Statistical Inference on Stochastic Dominance Efficiency. Do Omitted Risk Factors Explain the Size and Book-to-Market Effects?

Thierry Post

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2003-017-F&A
Publication status / version	March 2003
Number of pages	21
Email address corresponding author	gtpost@few.eur.nl
Address	Erasmus Research Institute of Management (ERIM) Rotterdam School of Management / Faculteit Bedrijfskunde Rotterdam School of Economics / Faculteit Economische Wetenschappen Erasmus Universiteit Rotterdam PoBox 1738 3000 DR Rotterdam, The Netherlands Phone: # 31-(0) 10-408 1182 Fax: # 31-(0) 10-408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website:
www.erim.eur.nl

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

REPORT SERIES
RESEARCH IN MANAGEMENT

BIBLIOGRAPHIC DATA AND CLASSIFICATIONS		
Abstract	This paper discusses statistical inference on the second-order stochastic dominance (SSD) efficiency of a given portfolio relative to all portfolios formed from a set of assets. We derive the asymptotic sampling distribution of the Post test statistic for SSD efficiency. Unfortunately, a test procedure based on this distribution involves low power in small samples. Bootstrapping is a more powerful approach to sampling error. We use the bootstrap to test if the Fama and French value-weighted market portfolio is SSD efficient relative to benchmark portfolios formed on market capitalization and book-tomarket equity ratio. During the late 1970s and during the 1980s, the market portfolio is significantly SSD inefficient, even if we use samples of only 60 monthly observations. This suggests that the size and book-to-market effects cannot be explained by omitted risk factors like higher-order central moments or lower partial moments.	
Library of Congress Classification (LCC)	5001-6182	Business
	5601-5689	Accountancy, Bookkeeping
	4001-4280.7	Finance Management, Business Finance, Corporation Finance
	HG 4027	Financial management: Statitstics
Journal of Economic Literature (JEL)	M	Business Administration and Business Economics
	M 41	Accounting
	G 3	Corporate Finance and Governance
	C 19	Statistical methods: other
	G 12	Asset pricing
European Business Schools Library Group (EBSLG)	85 A	Business General
	225 A	Accounting General
	220 A	Financial Management
	220 T	Quantitative methods for financial management
Gemeenschappelijke Onderwerpsontsluiting (GOO)		
Classification GOO	85.00	Bedrijfskunde, Organisatiekunde: algemeen
	85.25	Accounting
	85.30	Financieel management, financiering
	85.30	Financieel management, financiering
	31.73	Wiskundige statistiek
Keywords GOO	Bedrijfskunde / Bedrijfseconomie	
	Accountancy, financieel management, bedrijfsfinanciering, besliskunde	
	Effectenhandel, portfolio-analyse, stochastsiche methoden, bootstrap (statistiek)	
Free keywords	stochastic dominance, market efficiency, asset pricing, statistical inference, size and book-to-market effects	

Statistical Inference on Stochastic Dominance Efficiency

Do Omitted Risk Factors Explain the Size and Book-to-Market Effects?

THIERRY POST*

*Post is corresponding author: Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands, email: gtpost@few.eur.nl, tel: +31-104081428. This study forms part of a research program on stochastic dominance. Details on the program are available at the program homepage: *http://www.few.eur.nl/few/people/gtpost/stochastic_dominance.htm*. We appreciate the comments by Roger Bowden, Winfried Hallerbach, Haim Levy, Jan van der Meulen, Nico van der Sar and Jaap Spronk, as well as participants at the 30th EURO Working Group on Financial Modeling, the 29th annual meeting of the European Finance Association, and seminars of Center for Economic Studies (Catholic University of Leuven), Erasmus Center of Financial Research (Erasmus University Rotterdam) and Ghent Finance Center (Ghent University). Martijn van den Assem, Peter van der Spek and Pim van Vliet provided invaluable research assistance. Financial support by Tinbergen Institute, Erasmus Research Institute of Management and Erasmus Center of Financial Research is gratefully acknowledged. Any remaining errors are our own.

Statistical Inference on Stochastic Dominance Efficiency

Do Omitted Risk Factors Explain the Size and Book-to-Market Effects?

This paper discusses statistical inference on the second-order stochastic dominance (SSD) efficiency of a given portfolio relative to all portfolios formed from a set of assets. We derive the asymptotic sampling distribution of the Post test statistic for SSD efficiency. Unfortunately, a test procedure based on this distribution involves low power in small samples. Bootstrapping is a more powerful approach to sampling error. We use the bootstrap to test if the Fama and French value-weighted market portfolio is SSD efficient relative to benchmark portfolios formed on market capitalization and book-to-market equity ratio. During the late 1970s and during the 1980s, the market portfolio is significantly SSD inefficient, even if we use samples of only 60 monthly observations. This suggests that the size and book-to-market effects cannot be explained by omitted risk factors like higher-order central moments or lower partial moments.

STOCHASTIC DOMINANCE (SD) rules can analyze economic behavior under uncertainty without a parametric specification of the preferences of the decision-makers and the statistical distribution of the choice alternatives. The popular criterion of second-order SD (SSD) assumes only non-satiation and risk aversion for the preferences and it imposes minimal distribution assumptions. In financial economics, the problem of portfolio selection and portfolio evaluation is a potential application area for SD, because economic theory does not forward strong predictions to select a functional form for investor preferences and the asset distribution. In addition, large, high-quality data sets of asset returns are available (e.g., CRSP and Datastream data sets), so that SD rules can let the data 'speak for them selves'. Still, SD has seen minimal application for this problem. Rather, the focus has predominantly been on mean-variance analysis. Computational and statistical considerations may help to explain this situation.

For applying SD to empirical data, simple crossing algorithms have been developed that check in a pairwise fashion the difference of the empirical distribution functions (EDFs) of the choice alternatives (e.g. Levy (1992), App. A). Unfortunately, these algorithms are unable to deal with cases that involve infinitely many choice alternatives, such as the case where investors can fully diversify between assets. To circumvent this problem, Post (2003) derived a tractable linear programming test for SSD efficiency of a given portfolio relative to all portfolios formed from a set of assets.

Another practical complication is sampling error. SD is based on the full EDF rather than a finite set of sample statistics. In many cases, the EDF is a statistically consistent estimator for the true cumulative distribution function (CDF). However, in small samples, the EDF generally is very sensitive to sampling variation, which causes serious doubt about the reliability of SD applications that rely in a naïve way on the EDF without accounting for sampling error (see, e.g., Kroll and Levy (1980) and Post (2003), section III). Several steps have been made towards analyzing the sampling properties of SD tests. Still, some important questions remain unanswered. Beach and Davidson (1983), Dardanoni and Forcina (1999) and Davidson and Duclos (2000), among others, derive an analytical characterization of the asymptotic sampling distribution. Unfortunately, this literature deals with the pairwise comparison of a finite number of choice alternatives, and it is not immediately clear how to generalize the existing results to the case with full diversification possibilities. Post derived the asymptotic sampling distribution of his SSD test statistic for the special case where all assets have the same mean. Unfortunately, this special case is very unrealistic and the resulting distribution may lead to erroneous conclusions. Bootstrapping is another approach to sampling error. The powerful computer hardware and software currently available substantially reduces the computational burden associated with this approach. Still, the statistical goodness of the bootstrap for SSD efficiency has not been analyzed thus far.

The purpose of this paper is to fill the above gaps for statistical inference on SD efficiency. Section I introduces the notation, definitions and assumptions that will be used throughout the text. Section II derives the asymptotic sampling distribution for the Post test statistic under the true null of efficiency rather than the null of equal means. Subsequently, Section III analyzes the statistical size and power properties of a test procedure that uses this sampling distribution, as well as those of a bootstrap procedure. Next, Section IV presents an empirical application that tests if the Fama and French value-weighted market portfolio is SSD efficient relative to benchmark portfolios formed on market capitalization (size) and book-to-market equity ratio (BE/ME). Since SSD effectively considers the entire return distribution, the test results can help to determine if the size and BE/ME effects can be explained by the omission of risk factors. Finally, Section V summarizes our conclusions and presents directions for further research.

I. Preliminaries

We consider a single-period, portfolio-based model of investment that satisfies the following three assumptions:

1. Investors are nonsatiable and risk averse and they select investment portfolios to maximize the expected utility associated with the return of their investment portfolio. Throughout the text, we will denote utility functions by $u:\Re \rightarrow P$, $u \in U_2$, with U_2 for the class of strictly increasing and concave, continuously differentiable, von Neumann-Morgenstern utility functions, and P for a nonempty, closed, and convex subset of $\Re$.[1,2]

2. The investment universe consists of N assets, associated with returns $\boldsymbol{x} \in \Re^N$. Throughout the text, we will use the index set $\mathrm{I} \equiv \{1,\cdots,N\}$ to denote the different assets. The returns are serially independent and identically distributed (IID) random variables with a continuous joint cumulative distribution function (CDF) $G:\Re^N \rightarrow [0,1]$.

3. Investors may diversify between the assets, and we will use $\boldsymbol{l} \in \Re^N$ for a vector of portfolio weights. We consider the case where short sales are not allowed, and the portfolio weights belong to the portfolio possibilities set $\Lambda \equiv \{\boldsymbol{l} \in \Re^N_+ : \boldsymbol{e}^{\mathrm{T}} \boldsymbol{l} = 1\}$, with $\boldsymbol{e}$ for a unity vector with dimensions conforming to the rules of matrix algebra.[3]

Under these assumptions, the investors' optimization problem can be summarized as $\max_{l \in \Lambda} \int u(\boldsymbol{x}^{\mathrm{T}} l) dG(\boldsymbol{x})$. Post's (2003) test statistic is based on the first-order condition for this problem. Specifically, a given portfolio $t \in \Lambda$ is optimal for a given utility function $u \in U_2$ if and only if

$$\int u'(\boldsymbol{x}^{\mathrm{T}} t)(\boldsymbol{x}^{\mathrm{T}} t - x_i) dG(\boldsymbol{x}) \leq 0 \quad \forall i \in \mathrm{I}. \tag{1}$$

[4]

This naturally leads to the following measure for SSD efficiency:

$$x(t, G) \equiv \min_{u \in U_2} \left\{ \max_{i \in \mathrm{I}} \left\{ \int u'(\boldsymbol{x}^{\mathrm{T}} t)(\boldsymbol{x}^{\mathrm{T}} t - x_i) dG(\boldsymbol{x}) \right\} \right\}. \tag{2}$$

Definition 1 *Portfolio $t \in \Lambda$ is SSD efficient if and only if it is optimal for at least some $u \in U_2$, i.e., $x(t, G) = 0$.*

In practical applications, $G(\boldsymbol{x})$ generally is not known and information is limited to a discrete set of T time series observations. We assume that observations are independent random draws from the CDF. Throughout the text, we will represent the observations by the matrix $\mathrm{C} \equiv (\boldsymbol{x}_1 \cdots \boldsymbol{x}_T)$, with $\boldsymbol{x}_t \equiv (x_{1t} \cdots x_{Nt})^{\mathrm{T}}$. Since the timing of the draws is inconsequential, we are free to label the observations by their ranking with respect to the evaluated portfolio, i.e., $\boldsymbol{x}_1^{\mathrm{T}} t < \boldsymbol{x}_2^{\mathrm{T}} t < \cdots < \boldsymbol{x}_T^{\mathrm{T}} t$. Using the observations, we can construct the empirical distribution function (EDF):

$$F_{\mathrm{C}}(\boldsymbol{x}) \equiv \mathrm{card}\left\{ t \in \{1, \cdots, T\} : \boldsymbol{x}_t \leq \boldsymbol{x} \right\} / T, \tag{3}$$

with $\mathrm{card}\{\cdot\}$ for the number of elements of a set. Since the observations are serially IID distributed, $F_{\mathrm{C}}(\boldsymbol{x})$ is a consistent estimator for $G(\boldsymbol{x})$.

Our objective is to test the null hypothesis that a given portfolio $t \in \Lambda$ is SSD efficient, i.e., $H_0 : x(t, G) = 0$. For simplicity, we assume that the optimal utility function $u_G^* \equiv \arg\min_{u \in U_2} \left\{ \max_{i \in \mathrm{I}} \left\{ \int u'(\boldsymbol{x}^{\mathrm{T}} t)(\boldsymbol{x}^{\mathrm{T}} t - x_i) dG(\boldsymbol{x}) \right\} \right\}$ is unique. To accomplish this, we standardize utility such that $u(0) = 0$ and $u'(\boldsymbol{x}_T^{\mathrm{T}} t) = 1$.[5]

Post (2003) proposes $x(t, F_{\mathrm{C}})$ as an estimator for $x(t, G)$, and he derives the following linear programming formulation:

$$x(t, F_{\mathrm{C}}) = \min_{b \in \mathrm{B}, q} \left\{ q : \sum_{t=1}^{T} b_t (\boldsymbol{x}_t^{\mathrm{T}} t - x_{it}) / T + q \geq 0 \quad \forall i \in \mathrm{I} \right\}, \tag{4}$$

with $\mathrm{B} \equiv \left\{ b \in \Re_+^T : b_1 \geq b_2 \geq \cdots \geq b_T = 1 \right\}$. The optimal solution $b \in \mathrm{B}$, say b^*, represents the gradient vector $\nabla u_{F_{\mathrm{C}}}^*(t) \equiv (u'_{F_{\mathrm{C}}}(\boldsymbol{x}_1^{\mathrm{T}} t) \cdots u'_{F_{\mathrm{C}}}(\boldsymbol{x}_T^{\mathrm{T}} t))^T$ for the optimal utility function u_F^*. B represents the restrictions on the gradient vector that follow from the assumptions of nonsatiation and risk aversion and the standardization $u'(\boldsymbol{x}_T^{\mathrm{T}} t) = 1$.

II. Asymptotic sampling distribution

Since $\xi(\boldsymbol{\tau}, F_C)$ is a consistent estimator for $G(\boldsymbol{x})$ (see Section I), $\xi(\boldsymbol{\tau}, F_C)$ is a consistent estimator for $\xi(\boldsymbol{\tau}, G)$. However, $F_C(\boldsymbol{x})$ generally is very sensitive to sampling variation and the test results are likely to be affected by sampling error in a nontrivial way. The applied researcher must therefore have knowledge of the sampling distribution in order to make inferences about the true efficiency classification. Post (2003) derived the asymptotic sampling distribution of $\xi(\boldsymbol{\tau}, F_C)$ under the simplifying assumptions that all assets have the same mean, i.e., $H_1: E[\boldsymbol{x}] = \mu\boldsymbol{e}$, $\mu \in \Re$. H_1 gives a sufficient condition for the true null of efficiency, i.e., H_0. In fact, under the null, all portfolios $\lambda \in \Lambda$ are efficient, because they maximize the expected value of the utility function $u(x) = x$, i.e., the risk neutral investor is indifferent between portfolios with equal means. However, H_1 does not give a necessary condition for H_0, and rejection of H_1 generally does not imply rejection of H_0 and there is no guarantee that H_1 is sufficiently close to H_0. Hence, the sampling distribution H_1 may lead to erroneous conclusions for H_0.

Using $\mathbf{C} \equiv (\mathrm{I} - \boldsymbol{e}\boldsymbol{\tau}^{\mathrm{T}})$ and variance-covariance matrix $\Omega_0 \equiv \int (u_G'^{*}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\tau})^2 \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) dG(\boldsymbol{x})$, the following theorem summarizes the asymptotic sampling distribution under the true null:

THEOREM 1 *The p-value* $\Pr[\xi(\boldsymbol{\tau}, F_C) > y | H_0]$, $y \geq 0$, *asymptotically equals the integral* $\Gamma(y, \Sigma_0) \equiv (1 - \int_{\boldsymbol{z} \leq y\boldsymbol{e}} d\Phi(\boldsymbol{z} | \Sigma_0, \mathbf{0}))$, *with* $\Phi(\boldsymbol{z} | \Sigma_0, \mathbf{0})$ *for a N-dimensional multivariate normal distribution function with mean* $\mathbf{0}$ *and (singular) variance-covariance matrix* $\Sigma_0 \equiv (\mathbf{C}\Omega_0\mathbf{C}^{\mathrm{T}})/T$.

We may use this theorem by comparing the *p*-value for the observed value of $\xi(\boldsymbol{\tau}, F_C)$ with a predefined level of significance $a \in [0,1]$; we may reject efficiency if $\Gamma(\xi(\boldsymbol{\tau}, F_C), \Sigma_0) \leq a$. Alternatively, we may reject efficiency if the observed value of $\xi(\boldsymbol{\tau}, F_C)$ is greater than or equal to the critical value $\Gamma^{-1}(a, \Sigma_0)$ $\equiv \inf_{y \geq 0} \{y : \Gamma(y, \Sigma_0) \leq a\}$.

Computing *p*-values and critical values requires the unknown variance-covariance matrix Ω_0. We may estimate its elements $\omega_{0,ij}$, $i, j \in \mathrm{I}$, in a distribution-free and consistent manner using the sample equivalents

$$\hat{\omega}_{0,ij} \equiv \sum_{t=1}^{T} (\beta_t^{*2} (x_{it} - \sum_{t=1}^{T} x_{it}/T)(x_{jt} - \sum_{t=1}^{T} x_{jt}/T))/T. \tag{5}$$

The theorem subtly differs from Post's (2003) characterization of the sampling distribution under H_1. That characterization used the variance-covariance matrix $\Omega_1 \equiv \int \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} dG(\boldsymbol{x})$ in place of Ω_0. These two matrices are identical if the optimal utility function exhibits risk neutrality, i.e., $u_G^{*}(x) = x$ and hence $u_G'^{*}(x) = 1$. This reflects the replacement of the null of efficiency (H_0) by the null of equal means (H_1); under H_1, all portfolios are optimal for the risk neutral investor. Obviously, it is relatively simple to reject this null and hence the *p*-values and critical values under H_1 are likely to underestimate the true values under H_0.

III. Simulation experiment

This section analyzes the statistical size and power properties of various test procedures for SSD efficiency, including one that uses the asymptotic sampling distribution under the null of efficiency (H_0), as well as one that uses the bootstrap. For this purpose, we extend the simulation experiment of Post (2003, Section IIIC). The simulations involve 26 benchmark assets with a multivariate normal return distribution. The joint population moments are equal to the sample moments of the monthly excess returns of the one-month U.S. Treasury bill and the 25 Fama and French U.S. stock portfolios formed on market capitalization (size) and book-to-market equity ratio (BE/ME) during the sample period from July 1963 to October 2001. To provide some feeling for the data, Figure 1 shows a mean-variance diagram including the individual assets (the clear dots), as well as the mean-variance frontier for the case without the riskless asset (AB) and the case with the riskless asset (OP1B). The figure also includes the tangency portfolio (P1) and the equal weighted average of all 25 risky assets (P2). The tangency portfolio is efficient and we may analyze the size (=the relative frequency of Type I error) of a test procedure by the relative frequency of cases in which this portfolio is wrongly classified as inefficient. By contrast, the equal weighted portfolio is inefficient; it is possible to achieve a substantially higher mean given the standard deviation, and to achieve a substantially lower standard deviation given the mean. Hence, we may analyze the power (=one minus the relative frequency of Type II error) of a test procedure by its ability to correctly classify the equal weighted portfolio as inefficient; a Type II error occurs if the equal weighted portfolio is wrongly classified as efficient.

[Insert Figure 1 about here]

We assess the size and power of the following three alternative test procedures:

1. Procedure A uses Post's (2003) asymptotic sampling distribution under the null of equal means (H_1). Specifically, it rejects efficiency if and only if $x(t, F_C) \geq \Gamma^{-1}(a, \hat{S}_1)$.[6]

2. Procedure B uses the asymptotic sampling distribution under the null of efficiency, (H_0), as given in Theorem 1. Specifically, it rejects efficiency if and only if $x(t, F_C) \geq \Gamma^{-1}(a, \hat{S}_0)$.

3. Procedure C is a bootstrap procedure. Key to the success of the bootstrap is the selection of an appropriate approximation for the CDF. If the approximation is statistically consistent, then the bootstrap distribution gives a statistically consistent estimator for the original sampling distribution. Under the assumption that the return distribution is serially IID (see Section I), the EDF $F_C(\boldsymbol{x})$ is a

consistent estimator for the CDF $G(x)$. This suggests that bootstrap pseudosamples would be simply obtained by randomly sampling with replacement from the EDF. We generate 1,000 pseudo-samples $\hat{C}$ in this way and compute the test statistic $x(t, F_{\hat{C}})$ for each pseudo-sample. Subsequently, we compute the bootstrap *p*-value as the relative frequency of pseudo-samples in which the evaluated portfolio is classified as efficient, i.e., $x(t, F_{\hat{C}}) = 0$. Finally, we reject efficiency if and only if the bootstrap *p*-value falls below the level of significance (a).[7]

To assess the size and power of these procedures, we draw 1,000 random samples from the multivariate normal population distribution through Monte-Carlo simulation. For each random sample, we apply each of above three test procedures to the efficient tangency portfolio (P1) and the inefficient equal weighted portfolio (P2). For each procedure, we compute the size as the rejection rate for P1 and the power as the rejection rate for P2. This experiment is performed for a sample size (T) of 25 to 4,000 observations and for a significance level (a) of 2.5, 5, and 10 percent.

Figure 2 and Figure 3 show the size and power of the three test procedures. Again, $x(t, F_C)$ converges to $x(t, G)$, and we expect minimal Type I and Type II error in large samples. Indeed, for all procedures the size goes to zero and the power goes to unity as we increase the sample size. However, in small samples, there are substantial differences in size and power. For both asymptotic test procedures (Procedure A and B), the size is much lower than the nominal significance level. Presumably, this reflects the conservative nature of tests that are based on the least favorable distribution, i.e., that minimize Type I error (see the proof to Theorem 1). By contrast, the size of the bootstrap procedure in small samples is more comparable with the nominal level of significance (a). For the asymptotic sampling distribution, minimizing Type I errors comes at the cost of Type II errors, and we need large samples to obtain reasonable power. For example, using a ten percent significance level, Procedure A involves reasonable power (a rejection rate of about 60 percent) only for samples of at least 500 observations. Since H_0 is more general than H_1, Procedure B results in fewer rejections and hence less power than Procedure A. For example, using the ten percent significance level, Procedure B achieves a rejection rate of 60 percent only for samples of about 1400 observations. By contrast, the bootstrap involves substantially more power. For example, using the ten percent significance level, Procedure C yields a 60 percent rejection rate already for samples of 25 observations. Of course, this benefit has to be balanced against the additional computational burden associated with bootstrapping. However, this is not a major issue given the powerful computer hardware and software currently available.

[Insert Figure 2 about here]

[Insert Figure 3 about here]

IV. Size and BE/ME effects

The traditional mean-variance capital asset pricing model (CAPM) fares poorly in explaining observed stock returns. For example, the value-weighted market portfolio of risky assets seems highly mean-variance inefficient relative to portfolios formed on size and BE/ME. Related to this, market beta seems to explain only a small portion of the cross-sectional variation in average returns, while size and BE/ME appear to have substantial explanatory power (see, e.g., Fama and French (1992)).

One way to extend the mean-variance CAPM is by changing the maintained assumptions on investor preferences. If we do not restrict the shape of the return distribution, then mean-variance CAPM is consistent with expected utility theory only if utility takes a quadratic form.[8] Extensions can be obtained by using alternative classes of utility. For example, the three-moment CAPM, used by, e.g., Kraus and Litzenberger (1976) and Harvey and Siddique (2000), assumes a cubic utility function, which implies that investors care about the first three central moments of the return distribution (mean, variance and skewness). While altering the shape of the utility function, these extended models still assume a representative investor who holds the market portfolio. The most common approach to test these models is by testing the first-order optimality condition (or Euler equation) for the market portfolio. These conditions imply an exact linear relationship between assets' co-moments with the market portfolio. For example, the three-moment CAPM predicts an exact linear relationship between mean, co-variance and co-skewness.

A difficulty in changing the preference assumptions is the need to give a parametric specification of the functional form of the utility function. Unfortunately, economic theory gives minimal guidance for this purpose, and there is a substantial risk of specification error. For example, the three-moment CAPM ignores the central moments of order higher than three (e.g., kurtosis), as well as the lower partial moments (see, e.g., Bawa and Lindenberg (1977)), which generally cannot be expressed in terms of the first three central moments. Another problem associated with low order polynomials is the difficulty to impose restrictions on the derivatives that apply globally. For example, we cannot impose nonsatiation by restricting a quadratic function to be monotone increasing and we cannot impose risk aversion by restricting an increasing cubic function to be globally concave (see, e.g., Levy (1969)). To circumvent these problems, we may use the SSD efficiency test. Specifically, the market must be SSD efficient for all asset-pricing models that use a nonsatiable and risk-averse representative investor, and SSD inefficiency would imply that all such models would fail to rationalize the market portfolio, regardless of the functional form of the utility function.

To illustrate his test, Post (2003) tests if the Fama and French market portfolio is SSD efficient relative to the 25 Fama and French size and BE/ME portfolios. Interestingly, his results suggest that the market portfolio is significantly inefficient. In this section, we perform a more rigorous study of market efficiency. Post (2003) used a 460-month or 39-year sample period (July 1963 to October 2001) and he assumed that the asset return distribution remains unchanged for the entire period (apart from changes in the risk-free rate). However, there is substantial evidence that the return distribution (e.g., risk premiums, volatilities and correlation coefficients) varies through time. Hence, the observations generally are not serially IID random variables and the EDF is not a statistically consistent estimator for the CDF. Ideally, we would circumvent this problem by developing a test for *conditional* SSD efficiency that links the ex ante return distribution to the investors' (time-varying) information set. Unfortunately, the search for a satisfactory specification of the return dynamics is still far from accomplished. In fact, Ghysels (1998) finds that ill-specified conditional asset pricing models in many cases yield greater pricing errors than unconditional models. For this reason, we take another approach to account for time-variation. Specifically, we use a moving window analysis that applies the SSD test to a series of consecutive, short subsamples. Specifically, we consider 35 subsamples of 60 months, separated by 12-month intervals, beginning with July 1963 to June 1968 and ending with July 1997 to June 2002). In addition, we analyze 33 subsamples of 90 months, beginning with July 1963 to December 1970 and ending with July 1995 to December 2002, and 30 periods of 120 months, beginning with July 1963 to June 1973 and ending with July 1992 to June 2002. This approach is far less sensitive to time-variation, since it assumes that the distribution of excess returns is fixed for subsamples of 60, 90 or 120 months rather than for the full sample. Of course, this benefit comes at the cost of additional sampling error for small subsamples. Given the low power of the asymptotic test procedures in small samples, we therefore use the bootstrap procedure discussed in Section III (Procedure C). The simulation experiment in that section suggests that this procedure involves acceptable statistical size and power properties for samples of 60 to 120 observations.

For the purpose of comparison, we also apply a simple GMM-type test for mean-variance efficiency to each subsample. This test will help to determine the relative strength of the size and BE/ME effects in the different subsamples. If the market portfolio is mean-variance efficient, then the expected returns $\int \boldsymbol{x} dG(\boldsymbol{x})$ must equal the market risk premium $\int (\boldsymbol{x}^{\mathrm{T}} t) dG(\boldsymbol{x})$ times the market betas $\boldsymbol{b} \equiv (b_1 \cdots b_N)$. Hence, the null of mean-variance efficiency can be stated as $H_2 : e = \mathbf{0}$, with $e \equiv \int (\boldsymbol{x} - \boldsymbol{b}\boldsymbol{x}^{\mathrm{T}} t) dG(\boldsymbol{x})$ for the pricing errors. To test this null, we compute the sample pricing errors $\hat{e} \equiv \sum_{t=1}^{T} (\boldsymbol{x}_t - \hat{\boldsymbol{b}}\boldsymbol{x}_t^{\mathrm{T}} t) / T$, with $\hat{\boldsymbol{b}} \equiv (\hat{b}_1 \cdots \hat{b}_N)$ for the sample betas. Next, we construct the goodness measure $J \equiv \hat{e}^{\mathrm{T}} \hat{\mathrm{P}}^{-1} \hat{e}$, with $\hat{\mathrm{P}}$ for the sample variance-covariance matrix of the pricing errors. Under the null H_2, JT asymptotically obeys a chi-squared distribution with N degrees of freedom. Hence, we may compute asymptotic p-values for H_2 as $c_N^{-1}(J)$, with $c_N^{-1}(\cdot)$ for the inverse of the cumulative chi-squared distribution function with N degrees of freedom.[9] Figure 4 shows the resulting p-values. In 23 60-month subsamples (66 percent), the p-value is smaller than ten percent and the market is classified as mean-variance inefficient with at least 90 percent confidence. Similarly, for the 90-month subsamples, the market is classified as inefficient with at least 90 percent confidence in 21 cases (64 percent), and, for the 120-month subsamples, the market is inefficient 27 times (90 percent). Hence, for the large majority of subsamples, size and BE/ME effects occur. Clearly, the inefficiency classifications are not randomly distributed across the subsamples. The evidence against efficiency is especially strong during the late 1970s and during the 1980s. However, the evidence is weaker during the late 1960s and early 1970s and during the 1990s; all cases in which we cannot reject efficiency are concentrated in these periods.

[Insert Figure 4 about here]

Since SSD effectively considers the entire return distribution, the SSD efficiency test can help to determine if the size and BE/ME effects can be explained by the omission of higher order central moments and partial lower partial moments. Figure 5 shows the bootstrap p-values for the SSD efficiency test. The results are surprisingly similar to the results of the mean-variance efficiency test. In 22 60-month subsamples, 23 90-month subsamples and 29 120-month subsamples, the market portfolio is classified as

SSD inefficient with at least 90 percent confidence. Again, the evidence against efficiency is especially strong during the late 1970s and during the 1980s, and it is weak during the late 1960s and early 1970s and during the 1990s. Brief, for the subsamples in which the market portfolio is mean-variance inefficient, the market portfolio generally is also SSD inefficient, and no rational, nonsatiable and risk averse investor would hold this portfolio. This suggests that the size and BE/ME effects cannot be explained by omitted risk variables.

Our results do not solve the size and BE/ME puzzle; the results merely suggest that one possible explanation, the omission of risk factors, is unlikely to solve the puzzle. Several alternative explanations remain to be explored. For example, contrary to the predictions of representative investor models, investors (both individual and institutional) actually hold highly undiversified portfolios (see, e.g., Levy (1978)). Perhaps we have to move to models with heterogeneous investors and incomplete markets in order to understand the size and BE/ME effects.

[Insert Figure 5 about here]

V. Conclusions

1. We derive the asymptotic sampling distribution for the Post (2003) SSD test statistic. This distribution uses the least favorable distribution that minimizes Type I error at the cost of Type II errors. Hence, a test procedure based on this sampling distribution involves low size and power in small samples.

2. Contrary to Post (2003), we considered the sampling distribution under the null of SSD efficiency rather than the null of equal means (or equivalently the null that investors are risk neutral). The latter null is an unrealistic special case of the former null and the associated sampling distribution may lead to erroneous conclusions. Specifically, the null of equal means will underestimate the *p*-values and the critical values associated with the true null of SSD efficiency, and hence it involves lower size and power in small samples.

3. The bootstrap involves more power than a test procedure that is based on the asymptotic sampling distribution, while its size is closer to the nominal significance level. Hence, bootstrapping seems an interesting approach to sampling error, especially with the powerful computer hardware and software currently available.

4. We use the bootstrap to test if the Fama and French value-weighted market portfolio is SSD efficient relative to benchmark portfolios formed on size and BE/ME. During the late 1970s and during the 1980s, the market portfolio seems significantly mean-variance inefficient and size and BE/ME effects occur. In the same period, the market portfolio is also significantly SSD inefficient, even if we use samples of only 60 monthly observations. Since SSD effectively considers the entire return distribution, this suggests that the size and BE/ME effects cannot be explained by the omission of higher order central moments and partial lower partial moments.

5. Perhaps we have to move from models with a representative investor to models with heterogeneous investors who hold different, possibly highly undiversified portfolios, in order to solve the size and BE/ME puzzle. Dybvig and Ross (1982) have demonstrated that the SSD efficient set generally is not convex, and hence, there is no guarantee that the market portfolio is SSD efficient if different investors hold different portfolios of risky assets. Hence, a test for SSD efficiency of the market portfolio generally is not relevant in the context of a model with heterogeneous investors. Rather, we would need a test for 'SSD spanning' that tests if all traded assets are included in some SSD efficient portfolio (not necessarily with a weight that equals the relative market capitalization). Compared with the existing tests for mean-variance spanning (see, e.g., Huberman and Kandel, 1987), such a test would account for the full return distribution rather than its first two central moments only.

References

Bawa, Vijay S., and Eric B. Lindenberg, 1977, Capital market equilibrium in a mean-lower partial moment framework, *Journal of Financial Economics* 4, 189-200.

Beach, Charles and Russell Davidson, 1983, Distribution-free statistical inference with Lorenz curves and income shares, *Review of Economic Studies* 50, 723-735.

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

Dardanoni, Valentino and Antonio Forcina, 1999, Inference for the Lorenz curve orderings, *Econometrics Journal* 2, 49-75.

Davidson, Russell and Jean-Yves Duclos, 2000, Statistical inference for stochastic dominance and for the measurement of poverty and inequality, *Econometrica* 68, 1435-1464.

Dybvig, Phillip and Stephen Ross, 1982, Portfolio efficient sets, *Econometrica* 50, 1525-1546.

Efron, Bradley, 1987, Better bootstrap confidence intervals, *Journal of the American Statistical Association* 82, 171-200.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427-465.

Ghysels, Eric, 1998, On stable factor structures in the pricing of risk: Do time-varying betas help or hurt?, *Journal of Finance* 53, 549-573.

Harvey, Campbell R., and Akhtar Siddique, 2000, Conditional skewness in asset pricing tests, *Journal of Finance* 55, 1263-1295.

Huberman, G. and S. Kandel (1987): "Mean-Variance Spanning", *Journal of Finance* 42 (4), 873-888.

Kraus, Alan and Robert Litzenberger, 1976, Skewness preference and the valuation of risk assets, *Journal of Finance* 31, 1085-1100.

Kroll, Yoram and Haim Levy, 1980, Sampling errors and portfolio efficient analysis, *Journal of Financial and Quantitative Analysis* 15, 655-688.

Levy, Haim, 1969, Comment: A utility function depending on the first three moments, *Journal of Finance* 24, 715-719.

Levy, Haim, 1978, Equilibrium in an imperfect market: a constraint on the number of securities in a portfolio, *American Economic Review* 68, 643-658.

Levy, Haim, 1992, Stochastic dominance and expected utility: Survey and analysis, *Management Science* 38, 555-593.

Post, Thierry, 2003, Empirical Tests for Stochastic Dominance Efficiency, forthcoming in *Journal of Finance*.

Appendix

Proof to Theorem 1 Since $F_C(\boldsymbol{x})$ is a consistent estimator for $G(\boldsymbol{x})$ (see Section I), $u_{F_C}'^{*}$ converges to $u_G'^{*}$. Hence, $x(t, F_C)$ asymptotically behaves as the largest element of the vector $\mathbf{C}C\nabla u_G^{*}(t)/T =$ $(\sum_{t=1}^{T} u_G'^{*}(\boldsymbol{x}_t^{\mathrm{T}} t)(x_{1t} - \boldsymbol{x}_t^{\mathrm{T}} t)/T \cdots \sum_{t=1}^{T} u_G'^{*}(\boldsymbol{x}_t^{\mathrm{T}} t)(x_{Nt} - \boldsymbol{x}_t^{\mathrm{T}} t)/T)^{\mathrm{T}}$. Since the observations are serially IID, the vectors $\mathbf{C}\boldsymbol{x}_t u_G'^{*}(\boldsymbol{x}_t^{\mathrm{T}} t)$, $t = 1, \cdots, T$, are serially IID random vectors with mean $m \equiv \int u_G'^{*}(\boldsymbol{x}^{\mathrm{T}} t)(\boldsymbol{x}^{\mathrm{T}} t - \boldsymbol{x})dG(\boldsymbol{x})$ and variance-covariance matrix $\mathbf{C}W_0\mathbf{C}^{\mathrm{T}}$. Therefore, the Lindeberg-Levy central limit theorem implies that the vector $\mathbf{C}C\nabla u_G^{*}(t)/T$ obeys an asymptotically joint normal distribution with mean m and variance-covariance matrix $S_0 \equiv (\mathbf{C}W_0\mathbf{C}^{\mathrm{T}})/T$. Consequently, $x(t, F_C)$ asymptotically behaves as the largest order statistic of N random variables with a multivariate normal distribution, and $\Pr[x(t, F_C) > y | H_0] = 1 - \Pr[x(t, F_C) \le y | H_0]$ symptotically equals the multivariate normal integral $(1 - \int_{z \le ye} d\Phi(\boldsymbol{z} | S_0, m))$. To characterize the p-values under the null $H_0 : x(t, G) = 0$, we adhere to the statistical convention of using the least favorable distribution, i.e., the distribution $G(\boldsymbol{x})$ that maximizes the p-value under the null. The null can be stated equivalently as $H_0 : m \le 0$. In addition, $\int_{z \le ye} d\Phi(\boldsymbol{z} | S_0, m)$ is a decreasing function of m. Hence, the p-values are maximal if $m = 0$. Therefore, $\Pr[x(t, F_C) > y | H_0]$ asymptotically equals the multivariate normal integral $\Gamma(y, S_0) \equiv (1 - \int_{z \le ye} d\Phi(\boldsymbol{z} | S_0, \boldsymbol{0}))$.

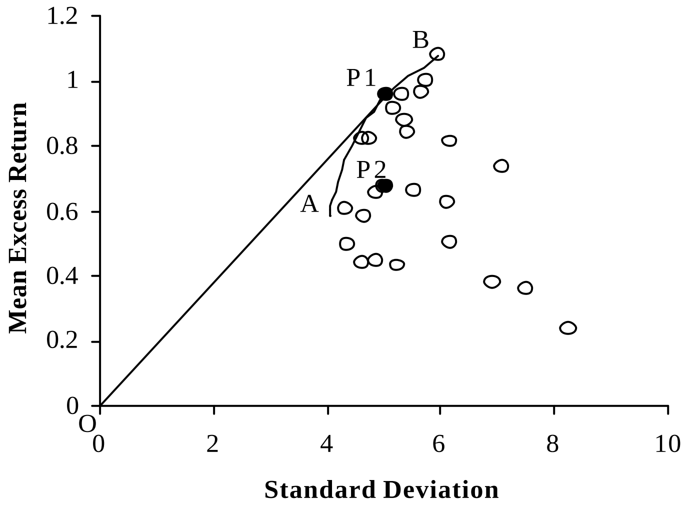

Figure 1. Mean-variance diagram 25 benchmark assets. Diagram for the mean excess returns and standard deviations of the 25 risky assets (the clear dots), as well as the efficient tangency portfolio (P1) and the inefficient equally weighted test portfolio (P2). The 25 assets obey a multivariate normal return distribution with joint population moments equal to the sample moments of the monthly excess returns of the 25 Fama and French benchmark portfolios. The curve AB represents the efficient frontier of risky assets without short selling. If we include the riskless asset, then OP1B represents the efficient frontier.

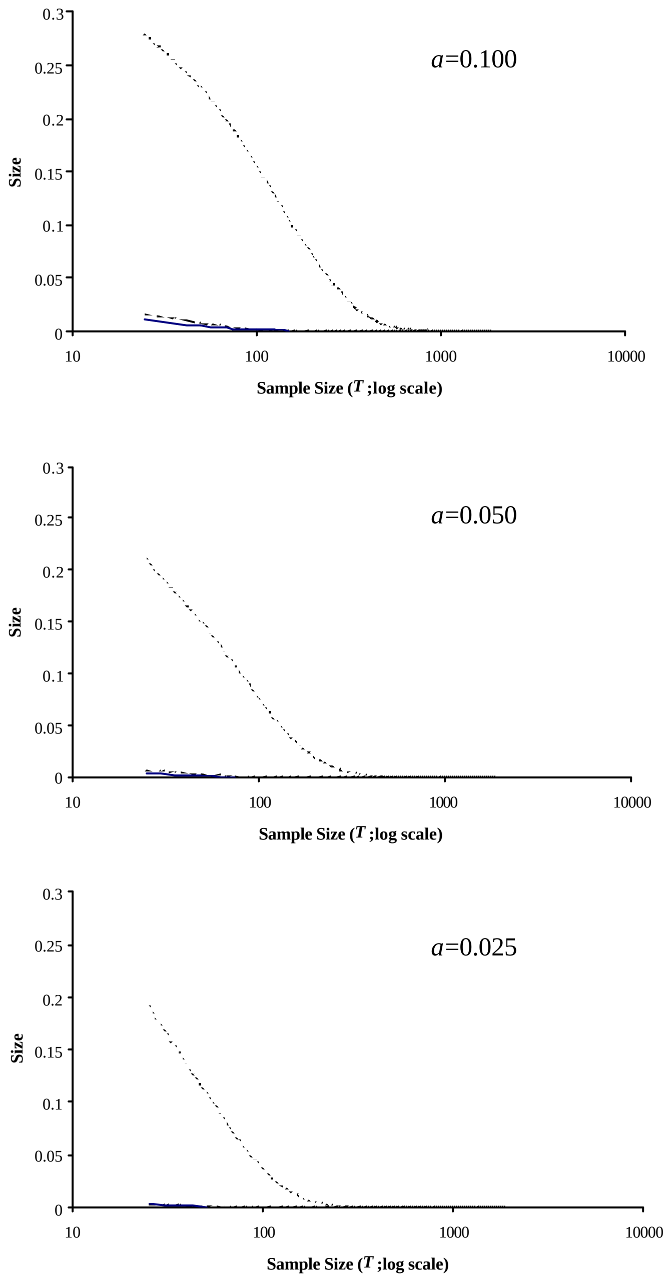

Figure 2. Size of competing test procedures. The figure shows the statistical size of the three competing test procedures. The dashed line shows the results for the procedure that uses the asymptotic sampling distribution under H_1 (Procedure A). Further, the solid line shows the results for the procedure that uses the asymptotic sampling distribution under H_0 (Procedure B). Finally, the dotted line shows the results for the bootstrap procedure (Procedure C). The figure displays the size for a sample size (T) of 25 to 4,000 and for a significance level (a) of 2.5, 5, and 10 percent. The results are based on 1,000 random samples from a multivariate normal distribution with joint moments equal to the sample moments of the monthly excess returns of the 25 Fama and French benchmark portfolios and the U.S. Treasury bill. Size is measured as the relative frequency of random samples in which the efficient tangency portfolio (P1) is wrongly classified as inefficient.

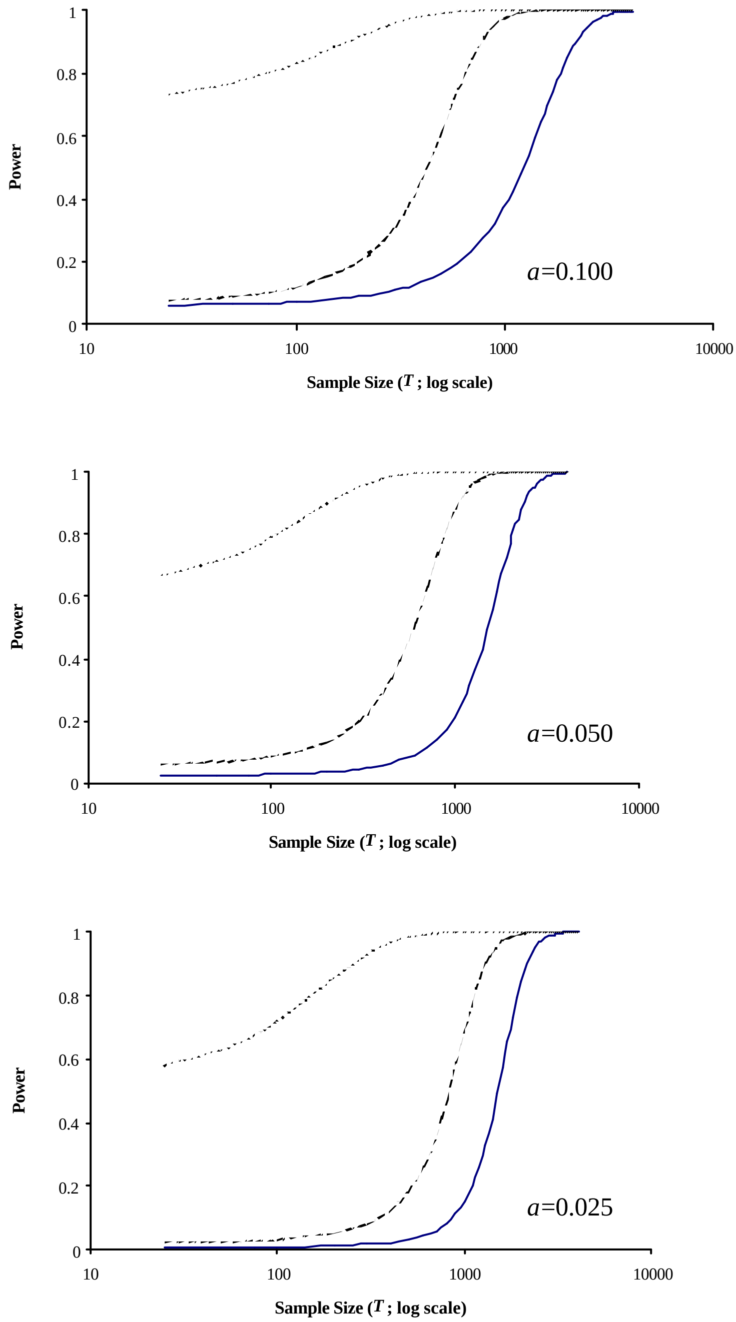

Figure 3. Power of competing test procedures. The figure shows the statistical power of the three competing test procedures. The dashed line shows the results for the procedure that uses the asymptotic sampling distribution under H_1 (Procedure A). Further, the solid line shows the results for the procedure that uses the asymptotic sampling distribution under H_0 (Procedure B). Finally, the dotted line shows the results for the bootstrap procedure (Procedure C). The figure displays the power for a sample size (T) of 25 to 4,000 and for a significance level (a) of 2.5, 5, and 10 percent. The results are based on 1,000 random samples from a multivariate normal distribution with joint moments equal to the sample moments of the monthly excess returns of the 25 Fama and French benchmark portfolios and the U.S. Treasury bill. Power is measured as the relative frequency of random samples in which the inefficient equally weighted portfolio (P2) is correctly classified as inefficient.

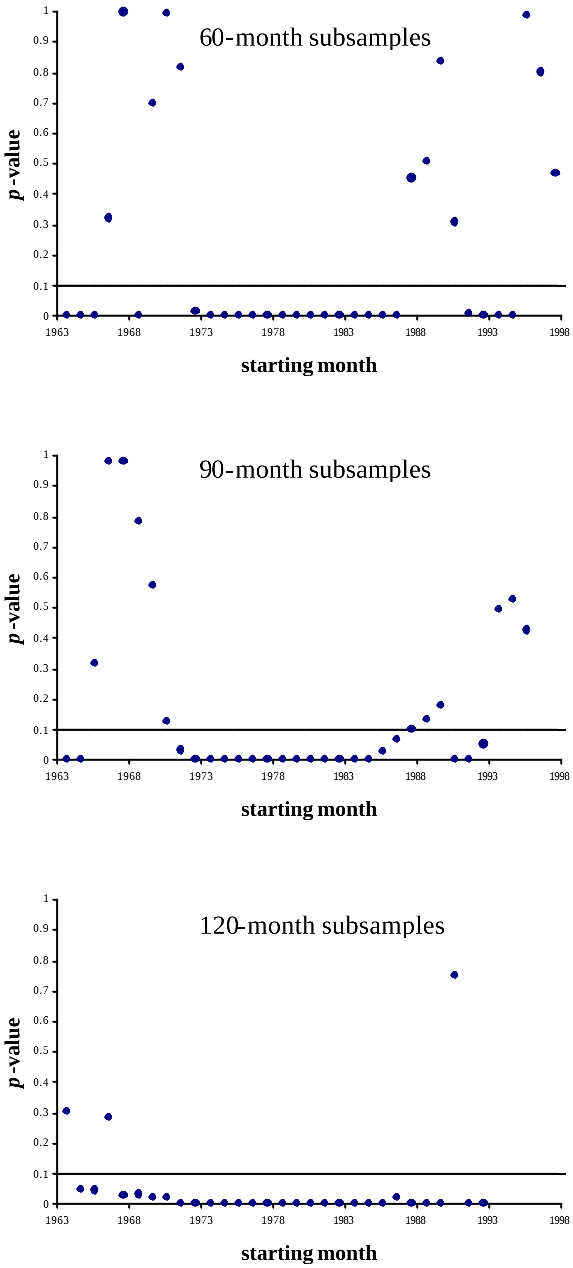

Figure 4. Asymptotic *p*-values for the null of mean-variance efficiency. The figure shows the asymptotic *p*-values for the null hypothesis of mean-variance efficiency for the Fama and French market portfolio relative to the 25 Fama and French benchmark portfolios formed on size and BE/ME and the one-month T-bill. Results are shown for 35 subsamples of 60 months, separated by 12 month intervals (beginning with Jul 1963-Jun 1968 and ending with Jul 1997-Jun 2002), 33 subsamples of 90 months (beginning with Jul 1963-Dec 1970 and ending with Jul 1995-Dec 2002), and 30 subsamples of 120 months (beginning with Jul 1963-Jun 1973 and ending with Jul 1992-Jun 2002). The *p*-values are computed as $c_N^{-1}(J)$, with $J \equiv \hat{e}^{\mathrm{T}}\hat{\mathrm{P}}^{-1}\hat{e}$ for the weighted average of the mean-variance CAPM sample pricing errors $\hat{e}$, weighted with the inverse sample variance-covariance matrix $\hat{\mathrm{P}}^{-1}$. We reject mean-variance efficiency of the market portfolio if and only if the asymptotic *p*-value falls below ten percent.

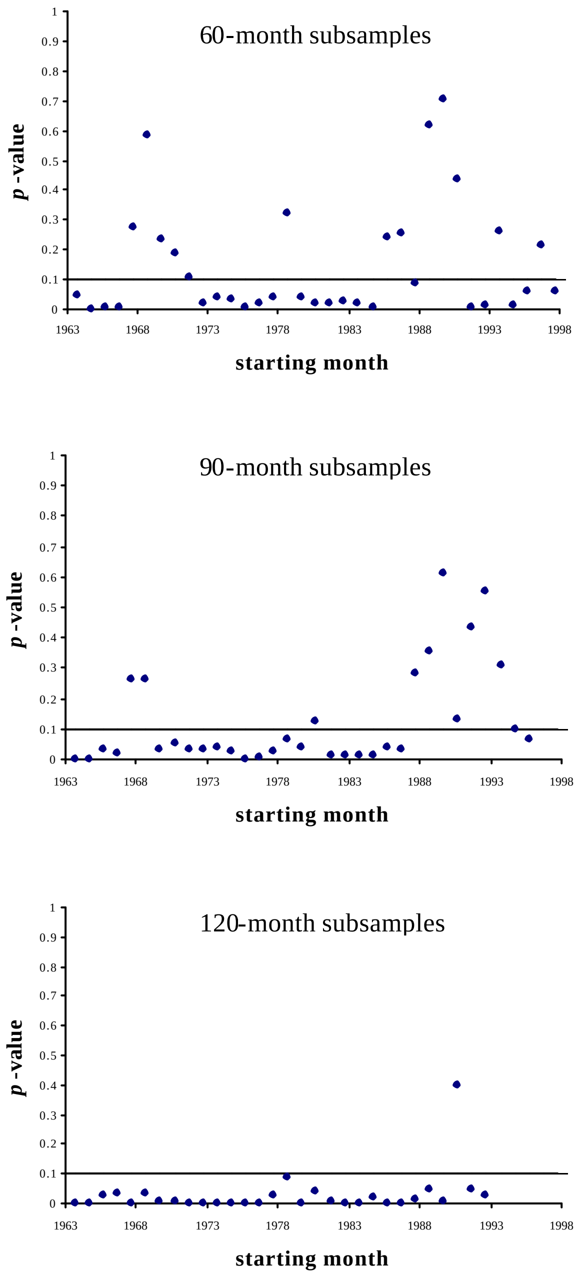

Figure 5. Bootstrap *p*-values for the null of SSD efficiency. The figure shows the bootstrap *p*-values for the null hypothesis of SSD efficiency for the Fama and French market portfolio relative to the 25 Fama and French benchmark portfolios formed on size and BE/ME and the one-month T-bill. Results are shown for 35 subsamples of 60 months, separated by 12 month intervals (beginning with Jul 1963-Jun 1968 and ending with Jul 1997-Jun 2002), 33 subsamples of 90 months (beginning with Jul 1963-Dec 1970 and ending with Jul 1995-Dec 2002), and 30 subsamples of 120 months (beginning with Jul 1963-Jun 1973 and ending with Jul 1992-Jun 2002). The *p*-values are computed using the BCa bootstrap method (see, e.g., Efron (1987)), which corrects for possible bias and skewness in the test statistic, and using 1,000 pseudosamples for each subsample. We reject SSD efficiency of the market portfolio if and only if the bootstrap *p*-value falls below ten percent.

Footnotes

[1] Throughout the text, we will use $\Re^N$ for an *N*-dimensional Euclidean space, and $\Re_+^N$ denotes the positive orthant. Further, to distinguish between vectors and scalars, we use a bold font for vectors and a regular font for scalars. Finally, all vectors are column vectors and we use $\boldsymbol{x}^{\mathrm{T}}$ for the transpose of $\boldsymbol{x}$.

[2] Post (2003) does not assume that the utility function is continuously differentiable, so as to allow for, e.g., piecewise linear utility functions. However, in practice, we typically cannot distinguish between a kinked utility function and a smooth utility function with rapidly changing marginal utility.

[3] The simplex Λ excludes short sales and it assumes that no additional restrictions are imposed on the portfolio weights. The SSD test is based on the first-order optimality conditions for optimizing a concave objective function over a convex set. In principle, the analysis is can be extended to a general polyhedral portfolio possibilities set, and hence it is possible to introduce short selling and to impose additional investment restrictions. We basically have to check whether there exists an increasing hyperplane that supports the extreme points of the portfolio possibilities set. One approach is to enumerate all extreme points and to include all extreme points as virtual assets.

[4] If an asset $i \in \mathrm{I}$ is included in the evaluated portfolio, i.e., $t_i > 0$, then the inequality $\int u'(\boldsymbol{x}^{\mathrm{T}} t)(\boldsymbol{x}^{\mathrm{T}} t - x_i) dG(\boldsymbol{x}) \le 0$ reduces to the equality $\int u'(\boldsymbol{x}^{\mathrm{T}} t)(\boldsymbol{x}^{\mathrm{T}} t - x_i) dG(\boldsymbol{x}) = 0$. Hence, if all assets are included, i.e., $t > 0$, then SSD efficiency reduces to $m \equiv \int u_G'^{*}(\boldsymbol{x}^{\mathrm{T}} t)(\boldsymbol{x}^{\mathrm{T}} t - \boldsymbol{x}) dG(\boldsymbol{x}) = \mathbf{0}$. Interestingly, this is the case where the probability of wrongly rejecting efficiency is maximal, i.e., the least favorable distribution used in the proof to Theorem 1.

[5] Since utility functions are unique up to a positive linear transformation, this standardization does not affect our results.

[6] We approximate $\Gamma^{-1}(a, \hat{\mathrm{S}}_l)$, $l = 0,1$, using Monte-Carlo simulation. Specifically, we generate 1,000 independent standard normal random vectors $\boldsymbol{w}_s \in \Re^{N-1}$, $s \in \{1, \cdots, 10{,}000\}$, using the RNDN function in Aptech Systems' GAUSS software. Next, each random vector $\boldsymbol{w}_s$ is transformed into a multivariate normal vector $\boldsymbol{z}_{l,s} \in \Re^N$ with variance-covariance matrix $\hat{\mathrm{S}}_l$, using $\boldsymbol{z}_{l,s} = \mathbf{C}\hat{\mathbf{D}}_l \boldsymbol{w}_s$, with $\hat{\mathbf{D}}_l \in \Re^{N \times N}$ for a lower triangular Cholesky factor of $\hat{\mathrm{W}}_l$, so $\hat{\mathrm{W}}_l = \hat{\mathbf{D}}_l \hat{\mathbf{D}}_l^{\mathrm{T}}$. Finally, $\Gamma^{-1}(a, \hat{\mathrm{S}}_l)$ is approximated by the (1-*a*)th percentile of the distribution of the largest elements of the transformed vectors $\boldsymbol{z}_s$, $s \in \{1, \cdots, 10{,}000\}$.

[7] To correct for possible bias and skewness in the test statistic, we use the BCa method (see, e.g., Efron (1987)). The raw bootstrap distribution of the SSD test statistic generally involves positive bias and skewness, and not correcting for bias and skewness lowers the bootstrap *p*-values and hence increases size and power.

[8] Less restrictive assumptions are obtained if we do restrict the shape of the return distribution; see, e.g., Berk (1997).

[9] This procedure is equivalent to a GMM estimation of the Euler equation for a quadratic utility function, but with the parameters of the utility function fixed such that the market risk premium equals the average excess return of the market portfolio.

Publications in the Report Series Research* in Management

ERIM Research Program: "Finance and Accounting"

2003

COMMENT, Risk Aversion and Skewness Preference
Thierry Post and Pim van Vliet
ERS-2003-009-F&A

International Portfolio Choice: A Spanning Approach
Ben Tims, Ronald Mahieu
ERS-2003-011-F&A

Portfolio Return Characteristics Of Different Industries
Igor Pouchkarev, Jaap Spronk, Pim van Vliet
ERS-2003-014-F&A

Statistical Inference on Stochastic Dominance Efficiency.
Do Omitted Risk Factors Explain the Size and Book-to-Market Effects?
Thierry Post
ERS-2003-017-F&A

A Multidimensional Framework for Financial-Economic Decisions
Winfried Hallerbach & Jaap Spronk
ERS-2003-021-F&A

---

* A complete overview of the ERIM Report Series Research in Management:
http://www.ers.erim.eur.nl

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship