*Journal of Statistical Computation and Simulation*,
2003, Vol. 73(1), pp. 71–82

Taylor & Francis
Taylor & Francis Group

# BINNING METHODOLOGY FOR NONPARAMETRIC GOODNESS-OF-FIT TEST

CHUNMING ZHANG* and BIN CHENG†

*Department of Statistics, University of Wisconsin, Madison, WI 53706-1685, USA*

*(Received 29 September 2001; In final form 22 April 2002)*

In this paper, we develop a simple procedure for goodness-of-fit test. One feature of its construction is the use of "binning" transform of the sample observations. This enables us to define suitably a design variable and a response variable and they are shown to follow asymptotically a nonparametric regression model, under which the regression function corresponds to the population density function. As a result, the problem of testing density is formulated as assessing the parametric fits of a regression function against its nonparametric alternatives. Many nonparametric tests based on curve estimation technique could be incorporated, and we shall extend here the tests (Fan *et al.*, 2001; Zhang, 2001) based on local polynomial smoother combined with nonparametric likelihood-ratio. The resulting procedure is capable of testing not only the conventional simple null hypotheses, but also certain types of composite null hypotheses. Simulation studies illustrate the power of this procedure.

*Keywords*: Goodness-of-fit; Binning procedure; Local polynomial regression; Generalized likelihood ratio test; Smoothing parameter

## 1 INTRODUCTION

For a given set of observations $X_1, \ldots, X_n$ from a distribution with a continuous distribution function $F$, one important task is directed at developing procedures that are useful for drawing statistical inferences about the underlying population. Namely, one wishes to test $F(x) = F_0(x)$, where $F_0(x)$ denotes the hypothetical distribution. For a simple null hypothesis where the form of $F_0(x)$ is completely specified, such as uniformity or normality, many existing tests could be employed in a straightforward manner. One could also carry out tests based on the transformed sample $F_0(X_1), \ldots, F_0(X_n)$, leading to tests for uniformity on the interval $[0, 1]$. However under many of the practical situations, only partial information about $F_0$ could be specified, in other words, the null hypothesis about $F_0$ is composite. Indeed, only a few proposals have been made for handling extensions to cover the case of composite null hypotheses.

There is a long list of literature on goodness-of-fit tests. This includes Pearson's Chi-square test, based on the $L_2$ distance between the null density and the histogram density estimator; Kolmogorov–Smirnov (KS) test and Cramér–Von Mises (CVM) test, based on the empirical

* Corresponding author. E-mail: cmzhang@stat.wisc.edu
† E-mail: bcheng@stat.wisc.edu

ISSN 0094-9655 print; ISSN 1563-5163 online 
DOI: 10.1080/0094965021000015440

distribution function; and Neyman's (1937) smooth test, based on the exponential family assumption of the underlying distribution. The power of Pearson's Chi-square test depends heavily on the number of histogram cells. Data driven versions of Pearson's Chi-square test for uniformity were studied in Bodgan (1995). The KS and CVM tests suffer from low power against alternatives containing high frequency components (Fan, 1996). Neyman's smooth test requires selecting the number of components, which acts as a smoothing parameter. Further development along this line focuses on seeking the number of components in a data-driven manner, which yields the adaptive Neyman's tests for uniformity (Ledwina, 1994; Kallenberg and Ledwina, 1995; Fan, 1996).

Kernel-based density estimation (KDE) introduced by Rosenblatt (1956) has also attracted several research efforts in tests for goodness-of-fit. As in the Neyman's smooth test, it is assumed implicitly that $F$ possesses a probability density $f$, and therefore the original testing problem becomes equivalent to checking

$$H_0: f = f_0 \quad \text{against} \quad H_1: f \neq f_0, \tag{1.1}$$

where $f_0$ corresponds to the probability density of $F_0$. Bickel and Rosenblatt (1973) proposed a test statistic based on the weighted $L_2$ distance between the KDE of $f$ and its expected value computed under the null hypothesis, in which $f_0$ is fully specified. A test based on the derivative of a KDE was considered in Huang (1997). One drawback of these kernel-based tests arises from the "boundary bias" problem well-known in kernel density estimation. To improve the performance of KDE, boundary kernel functions (Gasser and Müller, 1979) may be employed; however the resulting test procedure becomes complicated in both implementation and asymptotic analysis. Furthermore, in carrying out the tests, the tuning method for bandwidth parameter in KDE is not available as yet.

In this paper, we take a nonparametric regression model approach to the identification of a density function. With a "binning" transform defined in Section 2, we will see that a density function can be regarded as a nonparametric regression function. As a result, many tests constructed from nonparametric smoothing techniques could be applied, such as the kernel regression in Azzalini *et al.* (1989), Azzalini and Bowman (1993), and Härdle and Mammen (1993), among others. In this paper, we mainly concentrate on the tests developed by local polynomial smoother combined with "generalized likelihood-ratio" (Fan *et al.*, 2001) and with "multi-scale generalized likelihood-ratio" (Zhang, 2001), abbreviated as GLR test and MGLR test respectively. The purpose of the present paper intends to address three questions: how to test a composite null hypothesis, how to ameliorate boundary bias effects, and how to select smoothing parameter? For the first question, we shall see that our procedure benefits from the appealing feature that the null distribution of the GLR-type of tests based on local polynomial smoother is asymptotically free of the nuisance parameters, when the regression curve admits a polynomial structure. With regard to the second issue, gains can also be made from the superior behaviour of local polynomial approach at the edges of the sample space as compared to the kernel regression counterpart. Moreover, we could use the data-dependent optimal choice of smoothing parameter proposed in Zhang (2001). In this way, one could carry out simply the (M)GLR test for a density function. Computationally, the proposed new test is very simple to implement. The remaining key problem lies in bridging connections between density estimation and nonparametric regression, for which the "binning" strategy proves to be quite useful.

A similar binning idea was also considered in the adaptive-Neyman test (ANT) of Fan (1996), from which our testing procedure is essentially different. The goal of ANT is to assess the appropriateness of zero mean of a multivariate normal random vector, based on an orthogonal (Fourier) transform of the binned response vector, therefore no regression model or regression function will be incorporated in ANT.

The remainder of the paper is organized as follows. Section 2 contains the description of the binning transform and the proposed (M)GLR test procedure. Applications to goodness-of-fit tests for simple and composite null hypotheses are presented in Section 3, while in Section 4 we demonstrate the simulation studies on the powers of our proposed tests, in comparison to several other existing procedures. Concluding remarks are given in Section 5.

## 2 BACKGROUND

### 2.1 Binning Transform

We now illustrate how the binning procedure operates in goodness-of-fit. Assume that the true density function $f(x)$ has a bounded support, $\mathcal{I} = [0, 1]$, without loss of generality. Partition $\mathcal{I}$ into $N$ subintervals $\{\mathcal{I}_j, j = 1, \ldots, N\}$, of equal length $\Delta = 1/N$ for simplicity. Let $x_j$ be the center point of $\mathcal{I}_j$. Denote by $n_j$ the number of observations from the sample $\{X_i\}_{i=1}^n$ falling into the $j$th bin, namely, $n_j = \sum_{i=1}^n I(X_i \in \mathcal{I}_j)$, where $I(\cdot)$ represents the indicator function. Then it follows trivially that

$$(n_1, \ldots, n_N)^{\mathrm{T}} \sim \text{Multinomial}(n; p_1, \ldots, p_N), \quad \text{where } p_j = \int_{\mathcal{I}_j} f(x)\,\mathrm{d}x, \ \ j = 1, \ldots, N.$$

Provided that $N$ gets sufficiently large and hence the partition is finer, we could assume that the approximation, $p_j \approx f(x_j)\Delta$, holds with good accuracy. Setting $y_j = (n\Delta)^{-1} n_j$ will thus result in

$$E(y_j) = (n\Delta)^{-1} n p_j \approx f(x_j),$$

$$\mathrm{cov}(y_j, y_k) = \begin{cases} (n\Delta)^{-2} n p_j (1 - p_j) \approx (n\Delta)^{-1} f(x_j), & \text{if } j = k, \\ -(n\Delta)^{-2} n p_j p_k \approx -n^{-1} f(x_j) f(x_k), & \text{if } j \neq k. \end{cases}$$

Now we take $x_j$ as a predictor variable, and $y_j$ a response variable, respectively. Then the bivariate data $\{(x_j, y_j)\}_{j=1}^N$ can be reasonably assumed to follow a nonparametric regression model represented by

$$y_j = m_1(x_j) + \sigma_1(x_j)\varepsilon_j, \quad j = 1, \ldots, N, \tag{2.1}$$

where the regression function and variance function are expressed as

$$m_1(x) = f(x), \quad \text{and} \quad \sigma_1^2(x) = (n\Delta)^{-1} f(x), \tag{2.2}$$

whereas the errors $\varepsilon_j$ have zero mean, unit variance and are weakly-dependent. In this manner, the binning procedure transforms the original goodness-of-fit problem (1.1) into assessing the functional form of the smooth regression function under (2.1), the nonparametric regression model. If $N$ is chosen in a way that $n/N \to \infty$, then (2.2) suggests readily that regression model (2.1) possesses very high values of signal to noise ratio, and thus the regression function $f$ can be estimated and identified efficiently.

### 2.2 Test Statistics

For a nonparametric regression model with homoscedastic errors, many diagnostic tests for assessing the regression function have been developed based on nonparametric curve fitting technique, such as kernel smoothing, local polynomial regression (Wand and Jones, 1995), and smoothing spline (Eubank, 1999). It should be stressed that, in principle, it is possible

to integrate any of these tests into our present setup (2.1), in the event of constant variance function, which is equivalent to testing uniformity. However, for testing the null other than uniformity, application of this procedure will encounter difficulty. For this reason, we shall extend the idea of "generalized likelihood ratio" (Fan *et al.*, 2001) test, which possesses the desirable property of adaptation to heteroscedasticity.

To briefly describe the GLR test, let us begin by considering a general set-up of nonparametric regression model,

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \ldots, n, \tag{2.3}$$

with independent observations $\{(X_i, Y_i)\}_{i=1}^n$, of the predictor variable $X$ and the response variable $Y$. Assume that $X$ has a density function $p(x)$, with a bounded support $\Omega$. The errors in (2.3) are assumed to fulfill $E(\varepsilon_i|X_i) = 0$ and $\text{var}(\varepsilon_i|X_i) = 1$. Call $m(x) = E(Y|X = x)$, and $\sigma^2(x) = \text{var}(Y|X = x)$ the regression function and variance function, respectively. In particular, when the weak correlation structure of errors in model (2.1) is negligible, model (2.1) can be treated as a special case of (2.3). Back to model (2.3), suppose that we wish to test a hypothesis, asserting that

$$H_0\colon m(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_k x^k, \tag{2.4}$$

with the vector of unknown parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_k)^{\mathrm{T}}$. Under the null assumption above, the parameters can be estimated consistently, for instance, by least squares estimates. Call them $\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_k$. When the alternative holds, the unknown regression curve can be fitted nonparametrically. For example, the $q$th degree local polynomial estimate $\hat{m}_h(\cdot)$, as applied to estimate $m(\cdot)$ at a fitting point $x_0$, corresponds to the first component of the coefficient vector $(\beta_0, \beta_1, \ldots, \beta_q)^{\mathrm{T}}$ that minimizes $\sum_{i=1}^n \{Y_i - \beta_0 - \beta_1(X_i - x_0) - \cdots - \beta_q (X_i - x_0)^q\}^2 K\{(X_i - x_0)/h\}$, where $K(\cdot)$ and $h > 0$ are referred to as kernel function and bandwidth parameter respectively. The GLR statistic based on model (2.3) is constructed in terms of

$$\lambda_n(h) = \frac{n}{2} \log \frac{\text{RSS}_0}{\text{RSS}_1(h)}, \tag{2.5}$$

where $\text{RSS}_0 = \sum_{i=1}^n \{Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 X_i - \cdots - \hat{\alpha}_k X_i^k\}^2$, and $\text{RSS}_1(h) = \sum_{i=1}^n \{Y_i - \hat{m}_h(X_i)\}^2$. Certainly, in the event that $\boldsymbol{\alpha}$ at (2.4) is given, its true value will be used for obtaining $\text{RSS}_0$.

The finite sample performance of the GLR statistic and the effective data-based rule for bandwidth selection have been studied in Zhang (2001). Furthermore, based on the power considerations, a "multi-scale generalized likelihood ratio" (MGLR) statistic was proposed in the same paper. Under mild regularity conditions, it was shown there that, when the null hypothesis (2.4) holds and $q \geq k$,

$$\max_{1 \leq j \leq J} \frac{r_{\mathcal{K}}(\int \sigma^2(x)\,dx / \int \sigma^4(x)\,dx) E\{\sigma^2(X)\}\lambda_n(h_j) - [r_{\mathcal{K}} c_{\mathcal{K}} (\{\int \sigma^2(x)\,dx\}^2 / \int \sigma^4(x)\,dx) h_j^{-1} - (r_{\mathcal{K}} \mathcal{K}^2(0)/2nh_j^2)(\int \sigma^2(x)\,dx / \int \sigma^4(x)\,dx) \int (\sigma^2(x)/p(x))\,dx]}{\sqrt{2 r_{\mathcal{K}} c_{\mathcal{K}} (\{\int \sigma^2(x)\,dx\}^2 / \int \sigma^4(x)\,dx) h_j^{-1}}} \xrightarrow{\mathcal{L}} \max_{1 \leq j \leq J} Z_j, \tag{2.6}$$

where $\xrightarrow{\mathcal{L}}$ denotes converges in distribution. In this result, $\mathcal{K}$ denotes the equivalent kernel function induced from the $q$th degree local polynomial fit (refer to Appendix 1 for the expression of $\mathcal{K}$), with $c_{\mathcal{K}} = \mathcal{K}(0) - 2^{-1}\mathcal{K} * \mathcal{K}(0)$, and $r_{\mathcal{K}} = (\mathcal{K}(0) - 2^{-1}\mathcal{K} * \mathcal{K}(0))/(\int \{\mathcal{K}(t) - 2^{-1}\mathcal{K} * \mathcal{K}(t)\}^2\,dt)$, where $*$ denotes the convolution operator, and $(Z_1, \ldots, Z_J)^{\mathrm{T}} \sim N(0, \mathcal{R})$ is a

$J$-variate normal random vector with mean zero and correlation matrix $\mathcal{R}$, the entries of which rely solely on the kernel function and ratios between bandwidths $\{h_1, \ldots, h_J\}$, but not on the parameter $\boldsymbol{a}$. In such a case, the MGLR test rejects the null hypothesis (2.4) for large values of MGLR, namely, large values of $\max_{1\le j\le J} \mathrm{GLR}(h_j)$, where $\mathrm{GLR}(h_j)$ represents the standardized form of $\lambda_n(h_j)$ or, the fraction on the left-hand side of (2.6). It is readily observed that for $J = 1$, the MGLR test reduces to the GLR test. However, the MGLR test enjoys the adaptive feature that it is nearly as powerful as if a GLR test with a favorable bandwidth were used (Zhang, 2001). For convenience, the numerical work in this paper assumes the $K$ to be Epanechnikov kernel function used frequently in practice. The associated quantities $\mathcal{K}(0)$, $c_{\mathcal{K}}$ and $r_{\mathcal{K}}$, as well as the critical values of the MGLR test, were tabulated in Zhang (2001) (see Tab. I in Appendix 2 for Epanechnikov kernel, $\mathcal{K}(0)$, $c_{\mathcal{K}}$ and $r_{\mathcal{K}}$); $P$-value calibration was also described there in detail. Furthermore, under the null hypothesis (2.4), if the model (2.3) is homoscedastic with $\sigma^2(x) \equiv \sigma^2$ for all $x$, then whether the value of $\sigma^2$ is known or not, we shall establish from (2.6) that,

$$\max_{1\le j\le J} \frac{r_{\mathcal{K}}\lambda_n(h_j) - \{r_{\mathcal{K}}c_{\mathcal{K}}|\Omega|h_j^{-1} - (r_{\mathcal{K}}\mathcal{K}^2(0)/2nh_j^2)\int(1/p(x))\,dx\}}{\sqrt{2r_{\mathcal{K}}c_{\mathcal{K}}|\Omega|h_j^{-1}}} \xrightarrow{\mathcal{L}} \max_{1\le j\le J} Z_j, \tag{2.7}$$

where $|\Omega|$ represents the length of $\Omega$. Statements (2.6) and (2.7) provide useful insights into understanding the sampling distributions of our proposed goodness-of-fit test statistics.

## 3 GOODNESS-OF-FIT TEST

As exhibited in (2.2), the linkage between density and regression functions, allows us to test directly the density itself, under the nonparametric regression model (2.1). With the binning method described in Section 2.1 above, the equally-spaced design variable $x_j$ in (2.1) can be regarded as distributed uniformly on the interval [0, 1]. Bearing in mind that $\{(x_j, y_j)\}_{j=1}^N$ is the current set of observations to obtain $\mathrm{RSS}_0$, $\mathrm{RSS}_1$ and $\lambda_N$, one can therefore put $\Omega = [0, 1]$ and $p(x) \equiv 1$ in (2.6) and (2.7). There, the grid of bandwidths $\{h_1, \ldots, h_J\}$ is chosen according to the empirical rule proposed in Zhang (2001); that is, set $J = 3$, and take the bandwidth grid $\{1.5^{-1}h_0, h_0, 1.5h_0\}$, with $h_0 = \mathrm{std}(\{x_j\})N^{-2/(4q+5)}$, where $N^{-2/(4q+5)}$ stands for the optimal rate of bandwidth (Fan *et al.*, 2001) for nonparametric hypothesis testing. According to Section 2.1, the construction of $\{x_j\}$ does not rely on the configuration of $\{X_i\}$. This lends the empirical (constant) bandwidth easily to simple hand calculations. For our current applications to goodness-of-fit problem (1.1) and the induced model (2.1), we can also verify that, when $f_0$ is a polynomial function, the conclusion of (2.6) continues to hold with $\sigma^2(x)$ replaced by $(n\Delta)^{-1}f_0(x)$, leading to

$$\max_{1\le j\le J} \frac{r_{\mathcal{K}}\lambda_N(h_j) - \{r_{\mathcal{K}}c_{\mathcal{K}}h_j^{-1} - (r_{\mathcal{K}}\mathcal{K}^2(0)/2Nh_j^2)\}}{\sqrt{2r_{\mathcal{K}}c_{\mathcal{K}}h_j^{-1}\int f_0^2(x)\,dx}} \xrightarrow{\mathcal{L}} \max_{1\le j\le J} Z_j. \tag{3.1}$$

Notice in (3.1) the occurrence of $f_0$. In the following subsections, we will address two situations about, $f_0$, the null density function.

### 3.1 Simple Null Hypothesis

We first consider the case where $f_0$ is completely specified. In this instance, we offer two options of conducting tests for (1.1). In the first option, we require that $f_0(x)$ relates to $x$

via a given polynomial function. Then we can use directly MGLR test (3.1) for $H_0$: $m_1(x) = f_0(x)$, under the model,

$$y_j = m_1(x_j) + (n\Delta)^{-1/2} m_1^{1/2}(x_j)\varepsilon_j, \quad 1 \le j \le N. \tag{3.2}$$

Alternatively, the equivalent problem of testing uniformity, based on the transformed sample $X_i^* = F_0(X_i)$, $i = 1, \ldots, n$, amounts to testing for the hypothesis of no predictor effect in regression curve. That is to say, assess the adequacy of $H_0$: $m_1(x) \equiv 1$, under the model

$$y_j^* = m_1(x_j) + (n\Delta)^{-1/2} m_1^{1/2}(x_j)\varepsilon_j, \quad 1 \le j \le N, \tag{3.3}$$

where $\{y_j^*\}_{j=1}^N$ represent the binned responses when the binning procedure is applied to the set of transformed sample $\{X_i^*\}_{i=1}^n$. After this adjustment, one can directly proceed with (2.7) by replacing $\{(X_i, Y_i)\}_{i=1}^n$ with $\{(x_j, y_j^*)\}_{j=1}^N$. This step is identical to using (3.1), in which we have $|\Omega| = 1$, $p(x) \equiv 1$, and the null density of $X_i^*$ equals one. Compared with the first option of testing $f_0$ restricted to be a given polynomial, the second option provides more flexibility to accommodate various types of $f_0$.

### 3.2 Composite Null Hypothesis

We now consider more interesting cases where $f_0(x)$ may not be fully known, but is assumed to be a member of a parametric family of densities, such as,

$$f_0(x; \boldsymbol{\alpha}) = (\alpha_0 + \alpha_1 x + \cdots + \alpha_k x^k)^2, \tag{3.4}$$

for some unspecified parameters $\{\alpha_j\}$. In this context, the uniform density corresponds to either degree $k = 0$, or degree $k \ge 1$ but with $\alpha_1 = \cdots = \alpha_k = 0$. More generally, this family includes the "symmetric Beta densities" (Wand and Jones, 1995, p. 31) of the form

$$\left\{B\left(\frac{1}{2}, \ell + 1\right)\right\}^{-1} (1 - x^2)_+^{\ell}, \quad \ell = 0, 2, 4, \ldots, \tag{3.5}$$

where $x_+ = \max(x, 0)$, and $B(\cdot, \cdot)$ represents the beta function. In particular, a Gaussian density belongs to this family as the index $\ell$ above tends to infinity. Clearly, the null hypothesis above is composite and we are not aware of other conventional tests that can be applicable. In this case, the left-hand side of (3.1) contains the unknown quantity $f_0(x; \boldsymbol{\alpha})$ and thus can not be used as a test statistic. One may substitute $f_0$ by its consistent estimate, such as the kernel density estimate. However doing so will cause again the boundary bias problem, and hence deteriorate the power. An approach more effective is by means of "variance stabilizing transform" so as to achieve homoscedasticity in model (2.1). Naturally, the Poisson-type of link at (2.2) between mean and variance suggests the square-root transform $\sqrt{y_j}$. As a consequence, $\{(x_j, \sqrt{y_j})\}_{j=1}^N$ can be viewed as following a homoscedastic nonparametric regression model,

$$\sqrt{y_j} \approx m_2(x_j) + \sigma\varepsilon_j, \quad j = 1, \ldots, N, \tag{3.6}$$

where $m_2(x) = \sqrt{f(x)}$, and $\sigma > 0$ serves as a nuisance parameter. Under this model, one could then directly apply (2.7) to carry out tests for the presence of $\sqrt{f_0(x; \boldsymbol{\alpha})}$ or, equivalently, for the validity of a polynomial regression function. Hence, the asymptotic null distribution of (M)GLR will be free of the nuisance parameter $\boldsymbol{\alpha}$. Formally, the nonparametric regression formulation (3.6) for $\{(x_j, \sqrt{y_j})\}_{j=1}^N$ is justified by the following theorem.

THEOREM 1 *Let $f(x) > 0$ be bounded and continuous. Suppose that $N \to \infty$ and $Nn^{-1}\log(n) \to 0$, as $n \to \infty$. Then by changing a probability space, if necessary, we have the following stochastic representation:*

$$\sqrt{y_j} = \theta_j + \frac{\varepsilon_j}{2\sqrt{n/N}} + O_P\{n^{-1/2} + Nn^{-1}(\log n)^{1/2}\} \tag{3.7}$$

*uniformly in $j$, where $\{\varepsilon_j\}$ is a sequence of i.i.d. $N(0, 1)$ variables and $\theta_j = \sqrt{p_j N}$. Furthermore, if $f'(x)$ is continuous for $x \in [0, 1]$, and $f''(x)$ exists for $x \in (0, 1)$, then* (3.7) *becomes $\sqrt{y_j} = \sqrt{f(x_j)} + \varepsilon_j/(2\sqrt{n/N}) + O_P\{N^{-2} + n^{-1/2} + Nn^{-1}(\log n)^{1/2}\}$.*

Evidently, Theorem 1 reveals that, when the sample size $n$ and partition number $N$ grow sufficiently large, little information of $y_j$ will be lost from the square-root transform. The main ingredients of the proof are analogous to those used in Fan (1996, Theorem 4.1) and are thus omitted here.

The parameter $N$ controls the degree to which the data are binned to build the density function/regression function so that our proposed tests can be applied. Compared with the bandwidth parameter selectors familiar in kernel density estimation or nonparametric regression technique, this quantity $N$ appears relatively easier to select for practical usage. For instance, it is readily seen that the choice $N = O\{n^r(\log n)^s\}$, either taking $0 < r < 1$ and $s \geq 0$, or taking $r = 0$ and $s > 0$, satisfies the conditions of Theorem 1. Throughout our simulations described next, we shall set $N = n^{2/3}\log(n)$.

## 4 SIMULATIONS

To demonstrate the power comparison of our proposed (M)GLR test with other existing tests, we perform simulation studies. We shall first consider in Examples 1 up to 4, testing simple null hypotheses. To facilitate the comparisons, the sequences of alternatives are taken from those studied in Fan (1996, p. 682). The second option, as described in Section 3.1, will be adopted in our implementation. That is, the test statistic (2.7) combined with local linear fit ($q = 1$) will be applied to tests for the null hypothesis $H_0$: $m_1(x) \equiv 1$, under model (3.3). The empirical critical values are simulated at nominal level 5% and sample size $n = 200$, on 10,000 independent samples. In Figure 1, we illustrate the estimated density functions (using KDE method) of the GLR and MGLR test statistics under the null hypothesis of Example 1; similar plots for Examples 2–4 have also been obtained but are omitted herein. These graphs provide evidence that the asymptotic normal distributions at (2.7) are reflected in the finite-sample situation of model (3.3) with weakly dependent random noises. The empirical powers are estimated by the proportion of observed rejections in 1600 samples of size $n$. The power curves of GLR($h_j$), $j = 1, 2, 3$, and MGLR are displayed in Figure 2. For power comparison, the KS test, CVM test, ANT (Fan, 1996), and the $H_{15}$ test recommended in Bodgan (1995) are also included, where the choice for the number of bins used in ANT follows the suggestion of Fan (1996, p. 682).

*Example 1* Consider the problem

$$H_0\colon F = \text{Uniform}(-1, 1) \quad \text{versus} \quad H_1\colon F = F_\mu, \text{ with density}$$
$$F'_\mu(x) = 2^{-1} + \frac{2x(\mu - |x|)}{\mu^2 I(|x| < \mu)}, \quad 0 \leq \mu \leq 1.$$

The null hypothesis corresponds to the index $\mu = 0$.

This example serves to reflect how a test is capable of detecting local features.

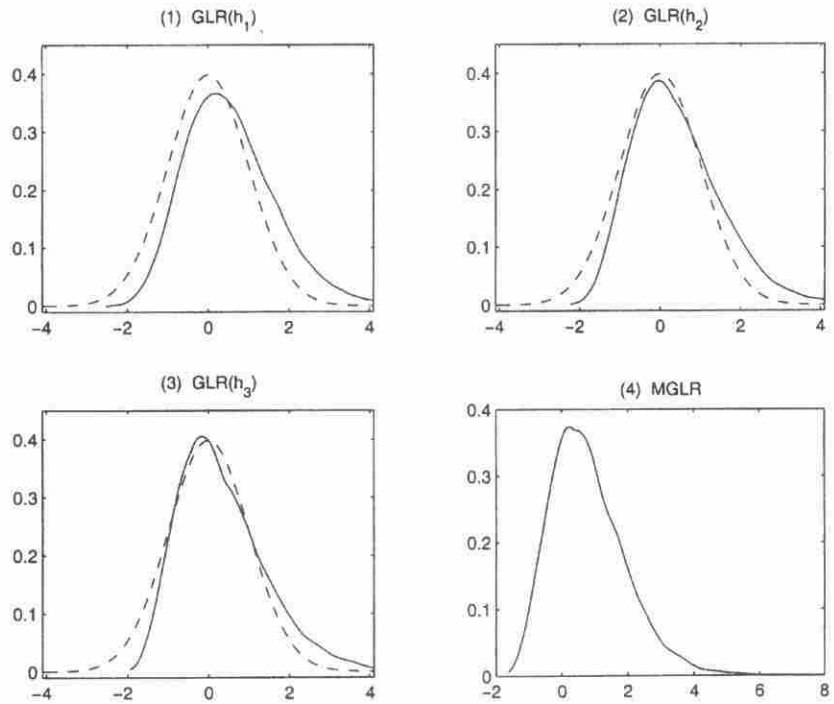


FIGURE 1 The estimated densities for the test statistics GLR($h_j$), $j = 1, 2, 3$, and MGLR test statistics at (2.7), for $n = 200$, based on 10,000 simulations. In panels (1)–(3), solid curve – the estimated density, based on kernel density estimation; dashed curve – standard normal density.

*Example 2* This example tests the global features with different frequencies:

$$H_0\colon F = \text{Uniform}(-1, 1) \quad \text{versus} \quad H_1\colon F = F_\sigma, \text{ with density}$$
$$F'_\sigma(x) = 2^{-1}\{1 + \sin(2\pi\sigma x)\}, \quad 0 \le \sigma \le 5.$$

The null hypothesis corresponds to $\sigma = 0$.

*Example 3* This example is designed to identify the normal scale mixture model

$$H_0\colon F = N(0, 1) \text{ versus}$$
$$H_1\colon F = 0.8N\left(0, \frac{1}{0.8 + 0.2\sigma^2}\right) + 0.2N\left(0, \frac{\sigma^2}{0.8 + 0.2\sigma^2}\right), \quad \frac{1}{8} \le \sigma \le 1.$$

When $\sigma = 1$, the alternative hypothesis agrees with the null.

*Example 4* Consider the normal mean mixture model

$$H_0\colon F = N(0, 1) \text{ versus } H_1\colon F = 0.7N\left(\frac{\mu}{0.7}, 1\right) + 0.3N\left(\frac{-\mu}{0.3}, 1\right), \quad 0 \le \mu \le 1.$$

When $\mu = 0$, the alternative hypothesis corresponds to the null.

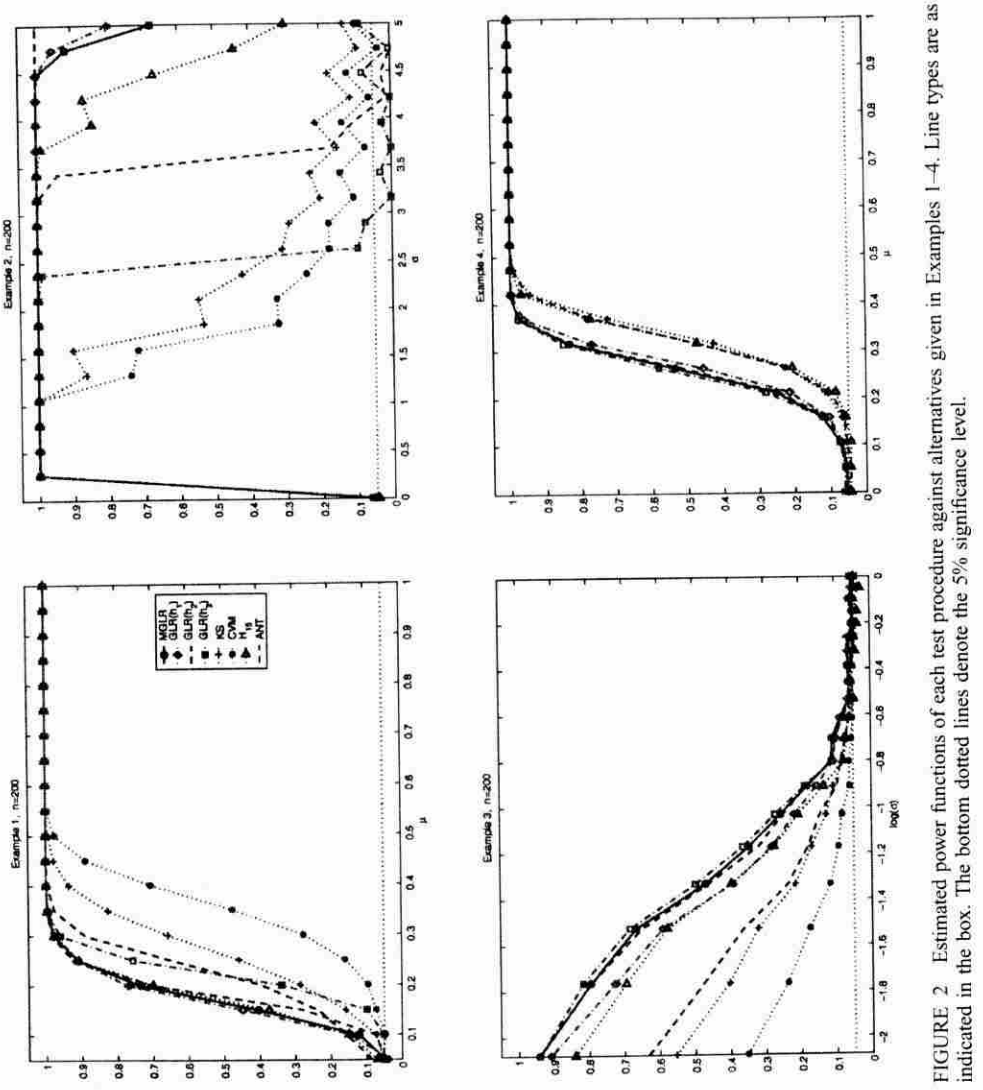


FIGURE 2 Estimated power functions of each test procedure against alternatives given in Examples 1–4. Line types are as indicated in the box. The bottom dotted lines denote the 5% significance level.

In summary, the power of the GLR test depends on the choice of bandwidth parameter. At each fixed alternative under consideration, the highest power of the three GLR tests occur in Examples 1 and 2 at the smallest bandwidth $h_1$, whereas the largest bandwidth $h_3$ in Examples 3 and 4. Nonetheless, the MGLR test performs always close to the best of the three GLR tests, which is GLR($h_1$) in Examples 1 and 2, and GLR($h_3$) in Examples 3 and 4. This *adaptive* feature is important because it means that, we could always take the MGLR test to avoid loss of power, although GLR tests demand different amount of smoothing in detecting alternatives of different patterns.

Moreover, examination of Figure 2 indicates that our proposed MGLR test outperforms the other types of existing tests, except in Example 2 where ANT is more powerful against alternatives of very high frequency.

Next we consider Example 5 which consists of composite null hypotheses. The empirical powers are evaluated in ways similar to those described above for simple null hypotheses.

*Example 5* Consider the composite null hypothesis

$$H_0: f(x) = (\alpha_0 + \alpha_1 x + \alpha_2 x^2)^2 \text{ versus}$$
$$H_1: f(x) = (1-\theta)(\alpha_0 + \alpha_1 x + \alpha_2 x^2)^2 + \theta\{1 + \sin(2\pi x)\}/2, \quad 0 \le \theta \le 1,$$

where $f(x)$ is supported on $[-1, 1]$, and the nuisance parameters $(\alpha_0, \alpha_1, \alpha_2)^{\mathrm{T}}$ are present. The null hypothesis corresponds to the case $\theta = 0$. Evidently, all the other existing tests mentioned above can not be exploited for this example. According to Section 3.2, this example is equivalent to detecting departures from the hypothesized regression curve $m_2(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ under model (3.6). To demonstrate the capacity of our test statistic (2.7), local quadratic fitting method ($q = 2$) will be conducted. Furthermore, the values, $\alpha_0 = -\alpha_2 = \sqrt{15/16}$ and $\alpha_1 = 0$, will be used in the alternatives from which simulated samples are to be generated. The resulting power comparisons are depicted in Figure 3. Once again, we observe that all GLR-type of test statistics achieve the given level of significance. Specifically, GLR($h_3$) offers the maximal power of GLR($h_j$), $j = 1, 2, 3$, and consistently outperforms MGLR. Nonetheless, the adaptive feature of the MGLR test continues to be reflected in this example.

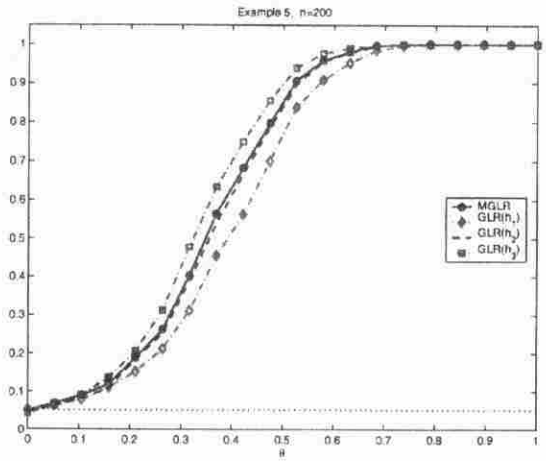


FIGURE 3 Estimated power functions of each test procedure against alternatives given in Example 5. The bottom dotted line denotes the 5% significance level.

## 5 CONCLUSION

We have demonstrated that the (M)GLR test in regression models performs reasonably well in tests for goodness-of-fit based on binned data. This procedure ameliorates the drawbacks of boundary bias problem arising from kernel density estimation approach, and benefits from the properties of local polynomial regression techniques, such as, data-driven selection of optimal smoothing parameter and allowing for extensions to certain types of composite null hypotheses. The simplicity, flexibility, adaptive feature, and competitive power of our proposed MGLR test indicate that it is a useful diagnostic tool. Extending our current scope of goodness-of-fit to multivariate setting and multiple-sample situation will be interesting topics for future research.

### *Acknowledgement*

The authors thank the associate editor and the referee for valuable comments that improved the presentation. Part of the work on this paper was performed while the first author was visiting Centre for Mathematics and its Applications at the Australian National University. The first author is most grateful to Peter Hall for financial support.

### *References*

Azzalini, A. Bowman, A. N. and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*, **76**, 1–11.

Azzalini, A. and Bowman, A. N. (1993). On the use of nonparametric regression for checking linear relationships. *J. Roy. Statist. Soc. Ser. B*, **55**, 549–557.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviation of density function estimates. *Ann. Statist.*, **1**, 1071–1095.

Bodgan, M. (1995). Data driven versions of Pearson's Chi-square test for uniformity. *J. Statist. Comput. Simul.*, **52**, 217–237.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* Marcel Dekker Inc., New York, Basel.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.*, **91**, 674–688.

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

Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions. In: Gasser, T. and Rosenblatt, M. (Eds.), *Smoothing Techniques for Curve Estimation.* Springer-Verlag, New York, pp. 23–68.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**, 1926–1947.

Huang, L. S. (1997). Testing goodness-of-fit based on a roughness measure. *J. Amer. Statist. Assoc.*, **92**, 1399–1402.

Kallenberg, W. C. M. and Ledwina, T. (1995). Consistency and Monte-Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Ann. Statist.*, **23**, 1594–1608.

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

Neyman, J. (1937). Smooth test for goodness of fit. *Skand. Aktuarietidskr*, **20**, 149–199.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Zhang, C. M. (2001). Adaptive tests of regression functions via multi-scale generalized likelihood ratios, *Technical Report 1026*, Dept. of Statistics, University of Wisconsin-Madison.

## APPENDIX 1

Set $\mu_j = \int t^j K(t)\,\mathrm{d}t, j = 0, 1, 2, \ldots$. Define the vector $\boldsymbol{e}_{1,q+1} = (1, 0, \ldots, 0)^{\mathrm{T}}$ of length $q+1$, and the $(q+1) \times (q+1)$ matrix $\boldsymbol{S}_q = (\mu_{j+l})_{0 \le j,l \le q}$. Then the equivalent kernel function $\mathcal{K}$ of the $q$th degree local polynomial fit is expressed in the form $\mathcal{K}(t) = \boldsymbol{e}_{1,q+1}^{\mathrm{T}} \boldsymbol{S}_q^{-1} (1, t, \ldots, t^q)^{\mathrm{T}} K(t)$, for $t \in \mathbb{R}$.

## APPENDIX 2

Epanechnikov kernel function is define by, $K(t) = 3/4(1 - t^2)$ if $|t| < 1$ and 0 otherwise.

TABLE I Values of $r_K$, $c_K$ and $\mathcal{K}(0)$ Induced from the $q$th Degree Local Polynomial Estimation with Epanechnikov Kernel Function (Excerpted from Zhang, 2001).

| | $q = 0$ or 1 | $q = 2$ or 3 | $q = 4$ or 5 |
|---|---|---|---|
| $r_K$ | 2.1153 | 1.9755 | 1.9336 |
| $c_K$ | 0.4500 | 0.7812 | 1.1043 |
| $\mathcal{K}(0)$ | 0.7500 | 1.4062 | 2.0507 |