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

Rejoinder: Nonparametric inference with generalized likelihood ratio tests*

Jianqing Fan · Jiancheng Jiang *

the date of receipt and acceptance should be inserted later

We are very grateful to the Editors, Maria Angeles Gil and Leandro Pardo, for organizing this stimulating discussion. We would like to take this opportunity to thank all discussants for their insightful and constructive comments regarding our paper, opening new avenues for the *GLR* tests. They have made valuable contributions to the understanding of various testing problems.

As stressed in our paper, we reviewed nonparametric inference using the *GLR* tests, laid down some interesting topics, and stressed the importance of the structured alternatives to maintain reasonable power, but we touched only the surface of this exciting field. We are very happy that the discussants responded part of our topics, offered some complementary views and new insights, and raised some interesting problems for further study.

The evolving statistical theory and practice in science and technology leads us to a vast discipline with many challenging statistical problems. The *GLR* test methods have been developed only for limited models based on limited smoothing platforms. We appreciate all efforts of discussnats that widen the scope of our paper.

Before going to technical aspects, let us response to the fundamental question raised by Professors Lafferty and Wasserman: Does hypothesis testing answer the right question? This depends certainly on the context. In the statistical learning or empirical model building that Professors Lafferty and Wasserman alluded to, we are sympathetic with their views. However, in many scientific endeavors, we have hypotheses in mind and we do need to answer the question if the model is supported by the data. For example, in the examples mentioned in section 1.1,

* 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

one may ask if the stock price dynamics follow the geometric Brown motion, if the dynamics contain jumps, if the Black-Formula or other asset pricing formulas are consistent with the market data. These structured modeling problems need nonparametric alternative hypotheses. They are the foundation of modern asset pricing theory and practice.

1 The Wilks phenomenon

The Wilks phenomenon exists in many situations. It is a nice property shared by many likelihood ratio tests and others. However, as pointed out by Professor Bickel, it does not exist in certain situations when the nuisance parameter η is not estimated efficiently. This is an important point for studying Wilks' phenomenon in other situations. We appreciate his general intuition on the Wilks phenomenon and his heuristic calculation for parametric models. We would also like to see more work on the Wilks phenomenon for nonparametric function space such as those mentioned at the end of Section 5.

We appreciate Professor Mammen's remark that the Wilks phenomenon is a great advantage of the *GLR* tests. We agree with him that this nice property is also shared by non- likelihood-ratio tests, such as the Wald tests or other discrepancy measures based tests. This is particularly the case for the simple univariate nonparametric regression problem. However, the issue is not as simple for more general problems. Even for testing problem (1.8) in the article, it is not clear if the discrepancy based tests would have the Wilks phenomenon, as the biases of estimating additive components $m_1(\cdot)$ and $m_2(\cdot)$ can depend on other components. As we admitted in the article, we do not expect the *GLR* tests to be a universal inference tool for all models. We indeed embrace other specific tests tailored for specific problems of interest. Since the idea of *GLR* tests is motivated from the likelihood principle, the *GLR* tests have wide applications and nice interpretations.

2 Power issues of *GLR* tests

The *GLR* tests achieve minimax power rates for detecting various smoothness alternatives, but this is at the price of losing power in a particular direction, as pointed out by Professor Bickel, that omnibus tests like the *GLR* test have smaller power in any particular direction. This is because the power of any goodness of fit test is poor against a local sequence of (contiguous) alternatives (Lehmann and Romano 2005, page 616).

By Lemma 14.3.1 and Theorem 14.6.1 of Lehmann and Romano (2005), any goodness-of-fit test has a preferred set of alternatives for which its power is locally high. This is also demonstrated heuristically in our response to the question raised by Professor Horowitz at the end of this section. Professor Mammen stressed the importance for determining a specific type of alternatives that a given test is appropriate to. However, it is difficult in general to determine such alternatives for a particular test. Naturally one seeks to design tests with high powers against a given finite dimensional set. The Neyman smooth tests, which can be regarded as the *GLR* tests, were constructed in this way.

Bickel, Ritov and Stoker (2006) gave an innovative framework for constructing tests with power at the root-n rate against selected important subclasses of alternatives. As elaborated at the end of this section, this kind of tests need also to pay prices somewhere. For example, they can have smaller power for detecting local deviations such as those in (2.4) below, as demonstrated in Fan (1996). On possibility is to combine, via Bonforroni's adjustment or randomization (at least theoretically), the *GLR* tests that are powerful in detecting the local deviations with more traditional nonparametric tests such as those constructed by Bickel, Ritov and Stoker (2006) that are powerful for global deviations to yield more omnibus tests. As far as rate of convergence is concerned, it will be root-n for each given direction, inheriting from the nonparametric tests in Bickel, Ritov and Stoker (2006), and enjoy rate optimality for detecting more broader class of function spaces, inheriting from the GLR test.

We are grateful to Professor Horowitz's point that the *GLR* tests can also be applied to the following regression model in econometrics:

$$Y_i = g(X_i) + U_i, \tag{2.1}$$

where $g(\cdot)$ is an unknown function and the unobserved U_i may have $E(U_i|X_i) \neq 0$. However, given the observed instrumental variable (IV) W_i, the conditional expectation $E(U_i|W_i) = 0$. He linked further the problem with the inverse problem such as deconvolution problem in classical statistical literature (Carroll and Hall, 1988; Zhang, 1990; Fan, 1991).

Consider the test problem raised by Professor Horowitz

$$H_0 : g(x) = G(x, \theta)$$

for a known function G and finite-dimensional parameters θ. Professor Horowitz pointed out that the test statistic τ_n in Hororwitz (2006) is consistent uniformly over a set of alternative hypotheses whose distance from the null hypothesis is $O(n^{-1/2})$, while the GLR tests would have rate $O(n^{-4/9})$. The class of alternatives that Horowitz (2006) considered is of form

$$G_n(x, \theta) = G(x, \theta_0) + n^{-1/2}\Delta(x). \tag{2.2}$$

This is the global deviation from the null hypothesis, which has a smaller class of alternatives than those targeted by the *GLR* test. Consider, for example, the alternative of form

$$G_n(x, \theta) = G(x, \theta_0) + a_n\Delta(x/a_n^{1/2}), \quad a_n \to 0. \tag{2.3}$$

which has a local deviation around $G(x, \theta_0)$. The difference has a bounded second derivative, but unlike (2.2), the second derivative of the difference in (2.3) may not go to zero. Our heuristic calculation shows that τ_n defined in Horowitz (2006) can detect alternatives with rate $a_n = O(n^{-1/4})$, slower than that of the GLR. In other words, τ_n targets at a smaller class of alternatives, having specific deviations of form (2.2) and is not omnibus among the class of functions having bounded second derivatives.

The heuristic goes as follows. From the definition of $S(z)$ given in Horowitz's discussion, the signal

$$Ea_n\Delta(X/a_n^{1/2})f_{XW}(z, W) = ca_n^{3/2}(1 + o(1)),$$

for some $c > 0$, which is an order smaller than that for the alternative of form (2.2):

$$G_n(x, \theta) = G(x, \theta_0) + b_n \Delta(x), \quad b_n \to 0, \tag{2.4}$$

when $a_n = b_n$. In the latter case, the signal is of order $O(b_n)$ using the same calculation. Now, the stochastic term of τ_n is of order $O_P(b_n n^{-1/2} + n^{-1})$ (cross-product and square term) for the alternative (2.4). Equating the signal with the noise $b_n^2 = O(b_n n^{-1/2} + n^{-1})$, we obtain that $b_n = O(n^{-1/2})$, the same as that obtained by Horowitz (2006). Now, employing the same calculation for the alternative models (2.3), we have $a_n^3 = O(a_n n^{-1/2} + n^{-1})$, yielding $a_n^{-1/4}$. In other words, τ_n can not detect the alternative of form (2.3) faster than $O(n^{-1/4})$. Hence, its minimax rate for detecting the alternatives with a bounded second derivative is no faster than $O(n^{-1/4})$. This argument applies to most of classical test statistics, targeting the alternative of form (2.2).

An excellent point raised by Professor Horowitz is to design a *GLR* test with the Wilks phenomenon and rate $n^{-1/2}$ for alternative of form (2.4). This is beyond the scope of our investigation.

3 Bandwidth Selection

We appreciate the contribution of Professor Müller on the choice of bandwidth for *GLR* tests, an alternative to the multiple-scale test (3.5) of the original article. The resulting test uses an average of the normalized *GLR* statistic (as opposed to not the maxima version) over a range of different bandwidths. As pointed out by Professor Müller that average test statistic may gain power over a large range of alternative hypotheses and be less sensitive to the nature of the alternatives. This is an interesting idea and further study is needed in order to understand the properties of such an averaging test and to compare it with the multiple-scale test.

We are grateful to the comment by Professor Cao on the choice of bandwidth to maximize the bootstrap estimated power of a test. This is a useful alternative to our approach. Our discretization method is only a quick-and-dirty method to implement (3.21), relying on the continuity of $\lambda_n(h)$ as a function of h. Nothing stops us from using a more refined grid.

4 Bias correction

We agree with Professor Mammen on the importance of bias correction. Our method is not to simply smooth on 'parametric residuals', though it indeed is for simple problems such as additive regression models. More generally, our idea is to re-parameterize the original problem so that the null hypothesis becomes testing if the functions of interest are zero. The advantage of this approach is that nonparametric estimates are usually unbiased under the new null hypothesis. The biases come from the estimation of parameters in the original problem, which usually admit the parametric rate. Hence the biases in the re-parameterized problem are significantly reduced to the parametric rate. One advantage of this method is that it is generally applicable when the original null hypothesis is parametric.

We appreciate Professor Efromovich's compliment on our bias correction idea and comment on more details on this topic. The best way to illustrate this via an example. Consider the additive model (4.33) with $D = 2$ and we wish to test

$$H_0 : m_1(x) = \exp(-\theta_1 x), \qquad m_2(x) = \theta_2 + \theta_3 x + \theta_4 x^2. \tag{4.1}$$

The normal-quasi likelihood function is

$$Q(\mathbf{m}, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \{Y_i - \alpha - m_1(X_{i1}) - m_2(X_{i2})\}^2.$$

The GLR statistic would have biases even under the null hypothesis since the nonparametric estimate such as the local linear estimate in RSS_1 has bias of order $O(h^2)$ even under H_0.

The bias correction is to reparametrize the problem. Let $\hat{\theta}$ be the estimate under the null hypothesis. Set

$$m_1^*(x) = m_1(x) - \exp(-\hat{\theta}_1 x), \qquad m_2^*(x) = m_2(x) - (\hat{\theta}_2 + \hat{\theta}_3 x + \hat{\theta}_4 x^2).$$

and translate the problem (4.1) as

$$H_0 : m_1^*(x) = 0, \qquad m_2^*(x) = 0. \tag{4.2}$$

With the reparametrization, the likelihood becomes

$$Q^*(\mathbf{m}^*, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \{Y_i^* - \alpha - m_1^*(X_{i1}) - m_2^*(X_{i2})\}^2,$$

where $Y_i^* = Y_i - \exp(-\hat{\theta}_1 X_{i1}) - (\hat{\theta}_2 + \hat{\theta}_3 X_{i2} + \hat{\theta}_4 X_{i2}^2)$. Let $\hat{m}_1^*$ and $\hat{m}_2^*$ be the nonparametric fit. Then, these nonparametric estimates have biases $O(n^{-1/2}h^2)$ under (4.1), since all parameters are estimated with rate $O(n^{-1/2})$. Applying *GLR* test to (4.2), we have

$$\hat{\sigma}_0^2 = n^{-1} \sum_{i=1}^{n} \{Y_i^* - \bar{Y}^*\}^2$$

and

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \{Y_i^* - \bar{Y}^* - \hat{m}_1^*(X_{i1}) - \hat{m}_2^*(X_{i2})\}^2.$$

Hence, the bias-corrected GLR test statistics is $n \log \hat{\sigma}_0 / \hat{\sigma}$. Note that $\hat{\sigma}$ differs from the one direct application of the GLR. It has smaller bias under the null hypothesis (4.1), as noted above.

5 *GLR* tests with estimating the disturbance distribution

Professors Cao, Efromovich, and Hall raised interesting topics on addressing the *GLR* tests. They would like to see that the error distribution is also infinite dimensional. We appreciate their offering nice ideas to this interesting problem, which are to first estimate the disturbance distribution under the null and alternative models and then to use the estimated disturbance distributions to define the likelihood ratio statistic, so that the true likelihood ratio can be estimated. Now let us call the resulting test the *estimated likelihood ratio test*. Further work along this direction is certainly worthy to develop. We are grateful to Professor Hall's further comments and reference on how to make this work with the stress of oversmoothing on the estimation of the error density. This reminisces Slutsky's theorem in construction of confidence intervals, and at the same time, makes the the distribution of disturbance as close to the parametric one as we can. It remains very interesting to investigate the power gain when the error distribution differs from normal, say.

A closely related idea is the empirical likelihood ratio test in Fan and Zhang (2004). This method compares the empirical likelihoods (Owen 1990, 2001) under the null and alternative models. It can avoid the bias caused by estimating the disturbance distribution if the null model is misspecified. In connection with Professor Hall's comment, it would be very interesting to see if the empirical likelihood ratio test in Fan and Zhang (2004) implicitly over smoothes the error distribution.

6 Robust *GLR* tests

We are grateful to Professors Carroll and Maity for providing insightful evidences about the parametric linear model that the Wilks phenomenon will hold in terms of actually computing p-values using the bootstrap methods designed for the *GLR* tests based on robust M-estimation. We do agree with them that this property can be carried over to nonparametric regression models. An ongoing research project by the second author on using the *GLR* test based on local M-estimation for time-varying coefficient models reveals that the bootstrap methods for robust *GLR* test share the Wilks phenomenon and is more powerful than the local least-square based *GLR* test when the underlying error distribution is deviated from the normal distribution.

We also appreciate their comments about the validity of the Wilks phenomenon. Inspecting our definition of the GLR test, the conditional distribution is given up to finite number of nuisance parameters. In that sense, the likelihood ratio test defined by Schrader and Hattmansperger (1980) corresponds to *GLR* with the error distribution of form $\exp(-\rho(\cdot/\sigma))$ with a known ρ and σ and the Wilks phenomenon continues to hold in the example given by Professors Carroll and Maity.

One interpretation of our last question in Section 5 in the current context is that if the error distribution is of form $\exp(-\rho(\cdot/\sigma))$ with a given $\rho(\cdot)$ and $\sigma = 1$ for the varying coefficient model (4.27) [or additive model (4.33)] of the main article, whether the Wilks phenomenon for the *GLR* test

$$\sum_{i=1}^{n} \rho(Y_i - \mathbf{A}(U_i, \hat{\boldsymbol{\beta}})^T \mathbf{X}_i) - \sum_{i=1}^{n} \rho(Y_i - \hat{\mathbf{A}}(U_i)^T \mathbf{X}_i) \tag{6.1}$$

holds for testing $H_0 : \mathbf{A}(u) = \mathbf{A}(u, \boldsymbol{\beta})$, i.e. whether the asymptotic distribution of the test statistics depends on $\boldsymbol{\beta}$. Similarly, for testing hypothesis (4.29) the main article, whether the *GLR* test statistic is asymptotically independent of the nuisance functions $a_{d+1}(\cdot), \cdots, a_p(\cdot)$. The same type of questions can be asked in the case that σ is unknown and for the additive model (4.33). In the former case, the *GLR* statistic admits slightly different form (the difference in (6.1) is replaced by the logarithm of their ratio).

7 Modelling nonparametric sparsity

The contributions by Professors Lafferty and Wasserman address the issue of fitting the SpAM

$$Y_i = \alpha + \sum_{d=1}^{D} \beta_d m_d(X_{di}) + \epsilon_i,$$

where $||\beta||_1 = \sum_{d=1}^{D} |\beta_d| \leq L$, $E(m_d(X_d)) = 0$, $\quad E(m_d^2(X_d)) = 1$, and $D = D_n$ is increasing with n such that $D_n > n$. This is a useful model selection technique for nonparametric additive model. We are happy to know that the Lasso estimation has good properties even if the model is wrong, and we would like to see this nice work. We also appreciate the conjecture of Lafferty and Wasserman that cross-validation followed by hypothesis testing might reduce risk without treating the model as truth. The problem as formulated by Professors Lafferty and Wasserman is in the spirit of empirical learning to identify significant variables, rather than the hypotheses that certain covariates have impact on the response variables. We believe that the *GLR* tests can provide a useful testing tool for such a nonparametric endeavor, particularly when the models are used in some disciplines where not all variables are treated the same.

8 Others

We agree with Professor Efromovich that the main conclusions of the *GLR* tests discussed in the paper can be extended to wavelet estimators as well as other smoothing platforms. We are also happy to see in Professor Müller's discussion that the *GLR* tests can be applied to determining the dimensionality of a function and even to functional data analysis. All these topics are interesting for further research.

Professor Cao correctly pointed out that the *GLR* tests in Section 4 are mostly derived from homoscedasticity models. Hence, it means to apply to these situations. For example, $\lambda_{n,1}$ is not designed for the logistic model. Using the general version of *GLR* tests, λ_n in (3.2), we arrive at the equation (1) of Professor Cao's comment. For heteroscedastic models, one can use either the empirical likelihood method (Fan and Zhang, 2004) or a modification in Fan and Jiang (2005, Section 7.2).

References

Carroll,R.J. and Hall,P.(1998) Optimal rates of convergence for deconvolving a density. *J. Amer. Statist.Assoc.*, **83** , 1184–1186, 1988.

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

Fan, J. (1991) On the optimal rate of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19**, 1257–1272.

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

Fan, J. and Zhang, J. (2004) Sieve empirical likelihood ratio tests for nonparametric functions. *The Annals of Statistics*, **32**, 1858-1907

Lehmann, E.L. and Romano, J. (2005) Testing Statistical Hypotheses (3rd Edition). Springer, New York

Owen, A. B. (1990) Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90-120

Owen, A. B. (2001) Empirical Likelihood. Chapman & Hall.

Schrader, R.M. and Hettmansperger, T. P. (1980) Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, 67, 93-101.

Zhang, C. H. (1990) Fourier methods for estimating mixing densities and distributions. *Ann. Statist.*, **18**, 806–831.