between the parametric and non-parametric inference results, Table 3 also reveals some differences between the SD and WD tests and the proposed SX_L and SP_L tests. One possible reason for the disagreement across these non-parametric tests could be the presence of cross-sectional disturbance correlations. Indeed, the SD and WD tests are not invariant to such correlations, whereas the new tests allow for cross-sectional dependencies just like the GRS test.

Table 4 shows the results for the Fama-French model. For the entire 38-year sample period, the results in Table 4 are in line with those for the single-factor model in Table 3. The standard J tests reject the null with very low p-values, whereas the distribution-free SX_L and SP_L tests are not significant. In the 5-year subperiods, we see some disagreements among the parametric tests. For instance, during 1/98–12/02 the J_1 and J_3 tests indicate non-rejections, while J_2 and J_4 point toward rejections of the null. The results for the last two 10-year subperiods resemble those for the entire sample period and the J tests depict a more coherent picture.

Table 4 shows that the SX_L and SP_L tests never reject the three-factor specification. Taken at face value, these results would suggest that the excess returns of the 10 size portfolios are well explained by the three Fama-French factors. This is entirely consistent with the non-rejections seen in Table 3 and it suggests that the size and the book-to-market factors play no role; i.e., the CAPM factor alone can price the 10 size portfolios.

Upon observing that the Fama-French model is never rejected by the non-parametric SX_L and SP_L tests with 10 test assets, one may be concerned about the ability of the new procedure to reject the null when the alternative is true. In order to boost power, we proceed next with a 50-fold increase in the number of test assets.

5.2 503 Individual Stocks

Tables 5 and 6 report the test results using the returns on 503 individual stocks. Here the J tests cannot be computed, since $N > T$. When compared to the test outcomes with the 100 size and book-to-market portfolios (reported in the online Appendix), the most striking result is that now for the entire sample period ($T = 456$) the preferred SP_L test no longer indicates a rejection of either the CAPM nor the Fama-French three-factor model. The SD and WD tests also agree with the non-rejection of the CAPM when moving from those 100 portfolios to individual stocks.

These results suggest that the excess returns on individual stocks are well explained by the CAPM, which in turn suggests that the size and book-to-market factors play no role in pricing this collection of assets. It also appears that creating portfolios on the basis of size and book-to-market biases the test outcomes toward a rejection of the model's validity. This finding with the newly proposed SP_L test is entirely consistent with the analysis in Lo and MacKinlay (1990), who show that 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; see also Berk (2000) for related theoretical analysis. Finally, it is interesting to note that this conclusion about the validity of the CAPM is also reached by Zhou (1993), Vorkink (2003), Gungor and Luger (2009), and Ray, Savin, and Tiwari (2009).

5.3 Extreme Observations

Looking back upon the results in Tables 3 and 4 with 10 size portfolios, one might think that the differences between the parametric J tests and the SX_L and SP_L tests is due to a lack of power by the latter when N is small. Yet another plausible reason for these differences is the adverse effect that a small number of extreme observations can have on the OLS estimates used to compute the J tests; see Vorkink (2003). To investigate that possibility we recompute the parametric tests with Winsorized data. This procedure has the effect of decreasing the magnitude of extreme observations but leaves them as important points in the sample.

Table 7 shows the results of the J tests with the 10 size portfolios when the full-sample returns are Winsorized at the 0.3%, 0.5%, 0.7%, and 1% levels. In the single-factor case (Panel A), the J tests cease to be significant at the 5% level with returns Winsorized at 0.3%. For the three-factor model (Panel B), the same pattern of increasing p-values occurs across Winsorization levels. These results clearly show that OLS-based inference can be very sensitive to the presence of even just a few extreme observations.

6 CONCLUSION

The beta-pricing representation of linear factor pricing models is typically assessed with tests based on OLS or GMM. In this context, standard asymptotic theory is known to provide a poor approximation to the finite-sample distribution of those test statistics, even with fairly large samples. In particular, the asymptotic tests tend to over-reject the null hypothesis when in fact it is true, and these size distortions grow quickly as the number of included test assets increases. So the conclusions of empirical studies that adopt such procedures can lead one to spuriously reject the validity of the asset pricing model.

Exact finite-sample methods that avoid the spurious rejection problem usually rely on strong distributional assumptions about the model's disturbance terms. A prominent example is the GRS test which assumes that the disturbances are identically distributed each period according to a multivariate normal distribution. Yet it is known from the empirical literature that financial asset returns are non-normal, exhibiting time-varying covariance structures and excess kurtosis. These stylized facts would put into question the reliability of any inference method which assumes that the cross-sectional distribution of disturbance terms is homogenous over time.

Another serious problem with standard inference methods has to do with the choice of how many tests assets to include. Indeed, if too many are included relative to the number of available time-series observations, the GRS test may lose all its power or may not even be computable. In fact, any procedure that relies on unrestricted estimates of the covariance matrix of regression disturbances will no longer be computable owing to the singularity that occurs when the size of the cross-section exceeds the length of the time series.

In this paper we have proposed a finite-sample test procedure that overcomes these problems. Specifically, our statistical framework makes no parametric assumptions about the distribution of the disturbance terms in the factor model. The only requirement is that the cross-section disturbance vectors be independent over time, conditional on the factors, and reflectively symmetric each period. The class of reflectively symmetric distributions includes elliptically symmetric ones, which are theoretically consistent with mean-variance analysis. Our non-parametric framework leaves open the possibility of unknown forms of time-varying non-normalities and many other distribution heterogeneities, such as time-varying covariance structures, time-varying kurtosis, etc.

The procedure is an adaptive one that first splits the sample to combine the assets into a single portfolio using weights based on the signs of estimated regression intercepts from a subsample. This approach formalizes the usual practice of forming portfolios to solve the problem of too many test assets. Of course, it could also be used in conjunction with an assumed form of the multivariate disturbance distribution to devise a parametric test. The Lehmann and Stein (1949) impossibility theorem, however, shows that if we wish to remain completely agnostic about heteroskedasticity, then the only valid tests are ones based on sign statistics. Even though some studies such as Affleck-Graves and McDonald (1989) report evidence showing the GRS test to be fairly robust to (some specified) deviations from normality, we find it hard to have faith in a parametric procedure whose assumptions are so obviously at odds with the empirical evidence. Moreover, our results show that the power of the new sign-based test procedure increases as either the time-series lengthens and/or the cross-section becomes larger. So the truly robust inference procedure developed here offers a very compelling way to assess the validity of linear factor pricing models, especially with a large number of test assets.

ACKNOWLEDGMENTS

We thank Lynda Khalaf, Antonio Diez de los Rios, two anonymous referees, an associate editor, and the editor Jonathan Wright for helpful comments. Earlier versions of this paper were presented at Northern Illinois University, the 2010 Meetings of the Midwest Econometrics Group, the 2010 International Conference on “High-Dimensional Econometric Modelling” at Cass Business School, and the 2011 North American Summer Meeting of the Econometric Society. 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. and B. McDonald. 1989. "Nonnormalities and tests of asset pricing theories." *Journal of Finance* 44: 889–908.

Bassett G. and R. Koenker. 1978. "Asymptotic theory of least absolute error regression." *Journal of the American Statistical Association* 73: 618–622.

Beaulieu, M.-C., Dufour, J.-M. and L. Khalaf. 2007. "Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: an exact simulation-based approach." *Journal of Business and Economic Statistics* 25: 398-410.

Berk, J. 2000. "Sorting out sorts." *Journal of Finance* 55: 407–427.

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

Boldin, M.V., Simonova, G.I. and Y.N. Tyurin. 1997. Sign-Based Methods in Linear Statistical Models. American Mathematical Society, Maryland.

Bossaerts, P. and P. Hillion. 1995. "Testing the mean-variance efficiency of well-diversified portfolios in very large cross-sections." *Annales d'Économie et de Statistique* 40: 93–124.

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

Chou, P.-H. and G. Zhou. 2006. "Using bootstrap to test portfolio efficiency." *Annals of Economics and Finance* 1: 217–249.

Coudin, E. and J.-M. Dufour. 2009. "Finite-sample distribution-free inference in linear median regressions under heteroscedasticity and non-linear dependence of unknown form." *Econometrics Journal* 12: S19–S49.

Duffee, G.R. 1995. "Stock returns and volatility: a firm-level analysis." *Journal of Financial Economics* 37: 399–420.

Duffee, G.R. 2001. "Asymmetric cross-sectional dispersion in stock returns: evidence and implications." Federal Reserve Bank of San Francisco Working Paper No. 2000-18.

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

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

Dufour, J.-M. and A. Taamouti. 2010. "Exact optimal inference in regression models under heteroskedasticity and non-normality of unknown form." *Computational Statistics and Data Analysis* 54: 2532–2553.

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

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

Ferson, W.E. and S.R. Foerster. 1994. "Finite sample properties of the Generalized Method of Moments in tests of conditional asset pricing models." *Journal of Financial Economics* 36: 29–55.

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

Grinblatt, M. and S. Titman. 1987. "The relation between mean-variance efficiency and arbitrage pricing." *Journal of Business* 60: 97–112.

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

Hsu, D.A. 1982. "A Bayesian robust detection of shift in the risk structure of stock market returns." *Journal of the American Statistical Association* 77: 29–39.

Huberman, G., Kandel, S. and R.F. Stambaugh. 1987. “Mimicking portfolios and exact arbitrage pricing.” *Journal of Finance* 42: 1–9.

Jobson, J.D. 1982. “A multivariate linear regression test for the Arbitrage Pricing Theory.” *Journal of Finance* 37: 1037–1042 .

Jobson, J.D. and B. Korkie. 1982. “Potential performance and tests of portfolio efficiency.” *Journal of Financial Economics* 10: 433–466.

Jobson, J.D. and B. Korkie. 1985. “Some tests of linear asset pricing with multivariate normality.” *Canadian Journal of Administrative Sciences* 2: 114–138.

Jouneau-Sion, F. and O. Torrès. 2006. “MMC techniques for limited dependent variables models: implementation by the branch-and-bound algorithm.” *Journal of Econometrics* 133: 479–512.

Kocherlakota, N.R. 1997. “Testing the Consumption CAPM with heavy-tailed pricing errors.” *Macroeconomic Dynamics* 1: 551–567.

Lehmann, E.L. and C. Stein. 1949. “On the theory of some non-parametric hypotheses.” *Annals of Mathematical Statistics* 20: 28–45.

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. and A.C. MacKinlay. 1990. “Data-snooping biases in tests of financial asset pricing models.” *Review of Financial Studies* 3: 431–467.

MacKinlay, A.C. and M.P. Richardson. 1991. “Using Generalized Method of Moments to test mean-variance efficiency.” *Journal of Finance* 46: 511–527.

Merton, R.C. 1973. “An intertemporal capital asset pricing model.” *Econometrica* 41: 867–887.

Newey, W.K. and K.D. West. 1987. “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.” *Econometrica* 55: 703–708.

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

Roll, R. 1979. "A reply to Mayers and Rice (1979)." *Journal of Financial Economics* 7: 391–400.

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

Savin, N.E. 1984. "Multiple hypothesis testing." In Handbook of Econometrics, Volume 2 (Griliches, Z. and M.D. Intriligator, eds.), North-Holland, Amsterdam.

Sentana, E. 2009. "The econometrics of mean-variance efficiency tests: a survey." *Econometrics Journal* 12: C65–C101.

Shanken, J. 1987. "A Bayesian approach to testing portfolio efficiency." *Journal of Financial Economics* 19: 195–215.

Shanken, J. 1996. "Statistical methods in tests of portfolio efficiency: a synthesis." In Handbook of Statistics, Vol. 14: Statistical Methods in Finance. (G.S. Maddala and C.R. Rao, eds.), North-Holland, Amsterdam.

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

Vorkink, K. 2003. "Return distributions and improved tests of asset pricing models." *Review of Financial Studies* 16: 845–874.

White, H. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica* 48: 817–838.

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

Table 1. Empirical size and power: 1-factor model

T	N	J_1	J_2	J_3	J_4	SD	WD	SX_L	SP_L
Panel A: Size									
60	10	22.7	33.7	22.9	27.1	5.0	5.1	0.7	1.0
	25	43.9	79.9	47.7	60.2	5.1	4.5	0.6	1.4
	50	46.4	95.3	86.2	82.7	4.9	4.7	0.3	0.6
	100	-	-	-	-	4.8	4.3	0.3	1.1
	200	-	-	-	-	5.1	3.6	0.5	1.2
120	10	18.9	22.7	18.9	18.5	5.0	4.2	0.7	1.4
	25	38.6	56.4	39.3	43.5	3.7	3.7	0.5	1.9
	50	65.2	93.3	68.9	75.7	4.1	4.3	0.5	1.6
	100	67.6	94.5	89.8	81.7	5.1	4.8	1.4	1.3
	200	-	-	-	-	4.0	3.9	1.5	1.7
Panel B: Size-corrected power									
60	10	18.7	18.7	18.7	21.4	14.2	17.2	1.5	3.3
	25	31.7	31.7	31.7	30.8	21.7	23.9	2.2	5.0
	50	34.6	34.6	34.6	22.7	31.4	34.7	5.0	9.0
	100	-	-	-	-	50.8	58.5	10.3	15.4
	200	-	-	-	-	74.4	82.6	19.8	28.2
120	10	24.1	24.1	24.1	24.7	29.7	30.9	6.2	10.8
	25	50.7	50.7	50.7	52.9	47.8	53.1	14.1	21.3
	50	78.9	78.9	78.9	74.9	71.2	77.3	30.4	38.3
	100	73.8	73.8	73.8	65.3	89.6	95.3	52.3	60.1
	200	-	-	-	-	99.0	99.6	78.6	84.1

NOTES: This table reports the empirical size (Panel A) and size-corrected power (Panel B) of the GRS test (J_1), the LR test (J_2), an adjusted LR test (J_3), a GMM-based test (J_4), a sign test (SD), a Wilcoxon signed rank test (SD), and the proposed SX_L and SP_L tests. The returns are generated according to a single-factor model with contemporaneous heteroskedastic disturbances. The pricing errors are zero under the null hypothesis, whereas $N/2$ pricing errors are set equal to -0.15 and the other half are set to 0.15 under the alternative. The nominal level is 0.05 and entries are percentage rates. The results are based on 1000 replications and the symbol "-" is used whenever a test is not computable.

Table 2. Empirical size and power: 3-factor model

T	N	J_1	J_2	J_3	J_4	SD	WD	SX_L	SP_L
Panel A: Size									
60	25	30.2	66.9	34.3	57.2	-	-	0.0	0.0
	50	29.5	99.8	86.5	92.4	-	-	0.0	0.1
	100	-	-	-	-	-	-	0.0	0.0
	200	-	-	-	-	-	-	0.0	0.0
	500	-	-	-	-	-	-	0.0	0.1
120	25	25.6	41.5	26.6	36.9	-	-	0.0	0.3
	50	49.4	89.3	55.8	72.7	-	-	0.0	0.3
	100	60.4	100.0	97.8	95.6	-	-	0.2	0.8
	200	-	-	-	-	-	-	0.4	0.6
	500	-	-	-	-	-	-	1.8	2.4
Panel B: Size-corrected power									
60	25	35.2	35.2	35.2	33.6	-	-	0.0	0.2
	50	24.1	24.1	24.1	23.8	-	-	0.0	1.4
	100	-	-	-	-	-	-	0.3	3.9
	200	-	-	-	-	-	-	0.6	12.3
	500	-	-	-	-	-	-	5.9	43.7
120	25	72.7	72.7	72.7	80.2	-	-	1.0	7.9
	50	90.5	90.5	90.5	90.1	-	-	7.2	19.4
	100	81.9	81.9	81.9	69.0	-	-	28.2	53.7
	200	-	-	-	-	-	-	65.8	88.6
	500	-	-	-	-	-	-	95.8	99.2

NOTES: This table reports the empirical size (Panel A) and size-corrected power (Panel B) of the GRS test (J_1), the LR test (J_2), an adjusted LR test (J_3), a GMM-based test (J_4), a sign test (SD), a Wilcoxon signed rank test (SD), and the proposed SX_L and SP_L tests. The returns are generated according to a 3-factor model with contemporaneous heteroskedastic disturbances. The pricing errors are zero under the null hypothesis, whereas $N/2$ pricing errors are set equal to -0.15 and the other half are set to 0.15 under the alternative. The nominal level is 0.05 and entries are percentage rates. The results are based on 1000 replications and the symbol "-" is used whenever a test is not computable.

Table 3. Tests of the CAPM with 10 size portfolios

Time period	J_1	J_2	J_3	J_4	SD	WD	SX_L	SP_L
38-year period								
1/73–12/10	**1.83**	**18.43**	**18.14**	**18.48**	**38.45**	**35.25**	208.77	0.06
	(0.05)	(0.05)	(0.05)	(0.04)	(0.00)	(0.00)	(0.92)	(0.97)
5-year subperiods and an 8-year subperiod								
1/73–12/77	0.56	6.46	5.71	5.24	16.80	15.09	197.79	4.02
	(0.83)	(0.77)	(0.84)	(0.87)	(0.08)	(0.12)	(0.64)	(0.14)
1/78–12/82	1.12	12.38	10.93	10.90	**54.93**	**42.28**	204.06	3.77
	(0.36)	(0.26)	(0.36)	(0.37)	(0.00)	(0.00)	(0.69)	(0.16)
1/83–12/87	0.81	9.20	8.12	8.75	5.46	5.14	128.48	1.24
	(0.62)	(0.51)	(0.61)	(0.55)	(0.85)	(0.88)	(0.81)	(0.55)
1/88–12/92	0.79	9.01	7.96	8.13	**21.60**	10.83	16.98	0.22
	(0.63)	(0.53)	(0.63)	(0.62)	(0.01)	(0.37)	(0.94)	(0.89)
1/93–12/97	1.08	12.00	10.60	12.64	2.26	1.85	37.58	0.72
	(0.39)	(0.29)	(0.38)	(0.24)	(0.99)	(0.99)	(0.86)	(0.69)
1/98–12/02	1.87	**19.43**	17.16	**18.66**	4.93	2.85	100.96	1.00
	(0.07)	(0.03)	(0.07)	(0.04)	(0.89)	(0.98)	(0.82)	(0.63)
1/03–12/10	1.73	17.84	16.54	17.45	10.16	7.37	45.42	0.03
	(0.09)	(0.06)	(0.09)	(0.06)	(0.42)	(0.68)	(0.93)	(0.98)
10-year subperiods								
1/73–12/82	0.73	7.85	7.39	7.49	**46.00**	**60.16**	1298.92	**15.01**
	(0.69)	(0.64)	(0.68)	(0.67)	(0.00)	(0.00)	(0.38)	(0.00)
1/83–12/92	1.48	15.29	14.40	14.59	11.00	8.30	112.25	0.54
	(0.15)	(0.12)	(0.15)	(0.14)	(0.36)	(0.60)	(0.87)	(0.77)
1/93–12/02	**2.07**	**20.93**	**19.71**	**20.17**	4.60	3.84	146.37	1.06
	(0.03)	(0.02)	(0.03)	(0.02)	(0.92)	(0.95)	(0.85)	(0.58)

NOTES: The results are based on value-weighted returns of 10 portfolios formed on size. The market portfolio is the value-weighted return on all NYSE, AMEX, and NASDAQ stocks and the risk-free rate is the 1-month Treasury bill rate. The numbers in parentheses are p-values and entries in bold represent cases of significance at the 5% level.

Table 4. Tests of the Fama-French model with 10 size portfolios

Time period	J_1	J_2	J_3	J_4	SD	WD	SX_L	SP_L
38-year period								
1/73–12/10	**2.04**	**20.49**	**20.18**	**20.21**	-	-	3480.59	3.57
	(0.03)	(0.02)	(0.03)	(0.02)			(0.82)	(0.47)
5-year subperiods and an 8-year subperiod								
1/73–12/77	0.39	4.62	4.08	5.37	-	-	475.03	4.01
	(0.94)	(0.91)	(0.94)	(0.86)			(0.75)	(0.41)
1/78–12/82	0.77	8.75	7.73	9.29	-	-	114.95	0.41
	(0.65)	(0.55)	(0.65)	(0.50)			(0.97)	(0.98)
1/83–12/87	1.10	12.24	10.82	10.29	-	-	128.47	2.04
	(0.37)	(0.26)	(0.37)	(0.41)			(0.97)	(0.78)
1/88–12/92	1.18	13.03	11.51	12.32	-	-	70.82	0.77
	(0.32)	(0.22)	(0.32)	(0.26)			(0.98)	(0.94)
1/93–12/97	**2.25**	**22.74**	**20.08**	**33.30**	-	-	143.55	5.58
	(0.03)	(0.01)	(0.02)	(0.00)			(0.93)	(0.24)
1/98–12/02	1.70	**17.92**	15.83	**18.93**	-	-	556.46	0.98
	(0.10)	(0.05)	(0.10)	(0.04)			(0.92)	(0.91)
1/03–12/10	1.65	17.01	15.76	**18.02**	-	-	257.68	0.42
	(0.10)	(0.07)	(0.10)	(0.05)			(0.95)	(0.98)
10-year subperiods								
1/73–12/82	0.19	2.09	1.97	2.14	-	-	255.08	1.86
	(0.99)	(0.99)	(0.99)	(0.99)			(0.96)	(0.77)
1/83–12/92	**1.98**	**20.11**	**18.94**	**19.92**	-	-	33.03	0.26
	(0.04)	(0.02)	(0.04)	(0.03)			(0.99)	(0.99)
1/93–12/02	**2.00**	**20.21**	**19.03**	**21.26**	-	-	79.26	0.03
	(0.04)	(0.02)	(0.03)	(0.01)			(0.99)	(0.99)

NOTES: The results are based on value-weighted returns of 10 portfolios formed on size, the returns on the three Fama-French factors, and the 1-month Treasury bill rate as the risk-free rate. The numbers in parentheses are p-values and entries in bold represent cases of significance at the 5% level. The symbol "-" is used whenever a test is not computable.

Table 5. Tests of the CAPM with 503 individual stocks

Time period	J_1	J_2	J_3	J_4	SD	WD	SX_L	SP_L
38-year period								
1/73–12/10	-	-	-	-	478.26	483.68	696.34	2.03
					(0.78)	(0.72)	(0.79)	(0.36)
5-year subperiods and an 8-year subperiod								
1/73–12/77	-	-	-	-	466.13	436.72	242.51	**6.22**
					(0.87)	(0.98)	(0.59)	(0.04)
1/78–12/82	-	-	-	-	**566.26**	522.36	106.39	0.66
					(0.02)	(0.26)	(0.80)	(0.72)
1/83–12/87	-	-	-	-	492.53	504.74	20.79	0.31
					(0.62)	(0.46)	(0.95)	(0.86)
1/88–12/92	-	-	-	-	496.40	500.58	35.88	0.05
					(0.57)	(0.52)	(0.90)	(0.97)
1/93–12/97	-	-	-	-	418.13	432.12	10.76	0.11
					(0.99)	(0.99)	(0.95)	(0.95)
1/98–12/02	-	-	-	-	429.46	353.28	205.36	4.74
					(0.99)	(1.00)	(0.71)	(0.09)
1/03–12/10	-	-	-	-	497.16	484.88	200.58	1.85
					(0.56)	(0.71)	(0.78)	(0.40)
10-year subperiods								
1/73–12/82	-	-	-	-	**594.33**	**600.21**	201.15	2.55
					(0.00)	(0.00)	(0.78)	(0.28)
1/83–12/92	-	-	-	-	511.40	503.47	28.39	0.24
					(0.38)	(0.48)	(0.96)	(0.88)
1/93–12/02	-	-	-	-	538.13	481.90	199.84	2.72
					(0.13)	(0.74)	(0.81)	(0.26)

NOTES: The results are based on the returns of 503 individual stocks traded on the NYSE, AMEX, and NASDAQ markets. The market portfolio is the value-weighted return on all NYSE, AMEX, and NASDAQ stocks and the risk-free rate is the 1-month Treasury bill rate. The numbers in parentheses are p-values and entries in bold represent cases of significance at the 5% level. The symbol "-" is used whenever a test is not computable.

Table 6. Tests of the Fama-French model with 503 individual stocks

Time period	J_1	J_2	J_3	J_4	SD	WD	SX_L	SP_L
38-year period								
1/73–12/10	-	-	-	-	-	-	2387.42	6.71
							(0.89)	(0.15)
5-year subperiods and an 8-year subperiod								
1/73–12/77	-	-	-	-	-	-	182.15	0.29
							(0.93)	(0.99)
1/78–12/82	-	-	-	-	-	-	463.17	2.49
							(0.81)	(0.67)
1/83–12/87	-	-	-	-	-	-	191.72	1.87
							(0.95)	(0.81)
1/88–12/92	-	-	-	-	-	-	246.14	1.12
							(0.88)	(0.90)
1/93–12/97	-	-	-	-	-	-	90.24	0.52
							(0.96)	(0.97)
1/98–12/02	-	-	-	-	-	-	642.42	2.61
							(0.91)	(0.65)
1/03–12/10	-	-	-	-	-	-	149.46	0.15
							(0.98)	(0.99)
10-year subperiods								
1/73–12/82	-	-	-	-	-	-	483.76	0.87
							(0.89)	(0.93)
1/83–12/92	-	-	-	-	-	-	1165.84	0.77
							(0.68)	(0.94)
1/93–12/02	-	-	-	-	-	-	775.66	0.73
							(0.93)	(0.95)

NOTES: These results are based on the returns of 503 individual stocks traded on the NYSE, AMEX, and NASDAQ markets, the returns on the three Fama-French factors, and the 1-month Treasury bill rate as the risk-free rate. The numbers in parentheses are p-values and entries in bold represent cases of significance at the 5% level. The symbol "-" is used whenever a test is not computable.

Table 7. Sensitivity of parametric tests to extreme observations

	0%	0.3%	0.5%	0.7%	1.0%
Panel A: CAPM					
J_1	**1.83**	1.54	0.89	0.64	0.59
	(0.05)	(0.12)	(0.54)	(0.77)	(0.81)
J_2	**18.43**	15.55	9.05	6.54	6.07
	(0.05)	(0.11)	(0.52)	(0.76)	(0.80)
J_3	**18.14**	15.31	8.91	6.44	5.97
	(0.05)	(0.12)	(0.54)	(0.77)	(0.81)
J_4	**18.48**	14.94	8.12	6.11	5.82
	(0.04)	(0.13)	(0.61)	(0.80)	(0.83)
Panel B: Fama-French model					
J_1	**2.04**	1.70	0.87	0.59	0.55
	(0.03)	(0.07)	(0.55)	(0.81)	(0.85)
J_2	**20.49**	17.13	8.91	6.05	5.63
	(0.02)	(0.07)	(0.54)	(0.81)	(0.84)
J_3	**20.18**	16.86	8.77	5.96	5.54
	(0.03)	(0.07)	(0.55)	(0.82)	(0.85)
J_4	**20.21**	16.31	8.16	5.63	5.10
	(0.02)	(0.09)	(0.61)	(0.84)	(0.88)

NOTES: This table shows the results of the parametric tests with the 10 size portfolios when the returns for the full sample period from January 1973 to December 2010 are Winsorized at various small levels. Panels A and B correspond to the 1- and 3-factor models, respectively. The numbers in parenthesis are p-values and bold entries represent cases of significance at the 5% level.

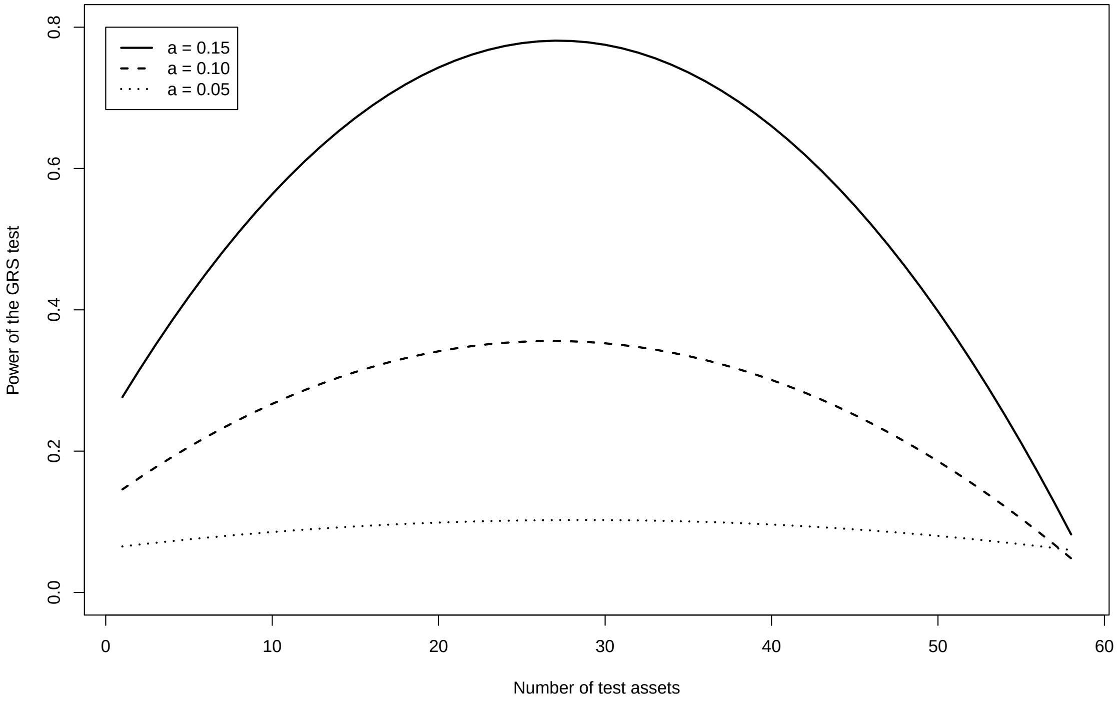

Figure 1. This figure plots the power of the GRS test as a function of the number of included test assets. The returns are generated from a 1-factor model with normally distributed disturbances. The sample size is $T = 60$ and the number of test assets N ranges from 1 to 58. The test is performed at a nominal 0.05 level. The higher power curves are associated with greater pricing errors.