THE UNIVERSITY of York

Discussion Papers in Economics

No. 2005/24

Goodness of Fit Tests via Exponential Series Density Estimation

by

Patrick Marsh

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Goodness of Fit tests via Exponential Series Density Estimation[1]

Patrick Marsh

Department of Economics

University of York

Heslington, York

YO10 5DD

tel. +44 1904 433084

fax +44 1904 433759

e-mail: pwnm1@york.ac.uk

July 25, 2005

[1]Thanks are due to Francesco Bravo, Giovanni Forchini, Grant Hillier, Les Godfrey, Peter Phillips and participants at the York Econometrics Workshops, July 2005.

Abstract

This paper explores the properties of a new nonparametric goodness of fit test, based on the likelihood ratio test of Portnoy (1988). It is applied via the consistent series density estimator of Crain (1974) and Barron and Sheu (1991). The asymptotic properties are established as trivial corollaries to the results of those papers as well as from similar results in Marsh (2000) and Claeskens and Hjort (2004).

The paper focuses on the computational and numerical properties. Specifically it is found that the choice of approximating basis is not crucial and that the choice of model dimension, through consistent selection criteria, yields a feasible procedure. Extensive numerical experiments show that the usage of asymptotic critical values is feasible in moderate sample seizes. More importantly the new tests are shown to have significantly more power than established tests such as the Kolmogorov-Smirnov, Cramér-von Mises or Anderson-Darling. Indeed, for certain interesting alternatives the power of the proposed tests may be several times that of the established ones.

1 Introduction

Testing whether a sample has a particular distribution, in other words the goodness of fit problem, has importance across many areas of applied statistics. Unsurprisingly therefore there are a very large number of suggested procedures. The problem can be formalised in that we have a sample $\{x_i\}_{i=1}^n$ and we wish to test the hypothesis that the x_i are identically distributed copies of a random variable X with known distribution function $P(x)$, i.e.

$$H_0 : x_i \sim iid\ X \quad ; \quad \Pr[X \leq x] = P(X). \tag{1}$$

By far the most common formal statistical procedures for testing (1) are those based upon the empirical distribution function, such as the Kolmogorov-Smirnov and the Cramér-von Mises tests, see for example Darling (1957). Both these tests are based upon distances of the empirical distribution function to the hypothesised distribution function. Refinements involve use of a weighting function, leading to a weighted measure of distance. Indeed for 50 years or so perhaps the preferred statistical procedure has been the weighted Cramér-von Mises, or the Anderson-Darling statistic, of Anderson and Darling (1952). A fuller historical perspective and details of the many other procedures can be found, for example, in Conover (1999). Stephens (1974) provides a Monte Carlo comparison of the powers of those tests based upon the empirical distribution function.

The purpose of this paper is to introduce a nonparametric goodness of fit test based upon the likelihood ratio test of Portnoy (1988). It is made nonparametric by utilising the exponential series density estimator of Crain (1974, 1976 and 1977) and also Barron and Sheu (1991). In common with two very related statistics, due to Marsh (2000) and Claeskens and Hjort (2004), the principle is to test via the ratio of the estimated density to that of the imposed null hypothesis. The difference lies in how the null is to be imposed. Claeskens and Hjort (2004) assume that the null density is uniform on $(0, 1)$ and so their ratio is just the estimated density. Marsh

(2000), on the other hand, utilises moment restrictions that the sample must satisfy under the null, but not the alternative. Here, we use, as the null density, that member of the, potentially infinite, exponential family which approximates the hypothesised density.

Since the properties of the exponential series estimator and of the infinite dimensional likelihood ratio test are well known, this paper concentrates upon the computational and numerical properties of the suggested procedure. The only theoretical results given here are a lemma dealing with the properties of the estimator and a theorem detailing the asymptotic distribution of the statistics under the null and fixed and local alternatives. Both can be trivially proved from existing results due to Portnoy (1988), Barron and Sheu (1991) and Claeskens and Hjort (2004).

First, this paper finds that in practice the dimension of the series density estimator need not be large. Consequently, this density estimator becomes a feasible basis upon which to build a test. Specifically, therefore, the choice of dimension may be data driven, in that we may apply a selection criterion over a relatively small subset of possible dimensions. For illustration the information criteria of Akaike (1974) and Schwarz (1978) are applied. In particular by using the precise form of the likelihood ratio here, as opposed to that of Claeskens and Hjort (2004), consistency of these criteria both under the null and alternative is assured. In addition it is found that the choice of basis, for example whether polynomial or trigonometric, for the approximating exponential is not crucial.

In terms of the properties of the proposed test statistics numerical comparisons are made with the established procedures. Both the Kolmogorov-Smirnov and the Cramér-von Mises tests and their weighted versions are used in the comparisons as are both fixed dimension and selected dimension versions of the proposed statistics. An extensive simulation study is carried out under the null to analyse the finite sample performance of asymptotic critical values. The performance of all statistics is broadly comparable, and thus not any basis upon which to choose. On the contrary though,

it is demonstrated that the power can be significantly higher for all versions of the proposed tests than for the established. In several cases of interest the power of the likelihood ratio tests may be two or three times that of any of the established tests.

As well as being more powerful the tests based upon the series density estimator enjoy another significant advantage. Supposing that the hypothesis is rejected then the applied researcher will still have available a consistent approximation. Indeed since this approximation is analytic rather than numerical, such as with a kernel based estimator, it may itself be readily be used for prediction or various probability calculations.

The plan for the rest of the paper is as follows. The next section summarises the pertinent theoretical properties of the density estimator and details the practical computational and numerical issues of choice of dimension and approximating basis. Similarly section 3 gives the asymptotic properties of the proposed tests and provides a detailed analysis of its numerical properties in a comparative size and power study. Section 4 concludes while all of the numerical results themselves are presented in an appendix.

2 Exponential Series Density Estimation

2.1 Theoretical Results

The procedures and tests of this paper are based upon the series density estimator introduced by Crain (1974) and further analysed by Crain (1976 & 1977) and Barron and Sheu (1991). Specifically we wish to estimate the density of a random variable x having distribution $P(x)$. Throughout we shall assume that the data $\{x_i\}_{i=1}^n$ are i.i.d. copies of the random variable x, which satisfies the following:

Assumption 1 *(i) Let x be defined on the bounded sample space (a,b), $a<b$ and both finite and with density*

$$p(x)=dP(x):\left\{(a,b)\rightarrow\mathbb{R},\int_a^b dP(x)=1,p(x)\geq 0\right\}.$$

(ii) The log-density of x satisfies

$$lp(x)=\ln\left[p(x)\right]\in W_2^r,$$

where W_2^r is the Sobolev space of functions, so that $lp^{(r-1)}(x)=\frac{d^{r-1}lp(x)}{dx^{r-1}}$ is absolutely continuous and $lp^{(r)}(x)$ is square integrable on (a,b) for all $r\geq 2$.

The density estimator of Crain (1974) is the limiting member of the exponential family, vis.

$$\lim_{m\rightarrow\infty}p_\theta(x)=p_0(x)\exp\left\{\sum\nolimits_{k=1}^m\theta_k\phi_k(x)-\varphi_m(\theta)\right\},\tag{2}$$

where in (2) the cumulant function is defined by

$$\varphi_m(\theta)=\log\int_a^b p_0(x)\exp\left\{\sum\nolimits_{k=1}^m\theta_k\phi_k(x)\right\}dx.\tag{3}$$

In (3) $\theta=(\theta_1,..,\theta_m)'\in\mathbb{R}^m$, $p_0(x)$ is a reference probability density function on (a,b) and the $\phi_k(x)$ are a set of linearly independent functions, forming a basis for a linear space S_m on (a,b). Choice of S_m, for example whether polynomials, trigonometric (and/or exponential) series and splines, will be the examined in numerical analysis to follow.

The density estimator itself is defined as follows. Given the i.i.d. sample $\{x_i\}_{i=1}^n$ the exponential series density estimator $p_{\hat{\theta}}(x)$ is the maximum likelihood estimator (mle) in the family (2). Formally,

$$\hat{\theta}=\arg\max_{\theta\in\mathbb{R}^m}\quad l(\theta)=\ln p_0(x)+\sum_{i=1}^n\sum_{k=1}^m\theta_k\phi_k(x_i)-n\varphi_m(\theta).\tag{4}$$

From (4) some key properties are immediately obtainable; first the score is given by

$$S(\theta)=\frac{dl(\theta)}{d\theta}=\sum_{i=1}^n\phi_k(x_i)-n\frac{d\varphi_m(\theta)}{d\theta},$$

while from (3) we have

$$\begin{aligned}\frac{d\varphi_m(\theta)}{d\theta} &= \frac{d}{d\theta}\left(\log\int_a^b p_0(x)\exp\left\{\sum_{k=1}^m \theta_k\phi_k(x)\right\}dx\right)\\ &= \frac{\frac{d}{d\theta}\left(\int_a^b p_0(x)\exp\left\{\sum_{k=1}^m \theta_k\phi_k(x)\right\}dx\right)}{\int_a^b p_0(x)\exp\left\{\sum_{k=1}^m \theta_k\phi_k(x)\right\}dx} = \int_a^b \phi(x)p_\theta(x)dx,\end{aligned}$$

where $\phi(x) = (\phi_1(x), .., \phi_m(x))'$. Thus the mle is the solution to the set of m estimating equations;

$$\int_a^b \phi(x)p_{\hat\theta}(x)dx = \bar\phi, \tag{5}$$

where $\bar\phi = n^{-1}\sum_{i=1}^n \phi(x_i)$. We can also define the Hessian,

$$H(\theta) = \frac{d^2 l(\theta)}{d\theta d\theta'} = n\frac{d^2\varphi_m(\theta)}{d\theta d\theta'} = \frac{d}{d\theta'}\int_a^b \phi(x)p_\theta(x)dx,$$

in the usual way for exponential models.

Indeed if m were fixed it is trivial to use these relations to derive asymptotic distributions for the standardised score and mle. At present though, the mle $\hat\theta$ has no obvious meaning in terms of the density being estimated, $p(x)$. However, since x and hence $\phi(x)$ are bounded then each element of $\bar\phi$ will obey a law of large numbers as $n \to \infty$, specifically

$$n^{-1}\sum_{i=1}^n \phi_k(x_i) \to_p \phi_k^0 < \infty \quad \forall\, k = 1, .., m. \tag{6}$$

From (6) we can therefore define a θ_0 which satisfies a set of equations, analogous to those in (5), as

$$\int_a^b \phi(x)p_{\theta_0}(x)dx = \phi_0, \tag{7}$$

where $\phi_0 = (\phi_1^0, .., \phi_m^0)'$. As a consequence we must consider the relationship between three points in the space of density functions, as defined by Assumption 1. We have the 'true' density $p(x)$, the approximating density $p_{\theta_0}(x)$ and the estimated density $p_{\hat\theta}(x)$. The first two densities are related via

$$\int_a^b \phi(x)p(x)dx = E_{p(x)}[\phi(x)] = \phi_0 = \int_a^b \phi(x)p_{\theta_0}(x)dx,$$

that is in terms of the basis $\phi(x)$, $p(x)$ and $p_{\theta_0}(x)$ have the same moments. On the other hand $p_{\theta_0}(x)$ and $p_{\hat{\theta}}(x)$ are related asymptotically via

$$\int_a^b \phi(x)p_{\theta_0}(x)dx = \phi_0 = \lim_{n\to\infty} \bar{\phi} = \lim_{n\to\infty} \int_a^b \phi(x)p_{\hat{\theta}}(x)dx,$$

that is heuristically (these results will be formalised in a lemma to follow) $p_{\theta_0}(x)$ is the limit of $p_{\hat{\theta}}(x)$.

We can analyse convergence of the density estimator in the following terms; consider a hyperplane of densities $\mathbb{C}^m$ defined by

$$\mathbb{C}^m = \left\{ q(x) : \int_a^b \phi(x)q(x)dx = \phi_0 \right\},$$

so that we have $p_{\theta_0}(x) \in \mathbb{C}^m$ while $p(x) \in \mathbb{C}^\infty$. Hence convergence on the triangle of densities follows from $\hat{\theta} \to_p \theta_0$ while $\mathbb{C}^m \to \mathbb{C}^\infty$ as respectively n and m tend to infinity.

This paper will consider goodness of fit tests which are based upon Portnoy's (1988) likelihood ratio test. Comparative tests, such as the Kolmogorov-Smirnov or Cramér-von Mises are based upon norms on the space of distributions (respectively the sup and L_2 norms) and convergence of the empirical distribution in those norms. Instead here we will exploit convergence of the exponential density with respect to relative entropy (or Kullback-Leibler distance), defined for densities satisfying Assumption 1 by

$$D(p_1|p_2) = \int_a^b \ln\left(\frac{p_1}{p_2}\right) p_1 dx.$$

Strictly speaking $D(p_1(x)|p_2(x))$ is not a norm, although we can trivially, if needed, construct $\Lambda(p_1, p_2) = (D(p_1|p_2) + D(p_2|p_1))$. Since we're interested in the convergence of the estimator $p_{\hat{\theta}}(x)$ to $p(x)$, then as in Barron and Sheu (1991) the following decomposition is central;

$$D(p_{\hat{\theta}}(x)|p(x)) = D(p_{\hat{\theta}}(x)|p_{\theta_0}(x)) + D(p_{\theta_0}(x)|p(x)). \tag{8}$$

In terms of the heuristic arguments above the vanishing of the first term in (8) reflects convergence of $\hat{\theta}$ to θ_0 while that of the second reflects $\mathbb{C}^m \to \mathbb{C}^\infty$. Specifically, these

results may be formulated in the following Lemma, which contains the pertinent results of Crain (1974) and Barron and Sheu (1991).

Lemma 1 *Let θ_0 be a solution of (7) then*
(i) $p_{\theta_0}(x)$ is the unique member of (2) in $\mathbb{C}^m$ and moreover,
(ii) as $m \to \infty$ the relative entropy (Kullback-Leibler divergence) of $p_{\theta_0}(x)$ to $p(x)$ is

$$D(p_{\theta_0}(x)|p(x)) = O_r\left(m^{-2r}\right),$$

where r is the 'smoothness' of the log-density $lp(x)$ as defined in Assumption 1.
(iii) Suppose that $m^3/n \to 0$ as $m, n \to \infty$, then the maximum likelihood estimator in the family (2), $p_{\hat{\theta}}(x)$, given by (5) converges, in relative entropy, to $p(x)$ according to

$$D(p_{\hat{\theta}}(x)|p(x)) = O_{pr}\left(m^{-2r} + \frac{m}{n}\right). \quad \blacksquare \tag{9}$$

Part (i) states the existence and uniqueness of θ_0 given the moment sequence ϕ_0 and therefore also implies the existence and uniqueness of the mle, $\hat{\theta}$. Part (ii) reflects the success with which we are able to approximate $p(x)$ with an (infinite) exponential, while part (iii) concerns our ability to estimate that exponential. Optimising the rate of convergence implies a rate of increase of $m = O(n^{\frac{1}{2r+1}})$ with a maximin rate, when $r = 2$, of $O(n^{1/5})$. On the other hand if it is known that $p(x)$ is analytic then m can grow arbitrarily slowly.

2.2 Computational Results

Although the primary aim of this paper is to propose and analyse a goodness of fit test based upon convergence of relative entropy, specifically the entropy $D(p_{\hat{\theta}}(x)|p_{\theta_0}(x))$, a secondary aim to assess the efficacy of the series density estimator itself. Supposing that the goodness of fit hypothesis (1) is rejected, then at least the estimator itself may be useful in its own right, whether for prediction or simple (approximate) probability calculations.

In order to implement the estimator, notice that the mle is given by (4), which we may rewrite as

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta\in\mathbb{R}^m} l(\theta) - \ln[p_0(x)] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{m}\theta_k(\phi_k(x_i)-\bar{\phi}_k) - n\log\int_a^b \exp\left\{\sum_{k=1}^{m}\theta_k(\phi_k(x_i)-\bar{\phi}_k)\right\}dx,\end{aligned} \tag{10}$$

and since at the mle the contribution of the first term of (10) is zero then

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^m} L_U(\theta) = \int_a^b \exp\left\{\sum_{k=1}^{m}\theta_k(\phi_k(x)-\bar{\phi}_k)\right\}dx. \tag{11}$$

Analogously, the approximate model $p_{\theta_0}(x)$ can be found, for any m by

$$\theta_0 = \arg\min_{\theta\in\mathbb{R}^m} L_R(\theta) = \int_a^b \exp\left\{\sum_{k=1}^{m}\theta_k(\phi_k(x)-\phi_k^0)\right\}dx.$$

Consequently and given a moment sequence - whether population or sample - the minimum argument of these functions can readily be found. In this paper all calculations were performed using Mathematica v.4 and its internal optimisation routine.

There are two issues of practical concern. The first relates to $p(x)$ and our ability to approximate it with $p_{\theta_0}(x)$. Closely related to that is the second, the choice of m, the dimension of θ, in any subsequent procedure based upon the density estimator. We can measure the efficacy of the approximate model, for any given $p(x)$, via evaluation of

$$D(p_{\theta_0}(x)|p(x)) = \int_a^b \left(\ln\left[\frac{p(x)}{p_0(x)}\right] - \sum_{k=1}^{m}\theta_k^0\phi_k(x) - \varphi_m(\theta)\right)p(x)dx. \tag{12}$$

Specifically and without loss of generality we will choose $(a,b) = (0,1)$ and $p_0(x) = 1$. Then for two choices of $p(x)$,

$$p^1(x) = 3x^2 \quad ; \quad p^2(x) = \frac{3\sqrt{x}}{2} \quad , \tag{13}$$

we chose two different bases

$$\phi_k(x) = Cos[\pi i x] \quad ; \quad \phi_k(x) = x^k, \tag{14}$$

and evaluated (12) for each combination for values of $m = 1, 3, 5.., 15$. The results are recorded in Table 1 in the appendix. From even this limited analysis two aspects are clear. As expected the entropy vanishes exponentially with m, so that choosing very large m, beyond say $m = 7$, has a very limited effect. The relative entropy using the trigonometric basis is perhaps 'smoother' than that of the polynomial. However, in practical terms there is very little to choose between them, at least given these densities. Notice also that since linear transformation of the basis $(\phi_1(x), .., \phi_m(x))$ would imply simply another member of (2). Thus there is no theoretical justification for, for example, orthonormalising the bases, or indeed taking any other linear transformation.

In fact the density functions have been chosen with care. They represent the densities of the cube root and the square of the cube root of a uniform random variable, respectively. Deliberately we have not chosen the uniform density for $p(x)$.The reason is that the uniform is a member of (2) but with $m = 0$. As a consequence the analysis would no longer be fully nonparametric. This turns out to be extremely important in terms of the density estimator, and the choice of m, whether we may consistently estimate the density $p(x)$.

We will consider two criteria for choosing m, the Akaike Information Criteria (AIC) of Akaike (1974) and the Bayesian Information Criteria (BIC) of Schwarz (1980). Labeling the respective optimal choice of m over a set of integers $\mathbb{M}$ given these criteria as $\hat{m}_A$ and $\hat{m}_B$, then given the log-likelihood in (4) and assuming $p_0(x) = 1$,

$$\begin{aligned} \hat{m}_A &= \arg\max_{m\in\mathbb{M}} \left(\sum_{i=1}^{n} \sum_{k=1}^{m} \hat{\theta}_k \phi_k(x_i) - n\varphi_m(\hat{\theta}) - m \right) \\ \hat{m}_B &= \arg\max_{m\in\mathbb{M}} \left(\sum_{i=1}^{n} \sum_{k=1}^{m} \hat{\theta}_k \phi_k(x_i) - n\varphi_m(\hat{\theta}) - \frac{m}{2} \ln n \right). \end{aligned} \tag{15}$$

Although, both the AIC and BIC are consistent, in the strict sense, for m only over

a finite set $\mathbb{M}$, see for example Haughton (1988), since for all θ

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\theta_k\phi_k(x_i) - n\varphi_m(\theta) = O(n) \quad \text{and} \quad m^3/n \to 0,$$

then both

$$\hat{m}_A \to \infty \quad \text{and} \quad \hat{m}_B \to \infty \quad \text{as} \quad n \to \infty, \bar{m} \to \infty, \bar{m}^3/n \to 0,$$

where $\bar{m} = \max(M)$. That is, asymptotically, either criterion will deliver a consistent density estimator.

To illustrate, for six sample sizes between 25 and 800 random samples $\{x_i\}_1^n$,were generated as i.i.d. copies of

$$X^3 = U \sim U(0,1) \quad \text{implying} \quad p(x) = 3x^2,$$

the polynomial basis functions $\phi_k(x) = x^k$ were chosen and the criteria given in (15) were maximised over the restricted set $\mathbb{M} = \{1,2,3,4,5\}$ and the estimated values $\hat{m}_A$ and $\hat{m}_B$, recorded. This was repeated for 5000 Monte Carlo replications and the proportions of outcomes of $\hat{m}_A$ and $\hat{m}_B$ for each member of $\mathbb{M}$ are given in Tables 2a and 2b, in the Appendix. In addition, for each sample size, the Monte Carlo sample averages for each are recorded. As expected the BIC tends to choose a slightly more parsimonious model for a given sample size. In parametric models this is viewed as an advantage. For a nonparametric density estimator this may not necessarily be the case, particularly if we are only optimising over a very restricted subset.

To return to the issue of not choosing $p(x) = 1$, suppose that instead we let $X \sim U(0,1)$. In this case, the solution to (7) is $\theta_0 = 0$, for every m. Since $\mathbb{M}$ can not include 0 then as $n \to \infty$ $\hat{m}_A$ and $\hat{m}_B$ can not converge to 0. Thus neither criterion can be consistent, at least under the null hypothesis. Moreover, since the primary aim is to provide a goodness of fit test, and since the density cannot be uniform under both the null and the alternative, this would imply very different properties of the estimator under the null and the alternative.

3 A Goodness of Fit Test

3.1 Theoretical Results

The proposed test is essentially the likelihood ratio test of Portnoy (1988) applied in the context of the exponential series estimator of Crain (1974) and Barron and Sheu (1991). That is, we transform the original goodness of fit hypothesis $H_0 : X \sim P(X)$, to

$$H_0 : \lim_{m\to\infty} X \sim P_{\theta_0}, \tag{16}$$

where $p_\theta(x) = dP_\theta$ and θ_0 is the unique solution to (7). The likelihood ratio for testing (16) is given by

$$\begin{aligned} \Lambda_m &= 2\log\left[\frac{p_{\hat{\theta}}(\underline{x})}{p_{\theta_0}(\underline{x})}\right] \\ &= 2n\left[\sum_{i=1}^{n}\sum_{k=1}^{m}(\hat{\theta}_k - \theta_0)\phi_k(x_i) - \left(\varphi_m(\hat{\theta}) - \varphi_m(\theta_0)\right)\right], \end{aligned} \tag{17}$$

where $\underline{x} = (x_1, ..., x_n)$, $\hat{\theta} = \left(\theta_1, .., \hat{\theta}_m\right)$ solves (5) while ϕ and φ_m are defined above. H_0 will be rejected in favour of any complimentary alternative for large outcomes of Λ_m.

Notice that the ratio given here differs subtly from the statistics proposed by Marsh (2000) and Claeskens and Hjort (2004) in the denominator. In the former a profile likelihood ratio test, in the spirit of Murphy and van der Vaart (1997), was proposed while the latter utilised a constant denominator, i.e. the uniform density. The first approach is unnecessarily complex for the current problem, while the second, as was indicated in the previous section, leads to potentially very different behaviour of the density estimator under the null and any alternative.

By utilising the form as in (17) the asymptotic results of both Portnoy (1988) Barron and Sheu (1991) may be employed directly. That is it is relatively trivial to establish the asymptotic distribution of the criterion and that the test will be

consistent against any complimentary fixed alternative,

$$H_1 : \lim_{m\to\infty} X \sim P_\theta, \quad \theta \neq \theta_0. \tag{18}$$

We can also define local alternatives, as in Claeskens and Hjort (2004), defined by

$$H_1^L : \lim_{m,n\to\infty \; ; \; m^3/n\to 0} X \sim P_{\theta_c}, \quad \theta_c - \theta_0 = c\sqrt{\frac{\sqrt{m}}{n}}, \tag{19}$$

where $c = \{c_j\}_{j=1}^n$and establish that the test has non trivial power against this local alternative. These results are presented, without proof, in the following theorem.

Theorem 1 *Suppose that data $\{x_i\}_{i=1}^n$ is generated such that Assumption 1 is satisfied, then;*

(i) Under the null hypothesis H_0 in (16)

$$\lim_{m,n\to\infty \; ; \; m^3/n\to 0} \frac{\Lambda_m - m}{\sqrt{2m}} \sim N(0,1) + o_p(1).$$

(ii) Under any complimentary alternative H_1 in (18), for any finite critical value k_α, of size $\alpha < 1$

$$\lim_{m,n\to\infty \; ; \; m^3/n\to 0} \Pr\left[\frac{\Lambda_m - m}{\sqrt{2m}} \geq k_\alpha\right] = 1.$$

(iii) Under the local alternative H_1^L in (19)

$$\lim_{m,n\to\infty \; ; \; m^3/n\to 0} \frac{\Lambda_m - m - m^{1/2}\sum_{j=1}^m c_j^2}{\left(2m + 4m^{1/2}\sum_{j=1}^m c_j^2\right)^{1/2}} \sim N(0,1) + o_p(1). \quad \blacksquare$$

3.2 Computational Results

As with the density estimator itself, at least theoretically, all is straightforward. The purpose of this section, however, is to highlight the ease of implementation of the test, and to compare its numerical performance with already established procedures.

First we will examine the usefulness of asymptotic critical values, in terms of their finite sample performance. The use of asymptotic critical values is not strictly necessary since the problem is distribution free and exact critical values are thus available

via Monte Carlo. However, direct numerical comparisons with some commonly used goodness of fit tests will yield some insights to the behaviour of the proposed statistics, under the null hypothesis. Specifically we shall compare the Monte Carlo size of asymptotic critical values of the likelihood ratio statistics with both the Kolmogorov-Smirnov and Cramér-von Mises statistics. These are given by

$$\begin{aligned} KS &= \max_i \sqrt{n}\,|F_n(x_i) - P(x_i)| \\ CM &= n\sum_{i=1}^{n} \left(F_n(x_i) - P(x_i)\right)^2, \end{aligned}$$

where $F_n(.)$ is the empirical distribution function and $P(.)$ is the hypothesised distribution.

For comparison we shall consider the finite sample size of critical values based upon three different asymptotic approximations. Specifically, these approximations are based upon

$$\begin{aligned} \lim_{m,n\to\infty} \lambda_m &= \frac{\Lambda_m - m}{\sqrt{2m}} \sim N(0,1) \\ \lim_{n\to\infty} \Lambda_m &\sim \chi^2_m \\ \lim_{n\to\infty} \Lambda^B_m &= \frac{\Lambda_m}{b_m} \sim \chi^2_m, \end{aligned}$$

where $b_m = E[\Lambda_m]/m$ is a Bartlett correction, see Lawley (1956), to the asymptotic chi-square likelihood ratio Λ_m. Since the goodness of fit problem is distribution free b_m can readily be calculated numerically via simulation.

Details of the experiments are as follows. Fixing $p(x) = 3x^2$ and choosing the polynomial basis samples of sizes $n = 50, 100, 200$ and 400 on $X \sim p(x)$ were generated 5000 times with likelihood ratios Λ_3, Λ_4 and Λ_5 constructed as in (17) and subsequently the Akaike and Schwarz criteria applied over $\mathbb{M} = \{3, 4, 5\}$ to give $\Lambda_{\hat{m}_A}$ and $\Lambda_{\hat{m}_B}$ respectively. Note that for the calculation of the Bartlett corrected statistics only the first 200 replications were used to calculate b_m. Likewise, in each replication, the criteria KS and CM were also calculated. Then for three different sizes, $0.1, 0.05$ and 0.01, the proportion of outcomes of these statistics exceeding the asymptotic

critical value was recorded. For the KS and CM statistics the asymptotic critical values tabulated in Anderson and Darling (1952) were employed. All of the Monte Carlo rejection proportions are contained in Tables 3a through 3f.

Although one could generate critical values for the likelihood ratio statistics (although since a grid over both m and n would be required this would be very time consuming) the tables do contain some useful information. First, one should dismiss the possibility of using normal critical values as allowing m to be large enough for these to be accurate is neither practical nor indeed warranted according to the Akaike and Schwarz criteria. On the other hand, the asymptotic chi-square versions fair far better with performance not dissimilar to that of the KS and CM. In particular using the Bartlett correction in this efficient way (if we were to use all 5000 replications then we might as well simply use them to obtain an exact critical value) proves useful for the smaller sample sizes.

The results in Tables 3a through 3f only establish that there is no basis for choosing between the likelihood ratio tests described in this paper and the established goodness of fit procedures in terms only of the properties of these tests when the null hypothesis is true. Consequently, we need to compare the power properties of the tests when the alternative is instead true.

To proceed suppose that the null hypothesis is that an independent sample $\{y_i\}_{i=1}^n$ is generated from a standard normal random variable Y, i.e.

$$H_0 : Y \sim N(0,1). \tag{20}$$

Thus define

$$X = \sqrt[3]{\Phi(Y)} \sim P(x) \quad ; \quad dP(x) = 3x^2,$$

where $\Phi(.)$ is the standard normal distribution function and apply the density estimator to the sample $\{x_i\}_{i=1}^N$, $x_i \sim iid\, X$, again using the polynomial basis. The powers of the likelihood ratio tests will be compared to those of the KS and CM tests as

well as weighted versions of these given in Anderson and Darling (1952), defined by

$$\begin{aligned} KS_\pi &= \max_i \sqrt{n} \left| \frac{(F_n(x_i) - P(x_i))}{\sqrt{P(x_i)\,(1 - P(x_i))}} \right| \\ AD &= CM_\pi = n \sum_{i=1}^{n} \frac{(F_n(x_i) - P(x_i))^2}{P(x_i)\,(1 - P(x_i))}, \end{aligned}$$

the weighted Cramér-von Mises being known as the Anderson-Darling statistic.

We shall only consider the powers of general goodness of fit tests, not any of the many available normality tests, such as those of Shapiro and Wilk (1965). There are two reasons for this. First the hypothesis in (20) is indicative and not the focus of the paper. Second it is difficult to fairly compare the powers of entirely nonparametric procedures, such as those based on the empirical distribution or a density estimator, with statistics designed with specific null distributions in mind. Indeed as Stephens (1974) comments on the results of Shapiro, Wilk and Chen (1968) when such comparisons are made the results may be misleading.

The power the tests under consideration will be compared under four sets of alternatives, as in

$$\begin{aligned} H_1^A &: Y \sim N\,(0.05 \times \mu, 1) \quad ; \quad \mu = 1, ..., 7 \\ H_1^B &: Y \sim N(0, (1 + 0.05 \times \mu)^2) \quad ; \quad \mu = 1, .., 7 \\ H_1^C &: Y \sim \frac{\chi_v^2 - v}{\sqrt{2v}} \quad ; \quad v = 5, 10, ..., 35 \\ H_1^D &: Y \sim \sqrt{\frac{v-2}{v}} t_v \quad ; \quad v = 3, 4, .., 9, \end{aligned} \tag{21}$$

where χ_v^2 and t_v represent, respectively, chi-square and Student-t random variables on v degrees of freedom. Moreover, since under each of these alternatives Y has a well defined density function we can define a point optimal likelihood ratio test, given by

$$PO_j = 2 \sum_{i=1}^{n} \ln \frac{f(y_i \mid H_1^j)}{\phi(y_i)}, \quad j = A, B, C, D,$$

where $\phi(y_i)$ is the standard normal density function. The power of the PO_j tests will then provide an absolute benchmark against which to judge that of the others.

The experiments proceeded as follows. Fixing $n = 200$ exact critical values for all of the tests were obtained via simulations under the null hypothesis, as described above. Using 5000 replications the rejection proportions for the likelihood ratio tests Λ_3, Λ_4 and Λ_5 as well as the information criteria based versions $\Lambda_{\hat{m}_A}$ and $\Lambda_{\hat{m}_B}$, for all four versions of the Kolmogorov-Smirnov and Cramér-von Mises tests (KS, KS_π, CM and AD) and for the point optimal tests PO_j, were simulated under every combination of alternatives given in (21). The results, for each set of alternatives, are presented in Tables 4a through 4d.

The alternatives in (21) were chosen so as to isolate, as far as is possible, alternatives which change the moments of Y one at a time. The exception being for H_1^C under which Y has kurtosis of $12/v$. Table 4a gives powers against changes in the mean. Under these alternatives the established procedures have, in fact, a slender advantage, particularly the Anderson-Darling statistic. Also most of the tests have powers which are a significant fraction of those for the point optimal test. However, the picture is very different for the other alternatives. If it is the variance which changes under the alternative then two features are obvious from Table 4b. First, all versions of the likelihood ratio test have powers significantly larger than the established tests, although the advantage over the AD statistic is less than it is over the others. Second, the power of all of the tests is lower in comparison with the point optimal.

The results are very similar for the skewed chi-square alternative, Table 4c. The powers of all the likelihood ratio tests are similar over m and hence so for the information criteria versions, they are also significantly higher than those of the established tests. In this case though it is the weighted Kolmogorov-Smirnov which performs the best amongst the four established tests. Again all tests have powers which are low compared to the point optimal. For the high kurtosis Student-t alternative the likelihood ratio tests are again significantly more powerful, equally so over all the established tests.

Overall it is clear that with the exception of the case where only the mean changes under the alternative all versions of the proposed likelihood ratio tests have powers which are significantly more powerful than established procedures. In many cases the powers are orders of magnitude higher. This seems to be coincident with cases where the power of the established nonparametric procedures are very low compared to the fully parametric, and therefore in this context infeasible, point optimal test.

In addition, the powers of the Λ_m tests are not sensitive to m, and so the powers of tests based upon the Akaike or Schwarz criteria are very similar. On the other hand, the relative powers of the established tests vary across alternatives. For example the weighted Kolmogorov-Smirnov has high power against skewed alternatives but is by far the worst against alternatives involving higher variances and lower power for the other cases. The Anderson-Darling statistic has more-or-less the opposite relative power characteristics. Consequently, in the absence of any information about the alternative we would not know which version of the established tests to use. However, either of the Akaike or Schwarz criterion likelihood ratio tests offers both consistency over various alternatives and for at least three of the alternatives it has significantly more power.

4 Conclusions

This paper has presented a nonparametric likelihood ratio test for the goodness of fit hypothesis based upon a consistent exponential series density estimator, by bringing together the results of Crain (1974), Portnoy (1988) and Barron and Sheu (1991). The test is very similar to ones provided by Marsh (2000) and Claeskens and Hjort (2004). However, it is simpler to use that the latter and has an advantage over the latter in terms of the consistency of selection criteria for the dimension.

Computationally it is shown that the procedure is feasible, since the dimension of the estimator need not be large and the choice of basis is not crucial. Indeed if the hypothesis is rejected the resultant parsimonious, analytic approximation may still,

in itself, be useful.

In terms of the numerical properties of the tests under the null hypothesis there is little basis for choosing. This is true, in particular since the distribution free nature of the problem implies exact critical values can easily be obtained. Under the alternative however the new tests may be significantly more powerful than tests based upon the empirical distribution function, included the favoured Anderson-Darling (1952) statistic. In addition, since the power properties of such tests are not relatively consistent, in the absence of information about the alternative the proposed tests would seem to have a clear power advantage.

References

Akaike, H. 1974. A new look at the statistical model identification. System identification and time-series analysis. IEEE Trans. Automatic Control AC-19, 716–723

Anderson, T.W. and D.A. Darling 1952. Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes. Annals of Mathematical Statistics, 23, 193-212.t

Barron, A.R. and C-H. Sheu 1991. Approximation of density functions by sequences of exponential families. Annals of Statistics, 19, 1347-1369.

Claeskens, G. and N.L. Hjort 2004. Goodness of fit via nonparametric likelihood ratios. Skandanavian Journal of Statistics, 31, 487-513.

Conover, W.J. 1999. Practical nonparametric statistics, John Wiley and Sons, New York.

Crain, B.R. 1974. Estimation of distributions using orthogonal expansions. Annals of Statistics, 2, 454–463.

Crain, B.R. 1976. More on estimation of distributions using orthogonal expansions. Journal of the American Statistical Association, 71, 741–745.

Crain, B.R. 1977. An information theoretic approach to approximating a probability distribution. SIAM Journal of Applied Mathematics, 32, 339–346.

Darling, D.A. 1957. The Kolmogorov-Smirnov, Cramér-von Mises tests. Annals of Mathematical Statistics, 28, 223-238.

D. M. A. Haughton, D.M.A. 1988. On the choice of a model to fit data from an exponential family. Annals of Statistics, 16, 342-355.

Lawley, D.N. 1956. A general method for approximating to the distribution of likelihood ratio criteria. Biometrika, 43, 295-303.

Marsh, P.W.N. 2000. Nonparametric likelihood ratio tests. Discussion paper in Economics, 00/56, University of York.

Murphy, S.A. and A.W. Van der Vaart 1997. Semiparametric likelihood ratio inference. Annals of Statistics, 25, 1471-1509.

Portnoy, S. 1988. Asymptotic behaviour of likelihood methods for exponential families when the number of parameters tends to infinity. Annals of Statistics, 16, 356-366.

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Shapiro, S.S. and M.B. Wilk 1965. An analysis of variance test for normality (complete samples). Biometrika, Vol. 52, 591-611.

Shapiro, S.S., M.B. Wilk and H.J. Chen 1968. A comparative study of various tests for normality. Journal of the American Statistical Society, 63, 1343-1372.

Stephens, M.A. 1974. EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69, 730-737.

Appendix

This appendix contains all of the tables of results for the experiments described in sections 2 and 3.

Table 1: Entropy $D(p_{\theta_0}(x)|p(x))$ for the densities in (13) and given bases (14) and increasing m.

$\phi_k(x)$	$p(x)$	m							
		1	3	5	7	9	11	13	15
$Cos[\pi kx]$	$p^1(x)$	.0186	.0029	.0020	.0018	.0018	.0017	.0017	.0016
	$p^2(x)$	.0132	.0045	.0024	.0020	.0017	.0015	.0012	.0011
x^k	$p^1(x)$	.0233	.0018	.0016	.0015	.0016	.0013	.0015	.0016
	$p^2(x)$	.0113	.0030	.0021	.0016	.0020	.0016	.0019	.0013

Table 2a: Proportions of outcomes and mean of $\hat{m}_A$.

n	m					$E[\hat{m}_A]$
	1	2	3	4	5	
25	0.344	0.213	0.256	0.146	0.041	2.328
50	0.132	0.210	0.287	0.236	0.135	3.029
100	0.015	0.106	0.318	0.321	0.240	3.666
200	0.000	0.012	0.117	0.383	0.488	4.347
400	0.000	0.000	0.022	0.370	0.608	4.587
800	0.000	0.000	0.002	0.201	0.797	4.796

Table 2b: Proportions of outcomes and mean of $\hat{m}_B$.

n	m					$E[\hat{m}_B]$
	1	2	3	4	5	
25	0.612	0.205	0.150	0.031	0.002	1.605
50	0.401	0.307	0.214	0.063	0.015	1.984
100	0.183	0.343	0.321	0.134	0.019	2.464
200	0.015	0.217	0.359	0.277	0.132	3.295
400	0.000	0.015	0.228	0.460	0.297	4.039
800	0.000	0.003	0.010	0.458	0.529	4.512

Table 3a: Rejection proportions of asymptotic critical values ($m = 3$).

	sample size	50	100	200	400
	sig. level				
λ_3	0.10	0.114	0.097	0.105	0.099
	0.05	0.083	0.067	0.077	0.066
	0.01	0.048	0.028	0.043	0.031
Λ_3	0.10	0.111	0.093	0.101	0.096
	0.05	0.063	0.045	0.059	0.041
	0.01	0.018	0.007	0.015	0.007
Λ_3^B	0.10	0.102	0.092	0.100	0.097
	0.05	0.057	0.043	0.059	0.047
	0.01	0.015	0.009	0.015	0.009

Table 3b: Rejection proportions of asymptotic critical values ($m = 4$).

	sample size	50	100	200	400
test	sig. level				
λ_4	0.10	0.097	0.099	0.101	0.079
	0.05	0.071	0.059	0.068	0.052
	0.01	0.032	0.022	0.030	0.021
Λ_4	0.10	0.094	0.093	0.095	0.094
	0.05	0.052	0.042	0.052	0.042
	0.01	0.014	0.007	0.012	0.007
Λ_4^B	0.10	0.096	0.107	0.110	0.105
	0.05	0.057	0.051	0.063	0.053
	0.01	0.015	0.009	0.017	0.012

Table 3c: Rejection proportions of asymptotic critical values ($m = 5$).

	sample size	50	100	200	400
test	sig. level				
	0.10	0.048	0.055	0.067	0.092
λ_5	0.05	0.027	0.030	0.042	0.055
	0.01	0.012	0.008	0.020	0.011
	0.10	0.091	0.088	0.093	0.105
Λ_5	0.05	0.035	0.036	0.048	0.053
	0.01	0.009	0.007	0.017	0.011
	0.10	0.109	0.110	0.101	0.104
Λ_5^B	0.05	0.061	0.060	0.058	0.054
	0.01	0.017	0.015	0.015	0.013

Table 3d: Rejection proportions of asymptotic critical values (AIC version).

	sample size	50	100	200	400
test	sig. level				
	0.10	0.102	0.089	0.084	0.087
$\lambda_{\hat{m}_A}$	0.05	0.073	0.056	0.056	0.054
	0.01	0.039	0.021	0.026	0.015
	0.10	0.104	0.092	0.095	0.100
$\Lambda_{\hat{m}_A}$	0.05	0.057	0.044	0.051	0.049
	0.01	0.016	0.007	0.015	0.009
	0.10	0.101	0.102	0.103	0.101
$\Lambda_{\hat{m}_A}^B$	0.05	0.057	0.050	0.061	0.053
	0.01	0.015	0.010	0.017	0.011

Table 3e: Rejection proportions of asymptotic critical values (BIC version).

	sample size	50	100	200	400
	sig. level				
$\lambda_{\hat{m}_B}$	0.10	0.108	0.095	0.096	0.095
	0.05	0.078	0.074	0.067	0.048
	0.01	0.043	0.036	0.034	0.016
$\Lambda_{\hat{m}_B}$	0.10	0.109	0.087	0.097	0.101
	0.05	0.057	0.044	0.055	0.050
	0.01	0.017	0.007	0.014	0.008
$\Lambda^B_{\hat{m}_B}$	0.10	0.101	0.096	0.103	0.097
	0.05	0.057	0.045	0.056	0.048
	0.01	0.015	0.008	0.013	0.010

Table 3f: Rejection proportions of asymptotic critical values (KS and CM tests)

	sample size	50	100	200	400
test	sig. level				
KS	0.10	0.065	0.083	0.089	0.090
	0.05	0.031	0.037	0.041	0.043
	0.01	0.006	0.007	0.007	0.008
CM	0.10	0.110	0.107	0.102	0.096
	0.05	0.048	0.053	0.048	0.048
	0.01	0.015	0.012	0.009	0.009

Table 4a: Power of all tests under $H_1^A : Y \sim N(0.05 \times \mu, 1)$

	μ						
Test	0.05	0.10	0.15	0.20	0.25	0.30	0.35
PO_A	0.145	0.384	0.631	0.873	0.974	0.996	1.000
Λ_3	0.086	0.191	0.409	0.641	0.862	0.969	0.992
Λ_4	0.081	0.169	0.382	0.614	0.843	0.943	0.981
Λ_5	0.075	0.165	0.373	0.609	0.831	0.938	0.976
$\Lambda_{\hat{m}_A}$	0.078	0.169	0.380	0.614	0.839	0.943	0.979
$\Lambda_{\hat{m}_B}$	0.082	0.179	0.394	0.627	0.850	0.955	0.985
KS	0.084	0.191	0.400	0.622	0.834	0.961	0.992
KS_π	0.057	0.076	0.159	0.297	0.487	0.722	0.867
CM	0.099	0.229	0.479	0.726	0.906	0.985	0.998
AD	0.108	0.263	0.524	0.780	0.937	0.991	0.999

Table 4b: Power of all tests under $H_1^B : Y \sim N(0, (1 + 0.05 \times \mu)^2)$

	μ						
Test	0.05	0.10	0.15	0.20	0.25	0.30	0.35
PO_B	0.247	0.629	0.886	0.978	0.997	0.999	1.000
Λ_3	0.102	0.276	0.485	0.764	0.911	0.978	0.996
Λ_4	0.097	0.259	0.478	0.751	0.910	0.984	0.998
Λ_5	0.093	0.259	0.488	0.760	0.913	0.983	1.000
$\Lambda_{\hat{m}_A}$	0.096	0.262	0.482	0.757	0.912	0.983	0.999
$\Lambda_{\hat{m}_B}$	0.099	0.266	0.481	0.759	0.911	0.981	0.997
KS	0.059	0.089	0.166	0.247	0.412	0.591	0.714
KS_π	0.050	0.058	0.061	0.061	0.065	0.128	0.243
CM	0.062	0.100	0.189	0.288	0.495	0.687	0.831
AD	0.074	0.175	0.367	0.632	0.847	0.955	0.993

Table 4c: Power of all tests under $H_1^C : Y \sim \frac{\chi^2(v)-v}{\sqrt{2v}}$

	v						
Test	35	30	25	20	15	10	5
PO_C	0.640	0.696	0.786	0.871	0.958	0.999	1.000
Λ_3	0.318	0.406	0.451	0.557	0.699	0.895	1.000
Λ_4	0.305	0.387	0.428	0.539	0.681	0.879	1.000
Λ_5	0.309	0.390	0.421	0.537	0.676	0.868	1.000
$\Lambda_{\hat{m}_A}$	0.308	0.390	0.427	0.540	0.681	0.875	1.000
$\Lambda_{\hat{m}_B}$	0.312	0.397	0.438	0.547	0.689	0.885	1.000
KS	0.143	0.162	0.184	0.223	0.272	0.369	0.662
KS_π	0.181	0.218	0.278	0.413	0.631	0.965	1.000
CM	0.147	0.165	0.183	0.221	0.285	0.388	0.739
AD	0.134	0.143	0.179	0.221	0.300	0.492	0.931

Table 4d: Power of all tests under $H_1^D : Y \sim \sqrt{\frac{v-2}{v}} t_v$

	v						
Test	9	8	7	6	5	4	3
PO_D	0.619	0.678	0.784	0.872	0.954	0.995	1.000
Λ_3	0.134	0.164	0.212	0.282	0.434	0.742	0.996
Λ_4	0.132	0.154	0.204	0.268	0.429	0.736	0.998
Λ_5	0.129	0.159	0.207	0.277	0.431	0.752	0.995
$\Lambda_{\hat{m}_A}$	0.131	0.158	0.206	0.274	0.431	0.745	0.996
$\Lambda_{\hat{m}_B}$	0.132	0.160	0.209	0.277	0.412	0.742	0.996
KS	0.079	0.085	0.102	0.132	0.196	0.376	0.908
KS_π	0.086	0.093	0.108	0.136	0.198	0.353	0.863
CM	0.073	0.079	0.092	0.117	0.184	0.378	0.961
AD	0.083	0.091	0.102	0.142	0.237	0.509	0.979