BANK OF CANADA
BANQUE DU CANADA
CELEBRATING 75 YEARS
CÉLÉBRONS 75 ANS

Working Paper/Document de travail
2010-36

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

by Sermin Gungor and Richard Luger

www.bank-banque-canada.ca

Bank of Canada Working Paper 2010-36

December 2010

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

by

Sermin Gungor[1] and Richard Luger[2]

[1]Financial Markets Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
sgungor@bankofcanada.ca

[2]Georgia State University

ISSN 1701-9397

Abstract

We develop a finite-sample procedure to test the beta-pricing representation of linear factor pricing models that is applicable even if the number of test assets is greater than the length of the time series. Our distribution-free framework leaves open the possibility of unknown forms of non-normalities, heteroskedasticity, time-varying correlations, and even outliers in the asset returns. The power of the proposed test procedure increases as the time-series lengthens and/or the cross-section becomes larger. This stands in sharp contrast to the usual tests that lose power or may not even be computable if the cross-section is too large. Finally, we revisit the CAPM and the Fama-French three factor model. Our results strongly support the mean-variance efficiency of the market portfolio.

JEL classification: C12, C14, C33, G11, G12
Bank classification: Econometric and statistical methods; Financial markets

Résumé

Les auteurs élaborent une procédure permettant de tester, en échantillon fini, la représentation des coefficients bêta donnés par les modèles linéaires d'évaluation factorielle, et ce, même si le nombre des actifs dépasse celui des valeurs de la série chronologique. Leur cadre autorise des formes inconnues de distribution autres que la loi normale ainsi que la présence d'hétéroscédasticité, de structures de corrélation variables dans le temps, voire de rendements aberrants. La puissance de la procédure s'accroît avec l'allongement de la série chronologique et la hausse du nombre des actifs. Cette propriété tranche avec les limites des tests habituels, qui perdent de leur puissance ou peuvent même devenir inexécutables si le nombre des actifs est trop élevé. Pour finir, les auteurs réexaminent le modèle d'évaluation des actifs financiers et le modèle trifactoriel de Fama et French. Leurs résultats indiquent clairement que le portefeuille de marché se situe sur la frontière efficiente dans le plan moyenne-variance.

Classification JEL : C12, C14, C33, G11, G12
Classification de la Banque : Méthodes économétriques et statistiques; Marchés financiers

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. Merton (1973) and Breeden (1979) develop models based on investor optimization and equilibrium arguments that also lead to multiple-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 riskfree 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 multiple-beta models. 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, empirical tests of beta-pricing representations can be severely affected and can lead to spurious 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, 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 provide truly exact, finite-sample tests.[1] 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 each period according to a multivariate normal distribution. As we already mentioned, there is ample evidence that financial returns exhibit non-normalities; for more evidence, see Fama (1965), Blattberg and Gonedes (1974), Hsu (1982), Affleck-Graves and McDonald (1989), and Zhou

[1]A number of Bayesian approaches have also been proposed. These include Shanken (1987), Harvey and Zhou (1990), and Kandel, McCulloch, and Stambaugh (1995).

(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 multiple-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 assumptions about the distribution of the disturbances in the 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 Dufour and Taamouti (2005, 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 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.[2]

Our proposed test procedure then exploits results from Coudin and Dufour (2009) on median regressions to construct confidence sets for the model coefficients by inverting exact sign-based statistics. A similar approach is used in Chernozhukov, Hansen, and Jansson (2009) to derive finite-sample confidence sets for quantile regression models. The motivation for using this technique 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 null hypothesis to be tested, 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

[2] The notorious noisiness of unrestricted sample covariances is a well-known problem in the portfolio management literature; see Michaud (1989), Jagannathan and Ma (2003), and Ledoit and Wolf (2003, 2004), among others.

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. Although not required for the proposed procedure, the vector $\boldsymbol{\varepsilon}_t$ is usually assumed to have well-defined first and second moments satisfying $E[\boldsymbol{\varepsilon}_t|\mathbf{f}_t] = \mathbf{0}$ and $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'|\mathbf{f}_t] = \boldsymbol{\Sigma}$, a finite $N \times N$ matrix.

Exact factor pricing implies that expected returns depend linearly on the betas associated with the factor portfolio returns:

$$E_t[\mathbf{r}_t] = \mathbf{B}\boldsymbol{\lambda}_K, \tag{2}$$

where $\boldsymbol{\lambda}_K$ is a $K \times 1$ vector of expected excess returns associated with $\mathbf{f}_t$, which represent market-wide risk premiums since they apply to 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.[3] 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, $\mathbf{a}$, 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 (3) that all the pricing errors are jointly equal

[3] A benchmark portfolio with excess returns r_p is said to be mean-variance efficient with respect to a given set of N test assets with excess returns $\mathbf{r}_t$ if it is not possible to form another portfolio of those N assets and the benchmark portfolio with the same variance as r_p but a higher expected return.

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 non-singular cross-sectional covariance matrix each period, conditional on the the $T \times K$ collection of factors $\mathbf{F} = [\boldsymbol{f}'_1, ..., \boldsymbol{f}'_T]'$. Under normality, the methods of maximum likelihood and ordinary least squares (OLS) yield the same unconstrained estimates of $\mathbf{a}$ and $\mathbf{B}$:

$$\hat{\mathbf{a}} = \bar{\mathbf{r}} - \hat{\mathbf{B}}\bar{\mathbf{f}}, \tag{4}$$

$$\hat{\mathbf{B}} = \left[\sum_{t=1}^{T}(\mathbf{r}_t - \bar{\mathbf{r}})(\mathbf{f}_t - \bar{\mathbf{f}})'\right]\left[\sum_{t=1}^{T}(\mathbf{f}_t - \bar{\mathbf{f}})(\mathbf{f}_t - \bar{\mathbf{f}})'\right]^{-1}, \tag{5}$$

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

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_t - \hat{\mathbf{a}} - \hat{\mathbf{B}}\mathbf{f}_t)(\mathbf{r}_t - \hat{\mathbf{a}} - \hat{\mathbf{B}}\mathbf{f}_t)'. \tag{6}$$

The GRS test statistic is

$$J_1 = \frac{T - N - K}{N}\left[1 + \bar{\mathbf{f}}'\hat{\boldsymbol{\Omega}}^{-1}\bar{\mathbf{f}}\right]^{-1}\hat{\mathbf{a}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{a}}, \tag{7}$$

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 test assets 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 (6) 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. 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 higher 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 H_0 in (3) 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 on the number of test assets. Our approach is motivated by results from classical non-parametric statistics that show that the *only* tests which yield reliable 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. This result is due to Lehmann and Stein (1949); see also Pratt and Gibbons (1981, p. 218), Dufour and Hallin (1991), and Dufour (2003). Next we present the statistical framework and then proceed to describe each step of the procedure.

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 remain symmetrically distributed each period. In what follows the symbol $\stackrel{d}{=}$ stands for the equality in distribution.

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

The diagonal (or reflective) symmetry condition in Assumption 1 can be equivalently expressed in terms of the density function as $f(\boldsymbol{\varepsilon}_t) = f(-\boldsymbol{\varepsilon}_t)$. Recall that a random variable v is symmetric around zero if and only if $v \stackrel{d}{=} -v$, so the symmetry assumption made here represents the most direct non-parametric extension of univariate symmetry. See Serfling (2006) for more on multivariate symmetry. The class of distributions encompassed by the diagonal symmetry condition 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; see Chamberlain (1983), Owen and Rabinovitch (1983), and Berk (1997). A random vector $\mathbf{V}$ is elliptically symmetric around the origin if its density function can be expressed as $|\boldsymbol{\Sigma}|^{-1/2} g(\mathbf{V}'\boldsymbol{\Sigma}^{-1}\mathbf{V})$ for some nonnegative scalar function $g(\cdot)$, where $\boldsymbol{\Sigma}$ is (proportional to) the covariance matrix. The class of elliptically symmetric distributions includes the well-known multivariate normal and Student-t distributions, among others. It is important to emphasize that the diagonal symmetry condition in Assumption 1 is less stringent than elliptical symmetry. For example, a mixture (finite or not) of distributions each one elliptically symmetric around the origin is not necessarily elliptically symmetric but it is diagonally symmetric. Note also that the distribution of $\mathbf{f}_t$ in (1) may be skewed thereby inducing asymmetry in the unconditional distribution of $\mathbf{r}_t$.

Assumption 1 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; see Fama (1965), Blattberg and Gonedes (1974), and Hsu (1982). In particular, the distribution of asset returns appears to have much heavier tails and is more peaked than a normal distribution. The following quote from Fama and MacBeth (1973, p. 619) emphasizes the importance of recognizing non-normalities:

> In interpreting [these] t-statistics one should keep in mind the evidence of Fama (1965) and Blume (1970) which suggests that distributions of common stock returns are "thick-tailed" relative to the normal distribution and probably conform better to nonnormal symmetric stable distributions than to the normal. From Fama and Babiak (1968), this evidence means that when one interprets large t-statistics under the assumption that the underlying variables are normal, the probability or significance levels obtained are likely to be overestimates.

The present framework leaves open not only the possibility of unknown forms of non-normality, but also heteroskedasticity and time-varying correlations 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 covariance structure of the disturbance terms could be any function of the common factors (contemporaneous or not). The simulation study below includes 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) points out, 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 naive portfolios of the test assets. In the spirit of weighted portfolio groupings, 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 H_0 is performed using only the returns on that portfolio observed over the second subsample. It is important to note that in the present framework this approach 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

$$\mathbf{r}_i^1 = a_i \boldsymbol{\iota} + \mathbf{F}^1 \mathbf{b}_i + \boldsymbol{\varepsilon}_i^1, \tag{8}$$

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

Assumption 2. *Only the first T_1 observations on $\mathbf{r}_t$ and $\mathbf{f}_t$ are used to compute the subsample estimates $\hat{a}_1, ..., \hat{a}_N$.*

This assumption does not restrict the choice of estimation method, so the subsample estimates $\hat{a}_1, ..., \hat{a}_N$ could be obtained by OLS, quasi-maximum likelihood, or any other method.[4] A well-known problem with OLS is that it is very sensitive to the presence of large disturbances and outliers. 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. For more on the efficiency of LAD versus OLS, see Glahe and Hunt (1970), Hunt, Dowling, and

[4] Of course, T_1 must at least be enough to obtain the estimates $\hat{a}_1, ..., \hat{a}_N$ by the chosen method.

Glahe (1974), Pfaffenberger and Dinkel (1978), Rosenberg and Carlson (1977), and Mitra (1987). The results reported below in the simulation study and the empirical application are based on LAD.

With the estimates $\hat{a}_1, ..., \hat{a}_N$ 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:

$$\hat{\omega}_i = \frac{\hat{a}_i}{|\hat{a}_1| + ... + |\hat{a}_N|} = sign(\hat{a}_i)\frac{|\hat{a}_i|}{|\hat{a}_1| + ... + |\hat{a}_N|}, \quad (9)$$

for $i = 1, ..., N$, and these weights are then used to find the T_2 returns of a portfolio computed as $y_t = \sum_i^N \hat{\omega}_i r_{it}$, $t = T_1 + 1, ..., T$. Note that having a zero denominator in (9) is a zero probability event in finite samples ($T < \infty$) when the disturbance terms are of the continuous type (as in Assumption 1). Let δ denote the sum of the weighted a_i's and set $\mathbf{x}_t = \mathbf{f}_t$.

Proposition 1. *Under H_0 and when Assumptions 1 and 2 hold, y_t is represented by the single equation*

$$y_t = \delta + \mathbf{x}_t'\boldsymbol{\beta} + u_t, \; \textit{for } t = T_1 + 1, ..., T, \quad (10)$$

where $\delta = 0$ and $(u_{T_1+1}, ..., u_T) \stackrel{d}{=} (\pm u_{T_1+1}, ..., \pm u_T)$, conditional on $\mathbf{F}$ and $\hat{\boldsymbol{\omega}}$.

Proof. The conditional expectation part of (10) follows from the common factor structure in (1), and the fact that δ is zero under H_0 is obvious. The independence of the disturbance vectors maintained in Assumption 1 implies that the T_2 vectors $\boldsymbol{\varepsilon}_t$, $t = T_1 + 1, ..., T$, are conditionally independent of the vector of weights $(\hat{\omega}_1, \hat{\omega}_2, ..., \hat{\omega}_N)$ given $\mathbf{F}$, since under Assumption 2 those weights are based only on the first T_1 observations of $\mathbf{r}_t$ and $\mathbf{f}_t$. Thus we see that given $\mathbf{F}$ and $\hat{\boldsymbol{\omega}}$,

$$\big(\hat{\omega}_1\varepsilon_{1t}, \hat{\omega}_2\varepsilon_{2t}, ..., \hat{\omega}_N\varepsilon_{Nt}\big) \stackrel{d}{=} \big(-\hat{\omega}_1\varepsilon_{1t}, -\hat{\omega}_2\varepsilon_{2t}, ..., -\hat{\omega}_N\varepsilon_{Nt}\big), \tag{11}$$

for $t = T_1 + 1, ..., T$. Let $u_t = \sum_i^N \hat{\omega}_i\varepsilon_{it}$. For a given t, (11) implies that $u_t \stackrel{d}{=} -u_t$, since any linear combination of the elements of a diagonally symmetric vector is itself symmetric (Behboodian 1990, Theorem 2). Moreover, this fact applies to each of the T_2 conditionally independent random variables $u_{T_1+1}, ..., u_T$. So, given $\mathbf{F}$ and $\hat{\boldsymbol{\omega}}$, the 2^{T_2} possible T_2 vectors

$$\big(\pm |u_{T_1+1}|, \pm|u_{T_1+2}|, ..., \pm|u_T|\big)$$

are equally likely values for $(u_{T_1+1}, ..., u_T)$, where $\pm|u_t|$ means that $|u_t|$ is assigned either a positive or negative sign with probability 1/2. □

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 $\sum_{i=1}^N \hat{a}_i$ and another one proposed in Gungor and Luger (2009) based on $\sum_{i=1}^N a_i$. Those approaches can suffer power losses depending on whether the a_i's tend to cancel out. Splitting the sample and applying the weights in (9) when forming the portfolio offsets that problem. Note that these weights do not correspond to any of the usual ones in mean-variance analysis since finding those requires an estimate of the covariance structure and that is precisely what we are trying to avoid. Furthermore, such estimates are not meaningful in our distribution-free context where possible forms of distribution heterogeneity are left completely unspecified. To see why the weights in (9) are reasonable, note that the sign component in the definition of $\hat{\omega}_i$ makes it more likely that all the intercept values in the equation describing $\hat{\omega}_i r_{it}$ will be positive under the alternative hypothesis. The component in (9) pertaining to the absolute values serves to give relatively more weight to the assets that seem to depart more from H_0 and to down weight those that seem to offer relatively less evidence against the null hypothesis.

3.3 Confidence sets

The model in (10) can be represented in matrix form as $\mathbf{y} = \delta\boldsymbol{\iota} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where the elements of $\mathbf{u}$ follow what Coudin and Dufour (2009) call a strict conditional "mediangale." Define a sign function as $s[x] = 1$ if $x > 0$, and $s[x] = -1$ if $x \leq 0$. The following result is an immediate consequence of the mediangale property.

Proposition 2. *Under Assumptions 1 and 2, the T_2 disturbance sign vector*

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

follows a distribution free of nuisance parameters, conditional on $\mathbf{F}$ and $\hat{\boldsymbol{\omega}}$. 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 $\Psi = \Psi(s(\mathbf{y} - \delta\boldsymbol{\iota} - \mathbf{X}\boldsymbol{\beta}); \mathbf{F})$, is also free of nuisance parameters (i.e. pivotal), conditional on $\mathbf{F}$. To see the usefulness of this result, consider the problem of testing $H_0(\delta_0, \boldsymbol{\beta}_0) : \delta = \delta_0, \boldsymbol{\beta} = \boldsymbol{\beta}_0$ against $H_1(\delta_0, \boldsymbol{\beta}_0) : \delta \neq \delta_0$ or $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. Under $H_0(\delta_0, \boldsymbol{\beta}_0)$, the statistic $\Psi(s(\mathbf{y} - \delta\boldsymbol{\iota} - \mathbf{X}\boldsymbol{\beta}); \mathbf{F})$ is distributed like $\Psi(\tilde{S}_{T_2}; \mathbf{F})$, conditional on $\mathbf{F}$. This means that appropriate critical values from the conditional distribution may be found to obtain a finite-sample test of $H_0(\delta_0, \boldsymbol{\beta}_0)$. For example, suppose that $\Psi(\cdot)$ is a non-negative function. The decision rule is then to reject $H_0(\delta_0, \boldsymbol{\beta}_0)$ at level α if $\Psi(s(\mathbf{y} - \delta\boldsymbol{\iota} - \mathbf{X}\boldsymbol{\beta}); \mathbf{F})$ is greater than the $(1 - \alpha)$-quantile of the simulated distribution of $\Psi(\tilde{S}_{T_2}; \mathbf{F})$.

Following Coudin and Dufour (2009), we consider two test statistics given by the quadratic forms

$$SX(\delta_0, \boldsymbol{\beta}_0) = s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{X}\boldsymbol{\beta}_0)'\mathbf{X}\mathbf{X}'s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{X}\boldsymbol{\beta}_0), \quad (12)$$

$$SP(\delta_0, \boldsymbol{\beta}_0) = s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{X}\boldsymbol{\beta}_0)'\mathbf{P}(\mathbf{X})s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{X}\boldsymbol{\beta}_0), \quad (13)$$

where $\mathbf{P}(\mathbf{X}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 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 disturbances that satisfy the mediangale property. It is interesting to note that (13) can be interpreted as a sign analogue of the F-test for testing the hypothesis that all the coefficients in a regression of $s(\mathbf{y} - \delta_0\boldsymbol{\iota} - \mathbf{X}\boldsymbol{\beta}_0)$ on $\mathbf{X}$ are zero.

An exactly distribution-free confidence set for δ and $\boldsymbol{\beta}$ can be constructed simply by inverting either (12) or (13). Consider the test statistic in (13) for example, and let c_α represent its one-sided α-level simulated critical value. A simultaneous confidence set, say $C_{1-\alpha}(\delta, \boldsymbol{\beta})$, with level $1-\alpha$ for δ and $\boldsymbol{\beta}$ is simply the collection of all values of $\delta_0, \boldsymbol{\beta}_0$ for which $SP(\delta_0, \boldsymbol{\beta}_0)$ is less than c_α. Note that the critical value c_α only needs to be computed once, since it does not depend on $\delta_0, \boldsymbol{\beta}_0$.

From the joint confidence set, it is possible to derive conservative confidence sets and intervals for general functions of the coefficients $\delta, \boldsymbol{\beta}$ using the projection method in Coudin and Dufour (2009); see also Abdelkhalek and Dufour (1998), Dufour and Jasiak (2001), Dufour and Taamouti (2005), and Chernozhukov, Hansen, and Jansson (2009) for other examples of this technique. To introduce the method, consider a non-linear function $g(\delta, \boldsymbol{\beta})$ of $\delta, \boldsymbol{\beta}$. It is easy to see that $(\delta, \boldsymbol{\beta}) \in C_{1-\alpha}(\delta, \boldsymbol{\beta}) \Rightarrow g(\delta, \boldsymbol{\beta}) \in g(C_{1-\alpha}(\delta, \boldsymbol{\beta}))$ so that $\Pr[(\delta, \boldsymbol{\beta}) \in C_{1-\alpha}(\delta, \boldsymbol{\beta})] \geq 1-\alpha \Rightarrow \Pr[g(\delta, \boldsymbol{\beta}) \in g(C_{1-\alpha}(\delta, \boldsymbol{\beta}))] \geq 1-\alpha$. This means that $g(C_{1-\alpha}(\delta, \boldsymbol{\beta}))$ is a conservative confidence set for $g(\delta, \boldsymbol{\beta})$; i.e., one for which the level is at least $1-\alpha$. In the special case when $g(\delta, \boldsymbol{\beta})$ is scalar, the interval

$$\left[\inf_{(\delta_0, \boldsymbol{\beta}_0) \in C_{1-\alpha}(\delta, \boldsymbol{\beta})} g(\delta_0, \boldsymbol{\beta}_0), \sup_{(\delta_0, \boldsymbol{\beta}_0) \in C_{1-\alpha}(\delta, \boldsymbol{\beta})} g(\delta_0, \boldsymbol{\beta}_0) \right]$$

satisfies

$$\Pr\left[\inf_{(\delta_0, \boldsymbol{\beta}_0) \in C_{1-\alpha}(\delta, \boldsymbol{\beta})} g(\delta_0, \boldsymbol{\beta}_0) \leq g(\delta, \boldsymbol{\beta}) \leq \sup_{(\delta_0, \boldsymbol{\beta}_0) \in C_{1-\alpha}(\delta, \boldsymbol{\beta})} g(\delta_0, \boldsymbol{\beta}_0) \right] \geq 1-\alpha.$$

Hence, a marginal confidence interval of the form $[\hat{\delta}_L, \hat{\delta}_U]$ for δ in model (10) can be found as

$$\begin{aligned} \hat{\delta}_L = & \operatorname*{argmin}_{(\delta_0, \boldsymbol{\beta}_0) \in \mathbb{R} \times \mathbb{R}^K} \delta_0, & \hat{\delta}_U = & \operatorname*{argmax}_{(\delta_0, \boldsymbol{\beta}_0) \in \mathbb{R} \times \mathbb{R}^K} \delta_0, \\ & \text{subject to} \quad SP(\delta_0, \boldsymbol{\beta}_0) < c_\alpha, & & \text{subject to} \quad SP(\delta_0, \boldsymbol{\beta}_0) < c_\alpha. \end{aligned} \tag{14}$$

Once the solutions in (14) are found, the null hypothesis $H_0 : \mathbf{a} = \mathbf{0}$ is rejected at level α

if zero is not contained in $[\hat{\delta}_L, \hat{\delta}_U]$, otherwise there is not sufficient evidence to reject it at that level of significance.

Searching over the $\mathbb{R} \times \mathbb{R}^K$ domain in (14) is obviously not practical and some restrictions need to be imposed. Here we perform that step by specifying a fine grid of relevant points $\mathcal{B}(\hat{\delta}_0, \hat{\boldsymbol{\beta}}_0)$ around LAD point estimates $\hat{\delta}_0, \hat{\boldsymbol{\beta}}_0$ and calculating $SP(\delta_0, \boldsymbol{\beta}_0)$ at each of those points.[5] An important remark about computation is that a single pass over the grid is enough to establish both the joint confidence set and the limits of the marginal confidence interval for δ. Note also that the grid search can be stopped and the null hypothesis can no longer be rejected at the α level as soon as zero gets included in the marginal confidence interval for δ.

More sophisticated global optimization methods could be used to solve for the limits of the marginal confidence interval. For instance, Coudin and Dufour (2009) make use of a simulated annealing algorithm (Goffe, Ferrier, and Rogers 1994). 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. For high dimensional cases, a better appraoch would be to follow Chernozhukov, Hansen, and Jansson (2009) and use Markov chain Monte Carlo methods to generate an adaptive set of grid points that explores the relevant region of the parameter space more quickly than the conventional grid search.

[5] This problem is naturally parallelizable on a multi-core computer; i.e., the grid can be split into subgrids and those can be searched over in parallel across multiple cores.

3.4 Summary of test procedure

Suppose that one wishes to use the SP statistic in (13). In a preliminary step, the distribution of that statistic is simulated to the desired degree of accuracy and a one-sided α-level critical value, c_α, is determined. The rest of the test procedure then proceeds according to the following steps.

1. The estimates $\hat{a}_i$ of a_i, $i = 1, ..., N$, are computed using the first subsample of observations, r_{it} and $\mathbf{f}_t$, $t = 1, ..., T_1$.

2. For each $i = 1, ..., N$, weights $\hat{\omega}_i$ are computed according to:

$$\hat{\omega}_i = \frac{\hat{a}_i}{|\hat{a}_1| + ... + |\hat{a}_N|},$$

   and T_2 returns of a portfolio are computed as $y_t = \sum_i^N \hat{\omega}_i r_{it}$, $t = T_1 + 1, ..., T$.

3. For each candidate point $(\delta_0, \boldsymbol{\beta}_0) \in \mathcal{B}(\hat{\delta}_0, \hat{\boldsymbol{\beta}}_0)$, the statistic $SP(\delta_0, \boldsymbol{\beta}_0)$ is computed. The limits of the marginal confidence interval, $\hat{\delta}_L$ and $\hat{\delta}_U$, are then found as:

$$\hat{\delta}_L = \underset{(\delta_0,\boldsymbol{\beta}_0)\in\mathcal{B}(\hat{\delta}_0,\hat{\boldsymbol{\beta}}_0)}{\operatorname{argmin}} \delta_0, \quad \text{subject to} \quad SP(\delta_0, \boldsymbol{\beta}_0) < c_\alpha,$$

$$\hat{\delta}_U = \underset{(\delta_0,\boldsymbol{\beta}_0)\in\mathcal{B}(\hat{\delta}_0,\hat{\boldsymbol{\beta}}_0)}{\operatorname{argmax}} \delta_0, \quad \text{subject to} \quad SP(\delta_0, \boldsymbol{\beta}_0) < c_\alpha.$$

4. The null hypothesis $H_0 : \mathbf{a} = \mathbf{0}$ is rejected if $0 \notin [\hat{\delta}_L, \hat{\delta}_U]$, otherwise it is accepted.

This procedure yields a finite-sample and distribution-free test of H_0 over the class of all disturbance distributions satisfying Assumption 1.

4 Simulation evidence

We present the results of some small-scale 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 test in (7). The other benchmarks are the usual likelihood ratio (LR) test, an adjusted LR test, and a test based on GMM. The latter is a particularly important benchmark here, since in principle it is "robust" to non-normality and heteroskedasticity of returns.

The LR test is based on a comparison of the constrained and unconstrained log-likelihood functions evaluated at the maximum likelihood estimates. The unconstrained estimates are given in (4), (5), and (6). For the constrained case, the maximum likelihood estimates are

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

The LR test statistic, J_2, is then given by

$$J_2 = T \left[\log |\hat{\mathbf{\Sigma}}^{*}| - \log |\hat{\mathbf{\Sigma}}|\right],$$

which, under the null hypothesis, follows an asymptotic chi-square distribution with N degrees of freedom, χ^2_N. As we shall see, the finite sample behavior of J_2 can differ vastly from what asymptotic theory predicts. Jobson and Korkie (1982) suggest an adjustment to J_2 in order to improve its finite-sample size properties when used with critical values from the χ^2_N distribution. The adjusted statistic is

$$J_3 = \frac{T - (N/2) - K - 1}{T} J_2,$$

which also follows the asymptotic χ^2_N distribution, under H_0.

MacKinlay and Richardson (1991) develop tests of mean-variance efficiency in a GMM framework. For the asset pricing model in (1), the GMM tests are based on the moments of the following $(K+1)N \times 1$ vector:

$$\mathbf{g}_t(\boldsymbol{\theta}) = \begin{pmatrix} 1 \\ \mathbf{f}_t \end{pmatrix} \otimes \boldsymbol{\varepsilon}_t(\boldsymbol{\theta}), \tag{15}$$

where $\boldsymbol{\varepsilon}_t(\boldsymbol{\theta}) = \mathbf{r}_t - \mathbf{a} - \mathbf{B}\mathbf{f}_t$. The symbol $\otimes$ refers to the Kronecker product. Here $\boldsymbol{\theta} = (\mathbf{a}', vec(\mathbf{B})')'$, where $vec(\mathbf{B})$ is an $NK \times 1$ vector obtained by stacking the columns of $\mathbf{B}$, one below the other, with the columns ordered from left to right. The model specification in (1) implies the moment conditions $E(\mathbf{g}_t(\boldsymbol{\theta}_0)) = \mathbf{0}$, where $\boldsymbol{\theta}_0$ is the true parameter vector. The system in (15) is exactly identified which implies that the GMM procedure yields the same estimates of $\boldsymbol{\theta}$ as does OLS applied equation by equation. The covariance matrix of the GMM estimator $\hat{\boldsymbol{\theta}}$ is given by $\mathbf{V} = [\mathbf{D}_0'\mathbf{S}_0^{-1}\mathbf{D}_0]^{-1}$, where $\mathbf{D}_0 = E[\partial \mathbf{g}_T(\boldsymbol{\theta})/\partial \boldsymbol{\theta}']$ with $\mathbf{g}_T(\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T} \mathbf{g}_t(\boldsymbol{\theta})$ and $\mathbf{S}_0 = \sum_{s=-\infty}^{+\infty} E[\mathbf{g}_t(\boldsymbol{\theta})\mathbf{g}_{t-s}(\boldsymbol{\theta})']$; see Campbell, Lo, and MacKinlay (1997, Chapter 5). The GMM-based Wald test statistic is

$$J_4 = T\hat{\mathbf{a}}' \left[\mathbf{R}\left(\hat{\mathbf{D}}'\hat{\mathbf{S}}^{-1}\hat{\mathbf{D}}\right)^{-1}\mathbf{R}'\right]^{-1} \hat{\mathbf{a}}, \tag{16}$$

where $\hat{\mathbf{D}}$ and $\hat{\mathbf{S}}$ are consistent estimators of $\mathbf{D}_0$ and $\mathbf{S}_0$, respectively, and $\mathbf{R} = (1, \mathbf{0}_K) \otimes \mathbf{I}_N$, with $\mathbf{0}_K$ denoting a row vector of K zeros and $\mathbf{I}_N$ as the $N \times N$ identity matrix. Note that the J_4 statistic cannot be computed whenever $(K+1)N$ exceeds T, since $\hat{\mathbf{S}}$ then becomes singular.

Our implementation of the proposed test procedure is computationally intensive owing to the numerical grid search we perform in Step 3. This is not overly costly for a single application of the procedure, but it does become prohibitive for a simulation study. For that reason, we restrict our attention to cases with $K = 1$ in model (1). For convenience, the single-factor specification is given again here as

$$r_{it} = a_i + b_i f_t + \varepsilon_{it}, \quad t = 1, ..., T, \ i = 1, ..., N, \tag{17}$$

in which case the null hypothesis is a test of the mean-variance efficiency of the given portfolio. The returns of the reference portfolio, f_t, follow a stochastic volatility process:

$$f_t = \exp(h_t/2)\epsilon_t \text{ with } h_t = \lambda h_{t-1} + \xi_t,$$

where the independent terms ϵ_t and ξ_t are both i.i.d. according to a standard normal distribution and the persistence parameter λ is set to 0.5. The b_i's 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 $SX(\delta_0, \boldsymbol{\beta}_0)$ and $SP(\delta_0, \boldsymbol{\beta}_0)$ are determined using 10,000 simulations. 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 are found via LAD. Finally, there are 1000 replications in each experiment.

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. Table 1 shows the power of the test procedure applied with the SX and SP statistics for various values of T_1/T, where $|a_i| = 0.20$, 0.15, and 0.10. Here $T = 60$ and $N = 100$ and the disturbance terms ε_{it} are drawn randomly from the Student-t distribution with ν degrees of freedom. We consider $\nu = 12$ and 6 to examine the effects of kurtosis on the power of the tests. As expected the results show that for any given value of T_1/T, the power increases as $|a_i|$ increases and decreases as the kurtosis of the disturbance terms increases. Overall, the results 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, the testing strategy represented by $T_1 = 0.4T$ is pursued in the remaining comparative experiments.

We also include in our comparisons two distribution-free tests proposed by Gungor and Luger (2009) that are applicable in the single-factor context. The building block of those tests is

$$z_{it} = \left(\frac{r_{i,t+m}}{f_{t+m}} - \frac{r_{it}}{f_t} \right) \times \frac{(f_t - f_{t+m})}{f_t f_{t+m}}, \tag{18}$$

defined for $t = 1, ..., m$, where $m = T/2$ is assumed to be an integer. The first test is based on the sign statistic

$$S_i = \frac{\sum_{t=1}^{m} 0.5(s[z_{it}] + 1) - m/2}{\sqrt{m/4}} \tag{19}$$

and the second one is based the Wilcoxon signed rank statistic

$$W_i = \frac{\sum_{t=1}^{m} 0.5(s[z_{it}] + 1)\text{Rank}(|z_{it}|) - m(m+1)/4}{\sqrt{m(m+1)(2m+1)/24}}, \tag{20}$$

where Rank($|z_{it}|$) is the rank of $|z_{it}|$ when $|z_{i1}|, ..., |z_{im}|$ are placed in ascending order of magnitude. Gungor and Luger (2009) show that a time-series symmetry condition ensures that both (19) and (20) have limiting (as $m \to \infty$) standard normal distributions. Under the further assumption that the disturbance terms are cross-sectionally independent, conditional on $(f_1, ..., f_T)'$, their sum-type statistics

$$SD = \sum_{i=1}^{N} S_i^2 \text{ and } WD = \sum_{i=1}^{N} W_i^2 \tag{21}$$

follow an asymptotic chi-square distribution with N degrees of freedom. Simulation results show that this approximation works extremely well and just like the test procedure proposed here, the SD and WD test statistics can be calculated even if N is large.

Tables 2 and 3 show the empirical size (Panel A) and power (Panel B) of the considered tests when $|a_i| = 0.15$ and $T = 60, 120$ and $N = 10, 25, 50, 100, 125$. The power results for the J_1, J_2, J_3, and J_4 are based on size-corrected critical values, since none of those tests are exact under the two specifications we examine. It is important to emphasize that size-corrected tests are not feasible in practice, especially under the very general symmetry condition in Assumption 1. 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.

The results in Table 2 correspond to the single-factor model where the disturbance terms ε_{it} are i.i.d. in both the time-series and the cross-section according to a Student-t distribution with 6 degrees of freedom. From Panel A, we see that the parametric J_1 and the distribution-free SD and WD tests behave well under the null with empirical rejection rates close to the nominal level. This finding for the GRS test is in line with Affleck-Graves and McDonald (1989) who present simulation evidence showing the GRS test to be fairly robust to deviations from normality. From Table 2, the (conservative) SX and SP tests are also seen to satisfy the level constraint in the sense that the probability of a Type I error remains bounded by the nominal level of significance.[6] The J_2, J_3, and J_4 tests, however, suffer massive size distortions as the number of equations increases.[7] When $T = 120$ and $N = 100$, the LR test (J_2) rejects the true null with an empirical probability of 100% and in the case of the adjusted LR test (J_3) that probability is still above 50%. Notice as well that the J_1, J_2, and J_3 are not computable when N exceeds T, and the GMM-based J_4 cannot even be computed here as soon as $2N$ exceeds T. (Those cases are indicated with "-" in the tables.)

In Panel B of Table 2, 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 increases. 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, Chapter 5). On the contrary, the power of the SD and WD tests and that of the new SX and SP tests increases monotonically with N.

[6]Following the terminology in Lehmann and Romano (2005, Chapter 3), we say that a test of H_0 has *level* α if the probability of incorrectly rejecting H_0 when it is true is *not greater than* α.

[7]This overrejection problem with standard asymptotic tests in multivariate regression models is also documented in Stambaugh (1982), Jobson and Korkie (1982), Amsler and Schmidt (1985), MacKinlay (1987), Stewart (1997), and Dufour and Khalaf (2002).

The second specification we consider resembles a stochastic volatility model and introduces dependence between the conditional covariance matrix and f_t. Specifically, we let $\varepsilon_{it} = \exp(\lambda_i f_t/2)\eta_{it}$, where the innovations η_{it} are standard normal and the λ_i's are randomly drawn from a uniform distribution between 1.5 and 2.5. 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. Panel A of Table 3 reveals that all the parametric tests have massive size distortions in this case, and these over-rejections worsen as N increases for a given T.[8] When $T = 120$, 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. As in the first example, Panel B shows the power of the distribution-free tests increasing with both T and N in this heteroskedastic case.

At this point, one may wonder what is the advantage of the new SX and SP tests since the SD and WD tests of Gungor and Luger (2009) seem to display better power in Panel B of Tables 2 and 3. Those tests achieve higher power because they eliminate the b_i's from the inference problem through the long differences in (18), whereas the new tests proceed by finding set estimates of those nuisance parameters. A limitation of the SD and WD tests, however, is that they are valid only under the assumption that the model disturbances are cross-sectionally independent. Table 4 reports the empirical size

[8]The sensitivity of the GRS test to contemporaneous heteroskedasticity is also documented in MacKinlay and Richardson (1991), Zhou (1993), and Gungor and Luger (2009).

of the those tests when the cross-sectional disturbances are multivariate Student-t with an equicorrelation structure. Specifically, the disturbances have zero mean, unit variance, and the correlation between any two disturbances is equal to ρ, which we vary between 0.1 and 0.5. The degrees of freedom parameter is equal to 12 in Panel A and to 6 in Panel B. The nominal level is 0.05 and we consider $T = 60$, 120 and $N = 10$, 100. We see from Table 4 that the SD and WD tests are fairly robust to mild cross-sectional correlation, but start over-rejecting as the equicorrelation increases and this problem is further exacerbated when T increases and more so when N increases. The second limitation of the SD and WD tests is that they are designed for the single-factor model and cannot be easily extended to allow for multiple factors. The new SX and SP tests are illustrated next in the context of a single- and a three-factor model.

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. 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 returns 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 were no sampling uncertainty, finding that the market portfolio is mean-variance efficient would trivially imply the validity of the three-factor model. We test both specifications with two sets of test assets comprising the returns on 10 portfolios formed on size and 100 portfolios formed on both size and book-to-market. All the data we use are from Ken French's online data library.[9] They consist of monthly observations for the period covering January 1965 to December 2009 (540 months) and the one-month U.S. Treasury bill is considered the risk-free asset. Figure 2 plots the excess returns of the stock market index over the full sample period. From that figure we see that this representative return series contains several extreme observations. For instance, the returns seen during the stock market crash of October 1987, the financial crisis of 2008, and at some other points in time as well are obviously not representative of normal market activity; we discuss the effects of extreme observations at the end of this section. 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 45-year period, we also examine nine 5-year and four 10-year subperiods. Here the choice of subperiods follows that in Campbell, Lo, and MacKinlay (1997), Gungor and Luger (2009), and Ray, Savin, and Tiwari (2009).

[9] We chose these well-known and readily available datasets so our findings can easily be reproduced.

5.1 10 size portfolios

Table 5 reports the CAPM test results based on the ten size portfolios. Columns 2–5 show the results of the parameteric J tests and columns 6 and 7 show the results of the non-parametric SD and WD tests. The numbers reported in parenthesis are p-values and the numbers in square brackets in the last two columns show the limits of 95% confidence intervals for δ based on the SX and SP statistics. The entries in bold represent cases of significance at the 5% level.

From Table 5 we see that the parametric J tests reject the null hypothesis over the entire sample period with p-values no more than 1%. In contrast, the p-values for the non-parametric SD and WD tests, 82% and 60% respectively, and the 95% confidence intervals for the new procedure with the SX and SP statistics clearly support the mean-variance efficiency of the market index.

In three of the nine 5-year subperiods, 1/65–12/69, 1/90–12/94, and 1/00–12/04, the J tests reject the CAPM specification with p-values less than 8%. The results of the non-parametric SD and WD statistics are consistent with those of the J tests in subperiods 1/65–12/69 and 1/00–12/04, although, they also reject the null in the subperiods 1/70–12/74 and 1/75–12/79. The new SX and SP tests, however, continue to indicate a non-rejection of the mean-variance efficiency hypothesis, except for SP in the 1/75–12/79 subperiod. The results for the 10-year subperiods are more in line with the findings for the entire sample period; i.e., the J tests tend to reject the null more often than the non-parametric tests. Out of those four subsamples, the J tests all agree on a a rejection during the last three ones, while the SD and WD tests reject the null only in the second subperiod.

Besides the obvious differences between the parametric and non-parametric inference results, Table 5 also reveals some differences between the SD and WD tests and the proposed SX and SP tests. One possible reason for the disagreement across these non-parametric tests could be the presence of cross-sectional disturbance correlations. As we saw in Table 4, the SD and WD tests are not robust to such correlations, whereas the new tests allow for cross-sectional dependencies just like the GRS test.

Table 6 shows the results for the Fama-French model. The format is essentially the same as in Table 5, except that here the SD and WD tests are not feasible in this three-factor specification. For the entire 45-year sample period, the results in Table 6 are very similar to those for the single-factor model in Table 5. The standard J tests reject the null with very low p-values, whereas the distribution-free SX and SP tests are not significant at the 5% level. In the 5-year subperiods, there is much disagreement among the parametric tests. In seven of the nine subsamples, the GMM-based Wald test (J_4) rejects the null with p-values smaller than 6%. The usual J_2 LR test rejects the null during the subperiods 1/80–12/84, 1/90–12/94, and 1/95–12/99, whereas the adjusted LR test (J_3) tends to reject only during the 1/80–12/84 subperiod with a p-value of 6%. Note that, only in the subperiod 1/80–12/84, the GRS J_1 test tends to reject the null with a p-value of 7% and agrees with the non-parametric SX and SP on a non-rejection for all the other 5-year intervals. The results for the 10-year subperiods resemble those for the entire sample period and the J tests depict a more consistent picture. Over the last two of those subperiods, all the J tests indicate a clear rejection with p-values less than 3%, while SX and SP maintain a non-rejection. For the first two 10-year subperiods, however, the Fama-French model is supported by all the test procedures.

Table 6 shows that the SX and SP 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 5 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 and SP statistics, one may be concerned about the ability of the new procedure to reject the null, when the alternative is true. But the fact that the single-factor model is rejected in the subperiod 1/75-12/79 suggests that the overall pattern of non-rejections is not necessarily due to low power. In order to increase the probability of rejecting the null hypothesis if it is indeed false, we proceed next with a tenfold increase in the number of test assets.

5.2 100 size and book-to-market portfolios

Tables 7 and 8 show the test results for the two considered factor specifications using return data on 100 portfolios formed on size and book-to-market. Note that with $N = 100$, none of the parametric tests are computable in the 5-year subperiods where $T = 60$.

Using the entire sample, the J tests decisively reject both factor specifications. In the single-factor case (Table 7), the inference results based on the SD and WD tests are in agreement with the parametric ones. In sharp contrast, the 95% confidence intervals derived from the SX and SP tests continue to support the mean-variance efficiency of the market portfolio.

In the nine 5-year subperiods, the SD and WD tests indicate a rejection in six of them and the SP test rejects the null only in the 1/75-12/74 subperiod, as it did with the 10 size portfolios (see Table 5). As mentioned before, such a difference among the non-parametric procedures could be due to the presence of cross-sectional disturbance correlations and as Table 4 shows, the overrejections by the SD and WD tests are far worse in large cross-sections. Using the 10-year subperiod data, the J_1, J_2, and J_3 tests become feasible and along with the SD and WD tests they mostly reject the null with very low p-values. In the same subperiods, SX and SP indicate non-rejection.

Table 8 presents the results for the Fama-French model. In the full 45-year period, the non-parametric SP test agrees with the parametric ones on a rejection. A possible reason for these rejections is the presence of temporal instabilities. Indeed, an implicit assumption of all the tests used here is that the parameters associated with each test asset are constant over time. That assumption is obviously less likely to hold over longer time periods; see Ang and Chen (2007) for evidence of time-varying betas over the long run.

Over the shorter 5-year periods, the three-factor model finds clear support with the SX and SP tests. Those results and the 10-year ones resemble the findings from Tables

5 and 6 suggesting that the CAPM factor explains well the risk premiums of the 100 size and book-to-market portfolios. It is interesting to note that this conclusion about the validity of the CAPM is also reached by Zhou (1993), Vorkink (2003), and Ray, Savin, and Tiwari (2009).

5.3 Extreme observations

A common theme in Tables 5–8 is the striking difference between the parametric and non-parametric inference results. A 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 9 shows the results of the J tests with the 10 size portfolios when the full-sample returns are winsorized at the 0.2%, 0.4%, 0.5%, 0.6%, 0.8%, and 1% levels. In the single-factor case (Panel A), the J tests cease to be significant at the 5% level with returns winosoized beyond 0.5%. For the three-factor model (Panel B), the same pattern occurs but at even smaller 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 it is in fact 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 be lead 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 that assumes that the disturbance terms are identically distributed each period according to a multivariate normal distribution. Yet it is known that financial asset returns are non-normal, exhibiting time-varying variances and excess kurtosis. These stylized facts would put into question the reliability of any inference method that assumes that the cross-sectional distribution of disturbance terms is homogenous over time. Another important issue 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 diagonally symmetric each period. The class of diagonally 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 based on a split-sample technique that is applicable even in large cross-sections. In fact, the power of the new test procedure increases as either the time-series lengthens and/or the cross-section becomes larger. The inference procedure developed here thus offers a potentially very useful way to assess linear factor pricing models.

References

Abdelkhalek, T. and J.-M. Dufour. 1998. "Statistical inference for computable general equilibrium models, with application to a model of the Maroccan economy." *Review of Economics and Statistics* 80: 520–534.

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

Amsler, C.E. and P. Schmidt. 1985. "A Monte Carlo investigation of the accuracy of multivariate CAPM tests." *Journal of Financial Economics* 14: 359–375.

Ang, A. and J. Chen. 2007. "CAPM over the long run: 1926–2001." *Journal of Empirical Finance* 14: 1–40.

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." *Journal of Business and Economic Statistics* 25: 398-410.

Behboodian, J. 1990. "Some characterization theorems on symmetry." *Computational Statistics and Data Analysis* 10: 189–192.

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

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.

Breeden, D.T. 1979. “An intertemporal asset pricing model with stochastic consumption and investment opportunities.” *Journal of Financial Economics* 7: 265–296.

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

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

Chernozhukov, V., Hansen, C. and M. Jansson. 2009. “Finite sample inference for quantile regression models.” *Journal of Econometrics* 152: 93–103.

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

Connor, G. and R.A. Korajczyk. 2009. ”Factor models in portfolio and asset pricing theory.” In Handbook of Portfolio Construction: Contemporary Applications of

Markowitz Techniques. (J. Guerard, ed.), London: Springer.

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: 19–49.

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 M. Hallin. 1991. "Nonuniform bounds for nonparametric t tests." *Econometric Theory* 7: 253–263.

Dufour, J.-M. and J. Jasiak. 2001. "Finite sample limited information inference methods for structural equations and models with generated regressors." *International Economic Review* 42: 815–843.

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 M. Taamouti. 2005. "Projection-based statistical inference in linear structural models with possibly weak instruments." *Econometrica* 73: 1351–1365.

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

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.

Fama, E.F. and J.D. MacBeth. 1973. "Risk, return, and equilibrium: empirical tests." *Journal of Political Economy* 81: 607–636.

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.

Glahe, F.R. and J.F. Hunt. 1970. "The small sample properties of simultaneous equation least absolute estimators vis-a-vis least squares estimators." *Econometrica* 38: 742–753.

Goffe, W.L., Ferrier, G.D. and J. Rogers. 1994. "Global optimization of statistical functions with simulated annealing." *Journal of Econometrics* 60: 65–99.

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.

Harvey, C.R. and G. Zhou. 1990. "Bayesian inference in asset pricing tests." *Journal of Financial Economics* 26: 221–254.

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.

Hunt, J.G., Dowling, J.M. and F.R. Glahe. 1974. "L_1 estimation in small samples with Laplace error distributions." *Decision Sciences* 5: 22–29.

Jagannathan, R. and T. Ma. 2003. "Risk reduction in large portfolios: why imposing the wrong constraints helps." *Journal of Finance* 58: 1651–1684.

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.

Kandel, S., McCulloch, R. and R.F. Stambaugh. 1995. "Bayesian inference and portfolio efficiency." *Review of Financial Studies* 8: 1–53.

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

Ledoit, O. and M. Wolf. 2003. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection." *Journal of Empirical Finance* 10: 603–621.

Ledoit, O. and M. Wolf. 2004. "Honey, I shrunk the sample covariance matrix." *Journal of Portfolio Management* 30: 110–119.

Lehmann, E.L. and J.P. Romano. 2005. Testing Statistical Hypotheses, Third Edition. Springer, New York.

Lehmann, E.L. and C. Stein. 1949. "On the theory of some non-parametric hypotheses." *Annals of Mathematical Statisitcs* 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. 1987. "On multivariate tests of the Capital Asset Pricing Model." *Journal of Financial Economics* 18: 341–372.

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.

Michaud, R. 1989. "The Markowitz optimization enigma: is 'optimized' optimal?" *Financial Analysts Journal* 45: 31–42.

Mitra, A. 1987. "Minimum absolute deviation estimation in regression analysis." *Computers and Industrial Engineering* 12: 67–76.

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

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

Pfaffenberger, R.C. and J.J. Dinkel. 1978. "Absolute deviations curve fitting: an alternative to least squares." In Contributions to Survey Sampling and Applied Statistics (H.A. David, ed.) Academic Press, New York.

Pratt, J.W. and J.D. Gibbons. 1981. Concepts of Nonparametric Theory. Springer, New York.

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

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.

Rosenberg, B. and D. Carlson. 1977. "A simple approximation of the sampling distribution of least absolute residuals regression estimates." *Communications in Statistics - Simulation and Computation* 6: 421–437.

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

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

Serfling, R.J. 2006. "Multivariate symmetry and asymmetry." In Encyclopedia of Statistical Sciences, Second Edition (S. Kotz, N. Balakrishnan, C. B. Read and B. Vidakovic, eds.), Wiley.

Shanken, J. 1987. “A Bayesian approach to testing portfolio efficiency.” *Journal of Financial Economics* 19: 195–216.

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.

Stambaugh, R.F. 1982. “On the exclusion of assets from tests of the two-parameter model: a sensitivity analysis.” *Journal of Financial Economics* 10: 237–268.

Stewart, K.G. 1997. “Exact testing in multivariate regression.” *Econometric Reviews* 16: 321–352.

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 power comparisons for various sample splits

T_1/T	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Panel A: t(12) Distribution							
$\lvert a_i \rvert = 0.20$							
SX	85.9	91.9	89.2	84.7	76.3	50.7	10.9
SP	95.5	97.7	98.3	97.6	97.1	81.7	34.1
$\lvert a_i \rvert = 0.15$							
SX	37.5	46.9	49.3	43.3	35.7	19.1	3.2
SP	56.9	69.8	67.9	65.8	60.1	39.1	12.1
$\lvert a_i \rvert = 0.10$							
SX	6.6	7.2	9.2	7.8	7.1	3.2	0.8
SP	12.5	16.9	16.5	14.9	14.8	7.7	2.2
Panel B: t(6) Distribution							
$\lvert a_i \rvert = 0.20$							
SX	71.5	81.9	81.4	76.0	62.8	36.6	7.5
SP	87.1	93.5	95.5	94.4	86.9	68.3	25.4
$\lvert a_i \rvert = 0.15$							
SX	24.7	32.6	32.9	31.1	25.3	12.0	1.9
SP	41.4	53.7	52.3	54.8	44.1	27.5	9.6
$\lvert a_i \rvert = 0.10$							
SX	4.0	6.2	4.5	4.2	4.5	2.2	0.3
SP	8.1	12.7	9.6	8.9	8.7	5.1	1.6

Notes: This table reports the empirical power (in percentages) of the proposed test procedure based on the SX and SP statistics in (12) and (13) for various sample splits, T_1/T. The sample size is $T = 60$ and the number of test assets is $N = 100$. The returns are generated according to a single-factor model with i.i.d. disturbances following a Student-t distribution with degrees of freedom equal to 12 (Panel A) or 6 (Panel B). The notation $|a_i| = a$ means that $N/2$ pricing errors are set as $a_i = -a$ and the other half are set as $a_i = a$. The nominal level is 0.05 and the results are based on 1000 replications.

Table 2

Empirical size and power comparisons with homoskedastic disturbances

T	N	J_1	J_2	J_3	J_4	SD	WD	SX	SP
Panel A: Size									
60	10	4.6	9.5	4.7	7.9	5.5	4.2	0.2	0.9
	25	4.3	32.1	5.6	14.9	4.4	5.2	0.4	1.4
	50	6.0	98.7	41.2	-	5.2	4.1	0.7	1.1
	100	-	-	-	-	4.6	2.9	0.5	0.9
	125	-	-	-	-	4.5	4.2	1.0	2.0
120	10	3.8	5.8	3.8	5.2	3.8	4.2	0.1	1.2
	25	5.1	12.0	5.2	7.8	4.2	3.4	0.5	1.0
	50	5.5	45.0	7.5	13.9	5.2	4.8	0.6	1.5
	100	4.0	100.0	53.5	-	4.2	3.7	0.8	1.8
	125	-	-	-	-	6.1	5.0	0.7	1.7
Panel B: Size-corrected power									
60	10	41.5	41.5	41.5	41.9	16.7	18.4	1.7	4.5
	25	55.5	55.5	55.5	52.7	26.0	25.3	5.7	11.0
	50	24.1	24.1	24.1	-	40.8	43.4	14.2	29.3
	100	-	-	-	-	61.6	69.0	37.2	55.9
	125	-	-	-	-	67.3	75.2	44.5	64.0
120	10	83.1	83.1	83.1	83.8	32.6	39.0	11.5	22.1
	25	97.0	97.0	97.0	97.4	55.9	62.4	33.2	51.4
	50	99.6	99.6	99.6	99.5	79.4	86.6	66.2	83.4
	100	91.1	91.1	91.1	-	96.4	98.2	94.4	98.7
	125	-	-	-	-	98.7	99.1	98.4	99.7

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- and SP-based tests. The returns are generated according to a single-factor model with i.i.d. disturbances following a $t(6)$ distribution. The pricing errors are zero under H_0, whereas $N/2$ pricing errors are set equal to -0.15 and the other half are set to 0.15 under H_1. 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

Empirical size and power comparisons with contemporaneous heteroskedastic disturbances

T	N	J_1	J_2	J_3	J_4	SD	WD	SX	SP
Panel A: Size									
60	10	23.2	32.8	23.6	26.6	5.9	5.3	0.4	0.9
	25	46.6	81.6	50.5	62.8	5.5	4.3	0.3	1.1
	50	50.6	97.4	90.2	-	3.9	4.4	0.9	2.1
	100	-	-	-	-	4.5	3.7	1.2	2.3
	200	-	-	-	-	5.4	4.2	1.1	2.6
120	10	19.0	23.1	19.1	18.5	4.9	4.4	0.3	1.5
	25	37.8	54.6	38.3	45.5	4.2	5.0	0.6	1.8
	50	67.8	92.8	72.2	78.2	4.8	3.9	1.1	2.4
	100	73.7	96.6	94.0	-	5.7	5.2	1.8	2.2
	200	-	-	-	-	6.0	4.9	1.6	1.9
Panel B: Size-corrected power									
60	10	14.9	14.9	14.9	15.0	14.0	15.9	2.0	4.2
	25	27.2	27.2	27.2	25.1	20.6	26.0	4.0	6.3
	50	32.3	32.3	32.3	-	28.6	35.4	6.7	11.6
	100	-	-	-	-	49.4	59.3	15.4	21.7
	200	-	-	-	-	72.4	80.3	29.9	38.8
120	10	23.0	23.0	23.0	24.0	26.8	31.2	7.3	10.8
	25	47.8	47.8	47.8	45.7	43.7	51.7	17.6	23.2
	50	78.6	78.6	78.6	73.6	69.4	78.9	36.0	44.2
	100	76.3	76.3	76.3	-	91.0	95.7	63.9	71.2
	200	-	-	-	-	99.6	99.8	88.6	91.9

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- and SP-based tests. The returns are generated according to a single-factor model with contemporaneous heteroskedastic disturbances. The pricing errors are zero under H_0, whereas $N/2$ pricing errors are set equal to -0.15 and the other half are set to 0.15 under H_1. 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 4

Empirical size under cross-sectional disturbance equicorrelation structure

	$N = 10$					$N = 100$				
ρ	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
Panel A: t(12) Distribution										
$T = 60$										
SD	5.4	6.4	7.3	8.0	8.2	5.6	8.5	14.5	16.4	17.5
WD	6.5	7.2	7.8	9.3	9.9	5.9	10.8	14.5	17.0	18.6
$T = 120$										
SD	4.8	6.0	6.8	8.1	8.9	8.2	12.4	14.8	17.5	19.6
WD	5.4	6.4	7.6	9.1	10.3	7.2	12.8	16.2	18.3	20.2
Panel A: t(6) Distribution										
$T = 60$										
SD	4.2	5.2	5.8	6.2	7.9	7.0	10.2	14.7	16.9	18.9
WD	3.8	5.3	6.9	7.8	9.4	7.8	13.1	16.1	19.0	20.3
$T = 120$										
SD	4.4	5.0	7.0	7.8	7.7	6.1	9.7	13.0	15.5	18.1
WD	4.3	4.1	5.3	6.7	8.3	7.0	12.1	15.2	18.3	20.3

Notes: This table reports the empirical size of a sign test (SD) and a Wilcoxon signed rank test (WD) when the single-model cross-sectional disturbances are multivariate Student-t with mean zero and the correlation between any two disturbances is equal to ρ. The degrees of freedom parameter is equal to 12 in Panel A and to 6 in Panel B. The nominal level is 0.05 and entries are percentage rates. The results are based on 1000 replications

Table 5

Tests of the CAPM with 10 size portfolios

Time period	J_1	J_2	J_3	J_4	SD	WD	SX	SP
45-year period								
1/65–12/09	**2.30**	**22.94**	**22.64**	**23.04**	5.91	8.31	[-1.34, 1.44]	[-0.40, 0.16]
	(0.01)	(0.01)	(0.01)	(0.01)	(0.82)	(0.60)		
5-year subperiods								
1/65–12/69	1.80	**18.80**	16.61	**18.42**	**24.00**	**34.80**	[-4.79, 5.20]	[-0.48, 1.37]
	(0.08)	(0.04)	(0.08)	(0.05)	(0.01)	(0.00)		
1/70–12/74	1.45	15.52	13.71	14.27	**34.93**	**34.94**	[-3.72, 6.27]	[-0.25, 2.30]
	(0.19)	(0.11)	(0.19)	(0.16)	(0.00)	(0.00)		
1/75–12/79	1.30	14.09	12.45	**18.50**	33.07	58.75	[-3.86, 6.13]	**[0.28, 1.72]**
	(0.26)	(0.17)	(0.26)	(0.05)	(0.00)	(0.00)		
1/80–12/84	1.17	12.84	11.34	11.99	3.07	5.79	[-4.87, 5.12]	[-0.45, 0.78]
	(0.33)	(0.23)	(0.33)	(0.29)	(0.98)	(0.83)		
1/85–12/89	1.27	13.85	12.23	12.60	9.60	4.33	[-4.73, 5.26]	[-0.31, 1.18]
	(0.27)	(0.18)	(0.27)	(0.25)	(0.48)	(0.93)		
1/90–12/94	1.80	**18.80**	16.61	**18.28**	8.93	14.01	[-1.45, 0.66]	[-0.24, 0.45]
	(0.08)	(0.04)	(0.08)	(0.05)	(0.54)	(0.17)		
1/95–12/99	1.60	16.94	14.96	**20.20**	9.47	7.29	[-3.75, 6.24]	[-1.10, 2.09]
	(0.14)	(0.08)	(0.13)	(0.03)	(0.49)	(0.70)		
1/00–12/04	1.93	**19.93**	17.61	**19.46**	**27.60**	16.80	[-4.66, 5.33]	[-0.88, 2.18]
	(0.06)	(0.03)	(0.06)	(0.03)	(0.00)	(0.08)		
1/05–12/09	1.54	16.39	14.48	15.42	8.93	6.57	[-5.07, 4.92]	[-1.07, 1.33]
	(0.15)	(0.09)	(0.15)	(0.12)	(0.54)	(0.77)		
10-year subperiods								
1/65–12/74	1.26	13.17	12.40	12.50	6.73	3.34	[-5.80, 4.19]	[-1.89, 0.05]
	(0.26)	(0.21)	(0.26)	(0.25)	(0.75)	(0.97)		
1/75–12/84	1.81	**18.45**	17.37	**23.18**	**40.87**	**56.62**	[-4.69, 5.30]	[-0.19, 0.85]
	(0.07)	(0.05)	(0.07)	(0.01)	(0.00)	(0.00)		
1/85–12/94	**2.29**	**22.84**	**21.51**	**22.18**	7.20	3.85	[-2.70, 2.65]	[-0.61, 0.62]
	(0.02)	(0.01)	(0.02)	(0.01)	(0.71)	(0.95)		
1/95–12/04	**2.30**	**22.94**	**21.60**	**21.61**	10.00	7.61	[-6.11, 3.88]	[-2.29, 0.66]
	(0.02)	(0.01)	(0.02)	(0.02)	(0.44)	(0.67)		

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 one-month Treasury bill rate. Columns 2–5 report the results for the parametric J test; columns 6–7 report the results for the non-parametric SD and WD statistics. The numbers in parentheses are the p-values. The results for the newly proposed procedure, SX and SP, are reported in columns 8 and 9. The 95% marginal confidence intervals of the intercept estimates are in square brackets. Entries in bold represent cases of significance at the 5% level.

Table 6
Tests of the Fama-French model with 10 size portfolios

Time period	J_1	J_2	J_3	J_4	SX	SP
45-year period						
1/65–12/09	**2.44**	**24.43**	**24.02**	**24.86**	[-0.59, 0.69]	[-0.03, 0.21]
	(0.01)	(0.01)	(0.01)	(0.01)		
5-year subperiods						
1/65–12/69	1.03	11.87	10.09	12.18	[-1.00, 0.96]	[-0.48, 0.40]
	(0.44)	(0.29)	(0.43)	(0.27)		
1/70–12/74	1.39	15.59	13.25	**19.55**	[-0.88, 1.08]	[-0.32, 0.44]
	(0.21)	(0.11)	(0.21)	(0.03)		
1/75–12/79	0.41	5.08	4.31	6.09	[-1.04, 0.92]	[-0.60, 0.48]
	(0.93)	(0.89)	(0.93)	(0.81)		
1/80–12/84	1.92	**20.55**	17.47	**28.76**	[-0.71, 1.29]	[-0.35, 0.61]
	(0.07)	(0.02)	(0.06)	(0.00)		
1/85–12/89	1.51	16.75	14.24	**19.07**	[-0.73, 1.27]	[-0.13, 0.67]
	(0.16)	(0.08)	(0.16)	(0.04)		
1/90–12/94	1.70	**18.56**	15.78	**21.81**	[-0.86, 1.10]	[-0.34, 0.54]
	(0.11)	(0.05)	(0.11)	(0.02)		
1/95–12/99	1.70	**18.50**	15.72	**31.27**	[-0.70, 1.30]	[-0.14, 0.66]
	(0.11)	(0.05)	(0.11)	(0.00)		
1/00–12/04	1.29	14.58	12.39	17.72	[-1.15, 0.81]	[-0.91, 0.81]
	(0.26)	(0.15)	(0.26)	(0.06)		
1/05–12/09	1.52	16.82	4.29	**19.54**	[-1.16, 0.80]	[-0.44, 0.20]
	(0.16)	(0.08)	(0.16)	(0.03)		
10-year subperiods						
1/65–12/74	1.06	11.35	10.50	12.52	[-1.04, 0.92]	[-0.16, 0.20]
	(0.40)	(0.33)	(0.40)	(0.25)		
1/75–12/84	0.95	10.16	9.39	12.28	[-1.13, 0.83]	[-0.45, 0.19]
	(0.50)	(0.43)	(0.50)	(0.27)		
1/85–12/94	**2.56**	**25.71**	**23.79**	**26.66**	[-0.94, 1.02]	[-0.18, 0.34]
	(0.01)	(0.00)	(0.01)	(0.00)		
1/95–12/04	**2.16**	**22.07**	**20.42**	**25.30**	[-1.18, 0.78]	[-0.54, 0.30]
	(0.03)	(0.01)	(0.03)	(0.00)		

Notes: The results are based on value-weighted returns of 10 portfolios formed on size, the returns on three Fama-French factors, and the one-month Treasury bill rate as the risk-free rate. Columns 2–5 report the results for the parametric J statistics and the p-values in the paranthesis. The results for the newly proposed distribution-free tests, SX and SP, are reported in columns 6 and 7. The 95% marginal confidence intervals of the intercept estimates are in square brackets. Entries in bold represent cases of significance at the 5% level.

Table 7

Tests of the CAPM with 100 size and book-to-market portfolios

Time period	J_1	J_2	J_3	J_4	SD	WD	SX	SP
45-year period								
1/65–12/09	**2.74**	**262.02**	**236.79**	**277.65**	**260.95**	**314.09**	[-0.92, 1.38]	[-0.12, 0.35]
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		
5-year subperiods								
1/65–12/69	-	-	-	-	**143.33**	**186.92**	[-4.68, 5.31]	[-0.35, 1.36]
					(0.00)	(0.00)		
1/70–12/74	-	-	-	-	**152.80**	**148.81**	[-3.90, 6.09]	[-0.36, 1.56]
					(0.00)	(0.00)		
1/75–12/79	-	-	-	-	**227.47**	**355.90**	[-3.03, 6.96]	**[0.33, 6.96]**
					(0.00)	(0.00)		
1/80–1/84	-	-	-	-	96.40	96.91	[-4.59, 5.40]	[-0.01, 0.71]
					(0.58)	(0.57)		
1/85–12/89	-	-	-	-	**164.00**	**156.98**	[-4.58, 5.41]	[-0.09, 0.68]
					(0.00)	(0.00)		
1/90–12/94	-	-	-	-	**156.67**	**185.98**	[-3.88, 0.83]	[-0.85, 0.25]
					(0.00)	(0.00)		
1/95–12/99	-	-	-	-	70.93	85.86	[-4.30, 5.69]	[-1.00, 1.57]
					(0.99)	(0.84)		
1/00–12/04	-	-	-	-	**219.47**	**201.90**	[-4.16, 5.83]	[-0.19, 1.47]
					(0.00)	(0.00)		
1/05–12/09	-	-	-	-	89.07	80.24	[-5.09, 4.90]	[-0.50, 0.50]
					(0.78)	(0.93)		
10-year subperiods								
1/65–12/74	1.33	**249.48**	**141.37**	-	121.27	104.95	[-5.59, 4.40]	[-1.43, 0.09]
	(0.24)	(0.00)	(0.00)		(0.07)	(0.35)		
1/75–12/84	0.94	**214.31**	**121.44**	-	**263.00**	**345.82**	[-4.56, 5.43]	[-0.83, 1.06]
	(0.60)	(0.00)	(0.07)		(0.00)	(0.00)		
1/85–12/94	1.75	**278.58**	**157.86**	-	**193.20**	**194.08**	[-1.58, 1.78]	[-0.12, 0.60]
	(0.08)	(0.00)	(0.00)		(0.00)	(0.00)		
1/95–12/04	**2.57**	**321.22**	**182.03**	-	**186.07**	**212.82**	[-5.39, 4.60]	[-1.47, 0.53]
	(0.01)	(0.00)	(0.00)		(0.00)	(0.00)		

Notes: The results are based on value-weighted returns of 100 portfolios formed on size and book-to-market. The market portfolio is the value-weighted return on all NYSE, AMEX, and NASDAQ stocks and the risk-free rate is the one-month Treasury bill rate. Columns 2–5 report the results for the parametric J test; columns 6–7 report the results for the distribution-free tests of SD and WD. The number in parentheses are the p-values. The results for the newly proposed procedure, SX and SP, are reported in columns 8 and 9. The 95% marginal confidence intervals of the intercept estimates are in square brackets. The symbol "-" is used whenever a test is not computable and entries in bold represent cases of significance at the 5% level.

Table 8

Tests of the Fama-French model with 100 size and book-to-market portfolios

Time period	J_1	J_2	J_3	J_4	SX	SP
45-year period						
1/65–12/09	**2.48**	**243.09**	**218.78**	**327.08**	[-0.14, 0.74]	**[0.06, 0.22]**
	(0.00)	(0.00)	(0.00)	(0.00)		
5-year subperiods						
1/65–12/69	-	-	-	-	[-0.88, 1.08]	[-0.16, 0.48]
1/70–12/74	-	-	-	-	[-1.15, 0.81]	[-0.35, 0.29]
1/75–12/79	-	-	-	-	[-0.93, 5.07]	[-0.93, 5.08]
1/80–12/84	-	-	-	-	[-1.00, 0.96]	[-0.32, 0.40]
1/85–12/89	-	-	-	-	[-0.94, 1.02]	[-0.22, 0.58]
1/90–12/94	-	-	-	-	[-1.03, 0.93]	[-0.51, 0.29]
1/95–12/99	-	-	-	-	[-0.63, 1.37]	[-0.15, 1.02]
1/00–12/04	-	-	-	-	[-1.37, 2.63]	[-1.37, 2.31]
1/05–12/09	-	-	-	-	[-1.01, 0.95]	[-0.45, 0.63]
10-year subperiods						
1/65–12/74	1.78	**297.18**	**163.45**	-	[-0.95, 1.01]	[-0.07, 0.25]
	(0.09)	(0.00)	(0.00)			
1/75–12/84	0.79	**191.69**	105.43	-	[-1.92, 2.04]	[-1.08, 1.68]
	(0.76)	(0.00)	(0.34)			
1/85–12/94	1.67	**285.83**	**157.21**	-	[-0.85, 1.11]	[-0.05, 0.39]
	(0.11)	(0.00)	(0.00)			
1/95–12/04	**2.23**	**317.49**	**174.62**	-	[-0.84, 1.12]	[-0.40, 0.56]
	(0.03)	(0.00)	(0.00)			

Notes: The results are based on value-weighted returns of 100 portfolios formed on size and book-to-market, the returns on three Fama-French factors, and the one-month Treasury bill rate as the risk-free rate. Columns 2–5 report the results for the parametric J statistics and the p-values in the paranthesis. The results for the newly proposed distribution-free tests, SX and SP, are reported in columns 6 and 7. The 95% marginal confidence intervals of the intercept estimates are in square brackets. The symbol "-" is used whenever a test is not computable and entries in bold represent cases of significance at the 5% level.

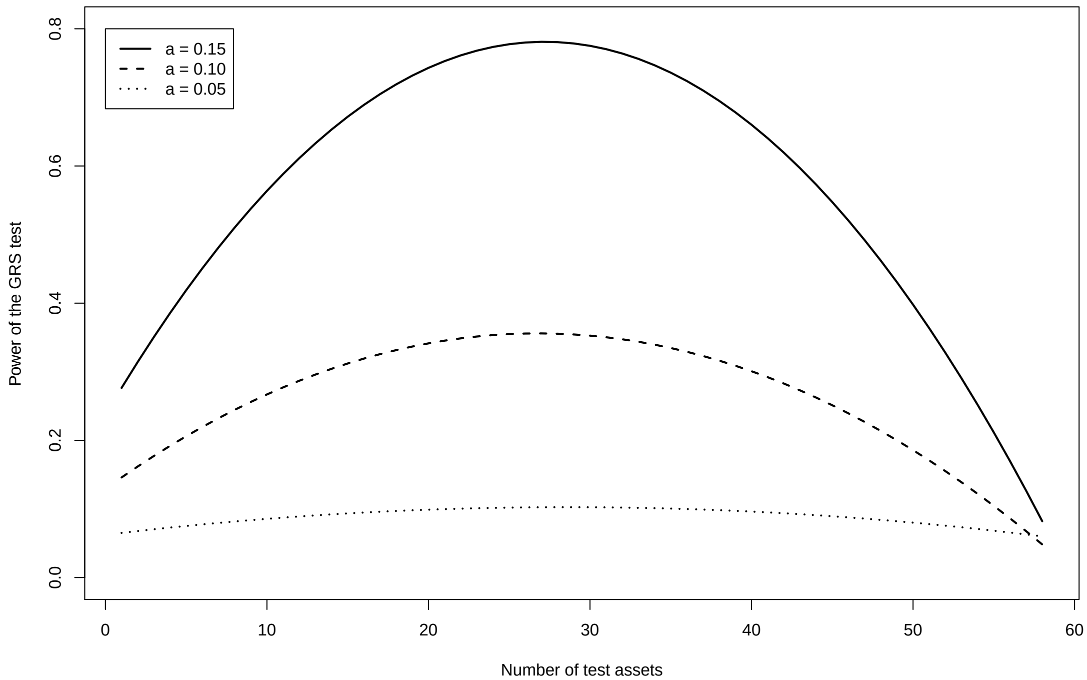
Table 9

Sensitivity of parametric tests to extreme observations

	0%	0.2%	0.4%	0.5%	0.6%	0.8%	1%
Panel A: CAPM							
J_1	**2.30**	**2.30**	**2.14**	**1.99**	1.71	1.37	1.53
	(0.01)	(0.01)	(0.02)	(0.03)	(0.07)	(0.19)	(0.12)
J_2	**22.94**	**22.96**	**21.41**	**20.00**	17.25	13.81	15.45
	(0.01)	(0.01)	(0.02)	(0.03)	(0.07)	(0.18)	(0.12)
J_3	**22.64**	**22.66**	**21.13**	**19.74**	17.02	13.63	15.25
	(0.01)	(0.01)	(0.02)	(0.03)	(0.07)	(0.19)	(0.12)
J_4	**23.04**	**23.00**	**21.22**	**19.69**	16.56	12.77	14.05
	(0.01)	(0.01)	(0.02)	(0.03)	(0.08)	(0.23)	(0.15)
Panel B: Fama-French model							
J_1	**2.44**	**2.40**	1.71	1.45	1.21	1.11	1.24
	(0.01)	(0.01)	(0.07)	(0.15)	(0.28)	(0.35)	(0.26)
J_2	**24.43**	**24.09**	17.30	14.72	12.26	11.22	12.62
	(0.01)	(0.01)	(0.07)	(0.14)	(0.26)	(0.33)	(0.24)
J_3	**24.02**	**23.68**	17.01	14.47	12.05	11.04	12.41
	(0.01)	(0.01)	(0.07)	(0.15)	(0.28)	(0.35)	(0.26)
J_4	**24.86**	**24.62**	**18.00**	15.40	12.60	11.70	13.69
	(0.01)	(0.01)	(0.05)	(0.12)	(0.25)	(0.31)	(0.30)

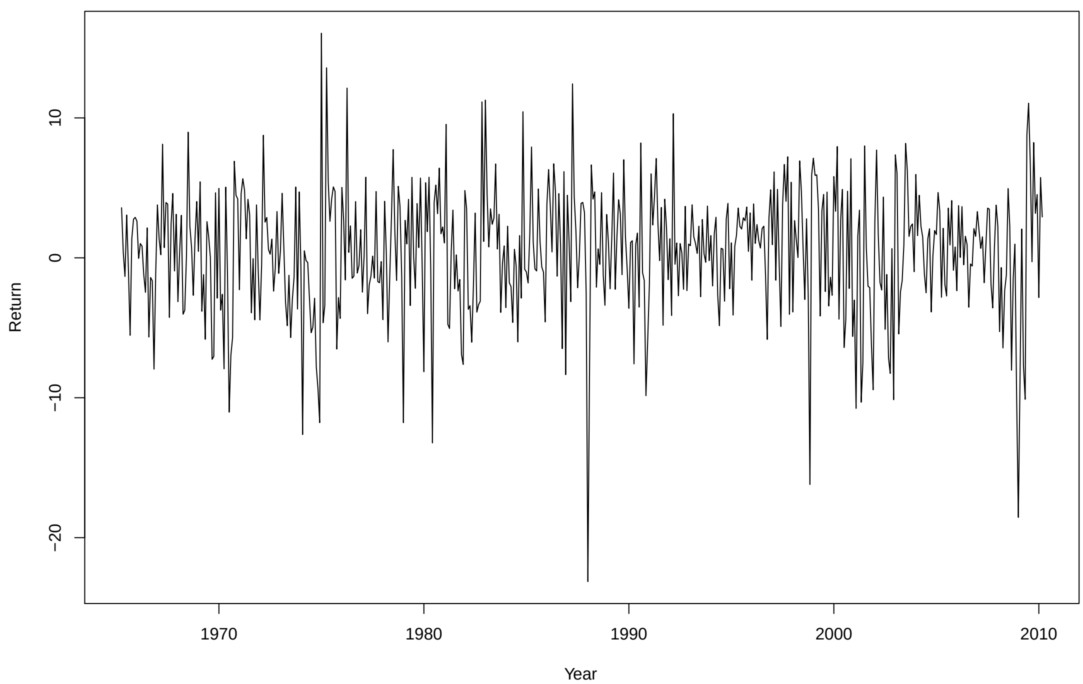
Notes: This table shows the results of the parametric tests with the 10 size portfolios when the returns for the full sample period from Janaury 1965 to December 2009 are winsorized at various small levels. Panels A and B correspond to the single- and three-factor models, respectively. The numbers in parenthesis are p-values and bold entries represent cases of significance at the 5% level.

Figure 1



Notes: 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 single-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.

Figure 2



Notes: This figure plots the monthly excess returns (in percentage) of a value-weighted stock market index of all stocks listed on the NYSE, AMEX, and NASDAQ for the period covering January 1965 to December 2009.