Frequentist-Bayes Lack-of-Fit Tests Based on Laplace Approximations

Jeffrey D. Hart*

Texas A&M University

Abstract

The null hypothesis that all of a function's Fourier coefficients are 0 is tested in frequentist fashion using as test statistic a Laplace approximation to the posterior probability of the null hypothesis. Testing whether or not a regression function has a prescribed linear form is one application of such a test. In contrast to BIC, the Laplace approximation depends on prior probabilities, and hence allows the investigator to tailor the test to particular kinds of alternative regression functions. On the other hand, using diffuse priors produces new omnibus lack-of-fit statistics.

The new omnibus test statistics are weighted sums of *exponentiated* squared (and normalized) Fourier coefficients, where the weights depend on prior probabilities. Exponentiation of the Fourier components leads to tests that can be exceptionally powerful against high frequency alternatives. Evidence to this effect is provided by a comprehensive simulation study, in which one new test that had good power at high frequencies also performed comparably to some other well-known omnibus tests at *low* frequency alternatives.

Keywords: Asymptotic distribution, BIC, Laplace approximation, Local alter-

natives, Microarrays, Nonparametric lack of-fit-tests, Score tests, Orthogonal series, Tidal heights data.

*Jeffrey D. Hart, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.

1 Introduction

Statistics that are weighted sums of squared Fourier coefficients have played a prominent role in the literature on nonparametric lack-of-fit tests. Such statistics arise naturally from the use of Gaussian likelihood ratio or score tests or from basing a test on squared error discrepancy. For example, the regression analog of the classic Neyman (1937) smooth test can be derived as a score test in a regression model with Gaussian errors. The reader is referred to Hart (1997) and Claeskens and Hjort (2004) for discussion of and some key references on the subjects of lack-of-fit and goodness-of-fit, respectively.

In this paper, we derive, from Bayesian principles, a lack-of-fit statistic that is a weighted sum of *exponentiated* squared Fourier coefficients. This statistic is subsequently used in frequentist fashion to test the fit of a linear regression model. It will be argued that one of our new tests has power properties that are highly competitive with those of a popular class of nonparametric, or omnibus, lack-of-fit tests. A fascinating aspect of this point becomes apparent from considering properties of the latter tests, which fall into two distinct categories: weighted sums of squared Fourier coefficients with nonrandom and random weights, respectively. Omnibus tests with nonrandom weights have largely been dismissed in favor of ones with (the right type of) random weights because the former tend to have good power only against certain types of alternatives; see, e.g., Eubank and Hart (1993), Eubank

(2000), and Inglot and Ledwina (2006) for a discussion of these points. The prime example of a statistic with nonrandom weights, and a favorite whipping boy, is the cusum statistic, which is a regression analog of the Cramér-von Mises goodness-of-fit statistic. To a good approximation, this statistic is

$$C_n = \frac{1}{\hat{\sigma}^2} \sum_{j=1}^{n} \frac{n\hat{\phi}_j^2}{j^2},$$

where $\hat{\phi}_1, \ldots, \hat{\phi}_n$ are sample Fourier coefficients arranged in order of increasing frequency and $\hat{\sigma}$ is a scale estimate. The popular explanation for the power deficiencies of C_n is that the higher frequency coefficients are unduly downweighted, meaning that C_n only has good power against low frequency alternatives. Our results offer an alternative explanation: C_n uses a relatively ineffective function of each Fourier coefficient. Our Bayesian point of view yields statistics of the form

$$B_n = \sum_{j=1}^{n} \rho_j \exp\left(\frac{n\hat{\phi}_j^2}{2\hat{\sigma}^2}\right),$$

where the weights $\rho_1, \rho_2, \ldots$ are related to prior probabilities. Remarkably, the special case of B_n with $\rho_j = j^{-2}$, $j = 1, \ldots, n$, has better overall power in a comprehensive simulation study than either C_n *or* a popular data-driven score test. Stated another way, our new test appears to have power comparable to that of a certain adaptive test *even though it is not itself adaptive.*

Recently there has been considerable interest in what some have termed "hybrid Bayes-frequentist" methods, i.e., methods that combine Bayesian and frequentist thinking; see, e.g., Bayarri and Berger (2004), Conrad, Botner, Hallgren and Pérez de los Heros (2003), Aerts, Claeskens and Hart (2004) and Chang and Chow (2005). Our proposed tests are examples of such hybrids, as they are derived from Bayesian principles but used in frequentist fashion. We shall refer to such tests as *frequentist-Bayes*. Good (1957) proposed a frequentist-Bayes test based on a Bayes factor.

The article of Aerts, Claeskens, and Hart (2004) appears to be the first to propose frequentist-Bayes lack-of-fit tests based on posterior probabilities, which is precisely the subject of the current paper. Our approach differs from that of Aerts, Claeskens, and Hart (2004) in two respects. First, we use the method of Laplace to approximate posterior probabilities whereas Aerts, Claeskens, and Hart (2004) use BIC. This difference turns out to have important implications in terms of power. The BIC analog of B_n takes the form $B_{n,BIC} = \sum_{j=1}^{n} \exp\left[n\hat{\phi}_j^2/(2\hat{\sigma}^2)\right]$. Aerts, Claeskens and Hart (2004) conclude that, in an overall sense, the power of this statistic is "rather poor." In contrast, the conclusion of our simulation study is that the overall power of B_n with $\rho_j = j^{-2}$ is quite good. This result shows that well-known deficiencies (Kass and Wasserman 1995) of BIC in the Bayesian world are mirrored by ones in the frequentist world. The second difference between our approach and that of Aerts, Claeskens and Hart (2004) is that we show explicitly how B_n arises naturally from a general Bayesian model for a function. In contrast, the derivation of $B_{n,BIC}$ in Aerts, Claeskens and Hart (2004) is based on so-called *singleton* models, i.e., ones in which one and only one Fourier coefficient is nonzero. Such models would rarely be used in function estimation, and hence provide a less appealing motivation for either $B_{n,BIC}$ or B_n than does our approach.

Our frequentist-Bayes tests are demonstrated to have good overall power properties. However, it is not the intent of this paper to argue that our tests are uniformly superior to any that have been previously proposed. Indeed, Janssen (2000) shows that, generally speaking, *any* nonparametric test has power that is flat on balls of alternatives except for those coming from a particular finite dimensional subspace. For this reason, no one omnibus test will ever be superior (in terms of power) to every other omnibus test. New omnibus tests should thus be judged in terms of their "overall" power properties and other factors, such as simplicity and how widely they

can be applied. We make the case that the tests proposed in this paper fare well on both these counts.

The literature on lack-of-fit has burgeoned in the last ten years. Many of the important references on the subject prior to 1997 may be found in the monograph Hart (1997). In addition to aforementioned articles, important work post-1995 includes, but is not limited to, that of Spokoiny (1996), Stute (1997), Dette and Munk (1998), Aerts, Claeskens and Hart (1999), Dette (1999), Aerts, Claeskens and Hart (2000), Dümbgen and Spokoiny (2001), Fan and Huang (2001), Fan, Zhang and Zhang (2001), Horowitz and Spokoiny (2001), Baraud, Huet and Laurent (2003), Guerre and Lavergne (2005), and Bickel, Ritov and Stoker (2006). All the approaches in these articles are frequentist in nature. Verdinelli and Wasserman (1998) proposed a purely Bayesian nonparametric goodness-of-fit test. Finally, worth special mention due to their fundamental nature are adaptive versions of the Neyman smooth test, which were introduced by Ledwina (1994) in the goodness-of-fit context. In this work Ledwina proposed that the Schwarz criterion, i.e., BIC, be used to choose the number of components in a Neyman smooth statistic. Inglot and Ledwina (2006) studied the analog of such tests in a regression setting. Kuchibhatla and Hart (1996) also investigated adaptive Neyman smooth tests in a regression context, but using Mallows' criterion (Mallows 1973) instead of BIC.

The next section introduces the model on which our results are based, and Section 3 derives a class of test statistics based on posterior probabilities. Cusum and score tests are contrasted with our approach in Section 4. Asymptotic distribution theory for the new tests is presented in Section 5, where it is shown that they can detect alternatives converging to H_0 at the rate $1/\sqrt{n}$. Section 6 addresses the problem of choosing a prior distribution over the alternatives to H_0. A simulation study is the subject of Section 7, and an example involving data from a microarray

analysis is provided in Section 8. The paper ends with some concluding remarks in Section 9.

2 A canonical model and the inference problem

We consider problems where it is of interest to test the null hypothesis that a function r is identical to 0. It is assumed that r is characterized by Fourier coefficients $\phi_1, \phi_2, \ldots$ and that the null hypothesis is equivalent to

$$H_0 : \phi_1 = \phi_2 = \cdots = 0. \tag{1}$$

The observed data are sample Fourier coefficients $\hat{\phi}_1, \ldots, \hat{\phi}_n$ that estimate $\phi_1, \ldots, \phi_n$, respectively. These data satisfy

*A*1. $\hat{\phi}_1, \ldots, \hat{\phi}_n$ are independent,

*A*2. $\hat{\phi}_j \sim N(\phi_{jn}, \sigma^2/n)$, $j = 1, \ldots, n$, and

*A*3. $\phi_{jn} = 0$, $j = 1, \ldots, n$, under the null hypothesis.

The scale parameter σ is allowed to be unknown. We will focus on nonparametric tests of H_0, i.e., tests that are consistent against virtually any alternative to (1) as $n \to \infty$.

To a good approximation, *A*1-*A*3 hold in a variety of problems, and are exact in the following canonical regression setting. Suppose we observe $Y_1, \ldots, Y_{n+p+1}$ from the model

$$Y_j = \mu(\boldsymbol{x}_j) + \epsilon_j, \quad j = 1, \ldots, n + p + 1, \tag{2}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n+p+1}$ are fixed, d-dimensional design points, and the unobserved errors $\epsilon_1, \ldots, \epsilon_{n+p+1}$ are independent and identically distributed as $N(0, \sigma^2)$. It is of

interest to test the null hypothesis

$$\mu(\boldsymbol{x}) = \theta_0 + \sum_{j=1}^{p} \theta_j \mu_j(\boldsymbol{x}) \equiv \mu_{\boldsymbol{\theta}}(\boldsymbol{x}), \tag{3}$$

where $\theta_0, \ldots, \theta_p$ are unknown parameters and $\mu_1, \ldots, \mu_p$ are known functions. In this case, $r \equiv \mu - \mu_{\boldsymbol{\theta}_0}$, where $\mu_{\boldsymbol{\theta}_0}$ is the best null approximation of μ. Starting from an arbitrary set of basis functions, one may use the Gram-Schmidt procedure to define $u_{1n}, \ldots, u_{nn}$ and

$$\hat{\phi}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i u_{jn}(\boldsymbol{x}_i), \quad j = 1, \ldots, n, \tag{4}$$

such that $\hat{\phi}_1, \ldots, \hat{\phi}_n$ satisfy $A1$-$A3$. Examples of basis functions that may be used are trigonometric functions, polynomials and wavelets.

Now suppose that model (2) holds but the errors $\epsilon_1, \ldots, \epsilon_{n+p+1}$ are merely assumed to be i.i.d. with mean 0 and finite variance σ^2. Then defining sample Fourier coefficients as in (4) preserves $A3$, and $A1$ and $A2$ continue to hold in an approximate sense as $n \to \infty$. It is also worth noting that these Fourier coefficients are uncorrelated, albeit not necessarily independent.

3 Derivation of test statistics

Let $\widehat{\boldsymbol{\phi}}_n$ denote the vector $(\hat{\phi}_1, \ldots, \hat{\phi}_n)$ of sample Fourier coefficients. To test the null hypothesis (1), we shall first propose a prior distribution for σ and the Fourier coefficients $\phi_1, \ldots, \phi_n$, and then compute the posterior probability, $\pi_0(\widehat{\boldsymbol{\phi}}_n)$, of H_0. One would be inclined to reject H_0 when the statistic $\pi_0(\widehat{\boldsymbol{\phi}}_n)$ is sufficiently small. A frequentist would determine the cutoff point for rejection by deriving the frequency distribution of $\pi_0(\widehat{\boldsymbol{\phi}}_n)$ under H_0 and then choosing an appropriate Type I error probability. Apparently, it is necessary to specify a prior probability, π_0, for H_0, but

it is easily checked that using $\pi_0(\widehat{\boldsymbol{\phi}}_n)$ in frequentist fashion is equivalent to a test that is invariant to the value of π_0.

A computable closed form for $\pi_0(\widehat{\boldsymbol{\phi}}_n)$ is possible only for selected prior distributions. We will thus approximate $\pi_0(\widehat{\boldsymbol{\phi}}_n)$ in a way that yields a closed form. Aerts, Claeskens and Hart (2004) used BIC to approximate $\pi_0(\widehat{\boldsymbol{\phi}}_n)$, but we will use a more refined approximation, namely the method of Laplace. To the author's knowledge, the only article to apply Laplace's approximation to goodness- or lack-of-fit testing is Bogdan (2001), who used the method to select the number of components in a Neyman smooth test. For details on using Laplace's method in a general Bayesian context, the reader is referred to de Bruijn (1970) and Tierney and Kadane (1986).

It will be assumed that $\phi_1, \ldots, \phi_n$ are a priori independent with

$$P(\phi_j = 0) = 1 - \pi_j, \quad j = 1, \ldots, n,$$

where $\pi_j < 1$ for all j and, given that $\phi_j \neq 0$, ϕ_j has density g, $j = 1, \ldots, n$. The scale parameter σ has prior π and is assumed to be a priori independent of the Fourier coefficients.

Before continuing we need to define some notation. For $m = 1, \ldots, n$, define $n_m = \binom{n}{m}$ and let $S_{m1}, \ldots, S_{mn_m}$ be the n_m subsets of $\{1, \ldots, n\}$ of size m. For each m and i, let $\bar{S}_{mi}$ be the elements of $\{1, \ldots, n\}$ that are not in S_{mi}, and let $\boldsymbol{\phi}_{mi}$ and $\widehat{\boldsymbol{\phi}}_{mi}$ denote the vectors $(\phi_{j_1}, \ldots, \phi_{j_m})$ and $(\hat{\phi}_{j_1}, \ldots, \hat{\phi}_{j_m})$, respectively, where $j_1 < \cdots < j_m$ are the elements of S_{mi}. Finally, for $i = 0, \ldots, n_m$ and $m = 1, \ldots, n$, define the integrals

$$\begin{aligned} I_{mi} &= \int_0^\infty \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \exp\left[-\frac{n}{2\sigma^2}\left(\sum_{j \in S_{mi}} (\hat{\phi}_j - \phi_j)^2 + \sum_{j \in \bar{S}_{mi}} \hat{\phi}_j^2\right)\right] \\ &\qquad \times \left[\prod_{j \in S_{mi}} g(\phi_j)\, d\phi_j\right] \sigma^{-n} \pi(\sigma)\, d\sigma. \end{aligned}$$

The posterior probability of H_0 may be expressed as

$$\pi_0(\widehat{\boldsymbol{\phi}}_n) = \frac{p_0(\widehat{\boldsymbol{\phi}}_n)}{p_{marg}(\widehat{\boldsymbol{\phi}}_n)},$$

where

$$p_0(\widehat{\boldsymbol{\phi}}_n) = I_{01} \prod_{j=1}^{n} (1 - \pi_j)$$

and

$$p_{marg}(\widehat{\boldsymbol{\phi}}_n) = p_0(\widehat{\boldsymbol{\phi}}_n) + \sum_{m=1}^{n} \sum_{i=1}^{n_m} \prod_{j \in S_{mi}} \pi_j \prod_{j \in \bar{S}_{mi}} (1 - \pi_j) I_{mi}.$$

We now use Laplace's method to approximate each integral I_{mi}:

$$I_{mi} \approx \frac{1}{\sqrt{2}} \left(\frac{2\pi}{n}\right)^{(m+1)/2} \hat{\sigma}_{mi}^{m+1-n} e^{-n/2} \prod_{j \in S_{mi}} g(\hat{\phi}_j) \pi(\hat{\sigma}_{mi}), \tag{5}$$

where

$$\hat{\sigma}_{mi}^2 = \sum_{j \in \bar{S}_{mi}} \hat{\phi}_j^2.$$

Here, we have used a common variant of the "pure" Laplace approximation in which the prior is evaluated at the maximum likelihood estimate of $\boldsymbol{\phi}_{mi}$ rather than at the posterior mode; see Kass and Raftery (1995).

Substituting (5) into $\pi_0(\widehat{\boldsymbol{\phi}}_n)$ yields our approximation, $\hat{\pi}_0$, of the posterior probability of H_0:

$$\hat{\pi}_0 = \frac{1}{1 + R_n},$$

where

$$R_n = \sum_{m=1}^{n} \sum_{i=1}^{n_m} \prod_{j \in S_{mi}} \frac{\pi_j}{(1 - \pi_j)} \left(\frac{2\pi}{n}\right)^{m/2} \hat{\sigma}_0^{n-1} \hat{\sigma}_{mi}^{m+1-n} \prod_{j \in S_{mi}} g(\hat{\phi}_j) \frac{\pi(\hat{\sigma}_{mi})}{\pi(\hat{\sigma}_0)}$$

and $\hat{\sigma}_0 \equiv \hat{\sigma}_{01}$. The frequentist test that rejects H_0 for small values of $\hat{\pi}_0$ is equivalent to one that rejects for large values of

$$\tilde{B}_n + E_n,$$

where

$$\tilde{B}_n = \hat{\sigma}_0 \sum_{i=1}^{n} \frac{\pi_i}{(1-\pi_i)} g(\hat{\phi}_i) \frac{\pi(\hat{\sigma}_{1i})}{\pi(\hat{\sigma}_0)} \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{1i}^2} \right)^{(n-2)/2}, \tag{6}$$

$E_n = \sum_{m=2}^{n} U_{nm}$ and, for $m = 2, \ldots, n$,

$$U_{nm} = \hat{\sigma}_0^m \left(\frac{2\pi}{n} \right)^{(m-1)/2} \sum_{i=1}^{n_m} \prod_{j \in S_{mi}} \left[\frac{\pi_j}{(1-\pi_j)} g(\hat{\phi}_j) \right] \frac{\pi(\hat{\sigma}_{mi})}{\pi(\hat{\sigma}_0)} \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{mi}^2} \right)^{(n-m-1)/2}.$$

It seems doubtful that the test statistic $\tilde{B}_n + E_n$ would ever be markedly more powerful than $\tilde{B}_n$. One way to see this is to consider what happens under $1/\sqrt{n}$ local alternatives to H_0. Theorem 1 shows that, under such alternatives and when $\pi_j = O(j^{-\alpha})$ for some $\alpha > 1$, $\tilde{B}_n$ converges in probability to a nongenerate random variable. Under the the same conditions it is straightforward to show that for each fixed $m \geq 2$, $U_{nm} = O_p(n^{-(m-1)/2})$ under $1/\sqrt{n}$ alternatives. With this result and the relative simplicity of $\tilde{B}_n$ as motivation, we shall consider only $\tilde{B}_n$ and modifications thereof in in the sequel. We also note that the multiplier $\hat{\sigma}_0$ seems of dubious benefit, and hence will subsequently be dropped.

From the standpoint of test power, the most important components in (6) are

$$\gamma_i = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{1i}^2} \right)^{(n-2)/2} = \left(1 + \frac{\hat{\phi}_i^2}{\hat{\sigma}_{1i}^2} \right)^{(n-2)/2}, \quad i = 1, \ldots, n.$$

The following remarks relevant to these components are in order:

R1. Under the null hypothesis each of γ_i is equal in distribution to $(1+F/n)^{(n-2)/2}$, where F has the F distribution with 1 and $n-1$ degrees of freedom.

R2. Under the null hypothesis, $\gamma_1, \ldots, \gamma_n$ are asymptotically independent and identically distributed as $\exp(\chi^2/2)$, where χ^2 has the χ^2 distribution with one degree of freedom.

R3. For any positive constant a, $\exp(ax) \geq (1+x)^a$ for all $x \geq 0$.

Remarks R2 and R3 imply that the following statistic is, asymptotically, both null-equivalent to and more powerful than $\tilde{B}_n/\hat{\sigma}_0$:

$$B_n = \sum_{i=1}^{n} \frac{\pi_i}{(1-\pi_i)} g(\hat{\phi}_i) \, \frac{\pi(\hat{\sigma}_{1i})}{\pi(\hat{\sigma}_0)} \exp\left(\frac{n\hat{\phi}_i^2}{2\hat{\sigma}_{1i}^2}\right). \tag{7}$$

The exponential terms in B_n are a common feature of tests based on the posterior probability of the null. When applying Laplace approximations to more general models, the statistic analogous to R_n will be a weighted sum of likelihood ratios $\hat{L}_j/\hat{L}_0$, where $\hat{L}_0$ and $\hat{L}_j$, $j = 1, \ldots, K$, are maximized likelihoods of the null and K alternative models, respectively. Writing $\xi_j = 2\log(\hat{L}_j/\hat{L}_0)$, we have

$$\frac{\hat{L}_j}{\hat{L}_0} = \exp\left(\frac{\xi_j}{2}\right),$$

and if the null is nested within each alternative, then under standard regularity conditions each ξ_j will have an asymptotic χ^2 distribution under the null hypothesis. In short, the "sum of exponentials" phenomenon can be attributed to two factors: (i) the use of a posterior probability to test H_0, and (ii) entertainment of more than two models. When the null is compared to just one other model, our frequentist-Bayes test is essentially the same as a likelihood ratio test.

The regression model in Section 2 that is isomorphic to our canonical, Fourier coefficients model is homoscedastic. If the errors in such a regression model have an unknown heteroscedastic structure, then the sample Fourier coefficients are no longer sufficient statistics. Given a parametric model for the error variances, one may nonetheless derive a frequentist-Bayes test statistic based on Laplace approximations. In this case the derivation would start from a likelihood written in terms of the observations $Y_1, \ldots, Y_{n+p+1}$.

4 Weighted sums and score statistics

Two types of test statistics have played a prominent role in the lack-of-fit literature: weighted sums of independent components and Gaussian-likelihood score tests. By a weighted sum, we mean a statistic of the form

$$S_n = \sum_{j=1}^{n} w_j (n\hat{\phi}_j^2/\hat{\sigma}^2),$$

where $w_1, w_2, \ldots$ are known positive constants with $\sum_{j=1}^{\infty} w_j < \infty$. A regression analog of a Neyman smooth statistic is

$$S_n(m) = \sum_{j=1}^{m} (n\hat{\phi}_j^2/\hat{\sigma}^2).$$

If $\hat{\sigma}^2$ is computed on the assumption that the null hypothesis is true, then $S_n(m)$ is a score statistic for testing (3) in the regression model (2). On the other hand, if $\hat{\sigma}^2$ is an estimate based on assuming that the alternative is true, then $S_n(m)$ arises from use of the *reduction method* in linear models; see, for example, Hart (1997, pp. 124-125). Here it is assumed that m is a priori fixed.

As an omnibus test, S_n is attractive since it is consistent against any alternative for which at least one ϕ_j is nonzero. Unfortunately, such tests are notorious for having poor power in "moderate-sized" samples for all but very low frequency alternatives. This is true even for the *cusum* test, which has weights $w_j = 1/j^2$, $j = 1, 2, \ldots$, that decrease to 0 at a fairly slow, algebraic rate (Hart 1997).

Suppose that the difference between the null and true functions has the form $r(x) = \sum_{j=1}^{m} \phi_j u_j(x)$. Then by a classic result of Lehmann (1959) a test based on $S_n(m)$ is uniformly most powerful among tests whose power functions depend only on $\sum_{j=1}^{m} \phi_j^2/\sigma^2$. However, due to its dependence on the smoothing parameter m, a score test cannot be regarded as an omnibus test. This "defect" can be repaired asymptotically by using a so-called *data-driven score test*, i.e., a test based on $S_n(\hat{m})$,

where $\hat{m}$ is selected from the data (Ledwina 1994, Kuchibhatla and Hart 1996, Lee and Hart 2000, Bogdan 2001 and Inglot and Ledwina 2006). The version of $S_n(\hat{m})$ that has received the most attention is $S_n(\hat{m}_{BIC})$, where $\hat{m}_{BIC}$ optimizes BIC.

The most interesting and potentially important result of the current paper may be stated as follows. In our comprehensive simulation study (Section 7), a particular weighted sum of *exponentiated* squared Fourier coefficients has substantially better power at higher frequency alternatives than a test based on $S_n(\hat{m}_{BIC})$, and power comparable to that of $S_n(\hat{m}_{BIC})$ at lower frequency alternatives. The good power properties of B_n are somewhat surprising because B_n is not adaptive, i.e., it has nonrandom weights. The version of B_n used in our simulation study has $\pi_j = 1/(1+j^2)$, meaning that it uses the same weights as the cusum test. In spite of the fact that each component $\exp[n\hat{\phi}_j^2/(2\sigma^2)]$ has infinite mean, it will be shown that, as $n \to \infty$, the statistic

$$\sum_{j=1}^{n} j^{-2} \exp\left(\frac{n\hat{\phi}_j^2}{2\hat{\sigma}^2}\right) \tag{8}$$

converges in distribution to a random variable that is finite with probability 1. This relative stability under the null and the fact that exponentiation "explodes" the effect of nonzero Fourier coefficients explains the good power properties of (8).

The exponentiation in B_n seems to be of fundamental importance since *it is not ad hoc, but rather a consequence of using a posterior probability to construct the test statistic.* This is evidenced by the discussion at the very end of Section 3.

5 Asymptotic distribution theory

We now consider the limiting distribution of B_n under both the null hypothesis and local alternatives that converge to the null at rate $1/\sqrt{n}$. Our local alternatives are

of the form

$$\phi_j = \frac{1}{\sqrt{n}}\,\beta_j, \quad n = 1, 2, \ldots, \; j = 1, \ldots, n. \tag{9}$$

A proof of the following theorem is sketched in the Appendix.

THEOREM 5.1. *Let $Z_1, Z_2, \ldots$ be i.i.d. standard normal random variables. Suppose that $\hat{\phi}_1, \ldots, \hat{\phi}_n$ are independent with $\hat{\phi}_j \sim N(\beta_j/\sqrt{n}, \sigma^2/n)$, $j = 1, \ldots, n$, where it is assumed that $\lim_{j\to\infty} \beta_j = 0$. Let $\pi_1, \pi_2, \ldots$ be defined as in Section 3 and suppose there exists $\delta < 1$ such that*

$$\sum_{j=1}^{\infty} \pi_j^{\delta} < \infty. \tag{10}$$

Assume also that the prior $\pi(x)$ is continuous at each positive x, and that g is bounded, and Lipschitz continuous in a neighborhood of 0. Then the statistic B_n defined by (7) converges in distribution to

$$B = g(0) \sum_{j=1}^{\infty} \frac{\pi_j}{(1-\pi_j)} \exp\left[(Z_j + \beta_j/\sigma)^2/2\right],$$

which is an almost surely convergent series.

Some remarks are in order concerning Theorem 5.1.

1. Consider the class $\mathcal{N}_n = \{M_1, \ldots, M_n\}$ of *nested* models. Here, model M_j is such that only the first j Fourier coefficients $\phi_1, \ldots, \phi_j$ are possibly nonzero, $j = 1, 2, \ldots$. Now, let $\hat{\pi}(\mathcal{N}_n)$ be the posterior probability of H_0 when the class of alternative models is $\mathcal{N}_n$. Then, as argued by Aerts, Claeskens and Hart (2004), the limiting power of tests based on $\hat{\pi}(\mathcal{N}_n)$ against the local alternatives (9) is completely determined by β_1. In particular, if $\beta_1 = 0$, then the asymptotic power is nil, i.e., it equals the test level. In contrast, Theorem 5.1 demonstrates that B_n can detect $1/\sqrt{n}$ alternatives whenever any Fourier coefficient β_j is nonzero.

2. A second remarkable aspect of Theorem 5.1 is its demonstration that *proper* prior probabilities π_j have a profound stabilizing effect on B_n. The BIC analog of B_n (studied in Aerts, Claeskens and Hart (2004)) is

$$B_{n,BIC} = \sum_{j=1}^{n} \exp\left(\frac{n\hat{\phi}_j^2}{2\sigma^2}\right).$$

In essence, $B_{n,BIC}$ is a special case of B_n with $\pi_j = 1/n$, $j = 1, \ldots, n$. The fact that the uniform prior is (asymptotically) *improper* entails that $B_{n,BIC}$ has to be carefully standardized to have a proper limiting distribution; see Aerts, Claeskens and Hart (2004) for the details. Furthermore, $B_{n,BIC}$ cannot detect $1/\sqrt{n}$ local alternatives, and against local alternatives it *can* detect, its power is completely determined by the largest Fourier coefficient.

Sufficient conditions for consistency of B_n against any *fixed* alternative with at least one $\phi_j \neq 0$ are (i) $\pi_j > 0$, $j = 1, 2, \ldots$, and (ii) g is strictly positive.

6 Choice of prior probabilities

The statistic B_n depends on the prior distribution via the prior density g, the probabilities π_j and the prior π for σ. A popular noninformative prior for a scale parameter σ is the improper prior $\pi(\sigma) = \sigma^{-1}$. If this prior is used in (6), then $\tilde{B}_n$ becomes

$$\tilde{B}_n = \hat{\sigma}_0 \sum_{i=1}^{n} \frac{\pi_i}{(1-\pi_i)} g(\hat{\phi}_i) \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{1i}^2}\right)^{(n-1)/2}, \tag{11}$$

which, in light of remarks R2 and R3, suggests that we use as a test statistic

$$B_n = \sum_{i=1}^{n} \frac{\pi_i}{(1-\pi_i)} g(\hat{\phi}_i) \exp\left(\frac{n\hat{\phi}_i^2}{2\hat{\sigma}_{1i}^2}\right). \tag{12}$$

Theorem 5.1 suggests that g does not play a very important role, since for local alternatives it only produces the multiplicative constant $g(0)$. On the other hand,

$\pi_1, \pi_2, \ldots$ are crucial. In situations where one suspects that the underlying function r is highly correlated with certain orthogonal functions, it would be judicious to place larger prior probabilities on the corresponding Fourier coefficients. Doing so will tend to increase the power of the resulting test if one's suspicions are justified.

In a nonparametric setting, very little is known about the underlying function. In such a case it would make sense to use vague probabilities. One possibility in this regard is to take $\pi_j = 1/2$ for each j. However, this yields the statistic $B_{n,BIC}$, which, as noted in the previous section, has no asymptotic power against $1/\sqrt{n}$ local alternatives. The problem with $\pi_j = 1/2$ is that it fails to reflect our knowledge that relatively few of a function's Fourier coefficients will be substantially different from 0. In what follows we will propose vague choices for prior probabilities in three different settings.

Functions of one variable: polynomial and trigonometric bases. Suppose that r is a function of a single variable, and that we represent it by linear combinations of polynomials or trigonometric functions. Furthermore, assume that $\phi_1, \phi_2, \ldots$ correspond to the basis functions arranged in order of increasing frequency. It seems natural that vague probabilities on the ϕ_js should decrease to 0 monotonically in frequency. Vagueness also suggests that the decrease to 0 should be quite slow. To ensure the desirable properties of Theorem 5.1, we need slightly more than summability of the π_js. Taking $\pi_j = 1/j^\alpha$ for any $\alpha > 1$ satisfies condition (10), and taking α fairly close to 1 will ensure vagueness of the prior probabilities.

In Section 7 we will consider a version of B_n in which g is a constant (i.e., the improper uniform prior) and $\pi_j = (1 + j^2)^{-1}$. In addition to being reasonably noninformative, these probabilities lead to the same weights, j^{-2}, as those used by the cusum test. Hence, differences in power between this version of B_n and the cusum test can be attributed solely to exponentiation of the squared Fourier

coefficients.

Functions of one variable: wavelet bases. In a wavelet analysis, there are $J = \log_2 n$ levels of frequency resolution, with each level consisting of a number of basis functions that are shifted versions of each other. This yields the well-known frequency *and* time localization that makes wavelets so attractive (Ogden 1997). Let $\phi_{j,k}$, $k = 0, 1, \ldots, 2^j - 1$ denote the true wavelet coefficients at resolution level j, $j = 0, 1, \ldots, J-1$. Here it seems reasonable to assign each of $\phi_{j,0}, \ldots, \phi_{j,2^j-1}$ the same prior probability of, say,

$$\pi_j = \frac{1}{(1 + 2^{j\alpha})}, \quad j = 0, 1, \ldots, J-1, \tag{13}$$

with $\alpha > 1$, in which case B_n takes the form

$$\sum_{j=0}^{J-1} 2^{-j\alpha} \sum_{k=0}^{2^j-1} \exp\left(\frac{n\hat{\phi}_{jk}^2}{2\sigma^2}\right).$$

The series $\sum_{j=0}^{J-1} 2^{-j(\alpha-1)}$ satisfies condition (10) since $\alpha > 1$, and hence the result of Theorem 5.1 is true for wavelet series when the π_js are defined as in (13).

Functions of several variables. For the sake of illustration, suppose that r is a function of two variables. The regression model of Section 2 with $d = 2$ is an example of this case. If one uses, say, a trigonometric basis, the sample Fourier coefficients will be a function of two frequency indices, say j and k, and may be written $\hat{\phi}_{jk}$, $j = 1, \ldots, \sqrt{n}$, $k = 1, \ldots, \sqrt{n}$, where for convenience we assume that $\sqrt{n}$ is an integer. The statistic B_n now has the form

$$B_n = \sum_{j=1}^{\sqrt{n}} \sum_{k=1}^{\sqrt{n}} \frac{\pi_{jk}}{(1 - \pi_{jk})} \exp\left(\frac{n\hat{\phi}_{jk}^2}{2\sigma^2}\right),$$

and we are faced with choosing the π_{jk}s. Assuming that each of the indices j and k is proportional to frequency, it seems reasonable to take as vague probabilities

$$\pi_{jk} = \frac{1}{1 + (jk)^\alpha}, \quad j = 1, \ldots, \sqrt{n}, \ k = 1, \ldots, \sqrt{n},$$

for some $\alpha > 1$. If we take $1/\alpha < \delta < 1$, then the series $\sum_{j=1}^{\sqrt{n}} \sum_{k=1}^{\sqrt{n}} (jk)^{-\delta\alpha}$ converges as $n \to \infty$, and hence Theorem 5.1 is applicable to this setting. It is worth noting, however, that a manifestation of the curse of dimensionality can be seen here. The maximum frequency that can be resolved in this case is $\sqrt{n}$, as opposed to n in the single variable case. The curse gets worse quickly with increasing dimension d, wherein the maximum observable frequency is $n^{1/d}$. When $n = 125$, for example, the upper bound on frequency is practically unlimited in the univariate case, but is a fairly limiting 5 when d is just 3.

7 A simulation study

Our simulations are limited to univariate regression, but are arguably fairly comprehensive in that setting. All statistics considered depend on an estimate of the scale parameter σ. Statistics relying on the score principle use an estimate of σ that assumes the null hypothesis to be true, whereas our frequentist-Bayes statistic uses estimates that are appropriate when some alternative is true. We would like our power comparisons to reveal differences in methods that are not due to differences in variance estimates. Hence, we will use the same scale estimator for all statistics, and this will be the one motivated by score tests. In terms of Fourier coefficients, this estimator has the form $\hat{\sigma}^2 = \sum_{j=1}^{n} \hat{\phi}_j^2$, which is unbiased for σ^2 under the null.

7.1 Testing for no effect

We first consider the model (2) in which $\epsilon_1, \ldots, \epsilon_{n+1}$ are i.i.d. $N(0, 1)$ and $x_j = (j - 1/2)/(n + 1)$, $j = 1, \ldots, n + 1$. The null hypothesis is the so-called no-effect

hypothesis, i.e., μ is identical to a constant. The sample Fourier coefficients are

$$\hat{\phi}_j = \frac{\sqrt{2}}{(n+1)} \sum_{i=1}^{n+1} Y_i \cos(\pi j x_i), \quad j = 1, \ldots, n.$$

Two different types of studies were performed. In the first, a "traditional" sort of study was done in which a large number (100,000) of data sets is generated from each of several models. In the second study, a single data set was generated from each of 100,000 randomly generated functions, and we consider how power is related to two crucial characteristics of the functions.

Four tests based on orthogonal series were considered in both studies. One test uses a special case of B_n having the form

$$B_n = \sum_{j=1}^{n} j^{-2} \exp\left(\frac{n\hat{\phi}_j^2}{2\hat{\sigma}^2}\right).$$

A test based on this statistic will be referred to as a *Bayes sum test*. A second test, to be called a *BIC score test*, uses an adaptive score statistic of the form $\sum_{j=1}^{\hat{m}} n\hat{\phi}_j^2/\hat{\sigma}^2$, where $\hat{m}$ maximizes $BIC(m) = \sum_{j=1}^{m} n\hat{\phi}_j^2/\hat{\sigma}^2 - m \log n$ over $m = 1, \ldots, n$. Finally, we also consider the cusum statistic

$$C_n = \sum_{j=1}^{n} \frac{1}{\hat{\sigma}^2} \frac{n\hat{\phi}_j^2}{j^2},$$

which has been investigated in the regression context by Buckley (1991) and Eubank and Hart (1993).

Both studies used sample size $n = 100$, and our first step was to obtain good approximations to critical values of each statistic. This was accomplished by generating one million random samples of size 100 from the standard normal distribution and computing all four statistics for each of the samples. Approximations to critical values for size 0.05 tests were 6.801, 5.471 and 4.583 for the Bayes sum, BIC score and cusum tests, respectively.

In the first study, the underlying function has the form

$$r(x) = \sqrt{2}\phi \cos(\pi m_0 x). \tag{14}$$

Here, the idea is to use a single cosine wave to investigate the effect of the frequency, m_0, on the power of each test. The frequencies considered were $m_0 = 1, 2, 3, 4$ and 7, and at each m_0 power was approximated for six values of $\lambda = n\phi^2$. One hundred thousand replications were performed at each combination of m_0 and λ. The results are given in Table 1. The only frequency at which the cusum test is not clearly inferior to the other three tests is the lowest one, $m_0 = 1$. A comparison of the Bayes sum and BIC score tests is conveyed graphically in Figure 1. Here, power as a function of m_0 is plotted for $\lambda = 1, 4$, and 16. At the lowest frequency, the Bayes sum test has, overall, a slight power advantage over BIC score. At the highest frequency, the Bayes sum test is never worse than and usually substantially better than the BIC score test. Only at some intermediate frequencies does BIC score have higher power than Bayes sum, and in these cases the difference in power is fairly small.

In our second study, each of 100,000 data sets of size $n = 100$ was generated as follows.

- A random sample, $\tilde{\phi}_1, \ldots, \tilde{\phi}_{100}$, was generated from $N(0, (0.2)^2)$.

- Fourier coefficients $\phi_1, \ldots, \phi_{100}$ were obtained by multiplying the quantities in the previous step by a random damping factor. Define, for each $m = 1, 2, \ldots, 100$,

$$d_j(m) = \exp\left[-\frac{1}{32}(j-m)^2\right], \quad j = 1, \ldots, 100. \tag{15}$$

  Define also the probability distribution

$$p_j = \frac{j^{-1.75}}{\sum_{k=1}^{100} k^{-1.75}}, \quad j = 1, \ldots, 100.$$

	m_0				
λ	1	2	3	4	7
	0.09465	0.06678	0.05781	0.05468	0.05052
0.5	0.08024	0.08777	0.05977	0.05151	0.04950
	0.10687	0.05469	0.04964	0.04959	0.04941
	0.14439	0.08585	0.07053	0.06197	0.05371
1	0.11736	0.13036	0.07253	0.05577	0.04786
	0.16513	0.05898	0.05173	0.05046	0.04852
	0.24602	0.13973	0.10249	0.08494	0.06123
2	0.19837	0.22430	0.10935	0.06580	0.04674
	0.28366	0.07186	0.05431	0.05062	0.04613
	0.45119	0.27411	0.19713	0.15755	0.10030
4	0.37752	0.41484	0.21041	0.10440	0.04559
	0.50065	0.10468	0.05933	0.05123	0.04696
	0.75328	0.56685	0.45151	0.37470	0.25338
8	0.68300	0.72343	0.46737	0.25229	0.04877
	0.79423	0.21180	0.07364	0.05437	0.04428
	0.96663	0.90040	0.83843	0.78130	0.65587
16	0.94684	0.95676	0.84476	0.64070	0.12307
	0.97568	0.518873	0.10811	0.05697	0.03910

Table 1: *Empirical Power of Size* 0.05 *Bayes Sum, BIC score and Cusum Tests.* The underlying function is $r(x) = \sqrt{2}\phi\cos(\pi m_0 x)$ and $\lambda = n\phi^2/\sigma^2$. For given λ and m_0, the upper row is the power of the Bayes sum test, and the second, third and fourth rows are powers of the Bayes nested, BIC score and cusum tests, respectively. Results are based on $n = 100$ and 100,000 replications. See the text for further details.

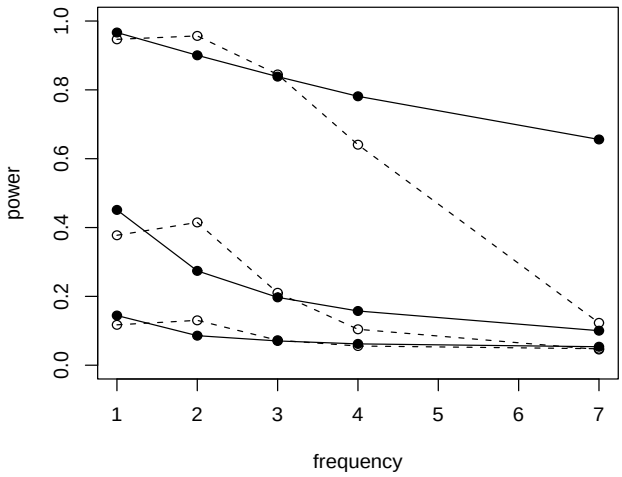

Figure 1: *Power Plots for Size* 0.05 *Bayes Sum and BIC score Tests.* Solid and dashed lines correspond to Bayes sum and BIC score, respectively. At any given m_0 and test, the powers from smallest to largest correspond to $\lambda = 1, 4$ and 16, respectively.

A value $\boldsymbol{m}$ was selected from this distribution, and Fourier coefficients were obtained by computing

$$\phi_j = d_j(\boldsymbol{m})\tilde{\phi}_j, \quad j = 1, \ldots, 100.$$

- Defining $r(x) = \sqrt{2}\sum_{j=1}^{100} \phi_j \cos(\pi j x)$, the data were

$$Y_i = r(x_i) + \epsilon_i, \quad i = 1, \ldots, 101,$$

where $\epsilon_1, \ldots, \epsilon_{101}$ were i.i.d. $N(0, 1)$.

Size 0.05 versions of the four tests were conducted for each of the 100,000 data sets, and the results are summarized in Figures 2-4. In each figure, estimated power is displayed as a function of the "size" and "frequency" of r, which are defined as follows:

$$\text{size} = \int_0^1 r^2(x)\, dx = \sum_{j=1}^{\infty} \phi_j^2$$

and

$$\text{frequency} \times \text{size} = \int_0^1 [r'(x)]^2 dx = \pi^2 \sum_{j=1}^{\infty} j^2 \phi_j^2.$$

The transformed size and frequency referred to in each plot are

$$\text{size}/1.12 \quad \text{and} \quad \log(\text{frequency})/11.5,$$

respectively. A log transform of frequency was used since the distribution of frequencies was highly right-skewed. For a given test, the 100,000 simulated data sets yield data (s_i, f_i, a_i), $i = 1, \ldots, 100,000$, where s_i and f_i are the transformed size and frequency, respectively, of the ith generated function, and

$$a_i = \begin{cases} 1, & \text{if } H_0 \text{ is rejected for data set } i, \\ 0, & \text{otherwise.} \end{cases}$$

A kernel smoother is used to estimate the regression of a_i on (s_i, f_i), and it is these smooths that are shown in Figures 2-4.

The results of this second part of our simulation study reinforce the conclusions of the first part. Interestingly, the red regions, i.e., those in which power is highest, appear to have a nested structure. The red region for the Bayes sum test contains that of the BIC score test, which contains that of the cusum test. The proportions of rejections out of all 100,000 tests conducted were 0.56108, 0.52500 and 0.37970 for the Bayes sum, BIC score and cusum tests, respectively. It should be noted that the manner in which our random functions were generated produced a much higher proportion of low than high frequency functions. The 75th and 90th percentiles of all transformed frequencies were 0.490 and 0.611. It follows that a simple average of power downweights the cases where the Bayes sum test has its biggest advantage over the others.

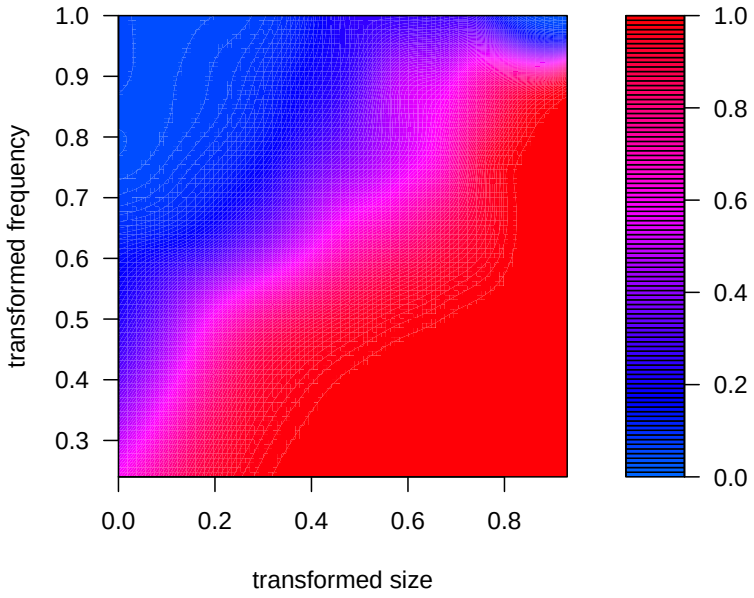

Figure 2: *Empirical Power Plot for Size 0.05 Bayes Sum Test.*

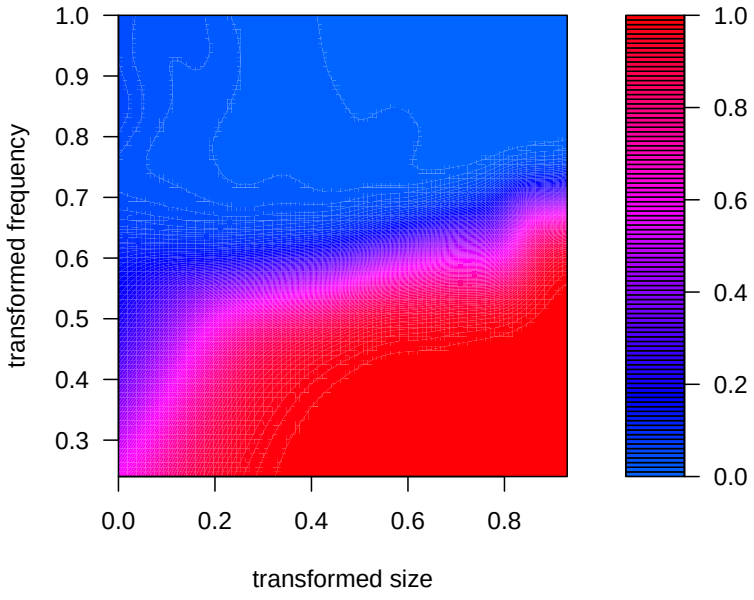

Figure 3: *Empirical Power Plot for Size 0.05 Data-driven Score Test.*

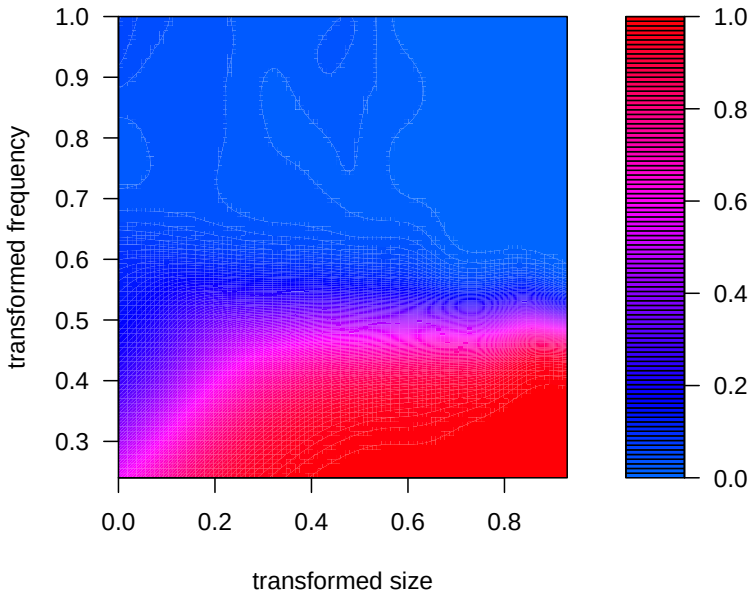

Figure 4: *Empirical Power Plot for Size 0.05 Cusum Test.*

7.2 Testing the fit of polynomials

We again use the model of Section 7.1 but now consider testing the null hypothesis that μ is a polynomial of specified degree. We consider two null hypotheses: μ is a straight line, and μ is quadratic. The basis functions are taken to be orthogonal polynomials, which guarantees that A1-A3 are true. We first do simulations under the null hypothesis (based on one million replications) to determine 0.05 level critical values for each of the tests considered in Section 7.1. This was done at each of the sample sizes $n = 50$, 100, 200 and 500. The critical values so-determined are given in Table 2.

Our power study parallels the second study of Section 7.1. Random functions were generated exactly as they were there. (This is appropriate since a function generated in this way is almost surely different from a polynomial.) One hundred

Table 2: *Approximate 0.05 level critical values for testing a straight line or quadratic null hypothesis.* The results are based on one million replications from the respective null models.

Null	n	Bayes sum	BIC score	Cusum
Linear	50	6.901	6.155	4.634
	100	6.927	5.530	4.593
	200	6.991	4.963	4.585
	500	6.901	4.342	4.533
Quadratic	50	7.289	6.363	4.720
	100	7.141	5.641	4.642
	200	7.054	4.988	4.586
	500	6.934	4.353	4.549

thousand replications were performed at each of the four sample sizes mentioned in the last paragraph. Rather than giving heat plots as we did previously, we simply report (in Table 3) the proportions of rejections at each sample size. The results parallel the analogous ones reported in the last paragraph of Section 7.1. The decrease in power with polynomial degree is to be expected since the L_2-discrepancy between a function and its best polynomial approximation decreases with polynomial degree.

7.3 Testing the fit of a tides model

The classic harmonic model (Cole 1997, p. 14) for a series of tidal heights $Y_1, \ldots, Y_n$ observed at evenly spaced time points $t = 1, \ldots, n$ has the form

$$Y_t = \theta_0 + \sum_{j=1}^{K} \beta_j \cos(a_j t + p_j) + \epsilon_t, \quad t = 1, \ldots, n, \tag{16}$$

where $a_1, \ldots, a_K$ are known *speeds*, $\beta_1, \ldots, \beta_K$ are unknown amplitudes, $p_1, \ldots, p_K$ are unknown phases, and $\epsilon_1, \ldots, \epsilon_n$ are unobserved, i.i.d., mean 0 random variables. Each cosine term is referred to as a *constituent*, and it is sometimes of interest to

Table 3: *Proportions of rejections in 0.05 level tests of straight line and quadratic null hypotheses.* For each n and type of null hypothesis, 100,000 random functions were generated as described in Section 7.1. The tests were applied to each resulting data set and the proportions of rejections recorded.

Null	n	Bayes sum	BIC score	Cusum
Linear	50	0.25527	0.27375	0.21021
	100	0.47137	0.46007	0.36222
	200	0.71594	0.66020	0.55255
	500	0.90939	0.84441	0.76883
Quadratic	50	0.22073	0.22954	0.19346
	100	0.40879	0.39968	0.32664
	200	0.64166	0.59546	0.50591
	500	0.86871	0.80200	0.72983

test the null hypothesis that a model containing a specified set of constituents is appropriate. The alternatives of interest might be that the tidal heights are subject to a slow drift or that some constituent is missing from the model.

The deterministic part of model (16) may be written

$$\begin{aligned}\theta_0 + \sum_{j=1}^{K} \beta_j \cos(a_j t + p_j) &= \theta_0 + \sum_{j=1}^{K} [\beta_j \cos(a_j t)\cos(p_j) - \beta_j \sin(a_j t)\sin(p_j)] \\ &= \theta_0 + \sum_{j=1}^{K} [\theta_{j1}\cos(a_j t) + \theta_{j2}\sin(a_j t)]\,.\end{aligned}$$

Since $a_1, \ldots, a_K$ are known, this makes it clear that model (16) is linear in $(\theta_{1j}, \theta_{2j})$, $j = 1, \ldots, K$, and hence the methods of the current paper can be used to test its adequacy.

A simulation was performed in which the null model was as follows:

$$\begin{aligned}Y_j &= \theta_0 + \theta_{11}\cos(2\pi\omega_1 \cdot 720x_j) + \theta_{12}\sin(2\pi\omega_1 \cdot 720x_j) + \theta_{21}\cos(2\pi\omega_2 \cdot 720x_j) + \\ &\qquad \theta_{22}\sin(2\pi\omega_2 \cdot 720x_j) + \epsilon_j, \quad j = 1, \ldots, 720, \end{aligned} \tag{17}$$

where $x_j = (j - 1/2)/720$, $j = 1, \ldots, 720$, $\epsilon_1, \ldots, \epsilon_{720}$ are i.i.d. standard normals,

$\omega_1 = 1/25.81933$ and $\omega_2 = 1/12.65835$. This model amounts to observing hourly tidal heights for a period of 30 days. The frequencies ω_1 and ω_2 correspond to two of the principal tidal constituents (Cole 1997, p. 10). Starting from the cosine functions $\cos(\pi jx)$, $j = 1, 2, \ldots$, the Gram-Schmidt process was used to define basis functions that are both orthornormal with respect to the design points and orthogonal to the functions of the null model. This ensures that the sample Fourier coefficients satisfy A1-A3.

Included in the simulations of this section is a test proposed in Section 3.4 of Inglot and Ledwina (2006). It is a variation of the BIC score test with improved power against high frequency alternatives. This is achieved by using a selection criterion that is a hybrid of BIC and AIC, which makes the criterion less likely to choose a very low dimensional model when the true model is high frequency. The modified criterion depends on a constant c that we take to be 2.4. We call this fourth test a "modified BIC score test."

An initial simulation with one million replications was performed under model (17) to determine 0.05 level critical values for the four tests of interest. The values obtained for the Bayes sum, BIC score, cusum and modified BIC score were 6.791, 4.216, 4.522 and 4.310, respectively. The first three of these are quite similar to those in Table 2 at $n = 500$, as would be expected.

Our power study considers alternatives of the form

$$\mu(x) = \mu_0(x) + a\sin(2\pi\omega_3 \cdot 720x), \tag{18}$$

where μ_0 is the null function and $\omega_3 = 1/23.93477$. This is a very high frequency alternative, but is in no way artificial in the context of the tides problem since it corresponds to a commonly occurring constituent (Cole 1997, p. 10). Estimated power (based on 100,000 replications) was determined for each of the four tests at each of the amplitudes $a = 0.2, 0.4, 0.6, 0.8, 1.0$. Results are given in Table 4.

a	Bayes sum	Modified BIC score	BIC score	Cusum
0.2	0.09170	0.06180	0.04164	0.04219
0.4	0.89534	0.12501	0.02441	0.02465
0.6	0.99997	0.29839	0.01025	0.01051
0.8	1	0.96817	0.00385	0.00383
1.0	1	0.99999	0.00090	0.00087

Table 4: *Empirical power of four 0.05 level tests for tides model.* Results are based on 100,000 replications from the alternative defined by (18).

The power of the Bayes sum test is exceptionally good in comparison to the other tests. The powers of the BIC score and cusum tests are extremely poor. The fact that these powers are actually lower than the test level is due to the fact that the null variance estimate is inflated by true Fourier coefficients that are nonzero. The power of the modified BIC score test is quite good at amplitiudes $a = 0.8$ and 1.0, but is poor relative to the Bayes sum test at $a = 0.4$ and 0.6. It is remarkable that the Bayes sum test is so much more powerful than the others in spite of the fact that it is not adaptive. This is testament to just how much exponentiation magnifies even small deviations from the null hypothesis.

8 An example involving microarrays

Here we apply a frequentist-Bayes test to data collected by the authors of Snijders, et al. (2001). The data are from a microarray experiment that measured genome-wide DNA copy number. The variable considered here is the ratio of dye intensities for test and reference samples at a given marker along a chromosome of interest. Each intensity is proportional to the number of marker copies. The reference samples are diploid, and hence each reference marker has only two copies. It is of interest to detect regions on a chromosome where the test samples may have more or fewer

copy numbers than the corresponding reference samples. Such variations in copy number are known to be correlated with cancer incidence; see, for example, Lucito, et al. (2000).

Data sets for four different chromosomes (gotten from cell line GM03563) are shown in Figure 5. In each graph, the horizontal axis is marker location and the vertical axis is the normalized average of three readings of $\log_2(\rho)$, where ρ is the aforementioned ratio of intensities (test over reference). The sample sizes are 135, 85, 171 and 109 for chromosomes 1, 3, 4 and 9, respectively. The Bayes sum test (as in Section 7) and the BIC score test were applied to each data set to test for constancy of expected $\log_2(\rho)$. In fact, the hypothesis of interest is that this expectation is identically 0. Neither of the tests used has any power against a simple shift alternative, and hence if the null hypothesis of constancy is not rejected, then one would still want to investigate the (perhaps unlikely) possibility that the true function is identical to a nonzero constant.

The variance σ^2 for a given data set $Y_1, \ldots, Y_n$ was estimated by

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} (0.809Y_{i-1} - 0.5Y_i - 0.309Y_{i+1})^2,$$

which is an asymptotically optimal estimator based on second differences (Hall, Kay and Titterington 1990). The same cosine basis as in Section 7 was used in constructing the two test statistics, which have the form

$$B_n = \sum_{j=1}^{n-1} j^{-2} \exp\left(\frac{n\hat{\phi}_j^2}{2\hat{\sigma}^2}\right)$$

and

$$S_n(\hat{m}) = \sum_{j=1}^{\hat{m}} \frac{n\hat{\phi}_j^2}{\hat{\sigma}^2},$$

where $\hat{m}$ is the maximizer of

$$BIC(m) = \sum_{j=1}^{m} \frac{n\hat{\phi}_j^2}{\hat{\sigma}^2} - m \log n, \quad m = 1, \ldots, n-1.$$

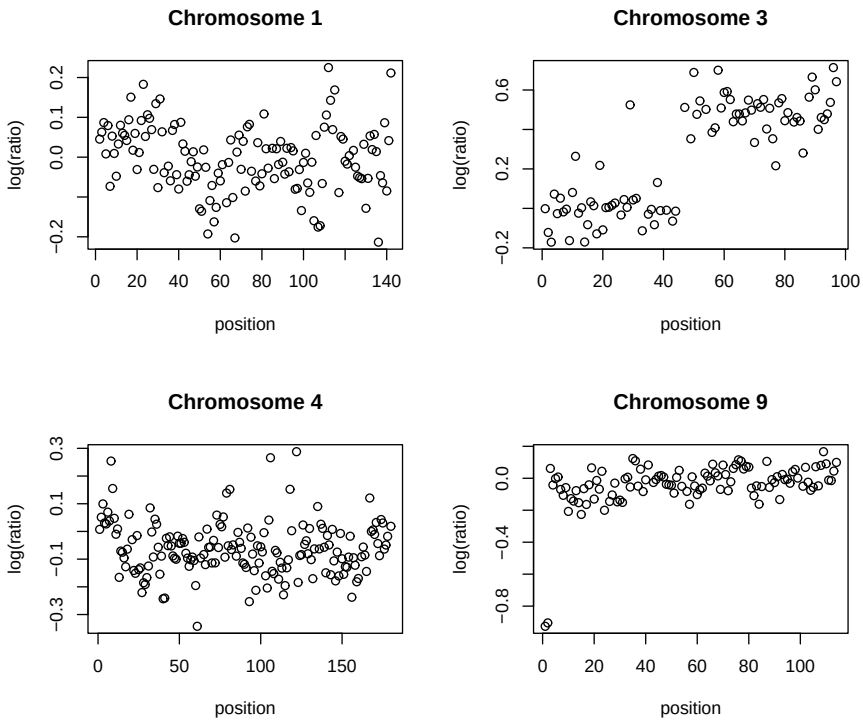

Figure 5: *Microarray data comparing DNA copy number in test and reference samples.*

Table 5: *P-values for Bayes sum and BIC score tests.* For a given data set and statistic, the first P-value is obtained assuming the data are normally distributed and the second by using the bootstrap, as explained in the text.

Data set	Bayes sum	BIC score
Chromosome 1	0.0003, 0.0006	0.0004, 0.0006
Chromosome 3	0, 0	0, 0
Chromosome 4	0.0155, 0.0146	0, 0
Chromosome 9	0, 0	0, 0.0002
Chromosome 9*	0, 0	0.0004, 0.0006

* The chromosome 9 data set with two outliers omitted.

Approximations to P-values were determined in two different ways: assuming normality and by use of the bootstrap. For the former approximation, a random sample of size n was generated from the standard normal distribution, where n is the sample size of the chromosome data in question. The Bayes sum and BIC score statistics were calculated from the data so-generated. This process was repeated independently 10,000 times, and P-values for B_n and $S_n(\hat{m})$ were approximated by comparison of each statistic with the appropriate empirical distribution of 10,000 values. The bootstrap approximation was carried out in exactly the same way, except that samples were drawn from the empirical distribution of the residuals

$$\hat{\epsilon}_i = Y_i - \bar{Y}, \quad i = 1, \ldots, n,$$

rather than from the normal distribution.

The P-values obtained from the process just described are given in Table 5. The two tests give very similar results for three of the four data sets. The P-values for the Bayes sum test were somewhat larger than those of the BIC score test in the case of chromosome 4, but the former test is still significant at level 0.016. The results obtained using the normality assumption were in basic agreement with those obtained using the bootstrap. It is worth noting that when the two obvious outliers

for chromosome 9 are deleted from the analysis, the estimated P-values for the BIC score test increase, while those of the Bayes sum test do not.

9 Conclusions

Frequentist-Bayes tests of the null hypothesis that all of a function's Fourier coefficients are 0 have been proposed. These methods use as test statistic a Laplace approximation to the posterior probability of the null hypothesis. This statistic is used in frequentist fashion, à la the proposal of Aerts, Claeskens and Hart (2004). The posterior probability is derived assuming a very general, nonparametric class of alternative models. Test statistics that are weighted sums of exponentiated squared Fourier coefficients arise naturally from asymptotic approximations of posterior probabilities. A version of such a sum with weights the same as those of a cusum test has excellent power properties in a simulation study. These results suggest that it is not necessary to use adaptive statistics depending on data-driven smoothing parameters in order to obtain an omnibus lack-of-fit test with good overall power properties. A simple weighted sum of independent Fourier components, as suggested by our Bayesian formulation, does the trick.

Additional work, both theoretical and numerical, should be done to investigate properties of the frequentist-Bayes tests. The limiting power results in this paper were based on $1/\sqrt{n}$ local alternatives. Results of Inglot and Ledwina (1996), Eubank (2000) and Kallenberg (2002) show that $1/\sqrt{n}$ local alternatives do not tell the whole story about asymptotic power of lack-of-fit tests. These articles use the device of Kallenberg (1983) known as *intermediate efficiency*. With this device, the level of a test tends to 0 as does the distance of an alternative from H_0. The relative efficiency of two tests is the limiting ratio of sample sizes at which the two tests have the same power. Eubank (2000) shows that, in testing for no effect of a

single regressor, the BIC score test has asymptotic intermediate efficiency equal to that of an optimal test. In the goodness-of-fit setting, Inglot and Ledwina (1996) and Kallenberg (2002) show that BIC score and AIC score tests, respectively, are asymptotically optimal with respect to intermediate efficiency. It is of interest to apply intermediate efficiency ideas to the tests proposed in this paper.

10 Appendix

Here we provide a partial proof of Theorem 5.1. A complete proof may be found at the author's website. We have the following decomposition:

$$B_n = A_n + \Delta_{n1} + \Delta_{n2} + \Delta_{n3},$$

where

$$A_n = g(0) \sum_{i=1}^{n} \frac{\pi_i}{(1-\pi_i)} \exp\left(\frac{n\hat{\phi}_i^2}{2\sigma^2}\right),$$

$$\Delta_{n1} = \sum_{i=1}^{n} \frac{\pi_i}{(1-\pi_i)} \left[g(\hat{\phi}_i) - g(0)\right] \frac{\pi(\hat{\sigma}_{1i})}{\pi(\hat{\sigma}_0)} \exp\left(\frac{n\hat{\phi}_i^2}{2\hat{\sigma}_{1i}^2}\right),$$

$$\Delta_{n2} = g(0) \sum_{i=1}^{n} \frac{\pi_i}{(1-\pi_i)} \left[\frac{\pi(\hat{\sigma}_{1i})}{\pi(\hat{\sigma}_0)} - 1\right] \exp\left(\frac{n\hat{\phi}_i^2}{2\hat{\sigma}_{1i}^2}\right)$$

and

$$\Delta_{n3} = g(0) \sum_{i=1}^{n} \frac{\pi_i}{(1-\pi_i)} \left[\exp\left(\frac{n\hat{\phi}_i^2}{2\hat{\sigma}_{1i}^2}\right) - \exp\left(\frac{n\hat{\phi}_i^2}{2\sigma^2}\right)\right].$$

By assumption, A_n is equal in distribution to

$$\tilde{A}_n = g(0) \sum_{j=1}^{n} \frac{\pi_j}{(1-\pi_j)} \exp\left[(Z_j + \beta_j/\sigma)^2/2\right].$$

Using the conditions imposed on π_j and β_j, $j = 1, 2, \ldots$, the fact that the distribution of $\exp(Z_j^2/2)$ has a regularly varying tail, and Theorem 2.1 of Cline (1983), it is straightforward to verify that A_n converges almost surely as $n \to \infty$. Theorem 2.1 of Cline(1983) relies heavily on Kolmogorov's three-series theorem.

Theorem 5.1 is now proven once it is shown that each of Δ_{ni} tends to 0 in probability as $n \to \infty$. The keys in doing so are (i) the fact that $\max_{1 \le j \le n} |Z_j| = O_p(\sqrt{\log n})$, and (ii) $A_n = O_p(1)$.

References

Aerts, M., Claeskens, G. and Hart, J. D. (1999). Testing the fit of a parametric function. *J. Amer. Statist. Assoc.* **94** 869–879.

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

Aerts, M., Claeskens, G. and Hart, J. D. (2004). Bayesian-motivated tests of function fit and their asymptotic frequentist properties. *Ann. Statist.* **32** 2580–2615.

Baraud, Y., Huet, S. and Laurent, B. (2003). Adaptive tests of linear hypotheses by model selection. *Ann. Statist.* **31** 225–251.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statist. Sci.* **19** 58–80.

Bickel, P. J., Ritov, Y. and Stoker, T. M. (2006). Tailor-made tests for goodness-of-fit to semiparametric hypotheses. *Ann. Statist.* **34** 721–741.

Bogdan, M. (2001). Data driven versions of Neyman's test for uniformity based on a Bayesian rule. *J. Statist. Comput. Simul.* **68** 203–222.

Buckley, M. J. (1991). Detecting a smooth signal: optimality of cusum based procedures. *Biometrika* **78** 253–262.

Chang, M. and Chow, S.-C. (2005). A hybrid Bayesian adaptive design for dose response trials. *J. Biopharmaceutical Statist.* **15** 677–691.

Claeskens, G. and Hjort, N. L. (2004). Goodness of fit via nonparametric likelihood ratios. *Scandinavian Journal of Statistics* **31** 487–513.

Cline, D. (1983). Infinite series of random variables with regularly varying tails, `http://www.stat.tamu.edu/~dcline/Papers/infiniteseries.pdf`. University of British Columbia, Institute of Applied Mathematics and Statistics, Technical Report #83-24.

Cole, G. M. (1997). *Water Boundaries*. New York: John Wiley & Sons, Inc.

Conrad, J., Botner, O., Hallgren, A. and Pérez de los Heros, C. (2003). Including systematic uncertainties in confidence interval construction for Poisson statistics. *Phys. Rev. D* **67** 012002.

De Bruijn, N. G. (1970). *Asymptotic Methods in Analysis*. Amsterdam: North-Holland.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist.* **27** 1012–1040.

Dette, H. and Munk, A. (1998). Validation of linear regression models. *Ann. Statist.* **26** 778–800.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing in the nonlinear structural errors-in-variables model. *Ann. Statist.* **29** 124–152.

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

Eubank, R. L. and Hart, J. D. (1993). Commonality of cusum, von Neumann and smoothing-based goodness-of-fit tests. *Biometrika* **80** 89–98.

Fan, J. and Huang, L.-S. (2001). Goodness-of-fit tests for parametric regression models. *J. Amer. Statist. Assoc.* **96** 640–652.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29** 153–193.

Good, I. J. (1957). Saddle-point methods for the multinomial distribution. *Ann. Math. Statist.* **28** 861–881.

Guerre, E. and Lavergne, P. (2005). Data-driven rate-optimal specification testing in regression models. *Ann. Statist.* **33** 840–870.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528.

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

Horowitz, J. L. and Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69** 599–631.

Inglot, T. and Ledwina, T. (1996). Asymptotic optimality of data driven Neyman's tests. *Ann. Statist.* **24** 1982–2019.

Inglot, T. and Ledwina, T. (2006). Data driven score tests for a homoscedastic linear regression model: asymptotic results. *Probability and Mathematical Statistics* **26** 41–61.

Janssen, A. (2000). Global power functions of goodness of fit tests. *Ann. Statist.* **28** 239–253.

Kallenberg, W. C. M. (1983). Intermediate efficiency, theory and examples. *Ann. Statist.* **11** 170–182.

Kallenberg, W. C. M. (2002). The penalty in data driven Neyman's tests. *Mathematical Methods of Statistics* **11** 323–340.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934.

Kuchibhatla, M. and Hart, J. D. (1996). Smoothing-based lack-of-fit tests: variations on a theme. *J. Nonparametr. Statist.* **7** 1–22.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.* **89** 1000–1005.

Lee, G. and Hart, J. D. (2000). Model selection criteria with data dependent penalty, with application to data-driven Neyman smooth tests. *Nonparametric Statistics* **12** 683–707.

Lehmann, E. (1959). *Testing Statistical Hypotheses.* New York: John Wiley & Sons.

Lucito, R., West, J., Reiner, A., Alexander, D., Esposito, D., Mishra, B., Powers, S., Norton, L. and Wigler, M. (2000). Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Research* **10** 1726–1736.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

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

Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis.* Boston: Birkhäuser.

Rayner, J. C. W. and Best, D. J. (1989). *Smooth Tests of Goodness of Fit.* New York: Oxford University Press.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. and Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29** 263–264.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498.

Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25** 613–641.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86.

Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.* **26** 1215–1241.