This article was downloaded by: [Texas A&M University Libraries]

On: 16 May 2013, At: 00:25

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Econometric Reviews

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/lecr20>

Information-Theoretic Distribution Test with Application to Normality

Thanasis Stengos^a & Ximing Wu^b

^a Department of Economics, University of Guelph, Guelph, Ontario, Canada

^b Department of Agricultural Economics, Texas A&M University, College Station, Texas, USA
Published online: 07 Jan 2010.

To cite this article: Thanasis Stengos & Ximing Wu (2009): Information-Theoretic Distribution Test with Application to Normality, *Econometric Reviews*, 29:3, 307-329

To link to this article: <http://dx.doi.org/10.1080/07474930903451565>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

INFORMATION-THEORETIC DISTRIBUTION TEST WITH APPLICATION TO NORMALITY

Thanasis Stengos¹ and Ximing Wu²

¹Department of Economics, University of Guelph, Guelph, Ontario, Canada

²Department of Agricultural Economics, Texas A&M University,
College Station, Texas, USA

□ We derive general distribution tests based on the method of maximum entropy (ME) density. The proposed tests are derived from maximizing the differential entropy subject to given moment constraints. By exploiting the equivalence between the ME and maximum likelihood (ML) estimates for the general exponential family, we can use the conventional likelihood ratio (LR), Wald, and Lagrange multiplier (LM) testing principles in the maximum entropy framework. In particular, we use the LM approach to derive tests for normality. Monte Carlo evidence suggests that the proposed tests are compatible with and sometimes outperform some commonly used normality tests. We show that the proposed tests can be extended to tests based on regression residuals and non-i.i.d. data in a straightforward manner. An empirical example on production function estimation is presented.

Keywords Distribution test; Maximum entropy; Normality.

JEL Classification C1; C12; C16.

1. INTRODUCTION

Testing that a given sample comes from a particular distribution has been one of the most important topics in inferential statistics and can be traced back to as early as Pearson's (1900) χ^2 goodness-of-fit test. Thanks to the prominent role of the central limit theorem in statistics, testing for normality has received an extensive treatment in the literature, see Thode (2002) for a comprehensive review on this topic. In this article, we present some alternative normality tests based on the method of maximum entropy (ME) density. The proposed tests are derived from maximizing the differential entropy subject to known moment constraints. By exploiting

Address correspondence to Ximing Wu, Department of Agricultural Economics, Texas A&M University, College Station, TX 77843-2124, USA; E-mail: xwu@ag.tamu.edu

the equivalence between ME and maximum likelihood (ML) estimation for the exponential family, we can use the conventional likelihood ratio (LR), Wald, and Lagrange multiplier (LM) testing principles in the maximum entropy framework. Hence, our tests share the optimality properties of the standard maximum likelihood based tests. Using the LM approach, we show that the ME method leads to simple yet powerful tests for normality. We propose some flexible maximum entropy densities characterized by a small number of generalized moments, which nest the normal density as a special case. The corresponding tests utilize some generalized moments that effectively capture deviations from the normal distribution. Our Monte Carlo simulations show that the proposed tests compare favorably and often outperform some commonly used tests in the literature, especially when the sample size is small. In addition, we show that the proposed method can be easily extended to: i) other distributions than the normal distribution; ii) regression residuals; iii) dependent and/or heteroskedastic data. Finally, we apply the proposed normality tests to residuals from a regression model of a production function using a benchmark dataset that has been extensively used in the literature.

The article is organized as follows. In the next section, we present the information theoretic framework on which we base our analysis. We then proceed to derive our normality tests and discuss their properties. In the following section we present some simulation results. Finally, before we conclude, we discuss some possible extensions and an empirical application. The appendix collects the proofs of the main results.

2. INFORMATION-THEORETIC DISTRIBUTION TEST

2.1. Maximum Entropy Density

According to Golan (2008), which provides an excellent review and synthesis of Information and Entropy Econometrics (IEE), "... , IEE is the subdiscipline of processing information from limited and noisy data with minimal a priori information on the data-generating process. In particular, IEE is a research that directly or indirectly builds on the foundations of Information Theory and the principle of [ME]."

Information entropy, the central concept of information theory, was introduced by Shannon (1949). Entropy is an index of disorder and uncertainty. Facing the fundamental question of drawing inferences from limited and insufficient data, Jaynes proposed the ME principle, which he viewed as a generalization of Bernoulli and Laplace's Principle of Insufficient Reason. The ME principle states that among all distributions that satisfy certain informational constraints, one should choose the one that maximizes Shannon's information entropy. According to Jaynes (1957), the ME distribution is "uniquely determined as the one which

is maximally noncommittal with regard to missing information, and that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters.” Shore and Johnson (1980) developed axiomatic foundations for this approach.

The ME density is obtained by maximizing the entropy subject to some moment constraints. Let x be a random variable distributed according to a probability density function (pdf) $f_0(x)$, and $X_1, X_2, \dots, X_n$ be an i.i.d. random sample from $f_0(x)$. The unknown density $f_0(x)$ is assumed to be continuously differentiable, positive on the interval of support (usually the real line if there is no prior information on the support of the density) and bounded. Suppose we maximize the entropy

$$\max_f : W = - \int f(x) \log f(x) dx,$$

subject to

$$\begin{aligned} \int f(x) dx &= 1, \\ \int g_k(x) f(x) dx &= \hat{\mu}_k, \quad k = 1, 2, \dots, K, \end{aligned}$$

where $g_k(x)$ is continuously differentiable and $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n g_k(X_i)$. This constrained optimization problem is called a *primal* formulation, where the operation is carried out with respect to the underlying density function $f(x)$. Alternatively, one can formulate this problem as an unconstrained optimization. The unconstrained objective function then takes the form

$$\begin{aligned} \max_{\theta_0, \dots, \theta_K} & - \int f(x) \log f(x) dx - \theta_0 \left(\int f(x) dx - 1 \right) \\ & - \sum_{k=1}^K \theta_k \left(\int g_k(x) f(x) dx - \hat{\mu}_k \right). \end{aligned}$$

This *dual* formulation offers several advantages. First, an unconstrained optimization is simpler, second the dimension of the optimization problem is reduced to K LMs, and third, that formulation allows a direct comparison with traditional likelihood methods.

It is straightforward to derive the solution, from the dual formulation,

$$f(x; \hat{\theta}) = \exp \left(-\hat{\theta}_0 - \sum_{k=1}^K \hat{\theta}_k g_k(x) \right), \quad (1)$$

where $\hat{\theta}_k$ is the LM associated with the k th moment constraint in the optimization problem, and $\hat{\theta}_0 = \log \left(\int \exp \left(- \sum_{k=1}^K \hat{\theta}_k g_k(x) \right) dx \right) < \infty$ ensures that $f(x; \hat{\theta})$ integrates to one. The maximized entropy $W = \hat{\theta}_0 + \sum_{k=1}^K \hat{\theta}_k \hat{\mu}_k$.

The ME density is of the generalized exponential family and can be completely characterized by the moments $Eg_k(x)$, $k = 1, 2, \dots, K$. We call these moments “characterizing moments,” whose sample counterparts are the sufficient statistics of the estimated ME density $f(x; \hat{\theta})$. A wide range of distributions belong to this family. For example, the Pearson family and its extensions described in Cobb et al. (1983), which nest normal, beta, gamma, and inverse gamma densities as special cases, are all ME densities with simple characterizing moments.

In general, there is no analytical solution for the ME density problem, and nonlinear optimization methods are required (Ornermite and White, 1999; Wu, 2003; Zellner and Highfield, 1988). We use Lagrange’s method to solve this problem by iteratively updating θ

$$\hat{\theta}_{(t+1)} = \hat{\theta}_{(t)} - \mathcal{H}^{-1}b,$$

where at the $(t+1)$ th stage of the updating, $b_k = \int g_k(x)f(x; \hat{\theta}_{(t)})dx - \hat{\mu}_k$ is the difference between the predicted and the empirical moment, and the Hessian matrix $\mathcal{H}$ takes the form

$$\mathcal{H}_{k,j} = \int g_k(x)g_j(x)f(x; \hat{\theta}_{(t)})dx, \quad 0 \leq k, j \leq K.$$

This updating scheme is essentially the Newton–Raphson algorithm. The positive-definiteness of the Hessian ensures the existence of a unique solution.¹

Given Eq. (1), we can also estimate $f(x; \theta)$ using ML. The maximized log-likelihood

$$\begin{aligned} l &= \sum_{i=1}^n \log f(x_i; \hat{\theta}) = - \sum_{i=1}^n \left(\hat{\theta}_0 + \sum_{k=1}^K \hat{\theta}_k g_k(x_i) \right) \\ &= -n \left(\hat{\theta}_0 + \sum_{k=1}^K \hat{\theta}_k \hat{\mu}_k \right) = -nW. \end{aligned}$$

¹Let $\gamma' = [\gamma_0, \gamma_1, \dots, \gamma_K]$ be a nonzero vector and $g_0(x) = 1$, we have

$$\gamma' \mathcal{H} \gamma = \sum_{k=0}^K \sum_{j=0}^K \gamma_k \gamma_j \int g_k(x)g_j(x)f(x, \theta)dx = \int \left(\sum_{k=0}^K \gamma_k g_k(x) \right)^2 f(x; \theta)dx > 0.$$

Hence, $\mathcal{H}$ is positive-definite.

Therefore, when the distribution is of the generalized exponential family, ML and ME estimates are equivalent provided that the sample counterparts of $\mu_k, k = 1, \dots, K$, are known. Moreover, they are also equivalent to the method of moments (MM) estimator. This ME/ML/MM estimator only requires the knowledge of sample characterizing moments.

Although ML and ME are equivalent in our case, there are some conceptual differences. For ML, the restricted estimates are obtained by imposing certain constraints on the parameters. In contrast, for ME, the dimension of parameters is determined by the number of moment restrictions imposed: the more moment restrictions, the more complex and thus the more flexible the distribution is. To reconcile these two methods, we note that a ME estimate with m moment restrictions has a solution of the form

$$f(x; \theta) = \exp\left(-\theta_0 - \sum_{k=1}^m \theta_k g_k(x)\right),$$

which implicitly sets $\theta_j, j = m+1, m+2, \dots$, to be zero. When we impose more moment restrictions, say, $\int g_{m+1}(x)f(x; \theta)dx = \hat{\mu}_{m+1}$, we let the data choose the appropriate value of θ_{m+1} .² In this sense, the estimate with more moment restrictions is in fact less restricted, or more flexible. ME and ML share the same objective function (up to a proportion) which is determined by the moment restrictions of the ME problem. Therefore, one can regard the ME approach as a method of model selection, which generates a ML solution.

2.2. Information Theoretic Estimators and Tests

The development of IEE, which is founded on information theory and the ME principle, is greatly affected by advances in statistics and econometrics. Maasoumi (1993), Ebrahimi et al. (1999), Bera and Biliias (2002), Golan (2002, 2007, 2008), and Golan and Maasoumi (2008) provide excellent reviews and a synthesis of IEE during the last century. In particular, Fig. 2.1 and 2.2 of Golan (2008) present a long-term and a short term history of IEE. Golan (2002, 2007) provide overviews based on special issues on IEE in *Journal of Econometrics* Vol. 107 and Vol. 138, respectively, while Golan and Maasoumi (2008) offer a review in a special issue of *Econometric Reviews*, Vol. 27. In the *Econometric Reviews* special issue on IEE, Golan and Maasoumi (2008) review the links between information

²Denote $\theta_m = [\theta_1, \dots, \theta_m]$. The only case that $\theta_{m+1} = 0$ is when the moment restriction $\int g_{m+1}(x)f(x; \theta_m)dx = \hat{\mu}_{m+1}$ is not binding, or the $(m+1)$ th moment is identical to its prediction based on the ME density $f(x; \theta_m)$ from the first m moments. In this case, the $(m+1)$ th moment contains no additional information that can further reduce the entropy.

measures and hypothesis testing as well as applications to dynamic models, Bayesian econometrics, empirical likelihood methods, and nonparametric econometrics. In this article, we will focus primarily on IEE and testing.

The early work that influenced the philosophy and approach of IEE include Pearson's work on goodness-of-fit measure and MM, Fisher's ML method, and later on Neyman and Pearson's Minimum chi-square method, and Sargan's Instrumental Variables method (see Bera and Biliias, 2002 and references therein). Hansen (1982) developed the general theory of the Generalized Method of Moments (GMM) which builds on all of the previous work. GMM recognizes while it does not specify the complete distribution of the data, the economic model does place restriction on population moment conditions. GMM thus bases its model construction and parameter estimation on population moment restrictions.

The development of GMM shares the same basic philosophy of some recent Information-Theoretic (IT) methods. At about the same time, the foundations of the empirical likelihood (EL) were established (Owen, 1988; Qin and Lawless, 1994). This method proposes a nonparametric likelihood method without assuming knowledge of the exact likelihood of the underlying data generating process. The connection of the GMM to IEE and IT was later established in some recent econometrics literature (Imbens et al., 1998; Kitamura and Stutzer, 1997). The EL method is further extended to the Generalized Empirical Likelihood (GEL) method (Imbens, 2002; Kitamura, 2006; Smith, 2004, 2005).

In a parallel and independent research, in the late 1980s and early 1990s, the ME method was generalized by Golan et al. (1996). This line of research develops an estimation method that is capable of handling ill-posed problem. It imposes minimal distributional assumption, can incorporate exact or stochastic moment conditions and incorporate prior information in a straightforward manner. This method, known as the Generalized Maximum Entropy (GME) estimator, provides a viable alternative to the GEL family estimators. One important advantage of the GME is that it is simpler to calculate, while the GEL family estimators are typically associated with difficult saddlepoint optimization problems.

A more recent addition to the IEE family estimators is the Bayesian Method of Moments (BMOM) by Zellner (1996). To avoid a likelihood function Zellner proposed to maximize the differential (Shannon) entropy subject to the empirical moments of the data. This yields the most conservative (closest to uniform) post data density. In that way the BMOM uses only assumptions on the realized error terms which are used to derive the post data density. To do so, the BMOM equates the posterior expectation of a function of the parameter to its sample value and chooses the posterior to be the (differential) ME distribution subject to that constraint.

Hypothesis testing based on the principle of ME has been discussed thoroughly since the original article of Jaynes (1957). Because the LMs θ_k take the value zero if the k th constraint is not binding, a *directional* test with respect to a given moment condition is equivalent to testing $\theta_k = 0$. In fact, this test principle is consistent with Neyman's (1937) smooth test, which appeared even earlier. This test is essentially an information-theoretic test in the sense that the test statistic is equivalent to the entropy of the uniform distribution under the null hypothesis.

Another important IT concept, the Kullback–Leibler Information Criterion (KLIC), is also commonly used in statistical estimations and inferences. Let f and g be two distributions with a common support. The KLIC is defined as

$$\int f(x) \log \frac{f(x)}{g(x)} dx,$$

which is non-negative and equals zero if and only if $f(x) = g(x)$ almost everywhere. Note that the KLIC is not a true distance as it is asymmetric and does not satisfy the triangle inequality. This criterion measures the discrepancy between two distributions. Since the KLIC is very sensitive to even a small discrepancy between two distributions, it is expected to perform well for estimations and tests on distributions. Haberman (1984) used the KLIC to select distributions satisfying a vector of moment conditions.

Starting with the seminal work by Owen (1988), the IT approach is further generalized to regressions and general inferences. In addition, a more general family of discrepancy measure is used. Consider two discrete distributions with common support $p = (p_1, \dots, p_n)$ and $q = (q_1, \dots, q_n)$. Cressie and Read (1984) proposed a family of discrepancy statistics

$$I_\lambda(p, q) = \frac{1}{\lambda(1 + \lambda)} \sum_{i=1}^n \left[\left(\frac{p_i}{q_i} \right)^2 - 1 \right],$$

which is indexed by a single parameter λ .

One can then define a family of estimators as

$$\min_{\pi, \theta} I_\lambda(1/n, \pi), \quad \text{subject to } \sum_{i=1}^n \psi(z_i, \theta) \pi_i = 0 \quad \text{and} \quad \sum_{i=1}^n \pi_i = 1, \quad (2)$$

where $E[\psi(z_i, \theta)] = 0$ is given moment conditions, and $1/n$ is the empirical distribution of the data. Instead of using the empirical distribution, estimator (2) reweighs observations such that they satisfy given moment conditions. Associated with the moment conditions is a

vector of Lagrangian multipliers, which essentially determines the *consistent* distribution π_i , $i = 1, \dots, n$. This estimator has several special cases of interest. For example, when $\lambda = 1$, it is the empirical likelihood estimator (Owen, 1988; Qin and Lawless, 1994), which is defined as

$$\min_{\pi, \theta} \sum_{i=1}^n \ln \pi_i, \quad \text{subject to } \sum_{i=1}^n \psi(z_i, \theta) \pi_i = 0 \quad \text{and} \quad \sum_{i=1}^n \pi_i = 1.$$

When $\lambda \rightarrow 0$, the estimator takes the form

$$\min_{\pi, \theta} \sum_{i=1}^n \pi_i \ln \pi_i, \quad \text{subject to } \sum_{i=1}^n \psi(z_i, \theta) \pi_i = 0 \quad \text{and} \quad \sum_{i=1}^n \pi_i = 1,$$

which is empirical tilting estimator corresponding to the KLIC (Imbens et al., 1998; Kitamura and Stutzer, 1997). When $\lambda = 2$, it is equivalent to the continuously updating GMM estimator of Hansen et al. (1996), defined by

$$\min_{\pi, \theta} \sum_{i=1}^n \frac{1}{n} (n^2 \pi_i^2 - 1), \quad \text{subject to } \sum_{i=1}^n \psi(z_i, \theta) \pi_i = 0 \quad \text{and} \quad \sum_{i=1}^n \pi_i = 1.$$

All these estimators are first order equivalent to the classical GMM estimator and enjoy certain higher order efficiency advantages.

IT overidentification tests follow naturally from the above generalized empirical likelihood estimators. Imbens et al. (1998) discussed three formulations. The Average Moment test compares the estimated moments to zero, which is similar to the J statistic of the classical GMM estimator. The LMs test is a direct test on the LMs of the minimization problem (2), which is the same approach as used in this study. Lastly, the Criterion Function test examines the discrepancy between the empirical distribution and the estimated distribution that satisfied given condition mean conditions. All three tests have asymptotic χ^2 distributions under the null hypothesis. Imbens et al. (1998) show that these IT tests have better small sample performances compared to the conventional overidentification GMM test.

In this article, we use the classical ME approach for distribution tests. Consider a M dimension parameter space Θ_M . Suppose we want to test the hypothesis that $\theta \in \Theta_m$, a subspace of Θ_M , where $m \leq M$. Because of the equivalence between ME and ML, we can use the traditional LR, Wald, and LM principles to construct test statistics.³ For $j = m, M$, let θ_j be the

³Imbens et al. (1998) discussed similar tests in the IT generalized empirical likelihood framework. The proposed tests differ from their tests, which minimize the discrete Kullback–Leibler information criterion (cross entropy) or other Cressie–Read family of discrepancy indices subject to moment constraints.

ML estimates in Θ_j, l_j and W_j be their corresponding log-likelihood and ME, we have

$$\begin{aligned}
 & -f(x; \theta_m) \log f(x; \theta_m) dx \\
 &= \int \left(\sum_{k=0}^m \theta_{m,k} g_k(x) \right) f(x; \theta_m) dx \\
 &= \sum_{k=0}^m \theta_{m,k} \int g_k(x) f(x; \theta_m) dx = \sum_{k=0}^m \theta_{m,k} \int g_k(x) f(x; \theta_M) dx \\
 &= \int \left(\sum_{k=0}^m \theta_{m,k} g_k(x) \right) f(x; \theta_M) dx = - \int f(x; \theta_M) \log f(x; \theta_m) dx.
 \end{aligned}$$

The fourth equality follows because the first m moments of $f(x; \theta_m)$ are identical to those of $f(x; \theta_M)$. Consequently, the log-likelihood ratio

$$\begin{aligned}
 \text{LR} &= -2(l_m - l_M) = 2n(W_m - W_M) \\
 &= -2n \left(\int f(x; \theta_m) \log f(x; \theta_m) dx - \int f(x; \theta_M) \log f(x; \theta_M) dx \right) \\
 &= 2n \left(\int f(x; \theta_M) \log f(x; \theta_M) dx - \int f(x; \theta_M) \log f(x; \theta_m) dx \right) \\
 &= 2n \int f(x; \theta_M) \log \frac{f(x; \theta_M)}{f(x; \theta_m)} dx,
 \end{aligned}$$

which is the Kullback–Leibler distance between $f(x; \theta_M)$ and $f(x; \theta_m)$ multiplied by twice the sample size. Hence if the true model $f(x; \theta_M)$ nests $f(x; \theta_m)$, the quasi-ML estimate $f(x; \theta_m)$ minimizes the Kullback–Leibler statistic between $f(x; \theta_M)$ and $f(x; \theta_m)$, as shown in White (1982).

If we partition $\theta_u = (\theta_m, \theta_{M-m}) = (\theta_{1u}, \theta_{2u})$ for the unrestricted model and similarly $\theta_r = (\theta_{1r}, 0)$ for the restricted model, then the score function

$$S(x; \theta_m, \theta_{M-m}) = \begin{pmatrix} \frac{\partial \ln f}{\partial \theta_m}(x; \theta_m, \theta_{M-m}) \\ \frac{\partial \ln f}{\partial \theta_{M-m}}(x; \theta_m, \theta_{M-m}) \end{pmatrix},$$

and the Hessian

$$\mathcal{H}(x; \theta_m, \theta_{M-m}) = \begin{pmatrix} \frac{\partial^2 \ln f}{\partial \theta_m \partial \theta'_m}(x; \theta_m, \theta_{M-m}) & \frac{\partial^2 \ln f}{\partial \theta_m \partial \theta'_{M-m}}(x; \theta_m, \theta_{M-m}) \\ \frac{\partial^2 \ln f}{\partial \theta_{M-m} \partial \theta'_m}(x; \theta_m, \theta_{M-m}) & \frac{\partial^2 \ln f}{\partial \theta_{M-m} \partial \theta'_{M-m}}(x; \theta_m, \theta_{M-m}) \end{pmatrix}.$$

We also partition similarly the inverse of the information matrix $\mathcal{J} = -E(H)$ as

$$\mathcal{J}^{-1} = \begin{pmatrix} \mathcal{J}^{11} & \mathcal{J}^{12} \\ \mathcal{J}^{21} & \mathcal{J}^{22} \end{pmatrix}.$$

The Wald test statistic is then defined as

$$\text{WALD} = n\hat{\theta}'_{2u}(\hat{\mathcal{J}}^{22})^{-1}\hat{\theta}_{2u},$$

and the LM test statistic is defined as

$$\text{LM} = \frac{1}{n} \sum_{i=1}^n \hat{S}(x_i; \hat{\theta}_{1r}, 0)' \hat{\mathcal{J}}^{-1} \sum_{i=1}^n \hat{S}(x_i; \hat{\theta}_{1r}, 0).$$

All three tests are asymptotically equivalent and distributed as χ^2 with $(M - m)$ degrees of freedom under the null hypothesis (see for example, Engle, 1984).

3. TESTS OF NORMALITY

In this section, we use the proposed ME method to derive tests for normality. Since the LR and the Wald procedures require the estimation of the unrestricted ME density, which in general has no analytical solution and can be computationally involved, we focus on the LM test, which enjoys a simple closed form.

3.1. Flexible ME Density Estimators

Suppose a density can be rewritten as or approximated by a sufficiently flexible ME density

$$f_0(x) = \exp\left(-\sum_{k=0}^2 \theta_k x^k - \sum_{k=3}^K \theta_k g_k(x)\right).$$

Two conditions are required to ensure that $f_0(x)$ is integrable over the real line. First, the dominant term in the exponent must be an even function; otherwise, $f_0(x)$ will explode at either tail as $|x| \rightarrow \infty$. Second, the coefficient associated with the dominant term, which is an even function by the first condition, must be positive; otherwise $f_0(x)$ will explode to ∞ at both tails as $|x| \rightarrow \infty$.

The LM test of normality amounts to testing whether $\theta_k = 0$ for $k=3, \dots, K$. In practice, only a small number of moments $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n g_k(x_i)$ are used for the test, especially when the sample size

is small. In this article, we consider three simple, yet flexible functional forms. To avoid scale effect, it is assumed that the data have been standardized throughout the text.

If we approximate $f_0(x)$ using the ME density subject to the first four arithmetic moments, the solution takes the form

$$f_1(x) = \exp\left(-\sum_{k=0}^4 \theta_k x^k\right).$$

This classical exponential quartic density was first discussed by Fisher (1922) and studied in the maximum entropy framework in Zellner and Highfield (1988), Ornermite and White (1999), and Wu (2003).

In practice, it is well known that the third and fourth sample moments can be sensitive to outliers. In addition to the robustness consideration, Dalén (1987) shows that sample moments are restricted by sample size, which makes higher order moments unsuitable for small sample problem. A third problem with the quartic exponential form is that this specification does not admit $\mu_4 > 3$ if $\mu_3 = 0$, where μ_i denotes the i th arithmetic moment. To see this point denote $\boldsymbol{\mu} = [\mu_1, \dots, \mu_4]$. Stohs (2003) shows that for the one-to-one mapping $\boldsymbol{\theta} = M(\boldsymbol{\mu})$, the gradient matrix $\mathcal{H}$ with $\mathcal{H}_{ij} = \mu_{i+j} - \mu_i \mu_j$, $1 \leq i, j \leq 4$, is positive definite and so is $\mathcal{H}^{-1}$. Denote $\mathcal{H}^{(4,4)}$ the lower-right-corner entry of $\mathcal{H}^{-1}$. It follows $\mathcal{H}^{(4,4)} > 0$. Consider a distribution with $\boldsymbol{\mu} = [0, 1, 0, 3]$, which are identical to the first four moments of the standard normal distribution. Clearly, $\theta_2 = 1/2$ and $\theta_1 = \theta_3 = \theta_4 = 0$. Suppose we introduce a small disturbance $d\boldsymbol{\mu} = [0, 0, 0, \delta]$, where $\delta > 0$. Since $d\boldsymbol{\theta} = -\mathcal{H}^{-1}d\boldsymbol{\mu}$, we have $d\theta_4 = -\mathcal{H}^{(4,4)}\delta < 0$. It then follows that $\theta_4 < 0$, which renders the approximation $f_1(x)$ nonintegrable.

Although $f_1(x)$ is rather flexible, the limitation discussed above precludes the applicability of the ME density to symmetric fat-tailed distributions, which occur frequently in practice, especially in financial data. Hence, we consider an alternative specification which can be motivated by the fat-tailed Student's t distribution. We note that the t distribution with r degrees of freedom has the density

$$T_r(x) = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{r\pi}\Gamma(\frac{r}{2})(1 + \frac{x^2}{r})^{(r+1)/2}} = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{r\pi}\Gamma(\frac{r}{2})} \exp\left\{-\frac{r+1}{2} \log\left(1 + \frac{x^2}{r}\right)\right\},$$

which can be characterized as an exponential distribution with a general moment $\log(1 + \frac{x^2}{r})$. Accordingly, we can modify the normal density by adding an extra moment condition that $E \log(1 + \frac{x^2}{r})$ equals its sample

estimate. The resulting general ME density

$$f_1'(x) = \exp\left(-\sum_{k=0}^2 \theta_k x^k - \theta_3 \log\left(1 + \frac{x^2}{r}\right)\right),$$

where $r > 0$. Since $\log\left(1 + \frac{x^2}{r}\right) = o(x)$, x^2 is the dominant term for all r , which implies that $\theta_2 > 0$ to ensure the integrability of $f_1'(x)$ over the real line. The presence of $\log\left(1 + \frac{x^2}{r}\right)$ allows the ME density to accommodate symmetric fat-tailed distributions.

To make the specification more flexible, we further introduce a term to capture skewness and asymmetry. One possibility is to use $\tan^{-1}(x)$ which is an odd function and bounded between $(-\pi/2, \pi/2)$. Formally, Lye and Martin (1993) derive the generalized t distribution from the generalized Pearson family defined by

$$\frac{df}{dx} = -\frac{(\sum_{k=1}^2 \theta_k x^k)f(x)}{(r^2 + x^2)}.$$

The solution takes the form

$$f_2(x) = \exp\left(-\sum_{k=0}^2 \theta_k x^k - \theta_3 \tan^{-1}\left(\frac{x}{r}\right) - \theta_4 \log(r^2 + x^2)\right), \quad r > 0.$$

Since the *degrees of freedom* r is unknown, we set $r = 1$, which allows the maximum degree of fat-tailedness.⁴ Alternatively, one can view r as the scale parameter and setting $r = 1$ is consistent with our standardization of the data. The alternative ME density is then defined as

$$f_2(x) = \exp\left(-\sum_{k=0}^2 \theta_k x^k - \theta_3 \tan^{-1}(x) - \theta_4 \log(1 + x^2)\right).$$

We further notice an *asymmetry* between $\tan^{-1}(x)$ and $\log(1 + x^2)$ in the sense that the former is bounded while the latter is unbounded. Therefore, we consider yet another alternative, wherein we replace $\log(1 + x^2)$ by $\tan^{-1}(x^2)$.⁵ We note that Park and Bera (Forthcoming) used the moment function $\tan^{-1}(x^2)$ to represent the peakedness of densities. Our third

⁴A t distribution with one degree of freedom is the Cauchy distribution, which has the fattest tails within the family of t distributions. See also Lye and Martin (1994) on the connection between testing for normality and the generalized Student t distribution. Premaratne and Bera (2005) also used the moment function $\tan^{-1}(x)$.

⁵We also tried $[\tan^{-1}(x)]^2$. The performance was essentially the same as that with $\tan^{-1}(x^2)$.

ME density is defined as

$$f_3(x) = \exp\left(-\sum_{k=0}^2 \theta_k x^k - \theta_3 \tan^{-1}(x) - \theta_4 \tan^{-1}(x^2)\right).$$

It is expected that $\tan^{-1}(x)$ and $\tan^{-1}(x^2)$ will mimic the behavior of x^3 and x^4 yet at the same time remain bounded such that $f_3(x)$ is able to accommodate distributions with exceptionally large skewness and kurtosis. Note that $f_3(x)$ is in spirit close to Gallant's (1981) flexible Fourier transformation where low-order polynomials are combined with a trigonometric series to achieve a balance of parsimony and flexibility. In Wu and Stengos (2005), we also consider $\sin(x)$ and $\cos(x)$ for flexible ME densities. Generally, using periodic functions like $\sin(x)$ and $\cos(x)$ requires rescaling the data to be within $[-\pi, \pi]$. Although in principle they are equally suitable for density approximations, we do not consider specifications with $\sin(x)$ and $\cos(x)$ in this study as rescaling the data to be within $[-\pi, \pi]$, rather than standardizing them, requires us to calculate the asymptotic variance under normality for each dataset.

The introduction of general moments offers a considerably higher degree of flexibility as we are not restricted to polynomials. Generally, by choosing general moments appropriately from distributions that are known to accommodate given moment conditions, we make the ME density more robust and at the same time more flexible. As an illustration, Fig. 1 shows ME approximations to a χ^2 distribution with five degrees of

FIGURE 1 Approximation of χ_5^2 distribution: true distribution (solid), f_1 (dash-dotted), f_2 (dotted), f_3 (dashed).

freedom by $f_1(x)$, $f_2(x)$ and $f_3(x)$. Although they have relatively simple functional forms, all three ME densities are shown to capture the general shape of the χ_5^2 density quite well.

3.2. Normality Tests

In this section we derive the LM tests for normality based on the ME densities $f_1(x)$, $f_2(x)$ and $f_3(x)$ presented in the previous section. When $\theta_3 = \theta_4 = 0$, all three densities reduce to the standard normal density.⁶ The information matrix of $f_1(x)$ under standard normality is

$$\mathcal{I}_1 = \begin{bmatrix} 1 & 0 & 1 & 0 & 3 \\ 0 & 1 & 0 & 3 & 0 \\ 1 & 0 & 3 & 0 & 15 \\ 0 & 3 & 0 & 15 & 0 \\ 3 & 0 & 15 & 0 & 105 \end{bmatrix},$$

and the score function under normality is $\widehat{S}_1 = n[0, 0, 0, \hat{\mu}_3, \hat{\mu}_4 - 3]$. It follows that the LM test statistic is

$$t_1 = \frac{1}{n} \widehat{S}_1' \mathcal{I}_1^{-1} \widehat{S}_1 = n \left(\frac{\hat{\mu}_3^2}{6} + \frac{(\hat{\mu}_4 - 3)^2}{24} \right).$$

This the familiar JB test of normality. Bera and Jarque (1981) derived this test as a LM test for the Pearson family of distributions, and White (1982) derived it as an information matrix test. More recently, Bontemps and Meddahi (2005) applied the Stein Equation to the mean of Hermite polynomials to arrive at the same test. Bai and Ng (2005), however, note that the convergence of $\frac{(\hat{\mu}_4 - 3)^2}{24}$ to its asymptotic distribution could be rather slow and the sample kurtosis can deviate substantially from its true value even with a large number of observations.

Instead of using the coefficients of skewness and kurtosis, whose small sample properties are unsatisfactory, we next consider tests based on alternative ME densities $f_2(x)$ and $f_3(x)$. Under normality, the information

⁶Shannon (1949) shows that among all distributions that possess a density function $f(x)$ and have a given variance σ^2 , the entropy $W = -\int f(x) \log f(x) dx$ is maximized by the normal distribution. The entropy of the normal distribution with variance σ^2 is $\log(\sqrt{2\pi e} \sigma)$. Vasicek (1976) uses this property to test a composite hypothesis of normality, based on a nonparametric estimates of sample entropy.

matrix of $f_2(x)$ takes the form

$$\mathcal{J}_2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0.5334532 \\ 0 & 1 & 0 & 0.6556795 & 0 \\ 1 & 0 & 3 & 0 & 1.2220941 \\ 0 & 0.6556795 & 0 & 0.4497009 & 0 \\ 0.5334532 & 0 & 1.2220941 & 0 & 0.5529086 \end{bmatrix},$$

and the score

$$\widehat{\mathbf{S}}_2 = n[0, 0, 0, \hat{\mu}_a, \hat{\mu}_b - 0.5334532],$$

where $\hat{\mu}_a = \frac{1}{n} \sum_{i=1}^n \tan^{-1}(X_i)$ and $\hat{\mu}_b = \frac{1}{n} \sum_{i=1}^n \log(1 + X_i^2)$. The corresponding LM test is given by

$$t_2 = \frac{1}{n} \widehat{\mathbf{S}}_2' \mathcal{J}_2^{-1} \widehat{\mathbf{S}}_2 = n(50.54269 \hat{\mu}_a^2 + 32.027545(\hat{\mu}_b - 0.5334532)^2).$$

Similarly, the information matrix for $f_3(x)$ under normality is

$$\mathcal{J}_3 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0.5299567 \\ 0 & 1 & 0 & 0.6556795 & 0 \\ 1 & 0 & 3 & 0 & 1.0664226 \\ 0 & 0.6556795 & 0 & 0.4497009 & 0 \\ 0.5299567 & 0 & 1.0664226 & 0 & 0.4839857 \end{bmatrix},$$

and the score

$$\widehat{\mathbf{S}}_3 = n[0, 0, 0, \hat{\mu}_a, \hat{\mu}_c - 0.5299567],$$

where $\hat{\mu}_c = \frac{1}{n} \sum_{i=1}^n \tan^{-1}(X_i^2)$. The LM test statistic is then computed as

$$t_3 = -\frac{1}{n} \widehat{\mathbf{S}}_3' \mathcal{J}_3^{-1} \widehat{\mathbf{S}}_3 = n(50.54269 \hat{\mu}_a^2 + 16.882261(\hat{\mu}_c - 0.5299567)^2).$$

The following theorem shows that all three tests are asymptotically distributed according to a χ^2 distribution with two degrees of freedom under normality.

Theorem 1. *Under the assumption that $E|x|^{4+\delta} < \infty$ for $\delta > 0$, the test statistics t_1 , t_2 , and t_3 are distributed asymptotically as χ^2 with two degrees of freedom under normality.*

The proof of Theorem 1 is presented in the Appendix.

4. SIMULATIONS

In this section, we use Monte Carlo simulations to assess the size and power of the proposed tests. Following Bai and Ng (2005), we consider some well known distributions such as the normal, the t and the χ^2 , as well as distributions from the generalized lambda family. The generalized lambda distribution, denoted by F_λ , is defined in terms of the inverse of the cumulative distribution $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2$, $0 < u < 1$. This family nests a wide range of symmetric and asymmetric distributions. In particular, we consider the following symmetric and asymmetric distributions:

- S1: $N(0, 1)$;
- S2: t distribution with 5 degrees of freedom;
- S3: $e_1 I(z \leq 0.5) + e_2 I(z > 0.5)$, where $z \sim U(0, 1)$, $e_1 \sim N(-1, 1)$, and $e_2 \sim N(1, 1)$;
- S4: F_λ , $\lambda_1 = 0$, $\lambda_2 = 0.19754$, $\lambda_3 = 0.134915$, $\lambda_4 = 0.134915$;
- S5: F_λ , $\lambda_1 = 0$, $\lambda_2 = -1$, $\lambda_3 = -0.8$, $\lambda_4 = -0.8$;
- S6: F_λ , $\lambda_1 = 0$, $\lambda_2 = -0.397912$, $\lambda_3 = -0.16$, $\lambda_4 = -0.16$;
- S7: F_λ , $\lambda_1 = 0$, $\lambda_2 = -1$, $\lambda_3 = -0.24$, $\lambda_4 = -0.24$;
- A1: lognormal: $\exp(e)$, $e \sim N(0, 1)$;
- A2: χ^2 distribution with 3 degrees of freedom;
- A3: exponential: $-\ln(e)$, $e \sim U(0, 1)$;
- A4: F_λ , $\lambda_1 = 0$, $\lambda_2 = 1$, $\lambda_3 = 1.4$, $\lambda_4 = 0.25$;
- A5: F_λ , $\lambda_1 = 0$, $\lambda_2 = -1$, $\lambda_3 = -0.0075$, $\lambda_4 = -0.03$;
- A6: F_λ , $\lambda_1 = 0$, $\lambda_2 = -1$, $\lambda_3 = -0.1$, $\lambda_4 = -0.18$;
- A7: F_λ , $\lambda_1 = 0$, $\lambda_2 = -1$, $\lambda_3 = -0.001$, $\lambda_4 = -0.13$;
- A8: F_λ , $\lambda_1 = 0$, $\lambda_2 = -1$, $\lambda_3 = -0.0001$, $\lambda_4 = -0.17$.

The first seven distributions are symmetric and the next eight are asymmetric, with a wide range of skewness and kurtosis coefficients as shown in Table 1. For each distribution, we draw 10,000 random samples of size $n = 20, 50, 100$, respectively, and compute the normality test statistics discussed above. For the sake of comparison, we also compute the commonly used Kolmogorov–Smirnov (KS) test. We note that the general-purpose KS test has very low power. Instead, we use the Lillie test, which is a special version of KS test tailored for the test of normality, see Thode (2002). All tests were computed based on standardized samples and all simulations were implemented in Matlab 6.5.

Table 1 reports the results of the normality tests at the 5% significance level. The first row reflects the size and the rest show the power of the tests. It is noted that all three moment-based tests are under-sized when $n = 20$ or 50. As n increases, the size of all tests converges to the theoretical level, and their powers generally increase.

TABLE 1 Size and power of normality test

		$n = 20$					$n = 50$				$n = 100$			
	μ_3	μ_4	t_1	t_2	t_3	KS	t_1	t_2	t_3	KS	t_1	t_2	t_3	KS
S1	0	3	0.02	0.03	0.04	0.05	0.03	0.04	0.04	0.05	0.05	0.05	0.05	0.05
S2	0	9	0.16	0.18	0.17	0.14	0.39	0.37	0.35	0.21	0.63	0.59	0.56	0.33
S3	0	2.5	0.01	0.02	0.04	0.05	0.00	0.07	0.09	0.07	0.00	0.15	0.17	0.10
S4	0	3	0.02	0.04	0.04	0.06	0.03	0.04	0.05	0.05	0.04	0.04	0.04	0.06
S5	0	6	0.17	0.19	0.19	0.15	0.39	0.38	0.36	0.22	0.62	0.61	0.58	0.35
S6	0	11.6	0.24	0.27	0.27	0.21	0.56	0.57	0.54	0.36	0.81	0.82	0.80	0.60
S7	0	126	0.32	0.36	0.36	0.30	0.69	0.71	0.70	0.53	0.92	0.93	0.92	0.79
A1	6.18	113.9	0.72	0.86	0.86	0.80	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
A2	2	9	0.48	0.67	0.68	0.60	0.96	0.99	0.99	0.96	1.00	1.00	1.00	1.00
A3	2	9	0.48	0.67	0.67	0.59	0.96	0.99	0.99	0.96	1.00	1.00	1.00	1.00
A4	5	2.2	0.02	0.18	0.23	0.21	0.08	0.73	0.77	0.52	0.79	0.99	0.99	0.89
A5	0.5	7.5	0.30	0.38	0.38	0.32	0.74	0.81	0.80	0.64	0.97	0.98	0.98	0.92
A6	2	21.2	0.30	0.35	0.34	0.29	0.66	0.69	0.68	0.53	0.91	0.93	0.92	0.81
A7	3.16	23.8	0.61	0.78	0.78	0.71	0.98	1.00	1.00	0.98	1.00	1.00	1.00	1.00
A8	3.8	40.7	0.64	0.81	0.81	0.75	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00

For all three sample sizes, t_2 and t_3 have comparable or higher power than t_1 for all distributions. For the thin-tailed S3, the power of t_2 and t_3 is considerably higher than t_1 . A similar pattern is observed for most of the asymmetric distributions in question. On the other hand, the power of the KS test is generally lower than the moment-based tests, except for S4. We note that S4 shares the same first four moments with the standard normal distribution. This distribution has also been investigated by Bera and John (1983), who showed that t_1 has “zero” power against S4. Although t_2 , t_3 , and the KS test do not depend on such moments, their power against S4 is essentially identical to that of t_1 . Similar results are reported in Bai and Ng (2005).

So far we have focused on comparing the performance of various tests based on their asymptotic critical values. To gain further insight into their small sample performance, in Table 2 we report the size-adjusted power of the tests. The critical values are calculated based on 100,000 repetitions of the tests in question under normality. The results are qualitatively close to those reported in Table 1. In particular, t_2 and t_3 continue to have higher power against S3, A2, A3, A4, A7, and A8 for $n = 20$.

Overall, our results suggest that the proposed tests compare favorably to some commonly used conventional tests, especially when the sample size is small.

TABLE 2 Size-adjusted power of normality test

		n = 20					n = 50				n = 100			
	μ_3	μ_4	t_1	t_2	t_3	KS	t_1	t_2	t_3	KS	t_1	t_2	t_3	KS
S1	0	3	—	—	—	—	—	—	—	—	—	—	—	—
S2	0	9	0.23	0.20	0.19	0.13	0.43	0.38	0.36	0.21	0.65	0.60	0.56	0.33
S3	0	2.5	0.02	0.04	0.05	0.05	0.01	0.09	0.10	0.07	0.01	0.16	0.18	0.10
S4	0	3	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.05
S5	0	6	0.23	0.21	0.20	0.14	0.42	0.40	0.37	0.22	0.64	0.62	0.58	0.34
S6	0	11.6	0.32	0.30	0.29	0.20	0.59	0.58	0.55	0.37	0.82	0.82	0.81	0.59
S7	0	126	0.41	0.39	0.38	0.28	0.72	0.73	0.71	0.53	0.93	0.93	0.93	0.79
A1	6.18	113.9	0.82	0.89	0.88	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
A2	2	9	0.62	0.73	0.72	0.58	0.98	0.99	0.99	0.96	1.00	1.00	1.00	1.00
A3	2	9	0.62	0.73	0.72	0.58	0.98	0.99	0.99	0.96	1.00	1.00	1.00	1.00
A4	5	2.2	0.06	0.26	0.29	0.20	0.17	0.76	0.79	0.52	0.88	0.99	0.99	0.89
A5	0.5	7.5	0.40	0.43	0.41	0.31	0.79	0.82	0.81	0.65	0.97	0.98	0.98	0.91
A6	2	21.2	0.39	0.38	0.37	0.28	0.70	0.70	0.69	0.54	0.92	0.93	0.93	0.81
A7	3.16	23.8	0.73	0.82	0.81	0.70	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00
A8	3.8	40.7	0.77	0.85	0.84	0.73	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00

5. EXTENSIONS

In addition to their simplicity, a major advantage of the proposed tests is their generality. In this section, we briefly discuss some easy-to-implement extensions of these tests.

Firstly, we note that we can incorporate higher order polynomials x^k for $k > 4$ and higher order trigonometric terms such as $\tan^{-1}(x^k)$ for $k > 2$. Usually, the addition of higher order terms will improve the approximation to the underlying distribution. However, we note that it does not necessarily improve the test. We experimented with adding x^5 and x^6 to $f_2(x)$ and $\tan^{-1}(x^3)$ and $\tan^{-1}(x^4)$ to $f_3(x)$ and derived tests based on four instead of two moment conditions.⁷ These alternative tests are distributed asymptotically according to a χ^2_4 distribution under normality. However, we note that their power is generally lower than that of tests based on two moment conditions. This is to be expected as the test statistics are distributed according to a noncentral χ^2 distribution under alternative non-normal distributions. For a given noncentrality parameter there is an inverse relationship between degrees of freedom and power, see Das Gupta and Perlman (1974).

Secondly, we can use the proposed method for other distributions than the normal. For example, the gamma distribution can be characterized as a ME distribution

$$f(x) = \exp(-\theta_0 - \theta_1 x - \theta_2 \log x), \quad x > 0.$$

⁷The first two moments are zero and one by standardization.

Because Ex and $E \log x$ are the characterizing moments for the gamma distribution, the presence of any additional terms in the exponent of $f(x)$ would reject the hypothesis that x is distributed according to a gamma distribution. Let $f_K(x) = \exp(-\theta_0 - \theta_1 x - \theta_2 \log x - \sum_{k=3}^K \theta_k g_k(x))$, the test of $\theta_k = 0$ for $k \geq 3$ is then the LM test for gamma distribution. The discussion in the previous section suggests that the natural candidates for $g_k(x)$ may include polynomials of x and $\log x$, and trigonometric terms of x and $\log x$.

Thirdly, we can generalize our tests to regression residuals within the framework of White and McDonald (1980). Consider a classical linear model

$$Y_i = Z_i \beta + \varepsilon_i, \quad i = 1, \dots, n. \quad (3)$$

Since the error term ε_i is not observed, one has to replace it with the residual $\hat{\varepsilon}_i$. The following theorem ensures that the test statistics computed from the residuals $\hat{\varepsilon}_i$ share the same asymptotic distribution as those from the true errors ε_i .

Theorem 2. *Assume the following assumptions hold:*

1. $\{Z_i\}$ is a sequence of uniformly bounded fixed $1 \times K$ vectors such that $Z'Z/n \rightarrow M_z$, a positive definite matrix, $\{\varepsilon_i\}$ is a sequence of iid random variables with $E\varepsilon_i = 0$, $E\varepsilon_i^2 = \sigma_i^2 < \infty$, and β is an unknown $K \times 1$ vector.
2. $E|\varepsilon_i|^{4+\delta} < \infty$ for $\delta > 0$.
3. The density of $\varepsilon_i, f(\varepsilon)$, is uniformly continuous, positive on the interval of support and bounded.

Let $\hat{\varepsilon}_i$ be the standardized residuals. Define $\hat{\mu}_3 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^3$, $\hat{\mu}_4 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^4$, $\hat{\mu}_a = \frac{1}{n} \sum_{i=1}^n \tan^{-1}(\hat{\varepsilon}_i)$, $\hat{\mu}_b = \frac{1}{n} \sum_{i=1}^n \log(1 + \hat{\varepsilon}_i^2)$, and $\hat{\mu}_c = \frac{1}{n} \sum_{i=1}^n \tan^{-1}(\hat{\varepsilon}_i^2)$. Then under normality, the test statistics

$$\begin{aligned} \hat{t}_1 &= n(\hat{\mu}_3^2/6 + (\hat{\mu}_4 - 3)^2/24) \sim \chi_2^2, \\ \hat{t}_2 &= n(50.54269\hat{\mu}_a^2 + 32.027545(\hat{\mu}_b - 0.5334532)^2) \sim \chi_2^2, \\ \hat{t}_3 &= n(50.54269\hat{\mu}_a^2 + 16.882261(\hat{\mu}_c - 0.5299567)^2) \sim \chi_2^2. \end{aligned}$$

The proof of Theorem 2 is presented in the Appendix.

Furthermore, for time series or heteroskedastic data, we can use the approach of Bai and Ng (2005) or Bontemps and Meddahi (2005). In general, for non-i.i.d. data, to test that the LMs associated with sample moments of $g_k(x)$ in the ME density are zero, we need to calculate a Heteroskedastic–Autocorrelation–Consistent (HAC) covariance matrix for

those moments, see Richardson and Smith (1993) on the use of HAC standard errors in testing for normality.⁸

Finally, as an illustration, we apply the proposed normality tests to regression residuals. We use data on the production cost of some U.S. electricity generating companies from Christensen and Greene (1976). We estimate a flexible cost function with 123 observations:

$$c = \beta_0 + \beta_1 q + \beta_2 q^2 + \beta_3 p_f + \beta_4 p_l + \beta_5 p_k + \beta_6 q p_f + \beta_7 q p_k + \beta_8 q p_l + \varepsilon,$$

where c is the total cost, q is the total output, p_f , p_l , and p_k are the price of fuel, labor, and capital, respectively, and ε is the error term. All variables are expressed in logarithmic form. It is expected that the distribution of the Ordinary Least Squares (OLS) residuals from a production function regression is skewed to the right due to the presence of firm specific, non-negative efficiency components in the error term. Nonetheless, the KS test fails to reject the normality hypothesis. On the other hand, all three LM tests reject the normality hypothesis with p -values of 0.03, 0.01, and 0.02, respectively.

6. CONCLUSION

In this article, we derive some general distributional tests from ME density methodology. The proposed tests are derived from maximizing the differential entropy subject to given moment constraints. By exploiting the equivalence between the ME and the ML estimates for the exponential family, we can use the conventional LR, Wald, and LM testing principles in the maximum entropy framework. Hence, our tests share the optimality properties of the standard ML based tests. In particular, we show that the ME approach leads to simple yet powerful LM tests for normality. We derive the asymptotic properties of the proposed tests and show that they are asymptotically equivalent to the popular Jarque–Bera test. Our Monte Carlo simulations show that the proposed tests have desirable small sample properties. They are comparable and often outperform some conventional tests for normality. In addition, we show that the proposed method can be generalized to tests for other distributions than the normal. Also, extensions to regression residuals and non-i.i.d. data are straightforward. In principle, the proposed methodology can be also applied to distributional tests for truncated distributions, as in Bera et al. (1984) and Lye and Martin (1998), something that we leave for future research.

⁸We thank a referee for this reference.

APPENDIX

Proof of Theorem 1

Proof. The assumption that $E|x|^{4+\delta} < \infty$ for $\delta > 0$ ensures the existence of $E\hat{\mu}_3$ and $E\hat{\mu}_4$. One can easily show that $\sqrt{n}\hat{\mu}_3 \sim N(0, 6)$ and $\sqrt{n}(\hat{\mu}_4 - 3) \sim N(0, 24)$ if x_i is iid and normally distributed (see for example, Stuart et al., 1994). Since $\text{cov}(\hat{\mu}_3, \hat{\mu}_4) = 0$, it follows that under normality

$$t_1 = n \left(\frac{\hat{\mu}_3^2}{6} + \frac{(\hat{\mu}_4 - 3)^2}{24} \right) \sim \chi_2^2.$$

Similarly, since $\tan^{-1}(x) = o(x)$, $\tan^{-1}(x^2) = o(x)$ and $\log(1 + x^2) = o(x)$ as $|x| \rightarrow \infty$, their expectations also exist under the assumption that $E|x|^{4+\delta} < \infty$ for $\delta > 0$. We then have $\sqrt{n}\hat{\mu}_a \sim N(0, 1/50.54269)$, $\sqrt{n}(\hat{\mu}_b - 0.5334532) \sim N(0, 1/32.027545)$, and $\sqrt{n}(\hat{\mu}_c - 0.5299567) \sim N(0, 1/16.8882261)$ under normality. In addition, since $\text{cov}(\hat{\mu}_a, \hat{\mu}_b) = 0$ and $\text{cov}(\hat{\mu}_a, \hat{\mu}_c) = 0$, it follows that under normality

$$t_2 = n(50.54269\hat{\mu}_a^2 + 32.027545(\hat{\mu}_b - 0.5334532)^2) \sim \chi_2^2,$$

$$t_3 = n(50.54269\hat{\mu}_a^2 + 16.8882261(\hat{\mu}_c - 0.5299567)^2) \sim \chi_2^2.$$

Proof of Theorem 2

Proof. Assumption 1 sets forth the classical linear model (except for the normality of ε_i) and ensures that $\hat{\beta}_n \xrightarrow{as} \beta_0$. Given Assumptions 1 and 2, one can show that $|\hat{\mu}_3 - \mu_3| \xrightarrow{as} 0$ and $|\hat{\mu}_4 - \mu_4| \xrightarrow{as} 0$ using Lemmas 1 and 2 of White and McDonald (1980). Using Corollary A of Serfling (1980, p. 19), one can show that since $\hat{t}_1 \xrightarrow{as} t_1$, $\hat{t}_1 \xrightarrow{d} t_1$ given Assumption 3. Since $t_1 \sim \chi_2^2$ by Theorem 1 in Section 3, we have $\hat{t}_1 \sim \chi_2^2$. Similarly, since $\tan^{-1}(x) = o(x)$, $\tan^{-1}(x^2) = o(x)$ and $\log(1 + x^2) = o(x)$ as $|x| \rightarrow \infty$, Assumptions 1 and 2 ensure that $|\hat{\mu}_a| \xrightarrow{as} 0$, $|\hat{\mu}_b - \mu_b| \xrightarrow{as} 0$, and $|\hat{\mu}_c - \mu_c| \xrightarrow{as} 0$. Using the similar arguments as the proof for $\hat{t}_1$, one can show that $\hat{t}_2 \xrightarrow{d} \chi_2^2$ and $\hat{t}_3 \xrightarrow{d} \chi_2^2$.

ACKNOWLEDGMENTS

We want to thank the associate editor, two anonymous referees, seminar participants at Penn State University, the 2004 European Meeting of the Econometric Society, and the 2004 Canadian Econometrics Study Group for comments. Financial support from SSHRC of Canada is gratefully acknowledged.

REFERENCES

- Bai, J., Ng, S. (2005). Tests for skewness, kurtosis and normality for time series data. *Journal of Business and Economic Statistics* 23(1):49–60.
- Bera, A., Biliias, Y. (2002). The MM, ME, ML, EL, EF, and GMM approaches to estimation: A synthesis. *Journal of Econometrics* 107:51.
- Bera, A., Jarque, C. (1981). Efficient tests for normality, heteroskedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters* 7:313–318.
- Bera, A., Jarque, C., Lee, L. F. (1984). Testing the normality assumption in limited dependent variable models. *International Economic Review* 25:563–578.
- Bera, A., John, S. (1983). Tests for multivariate normality. *Communication in Statistics, Theory and Methods* 12(1):103–117.
- Bontemps, C., Meddahi, N. (2005). Testing normality: A gmm approach. *Journal of Econometrics* 124(1):149–186.
- Christensen, L. R., Greene, W. H. (1976). Economies of scale in U.S. electric power generation. *Journal of Political Economy* 84:655–676.
- Cobb, L., Koppstein, P., Chen, N. (1983). Estimation and moment recursion relations for multimodal distributions of the exponential family. *Journal of American Statistical Association* 8(381):124–130.
- Cressie, N., Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B* 46:440–464.
- Dalén, J. (1987). Bounds on standardized sample moments. *Statistics and Probability Letters* 5:329–31.
- Das Gupta, S., Perlman, M. D. (1974). Power of the noncentral F test: effect of additional variates in Hotelling's t^2 test. *Journal of the American Statistical Association* 69:174–180.
- Ebrahimi, N., Maasoumi, E., Soofi, E. (1999). Ordering univariate distributions by entropy and variance. *Journal of Econometrics* 90:317–336.
- Econometric Reviews (2008). Special issues on IEE. 27:317–609.
- Engle, R. (1984). Wald, likelihood ratio and lagrange multiplier tests in econometrics. In: Grilliches, Z., Intriligator, M. D., eds. *Handbook of econometrics*, Vol. 3. North Holland: Elsevier.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* 222:309–68.
- Gallant, A. R. (1981). On the bias in flexible functional forms and an essentially unbiased form. *Journal of Econometrics* 15:211–241.
- Golan, A. (2002). Information and entropy econometrics – editor's view. *Journal of Econometrics* 107:1–16.
- Golan, A. (2007). Information and entropy econometrics – volume overview and synthesis. *Journal of Econometrics* 138:379–387.
- Golan, A. (2008). Information and entropy econometrics – a review and synthesis. *Foundations and Trends in Economics* 2:1–145.
- Golan, A., Maasoumi, E. (2008). Information theoretic and entropy methods: An overview. *Econometric Reviews* 27:317–328.
- Golan, A., Judge, G., Miller, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Chichester: John Wiley & Sons.
- Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics* 12:971–988.
- Hansen, L. P. (1982). Large sample properties of generalized methods of moments estimators. *Econometrica* 50:1029–1054.
- Hansen, L. P., Heaton, J., Yaron, A. (1996). Finite sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics* 14:262–280.
- Imbens, G. W., Spady, R. H., Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* 66:333–357.
- Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics* 20:493–507.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Review* 106:620–630.
- Kitamura, Y., Stutzer, M. (1997). An information theoretic alternative to generalized method of moments estimation. *Econometrica* 65:861–874.

- Kitamura, Y. (2006). Empirical likelihood methods in econometrics: theory and practice. Cowles Foundation Discussion Paper No. 1569.
- Lye, J. N., Martin, V. L. (1993). Robust estimation, nonnormalities, and generalized exponential distributions. *Journal of American Statistical Association* 88(421):261–267.
- Lye, J. N., Martin, V. L. (1994). Non-linear time series modelling and distributional flexibility. *Journal of Time Series Analysis* 15:65–84.
- Lye, J. N., Martin, V. L. (1998). Truncated distribution families. In: Creedy, J., Martin, V. L. eds. *Nonlinear Economic Models: Cross-sectional. Time Series and Neural Network Application*, Cheltenham, UK: Edward Elgar, pp. 47–68.
- Maasoumi, E. (1993). A compendium on information theory in economics and econometrics. *Econometrics Reviews* 12:137–181.
- Neyman, J. (1937). Smooth test for goodness of fit. *Scandinavian Aktuarial* 20:149–199.
- Ornermite, D., White, H. (1999). An efficient algorithm to compute maximum entropy densities. *Econometric Reviews* 18(2):141–167.
- Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75:237–249.
- Park, J., Bera, A. forthcoming, Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics* 150:219–230.
- Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of correlated systems of variables is such that it can be reasonably supposed to have arisen in random sampling. *Philosophical Magazine* 50(5):157–175.
- Premaratne, G., Bera, A. (2005). A test for symmetry with leptokurtic financial data. *Journal of Financial Econometrics* 3(2):169–187.
- Qin, J., Lawless, J. (1994). Empirical likelihood and general estimating functions. *Annals of Statistics* 22:300–325.
- Richardson, M., Smith, T. (1993). A test for multivariate normality in stock returns. *Journal of Business* 66:295–321.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons, Inc.
- Shannon, C. E. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.
- Shore, J. E., Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum crossentropy. *IEEE Transactions on Information Theory* 26(1):26–37.
- Smith, R. J. (2004). *GEL Criteria for Moment Condition Models*. Manuscript, University of Warwick.
- Smith, R. J. (2005). Local GEL Methods for Conditional Moment Restrictions. Working Paper, University of Cambridge, Cambridge, England.
- Stohs, S. (2003). *A Bayesian Updating Approach to Crop Insurance Ratemaking*. Ph.D. thesis, University of California at Berkeley.
- Stuart, A., Ord, K., Arnold, S. (1994). *Kendall's Advanced Theory of Statistics*. Vol. 2A, New York: Oxford University Press.
- Thode, H. (2002). *Testing for Normality*. New York: Marcel Dekker.
- Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B* 38:54–59.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50:1–26.
- White, H., McDonald, G. M. (1980). Some large-sample tests for nonnormality in the linear regression model. *Journal of American Statistical Association* 75:16–28.
- Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics* 115:347–354.
- Wu, X., Stengos, T. (2005). Partially adaptive estimation via maximum entropy densities. *Econometrics Journal* 8:352–366.
- Zellner, A., Highfield, R. A. (1988). Calculation of maximum entropy distribution and approximation of marginal posterior distributions. *Journal of Econometrics* 37:195–209.
- Zellner, A. (1996). Bayesian method of moments/instrumental variable (BMOM/IV) analysis of mean and regression models. In: Lee, J. C., Johnson, W. C., Zellner, A. eds. *Modeling and Prediction: Honoring Seymour Geisser*. Springer-Verlag, pp. 61–75.