

A Spline Lagrange Multiplier Test
of Goodness of Fit:
Some Preliminary Results

J. Huston McCulloch
Department of Economics
Ohio State University

E. Richard Percy, Jr.
Department of Economics
Ohio State University

June 2002

For presentation at the 8th international conference on Computing in Economics and Finance, Aix-en-Provence, France, June 27-29, 2002.

1. Introduction. Adler, Feldman and Taqqu (1998) preface their collection of papers with the observation that ever since information has been gathered, it has either been categorized as “good data” (translation: the investigator knew how to choose and perform the appropriate statistical tests) or “bad data” (that is, the observations did not conform to well-known and well-understood distributions, often having too many outliers or outliers that were too far from what was expected). This may lead to some studies not being completed at all, while others may interpret the data without taking advantage of the totality of information present.

The determination of distributional assumptions in econometric and financial models is an important concern. If the distribution of error terms is inconsistent with the assumed model, then the assumed model is misspecified. If a set of assumptions concerning error terms exists and is not used then estimates of a model’s parameters are needlessly inefficient.

Though there is a critical need in financial and economic models to match the right tool to the right distribution, the tests suggested herein are not restricted to those disciplines. This question is just as important in many other fields; wherever “bad data” reside, better tools are needed to transform them into their “better understood” foils.

Many often-used modeling techniques such as Ordinary Least Squares (OLS) and the Generalized Method of Moments (GMM) do not require the specification of the distribution of the error terms. Appealing to different versions of the law of large numbers, estimators of parameters using these techniques can be shown to be consistent. In addition, estimators can be shown to be consistent under certain moment conditions. Since some laws of large numbers depend only on the first moment, specification of a finite variance is not even required. However, Maximum Likelihood (ML) Estimators that exploit the properties of a particular distribution are not only consistent but also asymptotically efficient.

In some cases the investigator may be satisfied with a lesser level of relative efficiency in estimating the expected value of a random variable if the burden of searching for a more efficient estimation method is too difficult. However, consider the example of risk-averse agents making inferences concerning future values of a financial time series. With risk-neutral agents, it may be enough to estimate expected values of returns. However, with risk-averse agents it is well known that second and higher moments of distributions matter in the selection of an optimal investment portfolio. In addition, it is often desirable to place confidence limits on estimates of expected values, to calculate variances, and perhaps measures of skewness and kurtosis. To accomplish these goals, one should not use the classical methodologies, such as least squares for calculating means and variances conditional upon exogenous variables perhaps using an assumption that all error terms are to be assumed to be a random sample independently drawn from normal distributions with an identical yet unknown mean and variance, unless the assumptions of the model selected are at least approximately satisfied.

There are many tests that have been offered in the literature for determining whether an observed sample is likely to have been drawn from a normal distribution. There are also more robust, distribution-free or nonparametric tests that can be used. However, in many cases taking advantage of additional distributional information may lead to more efficient inferences and is to be preferred over the automatic use of nonparametric methods.

With some models previous work by others suggest distributions to be hypothesized. It is well known in financial literature that error terms of returns of many assets are leptokurtic, with an unusually high number of observations several standard deviations from the mean. This phenomenon suggests that it is inappropriate to make an assumption of normality in dealing with estimates arising from the use of such samples drawn from real world data. In addition it may be helpful to estimate parameters by something other than minimizing a quadratic form. As is well known, using least squares estimators is equivalent to using maximum likelihood estimators when the underlying error distribution is Gaussian. With other distributional assumptions, this relationship disappears.

Some applied practitioners show parameter estimates calculated both with and without observations that have residuals more than a given number of standard errors from zero. This throwing away of data (or, in

some cases, reducing some observations' distances from the median or reducing their impact on the model) without just cause should make theoreticians cringe. However, if the calculation methods are least-squares based and the error terms are distributed with a distribution that has an infinite variance, it may be that the truncated or a "Winsorized"¹ estimator actually has a greater probability of being within a given distance from the true parameter than the least squares estimator.

With financial time series, several non-Gaussian distributions have been suggested with the hope that one of these may be more appropriate in making inferences. Among these are stable distributions (also called stable Pareto-Lévy or stable Paretian distributions), which include the normal distribution as a special case. Other distributions that are considered as substitutes are mixtures of more than one normal distribution, generalized Student-*t* distributions and distributions that are mixtures of continuous distributions and discrete distributions which are used to account for sudden increases or decreases in a sample. Since the early 1980s Autoregressive Conditionally Heteroskedastic (ARCH) and Generalized Autoregressive Conditionally Heteroskedastic (GARCH) models have also been used to try to explain distributions of error terms that are not independent and identically distributed (IID).

Two additional reasons for attempting to determine the distribution of error terms follow. First, if a particular distribution is determined *not* to be the underlying distribution of the error terms, by implication, at least one of the necessary assumptions for that distribution must be false. This may lead to a new understanding of the observations and possibly a new theoretical model. Second, if a particular distribution *does* have a reasonable possibility of being the underlying distribution of the error terms, one can extrapolate to possible values that are not apparent in the sample but could occur in the future. That is, with a theoretical distribution, tail probabilities that are more remote than could be observed with the limited data can be estimated.

This paper proceeds as follows. Section 2 outlines the distributional conclusions and assumptions of previous studies on financial series. Section 3 offers a brief outline of a selected list of better-known influential goodness-of-fit tests (GFTs). Section 4 discusses the special problems that exist with goodness-of-fit tests when model parameters need to be estimated. Section 5 discusses a general Lagrange multiplier GFT, while Sections 6 and 7 introduce a new Cubic Spline Lagrange multiplier GFT and shows Pearson's test as being a special case of a spline GFT. Section 8 re-introduces the classical but lesser-known Neyman Ψ^2 GFT and suggests an LM test that is an alternative to Neyman's likelihood ratio test, with Sections 9 and 10 discussing the practicality of implementing such a test. Section 11 shows extensions of the GFTs to general distributions and cases with estimated model parameters. Section 12 outlines some finite sample properties. Section 13 explores the case that is primarily faced a distribution with unknown model parameters. A case study is performed in Section 14, while in Section 15 differences between competing distributions are investigated. Sections 16 and 17 discuss attributes of alternative GFTs, while Section 18 suggests some work to be completed in the future.

2. Determination of Error Distribution. There is a wide variety of opinion of the correct error distribution in many financial series. Most researchers rule out Gaussian distributions after any testing of skewness and kurtosis, although throughout the history of analysis, many have used them; for example, Fama (1976) has suggested normal distributions for monthly returns after previously (1965) being in the leptokurtic camp. Mandelbrot (1963), Samorodnitsky and Taqqu (1995), and McCulloch (1996) have suggested the use of stable distributions. A search for finite-variance leptokurtic distributions has included Blattberg and Gonedes (1974), Hagerman (1978), Perry (1983), and Boothe and Glassman (1987) investigating alternatives such as Student-*t* distributions. Praetz (1972) and Clark (1973) explored the possibility of a mixture of normal distributions. Among models with changing volatility, Bollerslev (1987) suggested the use of a Student-*t* distribution, Nelson (1991) tried a Generalized Error Distribution, and McCulloch (1985) used stable distributions.

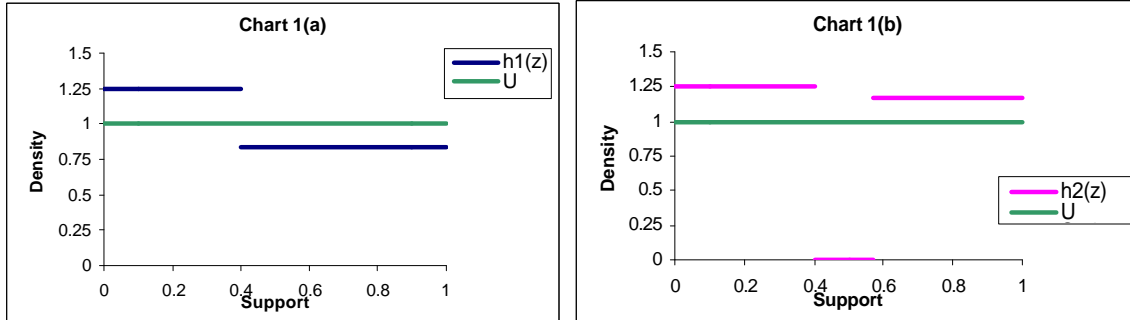
¹ Perhaps coined by John Tukey in honor of the biostatistician, Charles P. Winsor, who supposedly adopted the practice of replacing outliers with values closer to the median of the residual distribution, so that such outliers would have less impact on a model's parameters.

With this literature and appropriate goodness-of-fit tests (GFTs), there would seem to be a rich array of parametric distributions to choose from before one must resort to nonparametric procedures. A challenge that this paper is aimed at is choosing appropriate GFTs that can work well with all the above distributions.

3. Some Well-Known Goodness-of-Fit Tests. The Kolmogorov-Smirnoff statistic (KS) makes use of a transformation from a posited distribution to the distribution that is uniform on the unit interval.² KS is the largest vertical distance between the transformed empirical distribution function and a 45° line on the unit interval. KS is independent of the hypothesized distribution and critical values are dependent on n , however it requires knowledge of the true values of the parameters in a distribution. While this test statistic is sensitive to the single data value that is “farthest” away from the population cumulative distribution function (CDF), it does not directly take into account the relative deviations of the other observations. Andrews (1997) has offered a conditional K-S test that accounts for the parameter estimation effects. Still, this test is based on a single point of the empirical distribution.

The Cramér-von Mises test, again after transformation from the posited distribution to a uniform distribution, uses all the observations and is based on the integrated squared distance between the transformed empirical CDF and a 45° line. Since it is based on a distribution function and not the density function directly, some densities may tend to “fool” it. Consider the following example, adapted from McCulloch (1999), and pictured below in Charts 1(a) and 1(b), along with a uniform density on the unit interval:

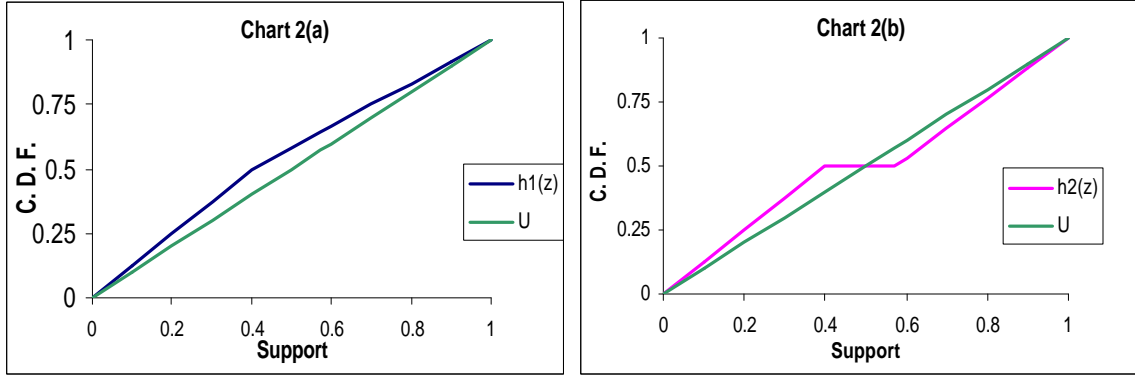
$$\text{Let } h_1(z) = \begin{cases} \frac{5}{4} & z \in [0, \frac{2}{5}] \\ \frac{5}{6} & z \in (\frac{2}{5}, 1] \\ 0 & \text{otherwise} \end{cases} \text{ and } h_2(z) = \begin{cases} \frac{5}{4} & z \in [0, \frac{2}{5}] \\ \frac{7}{6} & z \in (\frac{4}{7}, 1] \\ 0 & \text{otherwise} \end{cases}$$



The Uniform density on $[0,1]$ is shown for comparison. Clearly the function $h_1(z)$ is more nearly uniform than $h_2(z)$ from comparison of densities. The first function is the same distance as the second from the Uniform for every value except the range $(\frac{2}{5}, \frac{4}{7}]$; on this interval, the first function is closer to the Uniform. A look at the CDF's of these random variables will highlight a weakness in the Cramér-von Mises test.

The CDF of $h_2(z)$ is the same distance from the 45° degree line as is $h_1(z)$ everywhere except the interval $(\frac{2}{5}, \frac{4}{7}]$; on this interval, its distance from the Uniform is *smaller* than the distance of $h_1(z)$. Thus, the population integrated squared distance is smaller for $h_2(z)$ than for $h_1(z)$. Therefore, the Cramér-von Mises test would be less likely to reject $h_2(z)$ than $h_1(z)$ even though $h_2(z)$ departs more from the Uniform. Since, *a priori*, the investigator is not likely to know the type of departure from the hypothesized distribution; it seems that a reasonable property for a GFT is to be more sensitive to greater departures.

² Instead of using the sample, y_1, \dots, y_n , it uses the transformed sample, $F(y_1), \dots, F(y_n)$, where F is the posited distribution.



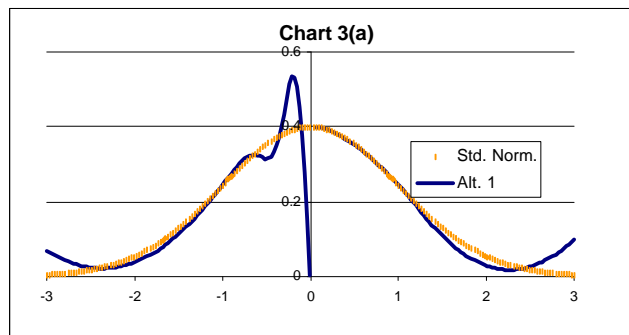
Another GFT is a transformation of the Pearson χ^2 test. Strictly speaking the Pearson test is designed to deal with discrete probability mass functions rather than continuous probability density functions. For a multinomial distribution with n observations the test is:

$$H_0: p_j = p_{j0}, j = 1, \dots, m+1^3 \text{ vs. } H_1: \text{Not } H_0$$

The Pearson statistic is $Q_m = \sum_{j=1}^{m+1} \left[\frac{(Obs_j - np_{j0})^2}{np_{j0}} \right]$ where Obs_j is the number of observations in the sample with the j^{th} value.

A continuous distribution can be transformed to a multinomial distribution with parameters $\{p_j\}$ by segmenting the support of the distribution into $m+1$ sub-supports or “bins” and calculating the population probability for each sub-support. It is not required that the p_j be equal. This statistic is easy to calculate, known to be asymptotically distributed as a χ^2 statistic with m degrees of freedom and can be used in a wide variety of situations. Its power is dependent both on the choice of m and the choice of $\{p_j\}$. It is intuitive that some power will be lost due to the reduction of information by grouping the data to test continuous distributions. This grouping has the effect of assigning an equal density to all possible values within each bin and also, perhaps, assigning nearby values that happen to be in different bins very different density values. The following example will highlight possible problems with this type of test.

Consider a test of whether n observations are from a standard normal distribution with $m+1 = 10$. For equiprobable bins, one would need the 9 breakpoints, $\Phi^{-1}\left(\frac{j}{10}\right)$, where Φ is the standard normal cumulative distribution function.

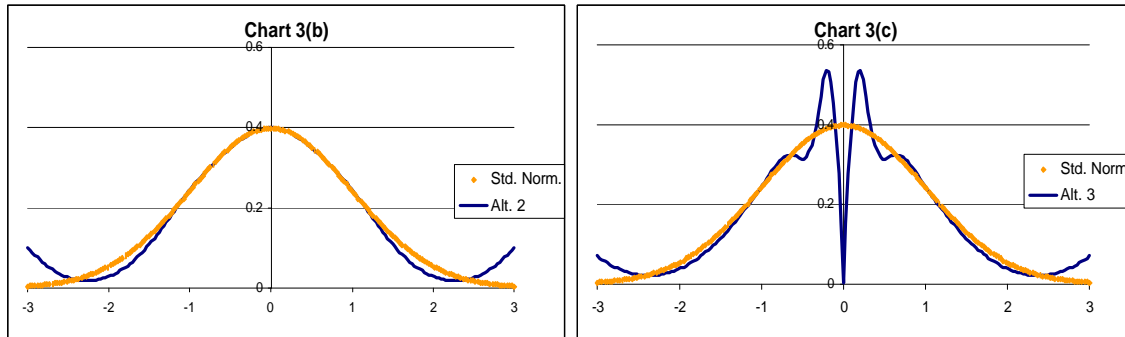


Above is a graph of a standard normal density and also a contrived density (Chart 3(a)) that integrates to $\frac{1}{2}$ on either side of zero; it also integrates to very close to 0.1 between consecutive standard normal deciles.

³ Here $m+1$ is used to facilitate future comparison with other tests. There are only m probability parameters being set since one of the parameters is constrained to be one minus the sum of the others.

Consequently, the Pearson χ^2 test has no more power against the alternative than the probability that it will reject the null hypothesis.

Also note that alternatives that are symmetric around zero based on the shown function on either the negative or positive support (Charts 3(b) and 3(c)) would be equally difficult to reject. Some different choices of $m+1$ may increase the power, but m will generally be a function of the sample size, so choice of q is not a dependable solution to this problem. The major concern in this study with the Pearson χ^2 test is that it will often be relatively insensitive to heavier tails that characterize many of the distributions of highest interest.



Please note that this is not a direct criticism of the Pearson χ^2 test since its use with discrete rather than continuous data is its strength, although many texts omit this.

4. Estimated Model Parameters. In most models the residuals are estimates of the unknown underlying errors. Roughly speaking, one can visually assess and infer a distribution of error terms from a histogram of residuals. However, a visual assessment should be reduced to a mathematical assessment, since histograms will necessarily differ from underlying distributions and what is desired is a determination of whether or not the histogram in question is *statistically significantly* different from the hypothesized theoretical distribution. For this task, GFTs must be designed to meet the varying needs of each situation.

In general, the error random variables to be tested are unobservable. With a standard regression model:

$$y_i = h(X_i; \mathbf{b}) + \mathbf{e}_i \quad (1)$$

is assumed to be true with $i = 1, \dots, n$, where y_i is the i^{th} observed dependent variable, X_i is a row vector (with dimension k) of known constants (or is uncorrelated with the vector of \mathbf{e} 's)⁴, \mathbf{b} is a k -vector of unknown coefficients, and \mathbf{e}_i is an unobservable random variable with some distribution, with $E(\mathbf{e}_i) = 0$ (or possibly the median of \mathbf{e}_i 's distribution is zero), $\Pr(\mathbf{e}_i < z) = F(z; \mathbf{g})$, $\mathbf{g} \in \mathbf{G}$, and \mathbf{e}_i is independent of \mathbf{e}_j if $i \neq j$. The function $h(X_i; \mathbf{b})$ could be linear or non-linear.

Under these assumptions we would like to test whether the vector $\mathbf{e} = [\mathbf{e}_1 \ \dots \ \mathbf{e}_n]$ is distributed according to the given function or if it has some other distribution. Typically, one must estimate \mathbf{b} and \mathbf{g} after which one can find a vector of residuals, e , that is an *estimate* of \mathbf{e} , rather than \mathbf{e} itself. If parameters are to be estimated from the data prior to the application of a GFT, standard tests are biased towards acceptance of the null hypothesis. See Mood, Graybill & Boes (1974), Bera and McKenzie (1986), and Bai (1997). DeGroot (1986) cites Chernoff and Lehmann's (1954) discovery that the use of maximum likelihood estimates, when testing whether a given distribution is normal, changes the asymptotic distribution of the test statistic under the null hypothesis in such a way as to result in smaller values. Given that larger values are necessary to reject the null hypothesis, this results in a greater than desired level of acceptance.

⁴ As usual, it is further assumed that the matrix whose rows are composed of the X_i 's are of rank k .

Intuitively, any “good” estimators of the parameters seek to fit the model as closely as possible. For example, with a classical linear regression and leptokurtic errors, the sum of the squared true errors will *almost surely* be greater than⁵ the sum of the squared residuals. This will tend to conceal the large errors that a test for leptokurtosis would be seeking.

Rayner and Best suggest a solution to the problem of testing for normality using residuals of equation (1) for a classical *linear* model, with the error terms assumed to be *iid* with the CDF given as $N(0, \sigma^2)$. By taking advantage of the familiar result from linear regression:

$$e = M\mathbf{e}, \text{ where } M = I - X(X'X)^{-1}X' \quad (2)$$

it is possible to simulate sets of residuals by generating random variables from a standard normal distribution. It is unnecessary to estimate σ^2 for most purposes because all the variables can be rescaled from σ^2 to unity. The matrix M can be calculated only once for a given model. Then as many simulations as are deemed appropriate can be used to accumulate realizations of the test statistic desired.

This method should be reasonable in many cases for its purpose, but some difficulties exist. If the random errors are not independent and a variance matrix Ω is known, M can be modified in the familiar way for generalized least squares. However, generally Ω must be estimated complicating any interpretation between the ε vector and the residuals. M has dimensions $n \times n$ and may be troublesome for especially large databases. This procedure may not be able to be extended in a straightforward manner to accommodate other situations that may arise such as non-linear regression or non-Gaussian error terms. So, the search remains for suitable alternate tests.

5. Lagrange Multiplier (LM) Test⁶ for a Uniform Distribution. Consider a random sample $x = (x_1, \dots, x_n)'$ from an unknown distribution $F(X)$. One would like to test:

$$H_0: X \sim U(0,1) \quad \text{vs.} \quad H_1: \text{Not } H_0$$

One could parameterize the alternative hypothesis in the following way:

$$H_1: X \sim G(z) \quad \text{where } G(z) = \int_0^z \left[1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(v) \right] dv, \quad 0 \leq z \leq 1.$$

To assure that $G(1) = 1$, $\{\mathbf{f}_j; j = 1, \dots, m\}$ is chosen so each perturbation function, \mathbf{f}_j , integrates to zero on the unit interval. For the set of alternative hypotheses not to contain redundant representations, $\{\mathbf{f}_j; j = 1, \dots, m\}$ must contain linearly independent elements.⁷ With no additional definition, the density associated with G can be written as:

$$g(z) = G'(z) = 1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(z). \quad \text{Also, } \Pr(x_i < z) = G(z). \quad (3)$$

It can also be seen that H_0 is nested in H_1 if one allows for $\mathbf{a}_j = 0$, $j = 1, \dots, m$. This nesting is what allows the use of a Lagrange multiplier statistic, since the parameter space of the null hypothesis is a subset of that of the alternate hypothesis.

Consider $\alpha = (\mathbf{a}_1, \dots, \mathbf{a}_m)' \neq (0, \dots, 0)$. Then for any choice of nonzero basis functions $\{\mathbf{f}_j\}$, $g(z)$ is a function different than the uniform density.⁸ For α near the origin in \mathfrak{R}^m , $g(z)$ can be seen as a perturbation

⁵ With a continuous error distribution, the probability is zero that the maximum likelihood estimates will, in fact, equal the true parameter values, so equality of the sums of squares has probability zero.

⁶ Such tests are also called “efficient score” tests, just “score” tests, or sometimes “Rao score” tests in honor of the first to suggest this type of test.

⁷ As will be discussed later, any set of linearly independent functions that integrate to zero may be chosen which will be sensitive to different departures with different power.

⁸ This is guaranteed by the linear independence of the elements of $\{\mathbf{f}_j; j = 1, \dots, m\}$.

of the uniform density,⁹ using perturbation functions, $\{\mathbf{f}_j; j = 1, \dots, m\}$. The condition $\int_0^1 g(z) dz = 1$ can easily be imposed by choosing $\{\mathbf{f}_j; j = 1, \dots, m\}$ such that $\int_0^1 \mathbf{f}_j(v) dv = 0$.

The likelihood function of interest is:

$$L(\boldsymbol{\alpha}; x) = \prod_{i=1}^n g(x_i; \mathbf{a}) \Rightarrow \log L(\boldsymbol{\alpha}; x) = \sum_{i=1}^n \log g(x_i; \mathbf{a}) = \sum_{i=1}^n \log \left(1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(x_i) \right).$$

The j^{th} first derivative, evaluated at $\boldsymbol{\alpha} = 0$, is:

$$\left. \frac{\partial \log L}{\partial \mathbf{a}_j} \right|_{\mathbf{a}=0} = \sum_{i=1}^n \frac{\mathbf{f}_j(x_i)}{1 + \sum_{k=1}^m \mathbf{a}_k \mathbf{f}_k(x_i)} \Bigg|_{\mathbf{a}=0} = \sum_{i=1}^n \mathbf{f}_j(x_i), \text{ so the transpose of the "score" vector of first}$$

derivatives, evaluated at $\boldsymbol{\alpha} = 0$, for the LM Statistic is:

$$s(0)' = \left(\sum_{i=1}^n \mathbf{f}_1(x_i), \dots, \sum_{i=1}^n \mathbf{f}_m(x_i) \right)$$

A typical element of the Hessian matrix is:

$$\frac{\partial^2 \log L}{\partial \mathbf{a}_j \partial \mathbf{a}_j} = \sum_{i=1}^n \frac{-\mathbf{f}_j(x_i) \mathbf{f}_j'(x_i)}{\left[1 + \sum_{k=1}^m \mathbf{a}_k \mathbf{f}_k(x_i) \right]^2}, \text{ so a typical element of the Fisher information matrix,}$$

evaluated at the null hypothesis, for the LM statistic is:

$$\begin{aligned} I_{jj}(0) &= -E_{\mathbf{a}=0} \left(\frac{\partial^2 \log L}{\partial \mathbf{a}_j \partial \mathbf{a}_j} \right) = -E_{\mathbf{a}=0} \left\{ \sum_{i=1}^n \frac{-\mathbf{f}_j(x_i) \mathbf{f}_j'(x_i)}{\left[1 + \sum_{k=1}^m \mathbf{a}_k \mathbf{f}_k(x_i) \right]^2} \right\} = E_{\mathbf{a}=0} \left[\sum_{i=1}^n \mathbf{f}_j(x_i) \mathbf{f}_j'(x_i) \right] \\ &= \sum_{i=1}^n E_{\mathbf{a}=0} [\mathbf{f}_j(x_i) \mathbf{f}_j'(x_i)] = \sum_{i=1}^n \int_0^1 \mathbf{f}_j(x_i) \mathbf{f}_j'(x_i) g(x_i; \mathbf{a}) \Big|_{\mathbf{a}=0} dx_i = n \int_0^1 \mathbf{f}_j(z) \mathbf{f}_j'(z) dz \end{aligned}$$

The LM statistic is $s(0)'I(0)^{-1}s(0)$. Since the null hypothesis is a point of dimension zero in \mathfrak{R}^m , this statistic is asymptotically distributed as a $\chi^2(m)$ as n increases to infinity.

The finite sample distribution for various values of n and m should be tabulated by Monte Carlo simulation. It is the intent of this further study to do so for the benefit of future researchers. For a given n , as m is increased, the power of the test relative to specific alternative distributions will increase. It is anticipated that m should be an increasing function of n . McCulloch (1971) suggests $m \approx \sqrt{n}$, whereas Li (1997) suggests $m = an^{2/5}$. There is expected to be a tradeoff between relative proximity to the $\chi^2(m)$ distribution and power of the test.

Two common estimators of the Fisher information matrix¹⁰ are the local information matrix (the negative of the empirical Hessian of the log likelihood function) and the Outer Product of the Gradient Estimator (OPG). Both estimators are consistent estimators for the Fisher information matrix. A typical element of the local information matrix is, evaluated at the null hypothesis is:

⁹Many features, methods of calculation, and inferences of the Pearson χ^2 test, the Neyman Ψ^2 test (to be discussed at the end of this section), and the proposed spline test are parallel. The differences in the tests are centered on the choice of a basis of perturbation functions.

¹⁰ See for example Davidson and MacKinnon (1993) or McCulloch (1999)

$$-\frac{\partial^2 \log L}{\partial \mathbf{a}_{j'} \partial \mathbf{a}_j} \Big|_{\mathbf{a}=0} = \sum_{i=1}^n \mathbf{f}_j(x_i) \mathbf{f}_{j'}(x_i)$$

The OPG estimator is also an empirical estimator. It is based on the contributions to the gradient matrix, a typical element of which is: $\frac{\partial \log L(\mathbf{a}, x_i)}{\partial \mathbf{a}_j} \Big|_{\mathbf{a}=0} = \mathbf{f}_j(x_i)$. The typical element of the OPG matrix is:

$$\sum_{i=1}^n \frac{\partial \log L(\mathbf{a}, x_i)}{\partial \mathbf{a}_j} \frac{\partial \log L(\mathbf{a}, x_i)}{\partial \mathbf{a}_{j'}} \Big|_{\mathbf{a}=0} = \sum_{i=1}^n \mathbf{f}_j(x_i) \mathbf{f}_{j'}(x_i).$$

Although in this case the estimators turn out to have the same form, that is not always the case. In general, since the Hessian and the OPG depend on the sample rather than the expectation, there is additional error included that inherently makes inferences poorer than if the Fisher information matrix can be computed.

There are two technical points to discuss. The first of these is a requirement that the parameter space for the null hypothesis is not on the boundary of the parameter space for the alternative hypothesis.¹¹ Intuitively, $\alpha = 0$ is an interior point of the parameter space, since at $\alpha = 0$, $g(z; \alpha) = 1$, for all values of z on the unit interval; and, evidently, under the null hypothesis, there is a local maximum at $\alpha = 0$. A more formal proof follows.

Proof. Since \mathbf{f}_j integrate to zero, $g(z; \alpha)$ will integrate to one regardless of the choice of α . So it is sufficient to show that a neighborhood exists around $\alpha = 0$ such that $g(z; \alpha) \geq 0$, for all z . Consider $\mathbf{t} = (t_1, \dots, t_m)$ with $t_j > 0$. For each value of z , $\mathbf{f}_j(z)$ is bounded above by 1 and bounded below by $-1/4$ ¹², $\forall j, \forall z$. So, a choice of t_j such that $|t_j| \leq \frac{1}{m}$, $j = 1, \dots, m$ will suffice. So, $g(z; \mathbf{t}) = 1 + \mathbf{t}'\boldsymbol{\varphi}(z)$

where $\boldsymbol{\varphi}(z) = (\phi_1(z), \dots, \phi_m(z))'$. This expands to $1 + \sum_{j=1}^m t_j \mathbf{f}_j(z)$ which cannot be greater than 2 nor less than 0. So, $g(z; \mathbf{t})$ for any such selected \mathbf{t} is a density. Next, it will be shown that the values of α that allow g to remain nonnegative are in a convex set in \mathfrak{R}^m . With that established, $g(z; \mathbf{t})$, for $\mathbf{t} \in [0, 1]$, must also be a density. So, the null hypothesis is an interior point of the unrestricted parameter space.

The second point of note is that $g(z; \alpha)$, as has been stated, is not a probability density function for some choices of α . Although care was taken in the construction of g so that it would integrate to one over the unit interval, some choices of α could cause g to be negative over some portion of that interval. Were we constructing a likelihood ratio statistic, this would be more troublesome, since the maximum likelihood estimate of α would have to be constrained to choices that allowed g to be a legitimate density function. It is expected that for many problems, an unconstrained¹³ maximum likelihood estimator may not even exist.¹⁴ However, the Lagrange multiplier statistic does not require calculation of the unconstrained maximum likelihood statistic. It merely requires a comparison of the gradient (roughly, slope) relative to the Hessian (roughly, curvature), evaluated at the null hypothesis. If the judgment is that the gradient is near enough to zero, then the null is not rejected. Nearby α 's to the $\alpha = 0$ point of the null hypothesis will

¹¹ See Rayner and Best (1989), p. 34.

¹² The upper bound is easy to calculate. The lower bound is determined by the following:

$\mathbf{f}_j(z) = -\int_0^1 \mathbf{y}_j(v) dv$ between 0 and $\frac{j-3}{m-2}$, and is evaluated to be $-\frac{1}{4} \left(z - \frac{j-3}{m-2} \right)^4 \Big|_{\frac{j-3}{m-2}}^1$ which must be greater than $-1/4$.

¹³ Since g is constructed to integrate to 1 over the unit interval, choices of α that allow g to be negative over regions of the unit interval that do not contain data allow the "likelihood" function to increase, possibly without limit, over regions that do contain data, which would cause the function to be unbounded. It should also be noted that some non-density g 's will cause the "likelihood" to be negative, if an odd number of observations occur in the region on which g is negative.

¹⁴ A constrained likelihood will exist depending on the choice of basis functions.

be densities as will be shown more formally relative to the previous technical point addressed. In fact, all the $g(z;\alpha)$ that are legal densities are near one another in the sense that the values of α that allow g to remain nonnegative are in a convex set in \mathfrak{R}^m .

Proof. Let $\alpha = (\alpha_1, \dots, \alpha_m)' \in \mathfrak{R}^m$, $\varphi(z) = (\phi_1(z), \dots, \phi_m(z))'$. Assume the contrary: at least one of the densities is not in a common convex region of \mathfrak{R}^m . Then there must be at least one function $g(z;\omega)$ that becomes negative at some point $z_0 \in [0,1]$, such that ω is a convex combination¹⁵ of ξ and ζ , where $g(z;\xi)$ and $g(z;\zeta)$ are nonnegative everywhere on the unit interval.¹⁶

$$\begin{aligned} \text{So, } g(z_0; t\xi + (1-t)\zeta) &= 1 + (t\xi + (1-t)\zeta)' \varphi(z_0) < 0 \\ \Rightarrow t\xi' \varphi(z_0) + (1-t)\zeta' \varphi(z_0) &< -1 \end{aligned}$$

$$\begin{aligned} \text{Since } g(z_0; \xi) = 1 + \xi' \varphi(z_0) &\geq 0 \text{ and } g(z_0; \zeta) = 1 + \zeta' \varphi(z_0) \geq 0, \\ t\xi' \varphi(z_0) + (1-t)\zeta' \varphi(z_0) &\geq -t - (1-t) = -1, \text{ which is a contradiction,} \end{aligned}$$

So, the assumption the $g(z;\omega)$ that becomes negative at some point $z_0 \in [0,1]$ is impossible and, thus, the densities are in a convex region of \mathfrak{R}^m .

6. Spline Lagrange Multiplier Test for a Uniform Distribution.

It is required that

$\int_0^1 \mathbf{f}_j(v) dv = 0$; one way of assuring this is to choose any set of functions $\{\mathbf{y}_j\}$ that are integrable over

$[0,1]$ and with $\mathbf{y}_j(z) = \Psi_j'(z)$, and define $\mathbf{f}_j(z) = \mathbf{y}_j(z) - \int_0^1 \mathbf{y}_j(v) dv$, since $\int_0^1 \left[\mathbf{y}_j(v) - \int_0^1 \mathbf{y}_j(u) du \right] dv =$

$$\left\{ \Psi_j(v) - v \left[\Psi_j(u) \right]_{u=0}^{v=1} \right\}_{v=0}^{v=1} = \Psi_j(1) - [\Psi_j(1) - \Psi_j(0)] - \Psi_j(0) = 0.$$

If we wanted to define a *cubic* spline, we could, for example, define

$$\mathbf{y}_j(z) = \begin{cases} z^j & j = 1, 2 \\ \max\left(z - \frac{j-3}{m-2}, 0\right)^3 & j = 3, 4, \dots, m \end{cases}$$

The bottom functional form can be visualized as translating the function z^{3+} , such that its origin is at each of the set of points $\{0, (m-2)^{-1}, 2(m-2)^{-1}, \dots, (m-3)(m-2)^{-1}, 1\}$, where

$$z^{3+} = \begin{cases} z^3 & z \geq 0 \\ 0 & z < 0 \end{cases}, \text{ the positive portion of } z^3.$$

The value, first derivative and second derivative of the \mathbf{y}_j 's are zero at $\frac{j-3}{m-2}$, where the positive portion of the function begins; so, the addition of a multiple of \mathbf{y}_j to a cubic function (or to a different cubic spline) results in a cubic spline. Thus, this set of \mathbf{y}_j 's form a basis for cubic splines with equidistant knots on $[0,1]$.¹⁷

It appears that, using the cubic spline basis, the Fisher information matrix can be calculated directly, so reliance on estimates in this case is unnecessary. The \mathbf{f} 's that are in the integrand of the typical element of this Fisher information matrix are cubic polynomials over part of their range and constant functions over the other part. So, it will only be necessary to integrate zero-degree, cubic and sextic polynomials. It is necessary to evaluate linear, quartic and septic polynomials at zero, one, and all knotpoints.

Some benefits and concerns of this type of basis will be presented in the next section.

¹⁵ $\omega = t\xi + (1-t)\zeta$ for some $t \in [0,1]$

¹⁶ ω, ξ , and ζ are all choices of $\alpha \in \mathfrak{R}^m$

¹⁷ A spline is a function of a given number m of piecewise polynomials (or some other general function) of a given degree n , each defined on a subset of a range, connected at $m-1$ points in the range, called nodes or knots, such that the values and derivatives up to degree $m-1$ of consecutive polynomials are identical at the nodes. Thus, cubic splines require values and first and second derivatives of consecutive cubic polynomials to be equal at the knots.

There is nothing magic about the selection of *cubic* splines over quadratic or linear splines. In fact even an exponentiated spline, using expressions of the form “ $\exp(\mathbf{a}_j \mathbf{f}_j)$ ” in the alternative hypothesis, may be a reasonable basis for a test. The choice against exponentiated splines is a matter of simplifying some of the calculations, since a normalizing constant that would be a possibly complex function of α would be necessary so that functions such as g would integrate to one over the unit interval. Quadratic and cubic splines are more aesthetic than linear splines in constructing likely alternative densities in that their knots are not discernible since the first derivatives of consecutive polynomials are equal. Cubic splines perhaps are to be preferred to quadratic splines since they are allowed to bend twice in a subinterval so they may be better at imitating the tails of some alternative distributions, but that characteristic may be at the expense of some other desired feature.

At this point, it can be noted that the Pearson χ^2 test is equivalent a zero-degree spline GFT for a continuous distribution. Recall the structure of the hypotheses:

$$H_0: p_j = p_{j0}, j = 1, \dots, m+1 \text{ vs. } H_1: \text{Not } H_0$$

If each of the p_{j0} are set to a constant p , then H_0 becomes the Uniform distribution. So a random sample, $x = (x_1, \dots, x_n)'$ from an unknown distribution $F(X)$ can be tested using:

$$H_0: X \sim U(0,1) \text{ vs. } H_1: X \sim G(z)$$

where $g(z)$ can be of the form of equation (3) in Section 5, H_1 :

$$g(z) = 1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(z), \text{ where, } \mathbf{f}_j(z) = \begin{cases} 1 & z \in \left[\frac{j-1}{m+1}, \frac{j}{m+1} \right) \\ -1 & z \in \left[\frac{m}{m+1}, 1 \right] \\ 0 & \text{otherwise} \end{cases} .$$

This ZSLM (Zero-degree Spline Lagrange Multiplier) test statistic would be formed in the same way as that of the CSLM, by using the score vector and Fisher information matrix indicated by the log likelihood function:

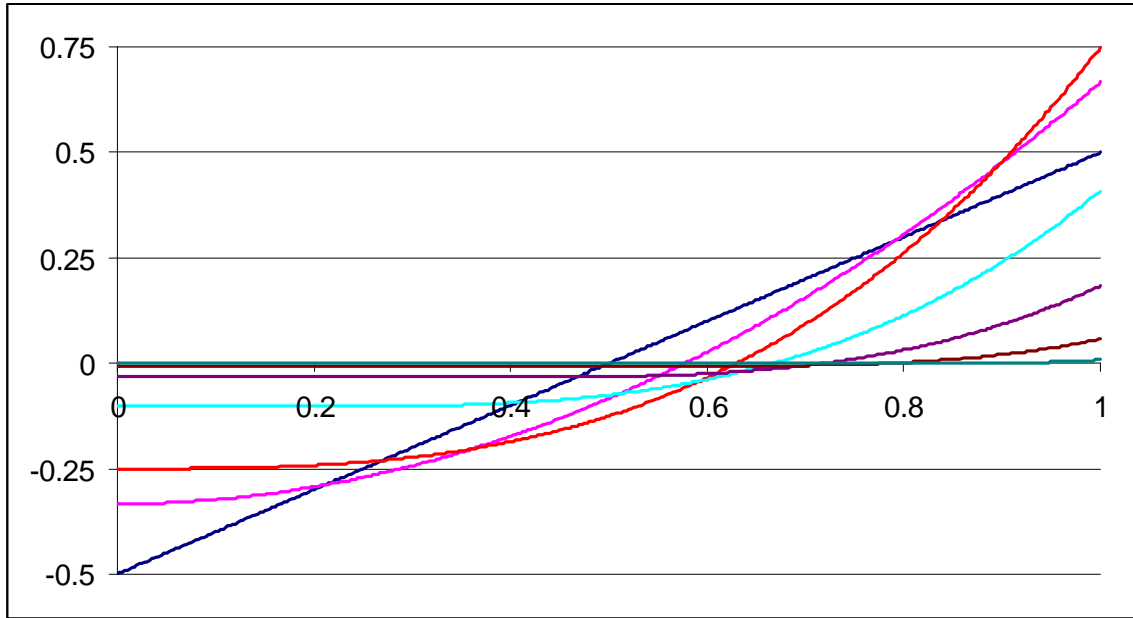
$$\text{Score: } s(0)' = \left(\sum_{i=1}^n \mathbf{f}_1(x_i), \dots, \sum_{i=1}^n \mathbf{f}_m(x_i) \right), \text{ a typical element of which would be}$$

$(n_j - n_{m+1})$, where n_j is the number of observations in the j^{th} bin, or in the interval $\left[\frac{j-1}{m+1}, \frac{j}{m+1} \right)$ and n_{m+1} is the number of observations in the last bin, or in the interval $\left[\frac{m}{m+1}, 1 \right]$.

A typical element of the Fisher information matrix is $I_{jj}(0) = n \int_0^1 \mathbf{f}_j(z) \mathbf{f}_j(z) dz$, so diagonal elements are $\frac{2n}{m+1}$ and off-diagonal elements are $\frac{n}{m+1}$. Such a matrix is easy to invert, with the inverse's diagonal elements being $\frac{m}{n}$ and the off-diagonal elements equal to $-\frac{1}{n}$.

7. B-Spline Basis. The simple cubic spline basis pictured below with seven members also has a Fisher information matrix that is poorly conditioned for inversion for large m . The seventh member is barely visible ranging from a minimum value of -0.0004 to a maximum of 0.0076 which would necessitate that its coefficient might be 2 to 3 orders of magnitude greater than a coefficient from one of the first few basis members. The difference in magnitude increases greatly as m increases.

Chart 4



Another basis for splines, typically called B-splines (see Judd, p. 227) and the one that is used in the test, is presented below. It has the advantage that all the basis members are of the same order of magnitude and the Fisher information matrix will be dominated by a strong diagonal and be nearly sparse. The linear spline matrix will be nearly tridiagonal, the quadratic spline matrix will have larger values on the main diagonal plus the four diagonals nearest the main diagonal, while the cubic spline matrix will have its largest values on the seven main diagonals. As such, this choice of bases is much better conditioned for inversion of Fisher information matrices and for accumulating the corresponding scores.

In general, B-splines of order k ($k=1$ corresponding to linear splines, $k=2$ corresponding to quadratic splines, ...) require $k+1$ basis functions for the first segment, with the requirement of adding one basis function for each additional segment. However, the splines that we are interested in have a requirement of integrating to zero over the unit interval. Consequently, we can do with one fewer basis function.

Linear B-spline functions look like “tent” functions increasing linearly from zero to a maximum from one knot point to the next, then decreasing from that maximum back to zero. Since this application requires the functions to integrate to zero on $[0, 1]$, these “tents” will be translated downward so that some of their range will be negative.

The functional form for the linear spline basis with m equal segments (and $m+1$ knots) is:

$$f_i^1(x) = y_i^1(x) - c_i^1, \quad i = 0, 1, \dots, m-1, \text{ where}$$

$$y_i^1(x) = \begin{cases} x - \frac{i}{m} & \text{if } \frac{i}{m} \leq x \leq \frac{i+1}{m} \\ \frac{i+2}{m} - x & \text{if } \frac{i+1}{m} \leq x \leq \frac{i+2}{m} \\ 0 & \text{otherwise} \end{cases} \text{ and } c_i^1 = \begin{cases} \frac{1}{m^2} & \text{if } i \in \{0, 1, \dots, m-2\} \\ \frac{1}{2m^2} & \text{if } i = m-1 \end{cases}.$$

It must be understood that the functions need not be defined outside the unit interval. For ease of exposition, that contingency is ignored. For example, the second segment of y_{m-1}^1 by the above definition is defined on the interval $[1, (m+1)/m]$ but is unnecessary for this application.

If the segments are unequal in length, one can substitute $\{x_0, x_1, \dots, x_{m-1}\}$ for $\{\frac{i}{m} \mid i = 0, 1, \dots, m-1\}$ in the above formula, where x_0 and x_m are zero and one while x_1, x_2, \dots, x_{m-1} are the desired knot points.

The formula for quadratic splines is a bit more complicated with three main segments per basis function. In addition, there must be one more function than in the linear basis for the same number of segments; so with m basis functions, one can describe only $m-1$ segments, and m knots.

$$\mathbf{f}_i^2(x) = \mathbf{y}_i^2(x) - c_i^2, \quad i = -1, 0, 1, \dots, m-2, \text{ where}$$

$$\mathbf{y}_i^2(x) = \begin{cases} \left(x - \frac{i}{m-1}\right)^2 & \text{if } \frac{i}{m-1} \leq x \leq \frac{i+1}{m-1} \\ \left(x - \frac{i}{m-1}\right)\left(\frac{i+2}{m-1} - x\right) + \left(\frac{i+3}{m-1} - x\right)\left(x - \frac{i+1}{m-1}\right) & \text{if } \frac{i+1}{m-1} \leq x \leq \frac{i+2}{m-1} \\ \left(\frac{i+3}{m-1} - x\right)^2 & \text{if } \frac{i+2}{m-1} \leq x \leq \frac{i+3}{m-1} \\ 0 & \text{otherwise} \end{cases} \text{ and}$$

$$c_i^2 = \begin{cases} \frac{4}{3(m-1)^3} & \text{if } i = -1 \& m-1 = 1 \\ \frac{5}{3(m-1)^3} & \text{if } i = -1 \& m-1 \in \{2, 3, \dots\} \\ \frac{6}{3(m-1)^3} & \text{if } i \in \{0, 1, \dots, m-4\} \& m-1 \in \{3, 4, \dots\} \\ \frac{5}{3(m-1)^3} & \text{if } i \geq 0 \& i = m-3 \\ \frac{1}{3(m-1)^3} & \text{if } i = m-2 \end{cases}$$

The formula for cubic splines has four main segments per basis function and can describe $m-2$ segments and the corresponding $m-1$ knots with m basis functions.

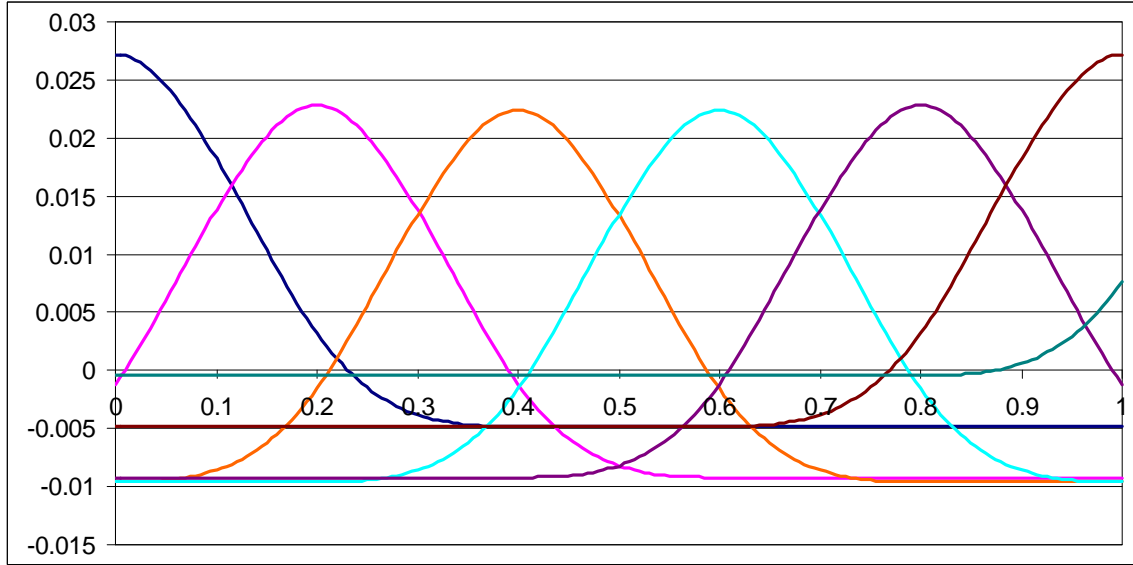
$$\mathbf{f}_i^3(x) = \mathbf{y}_i^3(x) - c_i^3, \quad i = -2, -1, 0, 1, \dots, m-3, \text{ where}$$

$$\mathbf{y}_i^3(x) = \begin{cases} \left(x - \frac{i}{m-2}\right)^3 & \text{if } \frac{i}{m-2} \leq x \leq \frac{i+1}{m-2} \\ \left(x - \frac{i}{m-2}\right)^2 \left(\frac{i+2}{m-2} - x\right) + \left(x - \frac{i}{m-2}\right)\left(\frac{i+3}{m-2} - x\right)\left(x - \frac{i+1}{m-2}\right) + \left(\frac{i+4}{m-2} - x\right)\left(x - \frac{i+1}{m-2}\right)^2 & \text{if } \frac{i+1}{m-2} \leq x \leq \frac{i+2}{m-2} \\ \left(x - \frac{i}{m-2}\right)\left(\frac{i+3}{m-2} - x\right)^2 + \left(\frac{i+4}{m-2} - x\right)\left(x - \frac{i+1}{m-2}\right)\left(\frac{i+3}{m-2} - x\right) + \left(\frac{i+4}{m-2} - x\right)^2 \left(x - \frac{i+2}{m-2}\right) & \text{if } \frac{i+2}{m-2} \leq x \leq \frac{i+3}{m-2} \\ \left(\frac{i+4}{m-2} - x\right)^3 & \text{if } \frac{i+3}{m-2} \leq x \leq \frac{i+4}{m-2} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } c_i^3 = \begin{cases} \frac{11}{4(m-2)^4} & \text{if } i \in \{-2, -1\} \& m-2 = 1 \\ \frac{12}{4(m-2)^4} & \text{if } i = -2 \& m-2 \in \{2, 3, \dots\} \\ \frac{22}{4(m-2)^4} & \text{if } i = -1 \& m-2 = 2 \\ \frac{23}{4(m-2)^4} & \text{if } i = -1 \& m-2 \in \{3, 4, \dots\} \\ \frac{24}{4(m-2)^4} & \text{if } i \in \{0, 1, \dots, m-6\} \& m-2 \in \{4, 5, \dots\} \\ \frac{23}{4(m-2)^4} & \text{if } i \geq 0 \& i = m-5 \\ \frac{12}{4(m-2)^4} & \text{if } i \geq 0 \& i = m-4 \\ \frac{1}{4(m-2)^4} & \text{if } i = m-3 \end{cases}$$

The first seven members of the basis for cubic B-splines defined above is pictured below:

Chart 5



Starting from the left, the first and sixth basis functions contain only two major segments (not including the constant segment). The second and fifth functions contain three major segments, the third and fourth functions are the only ones that contain all four major segments, while the seventh function contains only one major segment.

8. Neyman's Ψ^2 test. A GFT to which the CSLM is also closely related would be Neyman's Ψ^2 test.¹⁸ Neyman constructed an alternative hypothesis of order m (to a null of a uniform random variable on $[0,1]$)

to be $g_m(x_i; \alpha) = \exp\left\{\sum_{j=1}^m \mathbf{a}_j \mathbf{p}_j(x_i) - K(\mathbf{a})\right\}$, $0 < y < 1$, $m = 1, 2, \dots$ where $K(\alpha)$ is the constant necessary

for g_m to be a density, and the \mathbf{p}_j are orthonormal polynomials of degree j that integrate to zero on the unit interval. As in the general case, the null hypothesis is that $\alpha = 0$. With the exponentiation, there is no problem with $g_m(x_i; \alpha)$ taking negative values. The normalizing constant, which is a function of the entire α vector, may be replaced with another normalizing constant (e.g., $C(\alpha)$) outside and in front of the exponentiation function. Neyman's test statistic, which is asymptotically $\chi^2(m)$ is

$\Psi^2_m = \sum_{j=1}^m U_j^2$ where $U_j = \sum_{i=1}^n \frac{\mathbf{p}_j(y_i)}{\sqrt{n}}$. He expected that values of m of 4 or 5 would be sufficient to test a

large enough class of alternatives. Neyman's test statistic was a likelihood ratio test statistic rather than a Lagrange multiplier statistic. Since Neyman thought an m of 4 or 5 would be sufficient, it was not necessary in practice to compute $K(\alpha)$ for larger values of m . To change this to a Lagrange multiplier test with possibly larger values of m , which is concerned with perturbations only in the neighborhood of the null hypothesis, it will be convenient to simplify calculations by substituting a regular polynomial form in place of Neyman's exponentiated polynomial.

Using this structure, one definition could be:

$$g(z) = 1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(z), \text{ where, } \mathbf{f}_j = \begin{cases} z^j - \frac{1}{j+1} & z \in [0,1] \\ 0 & \text{otherwise} \end{cases}.$$

The corresponding LM test statistic would be formed in the same way as shown in Section 5, by using the score vector and Fisher information matrix indicated by the log likelihood function:

¹⁸ See Rayner and Best (1989), p.7 and p.46-48. The choices of indices, variable and parameter names have been changed to show the parallel with the CSLM and ZSLM.

Score: $s(0)' = \left(\sum_{i=1}^n f_1(x_i), \dots, \sum_{i=1}^n f_m(x_i) \right)$, a typical element of which would be $\sum_{i=1}^n x_i^j - \frac{n}{j+1}$. A typical element of the Fisher information matrix is $I_{jj}(0) = n \int_0^1 f_j(z) f_j'(z) dz$, or $n \int_0^1 (z^j - \frac{1}{j+1})(z^{j'} - \frac{1}{j'+1}) dz = \frac{njj'}{(j+j'+1)(j+1)(j'+1)}$.

One difference between the CSLM test and Neyman’s test is the expectation that the CSLM will detect differences that are local to a specific part of the unit interval. Because Neyman’s exponentiated polynomials were defined over the entire interval, each polynomial affected the likelihood of each data point. For this reason, polynomials may have to make more compromises since, in order to fit one point better, it may be necessary to fit other points worse. It remains for future work to determine the different levels of power for specific alternatives. The Pearson, Neyman, and spline tests are all asymptotically locally most powerful, unbiased tests against their design alternatives. So, it is expected that each test will work better for alternatives that are of the form determined by their respective perturbing functions. Tests with other bases of perturbing functions should be better for still other distributions.

9. Simple Polynomial Basis. For practical computations with most software using double precision with 32-bit processors, a basis of simple restricted polynomials, $\{x^m - (m + 1)^{-1}\}$, $m = 1, 2, \dots$, will likely be difficult to work with as m increases since the rows of the Fisher information matrix are nearly linearly dependent.

It is very easy to compute the cells of such matrices. Each cell is $\frac{ij}{(i+j+1)(i+1)(j+1)}$ where i is the row index and j is the column index. However, the determinants of the first 16 such matrices, using the determinant function of Microsoft Excel, are:

Table 1

1	2	3	4	5	6	7	8
0.083333	0.000463	1.65E-07	3.75E-12	5.37E-18	4.84E-25	2.74E-33	9.72E-43
9	10	11	12	13	14	15	16
2.16E-53	3.02E-65	2.73E-78	2.37E-92	-5E-107	-8E-121	-1E-135	-2E-150

The top rows indicate the number of columns (and rows) in the matrix and the bottom rows show the determinant. The determinant is getting ever smaller at a faster and faster rate. So, the hope of obtaining meaningful numerical inverses, without “infinite” precision, beyond the first few matrices is bleak.

As an example, the 8x8 matrix shows a non-dominant diagonal and little difference between the rows:

Table 2

0.083333	0.083333	0.075000	0.066667	0.059524	0.053571	0.048611	0.044444
0.083333	0.088889	0.083333	0.076190	0.069444	0.063492	0.058333	0.053872
0.075000	0.083333	0.080357	0.075000	0.069444	0.064286	0.059659	0.055556
0.066667	0.076190	0.075000	0.071111	0.066667	0.062338	0.058333	0.054701
0.059524	0.069444	0.069444	0.066667	0.063131	0.059524	0.056090	0.052910
0.053571	0.063492	0.064286	0.062338	0.059524	0.056515	0.053571	0.050794
0.048611	0.058333	0.059659	0.058333	0.056090	0.053571	0.051042	0.048611
0.044444	0.053872	0.055556	0.054701	0.052910	0.050794	0.048611	0.046478

Dividing the entries in each row by the average of the entries in that row yields the following near-singular matrix, which illustrates the ill-conditioning.

Table 3

1.30	1.30	1.17	1.04	0.93	0.83	0.76	0.69
1.16	1.23	1.16	1.06	0.96	0.88	0.81	0.75
1.07	1.18	1.14	1.07	0.99	0.91	0.85	0.79
1.00	1.15	1.13	1.07	1.00	0.94	0.88	0.82
0.96	1.12	1.12	1.07	1.02	0.96	0.90	0.85
0.92	1.09	1.11	1.07	1.03	0.97	0.92	0.88
0.90	1.07	1.10	1.07	1.03	0.99	0.94	0.90
0.87	1.06	1.09	1.07	1.04	1.00	0.95	0.91

10. Orthogonal Polynomial Basis. Alternatively, one can search for orthogonal polynomials so that the Fisher information matrix is diagonal with a uncomplicated inverse. A recursive formula for Legendre-type polynomials is shown below. The Legendre polynomials are typically defined on the range $[-1, 1]$, so a change of variable is necessary so that the resultant polynomials are orthogonal on our range of interest, $[0, 1]$.

Let $p_0(x) = 1$ and $p_1(x) = 2x - 1$. Then a recursive formula which will generate as many orthogonal polynomials as necessary on $[0, 1]$ is:

$$p_{m+1} = [(2m + 1)(2x - 1) p_m(x) - m p_{m-1}(x)] / (m + 1), \quad m = 1, 2, \dots,$$

with $\int_{[0,1]} [p_m(x)]^2 dx = (2m + 1)^{-1}$, while, as designed, $\int_{[0,1]} [p_m(x)] [p_k(x)] dx = 0$, if $m \neq k$.

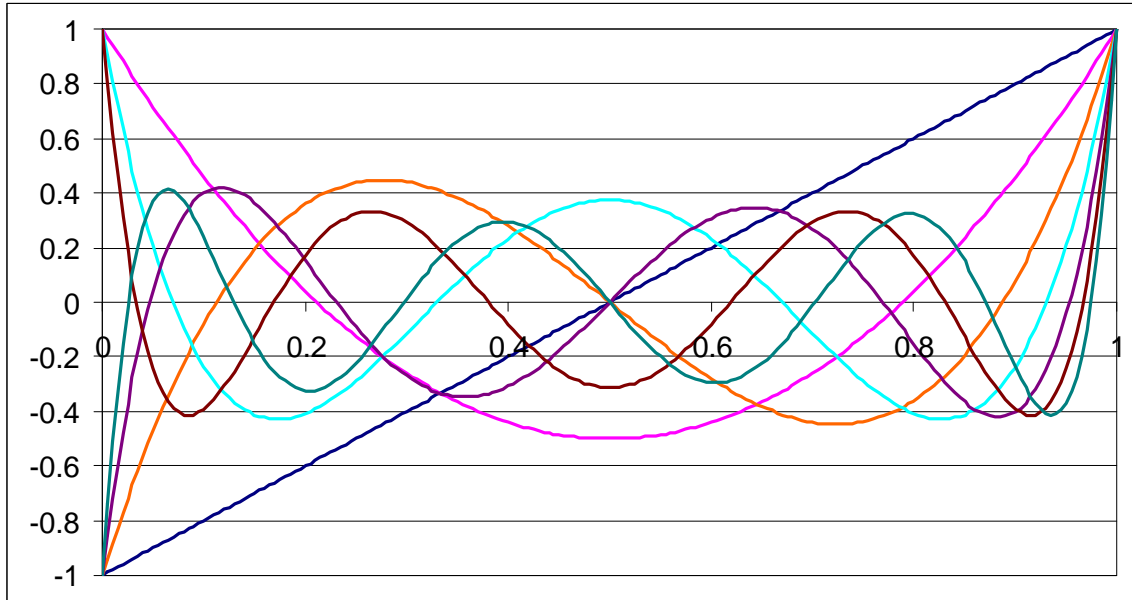
The first few such polynomials are:

$$\begin{aligned} p_2(x) &= 6x^2 - 6x + 1 \\ p_3(x) &= 20x^3 - 30x^2 + 12x - 1 \\ p_4(x) &= 70x^4 - 140x^3 + 90x^2 - 20x + 1 \\ p_5(x) &= 252x^5 - 630x^4 + 560x^3 - 210x^2 + 30x - 1 \\ p_6(x) &= 924x^6 - 2772x^5 + 3150x^4 - 1680x^3 + 420x^2 - 42x + 1 \end{aligned}$$

Some of the features of these polynomials are, as scaled, all the coefficients are integers, the sign of the lead coefficient is positive, with alternating signs thereafter, and the constant term is always ± 1 . Each function ranges between ± 1 on the domain $[0, 1]$, with $m - 1$ extrema between zero and one.

The graph below shows the first 7 members of the Neyman-Legendre type basis:

Chart 6



Although the Fisher information matrix is easy to compute, the other component of the Lagrange multiplier statistic, the score, can become problematic numerically as m increases because the relative magnitudes of the coefficients of the polynomials grows very rapidly. As an example, for $m = 24$, the coefficients of the 13th through 21st powers of x are on the order of 10^{16} , whereas the constant term still has a coefficient on the order of 10^0 . Since most software carries only 16 significant digits in its calculations, even with the use of Horner's rule of polynomial evaluation,¹⁹ the score, which is the sum of a number (equal to the sample size) of such polynomial evaluations can be expected to pick up some significant errors for large m .

However, one can still obtain reasonable numerical results for values of m beyond that which are obtainable with the simple polynomial basis.

11. Lagrange Multiplier Test for a General Completely Specified Distribution. The goal in this section is to show that the LM Test for any distribution with known parameters is the same as that for the uniform distribution.

To that end, consider a random sample $\mathbf{e} = (\mathbf{e}_1, \dots, \mathbf{e}_n)'$ from an unknown distribution. One would like to test:

$$H_0: \mathbf{e}_i \sim F(z) \quad \text{vs.} \quad H_1: \mathbf{e}_i \sim G(F(z))$$

where F is a completely specified distribution that is not $U(0,1)$, and $G(\cdot)$ is defined as before. First one can show that $F(\mathbf{e}_i) = u_i$, a random variable with a uniform distribution over the range $[0,1]$.²⁰ Conversely, if the u_i are not distributed uniformly over $[0,1]$, then the \mathbf{e}_i are not distributed according to F . So, we can test to see if the u_i are uniform, and this will be a test of the desired null hypothesis.

For the general distribution test, one can use the transformed random variables:

$u_i = F(\mathbf{e}_i)$. Under the alternative hypothesis,

$$\Pr(\mathbf{e}_i < z) = \Pr(F(\mathbf{e}_i) < F(z)) = \Pr(u_i < F(z)).$$

¹⁹The evaluation of a standard polynomial, $p_m(x) = c_m x^m + \dots + c_1 x + c_0$ by successive multiplications instead of exponentiation: $((\dots(c_m x + c_{m-1})x + c_{m-2})x + \dots + c_1)x + c_0$

²⁰The proof is well known but is included in the appendix for completeness.

Earlier, under the alternative hypothesis, $\Pr(u_i < v) = G(v)$ (substituting u_i for x_i and v for z), where G is the same G as defined in Section 5.

$$\Rightarrow \Pr(\mathbf{e}_i < z) = \Pr(u_i < F(z)) = G(F(z)).$$

The density associated with $G(F(z))$ is $g(F(z))f(z)$, by the chain rule, where $g(z) = 1 + \sum_{j=1}^m \mathbf{a}_j f_j(z)$ and $f = F'$.

The likelihood and log-likelihood functions are:

$$\begin{aligned} L(\alpha; \mathbf{e}) &= \prod_{i=1}^n g(u_i; \mathbf{a}) f(\mathbf{e}_i) \Rightarrow \log L(\alpha; \mathbf{e}) = \sum_{i=1}^n \log g(u_i; \mathbf{a}) f(\mathbf{e}_i) \\ &= \sum_{i=1}^n \log g(u_i; \mathbf{a}) + \sum_{i=1}^n \log f(\mathbf{e}_i). \end{aligned}$$

Since the second summation is constant relative to α , the first and second derivatives necessary to calculate the LM statistic are identical to those of the test for the uniform distribution. Thus, one can simply use the transformed observations, $F(\mathbf{e}_i)$, with the test for the uniform distribution. All tables and critical values that are suitable for the test of uniformity are also suitable for a general distribution.

12. Finite Sample Properties with a Completely Specified Distribution. The Lagrange multiplier statistic has a limiting asymptotic distribution that is Chi-squared with degrees of freedom equal to m , the number of perturbation parameters. Preliminary simulations suggest that for $n \geq 30$ (sample size) $m \geq 5$, and level of significance = 0.05, the convergence to the limiting distribution is quite rapid. At the alluded values of m , n , and test size, the 95th percentile of the simulated distributions could not be distinguished from the 95th percentile of a chi-squared random variable. If a lower level of significance is required, a higher sample size will be needed to use the chi-square approximation.

Following are some typical results from a simulation with 9999 repetitions²¹:

Table 4

$m = 5, n = 30$		Pearson	Neyman	Linear	Quad	Cubic	$(\chi^2_5)^{-1}(1-\text{size})$
Size							
0.0005	22.40	26.75	25.35	25.37	26.71	22.11	
0.0010	20.80	23.09	21.00	23.56	22.93	20.52	
0.0050	16.40	18.74	17.50	18.11	18.46	16.75	
0.0100	14.80	15.89	15.48	15.61	15.78	15.09	
0.0250	12.80	12.96	12.96	12.80	12.91	12.83	
0.0500	10.80	11.03	10.99	10.97	11.03	11.07	
0.1000	9.200	9.112	9.157	9.095	9.154	9.236	

²¹ In this study the number of simulation repetitions is consistently chosen to be $10^h - 1$ where h is an integer, rather than 10^h , so that the size (Type I error) of the tests that use the simulated results will be more accurate. Then for a given size (ξ), assuming that $10^h(1 - \xi)$ is an integer, the order statistic with the index $10^h(1 - \xi)$ can be used as the critical value. We could use 10^h instead and interpolate between the integers immediately above and below $(10^h + 1)(1 - \xi)$ to get to the mixed number $(10^h + 1)(1 - \xi)$. However, this involves one more calculation. It also involves an assumption that the c.d.f., which may be unknown, is linear, at least near where the critical order statistics are expected to be. Although the difference from linearity may be slight if the number of simulations is great enough, it is not necessary to make such an assumption with the proper selection of the number of repetitions.

$m = 5, n = 100$

Size	Pearson	Neyman	Linear	Quad	Cubic	$(\chi^2_5)^{-1}(1-\text{size})$
0.0005	23.12	23.58	23.65	24.80	24.29	22.11
0.0010	21.44	21.12	21.40	21.08	21.25	20.52
0.0050	17.00	17.13	16.81	17.02	17.05	16.75
0.0100	15.20	15.24	14.99	15.21	15.19	15.09
0.0250	12.80	12.94	12.97	12.74	12.89	12.83
0.0500	11.12	11.18	11.16	11.21	11.20	11.07
0.1000	9.200	9.333	9.307	9.343	9.327	9.236

$m = 10, n = 30$

Size	Pearson	Neyman	Linear	Quad	Cubic	$(\chi^2_{10})^{-1}(1-\text{size})$
0.0005	32.33	42.76	31.76	38.23	40.41	31.42
0.0010	29.40	36.19	30.93	32.62	35.93	29.59
0.0050	25.00	29.42	26.39	27.28	27.64	25.19
0.0100	22.80	26.24	24.24	24.25	24.95	23.21
0.0250	20.60	21.49	20.63	20.80	21.19	20.48
0.0500	18.40	18.74	18.34	18.56	18.79	18.31
0.1000	16.20	16.03	15.91	16.01	16.03	15.99

$m = 10, n = 100$

Size	Pearson	Neyman	Linear	Quad	Cubic	$(\chi^2_{10})^{-1}(1-\text{size})$
0.0005	32.88	36.86	34.12	34.28	35.34	31.42
0.0010	31.78	31.72	30.71	31.46	31.71	29.59
0.0050	26.28	26.46	25.95	25.69	25.84	25.19
0.0100	24.08	23.73	23.87	23.68	23.59	23.21
0.0250	20.78	20.73	20.67	20.64	20.79	20.48
0.0500	18.36	18.48	18.37	18.31	18.31	18.31
0.1000	15.94	16.20	16.11	16.26	16.23	15.99

With model parameters, the exact size in finite samples will be dependent on model characteristics including regressors, if any; however, the rapid convergence exhibited by the case with no model parameters suggests optimism that tests may likely be of approximately correct size when using a chi-squared distribution, with at least a moderate sample size.

13. LM Test for a General Distribution with Estimated Model Parameters. In this section, we seek to expand the scope of possible uses for the LM test. We still wish to consider a random sample $\mathbf{e} = (\mathbf{e}_1, \dots, \mathbf{e}_n)'$ from an unknown distribution. Again, we would like to test:

$$H_0: \mathbf{e}_i \sim F(z) \quad \text{vs.} \quad H_1: \text{Not } H_0.$$

However, now we do not know the full specification of F ; i.e., $F = F(z; \mathbf{g})$, so $\Pr(\mathbf{e}_i < z) = F(z; \mathbf{g})$, where \mathbf{g} is a vector of parameters describing the error distribution. For a Gaussian distribution, $\gamma = (\mu, \sigma^2)$; for a stable distribution, $\gamma = (a, b, c, d)$ ²²; for a generalized error distribution, γ would be a vector including a scale parameter and an exponent; and for the Student- t distribution, \mathbf{g} could be the scale and degrees of freedom, to name four examples.²³

To complicate matters just a bit more, we would like to explore the case in which our random sample, $\mathbf{e} = (\mathbf{e}_1, \dots, \mathbf{e}_n)'$, is a set of unobserved variables defined by a possibly non-linear regression form:

$$y_i = h(X_i; \mathbf{b}) + \mathbf{e}_i, \quad i = 1, \dots, n$$

²² More often, the stable parameters are known as $(\alpha, \beta, \gamma, \delta)$ or $(\alpha, \beta, c, \delta)$ but unused symbols are becoming scarcer, so the Latin letters are used to avoid notational abuse.

²³ The use of a likelihood function that contained conditional densities could allow the estimation of conditionally dependent error distributions such as ARIMA, ARCH, or GARCH distributions.

where y_i is the i^{th} observed dependent variable, X_i is a row vector of known constants (or is uncorrelated with the vector of ε 's), \mathbf{b} is a vector of unknown coefficients, with function $h(X_i; \mathbf{b})$ being possibly non-linear.

One could estimate $\theta = (\beta', \gamma)'$ by maximum likelihood and form estimates of the \mathbf{e}_i for testing as in the earlier tests with completely specified distributions. However, this would involve using residuals, without taking into consideration possible changes in the model parameters.

Instead, one could resort once again to a LM approach; but, this time, one will have to estimate all the parameters in θ simultaneously and evaluate LM statistic at the null hypothesis, $\alpha = 0$, based on the selected parameterized distribution, $F(z; \mathbf{g})$. If θ has dimension K and α still has dimension m , the LM statistic will indicate whether the m -dimensional gradient is significantly different than zero relative to the $(K+m) \times (K+m)$ dimensional Hessian. The potential improvement in the log likelihood function from its value at the null hypothesis is composed of the improvement due to the change in the error distribution measured by the change in α and the improvement due to the change in θ . Using a $\hat{\mathbf{q}}$ that is best suited to F_1 to test whether F_1 or F_2 is the better error distribution will bias a test towards F_1 , whereas using matched sets of $(\hat{\mathbf{q}}_1, F_1)$ and $(\hat{\mathbf{q}}_2, F_2)$ to determine which set better describes the data allows for a fairer test.

The score vector for the LM statistic is of dimension $K + m$, with the first K elements being zero, since these will measure the partial derivatives at the maximum likelihood estimates of θ . The Fisher information matrix will be of dimension $(K+m) \times (K+m)$, and may for some models be quite difficult to compute. For this endeavor, one may choose to estimate this by an alternate method, with consistent estimators based on the empirical Hessian or the OPG estimator.

Since the difference in dimension of the null and alternative hypotheses is m , once again, the LM test statistic will be asymptotically $\chi^2(m)$. The finite sample critical values will be dependent on the model and the specific regressors, but can be computed if need be by Monte Carlo simulations.

14. An Example with Model Parameters. To illustrate the test, we have used the monthly returns on the CRSP value-weighted index, including dividends, for the period 1/53-12/92 as described in McCulloch, 1997. The following model is estimated under the assumption of independent, identically distributed symmetric stable errors, by computing maximum likelihood estimates of the unknown parameters:

$$y_i = \mathbf{m} + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} S(a, 0, c, 0)$$

where S is the stable cumulative distribution function, a is a shape parameter and c is a scale parameter of a symmetric stable random variable. In the special case where $a = 2$, this random variable is a normal with mean zero and variance $2c^2$.²⁴ Please note that \mathbf{m} is a location parameter but not necessarily a mean, since the first moment of a non-Gaussian stable distribution with $a \leq 1$ does not exist. With the symmetric stable distribution, this parameter is always the median of the distribution.

Fitting the above model by maximum likelihood,²⁵ yields the following results:

²⁴ See the Appendix and Samorodnitsky and Taquq (1994) for more information about stable non-Gaussian random variables.

²⁵ This was accomplished using the SMSTRG Symmetric stable regression package available at <http://economics.sbs.ohio-state.edu/jhm/jhm.html>, as described in (McCulloch, "Linear Regression with Stable Disturbances," in R. Adler, R. Feldman, and M.S. Taquq, eds., A Practical Guide to Heavy Tails (1998).

Table 5

Symmetric Stable ML Estimates²⁶

a =	1.8450	se	0.0658
c =	2.7113		
log c =	0.9974	se	0.0433
log L =	-1364.7445		
μ =	0.6729	se	0.1840
n =	480 observations		

The results of the proposed GFTs, arbitrarily setting the number of parameters, m , equal to 12 for the sake of illustration, and using equidistant knots for the spline tests, yield:

$$H_0: f(\epsilon_i) \sim s(\mathbf{a}, 0, c, 0) \quad \text{vs.} \quad H_1: f(\epsilon_i) \sim G(S(a, 0, c, 0); \alpha), \text{ where the density associated with } G \text{ is}$$

$$g(S(a, 0, c, 0); \mathbf{a})s(a, 0, c, 0) = \left(1 + \sum_{j=1}^{12} \mathbf{a}_j \mathbf{f}_j(S(a, 0, c, 0)) \right) s(a, 0, c, 0) \text{ where } s(a, 0, c, 0) \text{ is the probability density function that corresponds to } S(a, 0, c, 0).$$

The results of such tests yield:

Table 6

	Statistic	$1 - \chi^2_{12}(\text{stat})$
Pearson	12.61	0.6321
Neyman-Legendre polynomial	17.23	0.3051
Linear Spline	17.39	0.2963
Quadratic Spline	18.26	0.2493
Cubic Spline	17.12	0.3119

With conventional levels of significance, one cannot reject the null hypothesis, that the errors are independent and identically distributed as a symmetric stable distribution. Note that this is not the same thing as accepting the null hypothesis. It may be that other well-known parametric distributions can fit the data as well, or that 480 observations are not sufficient to generate the power to reject the hypothesis of symmetric stable errors.

If we would consider the same test with another leptokurtic distribution, such as the generalized Student-t distribution,²⁷ we can get maximum likelihood estimates as follows:

²⁶ The application actually fit the natural logarithm of c , rather than c , so standard errors are applicable to $\log c$ rather than c itself.

²⁷ The following Student-t distribution is generalized so that it has a scale parameter, c :

$$f(x) = \int_{-\infty}^x \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right) c \sqrt{r\mathbf{p}}} \left(1 + \frac{t^2}{rc^2} \right)^{-\frac{r+1}{2}} dt, \text{ where } \Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx. \text{ The number of degrees of freedom is } r; \text{ in general, it does not need to be an integer.}$$

Table 7

Generalized Student-t²⁸ ML Estimates²⁹
 1/df = 0.1552 se 0.0413
 df = 6.4434
 c = 3.5310
 log c = 1.2616 se 0.0524
 log L = -1363.7234
 μ = 0.7164 se 0.1836
 n = 480 observations

Employing the same hypothesis testing procedure as before yields

Table 8

	Statistic	$1-\chi^2_{12}(\text{stat})$
Pearson	15.33	0.4278
Neyman-Legendre polynomial	15.36	0.4257
Linear Spline	17.81	0.2728
Quadratic Spline	18.76	0.2247
Cubic Spline	17.42	0.2945

So, the test does not reject the null hypothesis of a generalized Student-t distribution either.

Before exploring the difference and similarities between the maximum likelihood estimates of the symmetric stable parameters and the generalized Student-t parameters, it would be interesting to apply this test to the normal distribution. The normal or Gaussian distribution is a particular case of a symmetric stable distribution and is the limiting Student-t distribution as the number of degrees of freedom increases to infinity.

Table 9

Gaussian ML Estimates
 μ = 0.5554 se 0.1950
 σ^2 = 18.2487 se 1.1780
 n = 480 observations
 log L = -1378.0735

	Statistic	$1-\chi^2_{12}(\text{stat})$
Pearson	19.96	0.1315
Neyman-Legendre polynomial	30.55	0.0064
Linear Spline	26.88	0.0199
Quadratic Spline	26.62	0.0216
Cubic Spline	27.79	0.0152

So, for all but the Pearson tests, the hypothesis of Gaussian errors is rejected at conventional significance levels. The Pearson test simply looks at the proportion of observations falling in 12 equal probability regions of the distribution. When observing the residuals from the normal maximum likelihood estimation, only 26 out of the 480 observations (or 5.4%) lie more than 1.96 standard deviations from the mean. Since one would expect only 5% of the observations in this region under the null hypothesis, coupled with the

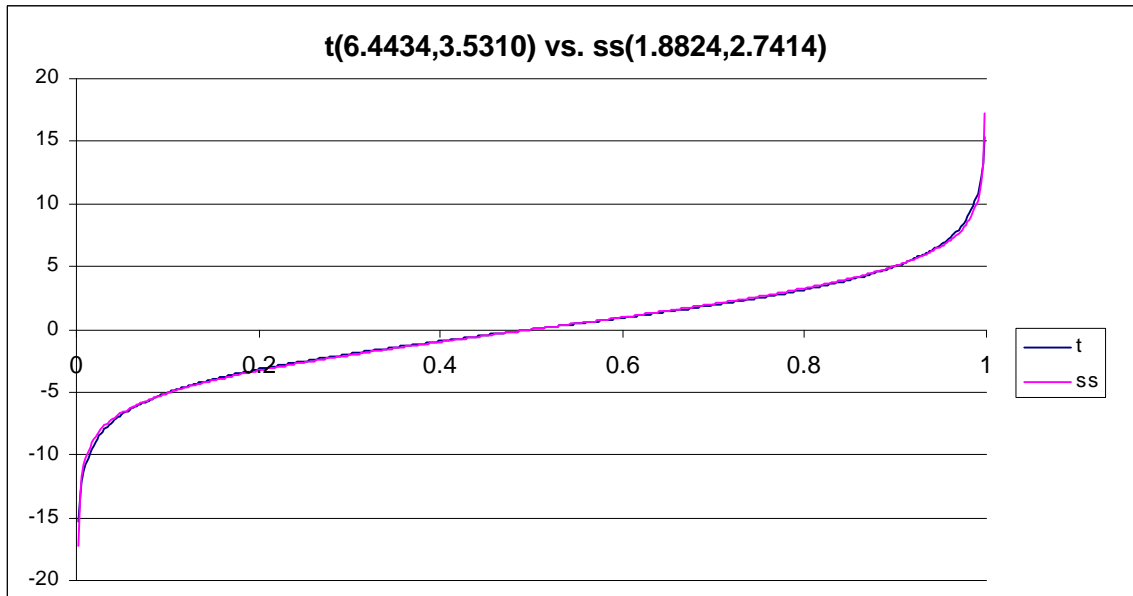
²⁸ The application actually fit the reciprocal of the degrees of freedom, so standard errors are applicable to the reciprocal rather than the estimated value of the degrees of freedom.

²⁹ Estimates were determined through a modification of SMSTRG, substituting the Student-t density and distribution.

fact that the Pearson test does not depend upon a residual's relative position in a region, this is not a surprising result.

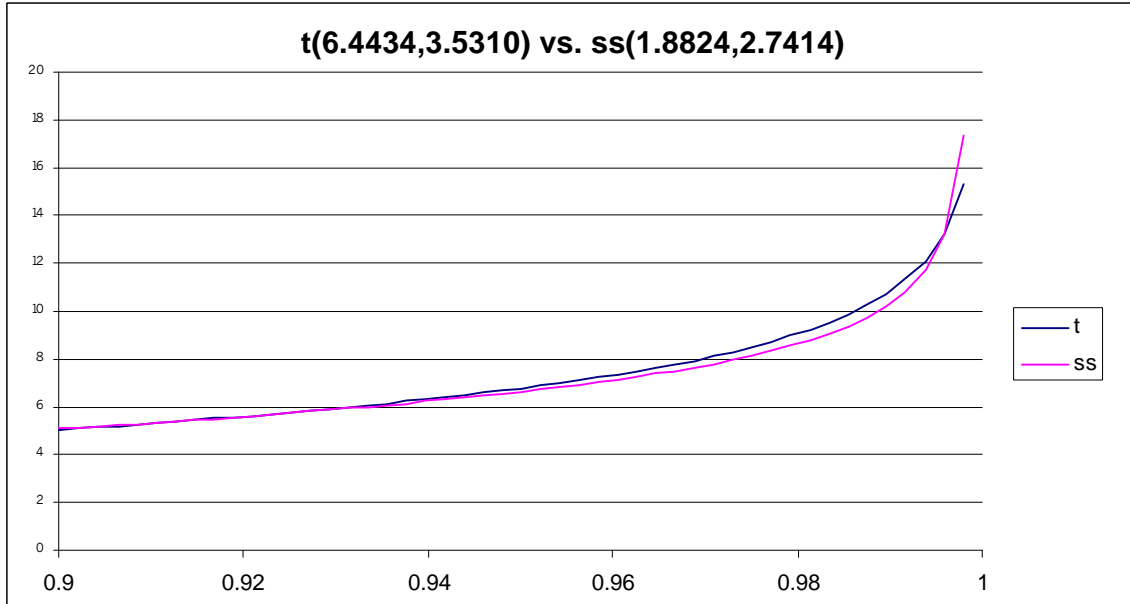
15. A Preliminary Investigation of Sensitivity. In the last section it was seen that neither symmetric stable nor Student-t errors could be rejected by the given test of the data. A comparison of the distribution functions of a symmetric stable distribution and a Student-t distribution with parameters determined by maximum likelihood estimation of the CRSP data shows that the distributions are very close. The following graph was assembled by choosing 480 probabilities from 1/481 to 480/481 for a comparison of typical samples from the two distributions of the same sample size as the data. The ordinates were determined by applying the respective inverse distribution functions to the vector of probabilities.

Chart 7



An expanded view of the tail region (48 points) allows the slight differences between the distributions to be discerned:

Chart 8



The next test uses generated series of 480 observations from a Student-t distribution with 6.4434 degrees of freedom, a scale factor 3.5310, with a location parameter of 0.7164. Not surprisingly, at this number of observations, similar tests to those previously employed do not allow summary rejection that the series of Student-t random variables were in fact a set of symmetric stable random variables. The maximum likelihood estimates of the earlier symmetric stable fitting of the CRSP data is included below for comparison.

Table 10

Series: Random Student-t Random Seed 04579384 (hex)		Series: CRSP data	
Symmetric Stable ML Estimates		Symmetric Stable ML Estimates	
a = 1.7814	se 0.0781	a = 1.8450	se 0.0658
c = 2.7441		c = 2.7113	
log c = 1.0094	se 0.0479	log c = 0.9974	se 0.0433
log L = -1384.4738		log L = -1364.7445	
μ = 0.6325	se 0.1875	μ = 0.6729	se 0.1840
n = 480 observations		n = 480 observations	
	Statistic	$1-\chi^2_{12}(\text{stat})$	
Pearson	11.84	0.6911	
Neyman-Legendre polynomial	6.32	0.9738	
Linear Spline	20.16	0.1657	
Quadratic Spline	32.94	0.0048	
Cubic Spline	42.41	0.0002	

Prior to commenting on the above statistics, it may be instructive to view another simulation:

Table 11

**Series: Random Student-t
Random Seed 9F3D29E9 (hex)**

Symmetric Stable ML Estimates
 a = 1.7820 se 0.0773
 c = 2.7914
 log c = 1.0265 se 0.0478
 log L = -1391.8818
 μ = 0.4416 se 0.1906
 n = 480 observations

	Statistic	$1-\chi^2_{12}(\text{stat})$
Pearson	21.55	0.1201
Neyman-Legendre polynomial	22.48	0.0959
Linear Spline	20.57	0.1511
Quadratic Spline	19.25	0.2027
Cubic Spline	17.65	0.2815

Clearly, in the second series, one cannot reject that the series is symmetric stable. The results from the first series are mixed, with strong rejections from the quadratic and cubic splines, but no rejections with the other tests.

Increasing the sample size, while holding m constant at 12, with the above random starting points to 2400 and 4800 still produced mixed results, while increasing the sample size to 10,000 produced rejections from all tests except the Pearson test. Preliminary illustrative results for sample size 10,000 are shown below in Table 12.

Table 12

**Series: Random Student-t
Random Seed 04579384 (hex)** **Series: Random Student-t
Random Seed 9F3D29E9 (hex)**

Symmetric Stable ML Estimates	Symmetric Stable ML Estimates
a = 1.8325 se 0.0141	a = 1.8627 se 0.0142
c = 2.6889	c = 2.6996
log c = 0.9891 se 0.0093	log c = 0.9931 se 0.0093
log L = -28417.7379	log L = -28289.9758
μ = 0.6668 se 0.0397	μ = 0.6904 se 0.0397
n = 10000 observations	n = 10000 observations

	04579384		9F3D29E9	
	Statistic	$1-\chi^2_{12}(\text{stat})$	Statistic	$1-\chi^2_{12}(\text{stat})$
Pearson	9.67	0.8397	9.58	0.8452
Neyman-Legendre polynomial	34.39	0.0030	39.08	0.0006
Linear Spline	32.91	0.0048	28.41	0.0192
Quadratic Spline	39.35	0.0006	35.97	0.0018
Cubic Spline	41.28	0.0003	38.73	0.0007

Although the data set in question has only 480 monthly returns, 50 years of daily returns would yield about 12,500 observations, so it is not unrealistic that one could observe sample sizes of 10,000 or even larger. When daily data is used the returns become less independent and less identically distributed since there is more apparent volatility clustering, day-of-the-week effects in both mean and scale, holiday effects, end-of-

year effects, among other complications. However, the method shown here of maximum likelihood estimation allows these extra considerations to be estimated without biasing the results.

Some studies use tick-by-tick Foreign Exchange rate data. At that frequency, transaction costs start to become a major consideration, so the returns are difficult to analyze, but now more than ever samples might have 100,000 or even 1,000,000 observations. So 10,000 may in some senses still be a “small” sample.

16. Comparison to Naïve Residual Tests. Many goodness-of-fit tests implicitly rely on residuals being distributed identically to the typically unknown error terms. Unless the model parameter terms are known with certainty, most parameters must be estimated. During that estimation parameters are chosen to fit the residuals as nearly as possible to the assumed error distribution. The suspected result is that residual tests will tend to be biased towards acceptance of the null hypothesis.

The symmetric stable case presented before can serve as an illustration. The first table below shows, in the left column, relative levels of significance when the hypothesis tests take into consideration possible model improvement by considering changes the model parameters as well as changes in the error distribution. The right column shows levels of significance when the estimated model parameters are taken as fixed and only changes in the error distribution are considered. The 90 tests are five tests each (Pearson, Neyman, and the three Spline tests) using from 3 to 20 free parameters to test the symmetric stable distribution.

Table 13

Frequency of Tests by Level of Significance		
Level of Significance	Informed Test	Naïve Test
≤ 0.05	3	1
0.05 – 0.10	1	0
0.10 – 0.15	0	1
0.15 – 0.20	9	1
0.20 – 0.30	24	8
0.30 – 0.40	32	30
0.40 – 0.50	8	28
0.50 – 0.60	4	14
0.60 – 0.70	3	1
0.70 – 0.80	4	3
0.80 – 0.90	0	1
0.90 – 1.00	1	2
N/A	1*	0

Average Level of Tests			
	Informed	Naïve	Difference
Pearson	0.45*	0.46*	0.01
Neyman-Legendre	0.23	0.42	0.19
Linear Spline	0.39	0.42	0.03
Quadratic Spline	0.33	0.40	0.07
Cubic Spline	0.31	0.41	0.10

*Eliminated a Pearson statistic that was negative; no Complement of the Inverse Chi-Squared Distribution statistic is available.³⁰

³⁰ The procedure employed uses a hybrid proxy for the Fisher information matrix which is not guaranteed to be positive definite. Some test-statistics can be negative. The next section will elucidate this matter.

The frequency table shows that the naïve test tends to give a higher level of significance. The average difference in the level of significance of the five types of tests is exhibited in the second part of the table. Again one can see a tendency for the naïve test to be more likely to accept the null hypothesis.

17. Comparison with Empirical Distribution Function Tests.³¹ Given a sample of size n , Y_1, \dots, Y_n , and the corresponding order statistics, $Y_{(1)}, \dots, Y_{(n)}$, the empirical distribution function (EDF) can be defined as follows:

$$EDF(y) = \begin{cases} 0 & -\infty < y < Y_{(1)} \\ \frac{i}{n} & Y_{(i)} < y < Y_{(i+1)}, \quad i = 1, 2, \dots, n-1 \\ 1 & Y_{(n)} < y < \infty \end{cases}$$

Many EDF tests have been around for half a century or more. It seems natural to compare the previous results to these tests. These tests have an assumption of a completely specified distribution. When the parameters have been determined by some optimization method such as maximum likelihood, as discussed in the previous section, inferences are less accurate.

The most common empirical distribution goodness-of-fit test is based on the Kolmogorov-Smirnoff (KS) statistic. It seeks to look at the largest single difference between the assumed distribution, $F(y)$ and the EDF, so it can be described as:

$$KS = \sup_y |EDF(y) - F(y)|$$

Two other common EDF tests are the Cramér-von Mises (CvM) statistic and an Anderson-Darling (AD) modification of CvM. The CvM statistic is the integrated squared difference between the EDF and the assumed distribution. The AD modification is based on the premise that one should examine the difference in the tails of the distributions more closely than the center of the distribution, which is accomplished by dividing by a function that takes its maximum value at the median of the distribution.

$$CvM = n \int_{-\infty}^{\infty} [EDF(y) - F^*(y)]^2 dy$$

$$AD = n \int_{-\infty}^{\infty} \frac{[EDF(y) - F^*(y)]^2}{F^*(y)[1 - F^*(y)]} dy$$

In the expressions above $F^*(y) = U[F(y)]$ where $U(y)$ is the distribution function for a uniform random variable on the unit interval. By using this transformation and sample size adjustment factors for KS and CvM, standard tables of critical values can be employed. In the table below, “Modified” refers to a function of the statistic and the sample size so that standard tables can be employed in determining the significance level of the statistic.³² Significance levels are calculated using reciprocal interpolation of the table in D’Agostino and Stephens (1986).

The results for the CRSP data follow:

Table 14

Kolmogorov-Smirnov	Student	Gaussian
Stable		

³¹ A source for the tables and descriptions of these EDF tests and others is D’Agostino and Stephens (eds.), Goodness-of-Fit Techniques, 1986, Chapter 4.

³² Modified $KS = KS \left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{nKS}} \right)$. Modified $CvM = \left(CvM - \frac{0.4}{n} + \frac{0.6}{n^2} \right) \left(1 + \frac{1}{n} \right)$. For $n \geq 5$, the AD statistic does not need any modification.

KS	0.039	0.039	0.053	
Modified	0.855	0.870	1.167	
Significance	>0.250		>0.250	0.129
Cramér-von Mises				
	Stable	Student	Gaussian	
CvM	0.099	0.097	0.305	
Modified	0.098	0.096	0.305	
Significance	>0.250		>0.250	0.129
Anderson-Darling				
	Stable	Student	Gaussian	
AD	0.793	0.813	2.046	
Significance	>0.250		>0.250	0.083

At conventional significance levels none of these tests would reject any of the posited distributions; however, the Gaussian tests appear to be in the weaker range of statistics indicating non-rejection.

18. More Accurate Results Expected in Future.

The application of Lagrange multiplier tests relies crucially on the Fisher information matrix or some estimate thereof. Especially with distributions such as the symmetric stable with no closed form for even its corresponding density function, numerical estimation of many of the components of the test statistic is necessary. When the Fisher information matrix is unavailable, consistent estimators such as the negative of empirical Hessian or the outer-product-of-the-gradient (OPG) estimator become likely candidates for substitution. When computing the Fisher information matrix the only stochastic aspect is the vector of maximum likelihood estimates. The use of the empirical Hessian or the OPG estimators inherently imparts more noise to the tests. And, per Davidson and McKinnon, the OPG estimator “often seems to be particularly poor.”³³ Thus, if possible, numerical estimation of the Fisher information matrix is the preferred option.

To determine the Fisher information matrix we require the expectation of the outer product of the gradients or the expectation of the negative Hessian.

Given the general nested hypotheses test for the case with a vector of unknown model parameters, θ ,

$$H_0: y_i \sim F(y; \theta) \text{ vs. } H_1: y_i \sim G[F(y; \theta)],$$

with the density for the alternative hypothesis being

$$g[F(y; \theta)] f(y; \theta), \text{ where } g(z) = 1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(z),$$

the likelihood and log likelihood functions are:

$$L(\mathbf{a}, \mathbf{q}; y) = \prod_{i=1}^n g[F(\mathbf{e}_i; \mathbf{q}); \mathbf{a}] f(y_i; \mathbf{q}) \quad \text{and} \quad \log L(\mathbf{a}, \mathbf{q}; y) = \sum_{i=1}^n \log g[F(y_i; \mathbf{q}); \mathbf{a}] + \sum_{i=1}^n \log f(y_i; \mathbf{q})$$

The first derivatives of the log likelihood are:

³³ Davidson and MacKinnon, Estimation and Inference in Econometrics, p. 266.

$$\frac{\partial \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{a}_j} = \frac{\partial}{\partial \mathbf{a}_j} \left(\sum_{i=1}^n \log \left(1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q})) \right) + \sum_{i=1}^n \log f(y_i; \mathbf{q}) \right) = \sum_{i=1}^n \frac{\mathbf{f}_j(F(y_i; \mathbf{q}))}{1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q}))} \text{ and}$$

$$\frac{\partial \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{q}_k} = \frac{\partial}{\partial \mathbf{q}_k} \left(\sum_{i=1}^n \log \left(1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q})) \right) + \sum_{i=1}^n \log f(y_i; \mathbf{q}) \right) =$$

$$\sum_{i=1}^n \frac{\sum_{j=1}^m \mathbf{a}_j \mathbf{f}'_j(F(y_i; \mathbf{q})) \frac{\partial F}{\partial \mathbf{q}_k}(y_i; \mathbf{q})}{1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q}))} + \sum_{i=1}^n \frac{\frac{\partial f}{\partial \mathbf{q}_k}(y_i; \mathbf{q})}{f(y_i; \mathbf{q})}$$

Evaluation at the maximum log likelihood of the restricted model, $\mathbf{a} = 0, \mathbf{q} = \hat{\mathbf{q}}$, yields:

$$\left. \frac{\partial \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{a}_j} \right|_{\substack{\mathbf{a}=0 \\ \mathbf{q}=\hat{\mathbf{q}}}} = \sum_{i=1}^n \mathbf{f}_j(F(y_i; \hat{\mathbf{q}})) \text{ and } \left. \frac{\partial \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{q}_k} \right|_{\substack{\mathbf{a}=0 \\ \mathbf{q}=\hat{\mathbf{q}}}} = \sum_{i=1}^n \frac{\frac{\partial f}{\partial \mathbf{q}_k}(y_i; \hat{\mathbf{q}})}{f(y_i; \hat{\mathbf{q}})}$$

The second derivatives of the log likelihood may not be necessary if the Fisher information matrix can be computed directly by taking the expectation of the outer product of the gradients. For completeness, if the calculation of the Hessian is necessary, the second derivatives are:

$$\frac{\partial^2 \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{a}_j \partial \mathbf{a}_j} = \frac{\partial}{\partial \mathbf{a}_j} \left(\sum_{i=1}^n \frac{\mathbf{f}_j(F(y_i; \mathbf{q}))}{1 + \sum_{h=1}^m \mathbf{a}_h \mathbf{f}_h(F(y_i; \mathbf{q}))} \right) = \sum_{i=1}^n \frac{-\mathbf{f}_j(F(y_i; \mathbf{q})) \mathbf{f}'_j(F(y_i; \mathbf{q}))}{\left[1 + \sum_{h=1}^m \mathbf{a}_h \mathbf{f}_h(F(y_i; \mathbf{q})) \right]^2},$$

$$\frac{\partial^2 \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{q}_k' \partial \mathbf{q}_k} = \frac{\partial}{\partial \mathbf{q}_k'} \left(\sum_{i=1}^n \frac{\sum_{j=1}^m \mathbf{a}_j \mathbf{f}'_j(F(y_i; \mathbf{q})) \frac{\partial F}{\partial \mathbf{q}_k}(y_i; \mathbf{q})}{1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q}))} + \sum_{i=1}^n \frac{\frac{\partial f}{\partial \mathbf{q}_k}(y_i; \mathbf{q})}{f(y_i; \mathbf{q})} \right) =$$

$$\sum_{i=1}^n \frac{\left[1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q})) \right] \left[\sum_{j=1}^m \mathbf{a}_j \left(\mathbf{f}'_j(F(y_i; \mathbf{q})) \frac{\partial F}{\partial \mathbf{q}_k'}(y_i; \mathbf{q}) \frac{\partial F}{\partial \mathbf{q}_k}(y_i; \mathbf{q}) + \mathbf{f}_j(F(y_i; \mathbf{q})) \frac{\partial^2 F}{\partial \mathbf{q}_k' \partial \mathbf{q}_k}(y_i; \mathbf{q}) \right) \right]}{\left[1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q})) \right]^2}$$

$$+ \sum_{i=1}^n \frac{\left[\sum_{j=1}^m \mathbf{a}_j \mathbf{f}'_j(F(y_i; \mathbf{q})) \frac{\partial F}{\partial \mathbf{q}_k}(y_i; \mathbf{q}) \right] \left[\sum_{j=1}^m \mathbf{a}_j \mathbf{f}'_j(F(y_i; \mathbf{q})) \frac{\partial F}{\partial \mathbf{q}_k'}(y_i; \mathbf{q}) \right]}{\left[1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q})) \right]^2} + \sum_{i=1}^n \frac{f(y_i; \mathbf{q}) \frac{\partial^2 f}{\partial \mathbf{q}_k' \partial \mathbf{q}_k}(y_i; \mathbf{q}) - \frac{\partial f}{\partial \mathbf{q}_k}(y_i; \mathbf{q}) \frac{\partial f}{\partial \mathbf{q}_k'}(y_i; \mathbf{q})}{[f(y_i; \mathbf{q})]^2}$$

and, first differentiating with respect to α_j , and then differentiating with respect to θ_k (for ease of calculation),

$$\frac{\partial^2 \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{q}_k \partial \mathbf{a}_j} = \frac{\partial}{\partial \mathbf{q}_k} \left(\sum_{i=1}^n \frac{\mathbf{f}_j(F(y_i; \mathbf{q}))}{1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q}))} \right) =$$

$$\sum_{i=1}^n \frac{\left[1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q})) \right] \mathbf{f}'_j(F(y_i; \mathbf{q})) \frac{\partial F}{\partial \mathbf{q}_k}(y_i; \mathbf{q}) - \mathbf{f}_j(F(y_i; \mathbf{q})) \sum_{j=1}^m \mathbf{a}_j \mathbf{f}'_j(F(y_i; \mathbf{q})) \frac{\partial F}{\partial \mathbf{q}_k}(y_i; \mathbf{q})}{\left[1 + \sum_{j=1}^m \mathbf{a}_j \mathbf{f}_j(F(y_i; \mathbf{q})) \right]^2}$$

These expressions simplify greatly at the maximized log likelihood of the restricted model, $\mathbf{a} = 0, \mathbf{q} = \hat{\mathbf{q}}$:

$$\left. \frac{\partial^2 \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{a}_j \partial \mathbf{a}_j} \right|_{\substack{\mathbf{a}=0 \\ \mathbf{q}=\hat{\mathbf{q}}}} = -\sum_{i=1}^n \mathbf{f}_j(F(y_i; \hat{\mathbf{q}})) \mathbf{f}'_j(F(y_i; \hat{\mathbf{q}}))$$

$$\left. \frac{\partial^2 \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{q}_k \partial \mathbf{q}_k} \right|_{\substack{\mathbf{a}=0 \\ \mathbf{q}=\hat{\mathbf{q}}}} = \sum_{i=1}^n \frac{f(y_i; \hat{\mathbf{q}}) \frac{\partial^2 f}{\partial \mathbf{q}_k \partial \mathbf{q}_k}(y_i; \hat{\mathbf{q}}) - \frac{\partial f}{\partial \mathbf{q}_k}(y_i; \hat{\mathbf{q}}) \frac{\partial f}{\partial \mathbf{q}_k'}(y_i; \hat{\mathbf{q}})}{[f(y_i; \hat{\mathbf{q}})]^2}$$

$$\left. \frac{\partial^2 \log L(\mathbf{a}, \mathbf{q}; y)}{\partial \mathbf{q}_k \partial \mathbf{a}_j} \right|_{\substack{\mathbf{a}=0 \\ \mathbf{q}=\hat{\mathbf{q}}}} = \sum_{i=1}^n \mathbf{f}'_j(F(y_i; \hat{\mathbf{q}})) \frac{\partial F}{\partial \mathbf{q}_k}(y_i; \hat{\mathbf{q}})$$

Thus, the full Fisher information matrix for the case with unknown model parameters is:

$$\begin{bmatrix} n \int_{-\infty}^{\infty} \frac{\frac{\partial f}{\partial \mathbf{q}_k}(y; \mathbf{q}) \frac{\partial f}{\partial \mathbf{q}_k'}(y; \mathbf{q})}{[f(y; \mathbf{q})]} dy & n \int_{-\infty}^{\infty} \mathbf{f}_j(F(y; \mathbf{q})) \frac{\partial f(y; \mathbf{q})}{\partial \mathbf{q}_k} dy \\ n \int_{-\infty}^{\infty} \mathbf{f}_j(F(y; \mathbf{q})) \frac{\partial f(y; \mathbf{q})}{\partial \mathbf{q}_k} dy & n \int_{-\infty}^{\infty} \mathbf{f}_j(F(y; \mathbf{q})) \mathbf{f}'_j(F(y; \mathbf{q})) f(y; \mathbf{q}) dy \end{bmatrix},$$

with each integral

being a typical entry in that quadrant of the matrix.

The off-diagonal elements can also be represented as $-n \int_{-\infty}^{\infty} \mathbf{f}'_j(F(y; \mathbf{q})) f(y; \mathbf{q}) \frac{\partial F(y; \mathbf{q})}{\partial \mathbf{q}_k} dy$, with the two different looking expressions being equal as can be determined by integration by parts.

It may be possible to perform the expectation integration for some distribution functions to get the Fisher information matrix directly. But, in any case it is always reasonable to get an empirical Hessian, by substituting residuals from the restricted estimate for the unknown error terms. All that is necessary to determine the mixed second derivatives is to (1) differentiate the chosen basis functions, (2) numerically evaluate the chosen distribution, using maximum likelihood estimates for parameters, at each residual, and (3) numerically differentiate the distribution function at each residual.

The preliminary results in Tables 6, 8, and 9 are based on a hybrid proxy for the information matrix. The hybrid matrix has some of its entries equal to those from the Fisher information matrix, some being entries from the negative of the empirical Hessian, and some being entries from the OPG estimator. Since a consistent matrix estimator consistently estimates each entry in the matrix, such a hybrid must also be a consistent matrix estimator.

The makeup of typical hybrid proxy follows:

$$\begin{bmatrix} -\sum_{i=1}^n \frac{\partial^2 \ln L(y_i)}{\partial \mathbf{q}_k \mathbf{q}_{k'}} & \sum_{i=1}^n \frac{\partial \ln L(y_i)}{\partial \mathbf{q}_k} \frac{\partial \ln L(y_i)}{\partial \mathbf{a}_j} \\ \sum_{i=1}^n \frac{\partial \ln L(y_i)}{\partial \mathbf{a}_j} \frac{\partial \ln L(y_i)}{\partial \mathbf{q}_k} & n \int_0^1 \left[-\sum_{i=1}^n \frac{\partial^2 \ln L(y_i)}{\partial \mathbf{a}_j \mathbf{a}_{j'}} \right] dy_i \end{bmatrix}_{\mathbf{q}=\hat{\mathbf{q}}_{\mathbf{a}=\hat{\mathbf{a}}}} ; \quad j, j' = 1, 2, \dots, m; k, k' = 1, \dots, K$$

The upper left portion of the matrix measures the curvature of the log likelihood with respect to the model parameters. For the stable and Student-t tests, a numerical estimate of empirical Hessian was used; for the Gaussian test, the actual Fisher information matrix was used.

The lower right portion of the matrix shows the curvature with respect to the perturbation parameters in the density of the alternate hypothesis. Given the basis functions chosen, exact analytical calculation of the Fisher information matrix was straightforward.

The off diagonal elements are estimated via OPG estimation, which perhaps is the noisiest of the three methods.

In addition to the noise, another inconvenient feature of this hybrid matrix is that it is not guaranteed to be positive definite. Even though the test statistics are asymptotically chi-squared, they can be negative in finite samples.

With numerical methods to estimate the Fisher information matrix these types of problems should be avoided.

19. Conclusion. General purpose Lagrange multiplier goodness-of-fit tests can be used with economic and financial data to probe the distribution underlying the generation of the data. Some parsimonious parametric distributions may be found that will aid inferences about levels of and relationships between economic variables. Thus, asymptotically consistent estimates of parameters are possible without either presuming normality of error terms or using solely nonparametric techniques. In that regard, these new procedures can offer new answers to old questions.

Unlike many goodness-of-fit tests, unknown model parameters can be estimated with the tests presented herein without prejudicing the tests. Since these tests rely on maximum likelihood techniques, they are asymptotically most powerful tests against their designed set of alternatives.

Spline models are more tractable than polynomial models with existing double precision software, but more work has to be done on power to determine whether tractability is offset by lower power in tests of interest.

An illustration with model parameters was presented for illustration of test properties and contrasted with some common goodness-of-fit tests. Rejection of the hypothesis of normal error terms was accomplished with the new tests but not with the old tests.

Further study is necessary to determine advantageous strategies in increasing power of the tests against particular alternatives. Uneven knot points and better estimation of Fisher information matrices are two such areas.

Appendix.

Stable random variables. Stable distributions are those that have the property of being *stable* under addition. The distribution of the sum of any number of independent random variables that are from a stable distribution will itself be stable. The logarithm of the characteristic function of each stable random variable

shares the same form.³⁴ Specifying the four-vector of parameters, (a,b,c,d) , can identify a particular stable distribution. The fourth parameter, d , is a location parameter and can, in a crude sense, be compared to some measure of average (mean, median, or mode); it is the mean for $a > 1$, and is the median and mode for $b = 0$. It can take on any real number as its value. The third parameter (c) is a scale parameter, and can be likened somewhat to a variance, standard deviation, or range; it can take on any non-negative number with a c of zero indicating a degenerate distribution. The second parameter, b , is a skewness index and can take on values between plus or minus one, inclusively; when $b > 0$ (< 0), the distribution is skewed to the right (left); when $b = 0$, the distribution is symmetric. The first parameter, a , is called the exponent and can take on values on the range $(0,2]$. When $a = 2$, this is the special case of the normal distribution. When $a \in (1,2)$, the mean of the distribution is d ; when $a \in (0,1]$, the mean of the distribution is infinite. The Cauchy distribution is a special case of the stable class of distributions and has $b = 0$ and $a = 1$. The density of stable distributions cannot be represented in closed form except in the case of the Normal, Cauchy, and Lévy³⁵, so the use of these distributions requires numerical applications such as are present at <http://economics.sbs.ohio-state.edu/jhm/jhm.html>, and <http://academic2.american.edu/~jpnolan> the home pages of J. Huston McCulloch and John P. Nolan, respectively.

Proof: Transformation of General Distribution to a Uniform over [0,1]. For any e_i drawn from the distribution, there exists a $u \in [0,1]$ such that $e_i = F^{-1}(u)$ ³⁶. Let u be distributed according to some unknown function Ξ . (Note for later use that $F^{-1}(u)$ is not defined for values outside $[0,1]$.)

$$\Pr(e_i \leq z) = F(z) \quad (*)$$

$$\Pr(e_i \leq z) = \Pr(F^{-1}(u) \leq z) = \Pr[F(F^{-1}(u)) \leq F(z)]^{37} = \Pr(u \leq F(z))$$

$$\Pr(u \leq F(z)) = \Xi(F(z)), \text{ by definition. This implies } \Xi(F(z)) = F(z), \text{ from } (*).$$

$$\text{Substituting } v \text{ for } F(z), \text{ we have } \Xi(v) = v. \text{ So, } \xi(v) = \Xi'(v) = 1.$$

Thus, ξ is a uniform density over $[0,1]$ and Ξ is the uniform distribution function over the same range.

³⁴ See McCulloch (1996)

³⁵ Stable parameters of $(\frac{1}{2}, 1, 1, 0)$.

³⁶ This is a bit informal, since not all distribution functions are strictly invertible. F is non-decreasing so it is invertible except in regions where the density is zero. However, in such regions, there will be no e_i for which we will require F^{-1} .

³⁷ Since F is a non-decreasing function.

Bibliography.

- Adler, Robert J., Raisa E. Feldman, and Murad S. Taqqu, Editors, A Practical Guide to Heavy Tails, Birkhäuser, 1998.
- Andrews, Donald W.K., "A Conditional Kolmogorov Test," *Econometrica*, 65, 1097-1128, 1997.
- Bai, Jushan, "Testing Parametric Conditional Distributions of Dynamic Models," MIT Department of Economics, 1997.
- Bera, Anil K. and Colin R. McKenzie, "Tests for Normal with Stable Alternatives," *Journal of Statistical Computation and Simulation*, 25, 37-52, 1986.
- Blattberg, R. and N. Gonedes, 1974, "A Comparison of the Stable and Student Distributions as Statistical Models for Stock Prices," *Journal of Business* 47, 244-280.
- Bollerslev, T., "A Conditional Heteroskedastic Time Series Model for Speculative Prices and Rates of Return," *Review of Economics and Statistics*, 69, 542-547, 1987.
- Boothe, Paul, and Debra Glassman, "The Statistical Distribution of Exchange Rates," *Journal of International Economics*, 22, 297-319, 1987.
- Borowski, E.J., and Borwein, J.M., The HarperCollins Dictionary of Mathematics, HarperPerennial, 1991.
- Chernoff, H. and E.L. Lehmann, "The Use of Maximum Likelihood Estiamtes in χ^2 Tests for Goodness of Fit," *Ann. Math. Statist.*, 25, 579-586, 1954.
- Clark, Peter K., "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices," *Econometrica*, 41, 135-155, 1973.
- Davidson, Russell, and James G. MacKinnon, Estimation and Inference in Econometrics, Oxford University Press, 1993.
- DeGroot, Morris, Probability and Statistics, Second Edition, Addison-Wesley, 1986.
- Fama, E., 1965, "The Behavior of Stock Prices," *Journal of Business* 38, 34-105.
- Fama, E., 1976, Foundations of Finance. New York: Basic Books.
- Fama, Eugene F., and Kenneth R. French, "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, 22, 3-25, 1988.
- Fama, Eugene F., and Kenneth R. French, "Business Conditions and Expected Returns on Stocks and Bonds," *Journal of Financial Economics*, 25, 23-49, 1989.
- Hagerman, Robert L., "More Evidence on the Distribution of Security Returns," *Journal of Finance*, 33, pp. 1213-1221, 1978.
- Judd, Kenneth L., Numerical Methods in Economics, The MIT Press, 1999.

- Li, Qi, "Semiparametric Estimation and Testing of Econometric Models," Department of Economics, University of Guelph (Canada), 1997.
- Mandelbrot, Benoit, "The Variation of Certain Speculative Prices," *Journal of Business*, 36, 394-419, 1963.
- McCulloch, J. Huston, "Numerical Approximation of the Symmetric Stable Distribution and Density," A Practical Guide to Heavy Tails, Adler, Feldman, & Taqqu (eds.), Birkhäuser, 1998, pp. 359-376.
- McCulloch, J. Huston, "Measuring the Term Structure of Interest Rates," *Journal of Business*, 44, 19-31, 1971.
- McCulloch, J. Huston, "Interest Risk Sensitive Deposit Insurance Premia: Stable ACH Estimates," *Journal of Banking & Finance*, Vol. 9, 137-156, 1985.
- McCulloch, J. Huston, "Financial Applications of Stable Distributions," Statistical Methods in Finance (Handbook of Statistics 14), G.S. Maddala and C.R.Rao, eds., Elsevier Science, Amsterdam, 1996.
- McCulloch, J. Huston, "Measuring Tail Thickness to Estimate the Stable Index α : A Critique," *Journal of Business & Economic Statistics*, January 1997, Vol. 15, pp. 74-81.
- McCulloch, J. Huston, "Linear Regression with Stable Disturbances," A Practical Guide to Heavy Tails, Adler, Feldman, & Taqqu (eds.), Birkhäuser, 1998, pp. 359-376.
- McCulloch, J. Huston, "A Spline Rao Score (LM) Test for Goodness of Fit: A Proposal," 1999.
- Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes, Introduction to the Theory of Statistics, Third Edition, McGraw-Hill, 1974.
- Nelson, D., "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59, 347-370, 1991.
- Perry, Philip R., "More Evidence on the Nature of the Distribution of Security Returns," *Journal of Financial and Quantitative Analysis*, 18, 211-221, 1983.
- Praetz, Peter D., "The Distribution of Share Price Changes," *Journal of Business*, 45, 49-55, 1972.
- Rayner, J.C.W., and D.J. Best, Smooth Tests of Goodness of Fit, Oxford University Press, 1989.
- Samorodnitsky, Gennady, and Murad S. Taqqu, Stable Non-Gaussian Random Processes: Stochastic Processes with Infinite Variance," Chapman and Hall, New York, 1994.
- Taylor, S., Modeling Financial Time Series, London: John Wiley & Sons, 1986.