Test (2007) 16: 445–447
DOI 10.1007/s11749-007-0081-7

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Peter J. Bickel

Published online: 6 November 2007

Abstract This is a very interesting paper reviewing the technique for testing semiparametric hypotheses using GLR tests. I'd like to supplement Fan and Jiang's review with some cautions and a somewhat different point of view.

1 The Wilks phenomenon

Rigorous results for smooth parametric models, see, for example, Bickel and Doksum (2005, Chap. 6) do say that, if $\hat{\theta}, \hat{\eta}$ or, equivalently, $(\hat{\theta}_{\hat{\eta}}, \hat{\eta})$ are MLE's, then $2(\ell(\hat{\theta}, \hat{\eta}) - \ell(\theta_0, \hat{\eta}_{\theta_0})) \Longrightarrow \chi^2_d$, where d is the dimension of Θ. But if $\hat{\eta}$ is not the MLE this result may fail to hold. In particular it will fail if, in the case of θ, η real, $E_{(\theta_0,\eta_0)} \frac{\partial \ell}{\partial \theta}(X_1, \theta_0, \eta_0) \cdot \frac{\partial \ell}{\partial \eta}(X_1, \theta_0, \eta_0) \neq 0$. More generally, if θ and η are infinite dimensional, the requirement is that the tangent spaces at (θ_0, η_0) of the models with $\theta = \theta_0$ kept fixed, $\overset{\circ}{\mathcal{P}}_{\eta}$, and $\eta = \eta_0$ kept fixed, $\overset{\circ}{\mathcal{P}}_{\theta}$, are orthogonal in $\mathrm{L}_2(P_{(\theta_0,\eta_0)})$—see Bickel et al. (1993). All of Fan and Jiang's examples satisfy this condition—appropriately generalized to the general dependent case—see Bickel and Kwon (2001). Murphy's does not but the estimator $(\hat{\theta}, \hat{\eta})$ that she uses is efficient, i.e., behaves like the MLE in nice parametric situations. We give a heuristic argument below why the Wilks phenomenon can only be expected if $\hat{\eta}$ is efficient or the tangent spaces are orthogonal.

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.

P.J. Bickel (✉)
Department of Statistics, University of California at Berkeley, 367 Evans Hall,
Berkeley, CA 94710-3860, USA
e-mail: bickel@stat.berkeley.edu

2 Power issues

A fact that Fan and Jiang allude to is that omnibus tests like the GLR test have very little power in any particular direction. The fact that they achieve minmax power rates for various smoothness classes is little comfort. Bickel et al. (2006), in view of this failure, espouse a different point of view. They show how to construct tests which have power at the $n^{-\frac{1}{2}}$ scale against selected subclasses of alternatives which are viewed as most important, but still achieve consistency against all alternatives—i.e., do not miss extremely strong evidence against the hypothesis which is inconsistent with one's prior views of the important alternatives. Of course, the power received is necessarily very small—see Lehmann and Romano (2005, pp. 617–621) for a nice explanation, and these tests are not minimax although I believe minimax versions of these which also exhibit power in a limited set of directions can be constructed. They do not exhibit the Wilks type of phenomenon but critical values can be set using bootstraps in the way Fan and Jiang describe.

3 Wilks phenomenon heuristic calculation

Here is a heuristic calculation when θ, η are both one-dimensional.

Using

$$\frac{\partial \ell}{\partial \theta}(\hat{\theta}_{\hat{\eta}}, \hat{\eta}) = 0 = \frac{\partial \ell}{\partial \eta}(\theta_0, \hat{\eta}_0),$$

we obtain after some manipulation that, under H,

$$2\Lambda \simeq n\big[(\hat{\theta}_{\hat{\eta}} - \theta_0)^2 I_{11} - I_{22}(\hat{\eta} - \eta_0)^2\big], \tag{1}$$

where $I = \left(\begin{smallmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{smallmatrix}\right)$ is the Fisher information at (θ_0, η_0). Further,

$$n^{-\frac{1}{2}} \frac{\partial \ell}{\partial \theta}(\theta, \hat{\eta}) \simeq I_{12}(\hat{\theta}_{\hat{\eta}} - \theta_0), \tag{2}$$

$$n^{-\frac{1}{2}} \frac{\partial \ell}{\partial \eta}(\theta_0, \eta_0) \simeq I_{22}(\hat{\eta}_0 - \eta_0). \tag{3}$$

Finally,

$$n^{\frac{1}{2}}(\hat{\eta} - \eta_0) \cong n^{\frac{1}{2}}(\hat{\eta}_0 - \eta_0) + \Delta,$$

where Δ is asymptotically $N(0, \tau^2)$ with $\tau^2 = 0$ iff $I_{12} = 0$. After some algebra and formally taking expectations we arrive at

$$E(2\Lambda) \simeq 1 - \frac{I_{12}^2}{I_{11} I_{22}} - \tau^2 \left(\frac{I_{12}^2}{I_{11}} - I_{22} \right). \tag{4}$$

We conclude from (4) that $E(2\Lambda) \simeq 1$ iff either $I_{12} = 0$ or

$$\tau^2 = \frac{I_{12}^2}{I_{11} I_{22}} \left(\div \frac{I_{12}^2 + I_{11} I_{22}}{I_{11}} \right)^{-1} = I^{22} - I_{22}^{-1}, \tag{5}$$

where $\|I^{ij}\| \equiv I^{-1}$. But (5) holds iff $\hat{\eta}$ is efficient, i.e., $(\hat{\theta}_{\hat{\eta}}, \hat{\eta})$ are equivalent to the MLE in regular situations. Note that in general the limit of $E(2\Lambda)$ here is a sum of weighted independent χ^2_1 variables. If Θ is infinite dimensional, failure of the Wilks' phenomenon means that $E(2\Lambda)/(2*)\mathrm{Var}(2\Lambda)$ does not converge to 1.

References

Bickel PJ, Doksum KA (2005) Mathematical statistics: basic ideas and selected topics, vol 1. Prentice Hall, Englewood Cliffs

Bickel PJ, Kwon J (2001) Inference for semiparametric models: some questions and an answer (with discussion). Stat Sin 11:863–960

Bickel PJ, Klaassen C, Ritov Y, Wellner J (1993) Efficient and adaptive estimation in semiparametric models. John Hopkins Press, Baltimore

Bickel PJ, Ritov Y, Stoker T (2006) Tailor-made tests for goodness of fit for semiparametric hypotheses. Ann Stat 34:721–741

Lehmann EL, Romano J (2005) Testing Statistical Hypotheses, 3rd edn. Springer, New York

Test (2007) 16: 448–449
DOI 10.1007/s11749-007-0082-6

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Peter Hall

Published online: 6 November 2007

This is a particularly engaging and motivating paper, which makes a number of innovative proposals. One of the interesting features is the authors' interpretation of "nonparametric" in their work. In the problems to which they devote most attention, the model for the mean is nonparametric under the alternative hypothesis but the model for the error distribution is distinctly finite-dimensional. Of course, this is a reasonable assumption in many settings, but one can change things around and consider hypothesis tests where the model for the error distribution is nonparametric.

It may be a little easier to contemplate the case where the model for the mean is parametric under both null and alternative hypotheses. However, this assumption is not essential. Quite generally, it is possible to address the case where the error distribution is estimated from residuals, becoming gradually richer and more complex as sample size increases, and at the same time to test the range of null and alternative hypotheses that Fan and Jiang consider so carefully.

The idea of estimating the disturbance distribution is encountered in a number of settings. See, for example, works of Linton (1993), who considered adaptive inference in ARCH models when the disturbance distribution is not known, and of Drost and Werker (2004), who suggested using nonparametric density estimation to compute empirical approximations to the score function in the Engle and Russell's (1998) conditional duration model. The secret to getting such an approach to work is to let the number of unknown parameters in the approximation to the disturbance distribution increase very slowly with sample size. For example, if kernel estimation is used, then the bandwidth should converge to zero quite slowly.

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.

P. Hall (✉)
Department of Mathematics and Statistics, Faculty of Science, The University of Melbourne, Victoria 3010, Australia
e-mail: P.Hall@ms.unimelb.edu.au

One might argue that the summation operation, which is employed when computing the log-likelihood ratio statistic, should dampen down at least some of the noise problems associated with relatively small bandwidths. However, this hope turns out to be denied in both theory or practice. On the other hand, using a relatively large bandwidth can give quite good results.

References

Drost FC, Werker BJM (2004) Semiparametric models. J Bus Econ Stat 22:40–50

Engle RF, Russell JR (1998) Autoregressive conditional duration: a new model for irregularly spaced transaction data. Econometrica 66:1127–1162

Linton O (1993) Adaptive estimation in ARCH models. Econ Theory 9:539–569

Test (2007) 16: 450–452
DOI 10.1007/s11749-007-0083-5

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Hans-Georg Müller

Published online: 6 November 2007

I would like to warmly congratulate Professors Fan and Jiang for their stimulating, lucid, and insightful account of the promising concept of generalized likelihood ratio tests, which they nicely demonstrate in a large variety of function estimation contexts. This seminal concept helps to fill the void that still exists regarding generally applicable and well-reasoned tools for inference in function space. It is of great importance to fill this void as in the absence of generally accepted and appealing tools for nonparametric inference, many practitioners will simply stay away from these methods, and therefore their great potential will not be fully realized. In conjunction with the Wilks phenomenon, the development of the generalized likelihood ratio tests has gone a long way towards a general and versatile theory of testing in function spaces. This will be particularly useful in those cases as considered in the examples where one cannot make use of semiparametric efficient approaches. Fan and Jiang cover an amazingly large array of important inference problems for which they demonstrate that the GLR test works.

In the following, I mention some of the thoughts that this very interesting paper generated—none of them may be new or compelling.

Recently, the likelihood ratio approach to testing has been revisited by various authors and alternative tests with better finite sample properties under certain complex alternatives have found renewed interest (Lehmann 2006, and the references cited therein). In these alternative tests, ratios of averages are considered rather than the ratio of maxima. Such alternatives to likelihood ratio tests may prove useful in functional settings. One suggestion that flows from this is to enter smoothing parameters simultaneously over a range of values into the test statistic, rather than re-

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.

H.-G. Müller (✉)
Department of Statistics, University of California, One Shields Ave., Davis, CA 95616, USA
e-mail: mueller@wald.ucdavis.edu

lying on a point estimate or otherwise selected single bandwidth. In this way, one might gain power over a larger range of alternatives, and the test may be less sensitive to the nature of the alternative. In a similar spirit, this suggests an alternative to the multiple-scale test (3.5), for which the maximization would be replaced by a (weighted) integral over a suitable domain of smoothing parameters, e.g.,

$$\int_{h\in[\alpha_n,\beta_n]} \big[\{r\lambda_n - \mu_n(h)\}/\sqrt{2\mu_n(h)}\,\big] w(h)\,dh.$$

One could consider choices such as

$$w(h) = \big[\{r\lambda_n - \mu_n(h)\}/\sqrt{2\mu_n(h)}\,\big]_+^2 \Big/ \int \big[\{r\lambda_n - \mu_n(h)\}/\sqrt{2\mu_n(h)}\,\big]_+^2\,dh,$$

staying close to the maximization idea. Generally, the sensitivity of the test to the bandwidth selection is of interest.

One potentially interesting application of the proposed test would be to determine the dimension of a function, which alternatively can be viewed as nonparametric model selection. Assume that ρ_j, $j = 1, 2, \ldots$, is a suitable orthonormal basis of the function space considered. Then a given function can be represented as $f(t) = \sum_j a_j \rho_j(t)$, and the question arises whether in fact the function belongs to an m-dimensional subspace for some m. One would then test $H_0 : f \in \{\sum_{j=1}^m b_j \rho_j \,|\, b_j \text{ arbitrary}\}$ versus $H_a : f$ cannot be represented with m components. The proposed tests could thus be put to use for stepwise nonparametric model selection, by successively increasing m until H_0 is not rejected, in analogy to forward variable selection in parametric models. This last m would then be the chosen dimension, with the a_j determined by local likelihood or least squares.

There are many potential applications of the GLR tests in Functional Data Analysis (FDA), an area with a clear shortage of available inference tools. Not much effort has gone so far into creating such tools, and current inference in FDA is mostly based on intuitively appealing test statistics whose distributions are obtained via bootstrap. An example is functional regression analysis. In the so-called linear model for the regression of a random response function Y on a random predictor function X, one postulates

$$E\big(Y(t)|X\big) = \mu_Y(t) + \int \big(X(s) - \mu_X(s)\big)\beta(s,t)\,ds,$$

where μ_Y and μ_X are the mean functions of X and Y, and β is the regression parameter function which has two arguments (Ramsay and Dalzell 1991). Under certain regularity assumptions, one can solve a functional least squares equation to obtain

$$\beta(s,t) = \sum_{j,k=1}^{\infty} \theta_{jk}\phi_j(s)\psi_k(t), \quad \theta_{jk} = \frac{E[A_j B_k]}{E[A_j^2]},$$

where (A_j, ϕ_j) and (B_k, ψ_k) are, respectively, functional principal component scores/eigenfunctions of X and Y (Yao et al. 2005).

The following two quite different inference problems arise within this model:

(1) Is the linear functional regression model appropriate? This model imposes a structural constraint on the relationship between Y and X which may or may not be satisfied. We note that this "linear model" only has nonparametric parts and not a single parametric component, indicating this "structural test" is of a different nature than the usual goodness-of-fit tests for parametric models. To address this question, a functional residual process was introduced in Chiou and Müller (2007). The underlying idea is quite similar to the "prewhitening" approach of Fan and Jiang. One can then use the properties of this residual process to devise some seemingly reasonable test statistics.

(2) If we believe the model is appropriate, is $\beta = 0$? Equivalently, does a regression relation exist? This latter question is quite straightforward in classical linear regression models. In the framework of the functional linear model, it is less so. The null hypothesis here can be equivalently expressed as $\theta_{jk} = 0,\ j, k = 1, 2, \ldots$. Again, properties of the residual process can be used, defined under the assumption of no regression relation.

For this second testing problem, the GLR test family could conceivably provide a viable approach. There are challenges: One does not observe directly any data generated by the function β, and the function β is two-dimensional.

References

Chiou J-M, Müller HG (2007) Diagnostics for functional regression via residual processes. Comput Stat Data Anal 51:4849–4863

Lehmann E (2006) On likelihood ratio tests. In 2nd Lehmann symposium. IMS lecture notes-monograph series, vol 49, pp 1–8

Ramsay J, Dalzell CJ (1991) Some tools for functional data analysis. J R Stat Soc Ser B 53:539–572

Yao F, Müller HG, Wang J-L (2005) Functional linear regression analysis for longitudinal data. Ann Stat 33:2873–2903

Test (2007) 16: 453–455
DOI 10.1007/s11749-007-0084-4

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Nonparametric sparsity

John Lafferty · Larry Wasserman

Published online: 6 November 2007

Abstract We discuss the issues raised by Fan and Jiang in the context of high dimensional models and argue that fitting sparse nonparametric models might be preferable to hypothesis testing.

Keywords Additive models · Lasso · Sparsity

1 Introduction

Fan and Jiang have shown that the generalized likelihood ratio statistic (GLR) is a practical method for testing hypotheses in nonparametric models. In particular, due to the Wilks phenomenon, the null distribution can be simulated by the bootstrap. In addition to being intuitively appealing, the method appears to be quite general.

In our discussion we raise the following questions: Does hypothesis testing answer the right question? Do the methods work in high dimensions?

2 Should we use hypothesis testing?

It is not clear that the hypothesis testing framework is answering the right question. Consider the case of additive models, as discussed by Fan and Jiang throughout, and

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.

J. Lafferty (✉)
Computer Science Department and Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA
e-mail: lafferty@cs.cmu.edu

L. Wasserman
Statistics Department and Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA
e-mail: larry@stat.cmu.edu

in detail in Sect. 4.5. We want to compare the model

$$\mathcal{M}_1: \quad Y = \mu_0 + \sum_{j=1}^{D} m_j(X_j) + \epsilon$$

to the model

$$\mathcal{M}_0: \quad Y = \mu_0 + \sum_{j=3}^{D} m_j(X_j) + \epsilon.$$

Fan and Jiang would cast this model comparison in terms of testing the null hypothesis that $m_1 = m_2 = 0$. Realistically, however, neither model can be expected to be correct. Moreover, in high dimensions, where D is large, how do we select which components to remove in the null hypothesis? There are 2^D different additive models of interest, each corresponding to a different subset of components being zeroed out.

What we really want to ask is: which of the exponentially many models yields the best predictor? There is no need to assume that any of the models is correct to answer this question. Framing the problem in this way shifts the thinking from testing to optimization.

Of course, there are some problems where the model is justified and the hypothesis test does answer a scientific question of interest. But these situations are rare. Indeed, the nonparametric methods discussed in the paper are used precisely when we do not want to assume a model. If one is willing to embrace nonparametric methods (and we are certainly sympathetic to this), does it make sense to then take a hypothesis testing approach which takes models as being truth?

If we adopt the more realistic stance that all models are wrong, we are on much safer ground. Predictive accuracy, rather than "finding the true model," then becomes the goal. This brings us to sparsity.

3 Sparsity

Even if we abandon the idea that the model is correct, it still makes sense to consider removing some terms from the additive model

$$Y_i = \alpha + \sum_{d=1}^{D} m_d(X_{di}) + \epsilon_i, \quad i = 1, \ldots, n.$$

In particular, if $D > n$, then we can often get much smaller risk by estimating some components with the zero function.

The approach of Ravikumar et al. (2007) is to cast the problem in an optimization framework called SpAM—sparse additive models. Rewrite the model as

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

where

$$\|\beta\|_1 = \sum_{d=1}^{D} |\beta_d| \leq L$$

and

$$\mathbb{E}\big(m_d(X_d)\big) = 0, \qquad \mathbb{E}\big(m_d^2(X_d)\big) = 1.$$

While this formulation is not convex, the problem can be recast as a convex optimization problem. The results of Ravikumar et al. (2007) show the following. The estimator that minimizes the sum of squares subject to the given constraints (and a smoothness constraint on each m_d) is given by the usual backfitting algorithm with a functional soft-thresholding step inserted at each iteration. For sufficiently small L, the solution is sparse: most $\widehat{m}_d$ are 0. If the model is correct and sparse, then the method is *sparsistent*: the method asymptotically finds the true, nonzero effects. But even if the model is wrong, the estimator satisfies a risk consistency property. Roughly speaking, it finds the best, sparse additive predictor.

The nice thing about SpAM is that it has good properties even if the model is wrong. And the results hold with D_n increasing with n and $D_n > n$. However, there may still be a role for hypothesis testing. If the tuning parameter L is chosen by cross-validation, then the method might overfit. We conjecture that cross-validation followed by hypothesis testing might lead to improved risk. This "screen and clean" approach has been used successfully in Wasserman and Roeder (2007) for linear models; it seems likely to be useful here too. But we emphasize that hypothesis testing can be used strictly to increase sparsity and reduce risk without treating the model as truth.

References

Ravikumar P, Lafferty J, Liu H, Wasserman L (2007) SpAM: sparse additive models. Adv Neural Inf Process Syst (NIPS) 21 (to appear)

Wasserman L, Roeder K (2007) Multistage variable selection: screen and clean. ArXiv: 0704.1139

Test (2007) 16: 456–458
DOI 10.1007/s11749-007-0085-3

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Raymond J. Carroll · Arnab Maity

Published online: 6 November 2007

1 Robust criterion functions and literature

We very much enjoyed reading this stimulating review. Generalized likelihood ratio testing for issues about such basic questions as to whether functions are constant is an appealing idea, and the bootstrap-under-the-null methodology for computing p-values makes these tests simple to implement. The methodology in Sect. 4.1 really is easy to apply and works extremely well.

Naturally enough, this review has focused on the additive model problems, but it is worth pointing out that Fan et al. (2001) also showed the same type of Wilks phenomenon for generalized linear models, and presumably only algebra prevents one from concluding that it holds for all likelihood problems.

At the very end of the paper, the authors ask the natural question about whether the Wilks phenomenon will hold if estimation and inference is based upon a robust criterion function, rather than upon least squares. We conjecture that the answer is no in terms of asymptotics, but yes in terms of actually computing p-values.

This issue has been partially addressed many years ago by Schrader and Hettmansperger (1980) in the parametric context. They considered the parametric linear model which we write here in general as

$$Y_i = m(X_i) + \epsilon_i.$$

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.

R.J. Carroll (✉) · A. Maity
Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843-3143, USA
e-mail: carroll@stat.tamu.edu

A. Maity
e-mail: amaity@stat.tamu.edu

They then proposed to replace the sum of squares criterion

$$\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2$$

by the robust criterion

$$\sum_{i=1}^{n}\rho[\{Y_i - m(X_i)\}/\sigma], \tag{1}$$

which they proposed to use as a likelihood function. They illustrated their work in the case that $\rho(\cdot)$ was the usual Huber function. They then pretended that the actual loglikelihood function was given by (1), replaced σ by a robust estimate $\widehat{\sigma}$, and computed a likelihood-ratio-type statistic, i.e.,

$$\Lambda = 2\sum_{i=1}^{n}\left[\rho\left\{\frac{Y_i - \widehat{m}_{\text{null}}(X_i)}{\widehat{\sigma}}\right\} - \rho\left\{\frac{Y_i - \widehat{m}_{\text{alt}}(X_i)}{\widehat{\sigma}}\right\}\right],$$

where here the subscripts "null" and "alt" refer to estimation under the null and alternative models, respectively.

Let $\psi(x)$ be the first derivative of $\rho(x)$, and let $\psi'(x)$ be the second derivative of $\rho(x)$. They showed that under the null hypothesis,

$$2\frac{E\{\psi'(\epsilon/\sigma)\}}{E\{\psi^2(\epsilon/\sigma)\}}\Lambda \Rightarrow \chi^2_{p-q},$$

where p and q refer to the number of parameters in the alternative and null model. Note that the constant

$$\frac{E\{\psi'(\epsilon/\sigma)\}}{E\{\psi^2(\epsilon/\sigma)\}} \tag{2}$$

depends on nuisance parameters and the distribution function of the errors ϵ, and hence the Wilks phenomenon fails.

2 Summary and conjecture

In summary, what Schrader and Hettmansperger showed is that the Wilks phenomenon does not hold if a robust criterion function is used for estimation and testing. However, they also showed that there was a simple fix, namely to multiply the generalized likelihood ratio statistic by (2). We conjecture that this means that if one applies the bootstrap procedure described in Sect. 4.1 by Fan and Jiang, then the p-values will be asymptotically correct anyway. The reason for this is that constants do not matter in the bootstrap procedure. Of course, all these calculations are in the homoscedastic case, and even if the conjecture is true, handling heteroscedastic cases may well have a different story.

Acknowledgements Our research was supported by grants from the National Cancer Institute (CA57030, CA104620), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

References

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

Schrader RM, Hettmansperger TP (1980) Robust analysis of variance based upon a likelihood ratio criterion. Biometrika 67:93–101

Test (2007) 16: 459–461
DOI 10.1007/s11749-007-0086-2

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Joel L. Horowitz

Published online: 6 November 2007

Fan and Jiang have provided a lucid exposition of a large and widely applicable class of generalized likelihood ratio (GLR) tests of parametric or nonparametric models against nonparametric alternatives. The tests have asymptotic distributions that are free of nuisance parameters and achieve the optimal rate of testing in some important cases. These are useful properties and make GLR tests attractive in many situations that arise frequently in applications. There are, however, settings in which the optimality of the GLR test is not clear. This discussion gives two examples that are important in econometrics and in which tests that are not asymptotically pivotal achieve a faster rate of testing than at least the most obvious versions of the Fan–Zhang approach. It would be useful to know whether there are less obvious versions that achieve a fast rate of testing and are asymptotically pivotal.

Suppose that data $\{Y_i, X_i : i = 1, \ldots, n\}$ on random variables (Y, X) are generated by the model

$$Y_i = g(X_i) + U_i, \tag{1}$$

where g is an unknown function and U_i is an unobserved random variable. The U_i's may be correlated with the X_i's and, in particular, $E(U_i|X_i)$ may not vanish. This situation arises frequently in economics. For example, suppose that Y_i is the wage of individual i and X_i is that individual's level of education. The U_i's typically include personal characteristics (called "ability") that influence the wage but are unobserved by the analyst. If high-ability individuals choose to obtain high levels of education, then $E(U_i|X_i)$ is an unknown function of X_i, and (1) does not identify g.

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.

J.L. Horowitz (✉)
Department of Economics, Northwestern University, Evanston, IL 60208-2600, USA
e-mail: joel-horowitz@northwestern.edu

One way of dealing with this problem is to assume the availability of data on a third variable, W, called an instrument, with the property that

$$E(U_i|W_i = w) = 0 \tag{2}$$

for each $i = 1, \ldots, n$ and all w in the support of W. Equations (1) and (2) together imply that

$$E(Y|W = w) = \int g(x) f_{X|W}(x|W = w)\, dx, \tag{3}$$

where $f_{X|W}$ is the density of X conditional on W. Under suitable conditions, (3) has a unique solution that identifies g. Methods for using (3) to form an estimator of g have been developed by Blundell et al. (2007), Darolles et al. (2006), Hall and Horowitz (2005), and Newey and Powell (2003). However, (3) is a Fredholm integral equation of the first kind and generates an ill-posed inverse problem (e.g., Kress 1999). Consequently, the fastest possible rate of convergence of an estimator of g is typically slow and may be logarithmic. In contrast, if g were known up to a finite-dimensional parameter, θ, then a $n^{-1/2}$-consistent, asymptotically normal estimator could be obtained by using the generalized method of moments (Hansen 1982).

Now consider testing the hypothesis $H_0 : g(x) = G(x, \theta)$ for some known function G and finite-dimensional parameter θ. It follows from (3) that this is equivalent to testing the hypothesis

$$E(Y|W = w) = \int G(x, \theta) f_{X|W}(x|W = w)\, dx \tag{4}$$

for some θ. To avoid technical complications that are not important for the argument here, assume that the density of (X, W) is known. Then the hypothesis (4) can be tested using the GLR method described by Fan and Jiang. The test statistic has an asymptotic distribution that is free of nuisance parameters and, under reasonable assumptions, the test has nontrivial power against alternatives whose distance from the null hypothesis is $O(n^{-4/9})$.

Now let f_{XW} denote the probability density function of (X, W) and define

$$S(z) = E\big\{\big[Y - G(X, \theta)\big] f_{XW}(z, W)\big\}.$$

It follows from (1) and (2) that under H_0, $S(z) = 0$ for every z in the support of X. Moreover, $S(z) \neq 0$ on a set of positive measure if H_0 is false. Therefore, a statistic for testing H_0 can be based on a sample analog of, say,

$$\int S(z)^2\, dz.$$

In particular, one can use the test statistic

$$\tau_n = \int S_n(z)^2\, dz,$$

where

$$S_n = n^{-1/2} \sum_{i=1}^{n} \big[Y_i - G(X_i, \hat{\theta}) \big] f_{XW}(z, W_i), \tag{5}$$

and $\hat{\theta}$ is, say, a generalized method of moments estimator of θ. If f_{XW} is unknown, as is usually the case in applications, then it can be replaced in (5) by a nonparametric estimator without changing the asymptotic distribution of τ_n under H_0 or alternative hypotheses. Horowitz (2006) shows that under H_0, τ_n is asymptotically distributed as a weighted sum of independent χ^2 variables with one degree of freedom. The weights depend on unknown population parameters, so the Wilks phenomenon does not hold. However, the τ_n test is consistent uniformly over a set of alternative hypotheses whose distance from the null hypothesis is $O(n^{-1/2})$. Thus, the asymptotically nonpivotal τ_n test has a faster uniform rate of testing than does the GLR test based on (4).

One can also consider testing the hypothesis $H_{0E} : E(U|X) = 0$ in (1) and (2). In econometrics, this is called an exogeneity test. Blundell and Horowitz (2007) show that this hypothesis can be tested using the statistic

$$\tau_{nE} = \int S_{nE}(z)^2 \, dz,$$

where

$$S_{nE}(z) = n^{-1/2} \sum_{i=1}^{n} \big[Y_i - \hat{G}(X_i) \big] \hat{f}_{XW}(z, W_i),$$

$\hat{G}(X_i)$ is a nonparametric estimator of $E(Y|X = X_i)$, and $\hat{f}_{XW}$ is a nonparametric estimator of the density of (X, W). As in the case of τ_n, τ_{nE} is not asymptotically pivotal, but it is consistent uniformly over a class of alternatives whose distance from H_0 is $O(n^{-1/2})$.

These results raise the question of whether there are GLR-type tests of H_0 and H_{0E} that have asymptotically pivotal statistics and uniform $n^{-1/2}$ rates of testing. Such tests, if they can be developed, would be very useful.

References

Blundell R, Horowitz JL (2007) A nonparametric test of exogeneity. Rev Econ Stud 74:1035–1058

Blundell R, Chen X, Kristensen D (2007) Semi-nonparametric IV estimation of shape-invariant Engel curves. Econometrica (forthcoming)

Darolles S, Florens J-P, Renault E (2006) Nonparametric instrumental regression. Working paper, GREMAQ, University of Social Science, Toulouse

Hall P, Horowitz JL (2005) Nonparametric estimation in the presence of instrumental variables. Ann Stat 6:2904–2929

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

Horowitz JL (2006) Testing a parametric model against a nonparametric alternative with identification through instrumental variables. Econometrica 74:521–538

Kress R (1999) Linear integral equations, 2nd edn. Springer, New York

Newey WK, Powell JL (2003) Instrumental variable estimation in nonparametric models. Econometrica 67:565–603

Test (2007) 16: 462–464
DOI 10.1007/s11749-007-0087-1

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Enno Mammen

Published online: 6 November 2007

Abstract In our discussion we make remarks on the range of the validity of Wilks phenomenon and address the question if one should always try to reach Wilks phenomenon. We discuss the need of bias corrections in test statistics, make some power considerations, and mention some open problems.

Keywords Nonparametric tests · Nonparametric curve estimation

Mathematics Subject Classification (2000) 62G10 · 62G08

1 Introduction

The article by Jianqing Fan and Jiancheng Jiang presents a nice introduction and overview on generalized likelihood ratio tests based on nonparametric curve estimators. By now, there is a huge amount on papers that use the comparison of nonparametric curve estimators with parametric, semiparametric, or more restricted nonparametric estimators as a starting point for a goodness-of-fit test. The article by Jianqing Fan and Jiancheng Jiang develops a link of these approaches to classical likelihood ratio tests and embed the overwhelming research into a promising framework. Our remarks are concerned with the range of the validity of Wilks phenomenon, the importance of Wilks phenomenon, bias corrections, power properties, and some open problems.

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.

E. Mammen (✉)
Department of Economics, University of Mannheim, L7, 3–5, 68131 Mannheim, Germany
e-mail: emammen@rumms.uni-mannheim.de

2 Comments

Wilks phenomenon and generalized likelihood ratio tests It is a great advantage of the generalized likelihood ratio tests that they fulfill the property of Wilks phenomenon, i.e., that their asymptotic null distribution does not depend on unknown parameters. But this property is not only valid for this class of tests. This property is shared by many tests that are not motivated by the likelihood principle. In the nonparametric regression model

$$Y_i = m(X_i) + \varepsilon_i \tag{1}$$

with homoscedastic mean zero errors ε_i, we get the Wilks phenomenon for a large class of tests T that are based on the difference of two estimators $\widehat{m}$ and $\widetilde{m}$ of m. The Wilks phenomenon holds as long as the bias terms in the difference of the estimators are asymptotically negligible for the asymptotic behavior of the test statistic. Besides generalized likelihood ratio tests, one can use test statistics that are asymptotically equivalent to weighted L_p norms of the difference $\widehat{m} - \widetilde{m}$, e.g., with $p = 1$, $p = 2$, or $p = \infty$. Other tests are based on norms of the difference of partial integrals of the estimators $\int^x [\widehat{m}(u) - \widetilde{m}(u)]\, du$. All these choices have been considered in the literature.

Is the Wilks phenomenon a must? In general, nonparametric models are not fully specified. Consider, e.g., model (1) for i.i.d. tuples (X_i, Y_i) without assuming that the conditional variance $\sigma^2(x) = E(\varepsilon_i^2 | X_i = x)$ is constant. There are at least two ways to proceed for this model. In the first approach one interprets $\sigma^2(x)$ as an additional nonparametric parameter of the model and uses the L_2 norm (or another measure) of $\widehat{\sigma}^{-1}(x)[\widehat{m}(x) - \widetilde{m}(x)]$ as test statistic, where $\widehat{m}$ and $\widetilde{m}$ are two fits of m and where $\widehat{\sigma}^2(x)$ is an estimator of $\sigma^2(x)$. The test statistic is asymptotically distribution free under some smoothness assumptions on the variance function σ^2. In the second approach one uses the L_2 norm (or another measure) of the unweighted difference $\widehat{m}(x) - \widetilde{m}(x)$ and uses, e.g., wild bootstrap or an estimate of the average error variance to get an asymptotically correct level. This approach works without smoothness conditions (or at least weaker conditions) on $\sigma^2(x)$, see Härdle and Mammen (1993). Both approaches make sense. The first one follows the motivation of the paper and may have better power properties. The second one needs less detailed model assumptions. This discussion carries over to more complicated models. Important examples are given by models with dependent errors where one is not willing to use more specific assumptions on the error structure. Then one may prefer tests that, as in the second approach, are not motivated by a likelihood principle.

How to correct for bias terms? An important issue in testing are bias corrections of nonparametric estimators. This is nicely discussed in the paper. I think that there exists no principal idea that always work. Some different ways of doing this have been proposed (that in the simple model (1) partially lead to the same procedure.) The smoothing of parametric residuals that has been described in this paper does not work for more complicated models. For such models, estimation of bias terms has been proposed. Another idea uses smoothing of degenerate data coming from parametric

fits without noise and it applies the resulting bias correction for test statistics that are motivated by approximations of likelihoods, see, e.g., Härdle et al. (2004) and Härdle et al. (1998), where testing in generalized partial linear and additive models is discussed. I think some further ideas are needed when coming to more complicated models.

Power properties The generalized likelihood ratio tests are not the only approaches that are motivated by asymptotic power considerations. There are many optimality criteria leading to different approaches. Thus it is important for which specific type of alternatives a given test is especially appropriate. Typically test statistics have an intuitive interpretation as distances from the hypothesis. It is important to check if the test statistic is really sensitive for deviations from the null hypothesis that the test statistic seems to measure. One illustrative example: if in model (1) one uses $T = \int [\widehat{m}(x) - \widetilde{m}(x)]^2 \, dx$ as a test statistic, where $\widehat{m}$ is a nonparametric kernel smoothing estimator and $\widetilde{m}$ is a parametric estimator, then this test is asymptotically equivalent to a purely parametric test that only checks for deviations in one parametric directions, see Härdle and Mammen (1993). Thus, in this case the just mentioned intuition is totally misleading.

Complications and future work I would like to mention some directions that need some further work. In this overview article mostly testing problems have been considered where parametric (or semiparametric) models have been tested against nonparametric alternatives. One can imagine much more involved testing problems. E.g., complications arise when nonparametric models are tested against more general nonparametric models. The complications may come from the fact that different degrees of complexity apply in the null and in the alternative function model. An example is given in Mammen and Sperlich (2007), where additive models are tested against models with interaction terms. There modifications of likelihood criteria must be used to get the test to work. Other examples are models that are not of the simple form: observation = nonparametric signal plus noise, e.g., nonparametric models that lead to inverse problems. A class of problems where is no hope anymore for the Wilks phenomenon are testing problems for testing shape restrictions. This is related to the fact that for parametric hypothesis given by inequalities on parameters no Wilks phenomenon holds. The distribution of the likelihood test statistic depends on the location on the null hypothesis, and its asymptotic limit is different from a χ^2-distribution.

References

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

Härdle W, Mammen E, Müller M (1998) Testing parametric versus semiparametric modelling in generalized linear models. J Am Stat Assoc 93:1461–1474

Härdle W, Huet S, Mammen E, Sperlich S (2004) Bootstrap inference in semiparametric generalized additive models. Econom Theory 20:265–300

Mammen E, Sperlich S (2007) Additivity tests based on smooth backfitting. Preprint

Test (2007) 16: 465–467
DOI 10.1007/s11749-007-0088-0

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Sam Efromovich

Published online: 6 November 2007

This is a wonderful paper with a wealth of information. I heartily thank Professors Fan and Jiang for preparing this manuscript, which, I believe, will provide inspiration in years to come.

Nonparametric curve estimation is in the center of the considered issues. I would like to begin with a brief presentation of several recent results in this area that will be instrumental in the discussion.

First, consider the heteroscedastic nonparametric regression model

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \quad i = 1, 2, \ldots, n. \tag{1}$$

Here the predictor X may be random or deterministic, the regression error ϵ and the predictor may be dependent or independent. Then, under several mild assumptions, with the main one being differentiability of the regression function $m(x)$, the scale (volatility) function $\sigma(x)$, and the design density $p(x)$ of the predictor, it is established in Efromovich (2005) that the density of regression errors can be estimated by a statistician as well as by an oracle which knows the sample of regression errors $\epsilon_1, \ldots, \epsilon_n$. Note that, in the case of dependent predictor and regression error, the marginal density of the error is estimated.

Second, for a sample of iid pairs of (Y, X), where again Y is the response and X is the predictor, there exists a data-driven estimator of the conditional density $f^{Y|X}(y|x)$ which is minimax when Y and X are dependent and independent. In the latter case

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.
The work was supported by NSF Grants DMS-0243606, DMS-0604558 and NSA Grant MDA904-07-1-0075.

S. Efromovich (✉)
Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75083-0688, USA
e-mail: efrom@utdallas.edu

the estimator automatically adjusts to the necessity of estimation of the univariate marginal density $f^Y(y) = f^{Y|X}(y|x)$ and thus performs dimension reduction. See details in Efromovich (2007).

Finally, a wide class of adaptive wavelet estimators for different statistical models, including regression and spectral density estimation, has been developed recently. In particular, multiwavelet estimators combine excellent properties of estimation of underlying functions and their derivatives with robustness toward different distributions of observations. See a discussion in Efromovich et al. (2004).

Now I am in a position to begin the discussion.

1. In considered regression settings, the null hypothesis is a classical parametric regression versus a nonparametric one. In Sect. 1.1, the authors present a smart example which shows why it is crucial to consider an alternative hypothesis which fits data at hand. But what will be if even a heteroscedastic nonparametric regression model (1) does not fit a dataset? The interested reader can find such an example in Efromovich (2007) which is a real practical dataset from civil engineering. In this case a conditional density approach may be a feasible alternative for developing GLR tests. Further, in general the analysis of an estimated conditional density may be instrumental in finding an appropriate alternative hypothesis which may be a homoscedastic, or heteroscedastic, or mixture, or any other type of regression.

2. In a regression setting, the assumption of normality of regression errors is a classical one, it implies nice formulas for likelihood ratio tests, it is easy to analyze, and, under a mild assumption, Wilk's phenomenon is robust toward deviations from the assumed normality. At the same time, it is clear that in many practical applications it can be beneficial to know an underlying distribution of the regression error. It is an interesting research topic to check feasibility of employing an estimated density of regression errors in GLR tests.

I also believe that using an estimated distribution in GLR tests will allow one to solve the raised, in Sect. 5, questions about a robust version of the tests in the presence of outliers.

3. Another interesting idea is to use estimates of the regression error density and/or conditional density in place of suggested bootstrapping procedures.

4. The authors primarily discuss estimation of Hölder functions. If a larger class of Besov functions is included in alternative hypotheses, then wavelet/multiwavelet nonparametric estimators can be used. It is reasonable to conjecture that all main conclusions of the discussed paper, made for classical kernel and spline estimators, will hold for wavelet estimators as well. Further, using multiwavelets may allow one to consider a wider set of alternative hypotheses including ones based on derivatives of underlying functions (or motivated by more general indirect problems).

Finally, I would like to note that a more detailed discussion, in the authors' rejoinder, of the bias-reduction approach will be appreciated by many readers. It is an excellent idea and a very useful tool in simplifying/understanding/justification statistical inferences, and it deserves to be explained in more detail.

References

Efromovich S (2005) Estimation of the density of regression errors. Ann Stat 33:2194–2227
Efromovich S (2007) Conditional density estimation in a regression setting. Ann Stat 35 (in press)
Efromovich S, Lakey J, Pereyra MC, Tymes N (2004) Data-driven and optimal denoising of a signal and recovery of its derivative using multiwavelets. IEEE Trans Signal Process 52:628–635

Test (2007) 16: 468–470
DOI 10.1007/s11749-007-0089-z

DISCUSSION

Comments on: Nonparametric inference with generalized likelihood ratio tests

Ricardo Cao

Published online: 6 November 2007

This paper gives a very nice survey on generalized likelihood ratio (GLR) statistics for model checking. The validity of the Wilks phenomenon is illustrated through a collection of problems that can be dealt with this general method.

The main idea behind the GLR tests is to compute the likelihood ratio for the null and alternative hypotheses, under certain assumptions on the data generating process (for instance, the error normality in regression models). Estimation of nuisance parameters (as the error variance) can be performed via maximum likelihood. The asymptotic null distribution of the GLR test that does not depend on populational quantities (the Wilks phenomenon), can be derived for a number of interesting setups. Under this general approach, one or both of the null and alternative models can be fully parametric, semiparametric, or nonparametric.

In the first two sections of the paper, some mention to empirical likelihood is missing. The nice book by Owen (2001) is a useful reference for the reader to have a complementary view about extending classical likelihood methods to non- and semiparametric contexts.

A crucial aspect in the derivation of closed formulas for GLR tests is the normality and homoscedasticity for the errors. As Fan and Jiang explain, the normal distribution for the errors is just an assumption to derive explicit expressions for the GLR tests, but the asymptotic null distribution of the tests can be often derived without such an assumption. However, as they recognize in Sect. 4.3, when deriving a GLR test for the spectral density, misspecification of the error distribution may imply an important loss in the power of the test.

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11749-007-0080-8.

R. Cao (✉)
Department of Mathematics, Universidade da Coruña, 15071 A Coruña, Spain
e-mail: rcao@udc.es

In fact, if the error distribution is assumed to be normal, the bootstrap method in Sect. 4.1 should be modified in order to incorporate this assumption in Step (3) of the resampling plan. Consequently, the $\hat{\varepsilon}_i^*$ should be drawn from the normal distribution with zero mean and variance given by the sample variance of the residuals.

If the ε_i are not assumed to be normal, but coming from a general common density, say g, an alternative equation to (4.2) can be easily computed:

$$\ell(\hat{m}_h) = \sum_{i=1}^{n} \log\big(g\big(Y_i - \hat{m}_h(X_i)\big)\big),$$

but, of course, this depends on the unknown density g. A feasible approach within such a nonparametric framework is to compute a kernel density estimator for g,

$$\hat{g}_{\hat{b}}(t) = \frac{1}{n\hat{b}} \sum_{i=1}^{n} K\left(\frac{t - \hat{\varepsilon}_i}{\hat{b}}\right),$$

and use it in the log-likelihood function

$$\tilde{\ell}(\hat{m}_h) = \sum_{i=1}^{n} \log\big(\hat{g}_{\hat{b}}\big(Y_i - \hat{m}_h(X_i)\big)\big),$$

where $\hat{b}$ is a suitable bandwidth selector for the sample of the residuals $\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i)$, $i = 1, 2, \ldots, n$. Of course, there is no closed formula for this nonparametric likelihood, but it can be maximized by means of numerical algorithms. This approach has been previously used by Hsieh and Manski (1987) in a regression context and Cao et al. (2003) in a time series context.

In regression-like models, as those studied in Sects. 4.1–4.5, it does not seem reasonable to make inferences about a parametric form for the conditional mean ($m(x) = E(Y|X = x)$), having assumed a very restrictive parametric form (a constant function) for the conditional variance ($\sigma^2(x) = \mathrm{Var}(Y|X = x)$).

An interesting particular case where this homoscedasticity assumption is not true is the binary response regression setup. In this case, Y takes the values 0 and 1, and $m(x) = E(Y|X = x) = P(Y = 1|X = x)$. In this framework, $\sigma^2(x) = \mathrm{Var}(Y|X = x) = m(x)(1 - m(x))$, and the conditional log-likelihood function is given by

$$\ell(m) = \sum_{i=1}^{n} \big[Y_i \log\big(m(X_i)\big) + (1 - Y_i) \log\big(1 - m(X_i)\big)\big].$$

If a parametric model is considered under the null hypothesis H_0: $m \in \mathcal{M}_\Theta = \{m_\theta(\cdot)/\theta \in \Theta\}$, the logarithm of the likelihood ratio test is

$$\ell(\hat{m}_h) - \ell(m_{\hat{\theta}}) = \sum_{i=1}^{n} \left[Y_i \log\left(\frac{\hat{m}_h(X_i)}{m_{\hat{\theta}}(X_i)}\right) + (1 - Y_i) \log\left(\frac{1 - \hat{m}_h(X_i)}{1 - m_{\hat{\theta}}(X_i)}\right)\right]. \quad (1)$$

Although the maximization of $\ell(m_\theta)$ in θ does not lead to a closed formula, it is clear that the normality and homoscedasticity assumption leading to $\lambda_{n,1}$ in (4.3) are not

realistic in this case. The behavior of the test $\lambda_{n,1}$ in (4.3) will be presumably poorer than that of (1).

The issue of optimal bandwidths for testing (mentioned in Sect. 3.3) deserves some attention on its own. The choice $h = h_0 1.5^j$ needs not to be close to optimal, since optimal bandwidths for estimation may be of different rate than optimal bandwidths for testing (whatever this means). A reasonable (although time consuming) approach for bandwidth selection in testing setups can be found in Remark 1 in Cao and Van Keilegom (2006). The basic idea is to select the bandwidth by maximizing a bootstrap estimation of the power.

In summary, in my view, GLR tests are very competitive procedures that share nice properties of classical LR tests (as the Wilks phenomenon). These methods can be improved by incorporating further nonparametric flexibility in the error structure, giving rise to adaptive LR tests that are closely related to empirical likelihood methods.

References

Cao R, Van Keilegom I (2006) Empirical likelihood tests for two-sample problems via nonparametric density estimation. Can J Stat 34:61–77

Cao R, Hart J, Saavedra A (2003) Nonparametric maximum likelihood estimators for AR and MA time series. J Stat Comput Simul 73:347–360

Hsieh DA, Manski CF (1987) Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. Ann Stat 15:541–551

Owen AB (2001) Empirical likelihood. Chapman & Hall/CRC, Boca Raton