THE UNIVERSITY of York

Discussion Papers in Economics

No. 2006/20

Data Driven Likelihood Ratio Tests for Goodness-to-Fit with Estimated Parameters

by

Patrick Marsh

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

Data Driven Likelihood Ratio Tests for Goodness-of-Fit with Estimated Parameters

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

3^{rd} October 2006

Abstract

This paper generalizes the goodness of fit tests of Claeskens and Hjort (2004) and Marsh (2006) to the case where the hypothesis specifies only family of distributions. Data driven versions of these tests are based upon the Akaike and Bayesian selection criteria. The asymptotic distributions of these tests are shown to be standard, unlike those based upon the empirical distribution function. Moreover, numerical evidence suggests that under the null hypothesis performance is very similar to tests such as the Kolmogorov-Smirnov or Anderson-Darling. However, in terms of power under the alternative, the proposed tests seem to have a consistent and significant advantage.

1 Introduction

Recently two papers, Claeskens and Hjort (2004) and Marsh (2006) have introduced new nonparametric likelihood ratio goodness-of-fit tests. Those tests are essentially based upon Portnoy's (1988) test applied in the context of an exponential series density estimator. The given tests are then the likelihood ratios of that nonparametric estimate to either a specified parametric density, in Claeskens and Hjort (2004), or the entropy minimizing approximate density, in Marsh (2006). There, as here, the choice of dimension of the approximate density is data driven via the selection criteria of Akaike (1974) and Schwarz (1978).

This paper extends the application of this testing principle to the general case in which only the general form of a family of distributions is hypothesized, with the parameters of that family unspecified. The most commonly used tests in this circumstance are those based upon the empirical distribution function (edf), such as the Kolmogorov-Smirnov (KS) and Cramér-von Mises (CM) procedures. However, it is well known that tests based on the edf are generally not asymptotically pivotal.

Indeed, as is clear from the analysis of Stephens (1976) and more recently Babu and Rao (2004), the asymptotic distribution of such tests may depend not only upon the hypothesized family of distributions, but also upon the particular parameters of that family. For example, simply to cover all permutations in the Gaussian family four sets of critical values are required, see Stephens (1974, 1976). Further critical values are required for other commonly tested families such as the Gamma (see Pettitt (1978)), the special cases of the Exponential with unknown mean (Lilliefors (1969)) and also with unknown mean and scale (Spinelli and Stephens (1987)).

The test considered in this paper is shown to be, under standard regularity conditions, asymptotically pivotal. That is the asymptotic distribution is the same within and across different hypothesized families. Moreover the asymptotics involved are standard. When appropriately standardized, with respect to the dimension of the approximating exponential, the test is asymptotically standard normal. Thus the

asymptotic distribution applies in all relevant cases, unlike that for edf based statistics. It is also straightforward to establish that the test is consistent against any fixed alternative. The paper also contains an exhaustive numerical study in which it is shown that the size properties of a version of the new test are very similar to those tests based on the edf. However, for cases involving both the Gaussian and Exponential null hypotheses the new test has a significant relative power advantage.

2 Preliminaries

Suppose that our sample $\underline{y}= \{y_i\}_{i=1}^n$are i.i.d. copies of a random variable Y having distribution

$$\Pr[Y \leq y] = F(y),$$

and we wish to test the goodness-of fit hypothesis

$$H_0 : F(y) = F(y;\beta), \tag{1}$$

where $F(.,.)$ is specified up to the unknown $k\times 1$ vector β. Let $\hat{\beta}_n$ denote the maximum likelihood estimator (mle) for β based upon the sample $\underline{y}$.

To proceed consider the (possibly composite) function $h = h(y;\beta) : \mathbb{R}\times\mathbb{R}^k \rightarrow (0,1)$ and suppose that the following assumption holds:

Assumption 1 :

i) The mle is uniformly consistent and

$$\sqrt{n}(\hat{\beta}_n - \beta) = O_p(1) \tag{2}$$

ii) Let $H_i(\beta) = \partial h(Y_i;\beta)/\partial\beta$ *and let* $\mathbf{B}$ *be a ball of radius* $\varepsilon/\sqrt{n}$ *for* $\varepsilon > 0$, *then*

$$\lim_{n\rightarrow\infty}\sup_{\beta\in\mathbf{B}}|H_i(\beta)| \leq c < \infty \quad \textit{for all } i = 1,..,n \tag{3}$$

(iii) Let m *be an asymptotic parameter satisfying* $m = o(n^{1/3})$. ■

On the basis of Assumption 1, if we define

$$x_i = h(y_i; \hat{\beta}_n),$$

then from a mean value expansion for $h(Y_i, .)$

$$h(y_i; \hat{\beta}_n) = h(y_i; \beta) + \left(\hat{\beta}_n - \beta\right)' H_i(\beta_i^*),$$

for each mean value β_i^* lying on a line segment joining $\hat{\beta}_n$ and β. We can write

$$x_i = h_i + e_i,$$

where x_i is as above, $h_i = h(y_i; \beta)$ is an i.i.d. random variable, having density $q(h)$, say, and from (2) and (3) of Assumption 1

$$e_i = e_i\left(F, h, \beta_i^*\right) = O_p(n^{-1/2}),$$

so the e_i are degenerate random variables depending upon the distribution F, the function h and the value β^*. By construction $h \in (0, 1)$, so that the random variables x_i, h_i and e_i are bounded as in

$$x_i, h_i \in (0, 1) \quad ; \quad e_i \in (-1, 1). \tag{4}$$

To proceed we shall define three vectors of raw sample moments up to and including the m^{th} lying in the sample space $\Phi \subset \mathbb{R}^m$, with

$$\bar{x} = n^{-1}\left\{\sum_{i=1}^{n} x_i^j\right\}_{j=1}^{m}, \ \bar{h} = n^{-1}\left\{\sum_{i=1}^{n} h_i^j\right\}_{j=1}^{m} \text{ and } \bar{e} = n^{-1}\left\{\sum_{i=1}^{n} e_i^j\right\}_{j=1}^{m}. \tag{5}$$

Analogously we can also define the population moments corresponding to these random variables as

$$\begin{aligned} \xi &= \left\{\xi_j\right\}_{j=1}^{m} \quad ; \quad \xi_j = \frac{\sum_{i=1}^{n} E[(x_i)^j]}{n} \leq \infty \\ \eta &= \left\{\eta_j\right\}_{j=1}^{m} \quad ; \quad \eta_j = E[(h_i)^j] \leq \infty \\ \varepsilon &= \left\{\varepsilon_j\right\}_{j=1}^{m} \quad ; \quad \varepsilon_j = \frac{\sum_{i=1}^{n} E[(\varepsilon_i)^j]}{n} \leq \infty, \end{aligned} \tag{6}$$

with the finiteness of these moments guaranteed by (4).

3 The Density Estimator

The properties of the exponential series density estimator were first detailed by Crain (1974), and refined and extended by Barron and Sheu (1991). Suppose that we have n observations on the i.i.d. h_i (for the purposes of this exposition we will simply assume that the density $q(h)$, defined on $(0,1)$, satisfies the conditions of Barron and Sheu (1991)). The exponential series density estimator of the density $q(h)$ is the maximum likelihood estimator in the (infinite) exponential family,

$$p_h(\theta) = \exp\left\{\sum_{j=1}^{m} \theta_j \phi_j(h) - \psi_m(\theta)\right\}, \tag{7}$$

where Θ is the m dimensional parameter space, the $\phi_j(.)$ are linearly independent functions spanning $\mathbb{R}^m$ and the cumulant function $\psi_m(\theta)$ is given by

$$\psi_m(\theta) = \ln \int_0^1 \exp\left\{\sum_{j=1}^{m} \theta_j \phi_j(h)\right\} dh.$$

For the sample $\underline{h} = (h_1, .., h_n)$, the density estimator is defined by

$$p_{\underline{h}}(\hat{\theta}) = \lim_{m,n\to\infty,\, m/n\to 0} \sup_{\theta\in\Theta\subset\mathbb{R}^m} \exp\left\{\sum_{j=1}^{m} \theta_j \sum_{i=1}^{n} \phi_j(h_i) - n\psi_m(\theta)\right\}. \tag{8}$$

Although there are a number of different choices for the functions $\phi_j(.)$ for simplicity we will assume that they are polynomials. The difference between the set-up here and that of the papers by Crain (1974) and Barron and Sheu (1991) is that we don't observe the h_i, but instead the x_i. We will also explicitly choose a polynomial basis with

$$\phi_j(h) = h^j.$$

As Marsh (2006) details, choosing polynomials over say trigonometric series has little numerical effect, in the known parameter case. Since, as far as the estimator itself is concerned, the effect of estimating nuisance parameters has only a small impact upon its numerical accuracy, according to evidence presented below, the same is true in this case also.

To apply the exponential series density estimator, note that the mle $\hat{\theta}$ given in (7) can be written as the solution to a set of m equations, as in

$$\int_0^1 h^j \, p_h\left(\hat{\theta}\right) dh = \bar{h} \quad ; \quad j = 1, .., m. \tag{9}$$

Analogously, and since the law of large numbers implies $n^{-1} \sum_{i=1}^n h_i^j \to_p \eta_{0j}$, then we can define a value θ_0 by

$$\int_0^1 h^j \, p_h\left(\theta_0\right) dh = \eta_0 \quad ; \quad j = 1, .., m, \tag{10}$$

with $\eta_0 = \left\{\eta_{0j}\right\}_{j=1}^m$ and $p_h\left(\theta_0\right)$ is the minimum entropy estimator for $p(h)$ in the exponential family. Finally, since in fact we observe $\{x_i\}_1^n$, then we define the value $\tilde{\theta}$ by

$$\int_0^1 h^j p_h\left(\tilde{\theta}\right) dh = \bar{x}. \tag{11}$$

Importantly, given the values $\bar{h}$, η and $\bar{x}$ in $\mathbb{R}^m$ the parameter values $\hat{\theta}$, θ_0 and $\tilde{\theta}$, are uniquely determined via the convexity of the exponential likelihood, see Barron and Sheu (1991).

To detail the asymptotic properties of the estimated density $p_h\left(\tilde{\theta}\right)$, suppose that we have chosen h so that the log-density $lp(h) = \log[p(h)]$ has $r-1$ absolutely continuous derivatives and that its r^{th} derivative, $d^r lp(h)/dh^r$ is square integrable, i.e. so that $lp(h) \in W_2^r$, the Sobolev space of functions on $[0,1]$. Also define the relative entropy for two densities p_1 and p_2 by

$$D[p_1 \,|\, p_2] = \int_0^1 \ln\left[\frac{p_1(h)}{p_2(h)}\right] p(h) dh,$$

then:

Theorem 1 *Let $\tilde{\theta}$ denote the estimated exponential parameter determined by (11) and suppose that the conditions in Assumption 1 are met, then*

$$D[p(h) \mid p_h(\tilde{\theta})] = O_{pr}\left(m^{-2r} + m/n\right). \quad \blacksquare$$

Theorem 1, proved in Appendix I, demonstrates that in terms of the density estimator itself the effect of observing $\{x_i\}_1^n$ rather than $\{h_i\}_1^n$ is asymptotically negligible, provided that the conditions of Assumption 1 are met. The rate of convergence of the estimator in the simpler case is at least of order $O_p\left(n^{-\frac{2r}{1+2r}}\right)$ when m is chosen optimally so that $m \propto n^{\frac{1}{1+2r}}$. It should not be surprising that the rate of convergence is unaffected when parameters are replaced by $\sqrt{n}$ uniformly consistent estimators.

In the following section the details of the proposed nonparametric likelihood ratio test will be given along with it's asymptotic distributions under both the null and alternative.

4 The Likelihood Ratio Test

As in the simpler goodness-of-fit case the proposed test is the likelihood ratio test of Portnoy (1988) applied via the density estimator of Crain (1974) and Barron and Sheu (1991). That is, we replace the goodness of fit hypothesis, with one within the (infinite) exponential family (7) i.e.,

$$H_0 : F(y) = F(y;\beta) \Rightarrow H_0 : h \sim \lim_{m\to\infty} P(\theta_0), \tag{12}$$

where $p_h(\theta_0) = dP(\theta_0)$ and θ_0 is the solution to (10). It is important to note that $P(\theta_0)$ represents the distribution of the transformed observations $x = h(y,\beta)$. However, given observations on $\{x_i\}_1^n$ rather than the unobserved $\{h_i\}_1^n$ Portnoy's (1988) test becomes

$$\begin{aligned}\Lambda_m &= 2\log\left[\frac{p_{\underline{x}}\left(\tilde{\theta}\right)}{p_{\underline{x}}\left(\theta_0\right)}\right] \\ &= 2n\left[\left(\tilde{\theta}-\theta_0\right)'\bar{x} - \left(\psi_m\left(\tilde{\theta}\right) - \psi_m\left(\theta_0\right)\right)\right],\end{aligned} \tag{13}$$

where $\tilde{\theta}$ is the solution to (11).

For every alternative distribution for Y there is a unique alternative distribution for h on $(0,1)$ and associated with that distribution will be another consistent density

estimator given by say, $p_h(\theta_1)$. In practice, of course, θ_1 will neither be known nor specified. None-the-less, as with analogous tests in the parametric exponential family, (12) will be rejected for large values of Λ_m.

The following Theorem, again proved in the appendix gives the asymptotic distribution of the likelihood ratio test statistic both under the null hypothesis (12) and also demonstrates consistency against fixed alternatives.

Theorem 2 *Suppose that we construct* $\{x_i\}_{i=1}^n$ *as described above and that the conditions required in Assumption 1 are met, then*
(i) Under the null hypothesis $H_0 : h \sim \lim_{m\to\infty} P(\theta_0)$,

$$\lim_{m,n\to\infty\ ;\ m=o(n^{1/3})} \frac{\Lambda_m - m}{\sqrt{2m}} \sim N(0,1) + o_p(1). \tag{14}$$

(ii) Under any complementary alternative $H_1 : h \sim Q \neq P(\theta_0)$*, then for any critical value* k_α *of size* $\alpha < 1$

$$\lim_{m,n\to\infty\ ;\ m=o(n^{1/3})} \Pr\left[\frac{\Lambda_m - m}{\sqrt{2m}} \geq k_\alpha\right] = 1. \quad \blacksquare \tag{15}$$

Theorem 1 applies for any distribution of Y satisfying Assumption 1 and any function $h(y;\beta)$ which has a density satisfying the conditions of Barron and Shue (1991). Consequently, the asymptotic distribution of the test is the same both across and within families of distributions. This is not so for those tests based on the edf.

As with the simplest case considered in Claeskens and Hjort (2004) and Marsh (2006) it is possible to demonstrate the existence of power against local alternatives of the form

$$H_0 : \lim_{m\to\infty} h \sim P(\theta_1) \quad ; \quad \theta_1 = \theta_0 + c\sqrt{\frac{\sqrt{m}}{n}}, \quad c'c = O(1).$$

However, since in this more general case the precise parameter values are irrelevant ascribing significance to such alternatives is difficult. More important will be the comparative numerical properties of the test, as described in the following section.

Before proceeding, as with the tests in both Portnoy (1988) and Marsh (2006), an approximation based upon (14), but utilizing the Chi-square distribution, as in,

$$\lim_{m,n\to\infty\ ;\ m=o(n^{1/3})} \Lambda_m \sim \chi^2(m),$$

will prove a more relevant approximation in finite samples. Moreover, in practical applications the remaining unresolved issue remaining is the choice of m.

Here we will employ the data driven model selection criteria of Akaike (1974) and Schwarz (1978). To implement these, define the set of integers $\mathbb{M} = \{1, 2, ..., \bar{m}\}$, and let the estimated dimensions based upon these criteria be $\hat{m}_A$ and $\hat{m}_B$, respectively, which satisfy

$$\begin{aligned} \hat{m}_A &= \arg\max_{m\in\mathbb{M}} \left[\log p_{\underline{x}}\left(\tilde{\theta}\right) - m\right] \\ \hat{m}_B &= \arg\max_{m\in\mathbb{M}} \left[\log p_{\underline{x}}\left(\tilde{\theta}\right) - m\log n\right]. \end{aligned} \tag{16}$$

As with the simpler case either criteria will deliver a consistent density estimator in the sense that if we allow $\bar{m}$ to grow but satisfying $\bar{m} = o(n^{1/3})$, then both $\hat{m}_A$ and $\hat{m}_B$ will diverge. Although in a finite family $\hat{m}_A$ will over-fit, this is not possible in this, asymptotically, infinite family.

5 Numerical Properties

5.1 Properties of the density estimator with estimated parameters

Before detailing the goodness-of-fit test in this circumstance we can enumerate precisely the effect that having to estimate unknown parameters has on the series density estimator of Crain (1974). The Kullback-Leibler distance (or relative entropy) from the true density $q(h)$ to the estimated density $p_h(\tilde{\theta})$ may be fully decomposed as

$$D[q(h) \mid p_h(\tilde{\theta})] = D[q(h) \mid p_h(\theta_0)] + D[p_h(\theta_0) \mid p_h(\hat{\theta})] + D[p_h(\hat{\theta}) \mid p_h(\tilde{\theta})]. \tag{17}$$

While the first term in (17) may be evaluated explicitly, the last two instead need to be simulated. To proceed consider two experiments,

$$\text{(i): } Y \sim Exp(1) \quad \text{and} \quad \text{(ii): } Y \sim N(0,1),$$

and define

$$h_i = \sqrt[3]{F(y_i;\beta)} \quad ; \quad x_i = \sqrt[3]{F(y_i;\hat{\beta})}, \tag{18}$$

where $F(y,\beta)$ is the distribution function of either case (i) or (ii) and $\hat{\beta}$ is the maximum likelihood estimator of the parameters of that distribution on the basis of the sample $\{x_i\}_1^n$. Choosing the cube root in (18) implies that the density of h is $q(h) = 3h^2$. Given this density for any dimension m the unique vector θ_0 can be found via (10). From that density the distance $D[p(h) \mid p_h(\theta_0)]$ can be explicitly calculated, the values for which are given in Table 1a. Notice that from a purely numerical perspective there is very little to be gained from choosing dimensions greater than say $m = 5$.

Given samples $\{h_i\}_1^n$ and $\{x_i\}_1^n$ the respective estimators $\hat{\theta}$ and $\tilde{\theta}$ can be determined from (9) and (11). From the associated density estimators the quantities $D[p_h(\theta_0) \mid p_h(\hat{\theta})]$ and $D[p_h(\hat{\theta}) \mid p_h(\tilde{\theta})]$ can then be calculated. Replicating these 5000 times over both experiments, for four samples sizes of $n = 50, 100, 200, 400$ and for dimensions of $m = 1, 2, 3, 4, 5$, these distances can be numerically evaluated via the Monte Carlo sample averages of the replicated distances.

These values are presented in Tables 1b (for experiment (i)) and 1c (for (ii)). From Table 1 it is clear that the overwhelmingly significant contribution to entropy, in (17), is $D[p_h(\theta_0) \mid p_h(\hat{\theta})]$ which measures our ability to estimate the approximate density $p_h(\theta)$. Both our ability to approximate $p(h)$ with $p_h(\theta)$ and the effect to of having to estimate β, i.e. using $\tilde{\theta}$ rather than $\hat{\theta}$ has a relatively small numerical impact. Thus we may concentrate on providing tests with good statistical properties, i.e. size and power, rather than on the basis of the density estimator itself.

5.2 Size and power properties

In this subsection we shall contrast the finite sample numerical behaviour of the proposed tests based on (13) with those commonly used tests based upon the edf, specifically the KS, CM and Anderson-Darling (AD) tests. These tests are given in either Stephens (1976) or Conover (1999), along with the relevant critical values for all the cases considered here. All of the experiments in this section are based upon the following procedure. The variables $\{y_i\}_i^n$ is generated from a particular distribution, $F(y;\beta)$, while samples lying in $(0,1)$ are then constructed from

$$x_i = \sqrt[3]{F(y_i;\hat{\beta})},$$

so that the relevant density is $q(h) = 3h^2$. For a given m, the test statistic Λ_m is constructed as described above.

The first set of experiments (based upon 5000 replications and for sample sizes of $n = 100, 200, 400$) concern testing the respective null hypotheses,

$$\begin{aligned} H_0^a &: Y \sim Exp[\mu_a], \\ H_0^b &: Y \sim N(\mu_b, \sigma^2), \end{aligned}$$

where μ_a, μ_b and σ^2 are to be estimated. Here we shall first consider the impact of choosing different values of m for the approximation. Data were generated using specified values of $\mu_a = 1$, $\mu_b = 0$ and $\sigma^2 = 1$. Tables 2a and 2b give the rejection frequencies of asymptotic critical values, at both the 5% and 10% significance level, based upon the $\chi^2(m)$ approximation for Λ_m and based upon tabulated values for the KS, CM and AD tests. Naturally, different critical values are required for the latter three tests for each of the hypotheses. For hypothesis H_0^a the values of $m = 6, 9, 12$ were chosen while for H_0^b, $m = 6, 9, 12, 15$ are used.

Two salient points are worth highlighting from Table 2. First it is clear that there are versions of the new test which have very similar size properties to those of the commonly used tests. Second is that to achieve that, the dimension m needs to be

quite large. This is particularly so for the second hypothesis, which involves an extra parameter to be estimated. Indeed for these cases, on the sole basis of finite sample size, it is necessary that m is larger than for the simpler hypotheses considered in Claeskens and Hjort (2004) and Marsh (2006). Commensurate with that is that the approximation is acceptable only for large sample sizes.

From the numerical work of Marsh (2006) and also that to be presented below, both of the data driven selection criteria in (16) tend to favour model dimensions significantly below those delivering acceptable finite sample size performance. Consequently, we shall also employ a numerical correction to the resulting statistics $\Lambda_{\hat{m}_A}$ and $\Lambda_{\hat{m}_B}$, based upon the principle of Bartlett correction. For a given family $F(y;\beta)$ and sample size n we can construct an estimator of the mean of Λ_m via re-sampling from $\tilde{y}_i \sim F(y;\hat{\beta})$ where $\hat{\beta}$ is the estimator obtained from $\{y_i\}_1^n$. Resampling say R times and constructing $\{\Lambda_{m,r}\}_{r=1}^R$we define, for any m,

$$\upsilon_{m,R} = \frac{1}{R}\sum_{r}^{R} \Lambda_{m,r},$$

so that as $n \to \infty$, $\hat{\beta} \to_p \beta$ and as $R \to \infty$, $\lim_m \upsilon_{m,R} \to_p \lim_{m,n} E[\Lambda_m] = m$. Consequently, we can construct a numerical Bartlett-type correction, giving

$$\frac{m\left(\frac{\Lambda_m}{\upsilon_R} - 1\right)}{\sqrt{2m}} \to_d N(0,1),$$

and with a Chi-square approximation

$$\lim_{m,n\to\infty\ ;\ m=o(n^{1/3})} \bar{\Lambda}_m = \frac{m\Lambda_m}{\upsilon_R} \sim \chi^2(m). \tag{19}$$

It is important to note that no higher-order claim is being made for the approximation in (19). Moreover, the resampling scheme proposed is not a Bootstrap. Specifically υ_R need only be calculated once within any family $F(y;\beta)$, for each m. Thus the computational burden is very low compared to a scheme which would resample for each replication of a simulation study. Although such schemes have been suggested, for example in Janssen, Swanepoel and Veraverbeke (2005), they were not actually

applied. Moreover, the evidence here suggests that such schemes, and the computational cost involved, are not necessary in order to get acceptable behaviour under the null hypothesis.

Consequently, all further experiments are performed on the basis of choosing the dimension m via (16), optimizing over $\mathbb{M} = \{1, 2, 3, 4\}$ and correcting the resulting statistics via (19) with $R = 250$. For each family, each correction factor $\upsilon_{m,R}$ took between 1 and 2 minutes to compute on a Pentium IV 3.0Ghz. The experiments presented in Tables 3a and 3b are entirely analogous those presented in Tables 2a and 2b, excepting that the sample sizes are now $n = 25, 50$ and 100. Tables 4a and 4b repeat these experiments but with data generated by $Exp[5]$ and $N[1, 2]$ random variables, respectively. Once again the tables indicate that the new test, albeit with this numerical Bartlett-type correction, has size broadly comparable with those commonly used tests. In fact for small samples perhaps the new tests performs slightly better, particularly so compared to the KS test. In addition there is nothing to choose between either of the dimension selection criteria.

The final set of experiments concern the relative power of all of the tests. First we shall test

$$H_0 : Y \sim Exp[1] \quad \text{vs.} \quad H_1 : Y \sim \Gamma(v, 1),$$

where $\Gamma(v, 1)$ denotes a Gamma random variable. Values for values of v ranging from -1.44 to 0.56 were used, with the null hypothesis implying that $v = 1$. Fixing the sample size at 100, simulated critical values were taken from the experiments used to generate Table 2a. Then those experiments were repeated, but with the data generated under the alternative hypothesis. The rejection frequencies of those simulated critical values are recorded in Table 5. Similar experiments were conducted for tests of

$$H_0 : Y \sim N(0, 1) \quad \text{vs.} \quad \begin{cases} H_1^a : Y \sim \chi^2(v) - v \\ H_1^b : Y \sim t(v) \end{cases},$$

where $\chi^2(v) - v$ and $t(v)$ are centered Chi-square and Student-t random variables, respectively. Rejection frequencies under the alternatives are presented in Tables 6a

and 6b, respectively, for ranges of values of v.

Under these alternatives there are now significant differences in the performances of the tests. While there is still nothing to choose between the dimension selection criteria, either produces a new test which is significantly more powerful than any of those based upon the empirical distribution function.

6 Conclusions

This paper has extended the nonparametric likelihood ratio goodness-of-fit tests of Claeskens and Hjort (2004) and Marsh (2006) to cover all cases involving estimated parameters. Specifically it has been demonstrated that having to estimate unknown parameters does not affect the rate of convergence of either the density estimator of Crain (1974) and Barron and Sheu (1991), nor of the associated likelihood ratio test of Portnoy (1988). Moreover it is therefore straightforward to so that the new test has an asymptotic distribution which depends upon neither the family being tested, nor the particular member of that family. Thus the test has significant theoretical advantages over those based upon the empirical distribution function.

Numerical evidence suggests that the impact of having to estimate parameters on the density estimator is quite negligible. In terms of the associated new test, initially it was found that the dimension of the model needed to be quite large in order that asymptotic critical values had usable finite sample properties. In order to deliver a more practical procedure data driven dimension selection criteria were employed. Utilizing a very simple and efficient numerical correction practical versions of the new test were demonstrated to have very competitive performance under different null hypotheses. Moreover, under relevant alternatives the new tests were shown to be significantly more powerful.

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. (1962). On the distribution of the two-sample Cramér-von Mises criterion. Annals of Mathematical Statistics, 33, 1148-1159.

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.

Babu, G. and C.R. Rao (2004). Goodness-of-fit tests when parameters are estimated. Sankhya, 66, 63–74.

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. Scandinavian 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.

Janssen, P., J. Swanepoel and N. Veraverbeke (2005). Bootstrapping modified goodness-of-fit statistics with estimated parameters. Statistics and Probability Letters, 111–121.

Lilliefors, H.W. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. Journal of the American Statistical Association, 64, 387-389.

Marsh, P.W.N. (2006). Goodness of fit tests via exponential series density estimation. Journal of Computational Statistics and Data Analysis, to appear.

Pettitt, A. N. (1978). Generalized Cramér-von Mises statistics for the gamma distribution. Biometrika, 65, 232–235.

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.

Spinelli, J.J. and M.A. Stephens (1987). Tests for exponentiality when origin and scale parameters are unknown. Technometrics, 29, 471–476.

Stephens, M.A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. Annals of Statistics, 4, 357–369.

Stephens, M.A. (1974). EDF statistics for Goodness of fit and some comparisons. Journal of the American Statistical Association, 69, 730-736.

Appendix I: Proofs

In order to avoid any ambiguity throughout this appendix the order of magnitude symbol $O(.)$ is defined by

$$a_{n,m} = O\left(b_{n,m}\right) \Longleftrightarrow \lim_{m,n\to\infty ;\ m^3/n\to 0} \frac{a_{n,m}}{b_{n,m}} \leq c_1 < \infty,$$

and analogously for the probabilistic versions $O_p(.)$ and $o_p(.)$. If the quantity under scrutiny does not depend upon the dimension m then the condition $m^3/n \to 0$ becomes redundant.

Proof of Theorem 1:

Consider the vectors given in (5)

$$\begin{aligned} \bar{x} &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i, ..., \sum_{i=1}^{n} x_i^m\right)' \\ \bar{h} &= \frac{1}{n}\left(\sum_{i=1}^{n} h_i, ..., \sum_{i=1}^{n} h_i^m\right)', \end{aligned}$$

and the Euclidean distance between them

$$\left|\bar{x} - \bar{h}\right| = \left|\frac{1}{n}\left(\sum_{i=1}^{n}\left(x_i - h_i\right), ..., \sum_{i=1}^{n}\left(x_i^m - h_i^m\right)\right)'\right|.$$

Taking a typical element, the j^{th}, and noting $x_i = h_i + e_i$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(x_i^j - h_i^j\right) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{s=0}^{j}\left(\frac{j!}{s!(j-s!)}h_i^{j-s}e_i^s - h_i^j\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{j}\frac{j!}{s!(j-s!)}h_i^{j-s}e_i^s.\end{aligned}$$

Since $h_i = O_p(1)$ and $h_i \in (0,1)$ while $e_i = O_p(n^{-1/2})$ and $e_i \in (-1,1)$ then

$$h_i^{j-s} = O_p(1) \quad \text{and} \quad e_i^s = O_p(n^{-s/2})$$

and so

$$\sum_{s=1}^{j}\frac{j!}{s!(j-s!)}h_i^{j-s}e_i^s = O_p(n^{-1/2}) \quad \text{for all } j,$$

and hence

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i^j - h_i^j\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{j}\frac{j!}{s!(j-s!)}h_i^{j-s}e_i^s = O_p(n^{-1/2}).$$

Consequently and from the definition of Euclidean distance we have,

$$\left|\bar{x} - \bar{h}\right| = \sqrt{\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i^j - h_i^j\right)\right)^2} = O_p\left(\sqrt{\frac{m}{n}}\right). \tag{20}$$

Considering now the population moment vector η, then from the triangle inequality we have

$$|\bar{x} - \eta| \leq \left|\bar{h} - \eta\right| + \left|\bar{x} - \bar{h}\right| = O_p\left(\sqrt{\frac{m}{n}}\right), \tag{21}$$

which follows from (20) and the same order of magnitude applies for the first distance, see Barron and Sheu (1991).

Extending the decomposition of the Kullback-Leibler divergence of Barron and Sheu (1991) we obtain,

$$D[p(h)|p_h(\tilde{\theta})] = D[p(x)|p_h(\theta_0)] + D[p_h(\theta_0)|p_h(\hat{\theta})] + D[p_h(\hat{\theta})|p_h(\tilde{\theta})]. \tag{22}$$

From Barron and Sheu (1991) the first two terms are, respectively, $O(m^{-2r})$ and $O_p(m/n)$. Application of Lemma 5 in Barron and Sheu (1991), which holds for any two values in $\mathbb{R}^m$, uniquely defined by equations as in (9) and (11) implies

$$O(D[p_h(\hat{\theta})|p_h(\tilde{\theta})]) = O_p\left(\left|\bar{h} - \bar{x}\right|^2\right) = O_p\left(\frac{m}{n}\right),$$

and hence

$$\begin{aligned} O(D[p(h)|p_h(\tilde{\theta})]) &= O\left(D[p(x)|p_h(\theta_0)) + O_p\left(\frac{m}{n}\right)\right. \\ &= O(m^{-2r}) + O_p\left(\frac{m}{n}\right), \end{aligned}$$

as required. ■

Proof of Theorem 2:

Part (i): To proceed we have defined

$$\Lambda_m = 2n\left[\left(\tilde{\theta} - \theta_0\right)' \bar{x} - \left(\psi_m\left(\tilde{\theta}\right) - \psi_m\left(\theta_0\right)\right)\right],$$

where $\tilde{\theta}$ solves (11), or alternatively

$$\psi'_m\left(\tilde{\theta}\right) = \left.\frac{\partial \psi_m\left(\tilde{\theta}\right)}{\partial \theta}\right|_{\theta=\tilde{\theta}} = \bar{x}. \tag{23}$$

Similarly the value θ_0 is defined by

$$\psi'_m\left(\theta_0\right) = \eta_0 = E(\bar{h}). \tag{24}$$

Since the exponential log-likelihood is strictly convex then the mapping

$$\theta(\eta) : \psi'_m\left(\theta\right) = \eta$$

is one-to-one between the parameter space Θ and sample space Φ and application of both (5.6) of Lemma 5 in Barron and Sheu (1991) and also (21) gives

$$O(|\tilde{\theta} - \theta_0|) = O\left(|\bar{x} - \eta_0|\right) = O_p\left(\sqrt{\frac{m}{n}}\right). \tag{25}$$

As a consequence of both (25) and (21) we have that

$$O(|\tilde{\theta} - \theta_0|) = O\left(|\hat{\theta} - \theta_0|\right) \quad \text{and} \quad O(|\bar{x} - \eta_0|) = O\left(|\bar{h} - \eta_0|\right).$$

Now let $U = V - E_{\theta_0}[V]$, $V \sim p_v\left(\theta_0\right)$, then the moment conditions (2.4) and (3.2) of Portnoy (1988) are trivially satisfied since here the elements of $V = \{v_j\}_{j=1}^m$ are bounded, i.e. $v_j \in (0, 1)$. Moreover the expansions provided in Portnoy (1988) in the

proofs of Theorem 3.1 and 3.2 apply for any two pairs of values in $\left(\tilde{\theta}, \theta_0\right)$ and $(\bar{x}, \eta_0)$ given the orders of error satisfy (25).

As there without loss of generality we may assume that the exponential family is parameterized such that

$$\psi'_m(\theta_0) = \eta_0 = E_{\theta_0}[\bar{h}] = 0 \quad \text{and} \quad \psi''_m(\theta_0) = Var_{\theta_0}[\bar{h}] = I_m, \tag{26}$$

that is it is $\bar{h}$ which is assumed to be standardized not $\bar{x}$. None-the-less using (23) and exploiting (26) we have expansions analogous to (3.5) and (3.6) of Portnoy (1988),

$$\begin{aligned} |\tilde{\theta} - \theta_0|^2 &= \left(\tilde{\theta} - \theta_0\right)' \bar{x} - \frac{1}{2} E_{\theta_0}\left[\left(\tilde{\theta} - \theta_0\right)' U\right]^2 + O_p\left(\frac{m^2}{n^2}\right) \\ \left(\tilde{\theta} - \theta_0\right)' \bar{x} &= |\bar{x}_0|^2 - -\frac{1}{2} E_{\theta_0}\left[\left(\left(\tilde{\theta} - \theta_0\right)' U\right)^2 \bar{x}' U\right] + O_p\left(\frac{m^2}{n^2}\right). \end{aligned}$$

Then arguments identical to those giving Theorem 3.1 of Portnoy (1988) yields,

$$|\tilde{\theta} - \theta_0 - \bar{x}| = O_p\left(\frac{m}{n}\right),$$

and consequently

$$\begin{aligned} \Lambda_m &= 2n\left[\left(\tilde{\theta} - \theta_0\right)' \bar{x} - \left(\psi_m\left(\tilde{\theta}\right) - \psi_m(\theta_0)\right)\right] \\ &= n\left[|\bar{x}|^2 - |\tilde{\theta} - \theta_0 - \bar{x}|^2 + \frac{1}{6} E_{\theta_0}\left(\left(\tilde{\theta} - \theta_0\right)' U\right)^3\right] + O_p\left(\frac{m^2}{n}\right), \end{aligned}$$

and so

$$\begin{aligned} \lambda_m &= \frac{\Lambda_m - m}{\sqrt{2m}} = \frac{n|\bar{x}|^2 - m}{\sqrt{2m}} + o_p(1) \\ &= \frac{n|\bar{h} + \bar{e}|^2 - m}{\sqrt{2m}} + o_p(1) \\ &\leq \frac{n|\bar{h}|^2 - m}{\sqrt{2m}} + \frac{n|\bar{e}|^2}{\sqrt{2m}} + o_p(1) \\ &= \frac{n|\bar{h}|^2 - m}{\sqrt{2m}} + o_p(1), \end{aligned}$$

the latter following since the elements e_i are degenerate random variables as $n \to \infty$. Since by construction $E(\bar{h}) = 0$ and $Var(\bar{h}) = I_m$, then application of the martingale

central limit theorem as in Portnoy given immediately the asymptotic distribution. ■

Part (ii): Suppose that under the fixed alternative the density of $\bar{h}$ is $q_A(h)$ and let θ_1 be the unique solution to

$$\int_0^1 h^j \, p_h\left(\theta_1\right) dh = \int_0^1 h^j \, q_A(h) dh \quad ; \quad j = 1, .., m.$$

Now $|\theta_0 - \theta_1| = O(\sqrt{m})$, while the mle is consistent for θ_1 in that $\left|\tilde{\theta} - \theta_1\right| = O_p\left(\sqrt{m/n}\right)$. Writing,

$$n\left(\tilde{\theta} - \theta_0\right)' \bar{x} = \sqrt{n} \sum_{j=1}^{m} (\tilde{\theta}_j - \theta_{0j}) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i^j,$$

where we have put $\tilde{\theta} = \left(\tilde{\theta}_1, .. \tilde{\theta}_m\right)'$ and $\theta_0 = (\theta_{01}, .., \theta_{0m})$, and noting that since the x_i^j are i.i.d., then for each j

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i^j = O_p(1),$$

and consequently

$$n\left(\tilde{\theta} - \theta_0\right)' \bar{x} = O_p\left(m\sqrt{n}\right). \tag{27}$$

Furthermore, from the definition of the exponential density $\psi_m\left(\theta_0\right)$ is $O(1)$, while $\psi_m(\tilde{\theta})$ is $O_p(1)$, and hence

$$n\left(\psi_m\left(\tilde{\theta}\right) - \psi_m\left(\theta_0\right)\right) = O_p(n). \tag{28}$$

Together (27) and (28) imply that under any fixed alternative

$$\Lambda_m = O_p\left(m\sqrt{n} + n\right),$$

and hence

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

Appendix II Tables

Table 1a: Kullback-Leibler distances

m	1	2	3	4	5	6	7
$D\left[q(h) \mid p_h(\theta_0)\right]$	.0186	.0115	.0028	.0023	.0008	.0007	.0007

Table 1b: Monte Carlo Kullback-Leibler distances for $Y \sim Exp[1]$

i) $D\left[p_h(\theta_0) \mid p_h(\hat{\theta})\right]$

n	50	100	200	400
m				
1	.0213	.0198	.0192	.0188
2	.0183	.0145	.0129	.0122
3	.0161	.0085	.0056	.0042
4	.0222	.0105	.0062	.0041
5	.0320	.0130	.0064	.0035

ii) $D\left[p_h(\hat{\theta}) \mid p_h(\tilde{\theta})\right]$

n	50	100	200	400
m				
1	.0010	.0008	.0005	.0003
2	.0023	.0014	.0011	.0005
3	.0036	.0015	.0011	.0004
4	.0040	.0018	.0010	.0006
5	.0049	.0019	.0012	.0004

Table 1c: Monte Carlo Kullback-Leibler distances for $Y \sim N(0,1)$

i) $D\left[p_h(\theta_0) \mid p_h(\hat{\theta})\right]$

n	50	100	200	400
m				
1	.0187	.0186	.0184	.0184
2	.0141	.0127	.0121	.0118
3	.0104	.0065	.0046	.0037
4	.0149	.0082	.0051	.0036
5	.0233	.0099	.0052	.0029

ii) $D\left[p_h(\hat{\theta}) \mid p_h(\tilde{\theta})\right]$

n	50	100	200	400
m				
1	.0043	.0016	.0013	.0007
2	.0060	.0025	.0014	.0006
3	.0084	.0032	.0017	.0009
4	.0099	.0038	.0020	.0010
5	.0133	.0046	.0021	.0011

Table 2a: Monte Carlo rejection frequencies under

$H_0^a : Y \sim Exp[1]$

size	0.10			0.05		
n	100	200	400	100	200	400
Λ_6	0.112	0.096	0.102	0.062	0.058	0.049
Λ_9	0.105	0.103	0.985	0.054	0.049	0.047
Λ_{12}	0.109	0.098	0.101	0.057	0.051	0.050
KS	0.081	0.092	0.103	0.041	0.049	0.051
CM	0.095	0.095	0.096	0.051	0.051	0.049
AD	0.093	0.095	0.100	0.050	0.051	0.049

Table 2b: Monte Carlo rejection frequencies under

$H_0^b : Y \sim N(0,1)$

size	0.10			0.05		
n	100	200	400	100	200	400
Λ_6	0.062	0.067	0.072	0.031	0.035	0.038
Λ_9	0.068	0.072	0.079	0.037	0.040	0.041
Λ_{12}	0.081	0.086	0.095	0.041	0.045	0.047
Λ_{15}	0.095	0.096	0.099	0.043	0.047	0.049
KS	0.076	0.088	0.099	0.036	0.042	0.044
CM	0.105	0.099	0.097	0.055	0.052	0.049
AD	0.101	0.097	0.098	0.053	0.047	0.048

Table 3a: Monte Carlo rejection frequencies under

$H_0 : Y \sim Exp(1)$

size	0.10			0.05			0.01		
n	25	50	100	25	50	100	25	50	100
$\bar{\Lambda}_A$	0.098	0.099	0.102	0.049	0.052	0.052	0.008	0.007	0.011
$\bar{\Lambda}_B$	0.097	0.093	0.096	0.046	0.044	0.050	0.008	0.008	0.011
KS	0.051	0.063	0.081	0.026	0.032	0.041	0.005	0.007	0.010
CM	0.098	0.108	0.095	0.051	0.056	0.051	0.014	0.012	0.012
AD	0.092	0.091	0.093	0.042	0.046	0.050	0.011	0.008	0.011

Table 3b: Monte Carlo rejection frequencies under

$H_0 : Y \sim N(0, 1)$

size	0.10			0.05			0.01		
n	25	50	100	25	50	100	25	50	100
$\bar{\Lambda}_A$	0.094	0.098	0.097	0.055	0.056	0.053	0.013	0.014	0.013
$\bar{\Lambda}_B$	0.093	0.096	0.096	0.047	0.053	0.049	0.014	0.012	0.010
KS	0.059	0.069	0.076	0.027	0.031	0.036	0.006	0.008	0.008
CM	0.127	0.114	0.108	0.068	0.059	0.055	0.017	0.015	0.011
AD	0.086	0.096	0.101	0.047	0.048	0.053	0.007	0.010	0.009

Table 4a: Monte Carlo rejection frequencies under

$H_0 : Y \sim Exp[5]$

size	0.10			0.05			0.01		
n	25	50	100	25	50	100	25	50	100
$\bar{\Lambda}_A$	0.103	0.103	0.102	0.053	0.054	0.054	0.013	0.013	0.012
$\bar{\Lambda}_B$	0.104	0.096	0.102	0.051	0.047	0.052	0.012	0.008	0.012
KS	0.051	0.070	0.078	0.028	0.031	0.040	0.006	0.010	0.009
CM	0.099	0.107	0.098	0.055	0.051	0.050	0.015	0.011	0.012
AD	0.094	0.096	0.096	0.047	0.046	0.050	0.013	0.011	0.010

Table 4b: Monte Carlo rejection frequencies under

$H_0 : Y \sim N(1,2)$

size	0.10			0.05			0.01		
n	25	50	100	25	50	100	25	50	100
$\bar{\Lambda}_A$	0.096	0.097	0.096	0.055	0.054	0.055	0.015	0.014	0.012
$\bar{\Lambda}_B$	0.091	0.095	0.099	0.048	0.045	0.050	0.011	0.012	0.009
KS	0.053	0.068	0.083	0.026	0.031	0.039	0.006	0.007	0.010
CM	0.125	0.111	0.105	0.061	0.058	0.055	0.016	0.012	0.012
AD	0.090	0.096	0.104	0.044	0.047	0.052	0.009	0.011	0.011

Table 5: Rejection frequencies at 5% level for tests of

$H_0 : Y \sim Exp(1)$ vs. $H_1 : Y \sim \Gamma(v,1)$ for $n = 100$.

v	$\bar{\Lambda}_A$	$\bar{\Lambda}_B$	KS	CM	AD
1.44	0.754	0.778	0.484	0.588	0.697
1.36	0.624	0.662	0.336	0.424	0.544
1.28	0.475	0.501	0.227	0.271	0.372
1.20	0.322	0.331	0.124	0.154	0.210
1.12	0.169	0.167	0.066	0.079	0.095
1.04	0.098	0.093	0.050	0.051	0.053
0.96	0.100	0.098	0.056	0.067	0.067
0.88	0.207	0.205	0.145	0.163	0.176
0.80	0.408	0.429	0.299	0.351	0.401
0.72	0.686	0.720	0.523	0.615	0.695
0.64	0.893	0.928	0.802	0.869	0.903
0.56	0.991	0.995	0.961	0.980	0.992

Table 6a: Rejection frequencies at 5% level for tests of $H_0 : Y \sim N(0,1)$ vs. $H_1 : Y \sim \chi^2(v)$ for $n = 100$.

v	$\bar{\Lambda}_A$	$\bar{\Lambda}_B$	KS	CM	AD
10	0.921	0.916	0.669	0.784	0.799
30	0.471	0.496	0.304	0.352	0.332
50	0.309	0.317	0.197	0.227	0.203
70	0.224	0.249	0.162	0.179	0.169
90	0.197	0.218	0.150	0.163	0.142
110	0.175	0.195	0.134	0.139	0.132

Table 6b: Rejection frequencies at 5% level for tests of $H_0 : Y \sim N(0,1)$ vs. $H_1 : Y \sim t(v)$ for $n = 100$.

v	$\bar{\lambda}_A$	$\bar{\lambda}_B$	KS	CM	AD
2	0.997	0.987	0.955	0.977	0.981
4	0.676	0.659	0.475	0.599	0.648
6	0.445	0.386	0.246	0.328	0.374
8	0.298	0.260	0.152	0.193	0.217
10	0.234	0.209	0.113	0.143	0.166
12	0.187	0.175	0.097	0.112	0.134