Discussion of the paper "Inference for Semiparametric Models: Some Questions and an Answer" by Bickel and Kwon *

Jianqing Fan
Department of Statistics
Chinese University of Hong Kong
AND Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

June 19, 2001

Bickel and Kwon are to be congratulated for this neat, insightful and stimulating paper on the general theory of semiparametric efficiency and for their successfully posing several important and challenge questions on semiparametric inferences. Semiparametric parametric models arise frequently in many applications. The interest in estimating certain principal parameters while imposing few assumptions on nuisance parameters gives rise to semiparametric models. The parameters of interest usually admit the similar interpretations to those in parametric models. Most of work focuses on efficient inferences on parameters of interest when semiparametric models are correctly specified. The question arises naturally how to validate whether a semiparametric model fits a given set of data, as asked by Bickel and Kwon. I welcome the opportunity to make a few comments and to provide additional insights.

1 Generalized likelihood ratio test

One of the most celebrated methods in parametric inferences is the maximum likelihood ratio test. It is intuitive and easily applicable due to the Wilks type of results. An effort of extending the scope of the likelihood ratio tests is the empirical likelihood (Owen 1988) and its various extensions. Yet, they can not be directly applied to hypothesis testing problems in multivariate semiparametric and nonparametric models.

In an effort to derive a generally applicable testing procedure for multivariate nonparametric models, Fan *etal.* (2001) proposed a generalized likelihood ratio test. The work is motivated by the fact that the nonparametric maximum likelihood ratio test may not exist. Further, even if it exists, it is not optimal even in the simplest nonparametric regression setting (see Fan *etal.*

Fan's research was partially supported by NSF grant DMS-0196041.

2001). Generalized likelihood ratio statistics, obtained by replacing unknown functions by reasonable nonparametric estimators rather than MLE as in parametric models, enjoy several nice properties to be outlined below.

As an illustration, consider the varying-coefficient model

$$Y = a_1(U)X_1 + \cdots + a_p(U)X_p + \varepsilon, \tag{1}$$

where Y is the response variable, $(U, X_1, \cdots, X_p)$ is the covariate vector independent of the random noise ε. Consider the problem of testing homogeneity

$$H_0 : a_1(\cdot) = \theta_1, \cdots, a_p(\cdot) = \theta_p. \tag{2}$$

For simplicity, assume further $\varepsilon \sim N(0, \sigma^2)$ [As demonstrated in Fan *etal.* (2001), the normality assumption is only used to motivate the procedure]. Given a random sample of size n, the likelihood under the null hypothesis can easily be obtained with parameters $\{\theta_j\}$ replaced by their MLE. Let $\ell_n(H_0)$ denote the log-likelihood under the null model. Under the more general model (1), the coefficient functions $a_1(\cdot), \cdots, a_p(\cdot)$ can easily be estimated by using, for example, a kernel method or local linear regression (Carroll *etal.*, 1998, Hoover *etal.*, 1998, Fan and Zhang, 1999). Using these estimated functions, one can easily form the likelihood under the general model (1), though it does not maximize the nonparametric likelihood. Let $\ell_n(H_1, h)$ denote the log-likelihood, where h is the bandwidth used in the local linear regression estimate of functions $a_1(\cdot), \cdots, a_p(\cdot)$. Then, the generalized likelihood ratio statistic is simply

$$T_n(h) = \ell(H_1, h) - \ell(H_0). \tag{3}$$

This generalized likelihood ratio test admits the same intuitive interpretation as the classical likelihood ratio test.

Fan *etal.* (2001) unveil the following Wilks phenomenon: The asymptotic null distribution of $T_n(h)$ is independent of nuisance parameters in the model under the null hypothesis and follows a χ^2-distribution (in a generalized sense) for testing homogeneity (2) versus (1). Thus, the P-values can easily be computed by either using the asymptotic distribution or simulations with parameter values taken to be the MLE under the null hypothesis. Further, they show that the resulting tests are asymptotically optimal in the sense of Ingster (1993).

The above Wilks phenomenon holds not only for testing parametric versus nonparametric hypotheses, but also for testing a nonparametric null hypothesis versus a nonparametric alternative hypothesis. As an example, Fan *etal*(2001) consider the problem of testing significance of variables

$$H_0 : a_1(\cdot) = a_2(\cdot) = \cdots = a_m(\cdot) = 0 (m \leq p).$$

The null hypothesis is still nonparametric because it involves nuisance functions $a_{m+1}(\cdot), \cdots, a_p(\cdot)$. Nevertheless, they show that the Wilks type of result continues to hold: the asymptotic null distribution is independent of these nuisance functions. Thus, the P-values can easily be computed by either using the asymptotic distributions or using simulations via fixing nuisance functions under the null hypothesis at their estimated values. These results were also extended to various other models.

The idea of the above generalized likelihood ratio method is widely applicable. It is easy to use because of the Wilks phenomenon and is powerful as it achieves the optimal rates for hypothesis testing. This encourages me to propose the generalized likelihood ratio test as a possible tool to the open question (D) posed by Bickel and Kwon.

2 Validating semiparametric models

To fix the idea, consider the test against the partially linear model

$$H_0 : Y = g(U) + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon. \tag{4}$$

Again, for simplicity, we assume that $\varepsilon \sim N(0, \sigma^2)$. Let $\hat{g}$ and $\hat{\beta}_1, \cdots, \hat{\beta}_p$ be the estimates based on a sample of size n, using for example the profile likelihood approach (see e.g. Speckman, 1988, Severini and Wong, 1992, and Carroll *etal.* , 1997). The profile likelihood gives semiparametric efficient estimator for parameters $\beta_1, \cdots, \beta_p$ and an optimal estimator for function g. With this, one can form the log-likelihood function under the null hypothesis, denoted by $\ell_n(H_0, h)$, where h is the bandwidth.

To test whether this model holds for a given data set, we need an alternative. Depending on the degree of prior belief on the model, one may consider the following possible alternative models:

1. **An additive model**:

$$H_{11} : Y = f_0(U) + f_1(X_1) + \cdots + f_p(X_p) + \varepsilon.$$

2. **A varying-coefficient model**:

$$H_{12} : Y = f_0(U) + f_1(U)X_1 + \cdots + f_p(U)X_p + \varepsilon.$$

3. **A full nonparametric model**:

$$H_{13} : Y = f(U, X_1, \cdots, X_p) + \varepsilon.$$

The unknown nonparametric functions in the above models can easily be estimated, using for example kernel and local linear estimators with bandwidth h (for additive model, one can use the backfitting algorithm as in Hastie and Tibshirani, 1990). Using the estimated nonparametric functions, one can form the nonparametric log-likelihood $\ell_n(H_{1j}, h)$ $(j = 1, 2, 3)$ as in Section 1 and the generalized likelihood ratio statistics:

$$T_{n,j}(h) = \ell_n(H_{1,j}, h) - \ell_n(H_0, h), \quad j = 1, 2, 3.$$

These form the generalized likelihood ratio test statistics for testing the semiparametric model (4) against the three nonparametric alternative models.

A few questions arise naturally. First of all, are the asymptotic null distributions for the test statistics independent of nuisance parameters in the null hypothesis? Secondly, do these test statistics achieve the optimal rates for hypothesis testing in the sense of Ingster (1993) and Spokoiny (1996)? Thirdly, what are the optimal rates for these three different alternatives?

In the additive model, Stone (1986) shows that one can estimate each additive component as well as the one-dimensional rate. Fan *etal.* (1998) strengthens the result further that one can estimate each additive component as well as if other components were known. The question then arises naturally if these kinds of results hold for hypothesis testing against the semiparametric model with the additive model as the alternative hypothesis.

3 Tests within semiparametric models

Suppose that we have validated a semiparametric model. Various inference problems arise within the semiparametric model. For example, under the partially linear model, one may wish to test if certain variables are statistically significant such as

$$H_0 : \beta_1 = \cdots = \beta_m = 0.$$

More generally, one may consider the linear hypothesis:

$$H_0 : \mathbf{A}\boldsymbol{\beta} = 0, \tag{5}$$

where $\mathbf{A}$ is a given matrix and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$. This is a semiparametric null hypothesis versus a semiparametric alternative hypothesis. The testing problem is usually handled by using the Wald-type of statistics,

$$W_n(h) = \hat{\boldsymbol{\beta}}^T \mathbf{A}^T (\mathbf{A}\hat{\Sigma}_h \mathbf{A}^T)^{-1} \mathbf{A}\hat{\boldsymbol{\beta}},$$

where $\hat{\Sigma}_h$ is the estimated covariance matrix of $\boldsymbol{\beta}$, which involves estimated nonparametric function $\hat{g}$ and depends on a certain smoothing parameter h.

Note that under the null hypothesis (5), the problem is still a partially linear model. Hence, its parameters can be estimated by using the profile likelihood approach the same way as that in (4). The generalized likelihood ratio statistics can be computed by substituting the semiparametric estimators under both null and alternative hypotheses into the likelihood function, using the same bandwidth. Let the resulting estimator be $T_n(h)$. The question then arises if the Wilks type of result holds. Between the two approaches $W_n(h)$ and $T_n(h)$, it remains to be seen which method is more powerful and which method gives a better approximation in terms of the size of the test.

For the partially linear model (1), one naive and simple approach is to use the partially linear structure to reduce the testing problem (5) to an approximate linear model. Let $(Y_i, U_i, X_{i1}, \cdots, X_{ip})$ be the random sample ordered according to the variable U. Then, by model (4),

$$\begin{aligned} Y_{2i+1} - Y_{2i} &= g(U_{2i+1}) - g(U_{2i}) + \beta_1(X_{2i+1,1} - X_{2i,1}) + \cdots \\ &\quad + \beta_p(X_{2i+1,p} - X_{2i,p}) + \varepsilon_{2i+1} - \varepsilon_{2i} \\ &\approx \theta_0 + \theta_1(U_{2i+1} - U_{2i}) + \beta_1(X_{2i+1,1} - X_{2i,1}) + \cdots \\ &\quad + \beta_p(X_{2i+1,p} - X_{2i,p}) + \varepsilon_{2i+1} - \varepsilon_{2i}. \end{aligned} \tag{6}$$

Note that the maximum distance between the spacing U_{2i+1} and U_{2i} is of order $O(n^{-1}\log n)$, when the density of U has a bounded support. Thus, the coefficients θ_0 and θ_1 in model (6) can be taken to be zero 0. However, we keep these two parameters in the model (6) to make the approximation more accurate. This kind of ideas appear independently in Yatchew (1997) and Fan and Huang (2001). By using the approximate linear model (6), the problem (5) becomes a linear hypothesis under the approximate linear model (6) and the F-test statistics can be employed. One naturally asks how effective this simple and naive method is, comparing with the above more sophisticated Wald-test and the generalized likelihood ratio test. Note that we lose the information contained in the data $\{Y_{2i+1}+Y_{2i}\}$, which itself follows approximately model (4). The data $\{Y_{2i+1}+Y_{2i}\}$ does not contain nearly as much information about $\boldsymbol{\beta}$ as $Y_{2i+1}-Y_{2i}$, since the former involves the nuisance function g. Thus, the efficiency based on model (6) should, intuitively, be at least 50%.

Note that the above test can be regarded as a generalized likelihood ratio test with a very rough estimate of g. In fact, for given $\boldsymbol{\beta}$, one estimates g by taking the average of two neighboring points:

$$\hat{g}(u) = 2^{-1}\{Y_{2i+1}+Y_{2i}-\beta_1(X_{2i+1,1}+X_{2i,1})+\cdots \\ +\beta_p(X_{2i+1,p}+X_{2i,p})\}, \quad \text{for } u \in (\tfrac{U_{2i-1}+U_{2i}}{2}, \tfrac{U_{2i+1}+U_{2i+2}}{2}].$$

Substituting $\hat{g}$ into the models on Y_{2i+1}, we obtain

$$Y_{2i+1}-Y_{2i} = \beta_1(X_{2i+1,1}-X_{2i,1})+\cdots+\beta_p(X_{2i+1,p}-X_{2i,p})+2\varepsilon_{2i+1}.$$

A similar equation is obtained by substituting $\hat{g}$ into the model on Y_{2i}:

$$Y_{2i+1}-Y_{2i} = \beta_1(X_{2i+1,1}-X_{2i,1})+\cdots+\beta_p(X_{2i+1,p}-X_{2i,p})-2\varepsilon_{2i}.$$

The above two equations contain basically the same information as the model (6). Note that the estimator $\hat{g}$ here is significantly undersmoothed, but nonetheless gives reasonable inferences on the parametric component. It is consistent with a point hinted in the paper by Bickel and Kwon.

After obtaining nonparametric estimate $\hat{g}$, researchers frequently ask if certain parametric model fits the nonparametric component. Namely, one wishes to test

$$H_0: g(u) = g(u,\theta).$$

Again, the generalized likelihood statistics can be constructed and its sampling properties need to be studied.

4 Choice of bandwidth

Bickel and Kwon raised the question how to select bandwidths for semiparametric models. If the primary interest focuses on parametric components, the selected bandwidth should not create excessive biases in the estimation of nonparametric components. The reason is that the biases in the estimation of nonparametric components can not be averaged out in the process of estimating parametric components, yet the variance in nonparametric estimates can be averaged out. This is evidenced in the approximate linear model (6), where g is estimated by the average of two neighboring points. If one wishes to choose a bandwidth that estimates well both parametric and nonparametric components simultaneously, a profile likelihood approach is needed, as demonstrated by Carroll *etal.*(1997). However, in semiparametric estimation problems such as the partially linear model (4), one can also employ a two step estimation scheme: choose a small bandwidth that efficiently estimates the parametric component and then treat the parametric component as if it were known and apply a nonparametric technique, with an optimally chosen bandwidth, to estimate the nonparametric component.

The problem of choosing an appropriate smoothing parameter arises also in the hypothesis testing problem. For each given bandwidth parameter h, one can regard the generalized likelihood test $T_n(h)$ [see e.g. (3)] as a proper test statistic. The question then becomes how to choose a good smoothing parameter that maximizes the power. The multi-scale test proposed in Fan (1996) appears to achieve good asymptotic power, as shown in Fan (1996) and Fan *etal.*(2001), though his formulation is in the frequency domain. The idea can simply be translated into the current setting. We refer to Zhang (2000) for some related work.

I have no intension to advocate using only the generalized likelihood ratio statistics for semiparametric and nonparametric inferences. In fact, very few properties are known about the generalized likelihood ratio statistics. Even worse, the generalized likelihood statistics do not suggest any fixed procedure for estimating nonparametric components. Much more additional work is needed beyond the work by Fan *etal.*(2001). In light of no generally applicable guideline for nonparametric and semiparametric testing problems, I outline some ideas, rather than some solutions, here in an attempt to address the model validation question raised by Bickel and Kwon and to stimulate some further research in this area.

Additional References

Carroll, R.J., Fan, J., Gijbels, I, and Wand, M.P. (1997). Generalized partially linear single-index models. *Jour. Ameri. Satist. Assoc.*, **92**, 477-489

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998). Nonparametric estimation via local estimating equations. *Jour. Ameri. Statist. Assoc*, **93**, 214-227.

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

Fan, J. and Huang, L. (2001). Goodness-of-fit test for parametric regression models. *Journal of American Statistical Association.* In press.

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

Hastie, T.J. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

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

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Royal Statist. Soc. B*, **50**, 413–436.

Severini, T.A. and Wong, W.H. (1992). Generalized profile likelihood and conditional parametric models. *Ann. Statist.*, **20**, 1768–1802.

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

Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.*, **14**, 590–606.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics Letters*, **57**, 135-143.

Zhang, C.M. (2000). "Adaptive tests of regression functions via multi-scale generalized likelihood ratios". Technical Report #1026, Department of Statistics, University of Wisconsin–Madison.