Bandwidth Selection for Semiparametric Averaged Derivatives

Y. Nishiyama

Graduate School of Environmental Studies, Nagoya University, Nagoya 464-8601, Japan (nishiyam@info.human.nagoya-u.ac.jp)

Abstract: Density-weighted averaged derivative estimator gives a computationally convenient consistent and asymptotically normally distributed (CAN) estimate of the parametric component of a semiparametric single index model. This model includes some important parametric models as special cases such as linear regression, Logit/Probit, Tobit and Box-Cox and other transformation models. First order asymptotic properties of this estimator have been investigated in several papers for a variety of data generating processes. This estimator involves a nonparametric kernel density estimate and thus it faces the problem of bandwidth selection as is usual with kernel based nonparametric or semiparametric inference. Bandwidth selection is practically an important problem and that minimizing the mean squared error (MSE) has been mainly studied in various kernel estimation methods. It is certainly a desirable method for point estimation, however when we are more interested in hypothesis testing or confidence interval estimation based on the asymptotic normality, it may not be the best selection. An alternative method may be to choose it such that it minimizes the normal approximation error. The author considered its global minimization in a previous joint paper. The purpose of this paper is to propose a more desirable bandwidth for these purposes by minimizing the normal approximation error only in the tail. We report the results of a Monte Carlo study comparing alternative bandwidths based on a Tobit model.

Keywords: Semiparametric averaged derivatives; Higher order asymptotic theory; Minimum normal approximation error; Bandwidth selection

1. INTRODUCTION

Semiparametric single index model handles the situation where mean response of a variable conditional on a vector of other variables depends on it only through its linear combination without specifying the functional form of the underlying distribution and regression function. This model includes some special cases important in econometrics such as linear regression, Tobit, Logit/Probit and Box-Cox and other transformation models. Several semiparametric estimators have been proposed for the parametric component of the model by various authors. Among them, the density-weighted averaged derivative estimator considered by Powell *et. al.* [1989] and Robinson [1989] is a practically convenient CAN estimator because it has an explicit formula and does not require any numerical optimization, though this does not attain the semiparametric efficiency bound.

This estimator involves a nonparametric kernel density estimate and thus it faces the problem of bandwidth selection as is usual with kernel based nonparametric or semiparametric inference. Bandwidth selection is practically an important problem and that minimizing the mean squared error (MSE) has been considered in various kernel based estimation methods. It is certainly a suitable method for point estimation, however when we are more interested in hypothesis testing or confidence interval estimation based on the asymptotic normality, it may not be the best way. Nishiyama and Robinson [2000] propose an optimal bandwidth which globally minimizes the normal approximation error for these purposes. However there is a possibility that we can derive an even more desirable one by locally minimizing the approximation error. The idea is as follows. For example, suppose we would like to perform a two-sided significance test of size 5% on a parameter. When we know the exact distribution of the test statistic, we simply use its 2.5% and 97.5% quantiles to determine the critical region, and they obviously give the correct size. However, when it is unknown, we often use the asymptotic distribution instead. Suppose it is the standard normal distribution. Then we take its 2.5% and

97.5% quantiles to construct the critical region, but the corresponding empirical size is not necessarily 5%. Whether the size distortion is small or not depends only on how close the quantiles of the exact distribution are to those of the normal, not the global distance between the exact and the asymptotic distributions. Therefore, we should pay more attention to tail approximation rather than global approximation. The purpose of this work is to propose a new bandwidth choice suitable for hypothesis testing and confidence interval estimation for the density-weighted averaged derivatives based on this idea. We compare the small sample performance of the minimum MSE bandwidth, one by Nishiyama and Robinson [2000] and the new one proposed here by a Monte Carlo study based on a Tobit model. The following section briefly explains the semiparametric single index model and its estimator. Section 3 reviews bandwidth selection methods for this estimator. Section 4 proposes a new bandwidth, while Section 5 reports the results of a Monte Carlo study.

2. SEMIPARAMETRIC SINGLE INDEX MODEL

For a $d\times 1$ variate X with density $f(x)$ and a scalar variate Y, we suppose the regression function $g(X)=E(Y\,|\,X)$ is known to have single index form

$$g(X)=G(\beta^{\tau}X)$$

where $G:R\to R$ and $d\times 1$ vector β are unknown and τ denotes transposition. This is a single index model and it is easily seen that parametric models such as linear regression model, Tobit model, Probit model and Box-Cox transformation model are its special cases. β is identifiable only up to scale so that we consider estimation of β up to scale. Given an estimate of β, G can be estimated by some nonparametric method assuming certain smoothness condition.

Powell *et.al.* [1989], Robinson [1989] and Cheng and Robinson [1994] propose to estimate

$$\overline{\mu}\equiv -E\{G'(\beta^{\tau}X)f(X)\}\beta,$$

a constant times β, where $G'(u)=dG(u)/du$, by density-weighted averaged derivatives

$$U=\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}h^{-d-1}K'\left(\frac{X_i-X_j}{h}\right)(Y_i-Y_j) \quad (1)$$

given a sample (Y_i,X_i^{τ}), $i=1,\ldots,n$. h is a positive constant decaying to zero as $n\to\infty$, $K:R^d\to R$ is a differentiable L-th order kernel function and $K'(u)=\partial K(u)/\partial u$. In case of independently and indentically distributed (iid) observations, Powell *et.al.* [1989] prove

$$U\xrightarrow{p}\overline{\mu}$$

and

$$\sqrt{n}(U-\overline{\mu})\xrightarrow{d}N(0,\Sigma)$$

where $\Sigma=E[(\mu_1-\overline{\mu})(\mu_1-\overline{\mu})^{\tau}]$, with $\mu_1=\{Y_1-g(X_1)\}f'(X_1)-g'(X_1)f(X_1)$. They also provide a consistent estimator for the asymptotic variance Σ. The $n^{-1/2}$ convergence rate of this estimator was thought to be surprising because it involves a nonparametric density derivative estimate with a significantly slower convergence rate than $n^{-1/2}$. Robinson [1989] and Cheng and Robinson [1994] investigate the asymptotic properties of this estimator in the case of weakly and long-range dependent observations respectively, the former showing $\sqrt{n}$-consistency and asymptotic normality, while the latter proving the possibility of non-standard asymptotic distribution.

Robinson [1995] obtains the Berry-Esseen bound of the estimator for iid observations, suggesting the possibility of $O(n^{-1/2})$ bound which prevails in parametric statistics. Nishiyama and Robinson [2000, 2001] further validate Edgeworth expansions for the estimator when suitably normalized and studentized under certain regularity conditions.

Some other semiparametric CAN estimators of the parametric component in this model are proposed by Stoker [1986], Härdle and Stoker [1989], Andrews [1991], Ichimura and Lee [1988], Newey and Stoker [1993] and Ichimura [1993].

3. BANDWIDTH SELECTION METHODS

Bandwidth selection problem always arises in kernel based nonparametric and semiparametric inference. A standard principle is the minimum MSE bandwidth selection, while Nishiyama and Robinson [2000] propose a different way minimizing the normal approximation error primarily for the purposes of hypothesis testing and confidence interval estimation.

3.1 Minimum Mean Squared Error Bandwidth Selection

For (1), we may choose h, the bandwidth, such that it minimizes the MSE,

$$h_{MSE} = \arg\min_h E\{(U - \bar{\mu})^{\tau}(U - \bar{\mu})\}. \quad (2)$$

This principle appears to be favorable for the purpose of point estimation. Minimum MSE bandwidth is proposed by Härdle and Tsybakov [1993] for the density-weighted averaged derivatives, and Powell and Stoker [1996] for slightly more general statistics. They derive leading terms of the MSE and minimize it with respect to the bandwidth. In both articles, it is shown that the order of the optimal bandwidth is $O(n^{-2/(2L+d+2)})$. The choice by (2) is infeasible because the expectation is taken with respect to the unknown distribution of (Y, X^{τ}), but the latter authors also provide a feasible choice.

3.2 Minimum Normal Approximation Error Bandwidth Selection

Minimum MSE bandwidth may not be appealing when we are more interested in the confidence interval estimation or hypothesis testing based on the asymptotic normality. In terms of the significance test, if the convergence rate to the normal is not sufficiently fast and the normal approximation is poor, it may cause a great size distortion. The Monte Carlo simulation based on a Tobit sample in Nishiyama and Robinson [2000] show that the distribution function of the averaged derivatives is not always well approximated by the normal in small sample (n=100, 400) depending on the bandwidth selection.

Nishiyama and Robinson [2000] propose to determine the bandwidth such that the global normal approximation error is minimized based on an Edgeworth expansion for the studentized estimator they established. Putting

$$F(z) = P(\hat{\sigma}^{-1}\lambda^{\tau}(U - \bar{\mu}) \le z),$$

where λ is any $d \times 1$ vector satisfying $\lambda^{\tau}\lambda = 1$ and $\hat{\sigma}^2$ is the jackknife estimator of the asymptotic variance of $\lambda^{\tau}(U - \bar{\mu})$, they validate an Edgeworth expansion

$$\sup_{\lambda^{\tau}\lambda=1}\sup_{z\in R} |F(z) - \tilde{F}(z)| = o(n^{-1/2} + n^{-1}h^{-d-2} + n^{1/2}h^{L}) \quad (3)$$

where

$$\tilde{F}(z) = \Phi(z) - \phi(z)[n^{1/2}h^{L}\kappa_1 - \frac{\kappa_2}{nh^{d+2}}z - \frac{1}{3n^{1/2}}\{(2z^2+1)\kappa_3 + 3(z^2+1)\kappa_4\}], \quad (4)$$

$\Phi(z)$ and $\phi(z)$ are respectively the distribution and density function of a standard normal variate, and κ_i, $i = 1,2,3,4$ are constants depending on the underlying distribution (see Nishiyama and Robinson [2000] for their definitions). Using this Edgeworth expansion, they propose to choose the bandwidth such that

$$\min_{h>0}\max_{z} |F(z) - \Phi(z)|. \quad (5)$$

They focus on the case when $(d+2)/2 < L < 2(d+2)$ under which the first two correction terms in (4) dominate the third one, taking into account that "too" large L tends to make the estimator rather unstable for smaller sample size. It is interesting that this saddle point problem has a simple closed form solution

$$h_{NR} = \left\{\frac{(d+2)^2\kappa_2^2}{L(L+d+2)\kappa_1^2}\right\}^{1/2(L+d+2)} n^{-3/2(L+d+2)}. \quad (6)$$

They also provide strongly consistent estimates of κ_i, $i = 1,2,3,4$, for the feasible versions of (3) and (4), which can be plugged in (6) for practical use. It is straightforward to show that the bandwidth minimizing the MSE of $\lambda^{\tau}U$, for any $d \times 1$ vector λ, is

$$h_{MSE} = \left\{\frac{(d+2)\kappa_2}{L\kappa_1^2}\right\}^{1/(2L+d+2)} n^{-2/(2L+d+2)} \quad (7)$$

by Lemmas 1, 11 and 12 of Nishiyama and Robinson [2000]. Therefore $h_{MSE} > h_{NR}$ asymptotically.

4. OPTIMAL BANDWIDTH SELECTION MINIMIZING THE NORMAL APPROXIMATION ERROR IN THE TAIL

The bandwidth selection (6) looks appealing and practically applicable to confidence interval estimation and hypothesis testing. The idea (5), however, may not be ideal for these purposes. Firstly, it only allows kernel order smaller than 2(*d*+2), when the bias may badly affect the inference. Secondly, when we would like to perform a two-sided significance test on a element of $\bar{\mu}$ based on the asymptotic normality, for instance, what we need is not the globally good approximation but the good approximation in the tail, or more precisely certain quantiles. This idea leads to choose the bandwidth such that $F(z)$ is close to $\Phi(z)$ at 100γ % and $100(1-\gamma)$ % quantiles of the standard normal distribution. One way is to choose h such that

$$h_{NAE} = \underset{h>0}{\arg\min}\ [\{\tilde{F}(z_\gamma) - \Phi(z_\gamma)\}^2 + \{\tilde{F}(z_{1-\gamma}) - \Phi(z_{1-\gamma})\}^2]. \qquad (8)$$

where z_γ is the 100γ % quantile of $\Phi(z)$.

PROPOSITION

(i) *(8) has a unique solution satisfying*

$$n\kappa_1^2 L h^{2L+2d+4} - \frac{4}{3}\{(2z_\gamma^2+1)\kappa_3 + 3(z_\gamma^2+1)\kappa_4\}h^{L+2d+4} - \left(\frac{\kappa_2 z_\gamma}{n}\right)^2 (d+2) = 0. \qquad (9)$$

(ii) *Putting*

$$Q(h) = [\{\tilde{F}(z_\gamma) - \Phi(z_\gamma)\}^2 + \{\tilde{F}(z_{1-\gamma}) - \Phi(z_{1-\gamma})\}^2],$$

we have

$$Q(h_{NAE}) < Q(h_{NR}) < Q(h_{MSE})$$

as $n \to \infty$.

The proof is omitted. It is notable that the order of this optimal bandwidth choice must be of exact order $n^{-\frac{3}{2(L+d+2)}}$, which is the same as the one by Nishiyama and Robinson [2000]. Therefore we also have $h_{MSE} > h_{NAE}$ asymptotically.

Other ways could be considered, for instance,

$$\min_h \{|\tilde{F}(z_\gamma) - \Phi(z_\gamma)| + |\tilde{F}(z_{1-\gamma}) - \Phi(z_{1-\gamma})|\}.$$

This optimization problem, however, may be inconvenient because the objective function involves the absolute values and requires numerical optimization, so that (8) looks more favorable than this.

In the case of one-sided test, it may be better to choose h such that, e.g.,

$$\min_h \{\tilde{F}(z_\gamma) - \Phi(z_\gamma)\}^2,$$

but we concentrate only on (8) in this work because this choice obviously does not suit confidence interval estimation.

5. MONTE CARLO STUDY

5.1 Monte Carlo Design

Consider the Tobit model

$$Y = (\beta^\tau X + \varepsilon)\, I(\beta^\tau X + \varepsilon \geq 0)$$

where $X = (X_1, X_2)$ is bivariate and *I*(.) is the indicator function. We took $(X^\tau, \varepsilon) \sim N(0, I_3)$, so that

$$g(x) = \beta^\tau x\{1 - \Phi(-\beta^\tau x)\} + \phi(-\beta^\tau x)\}$$

and $\bar{\mu} = -\beta/(8\pi)$, thus it satisfies the single index constraint. We set $\beta = (1,1)^\tau$. For the estimator, we used the normal density based higher order kernel in Robinson [1988] and employed three values of L, L=4,8 and 10.

We compare the nominal size of a hypothesis test with the empirical size to check out the size distortion in small samples of n=100, 400. We test for

$$H_0 : \bar{\mu}_1 = -1/(8\pi)$$

versus

$$H_1 : \bar{\mu}_1 \neq -1/(8\pi)$$

based on the asymptotic normality of U, where $\bar{\mu}_1$ is the first element of $\bar{\mu}$. We implemented a Monte Carlo simulation of the above two-sided test of size 5% with 10,000 replications. We used the consistent estimates of κ_i proposed by Nishiyama and Robinson [2000] to compute h_{MSE}, h_{NR} and h_{NAE}. For general comparison, we also implemented the simulation for h=0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.

5.2 Results

We first give the values of h_{MSE}, h_{NR} and h_{NAE} in Table 1. h_{NR} is reported only when L=4 because it is justified only when $(d+2)/L<L<2(d+2)$. Comparing the third and fourth columns, we see that the optimal bandwidths h_{MSE}, h_{NR} and h_{NAE} are smaller for the larger values of n in each of L=4, 8, 10. We also find that h_{MSE} is always larger than h_{NAE} for fixed n when L=8, 10, and $h_{MSE} > h_{NAE} > h_{NR}$ when L=4 . This is consistent with the discussion in the end of the previous section.

Table 1. Optimal bandwidths

		n=100	n=400
L=4	h_{MSE}	0.64	0.58
	h_{NAE}	0.59	0.51
	h_{NR}	0.51	0.44
L=8	h_{MSE}	1.05	0.84
	h_{NAE}	0.92	0.79
L=10	h_{MSE}	1.08	0.93
	h_{NAE}	0.98	0.81

Tables 2, 3 and 4 show the empirical size of the test for each bandwidth and sample size respectively for the cases of L=4, 8, 10. We found in general that larger bandwidth yields larger rejection probability. This is mainly because of the bias. It is easily seen that the size distortion is enormous when "too" large bandwidth is used. For example, the case h=1.0 and n=400 in Table 2 is the worst with the empirical size of 80%. The size distortion looks relatively less devastating for larger values of L(=8, 10) than the small L(=4) for h=0.4-1.0.

Table 2. Empirical size, L=4
(Nominal size=5%)

H	n=100	n=400
0.4	1.75%	3.13%
0.5	3.26%	5.50%
0.6	5.52%	9.05%
0.7	8.36%	15.69%
0.8	13.24%	29.22%
0.9	20.60%	52.41%
1.0	32.81%	80.23%
h_{MSE}	6.94%	8.51%
h_{NAE}	5.37%	5.97%
h_{NR}	3.51%	4.15%

Table 3. Empirical size, L=8
(Nominal size=5%)

h	n=100	n=400
0.4	0.74%	1.18%
0.5	1.33%	2.01%
0.6	2.12%	3.15%
0.7	2.95%	4.08%
0.8	4.01%	4.99%
0.9	5.01%	5.91%
1.0	6.09%	6.98%
h_{MSE}	7.23%	5.71%
h_{NAE}	5.23%	4.90%

h_{MSE} must be good for point estimation, but it does not perform well for hypothesis testing. When n=100, for instance, the empirical size is 6.94%, 7.23% and 13.01% for L=4, 8, 10 respectively. Typically it oversmoothes and the bias significantly affects the empirical size (Nishiyama and Robinson [2000] discuss that the dominant higher order term in the Edgeworth expansion is the bias related term when h_{MSE} is used).

Table 4. Empirical size, L=10

(Nominal size=5%)

h	n=100	n=400
0.4	0.81%	1.55%
0.5	1.35%	2.21%
0.6	2.23%	3.30%
0.7	3.24%	4.05%
0.8	4.12%	5.04%
0.9	5.50%	6.28%
1.0	7.13%	9.48%
h_{MSE}	13.01%	6.63%
h_{NAE}	7.04%	5.22%

h_{NAE}, on the other hand, seems to work quite well for all combinations of (L,n) except $(L,n)=(10,100)$ and clearly better than h_{MSE}. The poor result when $(L,n)=(10,100)$ may be partly because the estimation is typically less stable when higher order kernels are used.

When $L=4$, we also computed the empirical size under h_{NR}. From this Monte Carlo study, h_{NAE} seems to be better than h_{NR} when $n=100$, but it is not clear for $n=400$. However, h_{NAE} has an advantage over h_{NR} that the former can be used with any higher order kernel function, while the latter allows only the kernel functions satisfying $L<2(d+2)$.

We conclude that the simulation results support the use of h_{NAE} for the hypothesis testing in the sense that the empirical size is the closer to the nominal size of the test than h_{MSE} and that it allows any kernel function unlike h_{NR}. It is obvious that it also works well for confidence interval estimation.

6. REFERENCES

Andrews, D.W.K., Asymptotic normality of series estimators for nonparametric and semiparametric regression models, *Econometrica*, 59, 307-345, 1991.

Cheng, B. and P.M. Robinson, Semiparametric estimation for time series with long range dependence, *Journal of Econometrics*, 64, 335-353, 1994.

Härdle, W.A., and T.M. Stoker, Investigating smooth multiple regression by the method of averaged derivatives, *Journal of the American Statistical Association*, 84, 986-995, 1989.

Härdle, W.A., and A.B. Tsybakov, How sensitive are averaged derivatives?, *Journal of Econometrics*, 58, 31-48, 1993.

Ichimura, H., Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics*, 58, 71-120, 1988.

Ichimura, H. and L.F. Lee, Semiparametric least squares estimation of multiple index models: single equation estimation, in *Nonparametric and semiparametric methods in econometrics and statistics, Cambridge University Press*, 1988.

Newey, W.K. and T.M. Stoker, Efficiency of weighted averaged derivative estimators and index models, *Econometrica,* 61, 1199-1223, 1993.

Nishiyama, Y., and P.M. Robinson, Edgeworth expansions for semiparametric averaged derivatives, *Econometrica*, 68, 4, 931-979, 2000.

Nishiyama, Y., and P.M. Robinson, Studentization in Edgeworth expansions for estimates of semiparametric Index Models, in *Nonlinear Statistical Modeling*, 197-240, 2001.

Powell, J.L., J.H. Stock and T.M. Stoker, Semiparametric estimation of single index coefficients, *Econometrica,* 57, 1403-1430, 1989.

Powell, J.L. and T.M. Stoker, Optimal bandwidth choice for density-weighted averages, *Journal of Econometrics,* 75, 291-1316, 1996.

Robinson, P.M., Root-N-consistent semiparametric Regression, *Econometrica*, 56, 931-954, 1988.

Robinson, P.M., Hypothesis testing in semiparametric and nonparametric models for Econometric time series, *Review of Economic Studies*, 56, 511-534, 1989.

Robinson, P.M., The normal approximation for semiparametric averaged derivatives, *Econometrica*, 63, 667-680, 1995.

Stoker, T.M., Consistent estimation of Scaled coefficients, *Econometrica*, 54, 1461-1481 1986.