© Grace Scientific Publishing

Journal of
Statistical Theory and Practice
Volume 3, No. 3, September 2009

Informative Statistical Analyses Using Smooth Goodness of Fit Tests

O. Thas, *Department of Applied Mathematics, Biometrics and Process Control, Ghent University, 9000 Gent, Belgium. Email: olivier.thas@UGent.be*

J. C. W. Rayner, *School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia. Email: John.Rayner@newcastle.edu.au*

D. J. Best, *School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia. Email: John.Best@newcastle.edu.au*

B. De Boeck, *Department of Applied Mathematics, Biometrics and Process Control, Ghent University, 9000 Gent, Belgium. Email: Bert.DeBoeck@UGent.be*

Received: February 19, 2008 *Revised:* September 21, 2008

Abstract

We propose a methodology for informative goodness of fit testing that combines the merits of both hypothesis testing and nonparametric density estimation. In particular, we construct a data-driven smooth test that selects the model using a weighted integrated squared error (WISE) loss function. When the null hypothesis is rejected, we suggest plotting the estimate of the selected model. This estimate is optimal in the sense that it minimises the WISE loss function. This procedure may be particularly helpful when the components of the smooth test are not diagnostic for detecting moment deviations. Although this approach relies mostly on existing theory of (generalised) smooth tests and nonparametric density estimation, there are a few issues that need to be resolved so as to make the procedure applicable to a large class of distributions. In particular, we will need an estimator of the variance of the smooth test components that is consistent in a large class of distributions for which the nuisance parameters are estimated by method of moments. This estimator may also be used to construct diagnostic component tests.

The properties of the new variance estimator, the new diagnostic components and the proposed informative testing procedure are evaluated in several simulation studies. We demonstrate the new methods on testing for the logistic and extreme value distributions.

AMS Subject Classification: 62G07; 62G10.

Key-words: Generalised score test; Method of moment estimator; Nonparametric density estimation; Orthonormal polynomials.

* 1559-8608/09-3/$5 + $1pp – see inside front cover

1. Introduction

Smooth tests of goodness of fit have been proven very powerful for testing for a large collection of distributions. See for example Rayner and Best (1989) or Rayner, Thas and Best (2009b) for an overview. For many distributions the test statistic decomposes into squared components which are related to deviations between the sample moments and the moments of the hypothesised distribution. This allows an informative analysis in the sense that, at the rejection of the null hypothesis, the components give an indication of what moments are not in agreement with the hypothesised. Although this diagnostic property is supported by simulation studies for many common distributions (Rayner and Best, 1989; Rayner, Best and Mathews, 1995), there is no theoretical ground for this. By properly rescaling the components, however, it has been shown theoretically that the diagnostic property can be regained, at least asymptotically, but large sample sizes are needed in practice (Henze and Klar, 1996; Henze, 1997; Klar, 2000). Although the theory of Henze and Klar is fairly general, they particularly focus on component tests for distributions for which the maximum likelihood (MLE) and the method of moment estimators (MME) of the nuisance parameters are equal. (However, see the companion paper of Rayner, Best and Thas, (2009a).)

In this paper we present two contributions. First, we explore the diagnostic component tests for distributions for which MLE and the MME do not coincide. The core of the theory consists of a variance estimator that is consistent in a wide class of distributions when the nuisance parameters are estimated by means of MME. This particular variance estimator has not been studied by Henze and Klar.

In the second part of the paper we exploit the relation between data-driven smooth tests and nonparametric density estimation, resulting in a data-driven testing approach, that, at the rejection of the null hypothesis, results in a nonparametric density estimate that may be used to visually assess the sense in which the true distribution deviates from the hypothesised. The latter procedure may be very helpful in settings where no diagnostic component tests can be used.

This paper is organised as follows. In the next subsections of the introduction, more details on generalised smooth tests, diagnostic components, data-driven tests and nonparametric density estimation are provided. In Section 2 a variance estimator is introduced. The new informative data-driven procedure is the topic of Section 3, and this is illustrated on an example data set in Section 4. All simulation studies and examples involve testing for the logistic or the extreme value distribution. More details on these distributions are given in the Appendices A and B, respectively.

1.1. The full parametric null hypothesis

The one-sample goodness of fit problem is, perhaps, one of the oldest statistical problems. It tests the null hypothesis that the observations come from a hypothesised distribution. Let g and f denote the true and the hypothesised density functions, respectively. The latter is often indexed by a p-dimensional parameter vector, which may be known or unknown to the statistician. Usually the latter applies, and it thus has to be estimated from the data. The null hypothesis may then be expressed as

$$H_0 : g(x) = f(x;\boldsymbol{\beta}) \text{ for all } x \in \mathbb{R} \text{ and some } \boldsymbol{\beta} \in B, \tag{1.1}$$

where, without loss of generality, we have assumed that both density functions are defined over the whole real line, and where $B \subseteq \mathbb{R}^p$. At this time, we choose not to say anything about the alternative hypothesis, but later, in Section 1.4, we give more details on this. We will refer to the null hypothesis in (1.1) as the *full parametric null hypothesis*.

1.2. Generalised smooth tests

An order k smooth test may be constructed as explained in detail in Rayner and Best (1989, Chapter 6). In particular, a smooth test is a score test for testing $H_0 : \theta_1 = \ldots = \theta_k = 0$ against K : not H_0 in a smooth order k alternative,

$$g_k(x;\boldsymbol{\theta},\boldsymbol{\beta}) = C(\boldsymbol{\theta},\boldsymbol{\beta}) \exp\left\{ \sum_{j=1}^{k} \theta_j h_j(x;\boldsymbol{\beta}) \right\} f(x;\boldsymbol{\beta}), \tag{1.2}$$

where $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_k)$, $C(\boldsymbol{\theta},\boldsymbol{\beta})$ is a normalising constant, and $\{h_j\}$ is a set of orthonormal polynomials on f, that is, they satisfy the equalities

$$\int_{-\infty}^{+\infty} h_i(x;\boldsymbol{\beta}) h_j(x;\boldsymbol{\beta}) f(x;\boldsymbol{\beta}) dx = \delta_{ij},$$

$i, j = 0, 1, 2, \ldots$. We always take $h_0(x;\boldsymbol{\beta}) = 1$ for all $x \in \mathbb{R}$, and $h_1(x) = (x-\mu)/\sigma$, where μ and σ are the mean and the standard deviation of X under H_0. See Rayner, Thas and De Boeck (2008) for a convenient algorithm that generalises the Emerson (1968) recurrence relations for discrete distributions.

Another, but less common construction of smooth goodness of fit tests, results from starting from the order k smooth density given by

$$g_k(x;\boldsymbol{\theta},\boldsymbol{\beta}) = \left(1 + \sum_{j=1}^{k} \theta_j h_j(x;\boldsymbol{\beta}) \right) f(x;\boldsymbol{\beta}). \tag{1.3}$$

This model dates back to the Gram-Charlier series model; see, for example, Stuart and Ord (1994, Section 6.17 and the following sections and the references therein). As it was also considered by Barton (1953) it is sometimes referred to as the Barton model. It also occurs often in the literature on nonparametric density estimation where it is generally known as the orthonormal series density estimator (see e.g. Anderson and De Figueiredo (1980); Buckland (1992); Cencov (1962); Clutton-Brock (1990); Diggle and Hall (1986)). In nonparametric density estimation, usually f is the uniform density over $[0,1]$, and in the few occasions where f is not restricted to the uniform density, f is referred to as a 'parametric start' (Hjort and Glad, 1995), or a 'parametric key' (Buckland, 1992). In most of these references the density f is not parameterised by a nuisance parameter. An important advantage of density (1.3) is that no normalisation constant is needed, and it is an integratable function for bounded θ_j's. On the other hand, when k finite, it is not guaranteed to be a positive function. In Section 3.4 we will give more details on how this problem can be solved.

Baringhaus and Henze (1992), Kallenberg, Ledwina and Rafajlowicz (1997), and Mardia and Kent (1991) remarked that the density in (1.2) is not always well defined, because it is not guaranteed to be integrable, in which case the normalising constant is not well defined. Despite this theoretical problem, the score statistics that we will derive behave properly under the usual mild conditions.

The exact form of the order k smooth test statistic depends on how the nuisance parameter $\boldsymbol{\beta}$ is estimated. When MLE is used, Rayner and Best (1989, chapter 6) give details on how the test statistic is obtained. In the accompanying paper (Rayner, Best and Thas, 2009a) generalised smooth tests are discussed. Generalised smooth tests are basically generalised score tests (Boos, 1992; Hall and Mathiason, 1990), which are valid for the large class of *asymptotically linear estimators*, to which, among others, the MLE and the M-estimators, and thus also the MME belong. In this paper we will only use MME; the reason will become clear shortly. In the next paragraph we briefly discuss the construction of the generalised smooth test, as well as the estimation of $\boldsymbol{\beta}$ by means of the method of moments.

Suppose $\boldsymbol{\beta}$ is known, and let $X_1, \ldots, X_n$ denote a random sample of i.i.d. observations. The score statistic related to θ_j in (1.2) and (1.3) is given by

$$V_j(\boldsymbol{\beta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_j(X_i; \boldsymbol{\beta}).$$

Let $\tilde{\boldsymbol{\beta}}$ denote an asymptotically linear estimator, and $\boldsymbol{V}^t(\boldsymbol{\beta}) = (V_1(\boldsymbol{\beta}), \ldots, V_k(\boldsymbol{\beta}))$. The order k generalised smooth test statistic is then given by

$$\tilde{S}_k = \boldsymbol{V}^t(\tilde{\boldsymbol{\beta}}) \tilde{\boldsymbol{\Sigma}}_0^{-1} \boldsymbol{V}(\tilde{\boldsymbol{\beta}}),$$

in which $\tilde{\boldsymbol{\Sigma}}_0$ is a consistent estimator of $\boldsymbol{\Sigma}_0$, the asymptotic covariance matrix of $\tilde{\boldsymbol{V}} = \boldsymbol{V}(\tilde{\boldsymbol{\beta}})$ under the null hypothesis (1.1). Under this null hypothesis $\tilde{S}_k$ asymptotically has a χ^2_{k-p} null distribution. For more details we refer to Rayner, Best and Thas (2009a). In the present paper is it particularly important to realise that $\boldsymbol{\Sigma}_0$ is defined under the full parametric null hypothesis. In the next subsection we give more details on the nuisance parameter estimation by means of the method of moments.

1.3. Method of moment estimators

If a density function $f(.;\boldsymbol{\beta})$ is indexed by a p dimensional nuisance parameter $\boldsymbol{\beta}$, then its MME is defined as the $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ that makes $f(.;\boldsymbol{\beta})$ agree with the sample data in its first p moments. Since generalised smooth tests are generalised score tests, all nuisance parameters must be estimated under the null hypothesis. Suppose the central moments of the hypothesised distribution f are denoted by μ_{0j}, for $j > 1$, and the mean is denoted by μ_{01}. Here we write $\mu_{0j}(\boldsymbol{\beta})$ to stress that these moments depend on the nuisance parameter. The MME of $\boldsymbol{\beta}$ is then the solution to the estimation equations

$$\sum_{i=1}^{n} (X_i - \mu_{01}(\boldsymbol{\beta})) = 0$$

and

$$\sum_{i=1}^{n} \left\{ \mu_{0j}(\boldsymbol{\beta}) - (X_i - \mu_{01}(\boldsymbol{\beta}))^j \right\} = 0 \text{ for } j = 2, \ldots, p. \tag{1.4}$$

The construction of the orthonormal polynomials implies that the estimation equations (1.4) can be replaced with the equations $(j = 1, \ldots, p)$

$$\sum_{i=1}^{n} h_j(X_i; \boldsymbol{\beta}) = 0.$$

It now follows that the estimation equations are equivalent to $V_1(\tilde{\boldsymbol{\beta}}) = \ldots = V_p(\tilde{\boldsymbol{\beta}}) = 0$, that is, the first p components are exactly zero. Another consequence is that the $k \times k$ covariance matrix $\boldsymbol{\Sigma}_0$ is not well defined. We therefore agree to remove the first p terms from (1.2) and (1.3).

1.4. The diagnostic property

In some cases the matrix $\tilde{\boldsymbol{\Sigma}}_0$ is a diagonal matrix, say $\text{diag}(\tilde{\sigma}^2_{p+1}, \ldots, \tilde{\sigma}^2_k)$, so that a decomposition of $\tilde{S}_k$ follows, i.e.

$$\tilde{S}_k = \tilde{\boldsymbol{V}}^t \tilde{\boldsymbol{\Sigma}}_0^{-1} \tilde{\boldsymbol{V}} = \sum_{j=p+1}^{k} \frac{\tilde{V}_j^2}{\tilde{\sigma}_j^2} .$$

Moreover, under the full parametric null hypothesis, the standardised components, $\tilde{V}_j/\tilde{\sigma}_j$, all have an asymptotic standard normal distribution, and they are asymptotically independent. Klar (2000) showed that this decomposition always arises for distributions in which MLE and MME coincide.

The jth component, which is constructed from a polynomial of degree j, is a linear combination of contrasts between sample moments and the corresponding moments of the hypothesised distribution f up to the jth order. The latter moments depend on the nuisance parameter β, and when MME is used the first p contrasts in $\tilde{V}_j$ are exactly zero. When focussing on the first non-zero component, $\tilde{V}_{p+1}$, it would be tempting to conclude that when the null hypothesis is rejected due to a large $\tilde{V}_{p+1}/\tilde{\sigma}_{p+1}$, the true and the hypothesised distributions do not agree in the $(p+1)$th moment. Similar reasoning applies to higher order components, except that no more than one moment may be involved. However, as Henze and Klar (Henze and Klar, 1996; Henze, 1997; Klar, 2000) showed, the arguments just given are not correct. Since we will very often refer to their results, we use the abbreviation HK to refer to them and their papers just cited. Before we summarise their arguments, we first properly define a diagnostic component test.

A size α test based on the jth order component is said to be (asymptotically) diagnostic for the jth order moment when (1) it has (asymptotically) size α if and only if the true and the hypothesised distributions agree in the jth moment; and (2) it is consistent under the alternative that the true and the hypothesised distributions disagree in the jth moment.

The core of the argument of HK is based on the variance of $\tilde{V}_j$ which plays an important role as the hypothesis test is based on $\tilde{V}_j/\tilde{\sigma}_j$ and the standard normal distribution. Since

$\tilde{V}_j$ contains sample observations raised to the power j, its variance depends on moments of X up the order $2j$. Under the full parametric null hypothesis all moments are completely determined by the hypothesised distribution, and $\tilde{\sigma}_j^2$ is a consistent estimator of Var $\left[\tilde{V}_j\right]$. However, when the null hypothesis does not hold, deviations in moments up to the order $2j$ may influence the variance of $\tilde{V}_j$, and thus also the asymptotic distribution of $\tilde{V}_j/\tilde{\sigma}_j$. Even when the true and the hypothesised moments agree, a deviation in any of the other moments up to order $2j$ may alter the asymptotic distribution of $\tilde{V}_j/\tilde{\sigma}_j$, so that a test based on this component, using the standard normal distribution as a reference distribution, may have a power far greater than the expected nominal significance level α. The same argument may make the variance of $\tilde{V}_j$ so small that the component test using the standard normal as a null distribution has virtually no power at all.

For a large class of distributions HK solved the problem by *rescaling* the components. In particular, they suggest basing the hypothesis test on $\tilde{V}_j/\tilde{\sigma}_{Ej}$, where $\tilde{\sigma}_{Ej}^2$ is an estimator of the asymptotic variance of $\tilde{V}_j$ which is consistent under the *partial semiparametric null hypothesis* that g and f have equal jth moments. For the class of densities for which MLE and MME coincide, they propose the estimator

$$\tilde{\sigma}_{Ej}^2 = \frac{1}{n}\sum_{i=1}^{n} h_j^2(X_i;\tilde{\boldsymbol{\beta}}), \tag{1.5}$$

which is basically the empirical variance estimator of $\tilde{V}_j$. In Section 2 we give more theoretical details.

1.5. Data-driven smooth tests

A common criticism to smooth tests is that the order k has be be chosen a priori by the statistician. For a fixed order k, and using the terminology introduced in the previous section, we may say that an order k smooth test is only consistent for testing the *semiparametric null hypothesis* that f and g agree in their first k moments. In this sense the order k smooth test is not omnibus consistent.

In a series of papers, Ledwina and Kallenberg (Ledwina, 1994; Kallenberg and Ledwina, 1995, 1997) introduced data-driven smooth tests for which the order k is selected from the data by optimising the BIC model selection criterion. Their theory shows that these data-driven tests regain the omnibus consistency property. Many simulation studies have indicated that data-driven smooth tests have better overall power properties than fixed order k smooth tests. Their data-driven techniques, however, build upon the Neyman smooth test for uniformity; that is they work with the probability integral transformed observations so that they always use Legendre polynomials and they loose the interpretability of the components. More recently, Claeskens and Hjort (2004) suggested data-driven versions of the type of smooth tests discussed in Section 1.2. Starting from the order k smooth density of (1.2), they constructed both the likelihood ratio and the score tests. Most of the solutions they proposed are restricted to the case where the nuisance parameter $\boldsymbol{\beta}$ is assumed known. Some details on the unknown $\boldsymbol{\beta}$ case are presented in their section 6, but since they only considered MLE and they did not aim at producing interpretable components, we do not follow their theory exactly. They considered model selection based on the AIC, BIC and the BIG criteria.

1.6. Informative generalised smooth tests

The idea that we pursue here can be summarised as follows. Since the generalised smooth test statistics do not generally decompose into asymptotically independent components, and since the individual components are not necessarily diagnostic, we suggest using a fitted *improved density* estimate to see how the true density function might deviate from the hypothesised. A fitted improved density is simply a density of the form (1.2) or (1.3) with the unknown parameters replaced by estimates. We suggest only looking at the improved density estimate if the data-driven test results in the rejection of the null hypothesis. The relation between data-driven smooth tests and model selection now becomes very relevant. If some model selection rule gives that only a subset of the k θ parameters must be included, then it is sufficient to estimate these selected parameters and plot the corresponding fitted density. Thus AIC and BIC are very natural choices. However, these model selection criteria do not necessarily give good density estimates from an estimation point of view. In the literature about nonparametric density estimation, density estimates based on (1.3) are known as orthonormal series expansion. In this paper we propose a data-driven smooth test based on a criterion typically used for the selection of estimators: a weighted mean integrated squared error criterion (WISE). Our method has good properties and it gives no computational problems.

Since the WISE criterion requires a consistent estimator of the variance of $\tilde{V}_j$, we first propose an estimator different from those discussed by HK. Although this new variance estimator may also be used to construct asymptotically diagnostic component tests, we do not pursue this approach; simulation studies (not shown) have indicated that the convergence is so extremely slow that we could not recommend these rescaled component tests in practice.

2. A Consistent Variance Estimator

In this section we propose a consistent variance estimator for a class of distributions for which MLE and MME do not coincide. First, in Section 2.1, we give a formal construction of the semiparametric framework. The variance estimator is the topic of Section 2.2. Its performance is empirically investigated in Section 2.3.

2.1. The semiparametric framework

We first introduce a set $\mathcal{P}$ of proper densities, defined as ($m \geq 1$)

$$\mathcal{P}_m = \left\{ g \in \mathcal{D} : \int_{-\infty}^{+\infty} x^j g(x) dx < \infty, j = 1, \ldots, m \right\},$$

where $\mathcal{D}$ is the set of all continuous density functions defined over $(-\infty, +\infty)$. The order k semiparametric null hypothesis can now be formulated as

$$H_0^{SP} : g \in \mathcal{F}_0 = \left\{ g \in \mathcal{P}_{2k} : \mathrm{E}_g\left[h_1(X; \boldsymbol{\beta})\right] = \ldots = \mathrm{E}_g\left[h_k(X; \boldsymbol{\beta})\right] = 0, \boldsymbol{\beta} \in B \right\}.$$

Obviously $f \in \mathcal{F}_0$, but the set $\mathcal{F}_0$ also contains densities not consistent with f. Thus if $\boldsymbol{\beta}$ has to be estimated under the semiparametric null hypothesis, it is too restrictive to use f. Moreover, in the current semiparametric setting the function f is only used to generate the first k hypothesised moments. Therefore, we refer to f as the *moment generating density function*. A consequence of this discussion is that the MLE of $\boldsymbol{\beta}$ is meaningless in this setting. Since the semiparametric null hypothesis is about moments, HK argue that the method of moments estimator is the only sensible solution. Under the semiparametric null hypothesis, $\boldsymbol{\beta}$ is defined as the solution of

$$\mathrm{E}_g\left[h_j(X;\boldsymbol{\beta})\right] = 0 \quad (j = 1,\ldots,p; g \in H_0^{SP}). \tag{2.1}$$

This equation basically says that every p-dimensional vector $\boldsymbol{\beta}$ fixes the first p moments of f to those of g. By inverting this relation, we could say that every set of p moments of f, say $\mu_1,\ldots,\mu_p$, determines $\boldsymbol{\beta}$ uniquely for a given $g \in H_0^{SP}$. Let $\boldsymbol{\beta}(g)$ denote the parameter $\boldsymbol{\beta}$ that satisfies (2.1). The semiparametric null hypothesis can be restated as

$$H_0^{SP} : g \in \mathcal{F}_0 = \left\{g \in \mathcal{P}_{2k} : \mathrm{E}_g\left[h_{p+1}(X;\boldsymbol{\beta}(g))\right] = \ldots = \mathrm{E}_g\left[h_k(X;\boldsymbol{\beta}(g))\right] = 0\right\}.$$

We also introduce the r-th partial semiparametric null hypothesis,

$$H_{0;r}^{PSP} : g \in \mathcal{F}_{0;r} = \left\{g \in \mathcal{P}_{2r} : \mathrm{E}_g\left[\{X - \mu_1(\boldsymbol{\beta}(g))\}^r\right] = 0\right\}.$$

2.2. The variance estimator

The method of moments estimator $\tilde{\boldsymbol{\beta}}$ belongs, under suitable regularity conditions, to the class of *local asymptotical linear estimators*. We refer, for example, to van der Vaart (1998) for more technical details on locally asymptotically linear estimators.

The next theorem gives a consistent estimator of the variance of $\tilde{\boldsymbol{V}}_r$. Although HK did not use this estimator in their examples, the estimator can also be derived from the proof of theorem 2.1 in Klar (2000).

Theorem 2.1. *Assume $\tilde{\boldsymbol{\beta}}$ is a locally asymptotically linear estimator with representation*

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{b}(X_i;\boldsymbol{\beta}) + o_P(1), \tag{2.2}$$

where

$$\boldsymbol{b}(x;\boldsymbol{\beta}) = \left\{E_g\left[-\frac{\partial \boldsymbol{h}_\beta}{\partial \boldsymbol{\beta}}(X)\right]\right\}^{-1}\boldsymbol{h}_\beta(x;\boldsymbol{\beta}).$$

Let

$$w_r(x;\boldsymbol{\beta}) = h_r(x;\boldsymbol{\beta}) + \boldsymbol{b}^t(x;\boldsymbol{\beta})E_g\left[\frac{\partial h_r}{\partial \boldsymbol{\beta}}(X;\boldsymbol{\beta})\right], \tag{2.3}$$

and $\bar{w}_r(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} w_r(X_i;\boldsymbol{\beta})$. For all $r \in \{p+1,\ldots,k\}$, under the order k semiparametric null hypothesis, i.e. $g \in \mathcal{F}_0$, a consistent estimator of the asymptotic variance of $\tilde{V}_r$ is given *by*

$$\tilde{\sigma}_r^2(\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\left(w_r(X_i;\tilde{\boldsymbol{\beta}}) - \bar{w}_r(\tilde{\boldsymbol{\beta}})\right)^2. \tag{2.4}$$

Proof. First note that a Taylor series expansion of $\tilde{V}_r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_r(X_i;\tilde{\boldsymbol{\beta}})$ and the substitution of $\tilde{\boldsymbol{\beta}}$ with its representation (2.2) gives

$$\tilde{V}_r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h_r(X_i;\boldsymbol{\beta}) + \boldsymbol{b}^t(X_i;\boldsymbol{\beta})\mathrm{E}_g\left[\frac{\partial h_r(Y;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right] + o_P(n^{-1/2})\right). \tag{2.5}$$

Further note that

$$\tilde{V}_r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_r(X_i;\tilde{\boldsymbol{\beta}}).$$

We first consider $\boldsymbol{\beta}$ known.

Write

$$\tilde{\sigma}_r^2(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} w_r^2(X_i;\boldsymbol{\beta}) - \bar{w}_r^2(\boldsymbol{\beta}).$$

Then

$$\begin{aligned}
\mathrm{E}_g\left[\tilde{\sigma}_r^2(\boldsymbol{\beta})\right] &= \mathrm{E}_g\left[w_r^2(X;\boldsymbol{\beta})\right] - \mathrm{E}_g\left[\bar{w}_r^2(\boldsymbol{\beta})\right] \\
&= \mathrm{E}_g\left[w_r^2(X;\boldsymbol{\beta})\right] - \left(\mathrm{Var}_g\left[\bar{w}_r(\boldsymbol{\beta})\right] + \mathrm{E}_g\left[\bar{w}_r(X;\boldsymbol{\beta})\right]^2\right) \\
&= \mathrm{E}_g\left[w_r^2(X;\boldsymbol{\beta})\right] - \left(\frac{1}{n}\mathrm{Var}_g\left[w_r(X;\boldsymbol{\beta})\right] + \mathrm{E}_g\left[w_r(X;\boldsymbol{\beta})\right]^2\right) \\
&= \left(\mathrm{E}_g\left[w_r^2(X;\boldsymbol{\beta})\right] - \mathrm{E}_g\left[w_r(X;\boldsymbol{\beta})\right]^2\right) - \frac{1}{n}\mathrm{Var}_g\left[w_r(X;\boldsymbol{\beta})\right] \\
&= \frac{n-1}{n}\mathrm{Var}_g\left[w_r(X;\boldsymbol{\beta})\right].
\end{aligned}$$

Since, by (2.5), $\tilde{V}_r = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(w_r(X_i;\boldsymbol{\beta}) + o_P(n^{-1/2}))$, we have, as $n \to \infty$, $\mathrm{E}_g\left[\tilde{\sigma}_r^2(\boldsymbol{\beta})\right] - \mathrm{Var}_g\left[\tilde{V}_r\right] \to 0$. Moreover, since $\tilde{\boldsymbol{\beta}}$ is a $\sqrt{n}$ consistent estimator, as $n \to \infty$,

$$\mathrm{E}_g\left[\tilde{\sigma}_r^2(\tilde{\boldsymbol{\beta}})\right] \to \mathrm{Var}_g\left[\tilde{V}_r\right]. \qquad \square$$

From this proof it can be seen that it is particularly the $\frac{\partial h_r}{\partial\boldsymbol{\beta}}(x;\boldsymbol{\beta})$ term in the w_r function (2.3) that makes the new variance estimator different from $\tilde{\sigma}_{Er}^2$ in (1.5). The expectation of this term appears in the expansion (2.5) of $\tilde{V}_r$, and therefore also in the variance. It is exactly the expectation of this term that is assumed to be zero in (1.5). This happens for all distributions in the exponential family for which MLE and MME coincide, for example the normal, Poisson, exponential and the binomial distributions. Distributions for which the correction terms are necessary include the logistic, extreme value, Laplace, negative-binomial, beta-binomial and the generalised Pareto distributions.

The correction term makes the new variance estimator (2.4) slightly more complicated, particularly because this correction terms depends on the distributions and on the estimation function of the nuisance parameter. In Appendices A and B the partial derivatives $\frac{\partial h_r}{\partial \boldsymbol{\beta}}(x;\boldsymbol{\beta})$ are listed for the logistic and the extreme value distributions. HK suggested another solution: (1) compute Var $\left[\tilde{V}_r\right]$ and express it in terms of moments up to the order $2r$; (2) replace these moments by their empirical moment estimators.

2.3. Simulation study

In this section we empirically assess the validity of the new variance estimator in simulation studies. In the first series of simulations the asymptotic unbiasedness of the estimator is assessed, while in the second we investigate the diagnostic property of rescaled component tests, using the new variance estimator. Both the logistic and the extreme value distributions are considered.

2.3.1. The bias

Since the variance of the components should be estimated consistently under the semiparametric null hypothesis, which may, for example, only specify one particular moment, we have considered several distibutions in this simulation study.

- For the components of the generalised smooth test for the logistic distribution we have simulated from a logistic distribution; all moments are thus consistent with the semiparametric null hypothesis.
- For the components of the generalised smooth test for the extreme value distribution we have simulated from an extreme value distribution; all moments are thus consistent with the semiparametric null hypothesis.
- For both testing for the logistic and the extreme value distributions, we have further included:
  - ○ the normal distribution: it has the same skewness (symmetric) as the logistic distribution;
  - ○ the uniform distribution over $[0,1]$ has the same skewness (symmetric) as the logistic distribution;
  - ○ the exponential distribution: neither the third or the fourth moment agree with the logistic or the extreme value distribution;
  - ○ the gamma distribution with shape paramter γ: with increasing γ the gamma distribution becomes more symmetric, and with $\gamma = 5$ it has the same kurtosis as the logistic distribution, and with $\gamma = 3.08$ it has the same skewness as the extreme value distribution.

For samples sizes $n = 100$, $n = 500$ and $n = 1000$, $10,000$ Monte-Carlo simulation runs are performed. In each simulation run, the new and the simple empirical variance estimators for $\tilde{V}_3$ and $\tilde{V}_4$ are computed. Although $\tilde{\sigma}_E^2$ is not an appropriate estimator in the current setting, we only include it here to demonstrate the beneficial effect of the correction term in

(2.4). The averages of these estimates approximate the expected values of the estimators. Based on the $10,000$ simulation runs the true variances of $\tilde{V}_3$ and $\tilde{V}_4$ are approximated as the variance of the $10,000$ simulated components. These serve as the benchmark with which the approximated means of the variance estimators have to be compared so as to assess the bias.

The results for the logistic and the extreme value distributions are presented in Tables 2.1 and 2.2. We first discuss the results for the largest sample size considered, i.e. $n = 1000$. The expected values of the new variance estimator seem almost always to be very close to the true variance, for both the logistic and the extreme value distribution. The simpler estimator $\tilde{\sigma}_E^2$ performed clearly worse. However, for the fourth order component under the exponential distribution this estimator outperformed the new one, both for the logistic and the extreme value components. The variance of the fourth order logistic component was also better estimated by $\tilde{\sigma}_E^2$ under the very skewed gamma distributions with small γ parameter. From comparing the results for the smaller sample sizes ($n = 100$ and $n = 500$), we conclude that the convergence is very slow, and we conclude thus that a substantial bias persists unless the sample size is sufficiently large.

2.3.2. The diagnostic property

In this section we present the results of a limited simulation study in which the powers of the component tests are estimated under various alternatives so as to assess the diagnostic property of the rescaled components. As argued in Section 1.4, when ensuring that a component test is diagnostic the component should be properly rescaled before using it as a test statistic. We have set up a simulation study using the same alternatives as in the bias simulation study, but we will not present all results here. Since the convergences of the variance estimators are very slow, large sample sizes are needed before the (asymptotic) diagnostic property is true. This was also concluded by HK from their simulation studies. These large sample sizes, however, result in powers of approximately 100% under many interesting alternatives, so that a comparison becomes uninformative. Therefore, we present here only the results under alternatives that have at least one moment in agreement with one of the partial semiparametric null hypotheses. For these alternatives we expect the rescaled component tests to have size close to the nominal significance level.

Again all results are based on $10,000$ Monte Carlo simulation runs, but only sample size $n = 1000$ is considered. In each simulation run rescaled component tests based on both $\tilde{\sigma}_E^2$ and $\tilde{\sigma}^2$, and the unscaled component tests are all performed at the 5% level of significance. The latter is the MME based generalised smooth test, using the asymptotic variance under the full parameteric null hypothesis (see Appendices A and B for these variances). Since all tests are supposed to test a semiparametric null hypothesis, their null distributions must be computed under these semiparametric nulls. The theory presented in HK shows that the rescaled components, using any $\sqrt{n}$ consistent variance estimator, are asymptotically standard normally distributed under the appropriate (partial) semiparametric null hypothesis. Since we expect that the convergence to this limitting distribution is slow, we have tried some alternative methods, but none gave satisfactory results. For example, the bootstrap method proposed by Bickel, Ritov and Stoker (2006) gave worse results than the asymptotic standard normal quantiles. We therefore used the latter in our simulations.

Table 2.1. Averages of the variance estimates $\tilde{\sigma}_j^2$ and $\tilde{\sigma}_{Ej}^2$ of the components ($j = 3, 4$) for the logistic distribution (from 10,000 simulations). As a benchmark the approximate true variance of the components is also presented (Var $\left[\tilde{V}_j\right]$).

	$\tilde{\sigma}_j^2$		$\tilde{\sigma}_{Ej}^2$		Var $\left[\tilde{V}_j\right]$	
	$j = 3$	$j = 4$	$j = 3$	$j = 4$	$j = 3$	$j = 4$
			$n = 1000$			
			logistic			
$\sigma = 1$	0.97	0.86	0.96	0.88	1.03	0.98
$\sigma = 2$	0.97	0.83	0.95	0.85	1.03	0.92
			normal			
$\sigma = 1$	0.26	0.09	0.33	0.28	0.27	0.09
$\sigma = 2$	0.26	0.08	0.33	0.28	0.27	0.09
			uniform $[0, 1]$			
	0.09	0.01	0.29	0.22	0.09	0.00
			exponential			
$\gamma = 2$	2.40	16.63	8.27	30.47	2.94	25.01
			gamma			
$\gamma = 3$	0.84	2.53	1.95	2.83	0.92	3.03
$\gamma = 5$	0.59	1.25	1.18	1.25	0.62	1.40
$\gamma = 7.5$	0.48	0.74	0.86	0.73	0.50	0.81
			$n = 500$			
			logistic			
$\sigma = 1$	0.90	0.70	0.93	0.80	1.03	0.90
$\sigma = 2$	0.90	0.73	0.93	0.83	1.00	0.92
			normal			
$\sigma = 1$	0.26	0.08	0.33	0.28	0.27	0.09
			gamma			
$\gamma = 5$	0.54	0.99	1.13	1.09	0.61	1.26
			$n = 100$			
			logistic			
$\sigma = 1$	0.53	0.23	0.72	0.49	0.81	0.53
$\sigma = 2$	0.52	0.23	0.71	0.50	0.80	0.53
			normal			
$\sigma = 1$	0.22	0.06	0.32	0.28	0.25	0.08
			gamma			
$\gamma = 5$	0.31	0.30	0.85	0.58	0.47	0.72

Table 2.2. Averages of the variance estimates $\tilde{\sigma}_j^2$ and $\tilde{\sigma}_{Ej}^2$ of the components ($j=3,4$) for the extreme value distribution (from 10,000 simulations). As a benchmark the approximate true variance of the components is also presented (Var $\left[\tilde{V}_j\right]$).

	$\tilde{\sigma}_j^2$		$\tilde{\sigma}_{Ej}^2$		Var $\left[\tilde{V}_j\right]$	
	$j=3$	$j=4$	$j=3$	$j=4$	$j=3$	$j=4$
$n=1000$						
extreme value						
$b=1$	1.30	1.02	0.91	0.76	1.43	1.09
$b=2$	1.33	1.05	0.93	0.86	1.46	1.16
normal						
$\sigma=1$	0.29	1.63	1.77	3.35	0.30	1.67
$\sigma=2$	0.29	1.63	1.77	3.37	0.29	1.61
uniform $[0,b]$						
$b=1$	0.10	0.54	0.58	0.48	0.11	0.55
$b=2$	0.10	0.54	0.58	0.48	0.10	0.53
exponential						
$\gamma=1$	2.61	2.18	2.42	3.84	3.17	3.78
$\gamma=2$	2.59	2.07	2.39	3.42	3.17	3.38
gamma						
$\gamma=1$	2.55	2.10	2.35	3.76	3.08	3.75
$\gamma=2$	1.29	0.79	0.86	0.68	1.44	0.89
$\gamma=3.08$	0.91	0.75	0.63	0.47	0.99	0.77
$\gamma=4$	0.74	0.78	0.60	0.47	0.77	0.77
$n=500$						
extreme value						
$b=1$	1.11	0.97	0.85	0.70	1.33	1.01
$b=2$	1.09	0.96	0.84	0.69	1.33	1.04
gamma						
$\gamma=3.08$	0.79	0.74	0.60	0.45	0.94	0.76
$n=100$						
extreme value						
$b=1$	0.46	0.94	0.64	0.56	0.92	0.96
$b=2$	0.45	0.91	0.62	0.55	0.87	0.95
gamma						
$\gamma=3.08$	0.39	0.73	0.50	0.42	0.69	0.74

Table 2.3. Powers of the component tests ($\alpha = 5\%$) for the logistic distribution (number of rejections out of $10,000$ simulations).

	$\tilde{V}_j/\tilde{\sigma}_j$		$\tilde{V}_j/\tilde{\sigma}_{Ej}$		$\tilde{V}_j$	
$n = 1000$	$j = 3$	$j = 4$	$j = 3$	$j = 4$	$j = 3$	$j = 4$
			logistic			
$\sigma = 1$	607	1743	528	1111	474	346
$\sigma = 2$	585	1665	522	1040	509	335
			normal			
$\sigma = 1$	538	9975	271	9974	2	8283
$\sigma = 2$	572	9982	270	9981	1	8276
			uniform $[0, b]$			
$b = 1$	455	10000	4	10000	0	10000
$b = 2$	511	10000	5	10000	0	10000
			gamma			
$\gamma = 5$	10000	1744	9998	1606	10000	558

Table 2.4. Powers of the component tests ($\alpha = 5\%$) for the extreme value distribution (number of rejections out of $10,000$ simulations).

	$\tilde{V}_j/\tilde{\sigma}_j$		$\tilde{V}_j/\tilde{\sigma}_{Ej}$		$\tilde{V}_j$	
$n = 1000$	$j = 3$	$j = 4$	$j = 3$	$j = 4$	$j = 3$	$j = 4$
			extreme value			
$\sigma = 1$	1256	523	1316	1318	358	265
$\sigma = 2$	1234	480	1270	1264	390	303
			gamma			
$\gamma = 3.080$	1109	2098	1199	3724	175	677

The results for the logistic and the extreme value distributions are presented in Tables 2.3 and 2.4, respectively. For testing for the moments of the logistic distribution, the new third order rescaled component test seems to have good sizes, whereas all other third order test sizes are too small under the normal and uniform alternatives. For the fourth order components, even under the logistic full parametric null hypothesis, the sizes of the rescaled tests are too large. This is most likely due to the use of the standard normal quantiles as critical values. Despite the increased size, the new fourth order test seems to retain this size under the gamma distribution with $\gamma = 5$. When testing for the moments of the extreme value distribution, we see approximately the same behaviour, except that now the fourth order test has good size, and the size of the third order test is too large. Although this limited simulation study illustrates the necessity of correctly standardising the components, all rescaled tests seem to have very limited value in most practical settings where sample sizes are often much smaller than $n = 1000$. It is particularly this observation that motivates the informative goodness of fit procedure that is the topic of Section 3.

3. Data-Driven Smooth Tests and Nonparametric Density Estimation

3.1. Introduction

The simulation results presented in Section 2.3 have demonstrated that even the properly rescaled component tests do not possess the diagnostic property in small to moderately large samples. On the other hand it is generally known that (generalised) smooth goodness of fit tests have good power so that these tests are still often to be recommended in practice. The conclusion about the component tests only implies that at the rejection of the null hypothesis, nothing more informative can be said on how the true and the hypothesised distributions differ, at least not for most realistic sample sizes. In this section we propose a method that is related to data-driven smooth tests, as well as to nonparametric density estimation.

Data-driven smooth tests differ from ordinary smooth tests by selecting the order k using the data. In particular, a model selection criterion is first applied to the data, resulting in an 'estimated' order, which is subsequently used as the order of the smooth test. This order selection process affects the null distribution of the test statistic. See for example Ledwina (1994); Kallenberg and Ledwina (1997); Claeskens and Hjort (2004) for the theory of data-driven tests. The most common selection rules used in data-driven smooth testing are Akaike's Informatin Criterion (AIC, (Akaike, 1973, 1974)) and the Bayesian Information Criterion (BIC, (Schwarz, 1978)). More details on these criteria are given in Section 3.2.3. Despite the popularity of these selection rules, and their importance in statistical model selection in general, we suggest here using a criterion that originates from nonparametric density estimation. In Section 1.2, when we introduced the order k alternative (1.3), we mentioned that this particular form is also the basis of orthogonal series density estimators. By applying a criterion that aims at minimising the overall bias of the density estimate, and by plotting the resulting nonparametric density estimate, the statistician may use this density estimate as a basis for formulating conclusions. Moreover, since in this way the data-driven test and the density estimate are based on the same model selection criterion, no contradictory conclusion will arise.

In Section 3.2 more details on selection criteria are given. Our data-driven test is presented in Section 3.3, and in Section 3.5 the new procedure is evaluated in a simulation study.

3.2. Model selection criteria

3.2.1. The horizon and improved density estimates

In the previous section we said that a model selection criterion selects the 'order' of the smooth test, but more generally it may be used to select any number of components from an a priori speficied set. Let S_h be an index set of the form $\{p+1,\dots,k\}$, where $k > p$ is the maximal order one is prepared to consider. This index set is often called the *horizon*, which explains the S_h notation. In particular, we restrict our discussion to finite horizons, i.e. $k < \infty$. Let $S \subseteq S_h$ and let g_S denote the smooth density defined by

$$g_S(x;\theta_S,\boldsymbol{\beta}) = \left\{ 1 + \sum_{i \in S} \theta_i h_i(x;\boldsymbol{\beta}) \right\} f(x;\boldsymbol{\beta}). \tag{3.1}$$

where $\boldsymbol{\theta}_S^T = (\theta_i)_{i\in S}$. This definition allows for arbitrary index sets S, whereas g_k implies that all terms $\theta_i h_i(x;\boldsymbol{\beta})$ with $i \leq k$ are included. Finally, we also introduce the notation g_∞ for g_S with $S = \{p+1,\dots\}$, so that $g_\infty = g$.

Once a model S has been selected, the θ_j ($j \in S$) parameters in (3.1) can be estimated by $\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^n h_j(X_i;\tilde{\boldsymbol{\beta}})$. Substituting these, and the estimates of the nuisance parameters ($\tilde{\boldsymbol{\beta}}$) into (3.1) results in $g_S(x;\hat{\boldsymbol{\theta}}_S,\tilde{\boldsymbol{\beta}})$, which we refer to as the *improved density estimate*. Note that we actually do not necessarily use the most efficient estimation scheme here. First, the estimators $\hat{\theta}_j$ may be unbiased as demonstrated in Section 1.2, but they are not necessarally efficient. Moreover, they are conditional on $\tilde{\boldsymbol{\beta}}$, which is the MME of the nuisance parameter under the null hypothesis. It is likely that more efficient estimation is possible, but we aim here at a simple procedure from an implementation point of view. A similar estimation approach was also suggested by Claeskens and Hjort (2004), who referred to it as two-stage estimation. Buckland (1992) and Efron and Tibshirani (1996) discuss other estimation procedures. The improved density estimate is basically an orthogonal series nonparametric density estimate, with two major differences as to how it usually appears in the statistical literature. First, the 'parametric start' has nuisance parameters, and, second, the terms in S are only considered when the related data-driven test (see Section 3.3) rejects the null hypothesis.

3.2.2. Loss functions and the WISE model selection criterion

Model selection criteria often originate from loss functions. In the present context a loss function measures the discrepancy between the true and the selected model. It is a positive function that cannot increase with increasing complexity of the selected model, and which is zero if and only if the selected model is the true data-generating model. We denote the loss resulting from using g_S instead of g as $\Lambda(g,g_S) = \Lambda(g,g_S,\boldsymbol{\theta},\boldsymbol{\beta})$.

The basic idea behind the use of loss functions for model selection is that the model S should be chosen so that the loss function is minimized. This can be done if the parameter $\boldsymbol{\gamma}^T = (\boldsymbol{\beta},\boldsymbol{\theta})$ is replaced by an estimate, say $\hat{\boldsymbol{\gamma}}$, but this would almost always result in choosing the largest model among the models in $\mathcal{S}$. Thus this simple plug-in principle does not make much sense here. Moreover, using estimates would only result in the selection of the 'best' model for a given data set, whereas one wants to select a model that describes any other random sample from the same distribution g just as well as the sample used for model selection. The solution exists in using the expected loss $\mathrm{E}_g\left[\Lambda(g,g_S;\hat{\boldsymbol{\gamma}})\right]$ as a criterion. Here the expectation is taken over the estimators $\hat{\boldsymbol{\gamma}}$ with respect to the true distribution g of the sample observations. Since the expected loss typically contains unknown parameters, the actual model selection criterion is taken as an (asymptotic) unbiased estimator of the expected loss. This general idea is applied in the next paragraph.

Consider the *weighted integrated squared error* (weighted ISE),

$$\Lambda(g,g_S,\boldsymbol{\theta},\boldsymbol{\beta}) = \int_{-\infty}^{+\infty} \frac{(g(x)-g_S(x;\boldsymbol{\beta},\boldsymbol{\theta}))^2}{f(x;\boldsymbol{\beta})}dx, \tag{3.2}$$

where $1/f(x;\boldsymbol{\beta})$ serves as the weight function. The same loss function has been studied by Anderson and Figueiredo (1980) in the context of nonparametric density estimation, and by Eubank, LaRiccia and Rosenstein (1987) for goodness of fit testing. It is also known as Pearson's ϕ^2 divergence. Eubank, LaRiccia and Rosenstein (1987) showed that $\theta_j = \int_{-\infty}^{+\infty} h_j(x) g_S(x) dx$ is the minimiser of (3.2) for a fixed subset S. When g_S equals to the true density g, the estimator $\hat{\theta}_j$ is the plug-in estimator of θ_j, which is unbiased.

Let $\bar{S} = \{p+1, p+2, \ldots\} \setminus S$, and let g_∞ denote the expansion (1.3) with $k \to \infty$. Replacing g with g_∞, and $(\boldsymbol{\beta}, \boldsymbol{\theta})$ with $(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ in (3.2) gives

$$
\begin{aligned}
\Lambda(g_\infty, g_S, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}}) &= \int_{-\infty}^{+\infty} \frac{\left[\left(1 + \sum_{j=1}^{\infty} \theta_j h_j(x; \tilde{\boldsymbol{\beta}})\right) f(x; \tilde{\boldsymbol{\beta}}) - \left(1 + \sum_{j \in S} \hat{\theta}_j h_j(x; \tilde{\boldsymbol{\beta}})\right) f(x; \tilde{\boldsymbol{\beta}})\right]^2}{f(x; \tilde{\boldsymbol{\beta}})} dx \\
&= \int_{-\infty}^{+\infty} \frac{f^2(x; \tilde{\boldsymbol{\beta}})}{f(x; \tilde{\boldsymbol{\beta}})} \left(\sum_{j \in S} (\theta_j - \hat{\theta}_j) h_j(x; \tilde{\boldsymbol{\beta}}) + \sum_{j \in \bar{S}} \theta_j h_j(x; \tilde{\boldsymbol{\beta}}) \right)^2 dx \\
&= \sum_{j \in S} (\theta_j - \hat{\theta}_j)^2 + \sum_{j \in \bar{S}} \theta_j^2.
\end{aligned}
$$

The last step is basically Parseval's identity. To guarantee that Λ remains bounded, we require that the density g belongs to the ellipsoid

$$
\mathcal{G} = \left\{ g : \sum_{j=1}^{\infty} \theta_j^2 < \infty, \theta_j = \int_{-\infty}^{+\infty} h_j(x; \boldsymbol{\beta}) g(x) dx \right\}. \tag{3.3}
$$

For practical purposes, however, we may truncate the orthonormal expansion g_∞ of g at some specified maximal order, say k. We therefore redefine $\bar{S} = \{p+1, \ldots, k\} \setminus S$, and let $S_{\max} = \{p+1, \ldots, k\}$. We further write $\Lambda(g, g_S, \boldsymbol{\theta}, \boldsymbol{\beta})$ as $\Lambda(S, \boldsymbol{\theta}, \boldsymbol{\beta})$. When $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are replaced by their estimates $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\beta}}$, this loss measures how well the true density is estimated by the nonparametric density estimate $g_S(x; \tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ in terms of the weighted ISE.

This loss suffers from the disadvantage that it only measures how well the estimated g_S approximates g for the observed sample. To overcome this problem, the expected loss

$$
\mathrm{E}\left[\Lambda(S, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}})\right] = \sum_{j \in S} \mathrm{Var}\left[\hat{\theta}_j\right] + \sum_{j \in \bar{S}} \theta_j^2,
$$

is considered. This can be recognised as a weighted version of the mean ISE. The unweighted mean ISE is generally abbreviated as MISE; here we refer to the weighted MISE as WISE. The expected loss cannot be used as a criterion because it depends on unknown parameters. In practice, we therefore have to consider an unbiased estimator of the expected loss. Let $\tilde{\sigma}_j^2$ denote the estimator of $\mathrm{Var}\left[\tilde{V}_j\right]$ as proposed in Theorem 2.1, or, when MME and MLE coincide, the estimator (1.5) of Henze and Klar may be more convenient. Here we propose

$$
\hat{\Lambda}(S, \hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\beta}}) = \sum_{j \in S} \tilde{v}_j^2 + \sum_{j \in \bar{S}} \left(\hat{\theta}_j^2 - \tilde{v}_j^2\right), \tag{3.4}
$$

where $\tilde{v}_j^2 = \frac{1}{n} \tilde{\sigma}_j^2$ is an estimator of $\mathrm{Var}\left[\hat{\theta}_j\right] = \mathrm{Var}\left[\frac{1}{\sqrt{n}} \tilde{V}_j\right]$. Finally, we suggest replacing the expected loss estimator in (3.4) by a slightly modified version,

$$\hat{\Lambda}(S,\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\beta}}) = \sum_{j\in S}\tilde{v}_j^2 + \sum_{j\in\bar{S}}\left(\hat{\theta}_j^2 - \tilde{v}_j^2\right)_+, \tag{3.5}$$

where $(.)_+$ indicates $\max(0,.)$. This modification usually improves the risk estimate (Wasserman, 2005, Chapter 8). We refer to this model section criterion as the WISE criterion.

Now we have defined all the machinary required for our model selection criterion. From all models associated with all index subsets of $\{p+1,\ldots,k\}$, we define

$$\hat{S} = \text{ArgMin}_{S\subseteq\{p+1,\ldots,k\}}\hat{\Lambda}(S,\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\beta}}). \tag{3.6}$$

Thus $\hat{S}$ is such that $\hat{\Lambda}(\hat{S},\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\beta}})$ is smaller than all other $\hat{\Lambda}(S',\hat{\boldsymbol{\theta}},\tilde{\boldsymbol{\beta}})$ ($S'\subseteq\{p+1,\ldots,k\}$). Often model selection is restricted to searching within sequences of nested models, i.e. S is restricted to $\{p+1,\ldots,j\}$, for $j=p+1,\ldots,k$. In this case, the model selection starts with $j=p+1$, and an additional θ_j will be added to the model as long as $2\tilde{v}_j^2 < \hat{\theta}_j^2$. A similar selection rule was proposed by Diggle and Hall (1986) and Tarter (1976). This implementation is referred to as the *order selection* (OS) test; without this restriction the process is referred to as the *subset selection* (SS) test.

3.2.3. The AIC and BIC criteria

For completeness we also define the AIC and BIC criteria. Although they both are also related to loss functions, we present only the final criteria that are used in practice. The AIC for a model S is defined as

$$\text{AIC}_S = -2\log L_S(\hat{\boldsymbol{\theta}}_S) + 2|S|,$$

where $|S|$ is the number of components in S, and $L_S(\hat{\boldsymbol{\theta}}_S)$ is the maximised likelihood function, i.e. the likelihood function evaluated in the MLE $\hat{\boldsymbol{\theta}}_S$. Using the same notation, the BIC is given by

$$\text{BIC}_S = -2\log L_S(\hat{\boldsymbol{\theta}}_S) + |S|\log n.$$

It differs from AIC only in the complexity penalty term where the factor 2 has been replaced by $\log n$. The effect is that BIC penalizes more complex models more heavily, giving preference to simpler models.

From a computational point of view the AIC and BIC criteria are quite demanding, because they require MLE calculation of the parameters in (1.3). Although algorithms exist for finding the (approximate) MLEs (Buckland, 1992; Efron and Tibshirani, 1996), it is often suggested that $-2\log L_S(\hat{\boldsymbol{\theta}}_S)$ be replaced by $\tilde{S}_S$, or even by $\sum_{j\in S}\tilde{V}_j^2/\tilde{\sigma}_j^2$. Kallenberg and Ledwina (1997) argued that this substitution is allowed as the likelihood ratio test statistic ($-2\log L_S(\hat{\boldsymbol{\theta}}_S)$) and the score test statistic ($\hat{S}_S$, based on the MLE of $\boldsymbol{\beta}$) are locally equivalent. Although this local equivalence is theoretically correct, we think that this is against the rationale of using model selection techniques in the present setting. When the null hypothesis is not true, we hope to select a model that is close to the true distribution. This reasoning implies that the model selection criterion must also work well far away from the null hypothesis, and this is in contrast to the local setting under which the substitution with $\hat{S}_S$ is sustained.

3.3. The data-driven test

In this paper we adopt the data-driven testing framework of Claeskens and Hjort (2004), but instead of using MLE and likelihood ratio or score tests, we consider MME and generalised smooth tests.

Let $\tilde{\mathbf{\Sigma}}_S$ denote any $\sqrt{n}$ consistent estimator of $\mathbf{\Sigma}_S$, which is the asymptotic covariance matrix of $\tilde{\boldsymbol{V}}_S$. In a traditional full parametric setting, the consistency and the asymptotic variance are defined under the full parametric null hypohesis. They may also be replaced by the empirical variance estimators of HK, or, for distributions for which MLE and MME do not coincide, by the new variance estimator. The data-driven generalised smooth test is then defined as

$$\tilde{S}_{\hat{S}} = \tilde{\boldsymbol{V}}^t_{\hat{S}} \tilde{\mathbf{\Sigma}}^{-1}_{\hat{S}} \tilde{\boldsymbol{V}}_{\hat{S}},$$

where $\hat{S}$ is the selected model based on WISE (3.6).

In the following theorem we establish the limiting distribution of $\tilde{S}_{\hat{S}}$ under sequences of local alternatives which are defined as

$$g_n(x;\boldsymbol{\theta}_n,\boldsymbol{\beta}) = f(x;\boldsymbol{\beta})\left(1+\sum_{j=p+1}^{\infty}\theta_{jn}h_j(x;\boldsymbol{\beta})\right), \tag{3.7}$$

where $\boldsymbol{\theta}^t_n = (\theta_{p+1n},\ldots)$ and $\theta_{jn} = \frac{1}{\sqrt{n}}t_j$ $(j = p+1,\ldots)$ with non-zero $\boldsymbol{t}^t = (t_{p+1},\ldots)$. The densities are defined for all $\boldsymbol{t}$ so that g_n is a positive function. A similar sequence of local alternatives could have been constructed starting from a density of the form (1.2). The results would be equal because for large n the density (3.7) is a good first order Taylor approximation.

Theorem 3.1. *Suppose the maximal order k in the WISE selection criterion* (3.6) *is finite. Consider the sequences of local alternatives, given by g_n in* (3.7)*, and further assume that $g_n \in \mathcal{G}$. Let $\mathbf{\Sigma}_0$ denote the asymptotic covariance matrix of $\tilde{\boldsymbol{V}}^t = (\tilde{V}_{p+1},\ldots,,\tilde{V}_k)$ under the full parametric null hypothesis, and define $\boldsymbol{Z}^t = (Z_{p+1},\ldots,Z_k)$ as a zero mean multivariate normal variate with covariance matrix equal to $\mathbf{\Sigma}_0$. Let Q denote the index set defined as the minimiser of*

$$\Lambda_\infty(S) = \sum_{j\in\bar{S}}(Z_j+t_j)^2 + tr(\mathbf{\Sigma}_S) - tr(\mathbf{\Sigma}_{\bar{S}}),$$

over $S \subseteq \{p+1,\ldots,k\}$. Define $\boldsymbol{Z}_Q$ as the vector $(Z_j)_{j\in Q}$, $\boldsymbol{t}_Q$ as the vector $(t_j)_{j\in Q}$, and, similarly, $\mathbf{\Sigma}_S$ as the matrix built from the appropriate elements in $\mathbf{\Sigma}$. Then,

$$\tilde{S}_{\hat{S}} \xrightarrow{d} (\boldsymbol{Z}_Q+\boldsymbol{t}_Q)^t\,\mathbf{\Sigma}_Q^{-1}\,(\boldsymbol{Z}_Q+\boldsymbol{t}_Q). \tag{3.8}$$

Proof. The proof is similar to the method of proof used in Section 3.1 of Claeskens and Hjort (2004). The proof consists of two parts. First we prove that $\Lambda_\infty(S)$ and $n\hat{\Lambda}(S,\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\beta}})$ (given in (3.4)) are asymptotically equivalent in distribution. Based on this result it is straightforward to prove the convergence in (3.8).

Since $k<\infty$ is assumed, we only need convergence of finite dimensional random vectors. We use the following result. For the sequences of local alternatives, $\sqrt{n}\tilde{\boldsymbol{\theta}}=\tilde{\boldsymbol{V}}\xrightarrow{d}\boldsymbol{Z}+\boldsymbol{t}$. Since $\tilde{\sigma}_j^2=n\tilde{v}_j^2$ is a consistent estimator of the variance of $\tilde{V}_j=\sqrt{n}\tilde{\theta}_j$, we have that $n\tilde{v}_j^2$ converges in g_n probability to the appropriate diagonal element of $\boldsymbol{\Sigma}_0$. Hence, for all $S\subseteq\{p+1,\ldots,k\}$, as $n\to\infty$,

$$n\hat{\Lambda}(S,\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\beta}})\xrightarrow{d}\sum_{j\in\bar{S}}(Z_j+t_j)^2+\mathrm{tr}(\boldsymbol{\Sigma}_S)-\mathrm{tr}(\boldsymbol{\Sigma}_{\bar{S}})=\Lambda_\infty(S). \tag{3.9}$$

The convergence in (3.8) follows from the simultaneous convergence of all $\tilde{S}_S$ to $W_S\equiv(\boldsymbol{Z}_S+\boldsymbol{t}_S)^t\boldsymbol{\Sigma}_S^{-1}(\boldsymbol{Z}_S+\boldsymbol{t}_S)$ for all fixed index sets $S\subseteq\{p+1,\ldots,k\}$, and from the convergence in (3.9). Write

$$\begin{aligned}\tilde{S}_{\hat{S}} &= \sum_{S\subseteq\{1,\ldots,k\}} T_S\mathrm{I}\left[S=\hat{S}\right]\\ &\xrightarrow{d} \sum_{S\subseteq\{1,\ldots,k\}} W_S\mathrm{I}\left[S=Q\right]=W_Q.\end{aligned}$$

This completes the proof. □

The asymptotic null distribution of the data-driven test statistic follows immediately from this theorem by putting $\boldsymbol{t}=\mathbf{0}$. The next corallary relates the WISE based test with the AIC based data-driven test for distributions for which MME and MLE coincide. The latter is one of the tests considered by Claeskens and Hjort (2004).

Corollary 3.1. *Suppose that all assumptions required in Theorem* 3.1 *hold, and suppose that the hypothesised distribution f belongs to the class of distributions for which MME and MLE coincide. Then, under the sequence of local alternatives* (3.7)*, the WISE and AIC based data-driven tests are equivalent.*

Proof. First note that the Q that minimises $n\hat{\Lambda}(S,\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\beta}})$ is exactly equal to the maximiser of $\sum_{j\in S}n\tilde{\theta}_j^2-2n\sum_{j\in S}\tilde{v}_j^2$. The latter expression converges under the sequences of local alternatives in distribution to

$$\sum_{j\in S}(Z_j+t_j)^2-2\mathrm{tr}(\boldsymbol{\Sigma}_S). \tag{3.10}$$

When MME and MLE coincide, all diagonal elements of $\boldsymbol{\Sigma}_0$ are equal to 1, and thus $\mathrm{tr}(\boldsymbol{\Sigma}_S)=|S|$. □

Finally, note that we expect that our test behaves differently when the conditions of Corollary 3.1 do not hold. For example, when MME and MLE do not coincide, the variances $\boldsymbol{\Sigma}_S$ in the correction term of (3.10) are not necessarily equal to one. We expect thus that in order selection the WISE procedure will select more terms when their corresponding θ parameter estimators have variances smaller than one. The variances listed in Table 2.1 and 2.2 indicate that this may happen under some alternatives, and it does not for others.

3.4. Correcting the improved density estimate

In Section 3.1, where the rationale of the data-driven procedure has been given, we mentioned briefly that densities of the form (3.1) are not necessarally positive functions, and thus they do not always represent proper density functions. Gajek (1986) and Glad, Hjort and Ushakov (2003) proposed correction procedures that can be applied after model selection and parameter estimation. Here we only briefly explain the solution of Gajek (1986), as this fits very nicely in our WISE framework. Without loss of generality we will simplify the notation by not writing the dependence of densities and polynomials on the nuisance parameter β.

Gajek (1986) proposed a simple correction method based on theoretical arguments. Gajek described his method in a general way so that it is applicable to many types of non-bona fide density estimators. As a loss function he considered a weighted ISE,

$$\boldsymbol{\Lambda}(g, g_S; \hat{\boldsymbol{\gamma}}) = \int_{-\infty}^{+\infty} \left\{ g(x; \boldsymbol{\theta}) - g_S(x; \hat{\boldsymbol{\theta}}) \right\}^2 h(x) dx, \tag{3.11}$$

where $h(x)$ is a weight function satisfying

$$\int_{-\infty}^{+\infty} \{1/h(x)\} \, dx < \infty. \tag{3.12}$$

The Gajek-corrected density estimator g_S^c is then defined as

$$g_S^c(x; \hat{\boldsymbol{\theta}}) = f(x) \max \left\{ 0, 1 + \sum_{i \in S} \hat{\theta}_i h_i(x) \right\} - \frac{a}{h(x)}, \tag{3.13}$$

where a is such that $\int_{-\infty}^{+\infty} g_S^c(x; \boldsymbol{\theta}) dx = 1$. Gajek proposed a simple iterative algorithm to find a. He further proved that the expected loss of his corrected density estimator, which is a weighted MISE, is not larger than the weighted MISE of the uncorrected density estimator. Moreover, if the uncorrected estimator is consistent, then so is the corrected estimator. These are very important results, but the weight function h has an important role, particularly since it determines the meaningfulness of the expected loss function. Note that the WISE loss function corresponds with Gajek's loss function if $h(x) = 1/f(x)$, which clearly satisfies the condition of (3.12) for $\int_{-\infty}^{+\infty} \{1/h(x)\} dx = \int_{-\infty}^{+\infty} f(x) dx \equiv 1 < \infty$. The corrected density estimator of (3.13) now becomes

$$g_S^c(x; \hat{\boldsymbol{\theta}}) = f(x) \max \left\{ 0, 1 - a + \sum_{i \in S} \hat{\theta}_i h_i(x) \right\}, \tag{3.14}$$

where a is such that $\int_{-\infty}^{+\infty} g_S^c(x; \boldsymbol{\theta}, \beta) dx = 1$. Note that a appears in the linear part $1 - a + \sum_{i \in S} \hat{\theta}_i h_i(x)$ of which only the positive part contributes to g_S^c.

3.5. A simulation study

In this simulation study we investigate some properties of the data-driven tests based on the AIC, BIC and WISE selection criteria. Since we have introduced the WISE criterion so as to have a good improved density estimate after rejecting the null hypothesis, the bias and the variance of the density estimates are also estimated in this study. As the primary aim is still hypothesis testing, we also estimate the powers of the tests under the various alternatives. For assessing the bias we computed the WISE loss function (3.2) evaluated at the alternative g, with nuisance parameters fixed as in the distribution used in the simulations, and with $g_{\hat{S}}$ equal to the estimated improved density as selected by the selection criteria when this the corresponding test rejected the null hypothesis. Otherwise $f(x;\tilde{\boldsymbol{\beta}})$ was considered instead. We also included the unweighted ISE loss function, which is basically a simple least squares loss function. As a measure of the variance of the improved density estimator, we first computed the pointwise variances of the improved density estimates at 100 equally spaced points in the support over the $10,000$ Monte Carlo runs, and subsequently we have averaged these 100 variances, resulting in an overall measure for the variance. All tests are performed at the 5% level of significance. Since we only tested location-scale families, the null distributions were approximated by $100,000$ simulation runs under the null hypothesis. We present results for testing for the normal and the logistic distributions. As alternatives we considered densities (1.3) of order 5, indexed by the parameters θ_3, θ_4 and θ_5. All data-driven smooth tests are based on subset selection with horizon $S_h = \{3,4,5,6\}$, i.e. the maximal order is one higher than the maximal order of the alternatives considered.

The results for testing for the normal and logistic distributions are presented in Tables 3.1, 3.2, 3.3, and 3.4. These results show that for both the normal and the logstic case, the WISE based data-driven test has for most alternatives the smallest WISE, as expected. Also in terms of the unweighted ISE the new data-driven test has overall the best performance. The average bias of the improved density estimator based on the WISE data-driven test, seems again often to be slightly better as compared to the other two testing procedures. Note that particularly the BIC based data-driven test does worse than the other two tests in terms of all estimation quality measures considered. One possible explanation is that the penalty of the BIC criterion penalises more heavily for more complex models, resulting in underfitting.

Since the primary aim is still hypothesis testing, the powers are very important too. For both the logistic and the normal case, we conclude that the WISE data-driven test has very often the largest power among all three tests, and the BIC based test is the least powerful.

4. Example

Bain, Easterman and Engelhardt (1973) present data of a life-test of incandescent lamps. They consider the logistic distribution as a possible life-testing model. The observed failure times are 785, 855, 905, 918, 919, 920, 929, 936, 948 and 950. The same data set has also been used by Engelhardt (1975), who assumed that the data are well described by a logistic distribution. In this section, we formally test the null hypothesis that the data are sampled from a logistic distribution. Since nothing is known a priori about the variance of the distribution, we consider the two-parameter logistic distribution. For all smooth tests, bootstrap p-values are computed based on $10,000$ bootstrap runs.

Table 3.1. Estimated WISE and ISE measures for the three data-driven tests for normality (WISE, AIC and BIC based) under alternatives (1.3) indexed by θ_3, θ_4 and θ_5.

θ_3	θ_4	θ_5	WISE			ISE		
			WISE	AIC	BIC	WISE	AIC	BIC
0.3	0.0	0.0	0.0258	0.0312	0.0279	0.00032	0.00043	0.00036
0.0	0.3	0.0	0.0531	0.0547	0.0544	0.00112	0.00112	0.00114
0.0	0.0	0.3	0.0350	0.0389	0.0389	0.00080	0.00092	0.00094
0.3	0.3	0.0	0.0538	0.0535	0.0542	0.00103	0.00111	0.00108
0.3	0.0	0.3	0.0276	0.0269	0.0277	0.00020	0.00021	0.00021
0.0	0.3	0.3	0.0794	0.0860	0.0833	0.00179	0.00195	0.00190
0.3	0.3	0.3	0.0703	0.0734	0.0723	0.00120	0.00150	0.00136
0.3	0.3	-0.3	0.0988	0.0897	0.0998	0.00227	0.00220	0.00231
-0.3	0.0	0.0	0.0256	0.0307	0.0269	0.00033	0.00043	0.00035
0.0	-0.3	0.0	0.0389	0.0388	0.0381	0.00115	0.00116	0.00115
0.0	0.0	-0.3	0.0344	0.0378	0.0394	0.00077	0.00085	0.00092

Table 3.2. Estimated average biases, average variances and powers for the three data-driven tests for normality (WISE, AIC and BIC based) under alternatives (1.3) indexed by θ_3, θ_4 and θ_5.

θ_3	θ_4	θ_5	bias			variance			power (%)		
			WISE	AIC	BIC	WISE	AIC	BIC	WISE	AIC	BIC
0.3	0.0	0.0	-0.00024	-0.00025	-0.00024	0.00030	0.00031	0.00032	54.2	43.0	49.8
0.0	0.3	0.0	0.00139	0.00142	0.00143	0.00033	0.00035	0.00033	60.5	61.5	58.6
0.0	0.0	0.3	0.00001	0.00001	0.00000	0.00011	0.00011	0.00010	32.1	27.5	27.2
0.3	0.3	0.0	0.00166	0.00172	0.00169	0.00045	0.00038	0.00042	85.9	84.3	84.6
0.3	0.0	0.3	-0.00029	-0.00028	-0.00029	0.00025	0.00024	0.00027	27.5	30.8	32.8
0.0	0.3	0.3	0.00133	0.00139	0.00135	0.00034	0.00036	0.00034	76.2	73.9	71.4
0.3	0.3	0.3	0.00142	0.00156	0.00151	0.00052	0.00038	0.00045	70.2	73.2	69.2
0.3	0.3	-0.3	0.00094	0.00102	0.00098	0.00056	0.00039	0.00055	97.7	97.6	97.8
-0.3	0.0	0.0	-0.00025	-0.00026	-0.00025	0.00029	0.00030	0.00030	54.2	43.7	51.3
0.0	-0.3	0.0	0.00001	0.00001	0.00001	0.00002	0.00002	0.00000	24.9	27.3	9.7
0.0	0.0	-0.3	0.00003	0.00004	0.00005	0.00011	0.00014	0.00012	37.0	41.5	36.6

The p-values of the Anderson-Darling, Watson and Cramér-von Mises tests are 0.024, 0.046 and 0.046, respectively. Thus, at the $\alpha = 0.05$ level of significance, all tests suggest that the logistic distribution does not describe the data well. The order $k = 4$ generalised smooth test, using MME, gives a p-value of 0.027, and the individual third and fourth order components give p-values of 0.025 and 0.949, respectively. Again the null hypothesis is rejected, but as the component tests are not diagnostic, we may not relate them to moment deviations. For completeness we also give the p-values of the rescaled components, using the appropriate new variance estimator. The third and fourth order rescaled component tests

Table 3.3. Estimated WISE and ISE measures for the three data-driven tests for the logistic distribution (WISE, AIC and BIC based) under alternatives (1.3) indexed by θ_3, θ_4 and θ_5.

θ_3	θ_4	θ_5	WISE			ISE		
			WISE	AIC	BIC	WISE	AIC	BIC
0.5	0.0	0.0	0.04456	0.04465	0.04511	0.00099	0.00099	0.00100
0.0	0.5	0.0	0.03141	0.03141	0.03141	0.00079	0.00079	0.00079
0.0	0.0	0.5	0.02575	0.02615	0.02668	0.00065	0.00065	0.00067
0.5	0.5	0.0	0.07042	0.07042	0.07035	0.00166	0.00166	0.00166
0.5	0.0	0.5	0.00457	0.00450	0.00452	0.00006	0.00006	0.00006
0.0	0.5	0.5	0.05599	0.05699	0.05651	0.00141	0.00143	0.00142
0.5	0.5	0.5	0.03429	0.03448	0.03435	0.00082	0.00083	0.00083
0.5	0.5	-0.5	0.16308	0.16299	0.16320	0.00390	0.00390	0.00390
-0.5	0.0	0.0	0.04432	0.04451	0.04484	0.00098	0.00099	0.00099
0.0	-0.5	0.0	0.03111	0.03111	0.03110	0.00079	0.00079	0.00079
0.0	0.0	-0.5	0.02564	0.02606	0.02672	0.00064	0.00065	0.00067

Table 3.4. Estimated average biases, average variances and powers for the three data-driven tests for the logistic distribution (WISE, AIC and BIC based) under alternatives (1.3) indexed by θ_3, θ_4 and θ_5.

θ_3	θ_4	θ_5	bias			variance			power (%)		
			WISE	AIC	BIC	WISE	AIC	BIC	WISE	AIC	BIC
0.5	0.0	0.0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	70.2	68.1	65.3
0.0	0.5	0.0	-0.00584	-0.00583	-0.00583	0.00001	0.00001	0.00001	71.1	71.7	71.2
0.0	0.0	0.5	0.00002	0.00000	0.00001	0.00000	0.00000	0.00000	49.8	44.2	37.1
0.5	0.5	0.0	-0.00583	-0.00583	-0.00583	0.00000	0.00000	0.00000	90.6	90.6	91.7
0.5	0.0	0.5	0.00001	0.00000	0.00001	0.00000	0.00000	0.00000	36.4	36.7	37.8
0.0	0.5	0.5	-0.00579	-0.00583	-0.00580	0.00001	0.00001	0.00001	89.0	89.0	88.3
0.5	0.5	0.5	-0.00579	-0.00583	-0.00581	0.00001	0.00000	0.00001	76.4	77.2	76.6
0.5	0.5	-0.5	-0.00582	-0.00583	-0.00583	0.00000	0.00000	0.00000	99.4	99.5	99.4
-0.5	0.0	0.0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	72.4	68.8	66.7
0.0	-0.5	0.0	0.00583	0.00583	0.00583	0.00000	0.00000	0.00000	0.6	0.4	0.2
0.0	0.0	-0.5	-0.00002	-0.00001	-0.00002	0.00000	0.00000	0.00000	52.0	48.5	39.8

result in p-values 0.136 and 0.966, respectively. As compared to their unscaled versions, the significance of the third order terms has disappeared, but based on the conclusion of the simulation study of Section 2.3, we actually do not recommend using the results of these tests. Instead we continue with the data-driven test, using the new WISE criterion. We have chosen subset selection from the horizon inlcuding all terms up to order 5. The p-value of the data-driven test is 0.014, selecting the third and the fifth order terms. Figure 4.1 shows the histogram of the data, the fitted hypothesised and improved densities. The improved density estimate is skewed and it has a heavier tail as compared to the hypothesised logistic distribution.

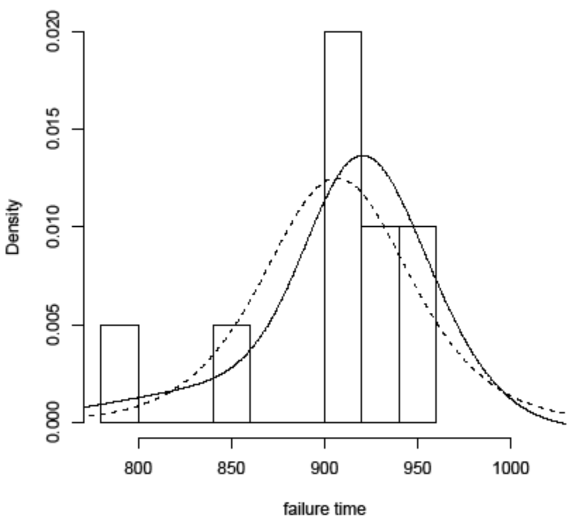

Figure 4.1. The histogram of the incandescent lamps data, the estimated hypothesised (dashed line) and the improved (solid line) densities

Appendix A. The Logistic Distribution

The density function of the two-parameter logistic distribution is given by

$$f(x;\boldsymbol{\beta}) = \frac{\exp(-(x-\mu)/\sigma)}{\sigma\,(1+\exp(-(x-\mu)/\sigma))^2} \quad \text{for} \quad -\infty < x < +\infty,$$

where $\boldsymbol{\beta}^t = (\mu, \sigma)$, and μ and σ are the location and scale parameter, respectively. When $\mu = 0$ and $\sigma = 1$, the distribution is referred to as the standard logistic distribution, which has density function $f_0(.)$. When only one of $\sigma = 1$ or $\mu = 0$ is assumed, one-parameter logistic distributions are obtained.

The MME is found by solving $V_1(\boldsymbol{\beta}) = V_2(\boldsymbol{\beta}) = 0$, resulting in

$$\tilde{\mu} = \bar{X} \text{ and } \tilde{\sigma} = \sqrt{3S^2}/\pi,$$

where S^2 is the sample variance. For this particular distribution the MME is not equal to the MLE. The use of MME rather than MLEs may result in some loss in efficiency, but, in this particular case, Lehmann (1999) has shown that the sample mean is an efficient estimator of μ.

Since the logistic distribution is a location-scale family, it is sufficient to know the orthonormal polynomials for the standard logistic distribution in terms of $z = (x-\mu)/\sigma$. The first five orthonormal polynomials for the standard logistic distribution are given by

$$h_0(z) = 1\,, \qquad h_1(z) = \frac{\sqrt{3}}{\pi} z$$

$$h_2(z) = \frac{3\sqrt{5}}{4\pi^2}(z^2 - \pi^2/3)$$

$$\begin{aligned} h_3(x) &= \frac{5\sqrt{7}}{12\pi^3}\left(x^3 - \frac{7}{5}\pi^2 x\right) \quad \text{and} \\ h_4(x) &= \frac{35}{64\pi^4}\left(x^4 - \frac{26}{7}\pi^2 x^2 + \frac{27}{35}\pi^4\right). \end{aligned}$$

The covariance matrix of $\tilde{\boldsymbol{V}}^t = (\tilde{V}_3, \tilde{V}_4)$ equals

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.064815 & 0 \\ 0 & 1.088889 \end{bmatrix}.$$

The partial derivates $\frac{\partial h_r}{\partial \boldsymbol{\beta}}(x;\boldsymbol{\beta})$, which are required for the computation of the new variance estimator, are listed next:

$$\begin{aligned}
\frac{\partial h_2}{\partial \mu}(x;\boldsymbol{\beta}) &= -6z/(\sqrt{3.2}\pi^2\sigma) \\
\frac{\partial h_3}{\partial \mu}(x;\boldsymbol{\beta}) &= -5\sqrt{7}\left(3z^2 - 4.2\frac{\pi^2}{3}\right)/(12\pi^3\sigma) \\
\frac{\partial h_4}{\partial \mu}(x;\boldsymbol{\beta}) &= -35\left(4z^3 - \frac{26}{7}\pi^2 2z\right)/(64\pi^4\sigma) \\
\frac{\partial h_5}{\partial \mu}(x;\boldsymbol{\beta}) &= -(0.0007112402z^4 - 0.0545973510z^2 + 0.4475789427)/\sigma \\
\frac{\partial h_6}{\partial \mu}(x;\boldsymbol{\beta}) &= -(7.520244\text{E-}5z^5 - 0.01045853z^3 + 0.2190961z)/\sigma \\
\frac{\partial h_7}{\partial \mu}(x;\boldsymbol{\beta}) &= -(6.821868\text{E-}6z^6 - 1.558928\text{E-}3z^4 + 0.06671607z^2 \\
&\quad -0.4159354)/\sigma \\
\frac{\partial h_2}{\partial \sigma}(x;\boldsymbol{\beta}) &= -6z^2/(\sqrt{3.2}\pi^2\sigma) \\
\frac{\partial h_3}{\partial \sigma}(x;\boldsymbol{\beta}) &= -5\sqrt{7}z\left(3z^2 - 4.2\frac{\pi^2}{3}\right)/(12\pi^3\sigma) \\
\frac{\partial h_4}{\partial \sigma}(x;\boldsymbol{\beta}) &= -35z\left(4z^3 - \frac{26}{7}\pi^2 2z\right)/(64\pi^4\sigma) \\
\frac{\partial h_5}{\partial \sigma}(x;\boldsymbol{\beta}) &= z\frac{\partial h_5}{\partial \mu}(x;\boldsymbol{\beta}) \text{ and } \frac{\partial h_6}{\partial \sigma}(x;\boldsymbol{\beta}) = z\frac{\partial h_6}{\partial \mu}(x;\boldsymbol{\beta}) \\
\frac{\partial h_7}{\partial \sigma}(x;\boldsymbol{\beta}) &= z\frac{\partial h_7}{\partial \mu}(x;\boldsymbol{\beta}),
\end{aligned}$$

where $\boldsymbol{\beta}^t = (\mu, \sigma)$ and $z = \frac{x-\mu}{\sigma}$.

Appendix B. The Extreme Value Distribution

The density function of the two-parameter extreme-value distribution is given by

$$f(x;\boldsymbol{\beta}) = \exp\left(-\frac{x-\mu}{\sigma} - \exp\left(-\frac{x-\mu}{\sigma}\right)\right) \text{ for } -\infty < x < +\infty,$$

where $\boldsymbol{\beta} = (\mu, \sigma)^t$, and μ and σ are the location and scale parameters, respectively.

The MME of $\boldsymbol{\beta}$ is the solution to $\tilde{V}_1 = \tilde{V}_2 = 0$, resulting in $\tilde{\mu} = \bar{X} - \frac{\sqrt{6}\gamma}{\pi}\tilde{\sigma}$, where γ is Euler's constant, and in which $\tilde{\sigma} = \frac{\sqrt{6}}{\pi}S$.

Next, the first five orthonormal polynomials for the standard extreme value distribution are given. Let $z = \frac{x-\mu}{\sigma} - \gamma$.

$$\begin{aligned} h_0(z) &= 1\,, \quad h_1(z) = \frac{\sqrt{6}}{\pi}z \quad \text{and} \\ h_2(z) &= 30\frac{z^2 - \frac{12\zeta(3)}{\pi^2}z - \frac{\pi^2}{6}}{\sqrt{110\pi^4 - \frac{1}{\pi^2}21600\zeta(3)^2}} \\ h_3(x) &= 0.1060499473z^3 - 0.4944037009z^2 - 0.219420091z \\ &\quad + 0.5583053486 \quad \text{and} \\ h_4(x) &= 0.02493263957z^4 - 0.2416834738z^3 + 0.2690771426z^2 \\ &\quad + 0.7769092008z - 0.2258795, \end{aligned}$$

where $\zeta(3)$ is Riemann's zeta function evaluated at 3, which is also known as Apéry's constant, which is approximately 1.20205690.

The variance-covariance matrix of $\tilde{\boldsymbol{V}}^t = (\tilde{V}_3, \tilde{V}_4)$ equals

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.585897122 & -0.4121139966 \\ -0.4121139966 & 1.291902227 \end{bmatrix}.$$

The partial derivates $\frac{\partial h_r}{\partial \boldsymbol{\beta}}(x;\boldsymbol{\beta})$, which are required for the computation of the new variance estimator, are listed next:

$$\begin{aligned} \frac{\partial h_3}{\partial a}(x;\boldsymbol{\beta}) &= -(0.1060499473(z-\gamma)^2 - 0.4944037009(z-\gamma) \\ &\quad -0.219420091)/b \\ \frac{\partial h_4}{\partial a}(x;\boldsymbol{\beta}) &= -(0.02493263979(z-\gamma)^3 - 0.2416834756(z-\gamma)^2 \\ &\quad +0.269077145(z-\gamma) + 0.7769092062)/b \\ \frac{\partial h_5}{\partial a}(x;\boldsymbol{\beta}) &= -(-0.765066730 - 0.462335263z + 0.480821661z^2 \\ &\quad -0.092490112z^3 + 0.004746706z^4)/b \end{aligned}$$

$$\begin{aligned}
\frac{\partial h_6}{\partial a}(x;\boldsymbol{\beta}) &= -(1.0327007209 - 0.0480989854z - 0.5440920752z^2 \\
&\quad +0.1947556972z^3 - 0.0219166650z^4 + 0.0007592173z^5)/b \\
\frac{\partial h_7}{\partial a}(x;\boldsymbol{\beta}) &= -(-0.9614699721 + 0.6159510209z + 0.4265597373z^2 \\
&\quad -0.2987314444z^3 + 0.0569116488z^4 - 0.0042051687z^5 \\
&\quad +0.0001046942z^6)/b \\
\frac{\partial h_r}{\partial b}(x;\boldsymbol{\beta}) &= z\frac{\partial h_r}{\partial a}(x;\boldsymbol{\beta}) \qquad \text{for} \quad r = 3,\ldots,7,
\end{aligned}$$

where $\boldsymbol{\beta}^t = (\mu,\sigma)$, $z = \frac{x-\mu}{\sigma}$ and $\gamma \equiv 0.5772156649$.

Acknowledgements

The first and the last author gratefully acknowledge the financial support from the IAP research network P6/03 of the Belgian Government (Belgian Science Policy).

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Inference Theory*, Petrov, B., Csàki, F. (editors), 267–281. Akadémiai Kiadó, Budapest.

Akaike, H., 1974. A new look at statistical model identification. *I.E.E.E. Trans. Auto. Control*, 19, 716–723.

Anderson, G., de Figueiredo, R., 1980. An adaptive orthogonal-series estimator for probability density functions. *Annals of Statistics*, 8, 347–376.

Bain, L., Easterman, J., Engelhardt, M., 1973. A study of life-testing models and statistical analyses for the logistic distribution. Technical Report ARL-73-0009, Aerospace Research Laboratories, Wright Patterson AFB.

Baringhaus, L., Henze, N., 1992. Limit distributions for Mardia measure of multivariate skewness. *Annals of Statistics*, 20, 1889–1902.

Barton D., 1953. On Neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. *Skandinavisk Aktuarietidskrift*, 36, 24–63.

Bickel, P., Ritov, Y., Stoker, T., 2006. Tailor-made tests of goodness of fit to semiparametric hypotheses. *Annals of Statistics*, 34, 721–741.

Boos, D., 1992. On generalized score tests. *The American Statistician*, 46, 327–333.

Buckland, S., 1992. Fitting density functions with polynomials. *Applied Statistics*, 41, 63–76.

Cencov, N., 1962. Evaluation of an unknown distribution density from observations. *Soviet. Math.*, 3, 1559–1562.

Claeskens, G., Hjort, N., 2004. Goodness of fit via non-parametric likelihood ratios. *Scandinavian Journal of Statistics*, 31, 487–513.

Clutton-Brock, M., 1990. Density estimation using exponentials of orthogonal series. *Journal of the American Statistical Association*, 85, 760–764.

Diggle, P., Hall, P., 1986. The selection of terms in an orthogonal series density estimator. *Journal of the American Statistical Association*, 81, 230–233.

Efron, B., Tibshirani, R., 1996. Using specially designed exponential families for density estimation. *Annals of Statistics*, 24, 2431–2461.

Emerson, P., 1968. Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics*, 24, 695–701.

Engelhardt, M., 1975. Simple linear estimation of the parameters of the logistic distribution from a complete or censored sample. *Journal of the American Statistical Association*, 70, 899–902.

Eubank, R., LaRiccia, V., Rosenstein, R., 1987. Test statistics derived as components of Pearson's phi-squared distance measure. *Journal of the American Statistical Association*, 82, 816–825.

Gajek, G., 1986. On improving density estimators which are not bona fide functions. *Annals of Statistics*, 14, 1612–1618.

Glad, I., Hjort, N., Ushakov, N., 2003. Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, 30, 415–427.

Hall, W., Mathiason, D., 1990. On large-sample estimation and testing in parametric models. *International Statistical Review*, 58, 77–97.

Henze, N., 1997. Do components of smooth tests of fit have diagnostic properties? *Metrika*, 45, 121–130.

Henze, N., Klar, B., 1996. Properly rescaled components of smooth tests of fit are diagnostic. *Australian Journal of Statistics*, 38, 61–74.

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

Kallenberg, W., Ledwina, T., 1995. Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Annals of Statistics*, 23, 1594–1608.

Kallenberg, W., Ledwina, T., 1997. Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association*, 92, 1094–1104.

Kallenberg, W., Ledwina, T., Rafajlowicz, E., 1997. Testing bivariate independence and normality. *Sankhyā, Series A*, 59, 42–59.

Klar, B., 2000. Diagnostic smooth tests of fit. *Metrika*, 52, 237–252.

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

Lehmann, E., 1999. *Elements of Large-Sample Theory*. Springer, New York.

Mardia, K., Kent, J., 1991. Rao score tests for goodness-of-fit and independence. *Biometrika*, 78, 355–363.

Rayner J., Best D., 1989. *Smooth Tests of Goodness-of-Fit*. Oxford University Press, New York.

Rayner, J., Best, D., Mathews, K., 1995. Interpreting the skewness coefficient. *Communications in Statistics - Theory and Methods*, 24, 593–600.

Rayner, J., Best, D., Thas, O., 2009a. Generalised smooth tests of goodness of fit. *Journal of Statistical Theory and Practice*, 3(3), 665–679. Accompanying paper.

Rayner, J., Thas, O., Best, D., 2009b. *Smooth Tests of Goodness of Fit*. Wiley, New York, USA.

Rayner, J., Thas, O., De Boeck, B., 2008. A generalised Emerson recurrence relation. *Australian and New Zealand Journal of Statistics*, 50, 235–240.

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

Stuart, A., Ord, J., 1994. *Kendall's Advanced Theory of Statistics*. Arnold / Halsted, London.

Tarter, M., 1976. An introduction to the implementation and theory of nonparametric density estimation. *The American Statistician*, 30, 105–112.

van der Vaart, A., 1998. *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wasserman, L., 2005. *All of Nonparametric Statistics*. Springer.