A Goodness-of-Fit Test Using Relative Entropy

Jin-Chuan Duan and Mike K.P. So*

(This Draft: June 2001)

Abstract

The use of a more general distribution function is often justified by treating the proposed distribution function as an alternative hypothesis. The suitability of such a distribution is then determined by the outcome of a nested test on a standard distributional assumption, which is typically a restricted version of the proposed distribution. Although such a practice is known to be conceptually flawed, it frequently occurs in the literature simply due to the lack of a better alternative. Although one can, in principle, use the Kolmogorov-Smirnov (KS) test to tackle such a testing problem, there are two difficulties associated with the KS test. First, it is hard to implement when the parameter value(s) is/are unknown. Second, the power of the KS test is relatively low. In this paper, we propose a generic parametric testing procedure based on a relative entropy construction. We demonstrate through a simulation study the superior performance of this testing procedure. We also implement this testing procedure on financial data using the popular GARCH model.

Key Words: kurtosis, skewness, relative entropy, likelihood ratio test, Kolmogorov-Smirnov test, Anderson-Darling test, GARCH.

*Duan, Rotman School of Management, University of Toronto and Department of Finance, Hong Kong University of Science & Technology, jcduan@rotman.utoronto.ca. So, Department of Information and Systems Management, Hong Kong University of Science and Technology. Duan acknowledges the support received as the Manulife Chair in Financial Services at University of Toronto. The authors thank the editor, the associate editor and an anonymous referee for valuable comments. Correspondence to: Mike So, Department of Information and Systems Management, Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, Hong Kong; E-mail: immkpso@ust.hk; Tel: 852-2358 7726; Fax: 852-2358 1946.

1 Introduction

In this paper, we propose a generic testing procedure for assessing the appropriateness of any given probability density assumption. The distributional assumption is critically important in many statistical applications. Take financial applications as examples. The knowledge of asset return distributions determines how investment decisions are made, and consequently the risk-return profile of a portfolio. In day-to-day risk management practice, financial institutions would like to control their risk exposure due to potential adverse market movements. Such a concern is in part motivated by self-interest and in part by governmental regulations. A reliable assessment of such exposures clearly depends on the knowledge of asset return distributions.

When a particular distribution is proposed for a modeling application, a classical way of testing the appropriateness of such an assumption is based on the empirical distribution function (EDF). Two well-known examples are the Kolmogorov-Smirnov (KS) test and the Andersen-Darling test (Pearson and Hartley, 1972; Stephens, 1974; Stephens, 1986). In a nutshell, the EDF test statistics measure the distance between the theoretical and empirical distributions. With known parameter value(s), the distributions of the test statistics have already been derived. For example, the distribution of the KS statistic can be determined by the fact that the properly constructed empirical process weakly converges to the Brownian bridge process. However in actual implementation of the EDF tests, the parameter values of the null distribution are unknown. There are two main problems associated with the EDF tests. First, the asymptotic distribution of the EDF test statistic will depend on the class of null distributions being tested. The sampling error associated with the unknown parameter(s) creates problems. In general, the way for handling the problem of unknown parameter value(s) relies on simulation to generate critical values for testing. For example, in the special cases of the normality and exponential distribution assumptions, Lilliefors (1967,1969) offered tables of critical values for the KS test. The second problem is that the asymptotic distribution of the test statistic may depend on the true values of the unknown parameters (Stephens, 1986). It becomes inconvenient to apply the EDF approach as one has to generate critical values for different true parameter values and null distributions

The application of the KS test to dynamic models can be significantly more complicated. If one knows the time series parameter values, then the observed data series can be filtered to obtain the empirical conditional distribution. However in reality, one does not have the knowledge of the time series parameter values, and thus faces the combined problem of unknown distribution and time series parameters. This difficulty in part leads to a common practice in the literature that a more general distributional assumption of interest is treated as the alternative hypothesis instead of the null, and it is considered accepted if its restricted version is rejected.[1] Such a practice is theoretically flawed. The error of accepting the alternative distribution is not controlled and has an unknown magnitude. In a recent paper, Bai (1998) proposed a test of conditional distribution for dynamic models along the line of the KS test. The test utilizes the martingale transformation of the Brownian bridge process

[1] See, for example, Bollerslev (1987), Nelson (1991), Fernandez and Steel (1998), Theodossiou (1998), among others. Some authors were fully aware of the questionable validity of such an approach and explicitly stated a caution in the paper.

proposed by Khmaladze (1981) to remove the effect of sampling errors due to the unknown parameter estimates. Bai's (1998) testing procedure has fairly general applicability even though it is restricted to the case of strictly increasing distribution functions due to the need for obtaining uniformly distributed random variates by transformation. Apart from this restriction, Bai's (1998) testing procedure, although complicated, has effectively solved the long-standing problem of unknown parameter values for the EDF tests.

Our test differs from the EDF tests in an important way. Our method is highly parametric and relies on the standard asymptotic likelihood ratio test. When testing a distributional assumption either in static or dynamic models, a parametric specification of the distribution function under the null hypothesis is automatically provided. If a generic suitable alternative class of distribution functions can be identified, then the testing problem becomes rather straightforward. This is indeed what our testing approach is all about. We show that for an arbitrary probability distribution (either discrete or continuous), we can construct a class of alternative distribution functions that wrap around the distribution function under the null hypothesis. Our construction relies on the principle of minimum relative entropy and employs some design instruments. Since the likelihood ratio test is known to be a uniformly most powerful test (asymptotically), our test is expected to be more powerful than the EDF tests.

We show in the paper how easily such a test of the distributional assumption can be constructed. Specifically, we identify a set of design instruments and use them to construct a minimum relative entropy probability density function around the density function under the null hypothesis. The choice of the design instrument(s) only depends on the density under the null hypothesis. We demonstrate by a simulation study that the KS and Andersen-Darling tests for normality are dominated by our test (in terms of power of rejection) by a wide margin. We apply this testing procedure to stock index returns in the GARCH model setting. The commonly used conditional distribution functions such as Student's t and exponential power distributions (even after allowing for asymmetry as in Fernández and Steel, 1998) are found to be inadequate for this financial data series. Moreover, we have conducted a simulation analysis to ascertain that the relative entropy goodness-of-fit test has the right size and sufficient power in a dynamic location-scale model setting. Although the discussion of this paper is mainly confined to testing the probability density function, tests for discrete distribution functions can be equally easy to construct.

The balance of the paper is organized as follows. Section 2 provides a brief overview of the theory about minimum relative entropy distribution. The procedures and some asymptotic properties for constructing the relative entropy goodness-of-fit test are given in Section 3. Section 4 discusses the application of the relative entropy goodness-of-fit test to location-scale models. Simulation results for comparing the performance of classical EDF tests and the relative entropy goodness-of-fit test are also reported. The application to financial data is presented in Section 5 where a simulation study is also provided. Finally, Section 6 concludes.

2 Minimum relative entropy distribution - a brief review

The study of entropy can be dated back to Shannon (1948) and Jaynes (1957a & b). Shannon (1948) defined the entropy associated with the probability density function $p(x)$ as

$$E_p = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx. \tag{1}$$

Using an axiomatic approach, Shannon (1948) justified why entropy is a measure of uncertainty associated with a probability distribution. As advocated in Shannon (1948) and Jaynes (1957a & b) the principle of maximum entropy favors a probability distribution with a higher entropy because it represents a higher uncertainty or greater disorder. If partial information is available, the principle calls for choosing a $p(x)$ by maximizing the entropy in (1) subject to the constraints implied by the information. The resulting probability distribution is called the maximum entropy distribution (MED).

The principle of maximum entropy was generalized by Kullback and Leibler (1951) into that of relative entropy, sometimes referred to as the Kullback-Leibler distance, which measures the distance between two probability density functions $p(x)$ and $q(x)$ by

$$D(p(x), q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx. \tag{2}$$

Relative entropy is not a true distance function because it neither satisfies the triangular inequality nor is a symmetric relationship. It, however, behaves somewhat like a distance measure because $D(p(x), q(x)) = 0$ if and only if $p(x) = q(x)$ and $D(p(x), q(x)) \geq 0$ for any $p(x)$ and $q(x)$ (see Cover and Thomas (1991)).

The concept of relative entropy can be used to derive the minimum relative entropy density (MRED). Suppose that one has a prior belief about the distribution, say, $q(x)$, and the true distribution must also satisfy a set of k moment restrictions. It is thus natural to seek for a probability density function $p(x)$ that is closest to the prior probability density function and satisfies the moment restrictions. In other words, the MRED is the one that solves the following minimization problem:

$$\min_{p(x)} D(p(x), q(x))$$

$$\text{subject to} \quad \int_{-\infty}^{\infty} p(x)\, dx = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} g_i(x) p(x)\, dx = c_i, \quad i = 1, \cdots, k.$$

Using the Frechet derivative, the solution can be derived to be:

$$p(x; \lambda_1, \cdots, \lambda_k) = q(x) \exp\left(\lambda_0 + \sum_{i=1}^{k} \lambda_i g_i(x)\right), \tag{3}$$

$$\lambda_0 = -\ln\left\{\int_{-\infty}^{\infty} q(x) \exp\left(\sum_{i=1}^{k} \lambda_i g_i(x)\right) dx\right\}, \tag{4}$$

$$\int_{-\infty}^{\infty} g_i(x) p(x; \lambda_1, \cdots, \lambda_k)\, dx = c_i, \quad i = 1, \cdots, k. \tag{5}$$

The MED can be viewed as a special case of the MRED by setting $q(x) = 1$, which is analogous to using a non-informative uniform prior in Bayesian analysis.

3 The relative entropy goodness-of-fit test

Our objective is to devise a generic testing procedure that can be used to test a given distributional assumption in either static or dynamic model setting. The strategy is to use the relative entropy concept to construct a class of alternative distributions around the null distribution. Specifically, if $q(x)$ is the null density function, the MRED, $p(x;\lambda_1,\cdots,\lambda_k)$, described in the preceding section will represent the class of alternative distributions, subject to the choice of k design instruments, $\{g_i(x); i=1,\cdots,k\}$. The principle guiding the choice of design instruments will be discussed later.

3.1 Basic idea

For an *i.i.d.* sequence of data, we are interested in the following null hypothesis.

$$H_0: \text{ The probability density function for } y_t \text{ is } f^0(y_t;\theta).$$

The null probability density function is characterized by a parametric class of functions with parameter θ. One main issue in testing a null hypothesis is the construction of a suitable alternative hypothesis H_1. In order to derive a nested test, a natural strategy is to choose an alternative distribution containing the null distribution as a special case. For example, if the null is a normal distribution, we can choose Student's t distribution as the alternative in an attempt to discriminate H_0 and H_1 by the kurtosis of the data.[2] Similarly, if the null is an exponential distribution, a natural alternative would be the gamma distribution. However, it is less obvious in deciding on an appropriate alternative distribution when the null is rather complicated; for example, the asymmetric distributions in Fernández and Steel (1998) and Theodossiou (1998). In this paper, we propose a generic way of obtaining a suitable distribution as the alternative. Our choice of the alternative distribution is obtained through the information-theoretic approach described in the previous section. Our general form of the alternative probability density function is thus given by

$$f^1(x;\theta,\lambda_1,\cdots,\lambda_k) = f^0(x;\theta)\exp\left(\lambda_0+\sum_{i=1}^{k}\lambda_i g_i(x)\right), \tag{6}$$

$$\lambda_0 = -\ln\left\{\int_{-\infty}^{\infty} f^0(x;\theta)\exp\left(\sum_{i=1}^{k}\lambda_i g_i(x)\right)dx\right\}. \tag{7}$$

This function is simply the MRED using the constraints defined by k restrictions on some design instruments, $g_i(x)$:

$$\int_{-\infty}^{\infty} g_i(x) f^1(x;\theta,\lambda_1,\cdots,\lambda_k)\,dx = c_i, \quad i=1,\cdots,k. \tag{8}$$

The function in the form of (6) was referred to as an augmented density with exponential carrier function by Chesher and Smith (1997). Their approach was, however, not based on

[2]Strictly speaking, the use of Student's t distribution as the alternative in testing the null hypothesis of normal distribution is problematic. This is because the null hypothesis corresponds to the case of infinite degrees of freedom for Student's t distribution. Standard testing procedures typically require the null hypothesis to reside in the interior of the parameter set defined for the alternative hypothesis.

the minimum relative entropy principle, and the augmented density function was intended for testing moment restrictions via the use of the likelihood ratio test.

Our generic construction of the alternative distribution gives rise to a natural nested test based on the likelihood ratio principle. In practice, we only need to compute the maximum likelihood estimates of the unknown parameters under H_0 ($\lambda_1 = \cdots = \lambda_k = 0$) and H_1. Depending on the number of restrictions used in the construction, the likelihood ratio test statistic (two times the difference in the log-likelihood functional values under H_1 and H_0) is distributed asymptotically as $\chi^2(k)$ under H_0 with k being its degrees of freedom. One can also apply Bartlett adjustments (see Barndorff-Nielsen and Cox (1984)) to the likelihood ratio test to yield a better $\chi^2(k)$ approximation of the test statistic when the sample is small.

Under H_1, the distribution of y_i has the form given in (6). Let $Y = (y_1, \cdots, y_n)'$. The log-likelihood function under H_0 is $L_0(Y;\theta) = \sum_{t=1}^{n} \ln f^0(y_t;\theta)$, whereas the log-likelihood function under H_1 is

$$L_1(Y;\theta,\lambda_1,\cdots,\lambda_k) = \sum_{t=1}^{n} \ln f^0(y_t;\theta) + n\lambda_0 + \sum_{i=1}^{k} \lambda_i \left(\sum_{t=1}^{n} g_i(y_t) \right). \tag{9}$$

Recall that λ_0 is not a free parameter and must satisfy equation (7). Differentiating $L_1(Y;\theta,\lambda_1,\cdots,\lambda_k)$ with respect to θ and λ_i yields the following system of likelihood equations:

$$\frac{\partial L_1(Y;\theta,\lambda_1,\cdots,\lambda_k)}{\partial \theta} = \sum_{t=1}^{n} \frac{1}{f^0(y_t;\theta)} \frac{\partial f^0(y_t;\theta)}{\partial \theta} + n\frac{\partial \lambda_0}{\partial \theta} = 0 \tag{10}$$

$$\frac{\partial L_1(Y;\theta,\lambda_1,\cdots,\lambda_k)}{\partial \lambda_i} = n\frac{\partial \lambda_0}{\partial \lambda_i} + \sum_{t=1}^{n} g_i(y_t) = 0, \quad i = 1, ..., k. \tag{11}$$

Using equation (7), we have

$$\int_{-\infty}^{\infty} \frac{\partial f^0(x;\theta)}{\partial \theta} \frac{f^1(x;\theta,\lambda_1,\cdots,\lambda_k)}{f^0(x;\theta)} dx = \frac{1}{n} \sum_{t=1}^{n} \frac{1}{f^0(y_t;\theta)} \frac{\partial f^0(y_t;\theta)}{\partial \theta}, \tag{12}$$

$$\int_{-\infty}^{\infty} g_i(x) f^1(x;\theta,\lambda_1,\cdots,\lambda_k) dx = \frac{1}{n} \sum_{t=1}^{n} g_i(y_t). \tag{13}$$

Note that the right-hand side of equation (13) is the sample counterpart of the moment restriction of the k-th design instrument. Not surprisingly, the maximum likelihood estimates for λ_i's are the ones that solve the empirical counterparts of the restrictions on the design instruments. Actually, no prior knowledge on the values of c_i's is needed. This is because c_i's have been reparameterized and replaced with λ_i's. The result in (13) is thus expected.

If the null distribution is the true data generating distribution, the restrictions in (13) would make $\lambda_i \approx 0$ for $i = 1, \cdots, k$. On the other hand, if H_0 is wrong, the values of λ_i's need to differ significantly from zero to make (13) hold, and the departure causes a rejection of the null hypothesis. The values for θ and λ_i's can be obtained by numerically solving equations (12), (13) and (7). The integration involved in these equations requires a numerical solution.

3.2 Choice of design instruments

The specific choice of design instruments is crucial in setting up the test. If $g_i(x)$ are selected in such a way that $f^1(x;\theta,\lambda_1,\cdots,\lambda_k)$ defines the same parametric distribution as $f^0(x;\theta)$,

one will encounter the problem that $f^1(x;\theta,\lambda_1,\cdots,\lambda_k)$ is not uniquely defined by a single set of parameters, i.e., lack of identification. For example, consider the normal density as the null, i.e.,

$$f^0(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{14}$$

Using x and x^2 as the design instruments implies that the alternative density is

$$f^1(x;\mu,\sigma^2,\lambda_1,\lambda_2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}+\lambda_0+\lambda_1 x+\lambda_2 x^2\right), \tag{15}$$

which again defines a normal distribution. An identifiability problem arises because four parameters are used to define a two-parameter normal distribution. In general, if the null distribution is from the exponential family, then any linear combination of the sufficient statistics should not be used as design instruments.

The identifiability requirement is still not sufficient for choosing design instruments. Again consider the normal density as the null. If one wants to detect the departure from H_0 using the skewness and kurtosis of the true distribution, it appears natural to select $k=2$ with $g_1(x)=x^3$ and $g_2(x)=x^4$.[3] But such a choice may lead to problems. Since the alternative density function becomes $\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}+\lambda_0+\lambda_1 x^3+\lambda_2 x^4)$, the tail behavior of the alternative probability density function is governed by $\exp(\lambda_2 x^4)$. This leading term becomes explosive if $\lambda_2>0$, which, of course, cannot even qualify as a distribution function. If one constrains λ_2 to be non-positive, the testing problem becomes a test of parameter value at the boundary, i.e., $\lambda_2=0$, which is a situation that the standard likelihood ratio test becomes invalid. Another problem is of numerical nature. If λ_2 is tentatively positive in the iterative numerical solution process, the norming constant approaches infinity and causes a breakdown of the numerical procedure. Using x^3 and x^4 as the design instruments for testing normality is not desirable even if $\lambda_2\leq 0$, because the alternative density function can only to have normal or thin tails.

Although the choice of $g_i(x)$ appears arbitrary, some general rules do apply. These general rules are fairly intuitive and practical. Essentially, these rules depend on the null distribution employed and the direction of departure in which one is interested. If one is interested in detecting excess kurtosis, then it is advisable to set $k=1$ and use $g_1(x)=|x|^{1.5}$ or $g_1(x)=\ln(1+x^2)$. The first $g_1(x)$ is motivated by the exponential power distribution described in Box and Tiao (1992), whereas the second $g_1(x)$ is guided by the t distribution. If one is interested in both skewness and excess kurtosis, then it is natural to consider $k=2$ with $g_1(x)=(x^+)^{1.5}$ and $g_2(x)=(x^-)^{1.5}$, where $x^+=\max(x,0)$ and $x^-=\max(-x,0)$ are the positive and negative parts of x, respectively. Suppose that one wants to test departure from the t distribution. It is natural to use the design instrument $g_1(x)=|x|^{1.5}$ as opposed to $g_1(x)=\ln(1+x^2)$ because the latter one will generate a t-distribution like alternative density function. By the same token, to test departure from the exponential power distribution, one should consider $g_1(x)=\ln(1+x^2)$ instead of $g_1(x)=|x|^{1.5}$. In short, the choice of design instruments is related to the characteristics of the null density function.

[3] Zellner and Highfield (1988) had previously applied a similar idea to use the first four moments to determine the maximum entropy distribution as an approximation to the marginal posterior distribution.

3.3 Asymptotic properties

To yield the asymptotic $\chi^2(k)$ distribution for the likelihood ratio test statistic, some regularity conditions (see Serfling (1980), pages 144-145 and Lehmann (1999), page 501) need to be satisfied. The most crucial one is the boundedness condition:

$$\left|\frac{\partial^3 \ln f^1(x;\beta)}{\partial\beta_a\partial\beta_b\partial\beta_c}\right| \leq H(x) \tag{16}$$

for all x, where $\beta = (\theta', \lambda_1, ..., \lambda_k)'$ is the vector of parameters in the alternative distribution, β_a, β_b, β_c are any elements of β and $E[H(x)] < \infty$.

To ensure that the boundedness condition is met, we use the truncated design instruments in the following form:

$$g_i^*(x) = \begin{cases} g_i(c_l) & \text{if } x < c_l \\ g_i(x) & \text{if } c_l \leq x \leq c_u \\ g_i(c_u) & \text{if } x > c_u \end{cases} \tag{17}$$

where $g_i(x)$ is any continuous real-valued function. The values for c_l and c_u can be set to have sufficient range of coverage by $g_i^*(x)$. The boundedness condition in (16) is satisfied if the truncated design instruments are used and the following conditions, exclusively on the null distribution function, are imposed:

$$\begin{aligned} E\left|\frac{\partial f^0(x;\theta)/\partial\theta_a}{f^0(x;\theta)}\right| &< \infty, E\left|\frac{\partial^2 f^0(x;\theta)/\partial\theta_a\partial\theta_b}{f^0(x;\theta)}\right| < \infty, \\ E\left|\frac{\partial^3 f^0(x;\theta)/\partial\theta_a\partial\theta_b\partial\theta_c}{f^0(x;\theta)}\right| &< \infty, \text{ and } E\left|\frac{\partial^3 \ln f^0(x;\theta)}{\partial\theta_a\partial\theta_b\partial\theta_c}\right| < \infty \end{aligned} \tag{18}$$

for any parameters θ_a, θ_b and θ_c in $f^0(x;\theta)$. The technical details are available from the authors upon request. Besides providing the theoretical foundation of applying asymptotic χ^2 percentiles in the relative entropy goodness-of-fit test, the use of truncated design instruments actually help solve the numerical problem described in Section 3.2. Furthermore, truncation removes the non-positivity restriction on λ_i so that the standard likelihood ratio test can be applied without worrying about the boundary value problem.

4 Application to location-scale models

In location-scale models, the null hypothesis for the distribution specification can be formulated as

$$H_0: \text{ The probability density function for } y_t \text{ is } \frac{1}{\sigma}f^0(\frac{y_t-\mu}{\sigma};\theta).$$

The null probability density function is characterized by a parametric class of functions with parameters μ, σ and θ. This density function is constructed, without loss of generality, from a standardized density function, $f^0(x;\theta)$, whose location and scale parameters are constrained to be 0 and 1, respectively. Naturally, the density function applies to the data sample under the null hypothesis will be the one restored with the location parameter μ and the scale parameter σ.

4.1 Detailed procedure

We rely on the MRED construction described in the preceding section to construct the distribution for y_t under the alternative hypothesis. Specifically, the alternative probability density relative to $f^0(x;\theta)$ is

$$p(x;\theta,\lambda_1,\cdots,\lambda_k) = f^0(x;\theta)\ \exp\left(\lambda_0 + \sum_{i=1}^{k}\lambda_i g_i(x)\right). \tag{19}$$

Since $p(x;\theta,\lambda_1,\cdots,\lambda_k)$ need not belong to the class of distributions with mean 0 and variance 1, we need to standardize the density function. A standard variable transformation yields the following alternative density function suitable for $\frac{y_t-\mu}{\sigma}$:

$$f^1(x;\theta,\lambda_1,\cdots,\lambda_k) = \Phi_2\ p(\Phi_1+\Phi_2 x;\theta,\lambda_1,\cdots,\lambda_k), \tag{20}$$

where

$$\Phi_1 = \int_{-\infty}^{\infty} xp(x;\theta,\lambda_1,\cdots,\lambda_k)dx, \quad \Phi_2 = \sqrt{\int_{-\infty}^{\infty} x^2 p(x;\theta,\lambda_1,\cdots,\lambda_k)dx - \Phi_1^2}. \tag{21}$$

Note that this alternative density function only has k more parameters because λ_0 is a norming constant. If $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0$, then $f^1(x;\theta,\lambda_1,\cdots,\lambda_k) = f^0(x;\theta)$. A specification test can now be constructed using the standard asymptotic likelihood ratio test:

$$2\left[L_1(\mathbf{Y};\hat{\theta},\hat{\lambda}_1,\cdots,\hat{\lambda}_k,\hat{\mu},\hat{\sigma}) - L_0(\mathbf{Y};\bar{\theta},\bar{\mu},\bar{\sigma})\right] \sim \chi^2(k), \tag{22}$$

where $L_0(\mathbf{Y};\bar{\theta},\bar{\mu},\bar{\sigma})$ is the maximum value of the log-likelihood function under the null hypothesis with

$$L_0(\mathbf{Y};\theta,\mu,\sigma) = \sum_{t=1}^{n}\left[\ln f^0\left(\frac{y_t-\mu}{\sigma};\theta\right) - \ln\sigma\right], \tag{23}$$

and $L_1(\mathbf{Y};\hat{\theta},\hat{\lambda}_1,\cdots,\hat{\lambda}_k,\hat{\mu},\hat{\sigma})$ is the maximum value of the log-likelihood function under the alternative hypothesis with

$$L_1(\mathbf{Y};\theta,\lambda_1,\cdots,\lambda_k,\mu,\sigma) = \sum_{t=1}^{n}\left[\ln f^1\left(\frac{y_t-\mu}{\sigma};\theta,\lambda_1,\cdots,\lambda_k\right) - \ln\sigma\right]. \tag{24}$$

For a given null probability density function and a set of design instruments, we can numerically obtain the maximum likelihood estimates under both the null and alternative hypotheses. Recall that the design instruments need to undergo truncation as described in equation (17) so that potential numerical problems can be avoided and the test does not have to deal with the boundary value problem. In location-scale models, we can set $c_l = -m$ and $c_u = m$ in the truncation operation. Since the null density function $f^0(x;\theta)$ has already been normalized to have mean 0 and variance 1, truncating at $-m$ and m corresponds to using the range of m standard deviations from the mean of y_t. The specific value of m affects the power of the proposed test. This is not surprising because a very small m does not permit the density function under the alternative hypothesis to differ in any meaningful

way from the density function under the null hypothesis. Therefore, it is important to choose m so that the interval $(c_l, c_u) = (-m, m)$ has a reasonably high coverage probability under $f^0(x;\theta)$. If the null distribution is parameter-free, this can be determined fairly easily. For example, in testing for normality where $f^0(x;\theta)$ is standard normal, the choice of $m = 2$ has a coverage probability of around 95%. Even if there is unknown parameter in the standardized null density, a suitable choice of m can usually be obtained by referring to the nature of the density function under the null hypothesis. For later tests of fat-tailed distributions in dynamic models, we set $m = 8$ to ensure sufficient coverage. Indeed, this choice of m implies a coverage probability of more than 99% for standardized t distribution with degrees of freedom as small as 3. In practice, one can use an arbitrary large m as long as no numerical problem is created. The power sensitivity to the value of m will be examined later.

4.2 Power analysis

We now conduct a simulation study to compare the power of the KS and Anderson-Darling tests with that of our testing procedure. We choose these two particular EDF tests because both are common normality tests with the Anderson-Darling test being known to be more powerful in detecting discrepancy in the tails (Stephens, 1974). In this study, we take Student's t distribution as the true distribution and test to see how frequently the null hypothesis of normal distribution can be rejected by different tests. We vary the degrees of freedom for Student's t distribution in generating the data sets but always fix the mean and variance at zero and one. We perform the test assuming no knowledge of the mean and variance. The critical values for the MRED test is based on the $\chi^2(k)$ percentiles and those for the KS and Anderson-Darling tests are taken from Table 4.7 of Stephens (1986, page 123).

For the simulation study, we report the results using only one design instrument, i.e., $k = 1$, but repeat for two different design instruments. For the first case, $g_1(x) = |x|^{1.5}$ and for the second $g_1(x) = \ln(1 + x^2)$. We pick these two design instruments for the reasons provided earlier in Section 3.2. For the results reported in Figures 1a - 1d, we have set $m = 2$; that is, $c_l = -2$ and $c_u = 2$. The power functions in these figures were calculated using $10,000$ Monte Carlo repetitions for each scenario. We consider two sample sizes: 100 and 500 and two sizes of test: 5% and 10%. The horizontal axis is the inverse of the degrees of freedom of Student's t distribution and the vertical axis is the rejection rate. A horizontal value of zero represents an infinite degrees of freedom, i.e., normality. The rejection rate in this case thus reflects the size of the test. Our results suggest that the simulated sizes are very close to their theoretical values. The power functions clearly indicate that the relative entropy goodness-of-fit test has a better power in comparison to the KS and Anderson-Darling tests. From Figure 1, the superior performance of the entropy goodness-of-fit test is clear for both sample sizes. Furthermore, there does not exist a noticeable difference between the two specific design instruments when the sample size is 500.

In the power analysis, we have used $m = 2$ to truncate the design instruments. Recall that m actually represents the coverage from minus m to positive m standard deviations for location-scale models. Although m only needs to be finite theoretically, its value may affect the performance of our relative entropy goodness-of-fit test. We now examine how sensitive the simulation results are to the choice of m and report the results in Table 1 using $10,000$ repetitions. We consider the sample size of 100. For either design instrument

analyzed and for either 5% or 10% test, the power of the goodness-of-fit test does not appear to be materially affected by the choice of m. In other words, the actual value of m is not particularly important as long as its value provides, say, one standard deviation coverage on either side of the mean.

5 An application to financial data

In this section, we implement the relative entropy goodness-of-fit test on the S&P 500 index return. Daily closing values from the first business day of 1989 to the last day of 1998 are obtained from Datastream. We then construct the continuously compounded return series (scaled up by a factor of 100). Specifically, $y_t = 100 \times \ln(S_t/S_{t-1})$, where S_t is the index value at time t. The model employed is a variant of the popular dynamic location-scale model known as the non-linear asymmetric GARCH (NGARCH) process proposed by Engle and Ng (1993):

$$y_t = \mu + \delta\sigma_t + \sigma_t\epsilon_t, \qquad \epsilon_t \sim D(0,1), \tag{25}$$

$$\sigma_t^2 = \beta_0 + \beta_1\sigma_{t-1}^2 + \beta_2\sigma_{t-1}^2(\epsilon_{t-1} - c)^2, \tag{26}$$

where $\{\epsilon_t; t = 1, 2, \cdots\}$ are *i.i.d.* random variables and $D(0,1)$ denotes a distribution with mean zero and variance one. The usual NGARCH model parameter restrictions $\beta_0 > 0,\ \beta_1 \geq 0, \beta_2 \geq 0, \beta_1 + \beta_2(1+c^2) < 1$ are imposed to ensure that the conditional variance is positive and stationary variance exists.

Two commonly used conditional distributions in modeling financial data series are Student's t distribution advocated by Bollerslev (1987) and the exponential power distribution (EPD) (also known as generalized error distribution) by Nelson (1991). The choice of either Student's t or the EPD is typically justified by testing the null hypothesis of conditional normality. To our knowledge, there has been no formal analysis in the literature, except that of Bai (1998), of treating Student's t or the EPD as the null hypothesis.[4] In other words, such a popular practice in the empirical finance literature has not been rigorously scrutinized. In light of the above, we considered Student's t distribution and EPD as the null hypothesis. Since the normal distribution is their special case, we have effectively include the normal distribution in our study.

Let the null distribution with mean 0 and variance 1 be denoted by $f^0(\epsilon;\theta)$, where θ represents parameter(s) other than mean and variance. The null hypothesis under our dynamic location-scale model can be formulated as

$$H_0 : \text{The probability density function of } \epsilon_t \text{ is } f^0(\epsilon_t;\theta).$$

In order to perform the relative entropy goodness-of-fit test on the conditional distribution assumption, we need to choose the design instrument(s), $g_i(x)$. Since Student's t distribution and the EPD are two competing fat-tailed distributions, we can select $g_i(x)$ by taking advantage of our knowledge on the nature of these two density functions. Consider

[4] In Bai (1998), Student's t distribution was assumed to have a known degrees of freedom, and specifically the value was set equal to 5. Such an assumption is clearly not a realistic scenario for applications. In principle, Bai's (1998) procedure can deal with the case of unknown degrees of freedom by properly constructing a martingale transformation. We can only guess that he made such a restriction for the purpose of demonstration.

first the case of using only one design instrument, i.e., $k = 1$. We use $g_1(x) = |x|^{1.5}$ if the null hypothesis is Student's t distribution because it produces an alternative density with the EPD-like feature (with the degrees of freedom being fixed at 1.5). If the null hypothesis is the EPD, we use $g_1(x) = \ln(1 + x^2)$ to produce an alternative density with Student's t-like feature. Such choices are natural and likely to have a higher power in discriminating between the null and alternative hypotheses. For $k = 2$, we set $g_1(x) = (x^+)^{1.5}$and $g_2(x) = (x^-)^{1.5}$ if the null hypothesis is Student's t distribution. Similarly, we set $g_1(x) = \ln(1 + (x^+)^2)$ and $g_2(x) = \ln(1 + (x^-)^2)$ if the null hypothesis is the EPD. Utilizing both x^+ and x^- in formulating the test enables us to detect departure from the null distribution due to asymmetry. In other words, it is a simple device to detect skewness.

The test results for the S&P 500 index return data are presented in Table 2. For $k = 1$, Student's t distribution is not rejected at the 5% level, whereas the EPD is rejected at the 5% level. If we use $k = 2$, Student's t distribution is also rejected at the 5% level. Since Student's t distribution is not rejected under $k = 1$, but rejected under $k = 2$, the results suggest that rejection may simply be due to asymmetry in the conditional distribution. This issue is further examined below.

Since there is evidence of asymmetry in the conditional distribution, we modify the null distributions to allow for some degree of skewness. Following Fernández and Steel (1998), we construct the skewed version of the probability density function for Student's t distribution and EPD. Their construction introduces an extra parameter γ that enables the conditional distribution to exhibit asymmetry. We repeat the relative entropy goodness-of-fit test for the asymmetric versions of the two null distributions. The results in the second panel of Table 2 indicate that neither the asymmetric t distribution nor asymmetric EPD is suitable for this index return series. Allowing for asymmetry in the form of Fernández and Steel (1998) cannot change the result of rejection when two design instruments are used to detect potential asymmetry.

We conduct a simulation study to gain some insights about the power of the relative entropy goodness-of-fit test in the setting of the NGARCH model. The parameter values for $\mu, \beta_0, \beta_1, \beta_2$ and c in (25) and (26) are assumed to be $0.045, 0.013, 0.903, 0.053$ and 0.756 respectively. These values are the parameter estimates obtained from applying the NGARCH model with the conditional normal distribution to the S&P 500 index return series from 1989 to 1998. We have removed parameter δ from this simulation study because its estimate is found to be statistically insignificant, and removing one parameter speeds up considerably our computer intensive simulation study. Since the parameter values associated with the GARCH model are fairly typical values found in empirical finance literature, we fix them in all simulations but treat them as unknown parameters. In other words, our focus is on the effect caused by different parameter values of the true conditional distribution and by the null hypothesis employed.

We simulate 500 data series each with 1000 and 3000 observations using the EPD and t conditional distributions. The EPD's degrees of freedom is set at $1.1, 1.2, 1.3, 1.5$ and 2, whereas the degrees of freedom for the t distribution is set at $4, 5, 6, 7$ and ∞. The choice of these parameter values reflects the range reported in our empirical analysis of the real data in the preceding section. We include the case of 2 for the EPD and ∞ for the t distribution because they correspond to the normality assumption. In the simulation study, we only assess the performance of using one design instrument. The choice of this function depends on the null hypothesis as described in the preceding section of empirical analysis. We consider two significance levels (α): 5% and 10%. The rejection rates under various scenarios are reported

in two panels of Table 3.

The first panel corresponds to the case of the true distribution function being the EPD for different degrees of freedom. When the null hypothesis is the EPD, the rejection rate should approximately equal the size of the test. The results in the case of 1000 observations indicate that the simulated sizes are slightly off for both $\alpha = 5\%$ and 10%. If the sample size is increased to 3000, the simulated sizes generally improve. When the null hypothesis is the t distribution, the rejection rate reflects the power of our relative entropy goodness-of-fit test. The results suggest that the smaller the degrees of freedom of the true distribution, the higher is the power of the test. Furthermore, the larger the sample size, the higher is the power. These properties are hardly surprising.

Notice that the true distribution implied by the EPD assumption becomes a normal distribution in the case of the parameter value equal to 2. The null hypothesis of t distribution thus contains the true distribution. In other words, the rejection rate in this case will again reflect the size of the test, and is thus omitted from the table.

The results for the case that the true distribution is Student's t are reported in the second panel of Table 3. When the null hypothesis is the t distribution, the rejection rate reflects the size (α) of the test. Our results in the case of 1000 observations show that they are close to their respective nominal values of 5% and 10%. When the sample size is increased to 3000, the simulated sizes again improve. Similar to the earlier result, a smaller degrees of freedom leads to a higher rejection rate when the null hypothesis is not the true distribution. A larger sample size increases the power of the test. Although the relative entropy goodness-of-fit test relies on the asymptotic distribution, these results indicate that for the sample size of 1000, the test has already performed reasonably well for a nonlinear time series model like the NGARCH process. Given that a sample size of 1000 or more is fairly typical for financial data, the proposed test is indeed a practical device. As indicated by the rejection rates in Table 3, the power of the relative entropy goodness-of-fit test is very high. In short, the two different fat-tailed distributions can be easily differentiated.

6 Conclusion

We have devised a simple parametric goodness-of-fit test that can be used to examine whether a given distributional assumption is compatible with the data sample. Our test utilizes the concept of minimum relative entropy to construct a suitable class of alternative distributions around the distribution function under the null hypothesis. In our construction, a natural set of design instruments is used in obtaining the minimum relative entropy distribution, and the choice of design instruments only depends on the nature of the distribution under the null hypothesis and the direction of departure of interest. The relative entropy goodness-of-fit test is parametric and derived directly from the likelihood ratio test. It differs significantly from the KS-type test that uses the empirical distribution as the alternative. As compared with the KS-type test, our test is easier to implement and is shown by a simulation study to have a higher power.

In terms of the methodological contribution, the relative entropy goodness-of-fit test offers one simple solution to the long-standing problem of testing a distributional assumption when a suitable class of alternative distributions is not readily available. From an application perspective, distributional assumption always plays a critical role. A simple and powerful testing procedure such as this can be immensely useful in discriminating among competing models.

References

[1] Bai, J. (1998), "Testing Parametric Conditional Distributions of Dynamic Models," Working paper, Massachusetts Institute of Technology.

[2] Barndorff-Nielsen, O.E. and Cox, D.R. (1984), "Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood ratio estimator," *Journal of the Royal Statistical Society* B, 46, 483-495.

[3] Bollerslev, T. (1987), "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return," *Review of Economics and Statistics*, 69, 542-547.

[4] Box, G. and Tiao, G. (1992), *Bayesian Inference in Statistical Analysis*, John Wiley and Sons.

[5] Chesher, A. and Smith, R.J. (1997), "Likelihood Ratio Specification Tests," *Econometrica*, 65, 627-646.

[6] Cover, T.M. and Thomas, J.A. (1991), *Elements of Information Theory*, John Wiley and Sons.

[7] Engle, R.F. and Ng, V. (1993), "Measuring and Testing the Impact of News on Volatility," *Journal of Finance*, 48, 1749-1778.

[8] Fernández C. and Steel, M.F.J. (1998), "On Bayesian Modeling of Fat Tails and Skewness," *Journal of the American Statistical Association*, 93, 359-371.

[9] Jaynes, E.T. (1957a), "Information Theory and Statistical Mechanics," *Physics Review*, 106, 620-630.

[10] Jaynes, E.T. (1957b), "Information Theory and Statistical Mechanics II," *Physics Review*, 108, 171-190.

[11] Khmaladze, E.V. (1981), "Martingale Approach in the Theory of Goodness-of-Tests," *Theory of Probability and its Applications*, 26, 240-257.

[12] Kullback, S. and Leibler, R.A. (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79-86.

[13] Lehmann, E.L. (1999), *Elements of Large-Sample Theory*, New York: Springer-Verlag.

[14] Lilliefors, H.W. (1967), "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, 62, 399-402.

[15] Lilliefors, H.W. (1969), "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown," *Journal of the American Statistical Association*, 64, 387-389.

[16] Nelson, D.B. (1991), "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59, 347-370.

[17] Pearson, E.S. and Hartley, H.O. (1972), *Biometrika Tables for Statisticians, Vol. 2*, New York: Cambridge University Press.

[18] Shannon, C.E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379-423.

[19] Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley and Sons.

[20] Stephens, M.A. (1974), "EDF Statistics for Goodness-of-fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730-737.

[21] Stephens, M.A. (1986), "Tests Based on EDF Statistics," in *Goodness-of-fit Techniques*, New York: Marcel Dekker.

[22] Theodossiou, P. (1998), "Financial Data and the Skewed Generalized T Distribution," *Management Science*, 44, 1650-1661.

[23] Zellner, A. and Highfield, R.A. (1988), "Calculation of Maximum Entropy Distributions and Approximation of Marginal Posterior Distributions," *Journal of Econometrics*, 37, 195-209.

Table 1. The power of testing normality using the relative entropy goodness-of-fit test for different m. The sample size is 100.

		Degrees of freedom of t distribution			
	α	15	10	7	4
			m = 1		
$g_1(x) = \|x\|^{1.5}$	5%	0.121	0.196	0.345	0.725
$g_1(x) = \ln(1+x^2)$	5%	0.123	0.200	0.348	0.729
$g_1(x) = \|x\|^{1.5}$	10%	0.190	0.277	0.434	0.786
$g_1(x) = \ln(1+x^2)$	10%	0.192	0.278	0.436	0.787
			m = 3		
$g_1(x) = \|x\|^{1.5}$	5%	0.095	0.158	0.293	0.675
$g_1(x) = \ln(1+x^2)$	5%	0.113	0.201	0.349	0.719
$g_1(x) = \|x\|^{1.5}$	10%	0.178	0.265	0.421	0.780
$g_1(x) = \ln(1+x^2)$	10%	0.196	0.301	0.461	0.802
			m = 4		
$g_1(x) = \|x\|^{1.5}$	5%	0.110	0.185	0.331	0.715
$g_1(x) = \ln(1+x^2)$	5%	0.113	0.201	0.349	0.720
$g_1(x) = \|x\|^{1.5}$	10%	0.191	0.285	0.447	0.798
$g_1(x) = \ln(1+x^2)$	10%	0.196	0.303	0.463	0.803

Table 2. Test results for the S&P500 daily return under different null conditional distributions in the NGARCH(1,1) model. The alternative probability density is the minimum relative entropy density (MRED) constructed around the null density function using k design instruments.

	Distribution	log-likelihood	LRT statistic	p-value
		Symmetric null		
Null:	t	-2848.77		
MRED:	t with $k = 1$	-2848.73	0.08	0.7757
	t with $k = 2$	-2844.67	8.20	0.0165
Null:	EPD	-2859.96		
MRED:	EPD with $k = 1$	-2850.11	19.70	0.0000
	EPD with $k = 2$	-2847.64	24.64	0.0000
		Asymmetric null		
Null:	Asy. t	-2847.70		
MRED:	Asy. t with $k = 1$	-2847.69	0.01	0.9392
	Asy. t with $k = 2$	-2838.52	18.35	0.0001
Null:	Asy. EPD	-2857.41		
MRED:	Asy. EPD with $k = 1$	-2847.04	20.75	0.0000
	Asy. EPD with $k = 2$	-2838.84	37.13	0.0000

Table 3. Power of the relative entropy goodness-of-fit test (with one design instrument). The numbers are the rejection rates in 500 repetitions under two true conditional distributions: EPD (the first panel) and t (the second panel) with different degrees of freedom.

		Degrees of freedom for EPD				
Null distribution	α	2	1.5	1.3	1.2	1.1
			n = 1000			
EPD	5%	0.080	0.074	0.070	0.066	0.060
	10%	0.140	0.134	0.128	0.120	0.126
t	5%		0.328	0.592	0.734	0.832
	10%		0.506	0.770	0.864	0.890
			n = 3000			
EPD	5%	0.072	0.064	0.074	0.080	0.078
	10%	0.128	0.108	0.108	0.116	0.112
t	5%		0.878	0.988	0.994	0.978
	10%		0.934	0.998	0.996	0.978

		Degrees of freedom for t				
Null distribution	α	∞	7	6	5	4
			n = 1000			
EPD	5%		0.548	0.656	0.782	0.896
	10%		0.670	0.760	0.870	0.932
t	5%	0.072	0.068	0.058	0.058	0.064
	10%	0.130	0.134	0.124	0.124	0.142
			n = 3000			
EPD	5%		0.952	0.984	0.994	0.998
	10%		0.978	0.992	0.994	0.998
t	5%	0.036	0.064	0.066	0.054	0.054
	10%	0.082	0.122	0.120	0.118	0.104

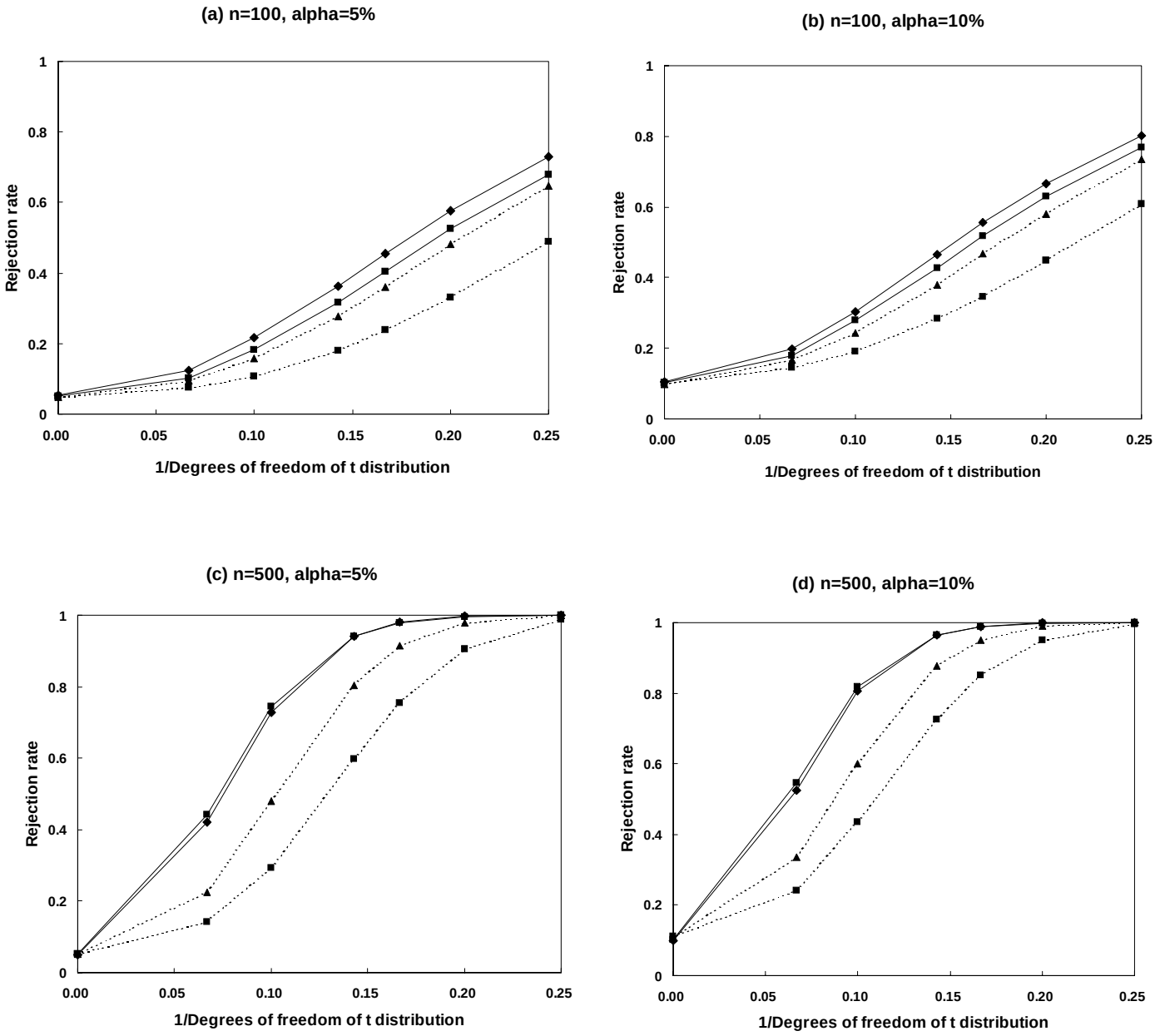

Figure 1. Power curve for the tests of normality based on 10,000 Monte Carlo repetitions:

, Kolmogorov-Smirnov test; , Anderson-Darling test; , MRED test, $g_1(x) = |x|^{1.5}$; , MRED test, $g_1(x) = \ln(1+x^2)$.