TESTING THE CAPITAL ASSET PRICING MODEL EFFICIENTLY UNDER ELLIPTICAL SYMMETRY: A SEMIPARAMETRIC APPROACH*

by

Douglas J Hodgson
University of Rochester, NY

Oliver Linton
London School of Economics and Political Science

Keith Vorkink
Brigham Young University, Provo, UT

Contents:

The Suntory Centre
Suntory and Toyota International Centres for Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020-7405 7686

Discussion Paper
No.EM/00/398
July 2000

* We would like to thank Pedro Gozalo and Gene Savin for helpful comments, Eugene Choo and Hyungsik Moon for research assistance and the National Science Foundation for financial support under CAREER grant SBR-9701959. Hodgson thanks CREST at INSEE for their hospitality while part of this research was being carried out.

Abstract

We develop new tests of the capital asset pricing model which 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. The elliptical symmetry assumption allows us to avoid 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; efficiency.
JEL Nos.: C14, C24, C13, C22.

1 Introduction

The capital asset pricing model (CAPM) posits that the expected excess return of any asset is linear in its covariance with the expected return on the market portfolio, see Sharpe (1964) and Lintner (1965).[1] This relationship is formalized in the following equation:

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

where $\beta_i = \text{cov}[R_i, R_M] / \text{var}[R_M]$, R_M is the 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 strongly restricting preferences.[2] A random variable u is elliptically symmetrically distributed if its density $p(u)$ can be written in the following fashion:

$$p(u) = (\det \Sigma)^{-1/2} g(u^T \Sigma^{-1} u), \tag{2}$$

for some function $g(\cdot)$ and positive definite, symmetric matrix Σ. Berk (1997) shows that elliptical symmetry is 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.

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

[2] See also Ingersoll (1987).

Suppose that we have the 'market model regression' in stacked form

$$r_t = \alpha + \beta r_{M,t} + u_t, \tag{3}$$

where r_t, α, β, and u_t are m-vectors of, respectively, excess returns on m portfolios of assets, intercept parameters, beta parameters, and regression disturbances. The regressor in each equation is the excess return on some measure of the market portfolio. Time series observations on the asset returns will be used to estimate the parameters of this system. The CAPM theory implies that the intercept vector α is a vector of zeros. This formulation of the CAPM hypothesis is employed by MacKinley (1987) and Gibbons, Ross, and Shanken (1989). The obvious approach to testing this null hypothesis is to estimate the unrestricted model (3) by ordinary least squares (OLS) and construct a Wald test using the point estimates of α and their estimated standard errors.

However, although such a test is valid, it may not be very powerful, since OLS is only a fully efficient estimator under a normality assumption on the errors[3]. As mentioned above, this assumption may not be a very good one, due to the presence of thick tails, which suggests that even more efficient estimates of (3) than OLS, and more powerful Wald tests, can be obtained by estimating the model by maximum likelihood for some non-Gaussian, thick-tailed likelihood. The problem then arises of specifying a parametric functional form for the likelihood function. In recent years, this problem has been addressed through the development of a new class of estimation methods based on approximating the unknown error distribution by estimates obtained through nonparametric smoothing. Semiparametric methods allow one to obtain robust and efficient estimators even in the absence of such parametric assumptions as multivariate normality of the errors. These semiparametric methods 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). The basic idea underlying such estimation procedures is to first estimate the model by some consistent method, such as OLS. The OLS residuals are then used to form a nonparametric kernel estimate of the unknown density of the disturbances. This estimate is then used as the basis of a fully efficient, "adaptive" estimator, that will be asymptotically equivalent to the maximum likelihood estimator. This estimator can then be used to form Wald tests, which will be more powerful than the Wald test formed using OLS. 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 (2000).

The problem alluded to in the preceding paragraph is often referred to as the "curse of dimen-

[3]OLS will be equivalent to generalized least squares (GLS) in this model since the regressor vector is identical across the equations of the system.

sionality" and is of particular relevance to our problem of efficiently estimating (3). See Silverman (1986, page 94) for a dramatic illustration of the effects of dimensionality on estimating a normal density at the origin. Although the semiparametric theory says that asymptotically these effects disappear when the properties of the parameter estimates are being considered, in even quite large samples they do not. If the number of assets in our system is at all large, then the application of the semiparametric methods referred to in the preceding paragraph becomes problematic and it would seem that we should fall back on parametric approaches such as OLS. However, if we exploit the elliptical symmetry assumption underlying the CAPM, then we have the opportunity to avert the curse of dimensionality. 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 of the advances in econometric estimation theory alluded to above, 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. As we can see from (2), the m-dimensional error distribution $p(u)$ is proportional to the *one*-dimensional function $g(\cdot)$. The implication is that we can obtain a nonparametric estimate of the former through the nonparametric estimation of the latter, which, being a one-dimensional nonparametric estimation problem, is not subject to the curse of dimensionality. This intuition is shown to be correct by Stute and Werner (1991).

The econometric contribution of the present paper is to develop adaptive estimators in a semiparametric linear seemingly unrelated regression (SUR) model, allowing for cointegrating regressions as well as standard stationary regressions, in which the error density is of unknown form. To overcome the curse of dimensionality, we focus on the restriction that the multivariate density is elliptically symmetric. Elliptical symmetry is also important for a number of statistical reasons, which makes our work transferable to a number of other problems. It allows for leptokurtic marginals, and, as mentioned above, a number of important thick-tailed distributions belong to this family [see Table 3.1 of Fang, Kotz, and Ng (1990)].[4] Similar semiparametric models have been explored previously in Bickel (1982), Jeganathan (1995) and Hodgson (1998a). These authors defined adaptive estimators of the identifiable parameters in various regression models. However, their proposed estimates do not exploit the dimensionality reduction implied by elliptical symmetry and consequently suffer serious "small sample" costs. What is required here is estimation of a multidimensional density function and its first derivative.

We find that accounting for the tail thickness present in daily stock returns in the nonparametric

[4]See Fernández, Osiewalski, and Steel (1995) for some generalizations of elliptical symmetry that are interesting from a statistical point of view.

manner described above will indeed have effects on our inference in the CAPM. Our results are discussed in detail in Section 5, but our basic finding is that the stock returns as modelled in (3) can be explained less by market returns and more by factors not accounted for in the basic model than may be suggested by OLS estimation. In particular, our semiparametric estimates yield estimates of β that are generally smaller, and estimates of α that are generally larger, than those obtained by OLS. These results obtain even when possible conditional heteroskedasticity in the regression disturbances is modelled nonparametrically or through use of a GARCH model - the implication being that thick tails are present even when we account for conditional heteroskedasticity and that our semiparametric adaptive methods will still pick up residual non-Gaussianity and therefore lead to improved inference. We find, in fact, that rejections of the CAPM in thick-tailed daily data are more likely to be obtained using the efficient estimator than they are when we use the inefficient OLS as the basis of a test statistic.

In Section 2, we introduce the cointegrated and non-cointegrated SUR models that we are interested in analyzing. Our application to the capital asset pricing model (CAPM) only involves the non-cointegrated model, but developing the theory for cointegrated models comes at little extra cost and has possible applications, such as forward unbiasedness tests [Phillips, McFarland and McMahon (1996)]. In Section 3, we outline a formula for computing an adaptive estimator under our assumptions. Section 4 discusses some further issues including boundary correction, bandwidth choice, transformation choice, and Beran's (1979) test for elliptical symmetry, Section 5 reports the results of our empirical analysis of the CAPM, while Section 6 investigates the performance of the estimator through a Monte Carlo simulation analysis. A mathematical appendix contains proofs. We use $\|A\| = \left(\mathrm{tr}A^TA\right)^{1/2}$ to denote the Euclidean norm of a vector or matrix A, while $\xrightarrow{P}$ denotes convergence in probability and $\Rightarrow$ signifies weak convergence of probability measures. We say that $X \sim MN(0,V)$ when X is mixed normal with (possibly) random covariance matrix V.

2 Models

In this section we consider the specification and estimation of a general seemingly unrelated regressions (SUR) model. The CAPM regression (3) falls within this class, but we consider a general SUR specification, also allowing for the possibility of cointegrating regressions. This is because the methodological contribution of the paper involves the introduction a new technique for the efficient estimation of such models, and the range of possible applications of the estimator is quite broad.

Consider the m-equation seemingly unrelated regression model

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

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 $x_{it}^T \in \mathbb{R}^{k_i}$ 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 may be either integrated of order one (I(1)) or stationary and ergodic. In either case, we assume that x_t and u_t are independent (i.e. that the regressors are strictly exogenous). In the stationary case, we assume that x_t has finite second moment, and in the I(1) case we assume that the spectral density function of Δx_t is finite and positive definite when evaluated at the origin. When the regressors are I(1), each of the m regressions is cointegrating. When the regressors are stationary, the asymptotic properties of least squares estimators are standard. Some variations on the basic model are of interest. Firstly, the regression matrix θ could be subject to nonlinear cross equation restrictions. This can be expressed by writing $\theta(\eta)$, for some vector η of deep parameters. Secondly, an important class of models for macroeconomics are stationary VAR's in which the regressor matrix is a lagged dependent variable.

We suppose that the error has a Lebesgue density $p(u)$ that is absolutely continuous with respect to Lebesgue measure. 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, Ω_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 $L_n(\theta)$. We define the weighting matrix δ_n, where $\delta_n = n^{-1/2} I_{m+k}$ if x_t are stationary and $\delta_n = \text{diag}\left[n^{-1/2} I_m, n^{-1} I_k\right]$ if x_t are integrated. These structures for δ_n are associated with the fact that the rate of consistency of estimators in non-cointegrated models is $n^{1/2}$, whereas in cointegrating regressions it is $n^{1/2}$ for intercept parameters and n for slope parameters. 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 δ_n^{-1}-consistent estimator $\widehat{\theta}$ that was obtained from the Gaussian likelihood. This approach to estimation apparently originates with R.A. Fisher, see Robinson (1988), and has been widely used in econometrics following Rothenberg and Leenders (1965). Under general conditions, this will be first order asymptotically equivalent to the maximum likelihood estimator (MLE), i.e.,

$$\delta_n^{-1}(\overline{\theta} - \theta_0) \Rightarrow MN(0, \mathcal{I}^{-1}),$$

where the asymptotic information matrix $\mathcal{I}$ is such that $\delta_n \left(\partial^2 L_n\left(\theta_0\right) / \partial\theta \partial\theta'\right) \delta_n \Rightarrow \mathcal{I}$. 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 $\Omega_p = \int \varphi(u)\varphi(u)^T p(u) du$, the information matrix of p. For the stationary model, the asymptotic information matrix is

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

while for the cointegrated model, it is

$$\mathcal{I} = \begin{bmatrix} \Omega_p & \Omega_p \int_0^1 M(r) dr \\ \int_0^1 M(r)^T dr \Omega_p & \int_0^1 M(r)^T \Omega_p M(r) dr \end{bmatrix},$$

where

$$M(r) = \begin{bmatrix} M_1^T(r) & & & 0 \\ & M_2^T(r) & & \\ & & \ddots & \\ 0 & & & M_m^T(r) \end{bmatrix}$$

and $M_i(r)$ is a k_i-dimensional Brownian motion with covariance matrix equal to the long run covariance matrix of Δx_{it}, for every $i = 1, \ldots, m$. Note that in the case of cointegration, $\mathcal{I}$ is random, hence the mixed normal limit theory.

We also 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

$$\widetilde{\theta} = \widehat{\theta} + \delta_n \widehat{\mathcal{I}}_n^{-1}(\widehat{\theta}) \widehat{\Delta}_n(\widehat{\theta}), \tag{4}$$

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 4 below. In particular,

$$\widehat{\Delta}_n(\widehat{\theta}) = -\delta_n \sum_{t=1}^{n} w_t' \widehat{\varphi}_t(\widehat{u}_t),$$

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.[5]

The value of using such an estimator rather then OLS is that it will downweight outlying observations in a robust and optimal manner, an important consideration when we are facing data that are drawn from a thick-tailed density, i.e. one that has high outlier probability. To illustrate the effect of choosing the score function φ, note that computing a maximum likelihood or pseudo-maximum likelihood estimator essentially involves choosing θ to set the score function $\Delta_n(\theta) = -\delta_n \sum_{t=1}^{n} w_t'\varphi(u_t)$ equal to zero. If we incorrectly specify a Gaussian likelihood, then the score $\varphi(u)$ will be proportional to u. Now, if the true density of u has thick tails, then with nonnegligible probability, we will obtain very large realizations of u ("large" according to any norm), which imply very large realizations of $\varphi(u)$, the presence of which will tend to distort the sum defining our estimator, and so distort the estimator. If thick tails are a potential problem, then one would prefer to use a more robust estimator, i.e. an estimator in which the function $\varphi(u)$ is not blowing up as u increases. For example, if a Student's distribution with κ degrees of freedom is specified, then $\varphi(u)$ will be proportional to $u/\left(1+\kappa^{-1}u^Tu\right)$, and if a logistic distribution is specified then it will be proportional to $u\left(1+2\exp(-u^Tu)\right)$. Arbitrary specification of such a pseudo-likelihood will lead to a downweighting of outliers that will produce a robust, but not necessarily optimal, estimator.

The attraction of our semiparametric approach is that our nonparametric score function $\widehat{\varphi}(u)$ is not chosen arbitrarily but is chosen based on a kernel estimate of the density of the disturbances. Thus it will downweight outliers in a way that reacts to the actual tail thickness exhibited in the empirical distribution of the data, and that does so in a manner that is optimal in the sense of delivering an asymptotically efficient estimator. Now, if the innovation density is from a thicker-tailed family than the Gaussian, then any finite sample will tend to exhibit more outliers (i.e. data points for which u^Tu is large) than it would if the distribution were Gaussian. This higher outlier frequency in the data will lead us to compute a nonparametric estimate of the error density that has thicker tails than

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

a Gaussian (the tails of the density estimate will be behaving more and more like those of the true density as our sample size increases). Consequently, the semiparametric estimator of our regression parameters that makes use of this nonparametric density estimate will tend to downweight outliers in a way approximating the downweighting that occurs with the correctly specified maximum likelihood estimator, the quality of the approximation increasing with sample size.

3 Estimation

The formula for an adaptive estimator given in (4) 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. From Muirhead (1982), the density of v, which we shall denote $h(v)$, is

$$h(v) = c_m v^{m/2-1} g(v),$$

where $c_m = \pi^{m/2}/\Gamma(m/2)$.

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, given 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 $J_\tau(z) = |\partial\tau^{-1}(z)/\partial z|$. 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:

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 $\widehat{v}_t = \widehat{\varepsilon}_t^T\widehat{\varepsilon}_t$.

2. Denoting by $K_{h_n}(\cdot)$ a kernel with bandwidth h_n, form the following estimates of the density of $\widehat{z}_t$ and its first derivative:

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

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 $\lambda(z) = (d/dz)^{-1}\rho^{1/2}(z)$.[6] Then estimate the score and information of $p(u)$ as follows:

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

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

4. Then define the score and information estimators for the model as

$$\widehat{\Delta}_n(\widehat{\theta}) = -\delta_n\sum_{t=1}^{n} w_t^T\widehat{\varphi}_t(\widehat{u}_t) \quad ; \quad \widehat{\mathcal{I}}_n(\widehat{\theta}) = \delta_n\sum_{t=1}^{n} w_t^T\widehat{\Omega}_p w_t\delta_n. \tag{5}$$

With these definitions, we can compute the adaptive estimator $\widetilde{\theta}$ given in (4) 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 Ω_p is finite and positive definite, that $\int_0^\infty v^{m/2}s(v)^2g(v)dv < \infty$, that the Lebesgue density $p(u)$ is absolutely continuous with respect to Lebesgue measure, that the regressors*

[6] 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.

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,

$$\delta_n^{-1}(\widetilde{\theta} - \theta) \Rightarrow MN(0, \mathcal{I}^{-1}), \tag{6}$$

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 $\zeta = 1/2m$, the condition implies $E[\left(\varepsilon^T \varepsilon\right)^{m/2-2}] < \infty$. 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 (5) is a consistent estimator of the asymptotic covariance matrix, so that $\widehat{\mathcal{I}}_n(\widehat{\theta}) - \mathcal{I} = o_p(1)$. This result is true even for cointegrated models, in which case $\mathcal{I}$ is random. 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 $\theta_\ell = r$, where r is some constant. Then we can compute the usual t-ratio, as follows:

$$\frac{\left(\delta_n^{-1}\right)_{\ell\ell}\left(\widetilde{\theta}_\ell - r\right)}{\sqrt{\left(\widehat{\mathcal{I}}_n^{-1}(\widehat{\theta})\right)_{\ell\ell}}} \Longrightarrow N(0,1)$$

under the null, where $\left(\delta_n^{-1}\right)_{\ell\ell}$ and $\left(\widehat{\mathcal{I}}_n^{-1}(\widehat{\theta})\right)_{\ell\ell}$ are the ℓ^{th} elements along the diagonals of δ_n^{-1} and $\widehat{\mathcal{I}}_n^{-1}(\widehat{\theta})$, respectively. If we want to test the joint hypothesis $\theta = r$ for the entire vector θ, where r is now a known $(m+k)-$vector of constants, we can compute the Wald statistic

$$\left[\delta_n^{-1}(\widetilde{\theta} - r)\right]' \widehat{\mathcal{I}}_n(\widehat{\theta}) \left[\delta_n^{-1}(\widetilde{\theta} - r)\right] \Longrightarrow \chi^2_{m+k}.$$

Note that these convergence results will hold regardless of whether the model is stationary or cointegrated.

(c) It is natural to ask how the present estimator will behave if the thick tails in the unconditional density of the errors are induced by some sort of conditional dependence, such as a multivariate GARCH model. A related question has been addressed in Hodgson (2000) within the context of adaptively estimating univariate time series regression models, and the following conjectures are based on Hodgson's (2000) findings. It should be possible to extend these findings to obtain a useful robustness result for our estimator in the case where the error process $\{u_t\}$ is uncorrelated but not necessarily

independent over time, and has an *unconditional* density which is elliptically symmetric. This would happen, for example, if the errors followed a multivariate GARCH process, had a conditional density which was elliptically symmetric, and had a conditional covariance matrix whose magnitude changed over time but whose covariance structure remained unchanged. In any event, if the unconditional density is elliptically symmetric, then the nonparametric score and information estimators $\widehat{\varphi}$ and $\widehat{\Omega}$ described above and used in our computation of the adaptive estimator should still consistently estimate the score and information of the unconditional density of the errors. Our one-step estimator will then have the same asymptotic distribution as the one-step iterative pseudo-MLE based on the true unconditional density of the errors. When the regressors are strictly exogenous, as we have assumed above, then the resulting estimator will have an asymptotic distribution which is identical to that which it would have if the i.i.d. assumption on the errors was correct. In other words, the distribution depends only on the unconditional density of the errors and is completely invariant to the presence of conditional heteroskedasticity. Furthermore, the standard error estimates and test statistics described in the preceding remark will be robust to the presence of conditional heteroskedasticity. When the strict exogeneity assumption on the regressors is relaxed, this robustness property no longer holds. It is still true that our one-step semiparametric estimator will have the same distribution as the one-step fully parametric estimator based on the true unconditional density, but it will now be the case that the latter estimator's asymptotic covariance matrix will have the "sandwich" structure characteristic of pseudo-MLE's in misspecified models (c.f. White (1982)). To construct robust standard errors in this case, we would require a consistent nonparametric estimator of the Hessian of the innovation density, since both the Hessian and OPG versions of the information will enter the asymptotic covariance matrix. The derivation of such a consistent Hessian estimator has not yet been considered in the literature and is a topic for future research.

4 Implementation Issues

In this section, we discuss some issues that arise in the implementation of our estimator. In subsection 1 we discuss a 'degrees of freedom' correction to our estimator of the information matrix; this improves the standard errors considerably. Subsection 2 describes Schuster's (1985) correction, a modification of our basic estimator which is undertaken to correct for the poor properties of the nonparametric density estimator in the neighbourhood of the origin which is due to the fact that we do not directly estimate the density of the innovations u_t but rather estimate the density of a transformation z_t whose support is only on the positive portion of the real line (see Stute and Werner (1991), for a discussion of this problem, known as the "volcano effect"). In subsection 3, we discuss the issue of bandwidth selection, and in subsection 4 we discuss the issue of transformation choice. Lastly, in

subsection 5 we discuss issues arising in the implementation of a test for elliptical symmetry due to Beran (1979).

4.1 Degrees of Freedom Correction

Our estimator of the information matrix, although consistent, has a finite sample upwards bias that therefore biases downwards our standard error estimates. We propose a simple degrees of freedom correction that generalizes the usual correction made to parametric standard errors. For example, we defined the estimator $\widehat{\Sigma}_u = (n-k-m)^{-1}\sum_{t=1}^{n}\widehat{u}_t\widehat{u}_t^T$ instead of $\widehat{\Sigma}_u = n^{-1}\sum_{t=1}^{n}\widehat{u}_t\widehat{u}_t^T$ following this practice. 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 $(\widehat{\gamma}'_t)^2/(\widehat{\gamma}_t)^2$ in (6) by

$$\frac{(\widehat{\gamma}'_t)^2 - \sum_s(\omega'_{nts})^2}{(\widehat{\gamma}_t)^2 - \sum_s(\omega_{nts})^2}. \tag{7}$$

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) for a discussion of this issue in semiparametric estimation.

4.2 Schuster's correction

The construction of $\widehat{\varphi}$ imposing elliptical symmetry uses one dimensional kernel estimates of the transformed variable z. For the no transformed specification where $z = \varepsilon^T\varepsilon$, the support will have the restriction $z \geq 0$. This additional information is not incorporated in the standard Parzen-Rosenblatt kernel estimator, $f_n(z) = n^{-1}h_n^{-1}\sum_{i=1}^{n} K((z-z_i)/h_n)$, which generates a downward bias in the density estimate at this boundary. For most standard choices of symmetric kernel, the density estimator $f_n(z)$ typically performs poorly on the right neighbourhood of zero. This bias arise because for points x_i in the right neighbourhood of 0, the contribution of x_i given by $n^{-1}h_n^{-1}K((x-x_i)/h_n)$to $f_n(x)$ extends to points $x \leq 0$ where $f(x) = 0$. A similar bias arise in the multivariate density estimates which imposes the elliptical symmetry restriction. This bias creates a volcano like contour in the density estimate. The overflow in weights beyond the lower support of 0 can be corrected by using an estimator which incorporates this additional support constraint information into $f_n(x)$.

Schuster (1985) 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.

4.3 Bandwidth Selection

The smoothing parameter used in the kernel estimation is chosen by rule-of thumb (ROT) methods generalizing the standard approach for density estimation suggested in Silverman (1986). In the first case we choose bandwidth to minimize a weighted mean integrated square error (MISE) of the kernel estimate of $\gamma(z)$, the density of the transformed random variable $z = \tau(\varepsilon^T \varepsilon)$, assuming that the underlying density is normal.[7] Of course, the true density is unknown, but the ROT provides an easy-to-compute benchmark. The ROT bandwidth will depend on the transformation used. We calculate the optimal bandwidth h_{opt} according to the MISE formula for the density γ of the transformed variable z. We assumed that $\varepsilon \sim N(0, I_m)$ and calculated the implied density function γ. We use the Gaussian kernel $K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$. The scoring equation used in our procedure requires that we estimate the ratio γ'/γ of the transformed variable z. The kernel estimate of the derivative of the density $\widehat{\gamma}'$ in the numerator will have a convergence rate which is slower compared to that of the density estimate $\widehat{\gamma}$ in the denominator. This slower convergence rate is likely to dominate in the estimate of this ratio. The second approach in parameterizing h_{opt} involves calculating the smoothing parameter that minimizes the weighted mean integrated squared error of the kernel estimate of the derivative of the density $\gamma'(z)$. This alternative criterion is appropriate given that the slower rate of convergence in the derivative estimate is likely to dominate that of the density estimate.

4.4 Choice of Transformation

The transformation choice ($z = \tau(v)$) is another input that merits discussion. There is a growing literature in statistics on using transformations to estimate univariate densities, see for example Wand, Marron, and Ruppert (1991). The situation here is slightly different because our original data are high dimensional and the object of ultimate interest is the parameter estimates, and so we have not followed precisely their recipes. We have experimented with many different transformations and have found that the interaction between our bandwidth choice method and the transformation is quite crucial. Recall that our transformation takes the form $z = (v^\zeta - 1)/\zeta$ with the problem being to choose ζ.

An obvious choice would be to set $\zeta = 1$, in which case the transformation is proportional to the identity. Above, we mention the setting $\zeta = m/2$ as a possible choice. To see the problem that such settings may cause, consider the example where the errors are independent standard normals. Then the distribution of the variable z is that of the χ^2_m, a distribution that changes, and, in particular, that becomes more widely dispersed, as the dimension of the system (m) increases. As the dimension, and

[7] We have used chi-squared density to do the weighting.

therefore the dispersion, increases, the ROT bandwidth increases. Therefore, the bias of our density estimate increases, and does so quite dramatically. The problem becomes even more severe when we choose ζ to be increasing in m. To counteract this effect, we suggest choosing a transformation yielding a random variable z whose dispersion is not increasing rapidly in m. One transformation that has such a property and that behaves well in Monte Carlo simulations is the logarithmic ($\zeta = 0$). There are many other possible transformations that will achieve a similar effect. We have arbitrarily chosen a number of such transformations and have arrived at the particular transformation choice used in our empirical analysis by running some Monte Carlo simulations over the range of dimensions used in our empirical analysis and choosing the transformation that worked best. Our choice turned out to be $\zeta = 1/2m$, which provides good results in a Monte Carlo exercise across different dimensions. Note that as the dimension increases with the limit ($m \to \infty$) we approach the log transformation.

4.5 Test for elliptical symmetry

In this section we describe a test for elliptical symmetry developed by Beran (1979) and discuss its implementation. Suppose we have a series of standardized regression residuals $\widehat{\varepsilon}_t = \widehat{\Sigma}^{-1/2}\widehat{u}_t$ for $t = 1, \ldots, n$, where $\widehat{u}_t$ are OLS residuals and $\widehat{\Sigma}$ is a consistent estimator of Σ. These residuals will be used in the construction of a test of the null hypothesis that the true underlying innovations $\{\varepsilon_t\}$ are i.i.d. draws from a spherically symmetric density. The test utilizes a couple of distinctive features of spherically symmetric random variables. The first is that the standardized random variable $\varepsilon_t / \|\varepsilon_t\|$ is uniformly distributed on the $m-1$-dimensional unit hypersphere. The second is that this standardized random vector, which we refer to as the "direction" of ε_t, is independent of the vector's "length", viz. $\|\varepsilon_t\|$.

We now describe the construction of the test, provide some intuition as to how it incorporates the aforementioned characteristics, and then state the test's asymptotic distribution under the null. We begin by ranking the distances $\{\|\widehat{\varepsilon}_t\|\}_{t=1}^n$ and dividing these ranks by $n+1$. Let $\{R_t\}_{t=1}^n$ denote these ranks. Note that the directional vector $\varepsilon / \|\varepsilon\|$ can be represented in terms of its polar coordinates $\Xi = (\xi_1, \ldots, \xi_{m-1})$ as follows:

$$\frac{\varepsilon}{\|\varepsilon\|} = \left(\cos(\xi_1), \sin(\xi_1)\cos(\xi_2), \ldots, \sin(\xi_1)\sin(\xi_2), \ldots, \sin\left(\xi_{m-1}\right)\right).$$

Define the coordinates of $\widehat{\varepsilon}_t / \|\widehat{\varepsilon}_t\|$ by Ξ_t. Let $\{a_k : k \geq 1\}$ be the family of functions orthonormal with respect to the Lebesgue measure on $[0,1]$ and orthogonal to the constant function on $[0,1]$. Furthermore, let $\{b_\ell : \ell \geq 1\}$ denote another family of orthonormal functions with respect to the uniform measure on $[0,\pi]^{p-2} \times [0, 2\pi)$ and orthogonal to the constant function on this domain [we use Legendre's polynomials as described in the Appendix]. Beran (1979) proposed a statistic of the

form

$$S_n = \sum_{k=1}^{K_n} \sum_{\ell=1}^{L_n} \left[\frac{1}{\sqrt{n}} \sum_{t=1}^{n} a_k\left(R_t\right) b_\ell\left(\Xi_t\right) \right]^2 .$$

If the innovations $\{\varepsilon_t\}$ have a spherically symmetric distribution, then S_n should be close to zero; otherwise, it should be far from zero. Using the fact that, under the null, R_t and Ξ_t both have uniform distributions, it follows from our assumptions on $\{a_k\}$ and $\{b_\ell\}$ that $E\left[a_k\right] = E\left[b_\ell\right] = 0$ for all k, ℓ. The independence of R_t and Ξ_t under the null furthermore implies that $E\left[a_k\left(R_t\right) b_\ell\left(\Xi_t\right)\right] = 0$ for all k, ℓ. The following proposition gives the asymptotic distribution of S_n.

Proposition 2 (Beran (1979)) . *Suppose that the functions* $\{a_k : k \geq 1\}$ *and* $\{b_\ell : \ell \geq 1\}$ *are differentiable and that:*

1. $\lim_{n\to\infty} K_n = \lim_{n\to\infty} L_n = \infty$
2. $\lim_{n\to\infty} n^{-1/2} K_n^{-1/2} L_n^{1/2} \sum_{k=1}^{K_n} \|a_k'\| = \lim_{n\to\infty} n^{-1/2} K_n^{1/2} L_n^{-1/2} \sum_{\ell=1}^{L_n} \|b_\ell'\| = 0$
3. $\lim_{n\to\infty} \frac{1}{n} L_n K_n = 0.$

Then, the null limiting distribution of $\left(2L_nK_n\right)^{-\frac{1}{2}}\left[S_n - L_nK_n\right]$ *as* $n \to \infty$ *is* $N(0,1)$.

This test of Beran (1979) is based on Fourier series expansion density estimators. Considering the fact that our estimation routine employs *kernel*-based density estimation, it would seem natural to employ a kernel-based test for elliptical symmetry. We are not immediately aware of the existence of such a test, although it would presumably be feasible to develop one, perhaps employing results on kernel-based goodness-of-fit tests for density functions as developed by Fan (1994).

5 Empirical CAPM Tests

5.1 Background

Many econometric tests of the CAPM were published shortly after the development of the theory and have consistently found their way into the finance literature ever since.[8] Early empirical work seemed to support the CAPM, see Black, Jensen, and Scholes (1972) and Fama and MacBeth (1973). The

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

primary methodology used in these early works was to perform cross-sectional regressions of mean returns on estimated betas (which were estimated from some preliminary time series regressions) and other putative variables and thus to test the linearity restriction of the theory. The main econometric problem with this approach is the errors in variables problem that arises from the first stage regressions; one approach to this was to group stocks together into portfolios thereby reducing the estimation error. By grouping according to some factor that might also affect returns, like size, one can improve the power of the test. Most modern tests of the CAPM applications have been based on the multivariate regression model, see for example Gibbons (1982) and Stambaugh (1982).

We group our data into a number of portfolios (m) according to size. Suppose that

$$r_{i,t} = \alpha_i + \beta_i r_{M,t} + u_{i,t}, \tag{8}$$

where $r_{i,t}$ represents the actual excess return for portfolio i in period t, and $r_{M,t}$ is the excess return on some 'proxy' market portfolio. We estimate this model for the m portfolios over a sample period of $t = 1, \ldots, T$. Rewriting (8) in vector form we obtain:

$$r_t = \alpha + \beta r_{M,t} + u_t \tag{9}$$

where r_t, α, $\boldsymbol{\beta}$, and u_t are all $m \times 1$ vectors. If there is some common source of return for the stocks within a given portfolio that is not due to market risk exposure, then that source of return will be found in the intercept (α). If the CAPM holds, then $\alpha = 0$, but the existence of additional returns implies $\alpha \neq 0$. The following null hypothesis on the parameters of (9)

$$H_0 : \alpha_i = 0 \quad i = 1, \ldots, m, \tag{10}$$

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

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

where $\tilde{\alpha}$ is our estimate of α, and $\widehat{\text{var}}(\tilde{\alpha})$ is an estimate of the asymptotic covariance matrix variance of $\tilde{\alpha}$. If this statistic deviates significantly from zero, we conclude that the CAPM does not fully explain the variations in returns.[9]

While early empirical work failed to reject the null hypothesis, subsequent work began to question its validity. For example, Basu (1977) found a relationship between price-earnings ratios and stock returns, while Banz (1981) found variation in stock returns due to market size that could not be explained by market variation. More recently, Fama and French (1992, 1993) found a relationship

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

between a firm's book value to market value ratio and stock returns as well as confirmed the Banz (1981) size anomaly.

Most empirical testing of the CAPM is based on using monthly returns while applications of the CAPM are found using many different horizons.[10] For example, event studies make extensive use of the CAPM. In most event studies daily data is used due to the importance of pinpointing the exact timing of some specific information arrival to the market. In these studies typically a market model of the CAPM is estimated using daily data and then using estimated parameters of the model abnormal returns are constructed and used in the event study. However, as stated earlier, it is generally believed that tests of the CAPM are best done using monthly returns. In the case of daily data, it is often argued that daily data are contaminated with the presence of market microstructure effects such as nonsynchronous trading which may bias the inferences of the hypothesis testing. An alternative explanation for this concentration on monthly returns may come from the observation that excess kurtosis is substantially greater in daily and weekly returns relative to monthly returns. Consequently, the power of Gaussian techniques will be much greater on monthly return data sets relative to daily returns. Our interest is to asses the validity of the CAPM on a data set of daily returns. We find that our methods provide sufficient power to reject the model on our data set while Gaussian methods fail to reject.

5.2 Elliptically Symmetric Returns: Adaptive Estimation and Tests

Our estimator in the previous section made the assumption that the residuals followed an elliptically symmetric distribution. However, to obtain the mean variance results discussed earlier returns, not residuals, are assumed to follow elliptically symmetric distributions. Assuming elliptically symmetric returns has implications that substantially differentiate this assumption from assuming thick-tailed residual distributions. Foremost, this assumption implies that conditional heteroskedasticity may exist. To see this, assume that $(r_t, r_{M,t})$ are jointly elliptically distributed with mean (μ) and variance (V)

$$u = \begin{pmatrix} \mu_r \\ \mu_M \end{pmatrix}, \quad V = \begin{pmatrix} V_{rr} & V_{rM} \\ V_{rM} & V_{MM} \end{pmatrix}.$$

Then the conditional variance of r_t on $r_{M,t}$ is as follows:

$$\text{var}\left(r_t | r_{M,t}\right) = k\left(r_{M,t}\right)\left(V_{rr} - V_{rM}V_{MM}^{-1}V_{rM}\right)$$

where $k\left(r_{M,t}\right)$ is some function of the quadratic $\left(r_{M,t} - \mu_M\right)' V_{MM}^{-1}\left(r_{M,t} - \mu_M\right)$ and is time varying for distributions within the elliptically symmetric family other than normality. The presence of

[10] One of the problems of the traditional CAPM is that no direction is given as to horizon with which the relation should hold. See Vorkink (1999) for a treatment of our estimation procedure on a data set of monthly returns.

conditional heteroskedasticity implies that some problems exist with OLS estimation methods as well as our proposed method. In the case of OLS, it will cause the standard errors of the OLS estimates to be biased, as was shown in Van Praag and Wessleman (1989). In the case of our estimator, the adaptive properties of our estimator will break down resulting from the possible high order dependence of the u_t on $r_{M,t}$. The abundant evidence supporting the presence of conditional heteroskedasticity in stock returns makes this issue even more relevant. 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 (2000) develop an estimator and theory for estimation in this setting.

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 stage. Under the assumption of elliptical symmetry this conditional heteroskedasticity should take the form

$$\text{var}\left(r_t|r_{M,t}\right) = k\left(r_{M,t}\right)\left(V_{rr} - V_{rM}V_{MM}^{-1}V_{rM}\right) \tag{11}$$

which we note is a function of the market return via the function $k\left(r_{M,t}\right)$. Because the functional form of $k\left(r_{M,t}\right)$ is not known we suggest nonparametrically estimating the conditional variance of r_t. To estimate this function we take 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:

$$\hat{\sigma}_i^2\left(r_{M,t}\right) = \frac{\sum_{i=1}^{n} K(\frac{r_{M,t}-r_{M,i}}{h_n})\hat{u}_i^2}{\sum_{i=1}^{n} K(\frac{r_{M,t}-r_{M,i}}{h_n})} \tag{12}$$

where $K(.)$ is a kernel weighting function as defined previously and h_n is the bandwidth. After $\hat{\sigma}_{i,t}$ are obtained we constructed the variables $r_{i,t}^w = \frac{r_t}{\hat{\sigma}_{i,t}}$ and then proceeded to estimate the given return model using the series $\left\{r_{i,t}^w\right\}_{t=1}^{T}$. We also performed tests of multivariate normality on the series of $r_{i,t}^w$ and found that kurtosis levels were inconsistent with assumptions of normality implying that kurtosis in returns is generated by a combination of time-varying second moments and thick-tailed time invariant distributions.

As an alternative weighting scheme we chose to estimate a GARCH model of conditional variance for the portfolio returns. These models are quite prevalent and have been found to be parsimonious

models that capture most features of the conditional variances in stock returns. We also provide results where returns are weighted by the estimated conditional variances of a Gaussian GARCH(1,1) as follows:

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

Lastly, for the case of the OLS estimates, one approach to correct for the bias present in the 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). He proposed a correction for the documented specification problems of standard asset pricing tests that follow from the regression estimation framework. He shows that a simple multiplicative correction to the J statistic will generalize that test statistic to the assumption of elliptically symmetric return distributions. The correction is a follows:

$$J^* = J * \eta^{-1} \overset{a}{\sim} \chi_N^2 \tag{2}$$

with

$$\eta = 1 + \upsilon, \quad \text{where } \upsilon = \frac{\kappa_x}{N(N+2)},$$

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 $J^* < J$ implying that correcting for elliptical symmetry will lead to fewer rejections of a pricing model than assuming normality.

5.3 Results

Our data set consists of returns taken from the CRSP data set of stock returns and includes daily observations from January 1996 through December 1997. We construct three portfolios by sorting firms according to size (market value), and test the CAPM on this set of portfolios.[11] On each trading day firms are placed into quartiles according to the NYSE firm size quartiles. Daily value-weighted returns are then constructed for the firms in each of the first three quartiles.[12] The returns on these three portfolios are then used to test the CAPM in the framework discussed above. As mentioned, there is strong evidence that returns are related to size which cannot be explained by the CAPM. However, we find that Gaussian methods do not have the power to reject the CAPM on these size sorted portfolios, whereas in other studies using monthly returns rejections do occur.

[11] Firms that are traded on the NYSE, NASDAQ and AMEX are included in this data set.

[12] We exclude the largest quartile from our exercise because of its similarity to our measure of the market portfolio return.

This provides an interesting environment for applying our adaptive estimation method, since it is possible that the associated efficiency gains may lead to a rejection of the CAPM on this data set.

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 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. However, an application of Beran's (1979) test of elliptical symmetry fails to reject the hypothesis that the excess returns are distributed elliptically symmetric at the 10% level as seen in Panel B of Table II.

We also include statistics for our two sets of weighted returns recalling that these excess returns that have been scaled by estimated conditional standard deviations. Hypothesis tests of normality are rejected on either set of returns although the returns weighted by the nonparametric estimate of the conditional standard deviation have smaller estimates of kurtosis. In fact, for the size 3 portfolio the Jarque-Bera test of normality fails to reject the hypothesis on this individual portfolio's return. However, when we look at the multivariate tests normality is strongly rejected while elliptical symmetry is not rejected. For the GARCH weighted returns the multivariate tests in Table II also lead to a rejection of normality and a failure to reject elliptical symmetry. These results imply that the nonnormalities in the unconditional distributions are not completely driven by time-varying second moments and perhaps that our nonparametric model does the better job of capturing the conditional heteroskedasticity. However, even after correcting for conditional heteroskedasticity, it would appear that efficiency gains are possible over Gaussian estimation.[13]

Table III reports the results of estimating (9) using both OLS and the adaptive methods using the unweighted returns. The OLS estimates are consistent with the empirical literature in that the estimates of β are positive and the estimates of α are close to zero relative to their standard errors. Panels B and C report the results of the adaptive estimations. Recall, we use the Box-Cox transformation with $\zeta = 1/2m$ in construction of the adaptive estimates, and use separate optimal MISE bandwidth parameters for estimating $\gamma(z)$ and $\gamma'(z)$. Panel B of Table III provides results of the adaptive estimates where the Gaussian kernel with Schuster's correction is used and Panel C of Table III provides the results of the adaptive estimates where the Gaussian kernel is used without Schuster's correction. 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%. Economically, this difference of 7% annual return is quite large. These differences may be driven by

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

outliers that influence the OLS point estimates substantially, but are downweighted in the adaptive procedure.

The adaptive estimates using either of the two kernels are nearly identical. The Gaussian kernel with Schuster's correction leads to slightly larger standard errors but the differences generally do not show up until the third digit. Given the little difference between the results using Schuster's correction and those that did not we conclude that the "volcano effect" has little influence in our example. 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. This reduction supports the properties of our estimator given the nonnormalities of the data. These efficiency gains are also supported by the simulation study reported below.

Tables IV and V report the results of estimating (9) using both OLS and the adaptive methods using the nonparametric conditional standard deviation weighted returns and the GARCH(1,1) weighted returns respectively. Weighting the returns implies that the estimated parameters do not have the same interpretation as in the case of the unweighted returns. However, we do find similar patterns to those in the unweighted return estimations. In general, the estimates of α are larger using the adaptive estimation procedure relative to OLS. We also find that estimates of β are lower using the adaptive method relative to OLS. With the larger intercepts and smaller β estimates, the adaptive method seems to imply that OLS attributes too large a portion of the portfolio return to market risk exposure and that a more correct model would reduce the market exposure (β) and increase the unexplained return (α). While this result is not consistent in all of our tested cases it does describe the general difference between the two methods.

Again, standard errors are smaller using the adaptive method with the reduction being 11% on average. Nonnormalities in the conditional return distributions, as evidenced in Tables I and II, allow the adaptive procedure to improve the efficiency of estimates relative to the Gaussian procedure.

Of primary interest is testing to see if the CAPM is consistent with our portfolio of returns. We list Wald test statistics which test the validity of the CAPM in Table VI. For the unweighted returns we find that none of the estimation methods lead to a rejection of the CAPM at the 10% level. However, the statistic constructed using either of the adaptive estimators nearly rejects the model as the p-value associated with the test statistic is slightly above the 0.10 level. These are substantially lower than the p-value associated with the OLS statistic of 0.89. The listed Wald test for the OLS estimates includes the correction for conditional heteroskedasticity as proposed by Zhou (1993). In conclusion, the adaptive estimator leads to a conclusion that is more consistent with the literature regarding the size (market value) effect on returns of greater horizons. Gaussian methods, while able to reject the model using monthly data, fail to have sufficient power to reject the model in our daily data set. We find that for the weighted returns these differences become even larger.

Table VI also reports mean-variance efficiency tests for the two sets of weighted returns. For both sets of weighted returns OLS cannot reject the CAPM while the adaptive methods strongly reject the model with p-values less than .01.

Our investigation into the validity of the CAPM on a stock returns with daily horizons led us to some interesting conclusions. First, we find that our method allows a rejection of the CAPM on our data set while Gaussian methods fail to reject the model. One may conclude that either additional risks are present in the daily data or that a disequilibrium model may best describe returns. Regardless, the CAPM is not consistent with stock returns even at daily horizons. Going back to our example of event studies, these results provide impetus for the use of some alternative model in the construction of abnormal returns. Some have previously argued using multi-factor models in connection with event study yet their arguments were somewhat loosely motivated by the results in the monthly return literature. Our results lend support to this notion, that risk in addition to market risk are present in stock returns even at short horizons, particularly as related to the market size of a firm. Consistent with the monthly literature, we find a small firm effect, or that firms with small market capitalizations earn excess returns relative to the CAPM predictions. Our method measures these excess returns to be substantially higher than Gaussian estimates, with the difference as large as 7% annually for firms in the smallest quartile.

Secondly, we find that Gaussian techniques tend to estimate exposure to systematic risk (β) for a given portfolio to be higher than the estimates using the adaptive method. This result was strikingly consistent across our data sets of unweighted returns and returns weighted by estimates of the conditional standard deviation.

6 Simulation Analysis

We also perform a Monte Carlo exercise to investigate the empirical properties of the estimator in a controlled environment. Davidson and MacKinnon (1994) provide a 'cookbook' graphical method of comparing simulation results through comparison of p-value plots and power-size plots. This method is an alternative to reporting results in standard table format, which provide only a few points on the distribution function and can be more difficult to interpret. This method also more easily indicates how sample size, degrees of freedom and other factors affect the performance of a given test statistic. In addition to the graphs, we also provide a few tables with some key simulation results.

To obtain the graphs we conduct a Monte Carlo experiment in which a large number of realizations of given test statistics J are generated using data under the null, and also data under some specified alternative distribution. We construct J from estimations using our adaptive estimator and from estimations using OLS. We label these statistics J and J_{OLS} respectively. We discuss the specifics of

the comparison below and refer the reader to Davidson and MacKinnon (1994) for a more complete discussion.

Step 1: Simulate data. We use the CAPM model to simulate the data for the portfolio regressions. For a given simulation the null and alternative data sets are constructed in the following manner. 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,

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

We chose three residual distributions from which to randomly draw $\breve{u}_{i,t}$: $t_{(3)}$, a mixed normal distribution, and a normal distribution. These were chosen to vary the degree of kurtosis found in the residuals. To construct errors from a multivariate t distribution we draw $z_m \sim N(0, I_m)$, where z is an $m \times 1$ vector and I_n is an $m \times m$ identity matrix. We also draw $c \sim \chi^2_v$ independent of z_m and let $\breve{u}_t = z_m / \left(\frac{c}{v}\right)^{\frac{1}{2}}$ which follows a multivariate t distribution with degree of freedom parameter v. To construct errors from a mixed normal distribution we draw U from a uniform distribution over $[0, 1]$ and again $z_m \sim N(0, I_m)$ as defined previously. If $U < (1 - \epsilon)$, then let $\breve{u}_t = \sqrt{\kappa_1} z_m$. Otherwise, we let $\breve{u}_t = \sqrt{\kappa_2} z_m$. The resulting $\breve{u}_t$ 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.[14]

We then add α_h, where $\alpha_h = 0$ under the null and draw $\alpha_h = .05$ to generate data under the alternative. The values for α_h were chosen by finding an approximate average of absolute intercepts from the CAPM estimations. We use the same residual in constructing both the alternative and null series because this method helps to isolate the results from randomness introduced by the simulations.

Step 2: For both the null and alternative data sets estimate the above model and then using the estimates construct J and the p-value under the null distribution for each statistic. We note that only the Gaussian kernel was used in constructing the adaptive estimates. We also note that 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

[14] We also scale the residual innovations with a factor that adjusts the second moments to be equivalent to the second moments of GLS residuals of our empirical exercise.

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 cdf 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 (1994) 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, differences in power result cannot be attributed to differences in size between two competing statistics.

6.1 Size Results

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 Table VII as well as the P-value plots in Figures 1, 3, and 5. In some cases the method appears to be undersized (Normality, $m = 4$; MN_2, $m = 4$). The trend appears that as the dimension increases the size of the adaptive tests declines. This could be a problem of our transformation choice and could potentially be corrected by fine tuning our selection method. These problems with size are not extremely large and the promising results we found in our simulations testing power seem to overshadow those here.

6.2 Power Results

The simulations testing power indicate that our estimator is picking up efficiency gains when kurtosis is present. Our results of these simulations are found in Table VIII as well as Figures 2, 4, 6, and 7.

For all the simulations using leptokurtokic distributions, we find that the adaptive procedure leads to a rejection of the alternative more often than OLS with the increase in rejection 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. One poor result is the simulations using the normal distributions where the power of the procedure is significantly lower than OLS indicating a strong dependence of our procedure on the presence of multivariate kurtosis.

In general, we find the simulation exercise to be promising regarding the properties of our estimator. We do suggest future simulation work that could help to fine tune the transformation selection as well as further tests of higher dimension and other nonnormal distributions.

A Appendix

A.1 Orthonormal Polynomials

The orthonormal polynomials used for our set of functions arise from solutions to what is called Legendre's differential equations of the form $(1-x^2)y'' - 2xy' + n(n+1)y = 0$. The general solution is given by $y = c_1 P_n(x) + c_2 Q_n(x)$, where the polynomials used are given by

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n$$

for some $x \in [-1, 1]$. These polynomials are suitably transformed to ensure orthogonality in the appropriate domain specified by the test statistics.

The family of differentiable orthonormal polynomials $\{a(z)_k : k \geq 1\}$ on the $[0, 1]$ domain for z are as follow:

$$\begin{aligned}
a(z)_1 &= \sqrt{3}(2z-1) \\
a(z)_2 &= \sqrt{5}(6z^2 - 6z + 1) \\
a(z)_3 &= \sqrt{7}\left(2z-1\right)\left(10z^2 - 10z + 1\right) \\
a(z)_4 &= \sqrt{9}(70z^4 - 140z^3 + 90z^2 - 20z + 1) \\
a(z)_5 &= \sqrt{\frac{11}{64}}\left(63\left(2z-1\right)^5 - 70\left(2z-1\right)^3 + 15\left(2z-1\right)\right) \\
a(z)_6 &= \frac{\sqrt{13}}{16}\left(231\left(2z-1\right)^6 - 315\left(2z-1\right)^4 + 105\left(2z-1\right)^2 - 5\right) \\
a(z)_7 &= \frac{\sqrt{15}}{16}\left(429\left(2z-1\right)^7 - 693\left(2z-1\right)^5 + 315\left(2z-1\right)^3 - 35\left(2z-1\right)\right).
\end{aligned}$$

The family of orthonormal polynomials $\{b(\xi)_m : m \geq 1\}$ are as follows:

$$
\begin{aligned}
b(\xi)_1 &= \sqrt{\frac{3}{\lambda\pi}}\left(\frac{2\xi}{\lambda\pi} - 1\right) \\
b(\xi)_2 &= \sqrt{\frac{5}{\lambda\pi}}\left(6\left(\frac{\xi}{\lambda\pi}\right)^2 - \frac{6\xi}{\lambda\pi} + 1\right) \\
b(\xi)_3 &= \sqrt{\frac{7}{\lambda\pi}}\left(\frac{2\xi}{\lambda\pi} - 1\right)\left(10\left(\frac{\xi}{\lambda\pi}\right)^2 - \frac{10\xi}{\lambda\pi} + 1\right) \\
b(\xi)_4 &= \sqrt{\frac{9}{\lambda\pi}}\left(70\left(\frac{\xi}{\lambda\pi}\right)^4 - 140\left(\frac{\xi}{\lambda\pi}\right)^3 + 90\left(\frac{\xi}{\lambda\pi}\right)^2 - 20\left(\frac{\xi}{\lambda\pi}\right) + 1\right) \\
b(\xi)_5 &= \sqrt{\frac{11}{64\lambda\pi}}\left(63\left(\frac{2\xi}{\lambda\pi} - 1\right)^5 - 70\left(\frac{2\xi}{\lambda\pi} - 1\right)^3 + 15\left(\frac{2\xi}{\lambda\pi} - 1\right)\right) \\
b(\xi)_6 &= \sqrt{\frac{13}{256\lambda\pi}}\left(231\left(\frac{2\xi}{\lambda\pi} - 1\right)^6 - 315\left(\frac{2\xi}{\lambda\pi} - 1\right)^4 + 105\left(\frac{2\xi}{\lambda\pi} - 1\right)^2 - 5\right) \\
b(\xi)_7 &= \sqrt{\frac{15}{256\lambda\pi}}\left(429\left(\frac{2\xi}{\lambda\pi} - 1\right)^7 - 693\left(\frac{2\xi}{\lambda\pi} - 1\right)^5 + 315\left(\frac{2\xi}{\lambda\pi} - 1\right)^3 - 35\left(\frac{2\xi}{\lambda\pi} - 1\right)\right),
\end{aligned}
$$

where $\lambda = 1$ for the range $\xi \in [0, \pi]$ and $\lambda = 2$ for the range $\theta \in [0, 2\pi]$.

A.2 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 $\Delta_n(\widehat{\theta}) = -\delta_n \sum_{t=1}^{n} w_t' \varphi(\widehat{u}_t)$. 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

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

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 $\Omega_p = \int \varphi(u)\varphi(u)^T p(u)du = \int \widetilde{\varphi}(\varepsilon)\widetilde{\varphi}(\varepsilon)^T (\det \Sigma)^{-1/2} f(\varepsilon)d\varepsilon$. 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:

$$\frac{g'}{g}(v) = s(v) + \tau'(v)\frac{\gamma'}{\gamma}\left\{\tau(v)\right\}.$$

It follows that we can use our kernel estimate of the score of γ to nonparametrically estimate of the score of g as follows:

$$\frac{\widehat{g}_t'}{\widehat{g}_t}(v) = s(v) + \tau'(v)\frac{\widehat{\gamma}_t'}{\widehat{\gamma}_t}\left\{\tau(v)\right\},$$

so that

$$\frac{\widehat{g}_t'}{\widehat{g}_t}(v) - \frac{g'}{g}(v) = \tau'(v)\left\{\frac{\widehat{\gamma}_t'}{\widehat{\gamma}_t}(\tau(v)) - \frac{\gamma'}{\gamma}(\tau(v))\right\}.$$

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

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

where $\alpha(v) = v\tau'(v)^2 J_\tau^{-1}\left\{\tau(v)\right\}$. 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 (1994), 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 4 satisfies

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

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 3 *Under the above Condition A,*

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

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

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

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

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

So

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

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

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

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] Amsler, C.E., and P. Schmidt 1985. A Monte Carlo invesitgation of the accuracy of multivariate CAPM tests. *Journal of Financial Economics* 14, 359-375.

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

[3] 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.

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

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

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

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

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

[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. 1994. Graphical methods for investigating the size and power of hypothesis tests. Working paper, Queens University.

[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] Fan, Y. 1994. Testing the goodness-of-fit of a parametric density function by kernel methods. *Econometric Theory* 10:316-356.

[20] Fang, K.-T., S. Kotz, and K.-W. Ng. 1990. *Symmetric Multivariate and Related Distributions.* London, Chapman and Hall.

[21] 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.

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

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

[24] 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.

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

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

[27] Hodgson, D.J. 2000. Partial maximum likelihood and adaptive estimation in the presence of conditional heterogeneity of unknown form. *Econometric Reviews* 19:176-206.

[28] Hodgson, D.J. and Vorkink, K. 2000. Semiparametric efficient estimation of GARCH-in-mean models and the conditional CAPM under elliptical symmetry. Unpublished, University of Rochester.

[29] Horowitz, J. 2000. Estimation of a generalised additve model with unknown link funciton. Forthcoming, *Econometrica.*

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

[31] Ingersoll, J. 1987. *Theory of Financial Decision Makin*g. Totowa, NJ: Rowan & Littlefield.

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

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

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

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

[36] 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.

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

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

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

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

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

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

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

[45] 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.

[46] 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.

[47] Robinson, P. M. (1988): "The Stochastic Difference between Econometric Statistics," *Econometrica*, 56, 531-548.

[48] Rothenberg, T.J., and C.T. Leenders (1964). Efficient estimation of simultaneous equation systems. Econometrica 32, 57-76.

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

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

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

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

[53] 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.

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

[55] 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.

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

[57] Vorkink, K. 1999. Improved power in testing asset pricing models, Unpublished manuscript, Brigham Young University.

[58] WAND, M.P., J.S. MARRON, AND D. RUPPERT. 1991. Transformations in Density Estimation (with discussion) *Journal of the American Statistical Association* 86, 343-361.

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

[60] WHITE, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50:1-25.

[61] ZHOU, G. 1993. Asset pricing tests under alternative distributions. *Journal of Finance* 48:1927-1942.