**TEST manuscript No.**
(will be inserted by the editor)

# Nonparametric inference with generalized likelihood ratio tests[⋆]

**Jianqing Fan · Jiancheng Jiang** [⋆]

the date of receipt and acceptance should be inserted later

**Abstract** The advance of technology facilitates the collection of statistical data. Flexible and refined statistical models are widely sought in a large array of statistical problems. The question arises frequently whether or not a family of parametric or nonparametric models fit adequately the given data. In this paper we give a selective overview on nonparametric inferences using generalized likelihood ratio (*GLR*) statistics. We introduce generalized likelihood ratio statistics to test various null hypotheses against nonparametric alternatives. The trade-off between the flexibility of alternative models and the power of the statistical tests is emphasized. Well-established Wilks' phenomena are discussed for a variety of semi- and non- parametric models, which sheds light on other research using *GLR* tests. A number of open topics worthy of further study are given in a discussion section.

**Keywords** Asymptotic null distribution · bootstrap · generalized likelihood ratio · nonparametric test · power function · Wilks' phenomenon

**Mathematics Subject Classification (2000)** 62G07 · 62G10 · 62J12

## 1 Introduction

Thanks to the efforts by many statisticians, led by Ronald A. Fisher, Jerzy Neyman, Egon S. Pearson and Samuel S. Wilks, there are several general applicable principles for parametric estimation and inferences. For example, in parametric estimation, one would use the maximum likelihood method when the full likelihood

[⋆] This invited paper is discussed in the comments available at:

[⋆] The work was supported by the NSF grants DMS-0354223, DMS-0532370 and DMS-0704337. The paper was initiated when Jiancheng Jiang was a research fellow at Princeton University.

J. Fan
Department of ORFE, Princeton University, Princeton, NJ 08544, USA
E-mail: jqfan@princeton.edu

J. Jiang
Department of Mathematics and Statistics, University of North Carolina at Charlotte, USA

function is available and use the least-squares, generalized method of moments (Hansen, 1982) or generalized estimation equations (Liang and Zeger, 1986) when only generalized moment functions are specified. In parametric inferences such as hypothesis testing and construction of confidence regions, the likelihood ratio tests, jackknife, and bootstrap methods (Hall, 1993; Efron and Tibshirani, 1995; Shao and Tu, 1996) are widely used. The likelihood principle appeared in literature as a term in 1962 (Barnard et al., 1962; Birnbaum, 1962), but the idea goes back to the works of R.A. Fisher in the 1920s (Fisher, 1922). Since Edwards (1972) championed the likelihood principle as a general principle of inference for parametric models (see also Edwards, 1974), it has been applied to many fields in statistics (see Berger and Wolpert, 1988) and even in the philosophy of science (see Royall, 1997).

For nonparametric models, there are also generally applicable methods for nonparametric estimation and modeling. These include local polynomial (Wand and Jones, 1995; Fan and Gijbels, 1996), spline (Wahba, 1990; Eubank, 1999; Gu, 2002) and orthogonal series methods (Efromovich, 1999; Vidakovic, 1999), and dimensionality reduction techniques that deal with the issues of the curse of dimensionality (e.g Fan and Yao, 2003, Chapter 8). On the other hand, while there are many customized methods for constructing confidence intervals and conducting hypothesis testing (Hart, 1997), there are few generally applicable principles for nonparametric inferences. An effort in this direction is Fan, Zhang and Zhang (2001), which extends the likelihood principle by using generalized likelihood ratio (*GLR*) tests. However, compared with parametric likelihood inference, the *GLR* method is not well developed, and corresponding theory and applications are available only for some models in regression contexts. Therefore, there is a great potential for developing the *GLR* tests, and a review of the idea of *GLR* inference is meaningful for encouraging further research on this topic.

Before setting foot in nonparametric inference, we review the basic idea of parametric inference using the likelihood principle.

## 1.1 Parametric inference

Suppose that the data generating process is governed by the underlying density $f(\mathbf{x};\theta)$, with unknown $\theta$ in a parametric space $\Theta$. The statistical interest lies in testing:

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta - \Theta_0$$

based on a random sample $\{\mathbf{x}_i\}_{i=1}^n$, where $\Theta_0$ is a subspace of $\Theta$. The null hypothesis is typically well formulated. For example, in the population genetics, sampling from an equilibrium population with respect to a gene with two alleles ("a" and "A"), three genotypes "AA", "Aa", and "aa" can be observed. According to the Hardy-Weinberg formula, their proportions are respectively

$$\theta_1 = \xi^2, \quad \theta_2 = 2\xi(1-\xi), \quad \theta_3 = (1-\xi)^2.$$

To test the Hardy-Weinberg formula based on a random sample, the null hypothesis is

$$\Theta_0 = \{(\theta_1, \theta_2, \theta_3) : \theta_1 = \xi^2, \quad \theta_2 = 2\xi(1-\xi), \quad \theta_3 = (1-\xi)^2, \quad 0 \le \xi \le 1\}. \quad (1.1)$$

For this problem, we have

$$\Theta = \{(\theta_1, \theta_2, \theta_3) : \theta_1 + \theta_2 + \theta_3 = 1, \quad 0 \le \theta_i \le 1\}.$$

For such a well formulated question, one can use the well-known maximum likelihood ratio statistic

$$\lambda_n = 2\{\max_{\theta \in \Theta} \ell(\theta) - \max_{\theta \in \Theta_0} \ell(\theta)\},$$

where $\ell(\theta)$ is the log-likelihood function for getting the given sample under the model $f(\mathbf{x}; \theta)$. This is indeed a very intuitive procedure: If the alternative models are far more likely to generate the given data, the null models should be rejected. The following folklore theorem facilitates the choice of the critical region: Under $H_0$ and regular conditions, $\lambda_n$ is asymptotically chi-square distributed with $k$ degrees of freedom, where $k$ is the difference of dimensions between $\Theta$ and $\Theta_0$.

An important fundamental property of the likelihood ratio tests is that their asymptotic null distributions are independent of nuisance parameters in the null hypothesis such as $\xi$ in (1.1). With this property, one can simulate the null distribution by fixing the nuisance parameters at a reasonable value or estimate. This property is referred to as *the Wilk phenomenon* in Fan, Zhang and Zhang (2001) and is fundamental to all hypothesis testing problems. It has been a folk theorem in the theory and practice of statistics and has contributed tremendously to the success of the likelihood inference.

In other examples, even though the null hypothesis is well formulated, the alternative hypotheses are not. Take the famous model for short-term interest rates as an example. Under certain assumptions, Cox, Ingersoll and Ross (1985) showed that the dynamic of short-term rates should follow the stochastic differential equation:

$$dX_t = \kappa(\mu - X_t)\, dt + \sigma X_t^{1/2}\, dW_t, \tag{1.2}$$

where $W_t$ is a Wiener process on $[0, \infty)$, and $\kappa$, $\mu$ and $\sigma$ are unknown parameters. This Feller process is called the CIR model in finance. The question arises naturally whether or not the model is consistent with empirical data. In this case, the null model is well formulated. However, the alternative model is not. To employ the parametric likelihood ratio technique, one needs to embed the CIR model into a larger family of parametric models such as the following constant elasticity of variance model (Chan et al., 1992):

$$dX_t = \kappa(\mu - X_t)\, dt + \sigma X_t^{\rho}\, dW_t, \tag{1.3}$$

and test $H_0 : \rho = 1/2$. The drawback of this approach is that the models (1.3) have to include the true data generating process. This is the problem associated with all parametric testing problems where it is implicitly assumed that the family of models $\{f(\mathbf{x}; \theta) : \theta \in \Theta\}$ contains the true one.

To see this more clearly, consider the question if variable $X$ (e.g. age) and $Y$ (salary) are related. If we embed the problem in the linear model

$$Y = \alpha + \beta X + \varepsilon, \tag{1.4}$$

then the problem becomes testing $H_0 : \beta = 0$. If unknown to us, the data generating process is govern by

$$Y = 60 - 0.1(X - 45)^2 + \varepsilon, \quad \varepsilon \sim N(0, 5^2), \tag{1.5}$$

with $X$ uniformly distributed on the interval $[25, 65]$, the null hypothesis will be accepted very often since slope $\beta$ in (1.4) is not statistically significant. Figure 1 presents a random sample of size 100 from model (1.5). The erroneous conclusion of accepting the null hypothesis is due to the fact that the family of models (1.4) does not contain the true one.

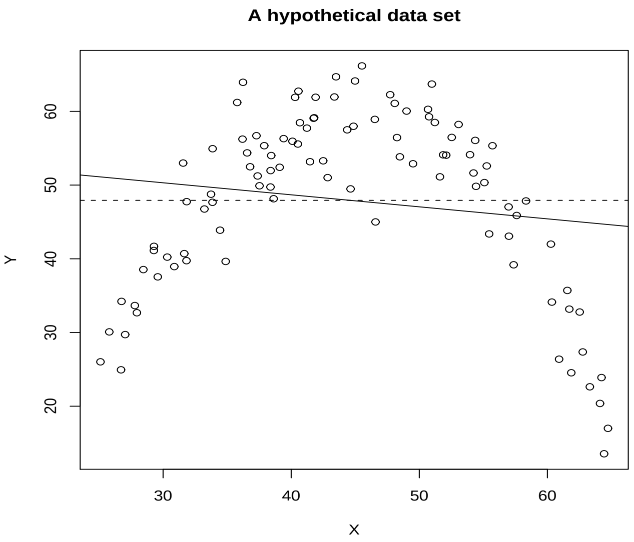

**Fig. 1** A simulated data set from (1.5) with sample size 100. When the alternative hypothesis does not contain the true model, the fit from the null model (dashed line) is statistically indistinguishable from the fit (solid line) from the alternative model (1.4). Erroneous conclusion is drawn from this analysis.

## 1.2 Nonparametric alternatives

The above discussion reveals that the family of alternative models should be large enough in order to make sensible inferences. In many hypothesis testing problems, while the null hypothesis is well formulated, the alternative one is vague. These two considerations make nonparametric models as attractive alternative hypothesis. In the hypothetical example presented in Figure 1, without knowing the data generating process, a natural alternative model is

$$Y = m(X) + \varepsilon, \tag{1.6}$$

where $m(\cdot)$ is smooth, while the null hypothesis is $Y = \mu + \varepsilon$. This is a parametric null hypothesis against a nonparametric alternative hypothesis. With such a flexible alternative family of models, the aforementioned pitfall is avoided.

For the interest rate modeling, a natural alternative to the CIR model (1.2) is the following one-factor model

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dW_t,$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are unspecified, but smooth functions. This is again the problem of testing a parametric family of models against a nonparametric family of alternative models. See Fan and Zhang (2003) for a study based on the *GLR* test.

The problems of nonparametric null against nonparametric alternative hypothesis also arise frequently. One may ask if the returns of stock contain jumps or if a discretely observed process is Markovian. In these cases, both the null and alternative hypotheses are nonparametric.

The problems of testing nonparametric null against nonparametric alternative hypotheses arise also frequently in statistical inferences. Consider, for example, the additive model (Hastie and Tibshirani, 1990):

$$Y = \alpha + \sum_{d=1}^{D} m_d(X_d) + \varepsilon \tag{1.7}$$

where $\alpha$ is an unknown constant, and $m_d$ are unknown functions, satisfying $E[m_d(X_{di})] = 0$ for identifiability. This nonparametric model includes common multivariate linear regression models. The question such as if the covariates $X_1$ and $X_2$ are related to the response $Y$ arises naturally, which amounts to testing

$$H_0 : m_1(\cdot) = m_2(\cdot) = 0. \tag{1.8}$$

This is a nonparametric null versus nonparametric alternative hypothesis testing problem, since under the null hypothesis (1.8), the model is still a nonparametric additive model:

$$Y = \alpha + \sum_{d=3}^{D} m_d(X_d) + \varepsilon \tag{1.9}$$

There are many techniques designed to solve this kind of problems. Many of them focused on an intuitive approach using discrepancy measures (such as the $L_2$ and $L_\infty$ distances) between the estimators under null and alternative models. See early seminal work by Bickel and Rosenblatt (1973), Azzalini, Bowman and Härdle (1989), and Härdle and Mammen (1993). They are generalizations of the Kolmogorov-Smirnov and Cramér-von Mises types of statistics. However, the approach suffers some drawbacks. First, choices of measures and weights can be arbitrary. Consider, for example, the null hypothesis (1.8) again. The test statistic based on discrepancy method is $T = \sum_{d=1}^{2} c_d \|\hat{m}_d\|$. One has to choose not only the norm $\|\cdot\|$ but the weights $c_d$. Second, the null distribution of the test statistic $T$ is unknown and depends critically on the nuisance functions $m_3, \ldots, m_D$. This hampers the applicability of the discrepancy based methods. Naturally, one would like to develop some test methods along the line of parametric likelihood ratio tests that possess Wilks' phenomenon to facilitate the computation of p-values.

We would like to note that it is possible to design some test statistics that tailored for some specific problems with good power. The question also arises naturally if we can come with a generally applicable principle for testing against nonparametric alternative models. The development of *GLR* statistics aims at a unified principle for nonparametric hypothesis testing problems.

## 1.3 Flexibility of models and power of tests

There are many ways to embed the family of null models into the alternative ones. Considering again testing if the CIR model (1.2) holds, in addition to the alternative parametric models (1.3) and nonparametric one-factor model (1.6), one can also consider an even larger family of nonparametric models such as the stationary Markovian model as the alternative models (Hong and Li, 2005; Aït-Sahalia, Fan and Peng, 2005). Figure 2 schematically illustrates the relationship among these three families of alternative models. In general, the larger the family of the alternative models, the more likely it includes the true model. On the other hand, the lower the power of omnibus tests. Therefore, it should not be surprised that the test in Fan and Zhang (2003) based on alternative models (1.6) is more powerful than the test constructed based only on the stationary Markovian assumption (Hong and Li, 2005; Aït-Sahalia, Fan and Peng, 2005), when the data are indeed generated from the one-factor model.

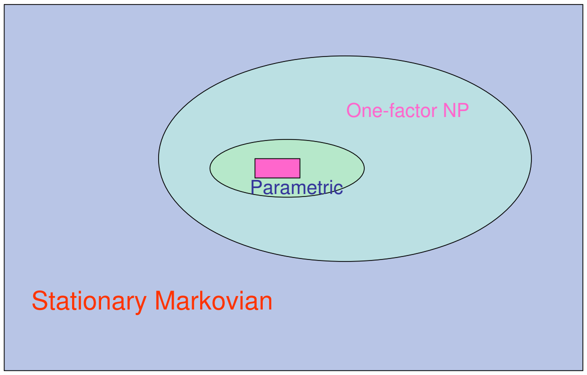

**Fig. 2** Schematic illustration of different families of alternative models for testing the theory of short-term interest model (1.2). The larger the family, the more omnibus the test and the less power in a particular direction.

A similar example is to test whether the covariates $X_1$ and $X_2$ are related to the response $Y$ as in (1.8). In this case, one can choose the alternative models such as the multiple regression model $Y = \beta_0 + \sum_{i=1}^{D} \beta_i X_i + \varepsilon$, the additive model (1.7), and the saturated nonparametric regression model $Y = f(X_1, \cdots, X_D) + \varepsilon$. The tests designated for the saturated nonparametric model necessarily have low power, while the tests with the multiple regression model as alternative can have no power when the true model is indeed nonlinear as illustrated in Figure 1.

### 1.4 Organization of the paper

In Section 2, we discuss some drawbacks of the naive extension of classical maximum likelihood ratio tests to nonparametric setting. Section 3 outlines the framework for the *GLR* tests. Several established Wilk's type of results for various nonparametric models are summarized in Section 4. In Section 5, we conclude this paper and set forth some open problems.

## 2 Naive extension of maximum likelihood ratio tests

Although likelihood ratio theory contributes tremendous success to parametric inference, there are few general applicable approaches for nonparametric inferences based on function estimation. A naive extension is the nonparametric maximum likelihood ratio test. However, nonparametric maximum likelihood estimation usually does not exist. Even if it exists, it is hard to compute. Furthermore, the resulting maximum likelihood ratio tests are not optimal.

The following example from Fan, Zhang and Zhang (2001) illustrates the above points and provides additional insights.

### 2.1 Problems with nonparametric maximum likelihood ratio tests

Suppose that there are $n$ data points $\{(X_i, Y_i)\}$ sampled from the following model:

$$Y_i = m(X_i) + \varepsilon_i, \;\; i = 1, \dots, n, \tag{2.10}$$

where $\{\varepsilon_i\}$ is a sequence of i.i.d. random variables from $\mathcal{N}(0, \sigma^2)$ and $X_i$ has a density with compact support, say $[0, 1]$. Assume that the parametric space is

$$\mathcal{F}_k = \Big\{ m \in L^2[0,1] : \int_0^1 [m^{(k)}(x)]^2 \, dx \le C \Big\},$$

for a given constant $C$. Consider the testing problem:

$$H_0 : m(x) = \alpha_0 + \alpha_1 x \;\text{ versus }\; H_1 : m(x) \neq \alpha_0 + \alpha_1 x. \tag{2.11}$$

Then the conditional log-likelihood function is

$$\ell(m, \sigma) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - m(X_i))^2. \tag{2.12}$$

Denote by $(\hat\alpha_0, \hat\alpha_1)$ the maximum likelihood estimator (MLE) under $H_0$, and $\hat m_{\text{MLE}}(\cdot)$ the MLE under $\mathcal{F}_k$ which solves the following minimization problem:

$$\min \sum_{i=1}^{n} (Y_i - m(X_i))^2, \;\text{ subject to }\; \int_0^1 m^{(k)}(x)^2 \, dx \le C.$$

Then $\hat m_{\text{MLE}}$ is a smoothing spline (Wahba, 1990; Eubank, 1999) with the smoothing parameter chosen to satisfy $||\hat m^{(k)}_{\text{MLE}}||_2^2 = C$. Define the residual sum of squares under the null and alternative as follows:

$$\text{RSS}_0 = \sum_{i=1}^{n} (Y_i - \hat\alpha_0 - \hat\alpha_1 X_i)^2, \;\; \text{RSS}_1 = \sum_{i=1}^{n} (Y_i - \hat m_{\text{MLE}}(X_i))^2.$$

Then the logarithm of the conditional maximum likelihood ratio statistic for the testing problem (2.11) is given by

$$\lambda_n = \ell(\hat{m}_{\mathrm{MLE}}, \hat{\sigma}) - \ell(\hat{m}_0, \hat{\sigma}_0) = \frac{n}{2} \log \frac{\mathrm{RSS}_0}{\mathrm{RSS}_1},$$

where $\hat{\sigma}^2 = n^{-1}\mathrm{RSS}_1$, $\hat{m}_0(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x$, and $\hat{\sigma}_0^2 = n^{-1}\mathrm{RSS}_0$.

Even in this simple situation where the nonparametric maximum likelihood exists, we need to know the constant $C$ in the parametric space $\mathcal{F}_k$ in order to compute the MLE. This is unrealistic in practice. Even if the value $C$ is granted, Fan, Zhang and Zhang (2001) demonstrated (see also §2.2) that the nonparametric maximum likelihood ratio test is not optimal. This is due to the limited choice of smoothing parameter in $\hat{m}_{\mathrm{MLE}}$, which makes it satisfy $||\hat{m}_{\mathrm{MLE}}^{(k)}||_2^2 = C$. In this case, the essence of the testing problem is to estimate the functional $\|m\|^2$ (we assume that $\alpha_0 = \alpha_1 = 0$ in (2.11) without loss of generality as they can be estimated faster than nonparametric rates). It is well-known that one smoothing parameter can not optimize simultaneously the estimate of the functional $\|m\|^2$ and the function $m(\cdot)$. See for example, Bickel and Ritov (1988); Hall and Marron (1988); Donoho and Nussbaum (1990); Fan (1991).

Now, if the parameter space is $\mathcal{F}_k' = \{m \in L^2[0,1] : \sup_{0\le x\le 1} |m^{(k)}(x)| \le C\}$ or more complicated space, it is not clear whether the maximum likelihood estimator exists and even if it exists, how to compute it efficiently.

The above example reveals that the nonparametric MLE may not exist and hence cannot serve as a generally applicable method. It illustrates further that, even when it exists, the nonparametric MLE chooses smoothing parameters automatically. This is too restrictive for the procedure to possess the optimality of testing problems. Further, we need to know the nonparametric space exactly. For example, the constant $C$ in $\mathcal{F}_k$ needs to be specified. The *GLR* statistics in Fan, Zhang and Zhang (2001), which replaces the nonparametric maximum likelihood estimator by any reasonable nonparametric estimator, attenuate these difficulties and enhances the flexibility of the test statistic by varying the smoothing parameter. By proper choices of the smoothing parameter, the *GLR* tests achieve the optimal rates of convergence in the sense of Ingster (1993) and Lepski and Spokoiny (1999).

## 2.2 Inefficiency of nonparametric maximum likelihood ratio tests

To demonstrate the inefficiency of nonparametric maximum likelihood ratio tests, let us consider a simpler mathematical model, the Gaussian white noise model, to simplify the technicality of mathematical proofs. The model keeps all important features of nonparametric regression model, as demonstrated by Brown and Low (1996) and Grama and Nussbaum (2002). Suppose that we have observed the whole process $Y(t)$ from the following Gaussian white noise model:

$$dY(t) = \phi(t)\, dt + n^{-1/2}\, dW(t),\ \ t \in (0,1), \qquad (2.13)$$

where $\phi$ is an unknown function and $W(t)$ is the Brownian process. By using an orthogonal series (e.g. the Fourier series) transformation, model (2.13) is equivalent to the following white noise model:

$$Y_i = \theta_i + n^{-1/2}\varepsilon_i, \;\; \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \;\; i = 1, 2, \ldots$$

where $Y_i$, $\theta_i$ and $\varepsilon_i$ are the $i$-th Fourier coefficients of $Y(t)$, $\phi(t)$ and $W(t)$, respectively. Now let us consider testing the simple hypothesis:

$$H_0 : \theta_1 = \theta_2 = \cdots = 0, \tag{2.14}$$

which is equivalent to testing $H_0 : \phi = 0$ under model (2.13).

Consider the parametric space $\mathcal{F}_k^* = \{\theta : \sum_{j=1}^{\infty} j^{2k}\theta_j^2 \leq 1\}$ where $k \geq 0$. When $k$ is a positive integer, this set in the frequency domain is equivalent to the Sobolev class of periodic functions $\{\phi : ||\phi^{(k)}|| \leq c\}$ for a constant $c$. Then under the parameter space $\mathcal{F}_k^*$, the maximum likelihood estimator is

$$\hat{\theta}_j = (1 + \hat{\xi} j^{2k})^{-1} Y_j,$$

where $\hat{\xi}$ is the Lagrange multiplier satisfying

$$\sum_{j=1}^{\infty} j^{2k}\hat{\theta}_j^2 = 1. \tag{2.15}$$

Under the null hypothesis (2.14), Lemma 2.1 of Fan, Zhang and Zhang (2001) shows that

$$\hat{\xi} = n^{-2k/(2k+1)} \Big\{ \int_0^{\infty} \frac{y^{2k}}{(1+y^{2k})^2}\, dy \Big\}^{2k/(2k+1)} \{1 + o_p(1)\}.$$

The maximum likelihood ratio statistic for the problem (2.14) is

$$\lambda_n^* = \frac{n}{2} \sum_{j=1}^{\infty} \Big(1 - \frac{j^{4k}\hat{\xi}^2}{(1+j^{2k}\hat{\xi})^2}\Big) Y_j^2.$$

Fan, Zhang and Zhang (2001) proved that the maximum likelihood ratio test, $\lambda_n^*$, can test consistently alternatives with a rate no faster than $n^{-(k+d)/(2k+1)}$ for any $d > 1/8$. Theorefore, when $k > 1/4$, by taking $d$ sufficiently close to $1/8$, the test $\lambda_n^*$ cannot be optimal according to the formulations of Ingster (1993) for hypothesis testing where an optimal test can detect alternatives converging to the null with rate $n^{-2k/(4k+1)}$. This is due to the restrictive choice of the smoothing parameter $\hat{\xi}$ of the MLE, which has to satisfy (2.15). *GLR* tests remove this restrictive requirement and allow one to tune optimally the smoothing parameter. By taking $\xi_n = cn^{-4k/(4k+1)}$ for some $c > 0$, the *GLR* test statistic defined by

$$\lambda_n = \frac{n}{2} \sum_{j=1}^{\infty} \Big(1 - \frac{j^{4k}\xi_n^2}{(1+j^{2k}\xi_n)^2}\Big) Y_j^2,$$

achieves the optimal rate of convergence for hypothesis testing (Theorem 3 of Fan, Zhang and Zhang, 2001).

The *GLR* test allows one to use any reasonable nonparametric estimator to construct the test. For the Sobolev class $\mathcal{F}_k^*$, another popular class of nonparametric estimator is the truncation estimator (see Efromovich, 1999):

$$\hat{\theta}_j = Y_j, \;\text{ for } j = 1, \cdots, m, \quad \hat{\theta}_j = 0, \;\text{ for } j > m,$$

for a given $m$. Then, the twice log-likelihood ratio [between this nonparametric estimator and the null estimator (2.14)] test statistic is the Neyman (1937) test

$$T_N = \sum_{i=1}^{m} nY_i^2.$$

In this sense, the Neyman test can be regarded as a *GLR* test. Note that the Neyman test is indeed the maximum likelihood ratio test for the problem (2.14) against the alternative hypothesis with constraints

$$\mathcal{F} = \{\theta : \theta_{m+1} = \theta_{m+2} = \cdots = 0\}.$$

With the choice of $m = cn^{2/(4k+1)}$, the Neyman test, which is a *GLR* test, can also achieve the optimal rate $n^{-2k/(4k+1)}$. This shows the versatility of the *GLR* tests.

### 2.3 Adaptive choice of smoothing parameter

There are many studies on the practical choice of the smoothing parameter $m$ for the Neyman test. See, for example, Eubank and Hart (1992), Eubank and LaRiccia (1992), Inglot and Ledwina (1996), and Kallenberg and Ledwina (1997). Fan (1996) introduced the following adaptive version of the Neyman test, called the adaptive Neyman test,

$$T_{AN}^* = \max_{1 \le m \le n} \sum_{i=1}^{m} (nY_i^2 - 1)/\sqrt{2m}, \tag{2.16}$$

and normalized it as

$$T_{AN} = \sqrt{2 \log\log n} T_{AN}^* - \{2 \log\log n + 0.5 \log\log\log n - 0.5 \log(4\pi)\}.$$

The adaptive choice of $m$ is indeed $\hat{m}$ that maximizes (2.16).

It was shown by Fan (1996) that under the null hypothesis (2.14),

$$P(T_{AN} < x) \to \exp(-\exp(-x)), \text{ as } n \to \infty,$$

and by Fan and Huang (2001) and Fan, Zhang and Zhang (2001) that the adaptive Neyman test can detect adaptively, in the sense of Spokoiny (1996), the alternatives with the optimal rate

$$(n^{-2} \log\log n)^{k/(4k+1)}$$

when the parameter space is $\mathcal{F}_k$ with unknown $k$.

## 3 Generalized likelihood ratio tests

Section 2 demonstrates convincingly that for testing against nonparametric alternatives, it is necessary to have flexible and good nonparametric estimators for computing the likelihood of generating the underlying data. We now describe a general framework for the generalized likelihood ratio tests.

### 3.1 GLR statistic

The basic idea of the *GLR* test can be transparently illustrated in terms of likelihood as follows. Let $\mathbf{f}$ be the vector of functions and $\boldsymbol{\eta}$ be the parameters in nonparametric or semiparametric models. Suppose that the logarithm of the likelihood of a given data set is $\ell(\mathbf{f}, \boldsymbol{\eta})$. Given $\boldsymbol{\eta}$, one has a good nonparametric estimator $\hat{\mathbf{f}}_{\boldsymbol{\eta}}$ of $\mathbf{f}$. The nuisance parameters $\boldsymbol{\eta}$ can be estimated by the profile likelihood, that is, to find $\boldsymbol{\eta}$ maximizing $\ell(\hat{\mathbf{f}}_{\boldsymbol{\eta}}, \boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$. This gives the maximum profile likelihood $\ell(\hat{\mathbf{f}}_{\hat{\boldsymbol{\eta}}}, \hat{\boldsymbol{\eta}})$, which is not the maximum likelihood since $\hat{\mathbf{f}}_{\boldsymbol{\eta}}$ is not an MLE.

Suppose we are interested in testing whether a parametric family $\mathbf{f}_\theta$ fits a given set of data with sample size $n$. Then the null hypothesis is

$$H_0 : \mathbf{f} = \mathbf{f}_\theta,\ \theta \in \Theta. \tag{3.17}$$

As argued before, we use the nonparametric model $\mathbf{f}$ as the alternative. Let $(\hat{\theta}_0, \hat{\boldsymbol{\eta}}_0)$ be the maximum likelihood estimator under the above null model, maximizing the likelihood function $\ell(\mathbf{f}_\theta, \boldsymbol{\eta})$. Then $\ell(\mathbf{f}_{\hat{\theta}_0}, \hat{\boldsymbol{\eta}}_0)$ is the maximum likelihood under the null. The *GLR* test statistic is simply defined as

$$\lambda_n = \ell(\hat{\mathbf{f}}_{\hat{\boldsymbol{\eta}}}, \hat{\boldsymbol{\eta}}) - \ell(\mathbf{f}_{\hat{\theta}_0}, \hat{\boldsymbol{\eta}}_0), \tag{3.18}$$

which calibrates the difference of log-likelihoods of producing the given data under the null and alternative models. Large values of $\lambda_n$ suggest rejection of the null hypothesis since the alternative family of models are far more likely to generate the data.

In general, the *GLR* test does not have to use the true likelihood. Like parametric inference, nonparametric inference generally does not assume underlying distributions are known. For example, in a parametric regression setting one can estimate the unknown parameters by maximizing a negative loss function or quasi-likelihood function $Q(\mathbf{f}_\theta, \boldsymbol{\eta})$. Then the *GLR* test statistic can be defined as

$$\lambda_n = Q(\hat{\mathbf{f}}_{\hat{\boldsymbol{\eta}}}, \hat{\boldsymbol{\eta}}) - Q(\mathbf{f}_{\hat{\theta}_0}, \hat{\boldsymbol{\eta}}_0).$$

In addition, the approach is also applicable to the cases with part unknown functions as nuisance parameters such as (1.8) or (1.9). The essence is to replace $\mathbf{f}_{\hat{\theta}_0}$ by a nonparametric estimate. We omit details.

Note that the *GLR* test does not require the concise knowledge of the nonparametric space. This relaxation extends the scope of applications and removes the impractical assumptions such as constant $C$ in (2.11) being known. Further, the smoothing parameter can be selected to optimize the performance of the *GLR* test.

### 3.2 What is Wilks' phenomenon?

A nice feature of the *GLR* tests is that for a host of statistical problems, they share Wilks' phenomenon as the traditional maximum likelihood ratio tests in testing problems with parametric nulls and alternatives. These are demonstrated in a number of papers such as the univariate nonparametric regression models and the varying-coefficient models in Fan, Zhang and Zhang (2001), the linear Gaussian process in Fan and Zhang (2004) for spectral density estimation, the varying-coefficient partly linear models in Fan and Huang (2005), the additive models in Fan and Jiang (2005), the diffusion models in Aït-Sahalia, Fan and Peng (2005), and the partly linear additive models in Jiang et al (2007). Corresponding results in depth will be given in Section 4. These results indicate that the Wilks phenomena exist in general.

By Wilks' phenomenon, we mean that the asymptotic null distributions of test statistics are independent of nuisance parameters and functions. Typically, the asymptotic null distribution of the *GLR* statistic $\lambda_n$ is nearly $\chi^2$ with large degrees of freedom in the sense that

$$r\lambda_n \overset{d}{\simeq} \chi^2_{\mu_n} \tag{3.19}$$

for a sequence $\mu_n \to \infty$ and a constant $r$, namely,

$$(2\mu_n)^{-1/2}(r\lambda_n - \mu_n) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

where $\mu_n$ and $r$ are independent of nuisance parameters/functions. They may depend on the methods of nonparametric estimation and smoothing parameters. Therefore, the asymptotic null distribution is independent of the nuisance parameters/functions. With this Wilks phenomenon, the advantages of the classical likelihood ratio tests are fully inherited: one makes a statistical decision by comparing likelihoods of generating the given data under two competing classes of models and the critical value can easily be found based on the known null distribution $\mathcal{N}(\mu_n, 2\mu_n)$ or $\chi^2_{\mu_n}$. Another important consequence of the results is that one does not have to derive theoretically the constant $\mu_n$ and $r$ in order to use the *GLR* tests, since as long as there is such a Wilks type of phenomenon, one can simply simulate the null distributions by setting nuisance parameters under the null hypothesis at reasonable values or estimates.

The above Wilks phenomenon is not a coincidence for the nonparametric model. In the exponential family of models with growing number of parameters, Portnoy (1988) showed the Wilks type of result in the same sense as (3.19), and Murphy (1993) revealed a similar type of result for Cox's proportional hazards model using a simple sieve method (piecewise constant approximation to a smooth function). While there is no general theory on the *GLR* tests, various authors have demonstrated that Wilks' phenomenon holds for many nonparametric models (see Section 4).

In order to better understand Wilks' phenomenon of *GLR* tests, we illustrate it numerically using a variant of the bivariate additive model from Fan and Jiang (2005).

A random sample $\{X_{1i}, X_{2i}, Y_i\}_{i=1}^n$ is generated from the bivariate model

$$Y = m_1(X_1) + m_2(X_2) + \varepsilon, \tag{3.20}$$

where $m_1(X_1) = 1 - 12X_1^2 + 5X_1^3$, $m_2(X_2) = \sin(\pi X_2)$, and the error $\varepsilon$ is distributed as $\mathcal{N}(0,1)$. The covariates are generated by the following transformation to create correlation:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \end{pmatrix},$$

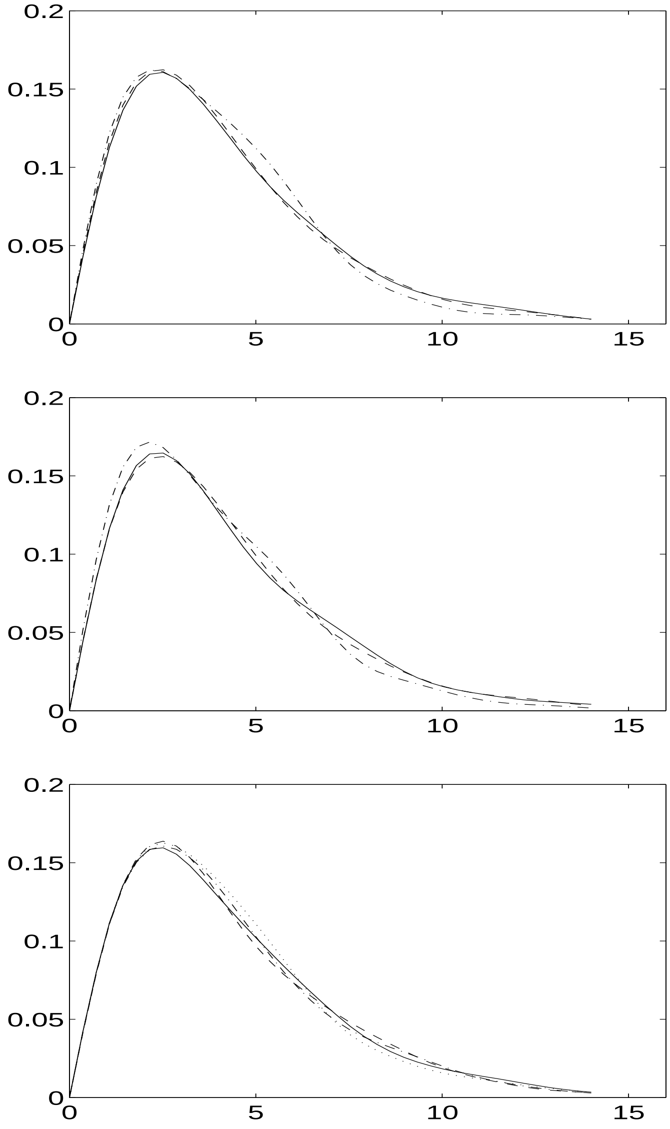

**Fig. 3** Results for Example 1. Estimated densities for the *GLR* statistics among 1000 simulations. Top Panel: with fixed $h_2 = h_{2,opt}$, but different bandwidths for $h_1$ ( solid – $h_1 = \frac{2}{3}h_{1,opt}$; dashed – $h_1 = h_{1,opt}$; dash-dotted – $h_1 = \frac{3}{2}h_{1,opt}$); Middle Panel: with different nuisance functions and optimal bandwidths $h_d = h_{d,opt}$ ( solid – $\beta = -1.5$; dashed – $\beta = 0$; dotted – $\beta = 1.5$); Bottom Panel: estimated densities for the *GLR* statistics under different errors (solid – normal; dashed – t(5) ; dotted – $\chi^2(5)$; dash-dotted – $\chi^2(10)$);

where $U_i \overset{iid}{\sim} U(-0.5, 0.5)$.

Consider the testing problem:

$$H_0 : \ m_2(x_2) = 0 \ \text{ versus } \ H_1 : m_2(x_2) \neq 0.$$

Since $m_1(\cdot)$ is an unknown nuisance function, this testing problem is actually a nonparametric null hypothesis against a nonparametric alternative hypothesis. Frequently, the backfitting algorithm with a smoothing method (Hastie and Tibshirani, 1990) is employed to estimate the additive components in (3.20). We here employ the backfitting algorithm with the local linear smoother using bandwidth $h_1$ for estimating $m_1$ and $h_2$ for estimating $m_2$ to construct a *GLR* statistic. See Section 4.5 for additional details.

To demonstrate Wilks' phenomenon for the *GLR* test, we use (i) three levels of bandwidth $h_1 = \frac{2}{3}h_{1,opt}$, $h_{1,opt}$, or $\frac{3}{2}h_{1,opt}$ with $h_2$ fixed at its optimal value $h_{2,opt}$, where $h_{d,opt}$ is the optimal bandwidth for the smoother on $m_d(\cdot)$ (see Opsomer, 2000); and (ii) three levels of nuisance function $m_1(X_1)$:

$$m_{1,\beta}(X_1) = \left[1 + \beta\sqrt{\text{var}(0.5 - 6X_1^2 + 3X_1^3)}\right] \ (0.5 - 6X_1^2 + 3X_1^3),$$

where $\beta = -1.5, \ 0, \ 1.5$. For the *GLR* test, we drew 1000 samples of 200 observations. Based on the 1000 samples, we obtained 1000 *GLR* test statistics. Their distribution is obtained via a kernel estimate with a rule of thumb bandwidth: $h = 1.06sn^{-0.2}$, where $s$ is the standard deviation of the normalized *GLR* statistics.

Figure 3 shows the estimated densities of the normalized *GLR* statistics, $r\lambda_n$, where $r$ is the normalization constant given in Section 4.5. As expected, they look like densities from $\chi^2$-distributions. The top panel of Figure 3 shows that the null distributions follow $\chi^2$-distributions over a wide range of bandwidth $h_1$. Note that different bandwidths $h_1$ gives different complexity of modeling the nuisance function $m_1$ and the results show that the null distributions are nearly independent of the different model complexity of the nuisance function $m_1$. This demonstrates numerically Wilks' phenomenon. Note that the degree of freedom depends on bandwidth $h_2$ but not on $h_1$, which reflects the difference of the complexity of $m_2$ under the null model and the alternative model.

The middle panel demonstrates also the Wilks type of phenomenon from a different angle: for the three very different choices of nuisance functions, the null distributions are nearly the same.

To investigate the influence of different noise distributions on the *GLR* tests, we now consider model (3.20) with different error distributions of $\varepsilon$. In addition to the standard normal distribution, the standardized $t(5)$ and the standardized $\chi^2(5)$ and $\chi^2(10)$ are also used to assess the stability of the null distribution of the *GLR* test for different error distributions. The bottom panel of Figure 3 reports the estimated densities of the normalized *GLR* statistics under the above four different error distributions. It shows that the null distributions of the tests are approximately the same for different error distributions, which again endorses Wilks' phenomenon.

### 3.3 Choice of smoothing parameter

The *GLR* statistic involves at least a parameter $h$ in smoothing the function $\mathbf{f}$. For each given smoothing parameter $h$, the GLR statistic $\lambda_n(h)$ is a test statistic.

This forms a family of test statistics indexed by $h$. In general, a larger choice of bandwidth is more powerful for testing smoother alternatives, and a smaller choice of bandwidth is more powerful for testing less smooth alternatives. Questions arise about how to choose a bandwidth for nonparametric testing and what criteria should be used.

Assume that the Wilks type of result (3.19) holds with degree of freedom $\mu_n(h)$. Inspired by the adaptive Neyman test (2.16) of Fan (1996), which achieves the adaptive optimal rate, an adaptive choice of bandwidth is to maximize the normalized test statistic

$$\hat{h} = \arg \max_{h \in [n^{-a}, n^{-b}]} \{r\lambda_n(h) - \mu_n(h)\}/\sqrt{2\mu_n(h)},$$

for some $a, b > 0$. This results in the following multiscale GLR test statistic in Fan, Zhang and Zhang (2001),

$$\max_{h \in [n^{-a}, n^{-b}]} \{r\lambda_n(h) - \mu_n(h)\}/\sqrt{2\mu_n(h)}. \qquad (3.21)$$

Like the adaptive optimality of the adaptive Neyman test, it is expected that the multi-scale *GLR* test possesses a similar optimality property. Indeed, Horowitz and Spokoiny (2001, 2002) demonstrated this kind of property for two specific models.

In practical implementations, one needs only to find the maximum in the multiscale *GLR* test (3.21) over a grid of bandwidths. Zhang (2003a) calculated the correlation between $\lambda_n(h)$ and $\lambda_n(ch)$ for some inflation factor $c$. The correlation is quite large when $c = 1.3$. Thus, a simple implementation is to choose a grid of points $h = h_0 1.5^j$ for $j = -1, 0, 1$, representing “small”, “right”, and “large” bandwidths. A natural choice of $h_0$ is the optimal bandwidth in the function estimation.

Another choice of bandwidth is to choose an optimal bandwidth to maximize the power of the *GLR* test over some specific family of alternative models. This problem has not been seriously explored in the literature. For practical implementations, the bandwidth used for curve fitting provides a reasonable starting point for the *GLR* tests, although it may not optimize the power. In general, there is no big difference in rates for the optimal bandwidths for estimation and testing. In fact, the optimal bandwidth for the local linear estimation of a univariate $\mathbf{f}$ is of order $O(n^{-1/5})$, and that for the *GLR* test is of order $O(n^{-2/9}) = O(n^{-1/5} \times n^{-1/45})$. Therefore, with the estimated optimal bandwidth $\hat{h}$ for estimation, one can employ the ad hoc bandwidth, $\hat{h} \times n^{-1/45}$, for the *GLR* test.

## 3.4 Bias correction

When $\mathbf{f}_\theta$ in the null hypothesis in (3.17) is not linear/polynomial, a local linear/polynomial fit will result in a biased estimate under the null hypothesis. Similarly, when the function $\mathbf{f}_\theta$ is not a spline function, the spline based smoothing results in the bias of the estimate under the null hypothesis. These affect the precision of null distribution of the *GLR* statistic and hence the reliability of statistical conclusion.

The aforementioned bias problem can be significantly attenuated as follows. Reparameterize the unknown functions as $\mathbf{f}^* = \mathbf{f} - \mathbf{f}_{\hat{\theta}_0}$. Then the test problem

(3.17) becomes testing

$$H_0 : \mathbf{f}^* = 0 \ \text{ versus } \ \mathbf{f}^* \neq 0$$

with the likelihood or more generally the quasi-likelihood function $Q^*(\mathbf{f}^*, \boldsymbol{\eta}) = Q(\mathbf{f}^* + \mathbf{f}_{\hat{\theta}_0}, \boldsymbol{\eta})$. Applying the *GLR* test to this reparameterized problem with the new quasi-likelihood function $Q^*(\mathbf{f}^*, \boldsymbol{\eta})$, one can eliminate the bias problem in the null distribution, since any reasonable nonparametric estimator will not have biases when the true function is zero. Let $(\hat{\mathbf{f}}^*_{\hat{\boldsymbol{\eta}}}, \hat{\boldsymbol{\eta}})$ be a profile estimator based on $Q^*(\mathbf{f}^*, \boldsymbol{\eta})$. Then the bias-corrected version of the *GLR* test is

$$T^* = Q^*(\hat{\mathbf{f}}^*_{\hat{\boldsymbol{\eta}}}, \hat{\boldsymbol{\eta}}) - Q^*(0, \hat{\boldsymbol{\eta}}_0) = Q(\hat{\mathbf{f}}^*_{\hat{\boldsymbol{\eta}}} + f_{\hat{\theta}_0}, \hat{\boldsymbol{\eta}}) - Q(f_{\hat{\theta}_0}, \hat{\boldsymbol{\eta}}_0)$$

The above idea is inspired by the prewhitening technique of Press and Tukey (1956) in spectral density estimation and the technique that was employed by Härdle and Mammen (1993) for univariate nonparametric testing. Indeed, for the univariate regression setting, our general method coincides with the test in Härdle and Mammen (1993). Our method is also related to the nonparametric estimator that uses a parametric start of Hjort and Glad (1995) and Glad (1998). Recently, Fan and Zhang (2004) and Fan and Jiang (2005) advocated the use of the bias reduction method in the study of testing problems for spectral density and additive models, respectively, when the null hypothesis is a parametric family.

## 3.5 Bootstrap

To implement a *GLR* test, we need to obtain the null distribution of the test statistic. Theoretically the asymptotic null distribution in (3.19) can be used in determining the p-value of a *GLR* statistic. However, this needs to derive its asymptotic null distribution. In addition, the asymptotic distribution does not necessarily give a good approximation for finite sample sizes. For example, from the asymptotic point of view, the $\chi^2_{\mu_n+100}$-distribution and the $\chi^2_{\mu_n}$-distribution are approximately the same since $\mu_n \to \infty$, but for moderate $\mu_n$, they are quite different. This means that a second order term is needed. Assume that the appropriate degree of freedom is $\mu_n + c$ for a constant $c$. Then when the bandwidth is large ($h \to \infty$), the local linear fit becomes a global linear fit, and the *GLR* test becomes the parametric maximum likelihood ratio test. Hence, $\lambda_n \to \chi^2_{2p}$ in distribution according to the classic Wilks type of result, where $2p$ denotes the difference of the degree of freedom difference under the null and alternative hypothesis. It is reasonable to expect that the degree of freedom $\mu_n + c \to 2p$ as $h \to \infty$. Since typically $\mu_n$ depends on $h$ in such a way that $\mu_n \to 0$ as $h \to \infty$, we have $c = 2p$. This is the calibration idea in Zhang (2003b). However, it also may not lead to a good approximation to the null distribution of $\lambda_n$, since in most of cases the bandwidth $h$ is not so big and the above calibration method may fail.

Thanks to Wilks' phenomenon for the *GLR* test statistic, the asymptotic null distribution is independent of nuisance parameters/functions under the null hypothesis. For a finite sample, this means that the null distribution does not sensitively depend on the nuisance parameters/functions. Therefore, the null distribution can be approximated by simulations, via fixing nuisance parameters/functions at their reasonable estimates. Since the resampling approximation is generally wild bootstrap, the resulting estimator of the null distribution is consistent. For different settings, some bootstrap approximations to the null distributions of *GLR* statistics have been studied. For details, see Section 4.

An additional advantage of the bootstrap method is that one does not need to derive the asymptotic null distribution first. As long as Wilks' phenomenon is believed to be true, it provides a consistent estimate of the null distribution.

### 3.6 Power

In various settings, it has been shown that the *GLR* tests are asymptotically optimal in the sense that they can detect alternatives with optimal rates for nonparametric hypothesis testing according to the formulation of Ingster (1993) and Lepski and Spokoiny (1999). The optimality here is not the same as the uniformly most powerful (UMP) test in the classic sense. In fact, for problems as complex as ours, no UMP test exists.

For the testing problem in (3.17), one may consider the following contiguous alternatives

$$H_{1n} : \mathbf{f} = \mathbf{f}_\theta + n^{-\gamma}\mathbf{g}_n,$$

where $\gamma > 0$ and $\mathbf{g}_n$ is an unspecified vector sequence of smooth functions in a large class of functions. The power of the *GLR* test under the above alternative has been investigated by several authors for different models, see for example Fan, Zhang and Zhang (2001) and Fan and Jiang (2005) among others. In general, it can be shown that when the local linear smoother is employed for estimating $\mathbf{f}$ and the bandwidth is of order $n^{-2/9}$, the *GLR* test can detect alternatives with the rate $\gamma = 4/9$ which is optimal according to Ingster (1993). Thus, the generalized likelihood method is not only intuitive to use, but also powerful to apply. This lends further support for the use of the generalized likelihood method.

## 4 Wilks' phenomena

In the last section, we introduced a general framework of the *GLR* test and its various implementation. The approach is general and can be used in many nonparametric testing problems. In the following, we present some established Wilks type of results for various models. This provides a stark evidence for the versatility of the results, which in terms supports the methodology.

### 4.1 Nonparametric regression

Consider the following nonparametric model

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1\ldots, n, \tag{4.22}$$

where $\varepsilon_i$ are iid random variables such that $E(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$. The univariate nonparametric regression model (4.22) is one of the simplest nonparametric models for understanding nonparametric techniques. It has been exhaustedly studied in the literature. Enormous papers have been devoted to the estimation and inference of the univariate regression model and it is impossible to mention all of the related references. See the references in the books mentioned in the second paragraph of the introduction for further details.

We now apply the *GLR* test to the testing problem

$$H_0 : m(x) = \alpha_0 + \alpha_1 x \quad \text{versus} \quad H_1 : m(x) \neq \alpha_0 + \alpha_1 x,$$

where $\alpha_0$ and $\alpha_1$ are unknown parameters. Using the local linear fit with a kernel $K$ and a bandwidth $h$, one can obtain the estimator $\hat{m}_h(\cdot)$ of the unknown function $m(\cdot)$ under the full model. If the error is normal, then the log-likelihood function $\ell(m, \sigma)$ is given by (2.12). Substituting the nonparametric estimator $\hat{m}_h(\cdot)$ into the likelihood function, we obtain the likelihood of generating the collected sample under the nonparametric model as

$$\ell(\hat{m}_h, \sigma) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\mathrm{RSS}_1, \tag{4.23}$$

where $\mathrm{RSS}_1 = \sum_{i=1}^n (Y_i - \hat{m}_h(X_i))^2$. Maximizing the above likelihood with respect to $\sigma^2$ yields $\hat{\sigma}^2 = n^{-1}\mathrm{RSS}_1$. Substituting the estimate into (4.23) gives the likelihood,

$$\ell(\hat{m}_h, \hat{\sigma}) = -\frac{n}{2}\log(\mathrm{RSS}_1) - \frac{n}{2}[1 + \log(2\pi/n)].$$

Denote by $(\hat{a}_0, \hat{\alpha}_1)$ the least squares estimator of $(\alpha_0, \alpha_1)$, and define $\mathrm{RSS}_0 = \sum_{i=1}^n (Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 X_i)^2$. Using a similar argument as above, we get the likelihood under $H_0$ as

$$\ell(\hat{m}_0, \hat{\sigma}_0) = -\frac{n}{2}\log(\mathrm{RSS}_0) - \frac{n}{2}[1 + \log(2\pi/n)].$$

Thus, the *GLR* statistic in (3.18) for the above testing problem is

$$\lambda_{n,1} = \ell(\hat{m}_h, \hat{\sigma}) - \ell(\hat{m}_0, \hat{\sigma}_0) = \frac{n}{2}\log(\mathrm{RSS}_0/\mathrm{RSS}_1). \tag{4.24}$$

Under the null hypothesis and certain conditions, if $nh^{3/2} \to \infty$, the Wilks type of result holds (Fan, Zhang and Zhang, 2001, Section 4.1)

$$r_K \lambda_{n,1} \overset{a}{\sim} \chi^2_{\mu_n}, \tag{4.25}$$

where $\mu_n = r_K c_K |\Omega|/h$ with $|\Omega|$ denoting the Lebesgue's measure of the support of $X$, $r_K = c_K/d_K$ with $c_K = K(0) - 0.5\|K\|^2$ and $d_K = \|K - 0.5K * K\|^2$.

The above result demonstrates that the *GLR* statistic obeys Wilks' phenomenon in this simple setup — the asymptotic null distribution is independent of any nuisance parameters, such as $\sigma^2$ and the density function of the covariate $X$. The normalization factor is $r_K$ rather than 2 in the parametric maximum likelihood ratio test. The degrees of freedom depend on $|\Omega|/h$, the difference of the effective number of parameters used under the null and alternative hypotheses. This can be understood as follows. Suppose that we partition the support of $X$ into equispaced intervals, each with length $h$. This results in $|\Omega|/h$ intervals. Hence, the difference of the number of parameters between the null and the alternative is approximately proportional to $|\Omega/h|$. Since the local linear smoother uses overlapping intervals, the effect number of parameters is slightly different from $|\Omega/h|$. The constant factor $r_K c_K$ reflects this difference.

Based on Wilks' phenomenon, the null distribution of the *GLR* statistic can be estimated by using the following conditional bootstrap method:

(1) Obtain the parametric estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ and nonparametric estimate $\hat{m}(x)$ under both the null and the alternative models. Fix the bandwidth at its estimated value $\hat{h}$ in the estimation stage.
(2) Compute the *GLR* test statistic $\lambda_{n,1}$ and the residuals $\hat{\varepsilon}_i$ from the nonparametric model (for a given data set, we are not certain whether the null model holds, so we use the fits from the larger alternative model, which is consistent under both classes of models).
(3) For each $X_i$, draw a bootstrap residual $\hat{\varepsilon}_i^*$ from the centered empirical distribution of $\hat{\varepsilon}_i$ and compute $Y_i^* = \hat{\alpha}_0 + \hat{\alpha}_1 X_i + \hat{\varepsilon}_i^*$. This forms a conditional bootstrap sample $\{X_i, Y_i^*\}_{i=1}^n$.
(4) Use the above bootstrap sample to construct the *GLR* statistic $\lambda_{n,1}^*$.
(5) Repeat Steps 3 and 4 $B$ times (say B=1,000) and obtain B values of the statistic $\lambda_{n,1}^*$.
(6) Use the B values in Step 5 to determine the quantiles of the test statistic under $H_0$. The p-value is simply the percentage of $\lambda_{n,1}^*$ values greater than $\lambda_{n,1}$.

The above resampling approximation method is the wild bootstrap. Using the same argument as in Fan and Jiang (2005), one can establish the consistency of the above conditional bootstrap estimation.

Consider now more generally the testing problem

$$H_0 : m(x) = m(x;\theta) \quad \text{versus} \quad H_1 : m(x) \neq m(x;\theta). \tag{4.26}$$

When $m(x;\theta)$ is non-linear, the local linear estimate is biased. The result (4.25) does not hold unless the bandwidth $h$ is small enough (e.g., $h = o(n^{-2/9})$; see Fan, Zhang and Zhang, 2001). The bias correction method in §3.4 is to apply the local linear smoother to the data $\{(X_i, Y_i - m(X_i;\hat{\theta})), i = 1, \cdots, n\}$ to obtain the estimator $\hat{m}_h^*(\cdot)$ using a kernel $K$ with the bandwidth $h$. When the null hypothesis in (4.26) holds, the conditional mean function of $Y_i - m(X_i;\hat{\theta})$ given $X_i$ is approximately zero and hence the local linear estimator does not introduce much biases. With the transformation outlined in §3.4, the log-likelihoods under the null and alternative models are the same as before, except $\mathrm{RSS}_0$ and $\mathrm{RSS}_1$ now replaced by

$$\mathrm{RSS}_0^* = \sum_{i=1}^n (Y_i - m(X_i;\hat{\theta}))^2, \qquad \mathrm{RSS}_1^* = \sum_{i=1}^n (Y_i - m(X_i;\hat{\theta}) - \hat{m}_h^*(X_i))^2.$$

The *GLR* statistic with bias correction now becomes (see also (4.24)) $\lambda_{n,1}^* = \frac{n}{2}\log(\mathrm{RSS}_0^*/\mathrm{RSS}_1^*)$. The result (4.25) continues to hold for the bias-corrected *GLR* statistic $\lambda_{n,1}^*$. The procedure is in the same spirit as that used in Härdle and Mammen (1993).

### 4.2 Varying-coefficient models

The varying-coefficient models arise in many statistical problems. They have been successfully applied to nonlinear time series models (Haggan and Ozaki, 1981; Chen and Tsay, 1993; Fan, Yao and Cai, 2003; Fan and Yao, 2003), the multidimensional nonparametric regression (Cleveland, Grosse and Shyu, 1991; Hastie and Tibshirani, 1993; Fan and Zhang, 1999) and generalized linear models (Kauermann and Tutz, 1999; Cai, Fan and Li, 2000). They have also been widely used in the analysis of longitudinal and functional data (Brumback and Rice, 1998; Carrol, Ruppert and Welsh, 1998; Hoover et al., 1998; Huang, Wu and Zhou, 2002; Fan and Li, 2004) and financial modeling (Härdle, Herwartz and Spokoiny, 2003; Mercurio and Spokoiny, 2004).

In the multiple nonparametric regression, the varying-coefficient model assumes

$$Y = a_1(U)X_1 + \cdots + a_p(U)X_p + \varepsilon,$$

where $\varepsilon$ is independent of covariates $(U, X_1, \ldots, X_p)$ and has mean zero and variance $\sigma^2$. It provides a useful tool for capturing possible nonlinear interactions among covariates $\mathbf{X}$ and $U$ and allows us to examine the extent to which the regression coefficients changes over the level of $U$. It effectively avoids the issue of the curse of dimensionality for multi-dimensional nonparametric regression.

Suppose we have a random sample $\{(U_i, X_{i1}, \ldots, X_{ip}, Y_i)\}_{i=1}^n$ from the above model. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\mathbf{A}(U) = (a_1(U), \ldots, a_p(U))^T$. Then the model can be rewritten as

$$Y_i = \mathbf{A}(U_i)^T \mathbf{X}_i + \varepsilon_i. \tag{4.27}$$

The unknown coefficient functions $a_j(\cdot)$ can be estimated by using local linear regression techniques. For any given $u_0$ and $u$ in a neighbourhood of $u_0$, it follows from the Taylor expansion that

$$a_j(u) \approx a_j(u_0) + a_j'(u_0)(u - u_0) \equiv a_j + b_j(u - u_0).$$

Using the data with $U_i$ around $u_0$, one can estimate the coefficient functions and their derivatives by the solutions to the following optimization problem:

$$\min_{a_j, b_j} \sum_{i=1}^n \Big[Y_i - \sum_{j=1}^p \{a_j + b_j(U_i - u_0)\} X_{ij}\Big]^2 K_h(U_i - u_0), \tag{4.28}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K$ is a kernel function, and $h$ is a bandwidth. Let $\{(\hat{a}_j, \hat{b}_j)\}$ be the resulting solutions. Then the local linear regression estimator is simply $\hat{a}_j(u_0) = \hat{a}_j,\ j = 1, \ldots, n$. This yields a nonparametric estimator under the full model (4.27) and the residual sum of squares under the nonparametric model

$$\mathrm{RSS}_1 = \sum_{i=1}^n (Y_i - \hat{\mathbf{A}}(U_i)^T \mathbf{X}_i)^2,$$

where $\hat{\mathbf{A}}(U) = (\hat{a}_1(U), \ldots \hat{a}_p(U))^T$.

In fitting the varying coefficient model (4.27), one asks naturally if the coefficients in $\mathbf{A}(u)$ vary really with $u$ and if certain covariates in $\mathbf{X}$ are related to the response $Y$. The former null hypothesis is parametric: $\mathbf{A}(u) = \boldsymbol{\beta}$, while the latter null hypothesis is nonparametric such as

$$H_0 : a_1(\cdot) = \cdots = a_d(\cdot) = 0, \tag{4.29}$$

in which the covariates $X_1, \cdots, X_d$ are not related to the regression function.

Let us consider testing the following parametric null hypothesis:

$$H_0 :\ \mathbf{A}(u) = \mathbf{A}(u, \boldsymbol{\beta}),$$

which is a generalization of null hypothesis $\mathbf{A}(u) = \boldsymbol{\beta}$. Following the framework in §3.1 which yields $\lambda_{n,1}$, we obtain the *GLR* test statistic

$$\lambda_{n,2} = \frac{n}{2} \log(\mathrm{RSS}_0/\mathrm{RSS}_1),$$

where $\mathrm{RSS}_0 = \sum_{i=1}^n (Y_i - \mathbf{A}(U_i, \hat{\boldsymbol{\beta}})^T \mathbf{X}_i)^2$ with $\hat{\boldsymbol{\beta}}$ being any root-n consistent estimator of $\boldsymbol{\beta}$ under $H_0$.

As unveiled in Fan, Zhang and Zhang (2001), under certain conditions, if $\mathbf{A}(u, \beta)$ is linear in $u$ or $nh^{9/2} \to 0$, then as $nh^{3/2} \to \infty$,

$$r_K \lambda_{n,2} \overset{d}{\simeq} \chi^2_{\mu_n}, \tag{4.30}$$

where $\mu_n = p\, r_K c_K |\Omega|/h$ with $|\Omega|$ being the length of the support of $U$, and $r_K$ and $c_K$ defined in (4.25). This Wilks' type of result allows one again to approximate the null distribution of $\lambda_{n,2}$ using a conditional bootstrap method similar to that for $\lambda_{n,1}$.

Now, let us consider the nonparametric null hypothesis (4.29). Under the null hypothesis, (4.27) is still a varying coefficient model: $Y = a_{d+1}(U) X_{d+1} + \cdots + a_p(U) X_p + \varepsilon$. Let $\hat{a}^0_{d+1}(\cdot), \cdots, \hat{a}^0_{d+1}(\cdot)$ be the local linear fit using the *same* kernel $K$ and the *same* bandwidth $h$ as those in fitting the full model (4.27). Denote by $\mathrm{RSS}_0^*$ the resulting sum of the squares, defined similarly to $\mathrm{RSS}_1$. Then, following the same derivation as before, the *GLR* test statistic for (4.29) is

$$\lambda_{n,3} = \frac{n}{2} \log(\mathrm{RSS}_0^*/\mathrm{RSS}_1).$$

For this nonparametric null hypothesis against nonparametric alternative hypothesis, Fan, Zhang and Zhang (2001) also demonstrated Wilks's phenomenon (Theorem 6 of the paper). If $h \to 0$ and $nh^{3/2} \to \infty$, then

$$r_K \lambda_{n,3} \overset{d}{\simeq} \chi^2_{d\, r_K c_K |\Omega|/h}.$$

Comparing it with (4.30), the degrees of freedom here are similar to the classical Wilks Theorem, in which each nonparametric function in $H_1$ is regarded as a parametric function with the number of parameters $r_K c_K |\Omega|/h$. This agrees with our intuition on the model complexity in nonparametric modeling. We would like to stress that the same kernel $K$ and the same bandwidth $h$ have to be used for fitting the null model and alternative model in order to have the Wilks phenomenon. In this way, the nuisance functions $m_{d+1}(\cdot)$, $\cdots$, $m_p(\cdot)$ are modeled with the same complexity under both the null and the alternative hypotheses, as in the parametric models.

We would also like to note that even though we use the normal error to derive the GLR statistic. The above results do not depend on the normality assumption of $\varepsilon$. In this case, one can regard the normal likelihood as the quasi-likelihood. Furthermore, Fan, Zhang and Zhang (2001) showed that the *GLR* test achieves the optimal rate of convergence for hypothesis testing, as formulated in Ingster (1993) and Lepski and Spokoiny (1999).

Note that when $p = 1$ and $X_1 \equiv 1$, model (4.27) becomes the nonparametric regression model. Therefore, the above remarks are applicable to the univariate nonparametric model discussed in §4.1.

### 4.3 Generalized varying-coefficient models

The generalized varying-coefficient models expand the scope of the applications from the normal-like of distributions to the exponential family of distributions, including the Binomial and Poisson distributions (Cai, Fan and Li, 2000). For example, one may wish to examine the extent to which the regression coefficients in the logistic regression vary with a covariate such as age or other exposure variables. The problem can be better handled by generalized varying-coefficient models, which enlarges the scope of applicability of the generalized linear models (McCullough and Nelder, 1989) by allowing the coefficients to depend on certain covariates.

Cai, Fan and Li (2000) proposed a local likelihood estimator to fit the nonparametric coefficients. Unlike the varying-coefficient model (4.27), the estimator of coefficient functions is implicit for generalized varying-coefficient models. Nevertheless, Fan, Zhang and Zhang (2001, Theorem 10) were able to demonstrate that the Wilks phenomenon for the *GLR* test continues to hold.

### 4.4 Varying-coefficient partially linear models

The *GLR* test has been successfully applied to various nonparametric inferences in Fan, Zhang and Zhang (2001). A natural question is if the approach is applicable to the semiparametric models with the focus on the hypotheses on parameter components instead of nonparametric functions. To address this issue, Fan and Huang (2005) appealed to the varying-coefficient partially linear models to illustrate the applicability of the *GLR* test to semiparametric inferences.

The varying-coefficient partially linear model admits the following form

$$Y = \boldsymbol{\alpha}^T(U)\mathbf{X} + \boldsymbol{\beta}^T\mathbf{Z} + \varepsilon, \tag{4.31}$$

where $\varepsilon_i$ is independent of $(U, \mathbf{X}, \mathbf{Z})$ and satisfies that $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ is a $q$-dimensional vector of unknown parameters, and $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_p(\cdot))^T$ is a $p$-dimensional vector of unknown coefficient functions. It allows one to explore the partial nonlinear interactions with $U$ and maintain the interpretability and explanatory power of parametric regression. The model has been studied by Zhang, Lee and Song (2002) and Li et al. (2002) in regression setting and by Lin and Carroll (2001a,b), Lin and Ying (2001), and Fan and Li (2004) for the analysis of longitudinal data. It is an extension of the varying coefficient model in §4.2 and a generalization of the partial linear model (see Härdle, Liang and Gao, 2000, and references therein).

There are many estimation methods for the above model. A semiparametrically efficient estimation is the profile least-squares approach in Fan and Huang (2005), which we now describe.

Assume that we have a random sample of size $n$, $\{(U_k, X_{k1}, \ldots, X_{kp}, Z_{k1} \ldots, Z_{kq}, Y_k), k = 1, \ldots, n\}$, from model (4.31). For any given $\boldsymbol{\beta}$, we can rewrite model (4.31) as

$$Y_k^* = \sum_{i=1}^{p} \alpha_i(U_k)X_{ki} + \varepsilon_k, \quad k = 1, \ldots, n,$$

where $Y_k^* = Y_k - \sum_{j=1}^q \beta_j Z_{kj}$. This is the varying-coefficient model considered in the previous section. Then the coefficient functions $\alpha_i$ can be estimated using (4.28) with $Y_i$ replaced by $Y_i^*$. After obtaining the estimate of $\boldsymbol{\alpha}(\cdot)$, say $\hat{\boldsymbol{\alpha}}(\cdot;\boldsymbol{\beta})$, we substitute $\hat{\boldsymbol{\alpha}}$ into (4.31) and get the following synthetic regression problem:

$$Y = \hat{\boldsymbol{\alpha}}^T(U;\boldsymbol{\beta})\mathbf{X} + \boldsymbol{\beta}^T\mathbf{Z} + \varepsilon.$$

Using the least squares to solve the above problem leads to the profile least-squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. The estimate of $\boldsymbol{\alpha}(\cdot)$ is simply $\hat{\boldsymbol{\alpha}}(\cdot;\hat{\boldsymbol{\beta}})$. For fast implementation, Fan and Huang (2005) used an iterative procedure to compute the profile estimator $\hat{\boldsymbol{\beta}}$.

We now consider testing significance of some of the parametric components in model (4.31), which leads to testing $H_0 : \beta_1 = \cdots = \beta_l = 0$, for $l \le q$. More generally, one may consider the linear hypothesis

$$H_0 : \mathbf{A}\boldsymbol{\beta} = 0 \ \text{ versus } \ H_1 : \mathbf{A}\boldsymbol{\beta} \neq 0,$$

where $\mathbf{A}$ is a given $l \times q$ full rank matrix. This is a semiparametric hypothesis versus another semiparametric hypothesis testing problem, and the conventional maximum likelihood ratio test cannot be applied, because the nonparametric MLEs for functions $\boldsymbol{\alpha}(\cdot)$ do not exist. A natural alternative is to relax the requirement on the estimate of function $\boldsymbol{\alpha}(\cdot)$ and use any reasonable nonparametric estimates to construct the *GLR* test in (3.18).

Suppose the error $\varepsilon \overset{d}{=} \mathcal{N}(0,\sigma^2)$. Then under model (4.31), the log-likelihood function is

$$\ell(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma) = -n\log(\sqrt{2\pi}\sigma) - \text{RSS}_1/(2\sigma^2),$$

where $\text{RSS}_1 = \sum_{i=1}^n [Y_i - \boldsymbol{\alpha}(U_i)^T\mathbf{X}_i - \boldsymbol{\beta}^T\mathbf{Z}_i]^2$. Substituting $\hat{\boldsymbol{\alpha}}(\cdot;\boldsymbol{\beta})$ into the above likelihood function, we get

$$\ell(\hat{\boldsymbol{\alpha}}(\cdot;\boldsymbol{\beta}),\boldsymbol{\beta},\sigma) = -n\log(\sqrt{2\pi}\sigma) - \text{RSS}(\boldsymbol{\beta})/(2\sigma^2), \tag{4.32}$$

where $\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^n [Y_i - \hat{\boldsymbol{\alpha}}(U_i;\boldsymbol{\beta})^T\mathbf{X}_i - \boldsymbol{\beta}^T\mathbf{Z}_i]^2$. Maximizing (4.32) with respect to $\boldsymbol{\beta}$ and $\sigma$ produces the profile likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2 = n^{-1}\text{RSS}_1$, where $\text{RSS}_1 = \text{RSS}(\hat{\boldsymbol{\beta}})$ is the residual sum of squares. Substituting these estimators into (4.32) yields the generalized likelihood under $H_1$

$$\ell(H_1) = -\frac{n}{2}\log(2\pi/n) - \frac{n}{2}\log(\text{RSS}_1) - \frac{n}{2}.$$

Similarly, maximizing (4.32) subject to constraint in $H_0$ yields the profile likelihood estimator for the null model. Denote by $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\alpha}}_0$ the resulting estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively. Then the generalized likelihood under $H_0$ is

$$\ell(H_0) = -\frac{n}{2}\log(2\pi/n) - \frac{n}{2}\log(\text{RSS}_0) - \frac{n}{2},$$

where $\text{RSS}_0 = \text{RSS}(\hat{\boldsymbol{\beta}}_0)$. According to the definition in (3.18), the *GLR* statistic is

$$\lambda_{n,4} = \ell(H_1) - \ell(H_0) = \frac{n}{2}\log\frac{\text{RSS}_0}{\text{RSS}_1}.$$

Under certain conditions, Fan and Huang (2005) proved that the asymptotic null distribution of $\lambda_{n,4}$ is the Chi-square distribution with $l$ degrees of freedom. This shows that the asymptotic null distribution is independent of the design density and the nuisance parameters $\sigma^2$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}(\cdot)$. Hence, the critical value can be computed either by the asymptotic distribution or by simulations with nuisance parameters' values taken to be reasonable estimates under $H_0$. It is also demonstrated that one can proceed to the likelihood ratio test as if the model were parametric.

In addition to the above parametric testing problem, one may also be interested in inference on nonparametric components. For example, consider the hypothesis testing problem

$$H_0 : \alpha_1(\cdot) = \alpha_1, \ldots, \alpha_p(\cdot) = \alpha_p,$$

where the functions $\alpha_k$ $(k = 1, \ldots, p)$ are unknown parameters. Since the parametric component can be estimated at a root-$n$ rate and regarded as known in nonparametric inference, the techniques and the results in the last section can be extended to the current model. In fact, let $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_p$ and $\tilde{\boldsymbol{\beta}}$ be the least-squares estimators under $H_0$, then the *GLR* statistic is defined as

$$\lambda_{n,5} = \frac{n}{2} \log \frac{\mathrm{RSS}(H_0)}{\mathrm{RSS}(H_1)},$$

where $\mathrm{RSS}(H_0) = \sum_{i=1}^n [Y_i - \sum_{j=1}^p \tilde{\alpha}_j X_{ij} - \tilde{\boldsymbol{\beta}}^T \mathbf{Z}_i]^2$, and $\mathrm{RSS}(H_1) = \mathrm{RSS}_1$ is the same as that for $\lambda_{n,4}$. As revealed in Fan and Huang (2005), under certain conditions, the Wilks type of result in (4.30) still holds for $\lambda_{n,5}$.

It is worthwhile to note again that the normality assumption is used merely to derive the *GLR* statistic. It is not needed for deriving the asymptotic properties. In other words, the *GLR* statistic can be regarded as that based on the normal quasi-likelihood when the error distribution is not normal.

## 4.5 Additive models

Additive models are an important family of structured multivariate nonparametric models. They model a random sample $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ by

$$Y_i = \alpha + \sum_{d=1}^{D} m_d(X_{di}) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{4.33}$$

where $\{\varepsilon_i\}$ is a sequence of independent and identically distributed random variables with mean zero and finite variance $\sigma^2$. For identifiability of $m_d(x_d)$, it is usually assumed that $E[m_d(X_{di})] = 0$ for all $d$.

The additive models, proposed by Friedman and Stuetzle (1981) and Hastie and Tibshirani (1990), have been widely used in multivariate nonparametric modeling. As all unknown functions are one-dimensional, the difficulty associated with the so-called "curse of dimensionality" is substantially reduced (see Stone (1985) and Hastie and Tibshirani (1990)). In fact, Fan, Härdle and Mammen (1998) showed that an additive component can be estimated as well as in the case where the rest of the components are known. This phenomenon was also revealed in Horowitz and Mammen (2004) and Jiang and Li (2007) using two-step estimation methods based on the least squares and the M-estimation, respectively. Various methods for estimating the additive functions have been proposed, including the marginal integration estimation method, the backfitting algorithm, the estimating equation method, Fourier, spline, and wavelet approaches, among others. Some additional references can be found in Fan and Jiang (2005) and Jiang et al (2007).

The backfitting algorithm is frequently used to estimate the unknown components in model (4.33) due to its intuitive and mathematical appeal and the availability of software. See, for example, Buja, Hastie and Tibshirani (1989) and Opsomer and Ruppert (1998). As the backfitting algorithm is frequently employed, determined efforts were made by Fan and Jiang (2005) to study the *GLR* test for the following hypothesis testing problem using the backfitting algorithm with the local polynomial smoothing technique to estimate nonparametric components:

$$\begin{aligned}&H_0:\ m_{D-d_0}(x_{D-d_0}) = \cdots = m_D(x_D) = 0\\ &\quad\text{versus}\quad H_1: m_{D-d_0}(x_{D-d_0}) \neq 0, \cdots, \text{or}\quad m_D(x_D) \neq 0,\end{aligned}\tag{4.34}$$

for some integer $d_0 \in \{0, 1, \ldots, D-1\}$. This amounts to testing the significance of the variables $X_{D-d_0}, \cdots, X_D$ in presence of nuisance functions $m_1(\cdot), \cdots, m_{D-d_0-1}(\cdot)$, which is a nonparametric null hypothesis against a nonparametric alternative hypothesis.

Since the distribution of $\varepsilon_i$ is unknown, we do not have a known likelihood function. Pretending that error distribution is normal, $\mathcal{N}(0, \sigma^2)$, the log-likelihood under model (4.33) is

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}\Big(Y_k - \alpha - \sum_{d=1}^{D} m_d(X_{dk})\Big)^2.$$

The unknown constant $\alpha$ can be estimated by the sample mean $\bar{Y}$ of $Y_i$. Under the alternative model $H_1$, based on the backfitting algorithm using the local polynomial smoothing technique as a building block, the additive components $m_d$ can be estimated. Denote by $\hat{m}_d$ the resulting estimator of $m_d$. Replacing the intercept $\alpha$ and the unknown function $m_d(\cdot)$ by $\hat{\alpha}$ and $\widehat{m}_d(\cdot)$ respectively leads to

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathrm{RSS}_1,$$

where $\mathrm{RSS}_1 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \sum_{d=1}^{D}\widehat{m}_d(X_{dk}))^2$. Maximizing over the parameter $\sigma^2$, we obtain a likelihood of the alternative model:

$$-\frac{n}{2}\log(2\pi/n) - \frac{n}{2}\log(\mathrm{RSS}_1) - \frac{n}{2}.$$

Therefore, up to a constant term, the log-likelihood of model (4.33) is taken as $\ell(H_1)$ $= -\frac{n}{2}\log(\mathrm{RSS}_1)$. Similarly, the log-likelihood for $H_0$ can be taken as $\ell(H_0)$ $= -\frac{n}{2}\log(\mathrm{RSS}_0)$, with $\mathrm{RSS}_0 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \sum_{d=1}^{D-1}\widetilde{m}_d(X_{dk}))^2$, and $\widetilde{m}_d(x_d)$ the estimator of $m_d(x_d)$ under $H_0$, using the same backfitting algorithm. Then the *GLR* test statistic in (3.18) is

$$\lambda_{n,6} = \ell(H_1) - \ell(H_0) = \frac{n}{2}\log\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1},$$

which compares the likelihood of the nearly best fitting in the alternative models with that under the null models if the error is normal.

Let $h_d$ be the bandwidth in smoothing $m_d$ using a local linear fit and $K$ be the kernel function. Let $|\Omega_d|$ be the Lebesgue measure of the support of the density $f_d(\cdot)$ of the covariate $X_d$. Put

$$\mu_n = c_K \sum_{d=D-d_0}^{D} \frac{|\Omega_d|}{h_d}, \qquad \sigma_n^2 = d_K \sum_{d=D-d_0}^{D} \frac{|\Omega_d|}{h_d}, \qquad r_K = \mu_n/\sigma_n^2,$$

where the constant $c_K$ and $d_K$ are the same as in (4.25). Under regularity conditions, Fan and Jiang (2005) established that

$$r_K \lambda_{n,6} \overset{d}{\simeq} \chi^2_{r_K \mu_n}. \tag{4.35}$$

Indeed, they showed a more general result with different order of local polynomial fit $p_d$ for the $d$-th component. The result (4.35) does not depend also on the normality assumption, which motivates the procedure.

The above result shows that Wilk phenomenon continues to hold. It also holds for testing the parametric null hypothesis. In this case, the bias correction method can be applied. See Fan and Jiang (2005). The asymptotic null distribution offers a method for determining approximately the $p$-value of the *GLR* test, even though one does not know how accurate it is. Fortunately, the Wilks phenomenon permits us to simulate the null distribution of the *GLR* test over a large range of bandwidths with nuisance functions/parameters fixed at their estimated values. See Fan and Jiang (2005) and §3.5 for the conditional bootstrap method.

It is worthwhile to note that unlike the degrees of freedom in (4.30) for the varying-coefficient models, the degrees of freedom for (4.35) are not additive, though the parameters $\mu_n$ and $d_n$ are. This is another difference from the classical likelihood ratio test.

In (4.34) one may check the significance of those components or if a family of parametric models (e.g. multiple linear regression models) fit the data. These two problems can be generalized as the following testing problems, validating if the $m_d$'s have parametric forms (for $d = D - d_0, \ldots, D$):

$$\begin{aligned} H_0' &: \; m_{D-d_0}(x_{D-d_0}) \in \mathcal{M}_{\Theta, D-d_0}, \ldots, m_D(x_D) \in \mathcal{M}_{\Theta, D} \quad \text{versus} \\ H_1' &: \; m_{D-d_0}(x_{D-d_0}) \notin \mathcal{M}_{\Theta, D-d_0}, \ldots, \text{or } m_D(x_D) \notin \mathcal{M}_{\Theta, D} \end{aligned}$$

where $\mathcal{M}_{\Theta,d} = \{m_\theta(x_d),\, \theta \in \Theta_d\}$ (for $d = D - d_0, \ldots, D$) are sets of functions of parametric forms, and the parameter space $\Theta_d$ contains the true parameter value $\theta_{0,d}$. In particular, if $d_0 = D - 1$, the problem is validating if a parametric family fits adequately the data. As shown in Jiang et al (2007), the Wilks type of result in (4.35) continues to hold for the above testing problem.

For the following partly linear additive model

$$Y_i = \mathbf{Z}_i^T \boldsymbol{\beta} + \sum_{d=1}^{D} m_d(X_{di}) + \varepsilon_i, \quad i = 1, \cdots, n,$$

as in Section 4.4, one may also test if $H_0 : A\boldsymbol{\beta} = 0$. A *GLR* test statistic similar to $\lambda_{n,4}$ can be developed using profile likelihood estimation, but it would involve much more complicated techniques to establish Wilks' phenomenon.

4.6 Spectral density estimation

Let $X_t$, for $t = 0, \pm 1, \pm 2, \ldots$, be a stationary time series with mean zero and autocovariance function $\gamma(u) = E(X_t X_{t+u})$ $(u = 0, \pm 1, \pm 2, \ldots)$. Then its spectral density is

$$g(x) = (2\pi)^{-1} \sum_{u=-\infty}^{\infty} \gamma(u) \exp(-iux), \quad x \in [0, \pi].$$

Based on the observed time series $X_1, \ldots, X_T$, one can construct the periodogram

$$I_T(w_k) = T^{-1} \sum_{t=1}^{T} |X_t \exp(-itw_k)|^2, \quad w_k = 2\pi k/T \quad (k = 1, \ldots, n, \quad n = [(T-1)/2]).$$

The periodogram is an asymptotically unbiased estimator of $g(x)$. However, it is not a consistent estimator of the spectral density (Brillinger, 1981, Chapter 5; Brockwell and Davis, 1991, Chapter 10). A consistent estimator of $g(x)$ can be obtained by locally averaging the periodograms. Most traditional methods are based on this approach; see for example Brillinger (1981).

Alternative estimators, such as the smoothed (log-)periodogram and Whittle likelihood-based estimator, have received much attention in the literature. For example, Wahba (1980) considered spline approximations to the log-periodogram using the least-square method; Pawitan and O'Sullivan (1994) and Kooperberg et al. (1995a,b) used Whittle's likelihood to estimate parameters in the spline models; Fan and Kreutzberger (1998) studied automatic procedures for estimating spectral densities, using the local linear fit and the local Whittle's likelihood; Jiang and Hui (2004) proposed a generalized periodogram and smoothed it using local linear approximations when missing data appeared.

Consider testing whether or not the spectral density of the observed time series $\{X_t\}_{t=1}^{T}$ belongs to a specific parametric family $g_\theta(\cdot) : \theta \in \Theta$. The problem can be formulated as testing the hypothesis $H_0 : g(\cdot) = g_\theta(\cdot)$ versus $H_1 : g(\cdot) \notin g_\theta(\cdot)$, which is equivalent to testing

$$H_0 : m(\cdot) = m_\theta(\cdot) \ \text{ versus } \ H_1 : m(\cdot) \notin m_\theta(\cdot),$$

where $m_\theta(\cdot) = \log g_\theta(\cdot)$.

There are several approaches to testing the above hypotheses for the spectral density. For example, the testing procedure in Paparoditis (2000) using the Priestley-Chao estimator and an $L_2$-distance, the testing approach without smoothing, as with Dzhaparidze's (1986, p. 273) test statistic based on a cumulative rescaled spectral density, and the Kolmogorov-Smirnov and Cramér-von Mises tests in Anderson (1993). These test methods directly compared various spectral estimation under the null and alternative hypotheses. As illustrated in Section 1, such discrepancy-based tests suffer from disadvantages over the *GLR* test.

Note that periodograms $I_T(w_k)$ are asymptotically exponentially distributed with mean $g(w_k)$ and asymptotically independent (see Fan and Yao, 2003):

$$(2\pi)^{-1} I_T(w_k) = g(w_k) V_k + R_n(w_k) \quad (k = 1, \ldots, n),$$

where $V_k$ $(k = 1, \ldots, n)$ are independently and identically distributed with the standard exponential distribution and $R_n(w_k)$ is a term that is asymptotically negligible. If we let $Y_k = \log\{I_T(w_k)/(2\pi)\}$ and $m(\cdot) = \log g(\cdot)$, then

$$Y_k = m(w_k) + z_k + r_k \ \ (k = 1, \ldots, n), \tag{4.36}$$

where $r_k = \log[1 + R_n(w_k)/g(w_k)V_k]$ and $z_k = \log V_k$. Therefore, $\{z_k\}_{k=1}^n$ are independently and identically distributed random variables with density function $f_z(x) = \exp\{-\exp(x)+x\}$, and $r_k$ is an asymptotically negligible term; see Lemma A1 of Fan and Zhang (2004). As shown in Davis and Jones (1968), the mean of $z_k$ is the Euler constant, which is $E(z_k) = C_0 = -0\Delta 57721$, and the variance is $\mathrm{var}(z_k) = \pi^2/6$. Let $Y_k^* = Y_k - C_0$ and $z_k^* = z_k - C_0$. Using (4.36) and ignoring the term $r_k$, we have the standard nonparametric regression model similar to (2.10)

$$Y_k^* = m(w_k) + z_k^* \ \ (k = 1, \ldots, n).$$

Then a similar *GLR* test statistic to $\lambda_{n,1}$ can be developed based on the least-squares estimation of $m(\cdot)$. Since the distribution of $z_k^*$ is not normal, the test based on the least-squares estimation does not fully use the likelihood information and cannot be powerful. Furthermore, the likelihood-based approach is more appealing as demonstrated in Fan and Zhang (2004).

For any given spectral density function, by ignoring the term $r_k$ in (4.36), we obtain the approximated log-likelihood function

$$\ell(H_1; m) = \sum_{k=1}^n [Y_k - m(w_k) - \exp\{Y_k - m(w_k)\}].$$

For any $x$, approximating $m(w_k)$ by the linear function $a + b(w_k - x)$ for $w_k$ near $x$, we obtain the local log-likelihood function

$$\sum_{k=1}^n [Y_k - a - b(w_k - x)\exp\{Y_k - a - b(w_k - x)\}]K_h(w_k - x). \tag{4.37}$$

The local maximum likelihood estimator $\hat{m}_{LK}(x)$ of $m(x)$ is $\hat{a}$ in the maximizer $(\hat{a}, \hat{b})$ of (4.37). Under the null hypothesis, the log-likelihood function of (4.36) would approximately be

$$\ell(H_0; \theta) = \sum_{k=1}^n [Y_k - m_\theta(w_k) - \exp\{Y_k - m_\theta(w_k)\}].$$

Its maximizer $\hat{\theta}$ would be the maximum likelihood estimate of $\theta$. Then by applying (3.18), a *GLR* test statistic can be constructed as

$$\begin{aligned}\lambda_{n,7} &= \ell(H_1; \hat{m}_{LK}) - \ell(H_0; \hat{\theta}) \\ &= \sum_{k=1}^n [\exp\{Y_k - m_{\hat{\theta}}(w_k)\} + m_{\hat{\theta}}(w_k) - \exp\{Y_k - \hat{m}_{LK}(w_k)\} - \hat{m}_{LK}(w_k)].\end{aligned}$$

Let $\mu_n = \pi c_K/h$ and $r_K$ as in (4.25). Then under certain conditions, as shown in Fan and Zhang (2004),

$$r_K \lambda_{n,7} \overset{d}{\simeq} \chi^2_{r_K \mu_n}.$$

This means that the Wilks phenomenon exists in spectral density estimation. It permits one to use the bootstrap to obtain the null distribution (see Section 3). Also a bias-corrected *GLR* statistic can be constructed using the method in §3.4. For details, see Fan and Zhang (2004).

## 4.7 Diffusion models with discrete jumps

Consider the following diffusion model:

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dW_t + J_t\,dN_t, \tag{4.38}$$

where the drift $\mu$ and the diffusion $\sigma$ are unknown. Assume that the intensity of $N$ and density of $J$ are unknown functions $\lambda(X_t)$ and $\nu(\cdot)$, respectively. If we believe that the true process is a jump-diffusion with local characteristics $(\mu, \sigma^2, \lambda, \nu)$, a specification test checks whether the functions $(\mu, \sigma^2, \lambda, \nu)$ belong to the parametric family

$$\mathcal{P} = \{(\mu(X_t, \theta_1), \sigma^2(X_t, \theta_2), \lambda(X_t, \theta_3), \nu(\cdot;\theta_4) \,|\, \theta_i \in \Theta_i, i = 1, \ldots, 4\},$$

where $\Theta_i$'s are compact subsets of $R^K$. This is equivalent to testing if there exist $\theta_i \in \Theta_i$ such that the following null model holds:

$$dX_t = \mu(X_t, \theta_1)\,dt + \sigma(X_t, \theta_2)\,dW_t + J_t\,dN_t. \tag{4.39}$$

In the case of jump-diffusions, the parametrization $\mathcal{P}$ corresponds to a parametrization of the marginal and transitional densities:

$$\{(\pi(\cdot, \theta), p(\cdot|\cdot, \theta)) \mid (\mu(\cdot, \theta_1), \sigma^2(\cdot, \theta_2), \lambda(\cdot, \theta_3), \nu(\cdot;\theta_4) \in \mathcal{P},\ \theta_i \in \Theta_i\}.$$

The null and alternative hypotheses are of the form

$$H_0 : p(y|x) = p(y|x, \theta) \quad \text{vs} \quad H_1 : p(y|x) \neq p(y|x, \theta),$$

with the inequality for some $(x, y)$ in a subset of non-zero Lebesgue measure. The GLR tests can be used to answer the above question. The work of Aït-Sahalia, Fan and Peng (2005) does not embed (4.39) directly into (4.38). Instead, they merely assume that the alternative model is Markovian with a transition density $p(y|x)$. See Figure 2.

Suppose the observed process $\{X_t\}$ is sampled at the regular time points $\{i\Delta, i = 1, \ldots, n+1\}$. Let $p(y|x)$ be the transition density of the series $\{X_{i\Delta},\ i = 1, \ldots, n+1\}$. For simplicity of notation, we rewrite the observed data as $\{X_i,\ i = 1, \ldots, n+1\}$. Then following Fan, Yao and Tong (1996), we can estimate $p(y|x)$ by

$$\hat{p}(y|x) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} W_n\Big(\frac{X_i - x}{h_1}; x\Big) K\Big(\frac{X_{i+1} - y}{h_2}\Big),$$

where $W_n$ is the effective kernel induced by the local linear fit.

Note that the logarithm of the likelihood function of the observed data $\{X_i\}_{i=1}^{n+1}$ is

$$\ell(p) = \sum_{i=1}^{n} \log p(X_{i+1}|X_i),$$

after ignoring the initial stationary density $\pi(X_1)$. It follows from (3.18) that the GLR test statistic compares the likelihoods under the null and alternative hypotheses, which leads to

$$\lambda_{n,8} = \ell(\hat{p}) - \ell(p(\cdot|\hat{\theta})) = \sum_{i=1}^{n} \log \hat{p}(X_{i+1}|X_i)/p(X_{i+1}|X_i, \hat{\theta}).$$

Noticing that the conditional density cannot be estimated well at the boundary region of $X$-variable, Aït-Sahalia, Fan and Peng (2005) introduced a weight function, $w$, to reduce the influences of the unreliable estimates, leading to the test statistic

$$T_0 = \sum_{i=1}^{n} w(X_i, X_{i+1}) \log \hat{p}(X_{i+1}|X_i)/p(X_{i+1}|X_i, \hat{\theta}).$$

Since the parametric and nonparametric estimators are approximately the same under $H_0$, $T_0$ can be approximated by a Taylor expansion:

$$\begin{aligned} T_0 \approx & \sum_{i=1}^{n} \frac{\hat{p}(X_{i+1}|X_i) - p(X_{i+1}|X_i, \hat{\theta})}{p(X_{i+1}|X_i, \hat{\theta})} w(X_i, X_{i+1}) \\ & - \frac{1}{2} \sum_{i=1}^{n} \Big\{ \frac{\hat{p}(X_{i+1}|X_i) - p(X_{i+1}|X_i, \hat{\theta})}{p(X_{i+1}|X_i, \hat{\theta})} \Big\}^2 w(X_i, X_{i+1}). \end{aligned}$$

To avoid complicated technicalities, they proposed the following $\chi^2$-test statistic

$$T_1 = \sum_{i=1}^{n} \Big\{ \frac{\hat{p}(X_{i+1}|X_i) - p(X_{i+1}|X_i, \hat{\theta})}{p(X_{i+1}|X_i, \hat{\theta})} \Big\}^2 w(X_i, X_{i+1})$$

and justified Wilks' phenomenon, that is, under certain conditions,

$$r_1 T_1 \overset{d}{\simeq} \chi^2_{r_1 \mu_1},$$

where

$$r_1 = \Omega_w \|W\|^2 \|K\|^2 / (\|w\|^2 \|W * W\|^2 \|K * K\|^2),$$

$$\mu_1 = \Omega_w \|W\|^2 \|K\|^2 / (h_1 h_2) - \Omega_x \|W\|^2 / h_1,$$

where for any function $f(\cdot)$, $\|f\|^2 = \int f^2(x)\,dx$, $\Omega_w = \int w(x, y)\,dx\,dy$, and $\Omega_x = \int E\{w(X, Y)|X = x\}\,dx$. This result again enables one to simulate the null distribution of the test statistic using the bootstrap method.

### 4.8 Others

The *GLR* tests are also applicable to many other situations. For example, Fan and Zhang (2004) studied the *GLR* tests using sieve empirical likelihood, which aims to construct the *GLR* test to adapt to unknown error distributions, including the conditional heteroscedasticity. Jiang et al (2007) extended Wilks' phenomenon to semiparametric additive models and studied the optimality of the tests.

## 5 Conclusion and outlook

As demonstrated above, the GLR test is a natural, powerful and generally applicable inference tool. Wilks phenomenon has been unveiled for a variety of models, which makes finite sample simulation feasible in determining the null distributions of *GLR* test statistics. However, compared with wide application of the parametric likelihood inference, the *GLR* inference tool is still underdeveloped. The Wilks type of results hold not only for the various problems that we have studied. They should be valid for nearly all regular nonparametric testing problems. There are many topics to be explored for the *GLR* tests, and we only scratched the surface of this exciting field.

In the following we list several open problems to conclude this article.

- *Application to other models.* The *GLR* test is so general that it can be used for checking if some nonparametric components in a statistical model admit certain parametric forms. Most established Wilks phenomenon is based on the *GLR* with normal quasi-likelihood. The likelihood based nonparametric models have barely been touched (except the results in §4.3). One example is to establish similar results to those in §4.5 for generalized additive models. The other example is to apply it to the Cox hazard regression models with additive nonparametric structure, or varying-coefficient form, or partially linear components.
- *Platform of smoothing.* Most of the Wilks phenomena of *GLR* statistics are unveiled by using the local polynomial smoothing as the platform of nonparametric estimation. Intellectually, one may ask whether these phenomena are shared by other smoothing methods, such as the polynomial splines or smoothing splines.
- *Choice of smoothing parameters.* The powers of *GLR* tests depend on the choice of smoothing parameters. It is important to develop the criteria and theoretic results for bandwidth selection in an attempt to optimize the powers of *GLR* tests, so that a data-driven selection of bandwidth can be established for optimizing the powers of *GLR* tests.
- *Optimality of multi-scale tests.* The multi-scale tests in (3.21) was proposed in Fan, Zhang and Zhang (2001). However, it is unknown if the resulting procedure (3.21) possesses the adaptive optimality. For the regression problems, Fan (1996), Spokoiny (1996), Fan and Huang (2001), Horowitz and Spokoiny (2001, 2002) established the adaptive optimality results. However, for most problems discussed in §4, such kind of adaptive optimality results are unknown. Further investigations are needed.
- *Implementation of multi-scale tests.* The multiple scale test (3.21) is hard to compute. In addition, its null distribution is harder to approximate. This leads Zhang (2003a) to replace the maximization by a discrete set of values. How to set such discrete grids? Are the null distributions easier to approximate? and how to approximate them?
- *Robustness.* Since the *GLR* test statistics $\{\lambda_{n,k}\}$ $(k = 1, \ldots, 8)$ are constructed using the local least squares or maximum likelihood estimation, they are not robust against outliers in the $Y$-space. It is interesting to investigate robust *GLR* tests, for example, with $\mathrm{RSS}_0$ and $\mathrm{RSS}_1$ replaced by their robust versions, but whether the Wilks phenomenon still exists remains unknown.

We have laid down a general blueprint for testing problems with nonparametric alternatives. While we have observed the Wilks phenomenon and demonstrated it for a few useful cases, it is impossible for us to verify the phenomenon for all nonparametric hypothesis testing problems. The Wilks phenomenon needs to be checked for other problems that have not been covered in this paper. In addition, most of the topics outlined in the above discussion remains open and are technically and intellectually challenging. More developments are needed, which will push the core of statistical theory and methods forward.

## References

Aït-Sahalia Y, Fan J, Peng H (2005) Nonparametric transition-based tests for Jump-diffusions. Unpublished manuscript.

Anderson TW (1993) Goodness of fit tests for spectral distributions. Ann. Statist. 21:830–847

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

Barnard GA, Jenkins GM, Winsten CB (1962) Likelihood inference and time series. J. Roy. Statist. Soc. Ser. A 125:321–372

Berger JO, Wolpert RL (1988) The Likelihood Principle, 2nd edn. The Institute of Mathematical Statistics, Haywood, CA.

Bickel PJ, Ritov Y (1988) Estimating integrated squared density derivatives: Sharp order of convergence estimates. Sankhyā, Ser. A 50:381–393

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

Birnbaum A (1962) On the foundations of statistical inference (with discussion). J. Amer. Statist. Assoc. 57:269–326

Brillinger DR (1981) Time Series. Data Analysis and Theory, 2nd edn. Holden-Day Series in Time Series Analysis. Holden-Day, Inc., Oakland, California

Brockwell PJ, Davis RA (1991) Time Series: Theory and Methods, 2nd edn. Springer, New York

Brown LD, Low M (1996) A constrained risk inequality with applications to nonparametric functional estimation. Ann. Statist. 24:2524–2535

Brumback B, Rice JA (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). J. Amer. Statist. Assoc. 93:961–994

Buja A, Hastie TJ, Tibshirani RJ (1989) Linear smoothers and additive models. Ann. Statist. 17:453–555

Cai Z, Fan J, Li R (2000) Efficient estimation and inferences for varying-coefficient models. J. Amer. Statist. Assoc. 95:888–902

Carrol RJ, Ruppert D, Welsh AH (1998) Nonparametric estimation via local estimating equations. J. Amer. Statist. Assoc. 93:214–227

Chan KC, Karolyi AG, Longstaff FA, Sanders AB (1992) An empirical comparison of alternative models of the short-term interest rate. J. Finance 47:1209–1227

Chen R, Tsay RJ (1993) Functional-coefficient autoregressive models. J. Amer. Statist. Assoc. 88:298–308

Cleveland WS, Grosse E, Shyu WM (1991) Local regression models. In: Chambers, JM, Hastie TJ (eds) Statistical Models in S. Chapman & Hall Computer Science Series. CRC Press, Inc., Boca Raton, Florida, pp. 309–376

Cox JC, Ingersoll JE, Ross SA (1985) A theory of the term structure of interest rates. Econometrica, 53:385–467

Davis HT, Jones RH (1968) Estimation of the innovation variance of a stationary time series. J. Amer. Statist. Assoc. 63:141–149

Donoho DL, Nussbaum M (1990) Minimax quadratic estimation of a quadratic functional. J. Complexity 6:290–323

Dzhaparidze K (1986) Parameter Estimation and Hypothesis Testing on Spectral Analysis of Stationary Time Series. Springer, New York

Edwards AWF (1972) Likelihood, 1st edn. Cambridge University Press, Cambridge

Edwards AWF (1974) The history of likelihood. Int. Statist. Rev. 42:9–15

Efromovich S (1999) Nonparametric Curve Estimation: Methods, Theory and Applications. Springer-Verlag, New York

Efron B, Tibshirani R (1995) An Introduction to the Bootstrap. Chapman and Hall, New York

Eubank RL (1999) Spline Smoothing and Nonparametric Regression, 2nd edn. Marcel Dekker, New York

Eubank RL, Hart JD (1992) Testing goodness-of-fit in regression via order selection criteria. Ann. Statist. 20:1412–1425

Eubank RL, LaRiccia VN (1992) Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. Ann. Statist. 20:2071–2086

Fan J (1991) On the estimation of quadratic functionals. Ann. Statist. 19:1273–1294

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

Fan J, Gijbels I (1996) Local Polynomial Modelling and its Applications. Chapman and Hall, London

Fan J, Härdle W, Mammen E (1998) Direct estimation of additive and linear components for high dimensional data. Ann. Statist. 26:943–971

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

Fan J, Huang T (2005) Profile Likelihood Inferences on semiparametric varying-coefficient partially linear models. Bernoulli 11:1031–1057

Fan J, Kreutzberger E (1998) Automatic local smoothing for spectral density estimation. Scand. J. Statist. 25:359–369

Fan J, Jiang J (2005) Nonparametric inference for additive models. J. Amer. Statist. Assoc. 100:890–907

Fan J, Li R (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J. Amer. Statist. Assoc. 99:710–723

Fan J, Yao Q (2003) Nonlinear Time Series: Nonparametric and Parametric Methods. Springer, New York

Fan J, Yao Q, Tong H (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika 83:189–206

Fan J, Yao Q, Cai Z (2003) Adaptive varying-coefficient linear models. J. Roy. Statist. Soc. Ser. B 65:57–80

Fan J, Zhang C (2003) A re-examination of diffusion estimations with applications to financial model validation. J. Amer. Statist. Assoc. 98:118–134

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

Fan J, Zhang J (2004). Sieve empirical likelihood ratio tests for nonparametric functions. Ann. Statist. 32:1858–1907

Fan J, Zhang W (1999) Statistical estimation in varying coefficient models. Ann. Statist. 27:1491–1518

Fan J, Zhang W. (2004) Generalized likelihood ratio tests for spectral density. Biometrika 91:195–209

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Phil. Trans. Royal Soc. Ser. A 222–326

Friedman JH, Stuetzle W (1981) Projection pursuit regression. J. Amer. Statist. Assoc. 76:817–823

Glad IK (1998) Parametrically guided non-parametric regression. Scand. J. Statist. 25:649–668

Grama I, Nussbaum M (2002) Asymptotic equivalence for nonparametric regression. Math. Methods Statist. 11:1–36

Gu C (2002) Smoothing Spline ANOVA Models. Springer, New York

Haggan V, Ozaki T (1981) Modeling nonlinear vibrations using an amplitude-dependent autoregression time series model. Biometrika 68:189–196

Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50:1029–1054

Hall P (1993) The Bootstrap and Edgeworth Expansion. Springer, New York

Hall P, Marron JS (1988) Variable window width kernel estimates of probability densities. Prob. Theory Rel. Fields 80:37–49

Härdle W, Herwartz H, Spokoiny V (2003) Time inhomogeneous multiple volatility modelling. J. Fin. Econometrics 1:55–95

Härdle W, Liang H, Gao J (2000). Partially Linear Models. Springer-Verlag, Heidelberg

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

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

Hastie TJ, Tibshirani RJ (1990) Generalized Additive Models. Chapman and Hall, New York

Hastie TJ, Tibshirani RJ (1993) Varying-coefficient models. J. Roy. Statist. Soc. Ser. B 55:757–796

Hjort N, Glad IK (1995) Nonparametric density estimation with a parametric start. Ann. Statist. 23:882–904

Hong Y, Li H (2005) Nonparametric specification testing for continuous-time models with applications to term structure of interest. Review of Financial Studies 18:37 –84

Hoover DR, Rice JA, Wu CO, Yang L-P (1998) Nonparametric smoothing estimates of time-varing coefficient models with longitudinal data. Biometrika 85:809–822

Horowitz JL, Mammen E (2004) Nonparametric estimation of an additive model with a link function. Ann. Statist. 32:2412–2443

Horowitz JL, Spokoiny GG (2001) An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. Econometrica 69:599–631

Horowitz JL, Spokoiny GG (2002) An adaptive, rate-optimal test of linearity for median regression models. J. Amer. Statist. Assoc. 97:822–835

Huang JZ, Wu CO, Zhou L (2002) Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika 89:111–128

Inglot T, Ledwina T (1996) Asymptotic optimality of data-driven Neyman's tests for uniformity. Ann. Statist. 24:1982–2019

Ingster Yu. I. (1993) Asymptotic minimax hypothesis testing for nonparametric alternatives I-III. Math. Methods Statist. 2:85–114; 3:171–189; 4:249–268

Jiang J, Hui YV (2004) Spectral density estimation with amplitude modulation and outlier detection. Ann. Inst. Statist. Math. 56:611–630

Jiang J, Li J (2007) Two-stage local M-estimation of additive models. Science in China, Ser. A, to appear

Jiang J, Zhou H, Jiang X, Peng J (2007) Generalized likelihood ratio tests for the structures of semiparametric additive models. Canad. J. Statist. 3 (to appear)

Kallenberg WCM, Ledwina T (1997) Data-Driven smooth tests when the hypothesis is composite. J. Ameri. Statist. Assoc. 92:1094–1104

Kauermann G, Tutz G (1999) On model diagnostics using varying coefficient models. Biometrika 86:119–128

Kooperberg C, Stone CJ, Truong YK (1995a) Rate of convergence for logspline spectral density estimation. J. Time Ser. Anal. 16:389–401

Kooperberg C, Stone CJ, Truong YK (1995b). Logspline estimation of a possibly mixed spectral distribution. J. Time Ser. Anal. 16:359–389

Lepski OV, Spokoiny VG (1999) Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. Bernoulli 5:333–358

Li Q, Huang CJ, Li D, Fu T-T (2002) Semiparametric smooth coefficient models. J. Bus. Econom. Statist. 20:412–422

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Lin DY, Ying Z (2001) Semiparametric and nonparametric regression analysis of longitudinal data (with discussions). J. Amer. Statist. Assoc. 96:103–126

Lin X, Carroll RJ (2001a) Semiparametric regression for clustered data using generalized estimating equations. J. Amer. Statist. Assoc. 96:1045–1056

Lin X, Carroll RJ (2001b) Semiparametric regression for clustered data. Biometrika 88:1179–1865

McCullough P, Nelder JA (1989) Generalized Linear Models, 2nd edn. Chapman and Hall, New York

Mercurio D, Spokoiny V (2004) Statistical inference for time-inhomogeneous volatility models. Ann. Statist. 32:577–602

Murphy SA (1993) Testing for a time dependent coefficient in Cox's regression model. Scand. J. Statist. 20:35–50

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

Opsomer J-D (2000) Asymptotic properties of backfitting estimators. J. Multivar. Anal. 73:166–179

Opsomer J-D, Ruppert D (1998) Fully automated bandwidth selection method for fitting additive models. J. Amer. Statist. Assoc. 93:605–619

Paparoditis E (2000) Spectral density based goodness-of-fit tests in time series models. Scand. J. Statist. 27:143–176

Pawitan Y, O'Sullivan F (1994) Nonparametric spectral density estimation using penalized Whittle likelihood. J. Amer. Statist. Assoc. 89:600–610

Portnoy S (1988) Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. Ann. Statist. 16:356–366

Press H, Tukey JW (1956) Power Spectral Methods of Analysis and Their Application to Problems in Airplane Dynamics. Bell Telephone System Monograph 2606.

Royall RM (1997) Statistical Evidence: A Likelihood Paradigm. Chapman and Hall, London

Shao J, Tu D (1996) The Jackknife and Bootstrap. Springer, New York

Spokoiny VG (1996) Adaptive hypothesis testing using wavelets. Ann. Statist. 24:2477–2498

Stone CJ (1985) Additive regression and other nonparametric models. Ann. Statist. 13:689–705

Vidakovic B (1999) Statistical Modeling by Wavelets. Wiley, New York

Wahba G (1980) Automatic smoothing of the log periodogram. J. Amer. Statist. Assoc. 75:122–132

Wahba G (1990) Spline Models for Observational Data. SIAM, Philadelphia

Wand MP, Jones MC (1995) Kernel Smoothing. Chapman and Hall, London

Zhang CM (2003a) Adaptive tests of regression functions via multi-scale generalized likelihood ratios. Canad. J. Statist. 31:151–171

Zhang CM (2003b) Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. J. Amer. Statist. Assoc. 98:609–628

Zhang W, Lee SY, Song X (2002) Local polynomial fitting in semivarying coefficient models. J. Multivar. Anal. 82:166–188