Centre de recherche sur l'emploi et les fluctuations économiques (CREFÉ)
Center for Research on Economic Fluctuations and Employment (CREFE)

Université du Québec à Montréal

Cahier de recherche/Working Paper No. 143

Testing the Capital Asset Pricing Model Efficiently Under Elliptical Symmetry : A Semiparametric Approach

Douglas J. Hodgson
UQAM

Oliver Linton
London School of Economics

Keith Vorkink
Brigham Young University

October 2001

Hodgson : Département des sciences économiques, Université du Québec à Montréal, C.P. 8888 Succursale Centre-Ville, Montréal, QC, Canada, H3C 3P8. Tel : 514-987-3000 (ext.4310). Email: hodgson.douglas-james@uqam.ca.

Linton: Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Email: lintono@lse.ac.uk.

Vorkink: Marriott School of Management, Brigham Young University, Provo, UT, USA, 84602. Tel: 801-378-1765. Email: kpv3@email.byu.edu.

For their helpful comments, we would like to thank John Geweke and two referees, Pedro Gozalo, Gene Savin, and participants in seminars at Toronto, Montreal, McGill, and the 1998 summer meetings of the North American and European Econometric Societies. We also thank Eugene Choo and Roger Moon for research assistance and the National Science for financial support. Hodgson thanks CREST at INSEE for their hospitality while part of this research was being carried out.

Sommaire:
Nous développons de nouveaux tests du modèle d'évaluation des actifs financiers (« CAPM ») qui tiennent compte de, et sont valides sous, l'hypothèse que les retours des actifs découlent d'un loi de probabilité elliptiquement symétrique. Cette hypothèse est nécessaire et suffisante pour la validité du CAPM. Notre test utilise un estimateur des paramètres du modèle qui a l'efficacité semiparamétrique quand on a un modèle de régression apparemment sans relation et qui a des erreurs qui suivent une loi elliptiquement symétrique. L'hypothèse de la symétrie elliptique nous permet d'éviter le problème d'estimer non-paramétriquement une fonction de haute dimension parce qu'on peut écrire la densité d'une loi elliptique comme une fonction d'une transformation unidimensionnelle de la variable aléatoire multidimensionnelle. La famille des lois elliptiquement symétriques inclue plusieurs lois leptokurtiques, donc elle est pertinente à des applications financières. Les bêtas obtenus avec notre estimateur sont plus bas que ceux qui sont obtenus en utilisant des moindres carrés, et sont moins compatibles avec le CAPM.

Abstract:
We develop new tests of the capital asset pricing model (CAPM) that take account of and are valid under the assumption that the distribution generating returns is elliptically symmetric; this assumption is necessary and sufficient for the validity of the CAPM. Our test is based on semiparametric efficient estimation procedures for a seemingly unrelated regression model where the multivariate error density is elliptically symmetric, but otherwise unrestricted. The elliptical symmetry assumption allows us to avert the curse of dimensionality problem that typically arises in multivariate semiparametric estimation procedures, because the multivariate elliptically symmetric density function can be written as a function of a scalar transformation of the observed multivariate data. The elliptically symmetric family includes a number of thick-tailed distributions and so is potentially relevant in financial applications. Our estimated betas are lower than the OLS estimates, and our parameter estimates are much less consistent with the CAPM restrictions than the corresponding OLS estimates.

Keywords: Adaptive estimation, capital asset pricing model, elliptical symmetry, semiparametric efficiency.

JEL classification: C22

1 Introduction

The mean-variance approach to asset pricing theory, initially investigated in work such as that of Tobin (1958) and Markowitz (1952, 1959), has great intuitive appeal and has the important practical advantage of greatly simplifying the modeling of asset returns. The principal result in this area is the capital asset pricing model (CAPM) of Sharpe (1964), Lintner (1965), and Mossin (1966), which posits that the expected excess return of any asset is linear in its covariance with the expected return on the market portfolio.[1] This relationship is formalized in the following equation:

$$E[R_i] = r_f + \beta_i \left(E[R_M] - r_f\right), \tag{1}$$

where R_i is the random rate of return on asset i, $\beta_i = \text{cov}[R_i, R_M]/\text{var}[R_M]$, R_M is the rate of return on the market portfolio and r_f is the risk-free rate, which is assumed to be observed in the Sharpe-Lintner version. Defining $r_i = E[R_i] - r_f$, equation (1) can be rewritten as $r_i = \beta_i r_M$.

The CAPM was originally derived under the assumption that either investors possess quadratic utility functions or that asset returns are normally distributed. Since quadratic utility functions have the intuitively unappealing property that they are decreasing at high consumption levels, the fact that the CAPM holds under normality for a much broader class of utility functions is comforting to proponents of the model. Unfortunately, there is a considerable amount of evidence that the assumption of normality is not an appropriate one for asset returns. There is a voluminous literature (dating back at least as far as Fama (1963, 1965) and Mandelbrot (1963)) documenting the excess thickness of the tails in asset return distributions relative to the normal. This tail thickness is associated with the tendency of asset returns to take values of extremely large magnitude with nonnegligible probability. Thus, it seems that we would need to fall back on the assumption of quadratic utility to justify the CAPM relationship (1). However, it has been shown that although, in the absence of strong restrictions on investor preferences, the assumption of normality is sufficient to generate (1), it is not necessary. In particular, Chamberlain (1983), Owen and Rabinovitch (1983), and most recently Berk (1997) show that (1) can be obtained under the assumption of elliptically symmetric return distributions without the strongly restricting preferences.[2] Berk (1997) shows that elliptical distributions are the most general distributional assumption that will imply the CAPM when agents maximize expected utility, that is, elliptical symmetry is both necessary and sufficient for the CAPM. The elliptically symmetric family contains the Gaussian distribution as a special case, but many well-known thick-tailed distributions also belong to this class - the Student t, logistic, and scale mixed-normal being examples.[3]

[1] The market portfolio is a value weighted portfolio of all assets in the market.

[2] See also Ingersoll (1987).

[3] See Fernández, Osiewalski, and Steel (1995) for some generalizations of ellyptical symmetry that are interesting

That asset returns may be non-normal can have important implications for the econometric implementation of the model, which often involves the estimation of a system of linear equations specified as a seemingly unrelated regressions (SUR) model (see, for example, MacKinlay (1987) and Gibbons, Ross, and Shanken (1989)). The standard estimator of this model is ordinary least squares (OLS), which will be fully efficient under normality, but will not be fully efficient if normality fails. In an innovative paper, Zhou (1993) considers implementation of OLS under possible non-normality, deriving a procedure to correct the size problems that may occur in CAPM tests if returns are elliptical but non-normal.

The contribution of the present paper is to derive an estimator of the SUR model that will be fully efficient in the presence of elliptical symmetry of general form. The resulting estimator will be more efficient than OLS and will yield more powerful CAPM tests than those of Zhou (1993) for example. The estimator we propose is semiparametric in nature, treating the true distribution of the data as being unknown (aside from the elliptical symmetry restriction) and is fully "adaptive" (Bickel (1982)), i.e., it will achieve the same asymptotic covariance matrix lower bound as would the maximum likelihood estimator if the distribution of the data were known.

Semiparametric methods such as we develop here employ nonparametric kernel smoothers to estimate the unknown distribution of the data and are well developed for single equation estimation problems, see for example Stone (1975), Bickel (1982), and Kreiss (1987). Some methods have also been proposed for multivariate data, see Bickel (1982) and Hodgson (1998b). However, there are problems with smoothing methods with high dimensional data: the estimates are hard to plot and interpret, and have slow convergence rates. For this reason, some intermediate structures are becoming increasingly popular, such as additive models in regression, see for example Horowitz (2001). This problem is often referred to as the "curse of dimensionality" and is of particular relevance to our problem of efficiently estimating the SUR system, since the semiparametric estimator requires the kernel estimation of a density whose dimensionality equals that of the system. However, if we exploit the elliptical symmetry assumption underlying the CAPM, then we have the opportunity to avoid the curse of dimensionality. This is because the density function of a vector-valued elliptical random variable can always be rewritten as the density function of a scalar random variable, regardless of the dimension of the vector. Owen and Rabinovitch (1983), in showing that the CAPM would hold under elliptical symmetry, also suggested that the possibility of elliptical symmetry should be taken into account in the formulation of econometric models of the CAPM. In recent years, it has become possible, due to some advances in econometric estimation theory, to incorporate the general assumption of elliptical symmetry into an econometric model without having to be more specific about the actual functional form of the distribution. The implication is that our computation of an adaptive

from statistical point of view.

estimator will always only require a one-dimensional nonparametric estimation problem, regardless of the size of the system, and so is not subject to the curse of dimensionality. This intuition is shown to be correct by Stute and Werner (1991).

In Section 2, we introduce the SUR model that we are interested in analyzing. In Section 3, we outline a formula for computing an adaptive estimator and give its asymptotic properties. Section 4 reports the results of our empirical test of the CAPM, while Section 5 investigates the performance of the estimator through a Monte Carlo simulation analysis. A mathematical appendix contains proofs. We use $\|A\| = \left(\mathrm{tr}\, A^T A\right)^{1/2}$ to denote the Euclidean norm of a vector or matrix A, while $\xrightarrow{P}$ denotes convergence in probability and $\xrightarrow{D}$ denotes convergence in distribution.

2 The SUR Model

In this section we consider the specification and estimation of a general seemingly unrelated regressions (SUR) model. The CAPM regression, implemented in Section 4, falls within this class. Consider the m-equation seemingly unrelated regression model

$$y_t = \alpha + x_t\beta + u_t := w_t\theta + u_t, \quad t = 1, \ldots, n, \tag{2}$$

where $y_t \in \mathbb{R}^m$, $\alpha \in \mathbb{R}^m$, $w_t = [I_m\ x_t]$, in which

$$x_t = \begin{bmatrix} x_{1t} & & & 0 \\ & x_{2t} & & \\ & & \ddots & \\ 0 & & & x_{mt} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \quad u_t = \begin{pmatrix} u_{1t} \\ \vdots \\ u_{mt} \end{pmatrix},$$

where $x_{it} \in \mathbb{R}^{k_i}$ and $\beta_i \in \mathbb{R}^{k_i}$ for every $i = 1, \ldots, m$, the full parameter vector is $\theta = \left[\alpha^T, \beta^T\right]^T \in R^{m+k}$, where $k = k_1 + \cdots + k_m$. The error terms $u_t \in \mathbb{R}^m$ are i.i.d., mean zero innovations with $E(u_t u_t^T) = \Sigma_u$. Here, the regressors x_t are assumed to be stationary and ergodic, and we assume that x_t and u_t are independent (i.e., that the regressors are strictly exogenous). The asymptotic properties of least squares estimators are standard under these assumptions.

We suppose that the error has a distribution that is absolutely continuous with respect to Lebesgue measure and has Lebesgue density $p(u)$. We shall assume that p is elliptically symmetric.

Definition. *An m-dimensional density function $p(u)$ is elliptically symmetric if it can be written in the form $(\det \Sigma)^{-1/2}\, g(u^T \Sigma^{-1} u)$ for some scalar density generating function $g(\cdot)$ and matrix Σ.*

The practical content of the elliptical symmetry restriction arises from the fact that the function g has only a scalar argument. Note that the matrix Σ is identified only up to a scalar multiple, as scale transformations in Σ can be incorporated into the function g. Without loss of generality, we shall use the normalization $\det(\Sigma) = 1$. Under this normalization, Σ is proportional to the covariance matrix of u, which we denote by Σ_u, so that $\Sigma_u = c\Sigma$, where $c = (\det \Sigma_u)^{1/m}$, i.e., $\Sigma = \Sigma_u/(\det \Sigma_u)^{1/m}$ [c.f. Kelker (1970) and Stute and Werner (1991)]. Also worth noting is the fact that the information matrix of p, Γ_p, is proportional to the inverses of these matrices [c.f. Mitchell (1989)].

If p were known, the log-likelihood for the data would be

$$L_n(\theta) = \sum_{t=1}^{n} \ln p(y_t - w_t\theta),$$

and a standard estimation method is to choose θ to maximize $L_n(\theta)$. One estimation strategy which avoids complicated nonlinear optimization associated with non-Gaussian p, is to use a two-step Newton-Raphson estimator $\overline{\theta}$ starting from a preliminary $\sqrt{n}$-consistent estimator $\widehat{\theta}$ that was obtained from the Gaussian likelihood (OLS, for example). This approach to estimation apparently originates with R.A. Fisher and has been widely used in econometrics. Under general conditions, this will be first order asymptotically equivalent to the maximum likelihood estimator (MLE), i.e.,

$$\sqrt{n}(\overline{\theta} - \theta_0) \xrightarrow{D} N(0, \mathcal{I}^{-1}),$$

where the asymptotic information matrix $\mathcal{I}$ is such that $n^{-1}\left(\partial^2 L_n(\theta_0)/\partial\theta\partial\theta'\right) \xrightarrow{P} \mathcal{I}$. In order to derive an expression for $\mathcal{I}$, we define $\varphi(u) = p'(u)/p(u)$, the m-dimensional score vector of p, and $\Gamma_p = \int \varphi(u)\varphi(u)^T p(u)du$, the information matrix of p. The asymptotic information matrix is

$$\mathcal{I} = \begin{bmatrix} \Gamma_p & E\left[\Gamma_p x_t\right] \\ E\left[x_t^T \Gamma_p\right] & E\left[x_t^T \Gamma_p x_t\right] \end{bmatrix}.$$

We use a Newton-Raphson iterative approach to estimation but must replace the unknown density p by a nonparametric estimator; thus our adaptive estimator $\widetilde{\theta}$ will have the form

$$\widetilde{\theta} = \widehat{\theta} + \widehat{\mathcal{I}}_n^{-1}(\widehat{\theta})\widehat{\Delta}_n(\widehat{\theta}), \tag{3}$$

where $\widehat{\Delta}_n$ and $\widehat{\mathcal{I}}_n$ are estimates of the first and second standardized derivatives of L_n respectively. Their computation is described in Section 3 below. In particular,

$$\widehat{\Delta}_n(\widehat{\theta}) = -\frac{1}{n}\sum_{t=1}^{n} w_t^T \widehat{\varphi}_t(\widehat{u}_t),$$

where $\widehat{\varphi}_t(\widehat{u}_t)$ is a consistent estimator of the m-dimensional score vector $\varphi(u_t)$, while $\widehat{u}_t = y_t - w_t\widehat{\theta}$. The standard approach to this problem is to use multivariate kernel estimates $\widehat{p}$ and $\widehat{p}'$ to construct

$\widehat{\varphi}$, with some observations possibly being trimmed, see Bickel (1982). Unfortunately, if m is large such estimates will have poor performance due to the curse of dimensionality, see Härdle and Linton (1994). We show how to construct a $\widehat{\varphi}_t(.)$ that takes advantage of our elliptical symmetry assumption and employs only one-dimensional smoothing operations.[4]

3 Estimation

The formula for an adaptive estimator given in (3) above presupposed the existence of consistent score and information estimators $\widehat{\varphi}_t$ and $\widehat{\mathcal{I}}_n$. In this section, we provide an algorithm for computing nonparametric estimates of these quantities while imposing the restriction that the errors $\{u_t\}$ have an elliptically symmetric distribution. Recall that the elliptical symmetry assumption allows us to reduce the dimensionality m of the density $p(u)$ to the dimension one of the function $g\left(u^T\Sigma^{-1}u\right) = g\left(\varepsilon^T\varepsilon\right)$, where $\varepsilon = \Sigma^{-1/2}u$ is a spherically symmetric random variable with density $f(\varepsilon) = g\left(\varepsilon^T\varepsilon\right) = g(v)$ where $v = \varepsilon^T\varepsilon$. We can thus obtain an indirect estimate of the density of u from a direct estimate of the density of the scalar random variable v. It may be preferable for computational reasons to directly estimate the density of the random variable $z = \tau(v)$, rather than that of v itself, and in our theory we allow for estimation of a general Box-Cox (1964) transformation $\tau(v) = (v^\zeta - 1)/\zeta$. We discuss our choice of ζ in our empirical and simulation work below. We will use direct kernel estimates of the density of z, denoted by $\gamma(z)$, to indirectly obtain consistent estimates of the score and information of p. By Theorem 2.1.2 of Casella and Berger (1990) we have

$$\gamma(z) = h(\tau^{-1}(z)) \cdot \left|\frac{\partial \tau^{-1}(z)}{\partial z}\right| = c_m \left[\tau^{-1}(z)\right]^{m/2-1} g(\tau^{-1}(z)) \cdot J_\tau(z),$$

where $h(v) = c_m v^{m/2-1} g(v)$ with $c_m = \pi^{m/2}/\Gamma(m/2)$ is the density of v, see Muirhead (1982), while $J_\tau(z) = |\partial\tau^{-1}(z)/\partial z|$. Thus, $g(v) = c_m^{-1} J_\tau^{-1}(\tau(v)) v^{1-m/2} \gamma(\tau(v))$. This gives us our desired expression for $g(v)$ - and hence for $f(\varepsilon)$ and $p(u)$ - in terms of $\gamma(z)$.

Our algorithm for estimating φ and $\mathcal{I}$ proceeds according to the following steps:

Step 1: First obtain $\widehat{\theta}$ (by ordinary least squares, for example) and define the associated OLS residuals $\{\widehat{u}_t\}_{t=1}^n$ and the standardized residuals $\{\widehat{\varepsilon}_t\}_{t=1}^n$, where $\widehat{\varepsilon}_t = \widehat{\Sigma}^{-1/2}\widehat{u}_t$, $\widehat{\Sigma} = \widehat{c}^{-1}\widehat{\Sigma}_u$, $\widehat{\Sigma}_u = (n-k-m)^{-1}\sum_{t=1}^n \widehat{u}_t\widehat{u}_t^T$, and $\widehat{c} = [\det \widehat{\Sigma}_u]^{1/m}$. Then compute the univariate transformed sequence $\{\widehat{z}_t\}_{t=1}^n$, where $\widehat{z}_t = \tau(\widehat{v}_t)$ with $\widehat{v}_t = \widehat{\varepsilon}_t^T\widehat{\varepsilon}_t$.

[4] As shown in Stute and Werner (1991) these procedures ensure density estimators whose pointwise rate of convergence is the one-dimensional rate.

Step 2: Compute leave-one-out kernel density and derivative estimators using data $\{\widehat{z}_t\}$, kernel $K_{h_n}(\cdot)$, and bandwidth h_n:

$$\widehat{\gamma}_t(z) = \frac{1}{n-1}\sum_{\substack{s=1\\ s\neq t}}^{n} K_{h_n}(z-\widehat{z}_s) \quad ; \quad \widehat{\gamma}'_t(z) = \frac{1}{n-1}\sum_{\substack{s=1\\ s\neq t}}^{n} K'_{h_n}(z-\widehat{z}_s).$$

Step 3: Introduce the following trimming conditions: (i) $\widehat{\gamma}_t(\widehat{z}_t) \geq d_n$; (ii) $|\widehat{z}_t| \leq e_n$; (iii) $|\lambda(\widehat{z}_t)| \leq b_n$; (iv) $|\rho^{1/2}(\widehat{z}_t)\widehat{\gamma}'_t(\widehat{z}_t)| \leq c_n\widehat{\gamma}_t(\widehat{z}_t)$, where $\rho(z) = v\tau'(v)J_\tau^{-1}(z)$ [recall that $v = \tau^{-1}(z)$] and $\lambda(z) = (d/dz)^{-1}\rho^{1/2}(z)$.[5] Then estimate the score and information of $p(u)$ as follows:

$$\widehat{\varphi}_t(\widehat{u}_t) = \begin{cases} \widehat{\Sigma}^{-1/2}\widehat{\varepsilon}_t\left[s(\widehat{v}_t) + \tau'(\widehat{v}_t)\frac{\widehat{\gamma}'_t}{\widehat{\gamma}_t}(\widehat{z}_t)\right] & \text{if (i) -- (iv) all hold} \\ 0 & \text{otherwise,} \end{cases}$$

where $s(v) = (1-m/2)v^{-1} - \frac{J'_\tau}{J_\tau}\{\tau(v)\}\tau'(v)$ and $\widehat{\Gamma}_p = \frac{1}{n}\sum_{t=1}^{n}\widehat{\varphi}_t(\widehat{u}_t)\widehat{\varphi}_t(\widehat{u}_t)^T$.

Step 4: Then define the score and information estimators for the model as

$$\widehat{\Delta}_n(\widehat{\theta}) = -\frac{1}{n}\sum_{t=1}^{n} w_t^T\widehat{\varphi}_t(\widehat{u}_t) \quad ; \quad \widehat{\mathcal{I}}_n(\widehat{\theta}) = \frac{1}{n}\sum_{t=1}^{n} w_t^T\widehat{\Gamma}_p w_t n^{-1}, \tag{4}$$

and compute the adaptive estimator $\widetilde{\theta}$ given in (3) above.

The important point to notice about this estimator is that it employs a direct kernel estimate of the density of the *univariate* process $\{z_t\}$ in order to arrive at score and information estimates of the *multivariate* process $\{u_t\}$.

We now state the main result of the paper, which is proved in the Appendix:

Theorem 1 *Suppose that Γ_p is finite and positive definite, that $\int_0^\infty v^{m/2}s(v)^2g(v)dv < \infty$, that the error distribution is absolutely continuous with respect to Lebesgue measure with Lebesgue density*

[5] These trimming conditions ensure consistency of our score estimator when a Gaussian kernel is being used, i.e., when K_{h_n} is a Gaussian kernel. For other kernels often employed in the literature [e.g., Schick's (1987) logistic kernel and the bi-quartic kernel], the necessary trimming conditions, if they differed at all from these, would be less stringent, so that these conditions will still be sufficient for consistency but may not be necessary. Simulation work reported by Hsieh and Manski (1987) and Hodgson (1998a) finds that, for a Gaussian kernel, the adaptive point estimate is not very sensitive to variation in the value of the trimming parameters, and that good results are obtained in practice when we trim as little as 1% of the observations.

$p(u)$, that the regressors x_t are strictly exogenous, and that the constants in (i)-(iv) satisfy $c_n \to \infty$, $e_n \to \infty$, $b_n \to \infty$, $h_n \to 0$, $d_n \to 0$, $h_n c_n \to 0$, $e_n h_n^{-3} = o(n)$, and $b_n h_n^{-3} = o(n)$. Then,

$$\sqrt{n}(\widetilde{\theta} - \theta) \xrightarrow{D} N(0, \mathcal{I}^{-1}), \tag{5}$$

i.e., the estimator $\widetilde{\theta}$ is adaptive.

Remarks. (a) The moment condition $\int_0^\infty v^{m/2} s(v)^2 g(v) dv < \infty$ is potentially restrictive; its implications for the moments of u will depend on the transformation $\tau(\cdot)$. For example, when the transformation is $\tau(v) = (v^\zeta - 1)/\zeta$ with either $\zeta = 0$, $\zeta = 1$, or $1/2m$, the condition implies $E[(\varepsilon^T \varepsilon)^{m/2-2}] < \infty$. However, when $\zeta = m/2$, there is no restriction on the moments of u.

(b) Note that the information matrix estimator $\widehat{\mathcal{I}}_n(\widehat{\theta})$ defined in (4) is a consistent estimator of the asymptotic covariance matrix, so that $\widehat{\mathcal{I}}_n(\widehat{\theta}) - \mathcal{I} = o_p(1)$. We can therefore use $\widehat{\mathcal{I}}_n(\widehat{\theta})$ in the construction of t-ratios and Wald statistics which will have respective standard normal and chi-squared asymptotic distributions. Let θ_ℓ and $\widetilde{\theta}_\ell$ be the ℓ^{th} elements of the θ and $\widetilde{\theta}$ vectors, respectively. Now suppose we wish to test the null hypothesis that $\theta_\ell = c$, where c is some constant. Then we can compute the usual t-ratio, as follows:

$$t = \frac{\sqrt{n}\left(\widetilde{\theta}_\ell - c\right)}{\sqrt{\left(\widehat{\mathcal{I}}_n^{-1}(\widehat{\theta})\right)_{\ell\ell}}},$$

where $\left(\widehat{\mathcal{I}}_n^{-1}(\widehat{\theta})\right)_{\ell\ell}$ is the ℓ^{th} elements along the diagonal of $\widehat{\mathcal{I}}_n^{-1}(\widehat{\theta})$. Under the null hypothesis, $t \xrightarrow{D} N(0,1)$. If we want to test the joint hypothesis $\tau(\theta) = 0$, where τ is a known $(m+k) \times 1$ vector of functions, we can compute the Wald statistic

$$W = n\tau(\widetilde{\theta})' \left[\dot{\tau}(\widehat{\theta})\widehat{\mathcal{I}}_n^{-1}(\widehat{\theta})\dot{\tau}(\widehat{\theta})'\right]^{-1} \tau(\widetilde{\theta})',$$

where $\dot{\tau}$ is the matrix of derivatives of τ with respect to θ. Under the null hypothesis $W \xrightarrow{D} \chi^2_{m+k}$.

(c) Our estimator of the information matrix, although consistent, has a finite sample upwards bias that therefore biases downwards our standard error estimates. In our empirical application, we employ a simple degrees of freedom correction. Write $\widehat{\gamma}_t' = \sum_s \omega_{nts}'$ and $\widehat{\gamma}_t = \sum_s \omega_{nts}$ for some weights ω_{nts}' and ω_{nts} implicitly defined in our estimation algorithm. We replace $(\widehat{\gamma}_t')^2$ and $(\widehat{\gamma}_t)^2$ in (4) by $(\widehat{\gamma}_t')^2 - \sum_s (\omega_{nts}')^2$ and $(\widehat{\gamma}_t)^2 - \sum_s (\omega_{nts})^2$ respectively. The correction terms $\sum_s (\omega_{nts}')^2$ and $\sum_s (\omega_{nts})^2$ consistently estimate the degrees of freedom bias terms (see Linton (1995)).

(d) We employ one-dimensional kernel estimates of the transformed variable z, which has a support restriction of $z \geq 0$. The kernel estimate will generally have a downward bias in the right neighborhood of zero. This bias arises because for points close to zero, the kernel smoother assigns positive weight extends to points $x \leq 0$ where $f(x) = 0$. The overflow in weights beyond the lower support of 0 can be corrected by applying a result of Schuster (1985), who offers a correction that incorporates this overflow to the region $z < c$, for finite c, back into the region $z \geq c$ by adding a mirror image term $n^{-1}h_n^{-1}K((z-2c+z_i)/h_n)$ to $n^{-1}h_n^{-1}K((z-z_i)/h_n)$. The resulting estimator for $z \geq c$ is given by

$$\widetilde{f}_n(z) = \frac{1}{nh_n}\sum_{i=1}^{n}\left[K\left(\frac{z-z_i}{h_n}\right) + K\left(\frac{z-2c+z_i}{h_n}\right)\right].$$

In our case, $c = 0$. Schuster (1985) also proves consistency and asymptotic normality results for this estimator.

(e) The advantage of the adaptive estimator over alternative estimators such as OLS is that, in the presence of thick-tailed errors, it will downweight outliers in an optimal manner, flexibly adapting to the tail behavior of the sample through the nonparametric score estimator. If the regression disturbances are not i.i.d., but have an unconditional distribution whose thick tails are induced by conditional heteroskedasticity, then an extension of results of Hodgson (2000) to our model should be possible if the regressors are strictly exogenous and the disturbances are uncorrelated with an elliptical unconditional density. Our nonparametric score estimator will consistently estimate the score of this unconditional error density and the distribution theory outlined above should follow, with the standard errors being asymptotically correct.

4 Empirical CAPM Tests

4.1 Background

There is an extensive empirical literature on the CAPM, with important early work by Black, Jensen, and Scholes (1972) and Fama and MacBeth (1973).[6] More recent work has employed the multivariate regression model introduced above, for example Gibbons (1982) and Stambaugh (1982). Estimating the model for m portfolios over a sample of length n, we have:

$$r_t = \alpha + \beta r_{M,t} + u_t,\ t = 1, \ldots, n, \tag{7}$$

where r_t is the m-vector of portfolio excess returns, α and $\boldsymbol{\beta}$ are m-dimensional parameter vectors, $r_{M,t}$ is the excess market return, and u_t is an m-vector of disturbances. If there is some systematic component of returns that is not due to market risk exposure, it will appear in the intercept (α). If

[6]See Campbell, Lo, and MacKinlay (1997) for a more comprehensive discussion of empirical tests of the CAPM.

the CAPM holds, then $\alpha = 0$, but the existence of additional returns implies $\alpha \neq 0$. The following null hypothesis on the parameters of (7)

$$H_0 : \alpha_i = 0 \quad i = 1, \ldots, m, \tag{8}$$

implies that no significant excess returns are present that cannot be explained by variation in the market return. We test this hypothesis by constructing a standard Wald test

$$J = \tilde{\alpha}' \, [\widehat{\mathrm{var}}(\tilde{\alpha})]^{-1} \, \tilde{\alpha},$$

where $\tilde{\alpha}$ is an estimate of α and $\widehat{\mathrm{var}}(\tilde{\alpha})$ estimates the asymptotic covariance matrix of $\tilde{\alpha}$.[7] Alternatively, one could test for the significance of additional regressors in (7). For example, Basu (1977) considers price-earnings ratios, Banz (1981) includes market size, and Fama and French (1992, 1993) consider a firm's book value to market value ratio as well as size.

4.2 Elliptically Symmetric Returns: Adaptive Estimation and Tests

In applying our estimator, some care should be taken regarding the possibility of conditional heteroskedasticity in the regression disturbances. The CAPM is derived under the assumption of elliptical symmetry in asset *returns*, which implies that the disturbances may possess conditional heteroskedasticity and higher order dependence with the regressors. The presence of conditional heteroskedasticity implies that some problems exist with both OLS and our estimator. In the case of OLS, the standard errors will be biased (Van Praag and Wesselman (1989)). Our semiparametric estimator loses its adaptivity property if there is high order dependence between u_t and $r_{M,t}$. We discuss some remedies below.

First, a parametric model of the conditional heteroskedasticity can be introduced into the model and estimation procedure. The parameter vector (θ) can be expanded to include the conditional heteroskedasticity model and estimation can proceed as in the previous section. However, in the context of regression models where second moments appear in the mean equation via 'β' the distribution theory of the estimator is much more difficult. Hodgson and Vorkink (2001) develop a semiparametric estimator for a multivariate GARCH-in-mean model such as that of Bollerslev, Engle, and Wooldridge (1988).

A second solution would be to use a procedure as proposed by White (1980) and correct for the conditional heteroskedasticity in the preliminary estimation step. This should purge any high order dependence between $r_{M,t}$ and u_t allowing the estimation theory as discussed in the previous section to be valid. A model for the conditional heteroskedasticity is required at the preliminary estimation

[7] See MacKinlay (1987) and Gibbons, Ross, and Shanken (1989) for a discussion of CAPM tests along these lines.

stage. We could proceed nonparametrically, estimating the conditional variance of equation i, for $i = 1, \ldots, m$, by taking the squares of the residuals from the preliminary regression $u_{i,t}^2$ and using kernels locally regress them on the contemporaneous market excess return $(r_{M,t})$ as defined below:

$$\hat{\sigma}_i^2\left(r_{M,t}\right) = \frac{\sum_{i=1}^n K(\frac{r_{M,t}-r_{M,i}}{h_n})\hat{u}_i^2}{\sum_{i=1}^n K(\frac{r_{M,t}-r_{M,i}}{h_n})}, \tag{9}$$

where $K(.)$ is a kernel weighting function and h_n is a bandwidth. We could then use $r_{i,t}^w = \frac{r_t}{\hat{\sigma}_{i,t}}$ and estimate the model using $\left\{r_{i,t}^w\right\}_{t=1}^n$. Multivariate normality tests on the series $\{r_{i,t}^w\}$ find excess kurtosis to be present (see below), implying that normality assumptions are not appropriate even after accounting for conditional heteroskedasticity. Alternatively, one could proceed parametrically by specifying the GARCH(1,1) model:

$$\sigma_{i,t}^2 = \mathbf{a}_{i,0} + \mathbf{a}_{i,1}\sigma_{i,t}^2 + \mathbf{a}_{i,2}u_{i,t}^2.$$

One approach to correcting for the bias present in the OLS standard errors is to use information from the unconditional distribution to correct for the conditional heteroskedasticity. As was noted earlier, if second moments are allowed to vary then the unconditional distribution will be thick-tailed. The degree of kurtosis in the unconditional distribution can be used to adjust variances as described in Zhou (1993), who shows that a simple correction of the Wald statistic will generalize it to allow elliptical returns, as follows:

$$J^* = J \cdot \eta^{-1} \xrightarrow{D} \chi_N^2,$$

where J is the standard Wald statistic, $\eta = 1+\kappa_x/(m(m+2))$, and κ_x is Mardia's (1970) multivariate measure of kurtosis. Under multivariate normality, $\upsilon = 0$ and $J^* = J$. However, when excess kurtosis exists, $\upsilon > 1$ and $J^* < J$.

4.3 Results

We use daily data on stock returns taken from the CRSP files and running from January 1996 to December 1997.[8] We construct three portfolios by sorting firms according to size (market value). On each trading day firms are placed into quartiles according to the NYSE firm size. Daily value-weighted returns are then constructed for the firms in each of the first three quartiles.[9] Our use of daily data is a bit unusual, but the CAPM itself says nothing about the length of the return period, and the question as to how well daily returns are approximated by the mean-variance model seems to us to be of no less intrinsic interest than the same question applied to monthly or annual returns,

[8]Firms that are traded on the NYSE, NASDAQ and AMEX are included.

[9]We exclude the largest quartile because of its similarity to our measure of the market.

for example. Daily returns tend to be more highly non-normal than returns over longer intervals (although some degree of non-normality is present even there), suggesting that our econometric methodology is particularly well suited to this question.[10] Applications to weekly and monthly data will be pursued in future work.

Tables I and II provide the summary statistics for the risk-free rate (30 day T-bill rate) $r_{f,t}$, the annualized return on the CRSP value-weighted market portfolio $r_{M,t}$, and annualized portfolio excess returns $r_t - r_{f,t}$. Multivariate normality is rejected using either the univariate kurtosis estimates or the Jarque-Bera (1980) tests performed on the individual series reported in Table I. The multivariate measures of kurtosis also reject normality as seen in Panel A of Table II. Panel B shows that Beran's (1979) test of elliptical symmetry fails to reject at the 10% level the null that excess returns are elliptical. We also consider returns weighted by estimated conditional standard deviations. Normality is rejected on either set of returns, although those weighted nonparametrically have smaller kurtosis. In fact, for the size 3 portfolio, the Jarque-Bera test fails to reject. However, when we look at the multivariate tests, normality is strongly rejected while elliptical symmetry is not rejected, for both the parametric and nonparametric conditional heteroskedasticity estimates.[11]

Table III reports the results of estimating (7) using unweighted returns. The OLS estimates of β are positive and of α are close to zero (relative to their standard errors). The adaptive estimates are computed using a Gaussian kernel with Schuster's (1985) correction and the Box-Cox transformation $z = \tau(v) = \left(v^{\zeta} - 1\right)/\zeta$, with $\zeta = 1/2m$.[12] We choose our bandwidth parameter by using separate optimal MISE rule-of-thumb (Silverman (1986)) bandwidths for $\gamma(z)$ and $\gamma'(z)$, respectively. In general, we find that the point estimates of α (β) using the adaptive estimator are greater (lesser) than their OLS counterparts. Some of the differences in the point estimates are substantial. For example, the adaptive method estimates that the unexplained return in the size 1 portfolio returns will be at least 12% while the OLS estimates are about 5%. The difference in standard errors between the adaptive procedures and the Gaussian methods is substantial. The reduction is 15% on average for the adaptive estimates. These efficiency gains also appear in the simulation study reported below.

Tables IV and V report the results of estimating (7) using the nonparametric and the GARCH(1,1) weighted returns, respectively. As in the unweighted return regression, we obtain estimates of α

[10] Unconditional GMM estimation of the model would use the moment conditions $E[u_t] = 0$ and $E[u_t r_{M,t}] = 0$, leading to the OLS point estimates, with possibly different standard errors. The resulting Wald statistic would have a lack of power similar to that of the OLS Wald statistic, due to the sensitivity of the estimator to thick tails in return distributions.

[11] Bollerslev (1987) and Nelson (1991) also find significant nonnormalities in GARCH standardized distributions.

[12] This transformation provided good results in Monte Carlo experiments (not reported) and increases smoothing as the dimension increases, with the limit as $m \to \infty$ being the natural log transformation. This increased smoothing reduces the bias of the nonparametric estimate as the dimension increases.

that are larger, and of β that are smaller, with the adaptive estimation than with OLS. We also find that estimates of β are lower using the adaptive method relative to OLS. The results suggest a return model that places less weight on market variation and more weight on additional factors. Standard errors again decline using the adaptive estimator, with the reduction being 11% on average.

We report Wald statistics of the zero-intercept null in Table VI. For the unweighted returns we find that none of the estimation methods lead to a rejection at the 10% level, although the adaptive estimator has a p-value only slightly greater than 0.10, as opposed to the OLS the p-value of 0.89 (the latter includes the Zhou (1993) correction). The adaptive estimator suggests the presence of a size (market value) effect on returns. The inefficient OLS estimator lacks sufficient power to reject the model, even in the presence of weighted returns. Table VI shows that, for both sets of weighted returns, OLS cannot reject while the adaptive estimator strongly rejects, with p-values less than 0.01.

Our results suggest the use of an alternative model of returns. Our results lend support to the use of a multi-factor model, possibly incorporating the market size of a firm. We find that firms with small market capitalizations earn excess returns relative to the CAPM predictions. Our estimator measures these excess returns to be substantially higher than OLS estimates, with the difference as large as 7% annually for firms in the smallest quartile.

5 Simulation Analysis

We now investigate the finite sample size and power properties of the Wald tests computed with the different estimators. We employ Davidson and MacKinnon's (1998) graphic method of comparing p-value plots and power-size plots. A large number of realizations of a given test statistic J are computed from data sets generated under the null and under a specified alternative hypothesis. We label the Wald statistics computed from the adaptive estimator and OLS as J and J_{OLS} respectively.

Step 1: In generating our simulated data sets, each return $\check{r}_{i,t}$ is constructed by taking the product of the market return $r_{M,t}$ and the estimated beta $\hat{\beta}_i$, and adding a randomly selected residual from some prespecified distribution,

$$\check{r}_{i,t} = \alpha_h + \hat{\beta}_i r_{M,t} + \check{u}_{i,t}.$$

The distributions from which we draw $\check{u}_{i,t}$ are Student t with 3 degrees of freedom, two mixed normals, and normal. To compute the mixed normals, define the uniform random variable $U \in [0, 1]$. If $U < (1 - \epsilon)$, then let $\check{u}_t = \sqrt{\kappa_1} z_m$. Otherwise, we let $\check{u}_t = \sqrt{\kappa_2} z_m$. The resulting $\check{u}_t$ will follow a mixed normal distribution. We set $\epsilon = .8$, $\kappa_1 = 0.65\,(MN_1)$, or $0.45\,(MN_2)$ and $\kappa_2 = 6$ in the simulations.[13] The intercept is set to $\alpha_h = 0$ or $\alpha_h = .05$, the latter being the approximate average absolute intercept from the empirical data. We use the same residual in constructing both the alternative and null series.

Step 2: For both the null and alternative data sets estimate the above model and then compute J and the p-value under the null distribution for each statistic. The Gaussian kernel was used in constructing the adaptive estimates. The p-value for the statistic constructed using the alternative data set also uses the null distribution to obtain the p-value.

Step 3: Repeat Steps 1 and 2 many times. We chose to simulate the data and statistics 1,000 times which should provide reasonable accuracy in the p-values we report.

Step 4: Given the simulated test statistics and their associated p-values, then calculate the empirical distribution function of the p-values generated by each statistic. This is obtained in the following manner. Recall that the p-value of a statistic ϖ_j is the probability of observing a value of the statistic more extreme than ϖ_j. Let $\hat{F}(x_i)$ represent the estimate of the c.d.f. of the p-values generated by a given statistic at the point x_i and define $p(\varpi_j)$ to be the p-value associated with statistic ϖ_j. Then $\hat{F}(x_i)$ is calculated using the following formula:

$$\hat{F}(x_i) = \frac{1}{W} \sum_{j=1}^{W} I\left(p(\varpi_j) > x_i\right),$$

where W is the number of simulations and I is an indicator function that is equal to one if the argument is true and zero otherwise. To generate $\hat{F}(x_i)$, it is recommended that a grid of values lying in the interval between 0 and 1 be chosen to save time and computer storage space. We chose the grid (X) to be the following: $X = \{0.001, 0.002, 0.003, \ldots, 1\}$ and obtained the associated $\hat{F}(X)$ for all statistics under both the null and alternative.

Once the empirical distributions of each statistic are generated they can be graphed to compare the size and power properties of the statistics. To compare size properties the following graph, entitled a p-value plot, is recommended. The plot is constructed by graphing of X_i versus $\hat{F}(X_i)$ for each of the statistics. A test with appropriate size would follow the 45° since this is the c.d.f. of any p-value distribution. When the graph is above (below) the 45° line the associated statistic is over (under) rejecting the null hypothesis.

To compare the power of two given test statistics, Davidson and MacKinnon (1998) recommend the graph entitled power-size plot. The power-size plots graph $\hat{F}^a(X_i)$ against $\hat{F}^n(X_i)$ where these stand for the empirical distributions of the p-values from a test statistic under the alternative and null respectively. When this line is plotted for a competing statistics any deviant size properties are removed by graphing $\hat{F}^n(X_i)$ on the x-axis. Because the actual size is used as the x-variable,

[13] We scale the errors so that their variances match those in our empirical data.

differences in power cannot be attributed to differences in size between two competing statistics.

Size The simulations indicate that the tests constructed with our estimator are, in general, well-sized. However, this is not true in all cases. We list the results of our simulations in the p-value plots in Figures 1, 3, and 5. In some cases the method appears to be undersized (normality, $m = 4$; MN_2, $m = 4$). It appears that as the dimension increases the size of the adaptive tests declines. This could be due to our transformation choice and could potentially be corrected by fine-tuning our selection method.

Power Power results are reported in Figures 2, 4, 6, and 7. For all the simulations using thick-tailed distributions, the adaptive estimator leads to more frequent rejections than OLS with the increase as great as 79% in one case (MN_1, $m = 4$). The power appears to increase with dimension (m) as seen in the Figures 2, 4, and 6. The adaptive estimator is less powerful if the errors actually are normal.

A Appendix

A.1 Proof of Theorem 1

To prove the adaptivity of $\widetilde{\theta}$ we must establish the following two convergence results:

$$\widehat{\Delta}_n(\widehat{\theta}) - \Delta_n(\widehat{\theta}) \xrightarrow{P} 0, \tag{A.1}$$

and

$$\widehat{\mathcal{I}}(\widehat{\theta}) - \mathcal{I} \xrightarrow{P} 0, \tag{A.2}$$

where $\Delta_n(\widehat{\theta}) = -n^{-1}\sum_{t=1}^n w_t^T \varphi(\widehat{u}_t)$. We can use arguments analogous to those of Bickel (1982), Linton (1993, p. 566), or Jeganathan (1995) to show that these results will hold provided

$$\int |\widehat{\varphi}_t(u) - \varphi(u)|^2 p(u) du \xrightarrow{P} 0. \tag{A.3}$$

We can show that (A.3) is equivalent to

$$\int_0^\infty v^{m/2} \left\{ \frac{\widehat{g}_t'}{\widehat{g}_t}(v) - \frac{g'}{g}(v) \right\} g(v) dv \xrightarrow{P} 0. \tag{A.4}$$

The proof of equivalence makes use of the facts that: $p'(u) = 2(\det \Sigma^{-1/2}) g'(u^T\Sigma^{-1}u)\Sigma^{-1}u$, $f'(\varepsilon) = 2g'(\varepsilon^T\varepsilon)\varepsilon = (\det\Sigma)^{1/2}p'(u)$, and $\varphi(u) = p'(u)/p(u) = p'(\Sigma^{1/2}\varepsilon)/p(\Sigma^{1/2}\varepsilon) = \Sigma^{-1/2}f'(\varepsilon)/f(\varepsilon) \equiv \widetilde{\varphi}(\varepsilon)$.

We also note here that $\text{-}_p = \int \varphi(u)\varphi(u)^T p(u)du = \int \widetilde{\varphi}(\varepsilon)\widetilde{\varphi}(\varepsilon)^T(\det\Sigma)^{-1/2}f(\varepsilon)d\varepsilon$. Since we are not interested in using direct nonparametric estimates of $g(v)$, but rather of $\gamma(z)$, we must state the convergence result (A.4) in terms of γ, γ', and their estimates. To do so, first note that it is easily shown that the following relationship exists between the scores of γ and g: $(g'/g)(v) = s(v) + \tau'(v)(\gamma'/\gamma)(\tau(v))$. It follows that we can use our kernel estimate of the score of γ to nonparametrically estimate of the score of g as follows: $(\widehat{g}_t'/\widehat{g}_t)(v) = s(v) + \tau'(v)(\widehat{\gamma}_t'/\widehat{\gamma}_t)(\tau(v))$, so that $(\widehat{g}_t'/\widehat{g}_t)(v) - (g'/g)(v) = \tau'(v)\{(\widehat{\gamma}_t'/\widehat{\gamma}_t)(\tau(v)) - (\gamma'/\gamma)(\tau(v))$. These calculations allow us to characterize the restrictions we must place upon $\widehat{\gamma}_t'/\widehat{\gamma}_t$ in order to ensure the consistency of $\widehat{g}_t'/\widehat{g}_t$ and hence of $\widehat{\varphi}_t$. Now we can write

$$\begin{aligned}
\int_0^\infty v^{m/2}\left\{\frac{\widehat{g}_t'}{\widehat{g}_t}(v) - \frac{g'}{g}(v)\right\}^2 g(v)dv &= \int_0^\infty v^{m/2}\tau'(v)^2\left\{\frac{\widehat{\gamma}_t'}{\widehat{\gamma}_t}(\tau(v)) - \frac{\gamma'}{\gamma}(\tau(v))\right\}^2 g(v)dv \\
&= \int_0^\infty \alpha(v)\left\{\frac{\widehat{\gamma}_t'}{\widehat{\gamma}_t}(\tau(v)) - \frac{\gamma'}{\gamma}(\tau(v))\right\}^2 \gamma(\tau(v))dv,
\end{aligned}$$

where $\alpha(v) = v\tau'(v)^2 J_\tau^{-1}\{\tau(v)\}$. Since $z = \tau(v)$ and $\rho(z) = \alpha(\tau^{-1}(z))$, we can rewrite the right hand side of the preceding equation as

$$\int_{-\infty}^{\infty} \rho(z)\left\{\frac{\widehat{\gamma}_t'}{\widehat{\gamma}_t}(z) - \frac{\gamma'}{\gamma}(z)\right\}^2 \gamma(z)dz. \tag{A.5}$$

Using the trimmed kernel estimator of γ'/γ described in Section 3 of the main text, we have now established that our whole argument hinges on showing that, under our specified trimming conditions, the integral in (A.5) converges to zero. We show below that the key assumption we must make is that the information of the density being estimated here be finite, i.e., that

$$\int \rho(z)\frac{[\gamma']^2}{\gamma}(z)dz < \infty. \tag{A.6}$$

Unfortunately, this inequality is stated in terms of the transformed random variable z and its density γ. We would like to know what this inequality implies in terms of primitive conditions on the density f (or, equivalently, g). Specifically, assuming that we are using a particular transformation τ, what conditions must f (or g) satisfy in order for this inequality to hold? It can be shown that (A.6) is implied by the moment conditions in the statement of the Theorem. As noted in the remark to the Theorem, the condition,

$$\int_0^\infty v^{m/2}s(v)^2 g(v) < \infty, \tag{A.7}$$

depends on our selection of a transformation τ, so that certain transformations may require us to place stronger moment conditions on our data generating process than others.

These results provide conditions under which the score of the error density in a multivariate model can be consistently estimated. We can then use standard methods (see Bickel (1982), Kreiss (1987), Linton (1993), Jeganathan (1995), etc.) to show that these error density score estimates can be used to consistently estimate the overall score for the model, the information matrix of the error density, and the information matrix of the model.

PROOF THAT (A.6) IS IMPLIED BY CONDITIONS OF THEOREM. The assumption that $p(u)$ has finite information is equivalent to assuming that $f(\varepsilon)$ has finite information, i.e., that $\int \left|\frac{f'}{f}(\varepsilon)\right|^2 f(\varepsilon)d\varepsilon < \infty$ so that $\int_0^\infty v^{m/2}\frac{[g']^2}{g}(v)dv < \infty$. The left hand side of (A.6) is $\int \rho(z)\frac{[\gamma']^2}{\gamma}(z)dz = \int_0^\infty \alpha(v)\frac{[\gamma']^2}{\gamma}\{\tau(v)\}dv$. We would like to express the right hand side of this equation as an integral in $\frac{[g']^2}{g}(v)dv$. To do so, note that

$$\frac{[\gamma']^2}{\gamma} = \left(\frac{\gamma'}{\gamma}\right)^2\gamma = \left\{\tau'(v)^{-1}\frac{g'}{g}(v) - \tau'(v)^{-1}s(v)\right\}^2 \left\{v^{m/2-1}J_\tau\{\tau(v)\}g(v)\right\}.$$

Our problem therefore reduces to deriving the conditions under which

$$\int_0^\infty \alpha(v)\left\{\tau'(v)^{-1}\frac{g'}{g}(v) - \tau'(v)^{-1}s(v)\right\}^2 \left[v^{m/2-1}J_\tau\{\tau(v)\}g(v)\right]dv < \infty.$$

But the left hand side of this inequality equals

$$\int_0^\infty v^{m/2}\frac{[g']^2}{g}(v)dv + \int_0^\infty \left\{v^{m/2}s(v)^2 - 2\frac{g'}{g}(v)s(v)v^{m/2}\right\}g(v)dv.$$

That this term is finite is a direct consequence of the assumptions of the Theorem, completing the proof.

■

We now show that, under our assumptions, the trimmed kernel estimator introduced in Section 3 satisfies

$$\int_{-\infty}^{\infty}\rho(z)\left\{\frac{\widehat{\gamma}_t'}{\widehat{\gamma}_t}(z) - \frac{\gamma'}{\gamma}(z)\right\}^2\gamma(z)dz \xrightarrow{P} 0 \qquad \text{(A.8)}$$

This will complete our proof of the Theorem. Our proof of (A.8) will follow the pattern of Lemma 4.1 of Bickel (1982), modifying it where necessary and using different conditions where necessary, to account for the difference between this model and his.

The following conditions are satisfied under our assumptions:

CONDITION A. (1) $\int \rho(z)\frac{[\gamma']^2}{\gamma}(z)dz < \infty$; (2) $\{z : |\lambda(z)| = \infty\}$ has Lebesgue measure zero, where $\lambda(z)$ is the anti-derivative of $\rho^{1/2}(z)$; (3) For all $\varepsilon > 0$, there exists $\mu > 0$ such that $\Pr\left\{\rho^{1/2}(z) > \mu\right\} < \varepsilon$.

REMARK. We have shown that Condition A(1) is a consequence of the moment conditions in the statement of the Theorem. Conditions A(2) and A(3) depend on the transformation $\tau(\cdot)$ and can be shown to be automatically satisfied for all Box-Cox transformations.

Our basic result is the following

Lemma 2 *Under the above Condition A,*

$$\int \left\{ q_t(z) - \frac{\rho^{1/2}(z)\gamma_t'(z)}{\gamma_t(z)} \right\}^2 \gamma_\sigma(z)dz \xrightarrow{P} 0.$$

PROOF. Let $\gamma_h(z) = (K_h * \gamma)(z)$ and $\gamma_h'(z) = (K_h * \gamma')(z)$, where $*$ denotes convolution, i.e., $(g * f)(z) = \int g(x)f(z-x)dx$. The pattern is similar to that of Lemma 6.1 in Bickel (1982), except that his equations (6.8) and (6.9) become

$$I_1 = \int_{ABCD} \rho(z) \left\{ \frac{\widehat{\gamma}_t'}{\widehat{\gamma}_t}(z) - \frac{\gamma_h'}{\gamma_h}(z) \right\}^2 \gamma_h(z)dz \tag{A.9}$$

$$I_2 = \int_{(ABCD)^c} \rho(z)\frac{[\gamma_h']^2}{\gamma_h}(z)dz \tag{A.10}$$

where $ABCD$ is the set where no trimming occurs.

So

$$E(I_1) \leq 2\left[\int_{ABCD} \rho(z)\gamma_h^{-1}(z)E\left[\widehat{\gamma}_t'(z) - \gamma_h'(z)\right]^2 dz + \int_{ABCD} c_n^2\gamma_h^{-1}(z)E\left[\widehat{\gamma}_t(z) - \gamma_h(z)\right]^2 dz\right] = o(1).$$

The second element of this sum is $o(1)$ exactly as in Bickel (1982). For the first element, things are different. This is where our new trimming condition (iii) comes in. The first term is less than or equal to

$$\int_{ABCD} \rho(z)\gamma_h^{-1}(z)\kappa_1 h^{-3}n^{-1}\gamma_h(z)dz = \int_{ABCD} \rho(z)\kappa_1 h^{-3}n^{-1}dz.$$

Since $|\lambda(z)| \leq b_n$, this expression is $o(1)$ because $b_n h_n^{-3} = o(n)$.

Now consider I_2. We have

$$E(I_2) \leq \int \rho(z)\frac{[\gamma_h']^2}{\gamma_h}(z)\left[\begin{array}{c} \Pr\left\{\left|\rho(z)^{1/2}\widehat{\gamma}_t'(z)\right| > c_n\widehat{\gamma}_t(z)\right\} + \Pr\left\{\widehat{\gamma}_t(z) < d_n, \gamma(z) > 0\right\} \\ +I\left[|z| > e_n\right] + I\left[|\lambda(z)| > b_n\right] \end{array}\right] dz.$$

The proof that $E(I_2) \xrightarrow{P} 0$ is modified little from Bickel's, except that we use Condition A(3) to ensure that the first probability in this expression converges to zero, and Condition A(2) to ensure that the second indicator function is equal to zero in the limit almost everywhere. One other modification is that we must show that $\int \rho(z) \frac{[\gamma_h']^2}{\gamma_h}(z) dz < \infty$. We can show that this holds for the class of transformations τ described in the main text due to our assumption that $\int \frac{\rho(z)\gamma'(z)^2}{\gamma(z)} dz < \infty$. ∎

Lemmas 6.2 and 6.3 of Bickel (1982) can be applied to our model to complete the proof of the Theorem.

∎

References

[1] Banz, R. 1981. The relation between return and market value of common stocks. *Journal of Financial Economics* 9:3-18.

[2] Basu, S. 1977. The investment performance of common stocks in relation the their price to earnings ratios: a test of the efficient market hypothesis. *Journal of Finance* 32:663-682.

[3] Beran, R. 1979. Testing for ellipsoidal symmetry of a multivariate density. *Annals of Statistics* 7:150-162.

[4] Berk, J. 1997. Necessary conditions for the CAPM. *Journal of Economic Theory* 73:245-257.

[5] Bickel, P.J. 1982. On adaptive estimation. *Annals of Statistics* 10:647-671.

[6] Black, F., M. Jensen, and M. Scholes. 1972. The capital asset pricing model: some empirical tests. In Jensen, M. (ed), *Studies in the theory of Capital Markets*, New York, Praeger.

[7] Bollerslev, T. 1987. A conditional heteroskedastic time series model for speculative prices and rates of returns. *Review of Economic and Statistics,* 69: 542-547.

[8] Bollerslev, T., R. Engle, and J. Wooldridge. 1988. A capital asset pricing model with time varying covariances. *Journal of Political Economy* 96: 116-131.

[9] Box, G. and D. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* 211-264.

[10] Campbell, J., W. Lo, and A. C. MacKinlay. 1997. *The Econometrics of Financial Markets*. Princeton: Princeton University Press.

[11] Casella, G. and R.L. Berger. 1990. *Statistical Inference*. Belmont, CA: Duxbury Press.

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

[13] Davidson, R. and J., MacKinnon. 1998. Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School* 66, 1-26.

[14] Fama, E. 1963. Mandelbrot and the stable Paretian hypothesis. *Journal of Business* 36: 420-429.

[15] Fama, E. 1965. The behaviour of stock market prices. *Journal of Business* 38:34-105.

[16] Fama, E., and K. French. 1992. The cross-section of expected returns. *Journal of Finance* 47:427-465.

[17] Fama, E., and K. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3-56.

[18] Fama, E., and J. MacBeth. 1973. Risk, return, and equilibrium: empirical tests. *Journal of Political Economy* 71:607-636.

[19] Fernández, C., J. Osiewalski, and M.F.J. Steel. 1995. Modelling and inference with v-spherical distributions. *Journal of the American Statistical Association* 90:1331-1340.

[20] Gibbons, M.R. 1982. Multivariate tests of financial models: A new approach. *Journal of Financial Economics*. 10:3-27.

[21] Gibbons, M., S. Ross, and J. Shanken. 1989. A test of the efficiency of a given portfolio. *Econometrica* 57:1121-1152.

[22] Härdle, W., and O.B. Linton. 1994. Applied nonparametric methods. In D.F. McFadden and R.F. Engle III (eds.), *The Handbook of Econometrics*, Vol. IV, pp. 2295-2339, North Holland.

[23] Hodgson, D.J. 1998a. Adaptive estimation of cointegrating regressions with ARMA errors. *Journal of Econometrics* 85:231-268.

[24] Hodgson, D.J. 1998b. Adaptive estimation of error correction models. *Econometric Theory* 14:44-69.

[25] Hodgson, D.J. 2000. Partial maximum likelihood and adaptive estimation in the presence of conditional heterogeneity of unknown form. *Econometric Reviews* 19:175-206.

[26] Hodgson, D.J. and K. Vorkink. 2001. Efficient estimation of conditional asset pricing models. University of Rochester.

[27] Horowitz, J. 2000. Estimation of a generalized additive model with unknown link function. Forthcoming, *Econometrica*.

[28] Hsieh, D.A. and C.F. Manski. 1987. Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Annals of Statistics* 15:541-551.

[29] Ingersoll, J. 1987. *Theory of Financial Decision Making*. Totowa, NJ: Rowan & Littlefield.

[30] Jarque, C.M. and A.K. Bera. 1980. Efficient tests for normality, heteroskedasticity, and serial independence of regression residuals. *Economics Letters* 6: 255-259.

[31] Jeganathan, P. 1995. Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* 11:818-887.

[32] Kelker, D. 1970. Distribution theory of spherical distributions and a location-scale generalization. *Sankhya A* 32:419-430.

[33] Kreiss, J.-P. 1987. On adaptive estimation in stationary ARMA processes. *Annals of Statistics* 15:112-133.

[34] Lintner, J. 1965. The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13-37.

[35] Linton, O. 1993. Adaptive estimation in ARCH models. *Econometric Theory* 9:539-569.

[36] Linton, O. 1995. Second order approximations in a partially linear regression model. *Econometrica* 63:1079-1113.

[37] MacKinlay, A. C. 1987. On multivariate tests of the CAPM. *Journal of Financial Economics* 18:342-372.

[38] Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36:394-419.

[39] Mardia, K.V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57:519-530.

[40] Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 12:77-91.

[41] Markowitz, H. 1959. *Portfolio Selection: Efficient Diversification of Investment.* New York: Wiley.

[42] Mitchell, A.F.S. 1989. The information matrix, skewness tensor and α-connections for the general multivariate elliptic distribution. *Annals of the Institute of Mathematical Statistics* (Tokyo) 41:289-304.

[43] Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34:768-783.

[44] Muirhead, R.J. 1982. *Aspects of Multivariate Statistical Theory.* New York: Wiley.

[45] Nelson, D. 1991. Conditional heteroskedasticity in asset returns: A new approach, *Econometrica.* 59: 347-370.

[46] 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.

[47] Phillips, P.C.B., J.W. McFarland, and P.C. McMahon. 1996. Robust tests of forward exchange market efficiency with empirical evidence from the 1920's. *Journal of Applied Econometrics* 11:1-22.

[48] Schick, A. 1987. A note on the construction of asymptotically linear estimators. *Journal of Statistical Planning and Inference* 16:89-105.

[49] Schuster, E.F. 1985. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods* 14:1123-1136.

[50] Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19:425-442.

[51] Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

[52] 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.

[53] Stone, C. 1975. Adaptive maximum likelihood estimation of a location parameter. *Annals of Statistics* 3:267-284.

[54] Stute, W. and Werner, U. 1991. Nonparametric estimation of elliptically contoured densities. In Roussas, G. (ed.), *Nonparametric Functional Estimation and Related Topics*, Kluwer Academic Publishers, pp. 173-190.

[55] Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25:65-86.

[56] Van Praag B. and A. Wesselman, 1987. Elliptical regression operationalized. *Economics Letters* 23:269-274.

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

[58] Zhou, G. (1993). Asset pricing tests under alternative distributions. *Journal of Finance*, 48:1927-1942.

Table I

Summary Statistics

J-B refers to the Jarque-Bera (1980) test for normality. $^{*}(^{\diamond})$Refers to a rejection of the hypothesis that the given moment is consistent with the Normal distribution at the .01 (.05) level.

Variable	Mean	Std. Dev.	min	max	Kurtosis	J-B
r_M	0.340	3.126	-23.776	14.707	35.442*	1326*
r_f	0.073	0.001	0.069	0.077	-0.598	2
Unweighted Excess Returns						
Size 1	0.1987	2.318	-21.155	9.955	73.614*	5799*
Size 2	0.233	2.725	-22.137	10.545	46.933*	2381*
Size 3	0.246	2.755	-22.637	8.553	45.945*	2293*
Nonparametric Conditional Std. Dev. Weighted Portfolio Excess Returns						
Size 1	0.201	0.840	-3.542	4.065	9.069*	82*
Size 2	0.204	0.939	-2.615	4.062	3.556*	12*
Size 3	0.324	1.537	-3.849	5.514	2.780$^{\diamond}$	2
GARCH(1,1) Conditional Std. Dev. Weighted Portfolio Excess Returns						
Size 1	0.173	1.372	-10.458	3.904	42.923*	2111*
Size 2	0.169	1.647	-13.932	4.713	50.660*	2826*
Size 3	0.204	2.176	-18.456	5.729	49.633*	2699*

Table II

Multivariate Tests of Normality and Elliptical Symmetry

The test statistics below are Mardia's (1970) multivariate kurtosis measure and

*Indicates a p-value less than .01. [1]Indicates a p-value greater than .1.

Panel A: Multivariate Kurtosis Test

	Size Portfolios
Unweighted Excess Returns	44.20*
Nonparametric Conditional Std. Dev. Weighted Excess Returns	16.34*
GARCH(1,1) Conditional Std. Dev. Weighted Excess Returns	33.69*

Panel B: Elliptical Symmetric Test (S_n)

	Size Portfolios
Unweighted Excess Returns	-0.866[1]
Nonparametric Conditional Std. Dev. Weighted Excess Returns	0.796[1]
GARCH(1,1) Conditional Std. Dev. Weighted Excess Returns	0.823[1]

Table III

Results of Estimation of CAPM

$$r_t = \alpha + \beta r_{M,t} + u_t$$

Panel A: OLS

	α		β_{im}	
Portfolio	Estimate	Std. Error	Estimate	Std. Error
Size 1	0.0530	0.0710	0.5418	0.0227
Size 2	0.0428	0.0707	0.7109	0.0226
Size 3	0.0351	0.0551	0.7889	0.0176

Panel B: Adaptive

	α		β_{im}	
	Estimate	Std. Error	Estimate	Std. Error
Size 1	0.1271	0.0596	0.4709	0.0190
Size 2	0.0760	0.0582	0.6472	0.0186
Size 3	0.0547	0.0447	0.7581	0.0142

Table IV

Results of Estimation of CAPM

Nonparametric Conditional Variance Weighted Returns

$$r_t^w = \alpha + \beta r_{M,t}^w + u_t$$

Panel A: OLS

	α		β_{im}	
Portfolio	Estimate	Std. Error	Estimate	Std. Error
Size 1	0.0781	0.0457	0.4723	0.0250
Size 2	0.0559	0.0458	0.6624	0.0259
Size 3	0.0559	0.0459	0.7623	0.0203

Panel B: Adaptive

	α		β_{im}	
	Estimate	Std. Error	Estimate	Std. Error
Size 1	0.1258	0.0416	0.4117	0.0228
Size 2	0.0751	0.0412	0.5900	0.0234
Size 3	0.0651	0.0412	0.7264	0.0183

Table V

Results of Estimation of CAPM

GARCH(1,1) Conditional Std. Dev. Weighted Returns

$$r_t^w = \alpha + \beta r_{M,t}^w + u_t$$

Panel A: OLS

	α		β_{im}	
Portfolio	Estimate	Std. Error	Estimate	Std. Error
Size 1	0.0655	0.0448	0.4678	0.0203
Size 2	0.0232	0.0449	0.6547	0.0211
Size 3	-0.0051	0.0450	0.7593	0.0169

Panel B: Adaptive

	α		β_{im}	
	Estimate	Std. Error	Estimate	Std. Error
Size 1	0.0806	0.0399	0.4172	0.0183
Size 2	0.0041	0.0409	0.6035	0.0194
Size 3	-0.0247	0.0397	0.7390	0.0151

Table VI

Mean-Variance Efficiency Tests

$$H_0: \quad \alpha_i = 0 \;\; i = 1, \ldots, m.$$

Under the null, J is distributed asymptotically $\chi^2(3)$. *We adjust this Jstatistic by the factor discussed in section 4.2 and Zhou (1993). P-values are in parentheses following the test statistics.

		J (p-value)
Unweighted Excess Returns		
	OLS*	0.52(0.92)
	Adaptive	5.68(0.12)
Nonparametric Weighted Excess Returns		
	OLS	3.30(0.35)
	Adaptive	11.05(0.01)
GARCH(1,1) Weighted Excess Returns		
	OLS	5.45(0.148)
	Adaptive	13.82(0.00)

Figure Information