Goodness of fit via nonparametric likelihood ratios

Gerda Claeskens and Nils Lid Hjort

Texas A&M University and University of Oslo

ABSTRACT. To test if a density f is equal to a specified f_0, one knows by the Neyman–Pearson lemma the form of the optimal test at a specified alternative f_1. Any nonparametric density estimation scheme allows an estimate of f, that is, of the proper location in the space of alternatives to f_0. This leads to estimated likelihood ratios. This article considers classes of goodness of fit tests constructed in this fashion. Properties are studied of tests which for the density estimation ingredient use log-linear expansions. Such expansions are either coupled with subset selectors like the AIC and the BIC regimes, or use order growing with sample size. Our tests are generalised to testing adequacy of general parametric models, and work also in higher dimensions.

The tests are related to but different from the 'smooth tests' which go back to Neyman (1937) and which have been studied extensively in recent literature. Our tests are large-sample equivalent to such smooth tests under local alternative conditions, but different and often better under non-local conditions. A weakness of the nested BIC scheme for choosing model order in this context is exposed.

KEY WORDS: *AIC, BIC, density estimation, goodness of fit, log-linear expansions, nonparametric likelihood ratio*

1. Background, motivation and summary

Let $X_1, \ldots, X_n$ be independent observations from a common density, and suppose it is required to test whether this density is equal to a specified f_0, against the nonparametric alternative that it is not. Of course there is a number of tests available for this situation, for example the Kolmogorov–Smirnov and Cramér–von Mises tests. This paper will discuss density-based omnibus goodness-of-fit tests based on estimated versions of likelihood ratio tests, incorporating nonparametric density estimation in a natural fashion. The methods will also be extended to the case of testing adequacy of parametric models.

To explain the basic idea, suppose for a moment that one envisages a specific alternative f to f_0. In that case the Neyman–Pearson lemma tells us that the optimal test procedure consists in rejecting f_0 when the ideal likelihood ratio statistic

$$\Lambda_n(f) = \prod_{i=1}^n f(X_i) \Big/ \prod_{i=1}^n f_0(X_i) \tag{1.1}$$

is large enough. But nonparametric density estimation strategies are available for producing an estimate $\widehat{f}$ of the unknown f. Hence

$$\Lambda_n(\widehat{f}) = \prod_{i=1}^n \widehat{f}(X_i)/f_0(X_i) \tag{1.2}$$

is a natural estimate of the underlying optimal $\Lambda_n(f)$, constructed without prior assumptions. In other words, if the null hypothesis is not correct, then $\Lambda_n(\widehat{f})$ directs itself adaptively towards the test statistic which would have been optimal at detecting this. In this light, $\Lambda_n(\widehat{f})$ appears to have a stronger omnibus motivation than other reasonable test statistics that have been or could be constructed, like

$$\int (\widehat{f}-f_0)^2\,\mathrm{d}x, \quad \int |\widehat{f}-f_0|\,\mathrm{d}x, \quad \max|\widehat{f}-f_0|/f_0^{1/2},$$

and similar. Such tests have been worked with in previous literature, and then typically employing kernel methods for estimation of the unknown f; see e.g. Bickel and Rosenblatt (1973), Hall (1984), Bowman (1992), Bowman and Foster (1993), and Anderson, Hall and Titterington (1994). Tests of the $\int(\widehat{f}-f_0)^2\,\mathrm{d}x$ type have been considered by Eubank and LaRiccia (1992) with additive expansion estimators for f.

Different density estimation schemes lead to different tests. In (1.2), tests of interest emerge by using a kernel estimate or a start-aided kernel type estimate of the Hjort and Glad (1995) variety for f. Here we choose to focus on estimators constructed via log-linear expansions, however, as they lead to a particularly revealing structure regarding both construction of tests and limit distributions. Specifically, consider

$$f_S(x\,|\,a) = f_0(x)c_S(a)^{-1}\exp\Big\{\sum_{j\in S} a_j\psi_j(x)\Big\} \tag{1.3}$$

for x in the interval of interest, where the ψ_j functions are chosen so as to be orthogonal and normalised w.r.t. f_0, and also orthogonal to the function $\psi_0=1$, that is, $\int f_0\psi_j\psi_k\,\mathrm{d}x = \delta_{j,k} = I\{j=k\}$. Also, S is a subset of the natural integers, like $\{1,\ldots,m\}$, and $c_S(a) = \int f_0\exp(\sum_{j\in S}a_j\psi_j)\,\mathrm{d}x$. Employing this model, the natural test statistic becomes

$$Z_n^* = 2\sum_{i=1}^n \log\frac{f_{S_n^*}(X_i\,|\,\widehat{a})}{f_0(X_i)} = 2n\Big\{\sum_{j\in S_n^*}\widehat{a}_j\bar{\psi}_j - \log c_{S_n^*}(\widehat{a})\Big\}, \tag{1.4}$$

where $\widehat{a}$ is arrived at via maximum likelihood in the particular model indexed by the selected set S_n^*, and where $\bar{\psi}_j = n^{-1}\sum_{i=1}^n\psi_j(X_i)$. Note that the density estimator involved is really nonparametric in situations where the index set S is allowed to grow in size with n.

To make this operational one has first of all to decide on a practical sequence of ψ_j functions, which of course can be done in several ways. The requirement besides orthogonality and normalisation w.r.t. f_0 is also, importantly, that $\int f_0\exp(\sum_j a_j\psi_j)\,\mathrm{d}x$ is finite for all sets of a_js in a neighbourhood of zero. Secondly an integral part of the problem is to decide on a suitable index set selector. It is to be noted that once such an index selection mechanism for S_n^* has been decided on, like the one following Akaike's information criterion (Akaike, 1974), the execution of the test is in principle an easy matter,

in that the required null hypothesis distribution can be obtained by simulation under f_0. This would be as easy and satisfactory as the alternative solution of determining the exact limiting distribution and then making a table based on simulations from this.

An alternative to the likelihood-ratio inspired test statistic Z_n^* is the score test, which here takes the particularly simple form

$$T_n^* = \sum_{j \in S_n^*} n\bar{\psi}_j^2. \tag{1.5}$$

This type of test has its origin in Neyman's 1937 paper on 'smooth tests'. The score test, in conjunction with the so-called Bayesian information criterion BIC for nested subsets (see Schwarz, 1978 and Rissanen, 1987), has been proposed by Ledwina (1994) and has been studied extensively since; see for example Inglot and Ledwina (1996) and references therein.

This article reports on a broad investigation into goodness-of-fit statistics of the type (1.4) and (1.5), including generalised versions useful for testing fit of parametric families. We might stress that our theoretical investigations are motivated not out of necessity for carrying out the tests as such, but to learn about performance properties and for purposes of comparison with other procedures.

Section 2 sets the local alternatives framework inside which our test statistics and several of their competitors may be studied, and provides initial large-sample results. It is proven there that Z_n^* and T_n^* are asymptotically equivalent tests, but only under local alternatives circumstances. The behaviour of Z_n^* and T_n^* is further studied in Sections 3 and 4, where the index set S_n^* is chosen in data-driven ways. We single out for special scrutiny the index set selectors of the AIC and BIC type, along with some other natural strategies, like that of searching for the most important coefficients. Our horizon is broader than that of the traditional setup of only nested submodels; specifically, we allow index sets to be chosen among all subsets within a given range. A strength of our framework and analysis is that quite general subset selection methods are allowed; we are able to characterise the limit behaviour of statistics Z_n^* and T_n^* not only for AIC and BIC type selected subsets, but for much broader classes.

Section 5 considers the situation in which there is a fixed alternative f to f_0. Here the Z_n^* and T_n^* statistics have different performances, and in fact the Z_n^* test can often be expected to perform better. Section 6 discusses extension of ideas to the case of testing adequacy of parametric families $f_0(x, \theta)$, with test statistics of the type

$$Z_n^* = 2 \sum_{i=1}^n \log \frac{f_{S^*}(X_i, \widehat{\theta} \,|\, \widehat{a})}{f_0(X_i, \widehat{\theta})}.$$

The machinery is amenable to testing any parametric model satisfying the usual conditions of regularity, also in higher dimensions. The structure of the tests and results about

them are particularly simple when testing adequacy of location and scale families, like the normal. The applications of the general theory to specific models in Section 7 include testing for multivariate normality. Finally in Section 8 a list of concluding remarks is offered, some pointing to further research work.

In the related context of lack of fit tests in regression, omnibus tests based on orthogonal series expansions are proposed by Eubank and Hart (1992) (see also Hart, 1997) and further generalised by Aerts, Claeskens and Hart (1999, 2000).

2. A nonparametric local alternatives framework

Below we establish a natural framework of local alternative densities, where it is possible to accurately determine the large-sample behaviour of several goodness of fit tests. In particular we shall see that the tests (1.4) and (1.5) are essentially equivalent for large n, under circumstances local to f_0. Some introductory results are also reached here that will be used in later sections.

2.1. Local alternatives. Suppose the real density at play is of the form

$$f = f_0 c(b/n^{1/2})^{-1} \exp\Big\{\sum_{j=1}^{\infty}(b_j/n^{1/2})\psi_j\Big\} \tag{2.1}$$

for certain constants b_j, defined for all b for which the integral $c(b/n^{1/2})$ is finite, writing $c(\beta)$ for $\int f_0 \exp\{\sum_{j=1}^{\infty}\beta_j\psi_j\}$. We work first with a fixed finite set S and consider

$$Z_{n,S} = 2\sum_{i=1}^{n}\log\frac{f_S(X_i\,|\,\widehat{a})}{f_0(X_i)} = 2n\Big\{\sum_{j\in S}\widehat{a}_j\bar{\psi}_j - \log c_S(\widehat{a})\Big\}. \tag{2.2}$$

Here $\widehat{a}$ maximises the likelihood under the model indexed by S, that is, it maximises $\sum_{j\in S}a_j\bar{\psi}_j - \log c_S(a)$. This function is concave, and the maximiser is also the unique solution to the equations

$$\bar{\psi}_j = \mu_j(a) \quad \text{for } j\in S,$$

where $\mu_j(a) = \partial\log c_S(a)/\partial a_j$ is the theoretical mean of $\psi_j(X)$ when X comes from the $f_S(\cdot\,|\,a)$ model.

Thus $Z_{n,S}$ is the classic log-likelihood ratio statistic for testing f_0, inside the parametric family indexed by a_js for $j\in S$. It is known that $Z_{n,S}$ tends to a noncentral chi-squared when the parameters of the model are $O(1/n^{1/2})$ away from their null values, as they are here, but there is an additional complication here in that all the $b_j/n^{1/2}$ parameters for $j\notin S$ are present, i.e. the true density is outside the finite-parametric model in question. Nevertheless, we will prove the following.

LEMMA 1. *Let S be a specified finite set of indexes. Under the local sequence of alternatives (2.1),*

$$Z_{n,S} \to_d \sum_{j\in S}(b_j+N_j)^2 \sim \chi^2_{|S|}\Big(\sum_{j\in S}b_j^2\Big). \tag{2.3}$$

Here the N_js are independent and standard normal, and $|S|$ denotes the number of j in S.

PROOF. The essence of the proof is that $n^{1/2}\widehat{a}_j$ is close to $n^{1/2}\bar{\psi}_j$, that these tend to independent normals $(b_j, 1)$, and that $n \log c_S(\widehat{a})$ is close to $\frac{1}{2}\sum_{j\in S} n\bar{\psi}_j^2$.

More formally, observe that $c_S(a) = 1 + \frac{1}{2}\sum_{j\in S} a_j^2 + O(\sum_{j\in S}|a_j|^3)$ for small a, and define the function

$$\begin{aligned} K_n(u) &= \sum_{i=1}^n \{\log f_S(X_i \mid u/n^{1/2}) - \log f_S(X_i \mid 0)\} \\ &= n\Big\{\sum_{j\in S}(u_j/n^{1/2})\bar{\psi}_j - \log c_S(u/n^{1/2})\Big\} \\ &= \sum_{j\in S}(n^{1/2}\bar{\psi}_j u_j - \tfrac{1}{2}u_j^2) + r_n(u) = K_{n,0}(u) + r_n(u), \end{aligned}$$

say, in which $r_n(u)$ goes pointwise to zero. The K_n function is concave, and the $n^{1/2}\bar{\psi}$ variable is bounded in probability. It follows from results in Hjort and Pollard (1994) that the maximiser of K_n, which is $n^{1/2}\widehat{a}$, is only $o_p(1)$ away from the maximiser of $K_{n,0}$, which is $n^{1/2}\bar{\psi}$; that is, $n^{1/2}(\widehat{a}_j - \bar{\psi}_j) \to_p 0$ for each j. And it is not difficult to prove that $n^{1/2}\bar{\psi}_j \to_d b_j + N_j$ under (2.1) conditions, with independent N_js, via the Lindeberg theorem. These matters combine to give

$$Z_{n,S} = 2n\sum_{j\in S}(\bar{\psi}_j^2 - \tfrac{1}{2}\bar{\psi}_j^2) + o_p(1) = \sum_{j\in S}(n^{1/2}\bar{\psi}_j)^2 + o_p(1),$$

with the required result. ■

Under the null hypothesis the distributional limit of Z_n is a $\chi^2_{|S|}$, of course, and this can be used to test $f = f_0$, provided the set S is selected in advance. Our tests are intended to be more genuinely nonparametric, however, and need the possibility of growing or data-driven subset selectors; see Sections 3 and 4. When the set S is pre-determined, the score test statistic (1.5) takes the simple form $T_{n,S} = \sum_{j\in S} n\bar{\psi}_j^2$, and it is clear from the proof of Lemma 1 that $Z_{n,S}$ and $T_{n,S}$ are asymptotically equivalent under (2.1) conditions. We also need to show that the two tests are close under broader (but still local) circumstances. The somewhat technical proof required for the following result, in which

$$M_m = \max_{j\le m}\|\psi_j\|, \quad \text{with } \|\psi_j\| = \max_x |\psi_j(x)|, \tag{2.4}$$

is placed in the Appendix.

LEMMA 2. *Consider local alternative densities of the form (2.1), and assume that $\sum_{j=1}^\infty |b_j|\|\psi_j\|$ is finite. Let m grow with n slowly enough to have $M_m m^2/n^{1/2} \to 0$. Then $Z_{n,S} - T_{n,S} \to_p 0$ for all subsets S contained in $\{1,\ldots,m\}$.*

3. Behaviour of tests using the AIC regime

The Akaike information criterion (AIC) method amounts to computing

$$\mathrm{AIC}_{n,S} = Z_{n,S} - C|S| = 2n\Big\{\sum_{j\in S}\widehat{a}_j\bar{\psi}_j - \log c_S(\widehat{a})\Big\} - C|S| \tag{3.1}$$

for each of a list of sets S of interest, and then stay with the submodel indexed by the set S_n^* which maximises the criterion. Here C is a constant bigger than 1. Akaike's *johoryo-tokeigaku* in its traditional form uses $C = 2$. This section studies the behaviour of tests of the type (1.4) and (1.5), when the set is selected by the AIC or closely related criteria.

3.1. AIC with all subsets within a finite horizon. Assume that at the outset all subsets S of $\{1, \ldots, m_0\}$ are allowed consideration, where m_0 is fixed. In addition to $\mathrm{AIC}_{n,S}$ we study its closely related score test version

$$\mathrm{AIC}_{n,S}^{(T)} = T_{n,S} - C|S| = \sum_{j \in S}(n\bar{\psi}_j^2 - C). \tag{3.2}$$

From Lemmas 1 and 2 it is clear that both AIC versions will tend in distribution to

$$\mathrm{AIC}_S = \sum_{j \in S}\{(b_j + N_j)^2 - C\} = Z_S - C|S| \tag{3.3}$$

under local alternatives (2.1). Let S_n^* and $S_n^*(T)$ be the sets chosen by the $\mathrm{AIC}_{n,S}$ and $\mathrm{AIC}_{n,S}^{(}T)$ criteria, respectively, with corresponding test statistics $Z_n^* = Z_{n,S_n^*}$ and $T_n^* = T_{n,S_n^*(T)}$ as in (1.4) and (1.5). The empty set is also allowed here, with corresponding $\mathrm{AIC}_{n,\emptyset} = 0 = \mathrm{AIC}_\emptyset$ and $Z_n^* = 0$. Define correspondingly S^* and Z^* for the limit experiment versions, in terms of the i.i.d. sequence $N_1, N_2, \ldots$ of standard normals.

It is now not difficult to derive the limit distribution of Z_n^* under (2.1) circumstances. In fact,

$$\begin{aligned} Z_n^* &= \sum_S Z_{n,S} I\{\mathrm{AIC}_{n,S} \text{ bigger than all other } \mathrm{AIC}_{n,S'}\} \\ &\to_d \sum_S Z_S I\{\mathrm{AIC}_S \text{ bigger than all other } \mathrm{AIC}_{S'}\} \\ &= \sum_S \Big[I\{S^* = S\} \sum_{j \in S}(b_j + N_j)^2 \Big] = Z^*, \end{aligned} \tag{3.4}$$

say. This happens since there is simultaneous convergence in distribution of all the finitely many $Z_{n,S}$ variables to the corresponding collection of Z_S variables, and Z_n^* is here being expressed as a finite sum of functions that are almost continuous in these variables, that is, the set of discontinuities has zero probability for the limit variables. There is also convergence $S_n^* \to_d S^*$.

Note that Z^* is a mixture of $\chi^2_{|S|}(\sum_{j \in S} b_j^2)$ variables, with probabilities

$$p_S(b) = \Pr_b\{S^* = S\} = \Pr\{Z_S - C|S| \text{ bigger than all other } Z_{S'} - C|S'|\}.$$

These are 2^{m_0} complicated but well-defined probabilities defined in terms of $b_j + N_j$ for $j = 1, \ldots, m_0$. In particular there is a point-mass at zero. That $Z^* = 0$ is equivalent to

having 0 bigger than all $2^{m_0}-1$ sums that can be formed of $(b_j+N_j)^2-C$ summands. But this is the same as having 0 bigger than each of the m_0 variables $(b_j+N_j)^2-C$. Hence

$$\Pr_b\{Z^*=0\}=p_\emptyset(b)=\prod_{j=1}^{m_0}\Gamma_1(C,b_j^2), \tag{3.5}$$

featuring the cumulative non-central $\chi_1^2(b_j^2)$ distribution functions.

There are at least two ways of constructing tests for $f=f_0$ based on the machinery developed. A deceptively simple-looking option is to reject the hypothesis if $Z_n^*>0$, with the threshold parameter $C=C_0$ adjusted to lead to a required significance level. By the arguments above, in concert with Lemma 2, one sees that not rejecting $f=f_0$ then is equivalent to having $n\bar\psi_j^2<C_0$ for each $j\le m_0$. The probability of this happening converges to (3.5), and with C_0 chosen such that $\Gamma_1(C_0,0)^{m_0}=0.95$, for example, the asymptotic level of the test becomes 0.05. It is also clear that the test for large n becomes equivalent to rejecting when $\max_{j\le m_0}|n^{1/2}\bar\psi_j|>C_0^{1/2}$. The limiting local power, under (2.1) conditions, is $1-\prod_{j=1}^{m_0}\Gamma_1(C_0,b_j^2)$.

A second approach is to operate with a fixed C, like the value 2 from the original AIC, and reject when $Z_n^*>z_0$, with this positive constant appropriately adjusted. It is not difficult to simulate from the limiting null distribution, which is that described in (3.4) but with the b_js set to zero, to find an appropriate z_0, for a fixed m_0. Limiting local power functions can also be studied via simulations from the (3.4) distribution, to compare with the $1-\prod_{j=1}^{m_0}\Gamma_1(C_0,b_j^2)$ found above for the first type of est.

The two types of test given here have certain parallels to ideas worked with earlier, but only in regression contexts and with a nested sequence of models, rather than as here where all submodels inside a certain range are allowed consideration. See comments in the following subsection.

3.2. AIC with the sequence of nested models. It is not generally possible to reach limit distribution results for Z_n^* if the set S_n^* in question is allowed to be picked from all possible finite subsets. This is because there will always be infinitely many indexes j at which $(b_j+N_j)^2-C$ is bigger than any given constant, so that the intended AIC_S number becomes unlimited.

If one wishes to allow subsets of $\{1,\ldots,m_0\}$ with a growing m_0, therefore, the list of subsets allowed must be restricted. The traditional and simplest solution is to work with the sequence of nested subsets, say $\{1,\ldots,m\}$. Thus consider

$$\mathrm{AIC}_{n,m}=Z_{n,m}-Cm=2n\Big\{\sum_{j=1}^m\widehat a_j\bar\psi_j-\log c_m(\widehat a)\Big\}-Cm\qquad\text{for } m=1,\ldots,m_0,$$

along with its sister version $\mathrm{AIC}_{n,m}^{(T)}=T_{n,m}-Cm=\sum_{j=1}^m(n\bar\psi_j^2-C)$. The test statistics are $Z_n^*=Z_{n,m_n^*}$ and $T_n^*=T_{n,m_n^*(T)}$, where m_n^* and $m_n^*(T)$ maximise the criteria $\mathrm{AIC}_{n,m}$

and $\mathrm{AIC}_{n,m}^{(T)}$, respectively. Let $m = 0$ correspond to the empty set and denote by m^* the maximiser of the limit experiment version $\mathrm{AIC}_m = Z_m - Cm$ for $m = 0, 1, \ldots$, where $Z_m = Z_{\{1,\ldots,m\}} = \sum_{j=1}^m (b_j + N_j)^2$.

To state the next result we need to start with a general approximation lemma, found via results of Götze (1991) and supplementary analysis as in the proof of Lemma 6 below, see our Appendix. The result is that

$$\begin{aligned}\Pr\{(n^{1/2}\bar\psi_1, \ldots, n^{1/2}\bar\psi_{m_0}) \in B\} = \Pr\{(N_1 + b_1, \ldots, N_{m_0} + b_{m_0}) \in B\} \\ + O\big(m_0^{5/4}/n^{1/2}\big)\end{aligned} \tag{3.6}$$

for all measurable convex sets B, provided $m_0 \geq 6$.

LEMMA 3. *Study local alternative densities of the form (2.1), where $\sum_{j=1}^\infty |b_j|\,\|\psi_j\|$ is finite, and assume that $m_0^{5/4}/n^{1/2} \to 0$ as m_0 grows with n. Then the probabilities $p_m(b) = \Pr\{m^* = m\} = \Pr\{\mathrm{AIC}_m$ bigger than all others$\}$ are well defined, and*

$$Z_n^* = Z_{n,m_n^*} \to_d \sum_{m=0}^\infty \Big\{ J_m \sum_{j=1}^m (b_j + N_j)^2 \Big\},$$

where J_m is indicator for the event that $Z_m - Cm$ is bigger than $Z_{m'} - Cm'$ for all other $m' \neq m$.

PROOF. We first note that if $m \leq m_0$ with a fixed m_0, then arguments above easily lead to

$$Z_n^* \to_d Z^*(m_0) = \sum_{m=0}^{m_0} J_m(m_0) Z_m,$$

where $J_m(m_0)$ is indicator for the event that $Z_m - Cm$ is bigger than $Z_{m'} - Cm'$ for all other m' inside $\{0, 1, \ldots, m_0\}$. The limit is again a mixture of $\chi^2_m(\sum_{j=1}^m b_j^2)$ variables. The same result holds for the score test version T_n^*.

For the growing m_0 case it follows from (3.6) that $\Pr\{\mathrm{AIC}_{n,m}$ bigger than all others$\}$ converges to $p_m(b)$. Since $C > 1$ the limiting probabilities are well defined. It follows from Lemma 2 that the limiting distribution of Z_n^* is as claimed. ■

Note that the probabilities $p_m(0)$ for the null case may be obtained explicitly via the generalised arc-sine distribution (Woodroofe, 1982), see Aerts, Claeskens and Hart (1999).

As in the previous subsection one can construct at least two different types of tests. The first test takes the simple form of rejecting if $Z_n^* > 0$, with C properly adjusted. Non-rejection of the null hypothesis in the limit experiment means observing $Z^* = 0$, which is equivalent to having all successive sums $\sum_{j=1}^m \{(b_j + N_j)^2 - C\}$ smaller than zero. The limiting local power function is $1 - p_0(b)$, which may be computed by simulation for different $b = (b_1, b_2, \ldots)$ of interest. The score test version takes the form of non-rejection only if all successive sums $\sum_{j=1}^m (n\bar\psi_j^2 - C)$ are smaller than zero.

The above test is similar to a method used in Hart (1997) in a traditional regression model and for more general regression contexts in Aerts, Claeskens and Hart (1999). Table 7.1 of Hart (1997) may be consulted for choices of C to attain a specified level of the test; in particular, C equal to 3.221, 4.179 and 6.745 corresponds to levels of respectively 0.10, 0.05 and 0.01. The typically used Akaike value of $C = 2$ corresponds in this special context to a significance level of 0.29. This type of test is referred to as an order selection test, which in this case is equivalent to rejecting when $\widetilde{Z} = \max_{j \geq 1} Z_{n,j}/j > C$. The score version is denoted by $\widetilde{T}$.

The second type of test keeps a fixed C value and rejects when Z_n^* exceeds a positive constant z_0, and is similar in spirit to tests used in regression models by Aerts, Claeskens and Hart (2000). For a fixed C it is not difficult to find z_0 via simulations from the limiting null distribution $\sum_{m=0}^{\infty} J_m Z_m$. For example, for $C = 2$, z_0 values 8.606, 13.829 and 27.234 correspond to levels of 0.10, 0.05 and 0.01 respectively.

REMARK. There is no reason to limit study to only the two strategies above; in particular the restriction to nested sets $\{1, \ldots, m\}$ only may be too severe. An extension which is simple in practice and promising in potential, but leads to somewhat more complicated mathematics when it comes to analysing its behaviour, is to use (3.1) or (3.2) again, but searching through all subsets $S \in \mathcal{S}(m_0)$, say. This is the set of all subsets of $\{1, \ldots, m_0\}$, where m_0 perhaps is small, plus all nested sets $\{1, \ldots, m\}$ for $m > m_0$. This could be particularly useful for alternatives that in addition to low order deviations also exhibit some higher order non-zero coefficients. ■

4. The BIC, the BIG, and growing sets

A competing model selection criterion to the AIC, also in the testing context, is the so-called Bayesian information criterion which in the present case takes the form

$$\mathrm{BIC}_{n,S} = Z_{n,S} - (\log n)|S|. \tag{4.1}$$

This section studies the behaviour of tests using the BIC criterion to select the set S. We discover that the BIC applied to nested models only, as commonly done, has disadvantages, and on the other hand that the 'all subsets' version of the BIC turns out to behave just as the different-looking 'all subsets' version of the AIC, for large n.

4.1. BIC with all subsets within a finite horizon. Because of the consistency of BIC as a model selection criterion, Ledwina (1994) proposed to exclude the empty set to avoid having the level of the test tending to zero for growing sample size; we also follow this approach.

Suppose as in Section 3.1 that all (non-empty) subsets inside $\{1, \ldots, m_0\}$ may be considered, where m_0 is fixed. Let S_n^* be the subset with maximal $\mathrm{BIC}_{n,S}$. Define analogously $\mathrm{BIC}_{n,S}^{(T)} = T_{n,S} - (\log n)|S|$ and $S_n^*(T)$ as the winning subset for the T_n tests. The test statistics in the end are Z_n^* and T_n^* as in (1.4) and (1.5), with sets S_n^* and $S_n^*(T)$. As a

consequence of not allowing the empty set, asymptotic distribution theory for tests where the model has been selected by BIC is quite different from that in the previous section.

LEMMA 4. *Under local conditions (2.1), the probability that a set S with two or more elements will be chosen by the BIC goes to zero as n grows. This is valid for both the* $\mathrm{BIC}_{n,S}$ *and the* $\mathrm{BIC}^{(T)}_{n,S}$ *criteria. Also, both Z_n^* and T_n^* tend in distribution to* $\max_{j\le m_0}(b_j+N_j)^2$.

PROOF. We give the demonstration in terms of the $\mathrm{BIC}^{(T)}_{n,S}$ criterion; that the same result then must hold also for the $Z_{n,S}$ tests follows from Lemmas 1 and 2.

Let S be a non-empty set not containing the index m. We shall show that $\{m\}\cup S$ will lose against $\{m\}$. This is because

$$\mathrm{BIC}^{(T)}_{n,\{m\}} - \mathrm{BIC}^{(T)}_{n,\{m\}\cup S} = |S|\log n - \sum_{j\in S} n\bar\psi_j^2,$$

which has no choice but to go to infinity in probability. This proves the first assertion.

The implication is that only the singletons $\{1\},\ldots,\{m_0\}$ can survive the BIC scrutiny when n grows. And of these the index $m^*=m$ is chosen with largest value of $(b_j+N_j)^2$. Thus

$$Z_n^* = \sum_S Z_{n,S} I\{S_n^*=S\} \to_d \sum_{m=1}^{m_0} Z_{\{m\}} I\{S^*=\{m\}\} = \max_{j\le m_0}(b_j+N_j)^2.$$

It is also clear that both Z_n^* and T_n^* are asymptotically equivalent to the test statistic $\max_{j\le m_0} n\bar\psi_j^2$, in the local framework (2.1). ■

REMARK. The model selection criteria AIC and BIC are at the outset quite different in spirit and execution, and in most situations give different results. But, surprisingly, in the present context of all possible subsets inside a limited horizon, the first type of AIC-based test (see Section 3.1) and the BIC-based test give exactly the same results for large n. Both schemes lead under local alternatives to test statistics asymptotically equivalent to $\max_{j\le m_0}|n^{1/2}\widehat\psi_j|$. The limiting local power is given in Section 3.1. ■

4.2. BIC with a sequence of nested models. Consider nested models $\{1,\ldots,m\}$ inside a limit m_0 which now is allowed to grow with n. The submodel with the largest $\mathrm{BIC}_{n,m} = Z_{n,m} - m\log n$ is chosen, with accompanying test statistic $Z_n^* = Z_{n,m_n^*}$, say. Analogously we define $T_n^* = \sum_{j=1}^{m_n^*(T)} n\bar\psi_j^2$, where $m_n^*(T)$ maximises $\mathrm{BIC}^{(T)}_{n,m} = T_{n,m} - m\log n$. This later 'simplification' inside the BIC scheme is commonly employed, see for example Inglot, Kallenberg and Ledwina (1997) and Bogdan (1999). It is formulated in these papers for use inside parametric families, but originated merely as a practical computational issue.

LEMMA 5. *Assume f is of type (2.1), and let m_0 grow with n slowly enough to have $m_0^{5/4}/n^{1/2}\to 0$. Then, with probability converging to 1, the BIC criterion for both Z_n^* and T_n^* picks out the first component only.*

PROOF. Let B_n be the event that $\mathrm{BIC}_{n,1}$ is bigger than all the other $\mathrm{BIC}_{n,m}$ numbers for $m=2,\ldots,m_0$, and let correspondingly C_n be the event that $\mathrm{BIC}^{(T)}_{n,1}$ is bigger than the

other $\mathrm{BIC}_{n,m}^{(T)}$ numbers for $m = 2, \ldots, m_0$. The task is to prove that $\Pr(B_n)$ and $\Pr(C_n)$ both go to 1. The strategy is to accomplish this via approximation to the simpler $\Pr(C_n^0)$, where C_n^0 is the limit experiment version that BIC_1 is bigger than all the other BIC_m numbers for $m = 2, \ldots, m_0$. Here $\mathrm{BIC}_m = \sum_{j=1}^m W_j$, with $W_j = (b_j + N_j)^2 - \log n$.

It follows in fact from the uniform approximation results of Götze (1991), as further discussed and worked with in the course of proving Lemma 6 below, in our Appendix, that $\Pr(C_n) = \Pr(C_n^0) + \rho_n$, where $|\rho_n| = O(m_0^{5/4}/n^{1/2})$. It also follows from approximations arrived at in the proof of Lemma 2 that $\Pr(B_n)$ and $\Pr(C_n)$ are close. By our growth restriction on m_0 it therefore suffices to prove $\Pr(C_n^0) \to 1$. We shall see that this takes place under the milder restriction $m_0/n^{1/2} \to 0$.

We are content to show that $\Pr(D_n^0) \to 1$, where D_n^0 is the event that each of the $W_2, \ldots, W_{m_0}$ variables are negative, since D_n^0 implies C_n^0. But a lower bound for $\Pr(D_n^0)$ is $(1-\lambda_n)^{m_0-1}$, where $1 - \lambda_n = \Pr\{(b+N)^2 < \log n\}$, in terms of a constant b bigger than all of the $|b_j|$. Analysis involving a classic approximation to the normal tail now leads to

$$\begin{aligned} 1 - \lambda_n &= \Phi((\log n)^{1/2} - b) + \Phi((\log n)^{1/2} + b) - 1 \\ &\doteq 1 - \frac{1}{(2\pi)^{1/2}} \frac{1}{n^{1/2}} \exp(-\tfrac{1}{2}b^2) \Big\{ \frac{\exp(b(\log n)^{1/2})}{(\log n)^{1/2} - b} + \frac{\exp(-b(\log n)^{1/2})}{(\log n)^{1/2} + b} \Big\}, \end{aligned}$$

where Φ is the cumulative standard normal. It follows that $(1-\lambda_n)^{m_0-1}$ indeed travels to 1 as long as $m_0/n^{1/2} \to 0$. ■

An important consequence of Lemma 5 is that $Z_n^* \to_d \chi_1^2(b_1^2)$ under local alternatives (2.1). For the class of densities f where $b_1 = 0$ but some of the other b_j are nonzero, the power of the deduced test is equal to the significance level. Since the probabilities $p_m(b)$ are non-zero for dimensions $m > 1$, the AIC based test is likely to outperform the nested sequence BIC for a large class of alternatives.

It is important to note that the performance of tests using BIC as a model selector can be drastically improved by not restricting attention to only nested model sequences, but rather allowing all subsets within a fixed dimension m_0. The fact that m_0 is fixed is not disturbing for practical matters, since it would correspond to typical use, and since it can be allowed to be arbitrarily large. Note that also here the mixture construction of all subsets of $\{1, \ldots, m_0\}$ (where m_0 is fixed) followed by a sequence of nested models is a worthwhile strategy.

4.3. The BIG criterion. As long as we work under local alternatives, the estimates $\widehat{a}_j$ of the model parameters are approximately independent with $N(a_j, 1/n)$ distributions. When trying to test whether all of them are zero it makes sense to hunt for and use the few coefficients with most influence. One strategy is therefore to compute $\widehat{a}_j$ for $j = 1, \ldots, m_0$, and use as test statistics

$$\mathrm{BIG}_{n,m} = Z_{n,B_{n,m}} = 2n \Big\{ \sum_{j \in B_{n,m}} \widehat{a}_j \bar{\psi}_j - \log \int f_0 \exp\Big(\sum_{j \in B_{n,m}} \widehat{a}_j \psi_j \Big) \, \mathrm{d}x \Big\},$$

where $B_{n,m}$ is the set of the m indexes j with biggest values of $|\widehat{a}_j|$.

The behaviour of this test statistic can be understood using Lemma 2, which implies that $\mathrm{BIG}_{n,m}$ is equivalent to the simpler version $\sum_{j\in B_{n,m}} n\bar\psi_j^2$ for large n, under local conditions (2.1). When m_0 is fixed,

$$\mathrm{BIG}_{n,m} \to_d \sum_{B_m} (b_j + N_j)^2,$$

where B_m is the random subset of $\{1,\ldots,m_0\}$ with the m biggest values of $(b_j+N_j)^2$. With $m=1$ this actually again reproduces the test statistic $\max_{j\le m_0} |n^{1/2}\bar\psi_j|$ which was seen to be large-sample equivalent to the tests using either the first type of AIC or the BIC inside all subsets within a finite m_0-horizon. With $m=2$ the test used becomes for large n the same as looking at the sum of the two largest $n\bar\psi_j^2$ contributions, and so on.

There is no limit distribution if m_0 here is allowed to grow beyond bounds. In that case some modifications would be needed for the test statistic, like tapering off higher order terms.

4.4. Local power for tests using growing m. One way of ensuring that the density estimators at work in (1.4) are really nonparametric, in the sense of being able to consistently estimate also densities that cannot be described by finitely many a_j parameters in (1.3), is to let the index set $S=\{1,\ldots,m\}$ grow slowly with n, without applying any further subset or order selector as in Sections 3.2 and 4.2. Thus let in this subsection

$$Z_n^* = 2n\Big\{\sum_{j=1}^m \widehat a_j\bar\psi_j - \log c_m(\widehat a)\Big\}, \tag{4.2}$$

where $c_m(a) = \int f_0 \exp(\sum_{j=1}^m a_j\psi_j)\,\mathrm{d}x$, and the $\widehat a_j$s are found by maximum likelihood inside the $(a_1,\ldots,a_m)$ model. To properly understand its behaviour, and to give recommendations for the choice of m as a function of n, we need to find its limiting null distribution and its local power characteristics. We also take an interest in the score test version $T_n^* = \sum_{j=1}^m n\bar\psi_j^2$.

By Lemma 1 we expect Z_n^* to be approximately a $\chi^2_m(B_m)$ under local alternatives (2.1), where $B_m = \sum_{j=1}^m b_j^2$. With growing m this would lead to limiting normality for $(Z_n^* - m - B_m)/(2m+4B_m)^{1/2}$; this can indeed be proved under the condition $M_m m^2/n \to 0$. However, this leads to a trivial asymptotic local power, since $\sum_{j=1}^\infty b_j^2$ is finite; in situations with a $\chi^2_m(\lambda_m)$, where one tests $\lambda_m=0$, one is only able asymptotically to detect alternatives which are at least $m^{1/2}$ away from zero. In other words, in the present situation, one would need B_m to grow like $m^{1/2}$, in order to have a non-trivial limiting local power result. These considerations lead us to study alternative densities of the form

$$f = f_0 c\big((m^{1/4}/n^{1/2})b\big)^{-1} \exp\Big\{\sum_{j=1}^\infty (m^{1/4}/n^{1/2})b_j\psi_j\Big\}. \tag{4.3}$$

The proof of the following result is found in the Appendix.

LEMMA 6. *Study local alternative densities of the form (4.3), where* $\sum_{j=1}^{\infty} |b_j| \max_x |\psi_j(x)|$ *is finite. If* m *grows with* n *slowly enough to have* $M_m m^{9/4}/n^{1/2} \to 0$, *then* $(Z_n^* - T_n^*)/m^{1/2} \to_p 0$. *If furthermore* $M_m m^{10/3}/n^{1/2} \to 0$, *then*

$$\frac{Z_n^* - m - m^{1/2}B_m}{(2m + 4m^{1/2}B_m)^{1/2}} \quad \text{and} \quad \frac{T_n^* - m - m^{1/2}B_m}{(2m + 4m^{1/2}B_m)^{1/2}} \quad \text{both tend to } \mathrm{N}(0,1),$$

where $B_m = \sum_{j=1}^{m} b_j^2$.

This result, which also implies that $(Z_n^* - m)/(2m)^{1/2}$ and $(T_n^* - m)/(2m)^{1/2}$ tend to $\mathrm{N}(B_\infty/\sqrt{2}, 1)$, where $B_\infty = \sum_{j=1}^{\infty} b_j^2$, is similar in spirit to Theorem 1 in Eubank and LaRiccia (1992). They worked with a different class of test statistics and considered additive expansions of densities, where we use the perhaps more appealing multiplicative expansions and the estimated likelihood ratio tests. It is fair to point out that the technical obstacles we encounter for Lemma 6, tackled in the Appendix, are by necessity more difficult than those met with Eubank and LaRiccia's additive expansions.

A test based on Z_n^* with significance level α must asymptotically be equivalent to rejecting when $Z_n^* > m + (2m)^{1/2} z_0$, where z_0 is the appropriate upper point of the standard normal. It follows from Lemma 6 that the limiting detection power against the (4.3) alternative becomes $\Phi(B_\infty/\sqrt{2} - z_0)$.

Comparing Z_n^* and T_n^* tests using AIC or BIC (Section 3, Section 4.1–2) with those using growing m (this subsection) is not an easy task. The former are able to detect alternatives a little bit closer to the null hypothesis (order $1/n^{1/2}$ away) than those alternatives which are detected by the latter (order $m^{1/4}/n^{1/2}$ away). The submodel selector versions of the tests must downweight higher order components in order to obtain the $1/n^{1/2}$ detection abilities, just as for Kolmogorov–Smirnov and Cramér–von Mises tests. The 'growing m tests', however, can often beat the former ones by converging more slowly but spreading out their power more evenly. A more careful analysis of this phenomenon, in a different but similar context, can be found in Eubank (2000); see also Inglot and Ledwina (1996).

5. Power at a fixed alternative

Assume now that the data come from a fixed density $f \neq f_0$. We shall study the approximate power of our various test statistics. Let $\xi_j = \mathrm{E}_f \psi_j(X)$ be the true mean of $\psi_j(X)$, and write $n^{1/2}(\bar\psi_j - \xi_j) \to_d V_j$, where these are multinormal with covariance structure say $k_{j,l} = \mathrm{cov}_f\{\psi_j(X), \psi_l(X)\}$.

5.1. Tests with fixed index set. Consider first the situation where $Z_{n,S}$ of (2.2) and $T_{n,S}$ of (2.4) operate with a fixed index set S. These are equivalent under the null hypothesis for large n, and in particular both tend to a $\chi^2_{|S|}$ distribution under f_0.

For the $T_{n,S}$ test, under f, it is clear that $T_{n,S}/n \to_p \alpha_T = \sum_{j\in S} \xi_j^2$, and it is not difficult to establish that

$$n^{1/2}(T_{n,S}/n - \alpha_T) \to_d \sum_{j\in S} 2\xi_j V_j \sim \mathrm{N}(0, \tau_T^2), \quad \text{where } \tau_T^2 = 4\xi_S^{\mathrm{t}} K_S \xi_S,$$

where the subset involved is indicated in the notation. Analysing the $Z_{n,S}$ test is somewhat more demanding. The estimators $\widehat{a}_j$ aim at certain least false parameter values $a_{0,j}$ defined as those making $f_S(x \,|\, a)$ closest to the f in question in the Kullback–Leibler sense. In other words, $a_{0,j}$ for $j \in S$ are those maximising $\sum_{j\in S} a_j\xi_j - \log c_S(a)$. Let $Z^0_{n,S}$ be the magical test statistic $2\sum_{i=1}^n \log\{f_S(X_i \,|\, a_0)/f_0(X_i)\}$ which 'knows' these $a_{0,j}$ values. Then one may show that

$$n^{1/2}(Z_{n,S}/n - Z^0_{n,S}/n) = n^{-1/2}\, 2\sum_{i=1}^n \log\{f_S(X_i \,|\, \widehat{a})/f_S(X_i \,|\, a_0)\} \to_p 0,$$

using the facts that $n^{-1}\sum_{i=1}^n \partial \log f_S(X_i \,|\, a_0)/\partial a$ vanishes in probability and that $n^{1/2}(\widehat{a} - a_0)$ has a limiting distribution. Hence

$$Z_{n,S}/n \to_p \alpha_Z = 2\int f \log(f_{a_0}/f_0)\,\mathrm{d}x = 2\Big\{\sum_{j\in S} a_{0,j}\xi_j - \log c_S(a_0)\Big\}$$

and

$$n^{1/2}(Z_{n,S}/n - \alpha_Z) \to_d \mathrm{N}(0,\tau_Z^2) \quad \text{where } \tau_Z^2 = 4a_{0,S}^{\mathrm{t}} K_S a_{0,S}.$$

These arguments and results lead to various power approximations for the two tests. The simplest of these takes the form

$$\Pr\{Z_{n,S} > \gamma_0\} = \Pr\{n^{1/2}(Z_{n,S}/n - \alpha_Z) > n^{1/2}(\gamma_0/n - \alpha_Z)\} \doteq \Phi(n^{1/2}\alpha_Z/\tau_Z),$$

with a similar expression for the $T_{n,S}$ test. Hence the $Z_{n,S}$ test is asymptotically stronger than the $T_{n,S}$ test when $\alpha_Z/\tau_Z > \alpha_T/\tau_T$, and vice versa.

A simple check was carried out for the case of $f = f_a$ with $a = (a_1, \ldots, a_m)$ of finite dimension, where in the above notation $a_{0,j}$ is a_j for $j \le m$ and zero for $j > m$. We used $\psi_j(x) = \sqrt{2}\cos(j\pi x)$ for testing uniformity on $(0,1)$ and could compute the necessary ξ and K for given a. It was quickly revealed that both cases $\rho > 1$ and $\rho < 1$ occur often, where ρ is the determining ratio $(\alpha_Z/\tau_Z)/(\alpha_T/\tau_T)$; in particular there can be no universal dominance of one test over the other.

Three simple experiments were performed in order to assess how relatively likely it is to encounter densities for which the Z test can be expected to outperform the T test. Simulation results are based on 1000 replications each. (i) With $m = 2$ and a_1, a_2 independently generated from the standard normal, giving a fair range of mostly unimodal densities on $(0,1)$, the Z test was better than the T test in about 38% of cases, with about 15% of ρ ratios falling inside $(1.0, 1.1)$ and about 23% above 1.1. (ii) For the broader class represented by $m = 5$, and again with the a_js drawn independently from the standard normal, the emerging densities are much more varied with up to three peaks or valleys. Inside this more varied class the ρ ratios also vary more, and the Z test wins more often, in fact in about 75% of the cases. These wins are often very clear, with about 20% of the

ρ values exceeding 2 and about 3% exceeding 3. (iii) Finally we investigated the case of independently drawn coefficients $a_j \sim \mathrm{N}(0, 1/j^2)$ for $j = 1, \ldots, 10$. This produces densities with some wiggliness to them but otherwise not with exaggerated freakish behaviour; in other words, varied densities that arguably may be considered 'not unlikely' for statistical practice outside the major parametric families. And in such cases the Z test was found to win asymptotically over the T in as much as about 93% of the cases. Most of these wins are also rather clear-cut, with about 27% of the ρ values above 2.

The precise proportions here are not important; the main point is the message conveyed that the Z test quite often can be expected to outperform the T test for large n, for densities likely to occur in practice. It is also comforting to observe that quite comparable results were reached for the case of the normalised Legendre polynomials as basis functions. The important consequence is that many of the score type 'smooth tests', whose theoretical properties have been investigated and found favourable in recent literature, see Ledwina (1994), Inglot and Ledwina (1996), Inglot, Kallenberg and Ledwina (1997) and references therein, often can be outperformed by their likelihood-ratio sister versions.

5.2. Growing or random index sets. Our general test procedures have been motivated by the hope that $\Lambda_n(\widehat{f})$ of (1.2) will be close to the Neyman–Pearson ratio $\Lambda_n(f)$ of (1.1). Various versions of this statement can be proved, under suitable assumptions, depending on the estimating strategy for $\widehat{f}$. A natural result to strive for would be closeness in the sense of

$$n^{-1} \log \Lambda_n(\widehat{f}) - n^{-1} \log \Lambda_n(f) = n^{-1} \sum_{i=1}^{n} \log\{\widehat{f}(X_i)/f(X_i)\} \to_p 0, \tag{5.1}$$

when data come from f. Note that $n^{-1} \log \Lambda_n(f) = n^{-1} \sum_{i=1}^{n} \log\{f(X_i)/f_0(X_i)\}$ goes to $\int f \log(f/f_0)\,\mathrm{d}x$, the Kullback–Leibler distances from f to f_0. Thus (5.1) says that the test statistic (1.2) succeeds in being close enough to the invisible Neyman–Pearson ratio to recover the same Kullback–Leibler distance for large n. One is not always guaranteed (5.1), since it requires stable closeness of $\widehat{f}/f$ to 1 also in areas where f is small, where e.g. kernel estimators might have problems. For expansion estimators, however, we have the following positive result.

PROPOSITION. *Assume f has a representation $f_0 c(a)^{-1} \exp(\sum_{j=1}^{\infty} a_j \psi_j)$, for a fixed set $a = (a_1, a_2, \ldots)$. Consider the mth order estimator $\widehat{f}_{n,m}$, which is the maximum likelihood density estimator based on $X_1, \ldots, X_n$, inside the model $f_m = f_0 c_m(b)^{-1} \exp(\sum_{j=1}^{m} b_j \psi_j)$, where $c_m(b) = \int f_0 \exp(\sum_{j=1}^{m} b_j \psi_j)\,\mathrm{d}x$. Then (5.1) holds for the $\widehat{f}_{n,m}$ sequence, under the minimal condition that m goes to infinity with n and $m < n$.*

PROOF. Let $\widehat{a} = (\widehat{a}_1, \ldots, \widehat{a}_m, 0, \ldots)$ be the maximum likelihood estimates inside the mth order model. These exist, with probability 1, provided only $n > m$; see comments in Crain (1977) and Barron and Sheu (1991). For sequences $b = (b_1, b_2, \ldots)$ in the set B

where $L(b) = \sum_{j=1}^{\infty} |b_j| \, \|\psi_j\|$ is finite, consider the function $K_n(b) = \sum_{j=1}^{m} b_j \bar{\psi}_j - \log c_m(b)$, where $m = m_n < n$ climbs towards infinity when n does. Note that

$$n^{-1} \sum_{i=1}^{n} \log\{\widehat{f}_{n,m}(X_i)/f_0(X_i)\} = \sum_{j=1}^{m} \widehat{a}_j \bar{\psi}_j - \log c_m(\widehat{a}) = K_n(\widehat{a}).$$

The K_n function is concave, and by dominated convergence its mean goes to $K(b) = \sum_{j=1}^{\infty} b_j \xi_j - \log c(b)$ for each $b \in B$. Also, its variance is bounded by $L(b)^2/n$. Hence K_n goes pointwise to K in probability as n grows. Via concavity this is sufficient to guarantee uniform convergence in probability on compact subsets of B, under the $\sum_{j=1}^{\infty} |b_j - b'_j| \, \|\psi_j\|$ metric. Note next that the maximiser of $K(b)$ is $b = a$. The maximiser $\widehat{a}$ of K_n goes in probability to the maximiser a of K, see convexity lemmas in Andersen and Gill (1982, Appendix) and Hjort and Pollard (1994). It follows from these facts that also the maximum of K_n must go in probability to the maximum of K. But $K_n(\widehat{a}) \to_p K(a)$ is seen to be equivalent to $n^{-1} \sum_{i=1}^{n} \log\{\widehat{f}_{n,m}(X_i)/f_0(X_i)\}$ having the same limit as $n^{-1} \sum_{i=1}^{n} \log\{f(X_i)/f_0(X_i)\}$. This proves (5.1). ■

One might similarly work towards proving (5.1) under suitable conditions with various subset selectors employed in (1.4), like with the AIC or BIC. Further results on the closeness of $\log \Lambda_n(\widehat{f})$ to $\log \Lambda_n(f)$ can be reached via a careful study of approximation precision of the best finite-parametric Kullback–Leibler approximation f_m to f; see Crain (1977), Barron and Sheu (1991) and Inglot and Ledwina (1996) for results of relevance. We do not pursue these themes further here.

5.3. Results of a simulation study. Below we illustrate certain aspects of the finite sample behaviour of several proposed tests for uniformity; see Section 7.5 for performance of tests for bivariate normality. The test statistics under comparison are the BIC data-driven likelihood ratio and score statistics Z^*, T^*, and the order selection statistics $\widetilde{Z}$ and $\widetilde{T}$. We consider both a nested model sequence, where the number of added terms is allowed to grow to 10, and the all subsets version, with a maximum of 5 added terms to the null model. The particular choices of where to cut off the series are not of much importance for power behaviour (see also Ledwina, 1994). Critical values at 5% are obtained by simulation of 30,000 datasets under the null model. The simulated power has for each case been calculated from 5,000 generated data sets under the alternative model in question.

In our first setting (a), we generated data from model (1.3), where f_0 is the uniform density on the unit interval, $S = \{1, 2, 3\}$, $a = (-1.2, -0.7, -0.6)$, and ψ_j is the jth order Legendre polynomial. We considered tests employing the Legendre polynomials, and a cosine system $\psi_j(x) = \sqrt{2}\cos(\pi j x)$. The sample size was either 25, 50 or 100. As expected for this setting, the polynomial basis functions perform better than tests using the cosine basis. Since the alternative function concentrates on the first three dimensions in the alternative models' space, in Table 5.1 we observe that the all subsets version slightly loses power in comparison to the nested sequence tests. In this setting, the AIC based order

selection tests have higher power than the BIC based tests, and the likelihood ratio test has higher power than the score test.

One should be careful to generalise these conclusions. In setting (b) the data are generated according to a $\frac{1}{2}:\frac{1}{2}$ mixture of a Beta$(0.5,1)$ and a Beta$(1,0.5)$ distribution. Here the score test has higher power than the likelihood ratio test, and the BIC based test is more powerful than the order selection tests. Differences in power are more pronounced for the Legendre basis than for the cosine system. It is interesting to observe that the all subsets version gives an improvement in power for the order selection tests, while the nested sequence here is preferred for the BIC based tests.

		Nested sequence				All subsets			
		Z^*(bic)	T^*(bic)	$\widetilde{Z}$	$\widetilde{T}$	Z^*(bic)	T^*(bic)	$\widetilde{Z}$	$\widetilde{T}$
(a) poly	n=25	83.94	74.33	86.30	82.43	80.67	74.90	67.31	66.85
	n=50	99.86	99.42	99.98	99.90	99.78	99.38	99.30	98.94
	n=100	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
cos	n=25	50.26	44.48	57.48	55.44	48.12	41.88	36.54	34.50
	n=50	89.48	88.76	95.38	94.86	87.86	86.04	85.10	84.68
	n=100	99.80	99.74	100.0	99.98	99.74	99.56	99.64	99.66
(b) poly	n=25	54.04	65.78	37.84	48.80	48.26	61.32	49.34	52.80
	n=50	76.84	84.46	66.32	74.92	73.46	81.10	75.98	78.22
	n=100	96.30	97.62	93.62	96.16	95.68	96.98	96.44	96.86
cos	n=25	42.06	44.20	25.24	26.48	36.84	40.54	33.00	31.90
	n=50	61.58	65.50	47.16	49.32	56.84	60.66	57.16	57.02
	n=100	88.58	89.88	82.76	83.60	86.68	88.48	87.96	88.10
(c) poly	n=25	33.72	36.52	13.92	16.26	29.76	40.10	29.88	33.52
	n=50	49.18	52.78	29.80	33.96	53.44	61.88	58.52	61.46
	n=100	82.56	84.28	76.38	79.38	88.20	90.86	90.86	91.82
cos	n=25	46.88	44.30	17.92	16.74	53.64	54.88	51.76	50.08
	n=50	70.18	68.88	46.68	46.92	85.10	86.20	87.08	86.98
	n=100	95.86	95.44	92.18	92.48	99.36	99.42	99.54	99.56

TABLE 5.1. *Simulated power results (as %) for uniformity tests. Estimated likelihood ratio (Z) and score tests (T), with order selected via AIC ($\widetilde{Z},\widetilde{T}$) or BIC ($Z^*,T^*$) using either a nested model search or all subsets selection. Basis functions are Legendre polynomials or a cosine system.*

In our third setting (c), data are generated according to the function $1+0.7\cos(4\pi x)$. Cosines are the best choice here. Powers of the likelihood ratio test and score based tests are comparable. Especially for the order selection tests, the all subsets version is definitely preferred to the nested sequence. The last two settings are taken from Ledwina (1994) where the score based BIC test is found to be superior to a variety of other classical tests for uniformity, such as Anderson and Darling's statistic and tests by Stephens (1974) and Neuhaus (1987, 1988).

The main conclusion is that the all subsets version improves on a nested model search

for several alternative densities, and does not lose much in situations where a nested model search is better.

6. Testing a parametric family

The hypothesis to be tested is that the density belongs to a parametric family $f_0(x,\theta)$, where θ is p-dimensional and traditional regularity conditions apply. In particular the family admits two continuous partial derivatives in θ in a neighbourhood around a focal point θ_0.

6.1. General two-stage approach. To describe our test statistics, let functions $\psi_j(x,\theta)$ be orthonormal w.r.t. $f_0(x,\theta)$, and for bounded index sets S consider the extended parametric model

$$f_S(x,\theta\,|\,a) = f_0(x,\theta)\exp\Big\{\sum_{j\in S} a_j\psi_j(x,\theta)\Big\}\Big/ c_S(a,\theta), \tag{6.1}$$

for (θ,a) around $(\theta_0,0)$, where $c_S(a,\theta) = \int f_0\exp(\sum_{j\in S} a_j\psi_j)\,\mathrm{d}x$. It is also a requirement of the basis functions that this integral is finite for all θ in a neighbourhood of θ_0 and for all a in a neighbourhood around zero. There are now several available options. In this subsection we consider the simplest one, at least from the point of view of implementation, which is to start with the maximum likelihood estimate $\widetilde\theta$ inside the f_0 family, and then proceed with finding the maximum likelihood estimators $\widetilde a$ in the (6.1) family considered as a model in a with given $\widetilde\theta$. This leads to a two-stage likelihood ratio statistic of the form

$$Z_{n,1,S} = 2\sum_{i=1}^n \log\{f_S(X_i,\widetilde\theta\,|\,\widetilde a)/f_0(X_i,\widetilde\theta)\} = 2n\Big\{\sum_{j\in S}\widetilde a_j\bar\psi_j(\widetilde\theta) - \log c_S(\widetilde a,\widetilde\theta)\Big\}. \tag{6.2}$$

This is as in (2.2), but with $\bar\psi_j(\widetilde\theta) = n^{-1}\sum_{i=1}^n\psi_j(X_i,\widetilde\theta)$ and with $\widetilde a$ computed conditionally on $\widetilde\theta$. Also consider the simpler score type test statistic

$$T_{n,1,S} = \sum_{j\in S} n\bar\psi_j(\widetilde\theta)^2. \tag{6.3}$$

Assume that the real density takes the form

$$f(x) = f_0(x,\theta_0)\exp\Big\{\sum_{j=1}^\infty (b_j/n^{1/2})\psi_j(x,\theta_0)\Big\}\Big/ c(b/n^{1/2},\theta_0), \tag{6.4}$$

where $c(a,\theta) = \int f_0\exp(\sum_{j=1}^\infty a_j\psi_j)\,\mathrm{d}x$ for coefficients for which the integral exists. Let G be the $p\times|S|$ matrix of elements $g_{i,j}(\theta_0) = \mathrm{E}_0 u_i(X,\theta_0)\psi_j(X,\theta_0)$, where $u(x,\theta) = \partial\log f_0(x,\theta)/\partial\theta$ is the score function of the model and E_0 indicates that X comes from f_0. Taking the derivative of $\int f_0\psi_j\,\mathrm{d}x = 0$ leads to $\int f_0(x,\theta)\partial\psi_j(x,\theta)/\partial\theta\,\mathrm{d}x = -g_j$, where g_j is the jth column of G.

PROPOSITION 1. *Let $J = J(\theta_0)$ be the information matrix of the f_0 model, and let (U_0, N) be a zero-mean $p + |S|$-dimensional normal vector with $\mathrm{Var}\, U_0 = J$, independent standard normal N_js, and $\mathrm{cov}(N_j, U_{0,i}) = g_{i,j}(\theta_0)$. Then, for local alternative densities (6.4),*

$$Z_{n,1,S} \text{ and } T_{n,1,S} \to_d Z_{1,S} = \sum_{j\in S}\Big(b_j - g_j^{\mathrm{t}} J^{-1} \sum_{k\in S} b_k g_k + N_j - g_j^{\mathrm{t}} J^{-1} U_0\Big)^2. \tag{6.5}$$

PROOF. Consider $\bar u = n^{-1}\sum_{i=1}^n u(X_i,\theta_0)$. The mean of $u(X,\theta_0)$ under (6.4) conditions is $n^{-1/2}\sum_{j\in S} b_j g_j$ plus smaller order terms, and use of the Lindeberg theorem leads to $n^{1/2}\bar u \to_d \sum_{j\in S} b_j g_j + U_0$. There is also simultaneous convergence with $n^{1/2}\bar\psi_j(\theta_0) \to_d b_j + N_j$, where the N_js are as described above, with and $\mathrm{cov}(N_j, U_{0,i}) = \mathrm{E}_0 u_i\psi_j = g_{i,j}(\theta_0)$. This leads first to $n^{1/2}(\widetilde\theta - \theta_0)$ being well enough approximated with $J^{-1}n^{1/2}\bar u$, which goes to $J^{-1}\sum_{j\in S} b_j g_j + J^{-1}U_0$, and next to

$$\begin{aligned} n^{1/2}\bar\psi_j(\widetilde\theta) &= n^{1/2}\bar\psi(\theta_0) + \Big(n^{-1}\sum_{i=1}^n \frac{\partial\psi_j(X_i,\theta_0)}{\partial\theta}\Big)^{\mathrm{t}} n^{1/2}(\widetilde\theta - \theta_0) + o_p(1) \\ &\to_d b_j + N_j - g_j^{\mathrm{t}} J^{-1}\Big(\sum_{k\in S} b_k g_k + U_0\Big). \end{aligned} \tag{6.6}$$

This gives the claimed limit distribution for $T_{n,1,S}$. One may also show that $Z_{n,1,S} - T_{n,1,S} \to_p 0$, as with the convexity arguments of Lemma 1. ■

REMARK. The limit variable (6.5) has a much simpler structure when the g_j vectors are zero. Such an orthogonality of basis functions w.r.t. the score functions can actually always be achieved. One uses a Gram–Schmidt procedure to make from the original $\psi_1(x,\theta), \psi_2(x,\theta), \ldots$ functions another sequence $\psi_1'(x,\theta), \psi_2'(x,\theta), \ldots$ functions which are orthogonal to the $u_i(x,\theta)$ functions, w.r.t. $f_0(x,\theta)$. When test statistics (6.2) or (6.3) are to be computed, one uses such $\psi_j'(x,\theta)$ functions constructed around the estimated point $\widetilde\theta$. This would be similar to a method invented in a different context by Khmaladze (1979). In this case, therefore, the limit distribution (6.5) reduces to the much simpler non-central χ^2 distribution (2.3). Another construction of tests with χ^2 type limits is given below. ■

6.2. Second general approach. A computationally more involved but nevertheless quite natural strategy is to use the full maximum likelihood solution $(\widehat\theta, \widehat a)$ inside the (6.1) family. This leads to

$$\begin{aligned} Z_{n,2,S} &= 2\sum_{i=1}^n \log\{f_S(X_i, \widehat\theta\,|\,\widehat a)/f_0(X_i,\widetilde\theta)\} \\ &= 2n\Big\{\sum_{j\in S}\widehat a_j\bar\psi_j(\widehat\theta) - \log c_S(\widehat a,\widehat\theta)\Big\} + 2\sum_{i=1}^n \log\{f_0(X_i,\widehat\theta)/f_0(X_i,\widetilde\theta)\}. \end{aligned} \tag{6.7}$$

The limiting distribution must be a non-central χ^2, by classical theory, if $b_j = 0$ for j outside S in (6.4). We wish to assess the distribution of $Z_{n,2,S}$ also in the wider local case, however, and also need more informative approximations in order to study the behaviour when the S set is growing or arrived at via a data-based selection criterion. This necessitates work summarised by the following. In addition to $\bar u = n^{-1}\sum_{i=1}^n u(X_i,\theta_0)$ and $\bar\psi = n^{-1}\sum_{i=1}^n \psi(X_i,\theta_0)$, define $\bar v = \bar\psi - G^{\rm t}J^{-1}\bar u$, a variable becoming independent of $\bar u$ in the limit.

PROPOSITION 2. *Under local (6.4) conditions the test statistic $Z_{n,2,S}$ is only $o_p(1)$ away from $Z^0_{n,2,S} = n\bar v^{\rm t}(I - G^{\rm t}J^{-1}G)^{-1}\bar v$, and the limit distribution is a non-central $\chi^2_{|S|}$ with excentricity parameter*

$$\Big(b_S - \sum_{j\in S} b_j G^{\rm t}J^{-1}g_j\Big)^{\rm t}(I - G^{\rm t}J^{-1}G)^{-1}\Big(b_S - \sum_{j\in S} b_j G^{\rm t}J^{-1}g_j\Big),$$

where b_S is the vector with b_j for $j\in S$.

PROOF. We rely on maximum likelihood asymptotics inside regular parametric families. The score function for the (6.1) model, when evaluated at the point $(\theta_0,0)$, is the $p\times|S|$ vector $(u(x,\theta_0),\psi(x,\theta_0))$. It has

$$\mathrm{Var}_0\begin{pmatrix} u(X,\theta_0)\\ \psi(X,\theta_0)\end{pmatrix} = \begin{pmatrix} J & G\\ G^{\rm t} & I\end{pmatrix} \quad \text{with inverse } \begin{pmatrix} J & G\\ G^{\rm t} & I\end{pmatrix}^{-1} = \begin{pmatrix} K_{00} & K_{01}\\ K_{10} & K_{11}\end{pmatrix},$$

say; in fact, $K_{11} = (I - G^{\rm t}J^{-1}G)^{-1}$, $K_{01} = -J^{-1}GK_{11}$ and $K_{00} = J^{-1} + J^{-1}GK_{11}G^{\rm t}J^{-1}$. It is now the case that

$$\begin{aligned}\widetilde\theta - \theta_0 &= J^{-1}\bar u + o_p(n^{-1/2}),\\ \widehat\theta - \theta_0 &= K_{00}\bar u + K_{01}\bar\psi + o_p(n^{-1/2}),\\ \widehat a &= K_{10}\bar u + K_{11}\bar\psi + o_p(n^{-1/2}).\end{aligned}$$

These linearisation approximations lead to $\widehat a \doteq K_{11}\bar v$ and to $\widehat\theta - \theta_0 \doteq J^{-1}\bar u - J^{-1}GK_{11}\bar v$, where the simplifying '$\doteq$' notation indicates here and below that the differences in question go to zero at the required speed. Also, by a Taylor approximation argument as with (6.6),

$$\bar\psi_j(\widehat\theta) = n^{-1}\sum_{i=1}^n \psi_j(X_i,\widehat\theta) = \bar\psi_j - g_j^{\rm t}(\widehat\theta - \theta_0) + o_p(n^{-1/2}),$$

which leads to $\bar\psi(\widehat\theta) \doteq \bar\psi - G^{\rm t}J^{-1}(\bar u - GK_{11}\bar v) = (I + G^{\rm t}J^{-1}GK_{11})\bar v$.

We may now approximate the first term of (6.7), also using $n\log c_S(\widehat a,\widehat\theta) = n\frac12\widehat a^{\rm t}\widehat a + o_p(1)$, which is seen to hold under (6.4) conditions. We find

$$2n\Big\{\sum_{j\in S}\widehat a_j\bar\psi_j(\widehat\theta) - \log c_S(\widehat a,\widehat\theta)\Big\} \doteq n\bar v^{\rm t}(2K_{11} + 2K_{11}G^{\rm t}J^{-1}GK_{11} - K_{11}^2)\bar v = n\bar v^{\rm t}K_{11}^2\bar v.$$

For the second term one finds

$$\begin{aligned} 2\sum_{i=1}^n \log\frac{f_0(X_i,\widehat\theta)}{f_0(X_i,\widetilde\theta)} &\doteq 2\sum_{i=1}^n \{u(X_i,\theta_0)^{\rm t}(\widehat\theta-\theta_0) - u(X_i,\theta_0)^{\rm t}(\widetilde\theta-\theta_0) \\ &\qquad + \tfrac12(\widehat\theta-\theta_0)^{\rm t} i(X_i,\theta_0)(\widehat\theta-\theta_0) - \tfrac12(\widetilde\theta-\theta_0)^{\rm t} i(X_i,\theta_0)(\widetilde\theta-\theta_0)\} \\ &\doteq 2n\{\bar u^{\rm t}(\widehat\theta-\theta_0) - \bar u^{\rm t}(\widetilde\theta-\theta_0) \\ &\qquad - \tfrac12(\widehat\theta-\theta_0)^{\rm t} J(\widehat\theta-\theta_0) + \tfrac12(\widetilde\theta-\theta_0)^{\rm t} J(\widetilde\theta-\theta_0)\}, \end{aligned}$$

writing $i(x,\theta)$ for the $p\times p$ second order derivative function of the log-density. In concert with previous approximations this leads to the second term being approximated with $-n\bar v^{\rm t}K_{11}G^{\rm t}J^{-1}GK_{11}\bar v$. Combining efforts,

$$Z_{n,2,S} \doteq n\bar v^{\rm t}(K_{11}^2 - K_{11}G^{\rm t}J^{-1}GK_{11})\bar v = n\bar v^{\rm t}K_{11}\bar v,$$

the required approximation. Some analysis, assisted by the Lindeberg theorem, shows that

$$n^{1/2}\bar v = n^{1/2}(\bar\psi - G^{\rm t}J^{-1}\bar u) \to_d b - G^{\rm t}J^{-1}\sum_{j\in S} b_j g_j + {\rm N}_{|S|}(0, I - G^{\rm t}J^{-1}G).$$

With the previously acquired approximation this implies the second part of the lemma. ■

For basis functions orthogonal to the score functions at θ_0, one has $G=0$, and, again, the test's limiting distribution equals that of (2.3).

A score-type approximation to the full likelihood-ratio statistic (6.7) is available, via the averages $\psi_j(\widetilde\theta)$ which rely only on the maximum likelihood estimates inside the $f_0(x,\theta)$ family. We saw in (6.6) that $\bar\psi_j(\widetilde\theta)$ becomes first order equivalent to $\bar\psi_j - g_j^{\rm t}J^{-1}\bar u$, in other words, $\bar\psi(\widetilde\theta)$ is only $o_p(n^{-1/2})$ away from $\bar\psi - G^{\rm t}J^{-1}\bar u = \bar v$. Thus

$$T_{n,2,S} = n\bar\psi(\widetilde\theta)^{\rm t}(I - \widetilde G^{\rm t}\widetilde J^{-1}\widetilde G)^{-1}\bar\psi(\widetilde\theta), \tag{6.8}$$

employing 'narrow' estimates $\widetilde G = G(\widetilde\theta)$ and $\widetilde J = J(\widetilde\theta)$, is a computationally simple approximation to $Z_{n,2,S}$, valid under local circumstances (6.4). One may also use the $\widehat\theta$ computed in the fuller model for the purpose of estimating G and J here, and yet other variations for these ingredients could involve jackknifing or bootstrapping. The (6.8) test should however not use $\widehat\theta$ in the ψ_j averages, since the limit distribution then would be different from the one given in the lemma.

The (6.8) statistic involves inversion of an $|S|\times|S|$ matrix, but the matrix identity

$$K_{11} = (I - G^{\rm t}J^{-1}G)^{-1} = I + G^{\rm t}(J - GG^{\rm t})^{-1}G,$$

easily proved under the condition that $J - GG^{\rm t}$ is positive definite, leads to a simpler equivalent form,

$$T_{n,2,S} = \sum_{j\in S} n\bar\psi_j(\widetilde\theta)^2 + n(\widetilde G\bar\psi(\widetilde\theta))^{\rm t}(\widetilde J - \widetilde G\widetilde G^{\rm t})^{-1}\widetilde G\bar\psi(\widetilde\theta), \tag{6.9}$$

where only a $p \times p$ matrix inversion is involved. Note also that $\sum_{j\in S} n\bar{\psi}(\widetilde{\theta})^2$ may serve as a simple conservative test statistic. For basis functions resulting in $G = 0$, $T_{n,1,S} = T_{n,2,S}$.

For a data-driven choice of S, the asymptotic null distribution may be obtained along the same lines as in Sections 2 and 3. Observe also that the Z and T tests, although asymptotically equivalent under (6.4) conditions, will differ significantly under the fixed alternative scenario, comparable to what we saw in Section 5.

7. Testing multivariate normality and other models

The apparatus developed in this article is very general, and can be applied to test the adequacy of any parametric family, subject to the usual conditions of regularity, also in higher dimensions. In this section the methodology is applied to construct explicit goodness of fit tests for some families.

7.1. Testing normality. We wish to test the hypothesis that data follow the normal density $\sigma^{-1}\phi(\sigma^{-1}(x-\mu))$, for suitable but unspecified (μ,σ), writing ϕ for the standard normal density. Let $\psi_1, \psi_2, \ldots$ be orthogonal and normalised w.r.t. ϕ, and consider encapsulating models of the type

$$f_S(x \,|\, a, \mu, \sigma) = \frac{1}{\sigma}\phi\Big(\frac{x-\mu}{\sigma}\Big) c_S(a)^{-1} \exp\Big\{\sum_{j\in S} a_j \psi_j\Big(\frac{x-\mu}{\sigma}\Big)\Big\}. \tag{7.1}$$

We focus first on the approach taken in 6.1, which is to keep $(\widetilde{\mu},\widetilde{\sigma})$ as the maximum likelihood estimators of location and scale under normality and then calculate $\widetilde{a}_j$ for $j \in S$ in the resulting $|S|$-parameter model. This is a practical and immediately interpretable solution, as one would see explicit corrections to the usual null model density estimate $\widetilde{\sigma}^{-1}\phi(\widetilde{\sigma}^{-1}(x-\widetilde{\mu}))$.

Note that the distribution of $Z_{n,1,S}$ and $T_{n,1}$ of (6.1) and (6.2) do not depend on (μ,σ), since they only feature $\bar{\psi}_j = n^{-1}\sum_{i=1}^n \psi_j(Y_i)$, where $Y_i = (X_i - \widetilde{\mu})/\widetilde{\sigma}$. It also means that the actual null distribution of $Z_{n,1,S}$, or for that matter also $Z_{n,1,S^*}$, where a precise algorithm has been decided on for selecting the index set S^*, can be found by simulation of standard normal data sets alone.

In the present model, the $2 \times |S|$ matrix G of Section 6 has elements $g_{1,j}/\sigma$ and $g_{2,j}/\sigma$, where $g_{1,j} = \int \phi(y) y \psi_j(y)\,\mathrm{d}y$ and $g_{2,j} = \int \phi(y) y^2 \psi_j(y)\,\mathrm{d}y$. Also, J is diagonal $(1/\sigma^2, 2/\sigma^2)$. The limiting null distribution for $Z_{n,1,S}$ and $T_{n,1,S}$ becomes that of

$$\sum_{j\in S}(N_j - g_{j,1}U_1 - \tfrac{1}{2}g_{2,j}U_2)^2,$$

where $\mathrm{cov}(N_j, U_1) = g_{1,j}$, $\mathrm{cov}(N_j, U_2) = g_{2,j}$, while $\mathrm{Var}\, U_1 = 1$ and $\mathrm{Var}\, U_2 = 2$. The mean of the limiting null distribution is $|S| - \sum_{j\in S}(g_{1,j}^2 + \frac{1}{2}g_{2,j}^2)$. The more general limit under local alternatives follows this pattern but with b_j parameters entering as in (6.6).

The second approach is as in Section 6.2, where $Z_{n,2,S}$ of (6.7) can be used, in addition to its simpler score test approximation

$$T_{n,2,S} = \sum_j n\bar\psi_j^2 + \begin{pmatrix} \sum_j g_{1,j}\bar\psi_j \\ \sum_j g_{2,j}\bar\psi_j \end{pmatrix}^{\rm t} \begin{pmatrix} 1-\sum_j g_{1,j}^2 & -\sum_j g_{1,j}g_{2,j} \\ -\sum_j g_{1,j}g_{2,j} & 2-\sum_j g_{2,j}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_j g_{1,j}\bar\psi_j \\ \sum_j g_{2,j}\bar\psi_j \end{pmatrix}.$$

This goes to a noncentral $\chi^2_{|S|}$ under local conditions. As explained in Section 6, there are certain advantages to working instead with a revised set of basis functions, which are orthogonal to the score function (y, y^2-1).

A simple technique for constructing orthonormal basis functions around a given f_0 is to let $\psi_j(x) = \gamma_j(F_0(x))$, where $1, \gamma_1, \gamma_2, \ldots$ are orthonormal with respect to the uniform distribution on $(0,1)$. In addition to $\int f_0\psi_j\psi_k = \delta_{j,k}$, one has $\int f_0 \exp(\sum_j a_j\psi_j)\,{\rm d}x = \int_0^1 \exp(\sum_j a_j\gamma_j)\,{\rm d}y$, which makes it easier to check the requirement of finiteness of the integral for a_js in a neighbourhood around zero. Choices for the γ_js include the normalised Legendre polynomials, employed for a similar purpose already in Neyman (1937), and the cosine functions $\sqrt{2}\cos(j\pi x)$.

REMARK. For the particular case of the normal there is also another attractive possibility, exploiting scaled and exponentially modified Hermite polynomials. Let $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = (x^2-1)/\sqrt{2!}$, $H_3(x) = (x^3-3x)/\sqrt{3!}$ and so on be the normalised Hermite polynomials. They may not be used as ψ_j functions in the present context in that $\int \phi\exp(a_jH_j)\,{\rm d}x$ too easily becomes infinite. But, for any $c>0$,

$$\int c\phi(cx)H_j(cx)H_k(cx)\,{\rm d}x = \int \phi(x)c\exp\{-\tfrac12(c^2-1)x^2\}H_j(cx)H_k(cx)\,{\rm d}x = \delta_{j,k},$$

which means that we are free to use $\psi_j(x) = c^{1/2}\exp\{-\frac14(c^2-1)x^2\}H_j(cx)$. For $c>1$ these functions are orthonormal w.r.t. ϕ and bounded, which means that $\int\phi\exp(\sum_j a_j\psi_j)\,{\rm d}x$ will be finite as long as $\sum|a_j|\max_x|\psi_j(x)|$ is finite. ■

7.2. The multivariate normal. Suppose we wish to test whether the data are coming from a d-variate normal distribution $f_0(x,\mu,\Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\{-\frac12(x-\mu)^{\rm t}\Sigma^{-1}(x-\mu)\}$, where Σ is a positive definite $d\times d$ matrix. Basis functions for multivariate models may easily be constructed as products of univariate basis functions. Alternatively, as above, we might take $\psi_j(x,\mu,\Sigma) = \gamma_j(\Phi(y_1),\ldots,\Phi(y_d))$ where $(y_1,\ldots,y_d)^{\rm t} = \Sigma^{-1/2}(x-\mu)$ while the γ_js for example may be products of cosine functions.

Limiting distributions of the likelihood ratio test for multivariate normality are given in Sections 6.1 and 6.2. For this model we arrive at a $\frac12(d^2+3d)\times|S|$ matrix G with (i,j)th element $g_{i,j} = |\Sigma|^{-1/2}\int\phi(y)y_i\psi_j(y)\,{\rm d}y$ for $i\le d$, and

$$g_{i,j} = \tfrac12\int y^{\rm t}\Sigma^{-1/2}E_i\Sigma^{-1/2}y\phi(y)\psi_j(y)\,{\rm d}y$$

for $i>d$, where E_i denotes a matrix of zeros, except for the (r,s) and (s,r)th elements, which equal one. Here (r,s) refers to the row and column indices in the original matrix Σ

of the $(i-d)$th element of vech(Σ). The orthogonality of mean and variance components results in a block diagonal Fisher information matrix $J = \mathrm{diag}(J_\mu, J_\Sigma)$, where $J_\mu = \Sigma^{-1}$ and $(J_\Sigma)_{i,j} = \frac{1}{2}\mathrm{tr}(\Sigma^{-1}E_i\Sigma^{-1}E_j)$. Let $g_{\mu,j}$ and $g_{\Sigma,j}$ be the corresponding subvectors of g_j. The null distribution of $Z_{n,1,S}$, for example, reduces to that of

$$\sum_{j\in S}\big(N_j - g_{\mu,j}^{\mathrm{t}}J_\mu^{-1}U_{0,\mu} - g_{\Sigma,j}^{\mathrm{t}}J_\Sigma^{-1}U_{0,\Sigma}\big)^2,$$

where $(U_{0,\mu}^{\mathrm{t}}, U_{0,\Sigma}^{\mathrm{t}})^{\mathrm{t}}$ is a mean zero normally distributed random vector with variance J and $\mathrm{cov}(U_{0,\mu}, N) = G_\mu$, $\mathrm{cov}(U_{0,\Sigma}, N) = G_\Sigma$. Also, N is an $|S|$-dimensional standard normal random variable with components N_j, and the $d \times |S|$ matrix G_μ consists of the first d rows of G, the elements of which are explicitly given above. The remaining part of G is G_Σ. The score test $T_{n,1,S}$ is simply as given in (6.3). The likelihood ratio statistic $Z_{n,2,S^*}$ can be readily computed, once a set S^* is decided upon. Its score version $T_{n,2,S^*}$, defined in (6.8), is calculated as

$$T_{n,2,S^*} = n\bar\psi(\widetilde\mu,\widetilde\Sigma)^{\mathrm{t}}(I - \widetilde G_\mu^{\mathrm{t}}J_\mu^{-1}\widetilde G_\mu - \widetilde G_\Sigma^{\mathrm{t}}J_\Sigma^{-1}\widetilde G_\Sigma)^{-1}\bar\psi(\widetilde\mu,\widetilde\Sigma).$$

In both matrices $\widetilde G_\mu$ and $\widetilde G_\Sigma$, the variance matrix Σ is estimated using the multivariate normal density $f_0(x,\mu,\Sigma)$.

In multiple dimensions, for the nested sequence type of tests, the order in which the terms enter the sequence becomes even more important. Taking all subsets up to a finite, pre-specified number m_0 is still possible, but this might lead to a very large number if a reasonably large number in each direction is wanted. A compromise strategy between the all subsets selection and a nested sequence, as already noted at the end of Section 3, might be particularly advantageous. Still many other options exist. As in Aerts, Claeskens and Hart (2000), one could construct a nested sequence of models, by adding not one, but several components at a time. This slightly changes the asymptotic distribution results.

7.3. General location and scale families. Only minor changes to the results of Sections 7.1 and 7.2 apply for location-scale families more general than the normal densities, say of the form $f_0(\Sigma^{-1/2}(x-\mu))|\Sigma|^{-1/2}$. Formulae for J and G change accordingly, of course.

7.4. Testing a small smooth family. Let m be fixed and perhaps small, and let $\psi_1,\ldots,\psi_m$ be orthonormal w.r.t. some density f_0. The family of densities $f_a(x) = f_0(x)c_m(a)^{-1}\exp\{\sum_{j=1}^m a_j\psi_j(x)\}$ is an attractive model of the exponential type. Our machinery is now available to test whether data are adequately modelled in this way. Fill in more orthonormal basis functions ψ_j for $j > m$. Test statistics $Z_{n,1,S}$ and $T_{n,1,S}$ of Section 6.1, where S is a subset of $\{m+1,\ldots,\}$, are computed via likelihood estimates $\widetilde a$ in the smaller model, and have the simple limit distribution $\sum_{j\in S}(b_j+N_j)^2$ under local alternatives (2.1). This follows from (6.6) in that the G matrix in question is equal to zero. The same happens with test statistics $Z_{n,2,S}$ and $T_{n,2,S}$ defined in Section 6.2, and in fact $T_{n,2,S}$ is identical to $T_{n,1,S} = \sum_{j\in S} n\bar\psi(\widetilde a)^2$.

7.5. Results of a simulation study. We will test for bivariate normality comparing various versions of the score statistic: with order chosen by the classical AIC regime, $C = 2$, by BIC, and via the order selection principle. Critical values at the 5% level are obtained via a simulation of size 30,000 under the standard bivariate normal distribution. Legendre polynomials are orthogonalised with respect to the score vector. Not only this simplifies the test statistic, it also implies that there is no point in including basis functions of order two or less. Bogdan (1999) already hints about excluding some of the lowest order terms. Our model sequence starts with adding cubic terms, followed by interactions, up to a total of 14 additional terms: $x_1^3, x_2^3, x_1^2x_2, x_1x_2^2, x_1^4, x_2^4, \ldots$ Of course, numerous different variations could have been chosen. For comparison reasons, the simulation settings are taken from Bogdan (1999), where she compares a large number of tests. Our test statistic $T^*(\text{bic})$ differs from her $W_{S(5)}$ in that we do not need to include the interaction term x_1x_2, and we start penalising the smallest model with constant 1, instead of 5. In setting (a), data are generated from a $3:1$ mixture of a standard normal and two $\mathrm{N}(3,3)$ random variables with covariance 2.7. Setting (b) chooses two independent $\mathrm{Beta}(2.5, 1.5)$ random variables, and in (c) the alternative consists of two independent uniform random variables on the $(0,1)$ interval. Results are shown in Table 7.1. As in previous cases there is no clear winner among the tests studied, but simulated powers exceed those of classical tests applied in the same situations; see Bogdan (1999).

		$T^*(C=2)$		$T^*(\text{bic})$		$\widetilde{T}$	
		nested	all sets	nested	all sets	nested	all sets
(a)	n=25	88.56	97.62	96.40	97.24	91.36	95.92
	n=50	63.46	82.14	83.90	82.88	63.94	76.16
(b)	n=25	24.48	30.56	31.66	29.62	33.82	25.40
	n=50	22.88	24.00	28.46	26.26	23.04	19.10
(c)	n=25	32.00	32.14	28.16	32.06	21.82	30.50
	n=50	27.10	25.96	28.16	27.52	17.56	20.78

TABLE 7.1. *Simulated power results (as %) for bivariate normality tests. Order selected via a nested model search, or all subset selection, employing AIC with $C = 2$, BIC, or the order selection test $\widetilde{T}$, using Legendre polynomials.*

8. Concluding comments

8.1. Chi squared tests revisited. Our general strategy for testing $f = f_0$ has been to use $Z_n = 2\sum_{i=1}^n \log\{\widehat{f}(X_i)/f_0(X_i)\}$, for different choices of $\widehat{f}$. Consider the histogram estimator based on cells $C_1, \ldots, C_m$, which uses $N_j/(nh_j)$ to estimate f in a cell C_j, with N_j the number of points caught in C_j. It is comforting to see that our general apparatus then leads to statistics which are close approximations to well-known statistics.

To verify this, express Z_n as $2\sum_{j=1}^m N_j\{\log(\widehat{p}_j/h_j) - \log \bar{f}_{0,j}\}$, where $\widehat{p}_j = N_j/n$ and $\bar{f}_{0,j}$ is the geometrical mean of the $f_0(X_i)$ for which $X_i \in C_j$. When this is approximated with $p_{0,j}/h_j$, where $p_{0,j} = \int_{C_j} f_0 \,\mathrm{d}x$, one finds that Z_n is close to $Z_n' =$

$2\sum_{j=1}^m N_j \log(\widehat{p}_j/p_{0,j})$, the log-likelihood ratio statistic for testing whether the vector of p_js is equal to that of $p_{0,j}$s. As is well known, both Z_n' and its further approximation $Z_n'' = \sum_{j=1}^m n(\widehat{p}_j - p_{0,j})^2/p_{0,j}$ are asymptotically χ^2_{m-1} distributed under the null hypothesis.

8.2. Matching the performance of Cramér–von Mises tests. Under local alternative conditions (2.1) and with appropriate growth conditions on m we have seen that the vector of $n^{1/2}\bar{\psi}_j$s goes to that of $b_j + N_j$ (with notation as in Lemma 1). In generalisation of the plain score test (1.5), consider $U_n = \sum_{j=1}^m \lambda_{n,j} n\bar{\psi}_j^2$ for suitable constants $\lambda_{n,j}$. If these are chosen so as to converge to a sequence of λ_js with finite sum, then $U_n \to_d U = \sum_{j=1}^\infty \lambda_j (b_j + N_j)^2$ under mild conditions; see e.g. the strong approximation result (3.6) above. This limit is precisely of the form reached for the Cramér–von Mises statistic, and also for several related tests; see e.g. Shorack and Wellner (1986) and Hall (1985). Hence any of these will have a competitor of our type U_n which will match it precisely in large-sample performance.

8.3. Goodness of fit versus an infinite-dimensional normal testing problem. Consider a statistical experiment where a full sequence of independent normal variables $V_j \sim \mathrm{N}(b_j, 1)$ is observed, and for which it is only known that $\sum_{j=1}^\infty |b_j|$ is finite. Assume that it is required to test the hypothesis that every $b_j = 0$ versus the alternative that at least one of them is non-zero. Our article has demonstrated in various ways that the general large-sample goodness-of-fit problem, with a nonparametric alternative to the null hypothesis, must have precisely this structure. There might be precise formulations of this equivalence statement in the tradition of comparison of experiments, e.g. in the style of Nussbaum's (1996) result comparing density estimation with Gaußian white noise problems.

This asymptotic equivalence also invites performance comparisons between different tests to be made directly in the limit experiment. This represents a significant simplification. We have seen in Sections 3 and 4 that two rather different schemes, related to respectively AIC and BIC, become equivalent to the test $\max_{j\le m_0} |V_j|$, with power function $1 - \prod_{j=1}^{m_0} \Gamma_1(C_0, b_j^2)$, with $\Gamma_1(C_0, 0)^{m_0} = 1 - \alpha$ in terms of the wished for significance level α. The versions which use BIC with nested submodels have been seen in Section 4.2 to be equivalent to the potentially very weak test $|V_1|$. The BIG-related schemes of Section 4.3 would correspond to tests of the form $\sum_{j\in B_m} V_j^2$, where B_m is the subset of $\{1, \ldots, m_0\}$ with the m biggest values of $|V_j|$, and so on. Each subset selector corresponds to a well-defined test rule in the $(V_1, V_2, \ldots)$ experiment, and power functions can be computed and compared by simulation. For example, the nested AIC regime corresponds to the rule $\sum_{m=0}^\infty J_m(V_1^2 + \cdots + V_m^2)$, where J_m is indicator for $Z_m - Cm$ being bigger than all other $Z_{m'} - Cm'$, and $Z_m = \sum_{j=1}^m V_j^2$.

8.4. Which basis functions? The apparatus we have developed works of course for any sequence of orthogonal basis functions $\psi_1, \psi_2, \ldots$. These might also be re-ordered, though there is often a canonical way of listing them. In addition to the cosine and

Legendre functions, used in our simulation studies, one might use splines with equally spaced knots; see comments in Barron and Sheu (1991). Regarding the practical question of which sequence to use, there can be no universal dominance result; tests using the cosine functions will be stronger than those using the Legendre functions for one set of fs and weaker for the complementary set. For envisaged alternatives, one may compute the determining ratios α_Z/τ_Z, see Section 5.1, for each of the basis systems. One may also actually estimate this ratio, via a nonparametric density estimate, for each basis system considered, and then in the end use the system which has the biggest ratio estimate.

8.5. Log-linear expansion density estimators. As a side product of our models and methods, one may put forward the log-linear expansions as bona fide density estimators, worthy of further separate study. For example,

$$\widehat{f}(x) = \frac{1}{\widetilde{\sigma}}\phi\Big(\frac{x-\widetilde{\mu}}{\widetilde{\sigma}}\Big)c_{S^*}(\widetilde{a})^{-1}\exp\Big\{\sum_{j\in S^*}\widetilde{a}_j\psi_j\Big(\frac{x-\widetilde{\mu}}{\widetilde{\sigma}}\Big)\Big\},$$

in the notation of Section 7.1, and perhaps with S^* decided upon by AIC, would have motivation and ambition similar to that of the multiplicative estimators developed in Hjort and Glad (1995).

8.6. Mixing over candidate models. It is clear that our framework and methodology allow quite general subset selection regimes when choosing the set $S = S^*$ for use in $Z_{n,S}$ or $T_{n,S}$, not only those selected via the AIC or the BIC. This may actually be generalised further, to form sensible test averages of the type say $\sum_{j=1}^k w_{n,j}Z_{n,S_j}$, over candidate subsets $S_1,\ldots,S_k$, with weights $w_{n,1},\ldots,w_{n,k}$ dictated by the data. Theory developed in Hjort and Claeskens (2003) and Claeskens and Hjort (2003) make this possible. Among possible weight schemes are the 'smoothed AIC' weights discussed in these papers.

Appendix

Here we give proofs of Lemmas 2 and 6. As a preamble to these, let f be a local alternative density of the form (2.1), and assume that $L = \sum_{j=1}^\infty |b_j|\,\|\psi_j\|$ is finite. Then

$$\begin{aligned}
\mathrm{E}\psi_j(X) &= \frac{b_j/n^{1/2}+O(L^2/n)}{1+O(L^2/n)} = b_j/n^{1/2}+O(1/n),\\
\mathrm{cov}\{\psi_j(X),\psi_k(X)\} &= \frac{\delta_{j,k}+O(L/n^{1/2})}{1+O(L^2/n)} - b_jb_k/n + O(1/n^{3/2}) = \delta_{j,k}+O(1/n^{1/2}).
\end{aligned} \tag{A.1}$$

These are reached working with the integrals involved, one ingredient being that $\sum_{j=1}^\infty b_j^2$ is finite. This comes from $\sum_{j=1}^m b_j^2 = \int f_0(\sum_{j=1}^m b_j\psi_j)^2\,\mathrm{d}x \le L^2$ for each m, which shows that the denominator $c(b/n^{1/2})$ involved is $1+O(L^2/n)$ with $O(L^2/n)$ term smaller in size than the $O(1/n^{1/2})$ terms involved in the numerators.

Proof of Lemma 2. In Lemma 1 a crucial ingredient was the $\frac12\sum_S a_j^2 = \frac12\|a\|_S^2$ approximation to $\log c_S(a)$. Presently we need a somewhat more careful assessment of this

approximation. Start out writing $\log(1+z) = z + \varepsilon(z)$, where an easily derived bound, sufficient for our purposes, is $|\varepsilon(z)| \le z^2$ for $|z| \le \frac12$. Thus

$$\log c_S(a) = \log\Big\{1 + \tfrac12 \sum_{j\in S} a_j^2 + d_S(a)\Big\} = \tfrac12 \|a\|_S^2 + e_S(a),$$

where $e_S(a) = d_S(a) + \varepsilon\big(\frac12\|a\|_S^2 + d_S(a)\big)$ and

$$d_S(a) = \int f_0 \Big\{\exp\Big(\sum_{j\in S} a_j\psi_j\Big) - 1 - \Big(\sum_{j\in S} a_j\psi_j\Big) - \tfrac12\Big(\sum_{j\in S} a_j\psi_j\Big)^2\Big\}\,\mathrm{d}x.$$

Noticing that $|\exp(y) - (1 + y + \frac12 y^2)| \le \frac16 |y|^3 \exp(|y|)$, we derive

$$\begin{aligned} |d_S(a)| &\le \tfrac16 \int f_0 \Big|\sum_{j\in S} a_j\psi_j\Big|^3 \exp\Big(\Big|\sum_{j\in S} a_j\psi_j\Big|\Big)\,\mathrm{d}x \\ &\le \tfrac16 \sum_{j\in S} |a_j| M_m \int f_0 \Big|\sum_{j\in S} a_j\psi_j\Big|^2 \exp\Big(\sum_{j\in S} M_m|a_j|\Big)\,\mathrm{d}x \qquad (A.2) \\ &= \tfrac16 M_m \sum_{j\in S} |a_j| \sum_{j\in S} a_j^2 \exp\Big(M_m \sum_{j\in S} |a_j|\Big). \end{aligned}$$

This will help us pass two separate technical hurdles below.

First reconsider the concave function K_n used in the proof of Lemma 1, which was expressed as a simpler concave function $K_{n,0}$ plus a remainder term $r_n = r_{n,S}$, for which we now need a more precise bound. One finds in fact that $K_n(u) = K_{n,0}(u) - n e_S(u/n^{1/2})$. For reasons becoming apparent below we need $r_{n,S}(n^{1/2}\bar\psi)$ to go to zero in probability, which by the above translates into demonstrating that

$$n d_S(\bar\psi) + n\varepsilon\big(\tfrac12\|\bar\psi\|_S^2 + d_S(\bar\psi)\big) \to_p 0. \qquad (A.3)$$

Write $n^{1/2}\bar\psi_j = b_j + N_{n,j}$. It follows from (A.1) that the $N_{n,j}$s have means of size $O(L^2/n^{1/2})$, covariances of size $O(L/n^{1/2})$ going to zero, and variances of the type $1 + O(L/n^{1/2})$ and therefore going to 1. Hence $R_{m,2} = \sum_{j\in S} |b_j + N_{n,j}|^2$ is such that $R_{m,2}/|S|$ has mean bounded in $|S|$, implying $R_{m,2} = O_p(|S|)$. Similarly, $R_{m,1} = \sum_{j\in S} |b_j + N_{n,j}|$ has $\mathrm{E}(R_{m,1}/|S|)$ bounded in $|S|$, so that $R_{m,1} = O_p(|S|)$ too. For any $S \in \{1,\ldots,m\}$ this gives

$$|d_S(\bar\psi)| \le \tfrac16 \frac{M_m}{n^{3/2}} \sum_{j\in S} |n^{1/2}\bar\psi| \sum_{j\in S} n\bar\psi_j^2 \exp\Big\{\frac{M_m}{n^{1/2}} \sum_{j\in S} |n^{1/2}\bar\psi_j|\Big\} = O_p\Big(\frac{M_m m^2}{n^{3/2}} \exp\Big(\frac{M_m m}{n^{1/2}}\Big)\Big).$$

This takes care of the first term of (A.3), by the stipulated growth condition. As for the second term, $\frac12\|\bar\psi\|_S^2$ is of order $O_p(m/n)$ and $d_S(\bar\psi)$ of order $O_p(M_m m^2/n^{3/2})$, making their sum less than $\frac12$ with probability going to 1, which means that the $|\varepsilon(z)| \le |z|^2$

inequality applies. The second term is therefore seen to be dominated by a variable of order $O_p(m^2/n + M_m^2 m^4/n^2)$, which also goes to zero.

We are then in a position to accurately approximate $Z_{n,S}$ with $T_{n,S}$. Write $\widehat{a}_j = \bar{\psi}_j + \varepsilon_{n,j}/n^{1/2}$ for $j \in S$. It was a consequence of the proof of Lemma 1 that the $\varepsilon_{n,j} \to_p 0$ in case of a fixed m, but now the horizon is becoming broader with n. An application of the nearness-of-argmax lemma of Hjort and Pollard (1994) yields

$$\Pr\Big\{\sum_{j\in S} \varepsilon_{n,j}^2 \geq \delta^2\Big\} \leq \Pr\{\Delta_n(\delta) \geq \tfrac{1}{2}\delta^2\},$$

in which $\Delta_n(\delta) = \max_{\|v\|\leq\delta} |r_n(n^{1/2}\bar{\psi} + v)|$. But by a slight extension of the arguments above this variable goes to zero in probability. Combining this with

$$Z_{n,S} = 2n\Big\{\sum_{j\in S} \widehat{a}_j\bar{\psi}_j - \tfrac{1}{2}\sum_{j\in S}\widehat{a}_j^2 - e_S(\widehat{a})\Big\} = \sum_{j\in S} n\bar{\psi}_j^2 - \sum_{j\in S}\varepsilon_{n,j}^2 - 2ne_S(\widehat{a}),$$

it remains only to show that the last term goes to zero in probability. By arguments used above this is the same as showing

$$2nd_S(\widehat{a}) + 2n\varepsilon\Big(\tfrac{1}{2}\sum_{j\in S}\widehat{a}_j^2 + d_S(\widehat{a})\Big) \to_p 0.$$

A bound on the first term is found to be

$$\tfrac{1}{3}(M_m/n^{1/2})\sum_{j\in S}|n^{1/2}\widehat{a}_j|\sum_{j\in S}|n^{1/2}\widehat{a}_j|^2\exp\Big\{(M_m/n^{1/2})\sum_{j\in S}|n^{1/2}\widehat{a}_j|\Big\},$$

and this goes to zero in probability for precisely the same reasons as above. The second term can also be handled by arguments parallelling those used in connection with the second term of (A.3), again exploiting the fact that $\sum_{j\in S}\varepsilon_{n,j}^2 \to_p 0$. ■

Proof of Lemma 6. The plan is to show that (i) $(Z_n^* - T_n^*)/m^{1/2} \to_p 0$ under the $M_m m^{9/4}/n^{1/2} \to 0$ condition, and then demonstrating (ii) that T_n^* can be approximated with a non-central χ^2 well enough to imply that $(T_n^* - m - m^{1/2}B_m)/(2m + 4m^{1/2}B_m)^{1/2}$ tends to the standard normal, under the $M_m m^{10/3}/n^{1/2} \to 0$ condition.

To the first end, write $\varepsilon_{n,j} = n^{1/2}(\widehat{a}_j - \bar{\psi}_j)$ for $j \leq m$ and note from the proof of Lemma 2 that

$$Z_n^* = T_n^* - \sum_{j=1}^m \varepsilon_{n,j}^2 - 2ne_m(\widehat{a}),$$

where we write e_m and d_m for the e_S and d_S of the proof of Lemma 2 corresponding to the full set $S = \{1,\ldots,m\}$. It suffices for the first part of the proof to show that $\|\varepsilon_n\|^2/m^{1/2} \to_p 0$ and that $ne_m(\widehat{a})/m^{1/2} \to_p 0$.

With a little work, the nearness-of-argmax lemma of Hjort and Pollard (1994) gives

$$\Pr\Big\{\sum_{j=1}^m \varepsilon_{n,j}^2/m^{1/2} \geq \delta^2\Big\} \leq \Pr\{\Delta_n(m^{1/4}\delta) \geq \tfrac{1}{2}m^{1/2}\delta^2\},$$

where

$$\Delta_n(m^{1/4}\delta) = \max_{\|v\| \le m^{1/4}\delta} |r_n(n^{1/2}\bar\psi + v)|.$$

We must show that $\Delta_n(m^{1/4}\delta)/m^{1/2} \to_p 0$. Using $r_n(u) = -ne_m(u/n^{1/2})$ this translates into showing

$$\{nd_m(\bar\psi + v/n^{1/2}) + n\varepsilon(\tfrac12\|\bar\psi + v/n^{1/2}\|^2 + d_m(\bar\psi + v/n^{1/2}))\}/m^{1/2} \to_p 0, \qquad (A.4)$$

uniformly over $\|v\| \le m^{1/4}\delta$. This can be worked with using appropriate careful extensions of arguments used to prove Lemma 2. Analysis parallelling the one that lead to approximations (A.1) shows that if we write $n^{1/2}\bar\psi_j = b_j m^{1/4} + N_{n,j}$, then $N_{n,j}$ has mean $O(L^2 m^{1/2}/n^{1/2})$ and variance $1+O(Lm^{1/4}/n^{1/2})$ while the covariances are $O(Lm^{1/4}/n^{1/2})$. These facts imply

$$\sum_{j=1}^m n^{1/2}|\bar\psi_j + v_j/n^{1/2}| = \sum_{j=1}^m |b_j m^{1/4} + N_{n,j} + v_j| = O_p(m^{5/4}),$$
$$\sum_{j=1}^m n|\bar\psi_j + v_j/n^{1/2}|^2 = \sum_{j=1}^m |b_j m^{1/4} + N_{n,j} + v_j|^2 = O_p(m^{6/4}),$$

for all v of length bounded by $m^{1/4}\delta$. Using the (A.2) bound,

$$|nd_m(\bar\psi + v/n^{1/2})|/m^{1/2} = O_p(M_m m^{5/4} m^{6/4}/n^{1/2})/m^{1/2} = O_p(M_m m^{9/4}/n^{1/2}),$$

which goes to zero by the stipulated condition. The second term of (A.4) can be dealt with similarly, and is in fact smaller in size than the first one.

To show $ne_m(\widehat a)/m^{1/2} \to_p 0$ it suffices by arguments used to prove Lemma 2 to demonstrate

$$nd_m(\widehat a)/m^{1/2} + n\varepsilon(\tfrac12\|\widehat a\|^2 + d_m(\widehat a))/m^{1/2} \to_p 0.$$

This is quite similar to the above. One may show that $\sum_{j=1}^m n^{1/2}|\widehat a_j| = O_p(m^{5/4})$ and $\sum_{j=1}^m n\widehat a_j^2 = O_p(m^{6/4})$, and via (A.2) the crucial condition for convergence to zero is again $M_m m^{9/4}/n^{1/2} \to 0$.

We have therefore confirmed that $Z_n^* = T_n^* + \eta_n$ with η_n small enough, and it remains to show that T_n^* has the required limit distribution. For this second part of the proof, let ξ_n and Σ_n be the mean vector and variance matrix of the m-vector $\psi(X_i) = (\psi_1(X_i), \ldots, \psi_m(X_i))^{\rm t}$, and consider i.i.d. vectors $Y_i = \Sigma_n^{-1/2}(\psi(X_i) - \xi_n)$. By efforts above, $\Sigma_n = I + O(m^{1/4}/n^{1/2})$ and $n^{1/2}\xi_{n,j} = b_j m^{1/4} + O(m^{1/2}/n^{1/2})$. We find from this that

$$\begin{aligned}\|Y_i\|^2 &= (\psi(X_i) - \xi_n)^{\rm t}\Sigma_n^{-1}(\psi(X_i) - \xi_n)\\ &= \{1 + O(m^{1/4}/n^{1/2})\}\sum_{j=1}^m \{\psi_j(X_i) - \xi_{n,j}\}^2 = O(mM_m^2),\end{aligned}$$

which implies

$$\mathrm{E}\|Y_i\|^3 = O(m^{1/2}M_m)\,\mathrm{E}\|Y_i\|^2 = O(m^{3/2}M_m).$$

Result (1.5) of Götze (1991) for approximating the distribution of the normalised sum $n^{-1/2}\sum_{i=1}^n Y_i = \Sigma_n^{-1/2}n^{1/2}(\bar{\psi} - \xi_n)$ with that of $N = (N_1, \ldots, N_m)^{\mathrm{t}}$, where these are independent standard normals, implies

$$\Pr\{\Sigma_n^{-1/2}n^{1/2}(\bar{\psi} - \xi_n) \in B\} = \Pr\{N \in B\} + \rho_{n,1}(B), \tag{A.5}$$

where $|\rho_{n,1}(B)| = O(m/n^{1/2})$ for all measurable convex sets B, provided that $m \geq 6$. This leads to

$$\Pr\{n^{1/2}\bar{\psi} \in n^{1/2}\xi_n + B\} = \Pr\{N \in \Sigma_n^{-1/2}B\} + \rho_{n,2}(B) = \Pr\{N \in B\} + \rho_{n,2}(B) + \rho_{n,3}(B),$$

where $\rho_{n,2}(B) = \rho_{n,1}(\Sigma_n^{-1/2}B)$ and $|\rho_{n,3}(B)| \leq aLm^{5/4}/n^{1/2}$, for a finite constant a, in that the elements of $\Sigma_n^{-1/2}$ are at most a finite constant times $Lm^{1/4}/n^{1/2}$ away from those of the $m \times m$ identity matrix. This further leads to $\Pr\{n^{1/2}\bar{\psi} \in C\} = \Pr\{N + n^{1/2}\xi \in C\} + \rho_{n,4}(C)$, where $\rho_{n,4}(C) = O(m^{5/4}/n^{1/2})$ for all C. ■

References

Aerts, M., Claeskens, G. and Hart, J.D. (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association* **94**, 869–879.

Aerts, M., Claeskens, G. and Hart, J.D. (2000). Testing lack of fit in multiple regression. *Biometrika* **87**, 405–424.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**, 1100–1120.

Anderson, N.H., Hall, P. and Titterington, D.M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis* **50**, 41–54.

Barron, A.R. and Sheu, C. (1991). Approximations of density functions by sequences of exponential families. *Annals of Statistics* **19**, 1347–1369.

Bhattacharya, R.N. and Ranga Rao, R. (1976). Normal approximation and asymptotic expansions. Wiley, New York.

Bickel and Rosenblatt (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics* **1**, 1071–1095. Corrigendum, ibid. **3**, 1370.

Bogdan, M. (1999). Data driven smooth tests for bivariate normality. *Journal of Multivariate Analysis* **68**, 26–53.

Bowman, A.W. (1992). Density based tests for goodness-of-fit. *Journal of Statistical Computation and Simulation* **40**, 1–13.

Bowman, A.W. and Foster, P.J. (1993). Adaptive smoothing and density-based tests of multivariate normality. *Journal of the American Statistical Association* **88**, 529–537.

Claeskens, G. and Hjort, N.L. (2003). The focussed information criterion [with discussion]. *Journal of the American Statistical Association* [to appear].

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

Eubank, R.L. (2000). Testing for no effects by cosine series methods. *Scandinavian Journal of Statistics* **27**, 747–763.

Eubank, R.L. and LaRiccia, V.N. (1992). Asymptotic comparison of Cramér–von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Annals of Statistics* **20**, 2071–2086.

Eubank, R.L. and Hart, J.D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics* **20**, 1412–1425.

Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Annals of Probability* **19**, 724–739.

Hall, P. (1984). Central limit theorem for integrated squared error for multivariate nonparametric density estimator. *Journal of Multivariate Analysis* **50**, 41–54.

Hall, P. (1985). Tailor-made tests of goodness of fit. *Journal of the Royal Statistical Society, Series B* **47**, 125–131.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* Springer-Verlag, New York.

Hjort, N.L. and Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *Annals of Statistics* **23**, 882–904.

Hjort, N.L. and Pollard, D. (1994). Asymptotics for minimisers of convex processes. Statistical Research Report, University of Oslo.

Hjort, N.L. and Claeskens, G. (2003). Frequentist model average estimators [with discussion]. *Journal of the American Statistical Association* [to appear].

Inglot, T., Kallenberg, W.C.M. and Ledwina, T. (1997). Data driven smooth tests for composite hypotheses. *Annals of Statistics* **25**, 1222–1250.

Inglot, T. and Ledwina, T. (1996). Asymptotic optimality of data-driven Neyman's tests for uniformity. *Annals of Statistics* **24**, 1982–2019.

Khmaladze, E.V. (1979). The use of ω^2 tests for testing parametric hypotheses. *Theory of Probability and its Applications* **24**, 283–301.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association* **89**, 1000–1005.

Lehmann, E. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

Neuhaus, G. (1987). Local asymptotics for linear rank statistics with estimated score functions. *Annals of Statistics* **15**, 491–512.

Neuhaus, G. (1988). Addendum to: 'Local asymptotics for linear rank statistics with estimated score functions'. *Annals of Statistics* **16**, 1342–1343.

Neyman, J. (1937). 'Smooth' test for goodness of fit. *Skandinavisk aktuarietidskrift* **20**, 149–199.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Annals of Statistics* **24**, 2399–2430.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society* **B 49**, 223–239.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes With Applicatons to Statistics*. Wiley, New York.

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

Woodroofe, M. (1982). On model selection and the arc-sine laws. *Annals of Statistics* **10**, 1182–1194.