Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Adjusted Confidence Bands in Nonparametric Regression

Guoyi Zhang [a] & Yan Lu [b]

[a] Department of Mathematics and Statistics, Arizona State University, Tempe, Arizona, USA

[b] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA

Version of record first published: 03 Jan 2008.

To cite this article: Guoyi Zhang & Yan Lu (2007): Adjusted Confidence Bands in Nonparametric Regression, Communications in Statistics - Simulation and Computation, 37:1, 106-113

To link to this article: http://dx.doi.org/10.1080/03610910701723930

Communications in Statistics—Simulation and Computation®, 37: 106–113, 2008

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910701723930

Taylor & Francis
Taylor & Francis Group

Regression Analysis

Adjusted Confidence Bands in Nonparametric Regression

GUOYI ZHANG[1] AND YAN LU[2]

[1]Department of Mathematics and Statistics, Arizona State University, Tempe, Arizona, USA
[2]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA

Suppose we have $\{(x_i, y_i)\}$ $i = 1, 2, \ldots, n$, *a sequence of independent observations. We wish to find approximate* $1 - \alpha$ *simultaneous confidence bands for the regression curve. Many previous confidence bands in the literature have practical difficulties. In this article, the local linear smoother is used to estimate the regression curve. The bias of the estimator is considered. Different methods of constructing confidence bands are discussed. Finally, a possible method incorporating logistic regression in an innovative way is proposed to construct the bands for random designs. Simulations are used to study the performance or properties of the methods. The procedure for constructing confidence bands is entirely data-driven. The advantage of the proposed method is that it is simple to use and can be applied to random designs. It can be considered as a practically useful and efficient method.*

Keywords Confidence bands; Local linear smoother; Nonparametric regression.

Mathematics Subject Classification 62G15.

1. Introduction

Nonparametric regression models have received considerable attention in recent years. Many of the major issues such as bandwidth selection, kernel functions, minimax efficiencies, and best uniform convergence rates are well studied. Confidence bands in nonparametric regression have been studied for a long time. Unfortunately, current methods of constructing confidence bands in nonparametric regression still have some practical difficulties. In this article, we will develop a new method to construct confidence bands for random designs in nonparametric regression.

Consider the general nonparametric regression model

$$y_i = m(x_i) + \sigma(x_i)\varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

Received July 6, 2005; Accepted May 25, 2007
Address correspondence to Guoyi Zhang, Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85281, USA; E-mail: guoyi.zhang@asu.edu

where $\{\varepsilon_i\}$ is a sequence of independent, identically distributed, random variables with $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = 1$; $m(\cdot)$ is an unknown smooth regression curve; $\{(x_i, y_i)\}$ is a sequence of observations; and x_i has a density function $f(\cdot)$. In this article, we assume $\sigma(\cdot)$ is a constant function. In general, it could be any unknown smooth function. Without loss of generality, we assume that $x_i \in [0, 1]$, $i = 1, 2, \ldots, n$.

Given $\alpha \in (0, 1)$, to construct a confidence band, we need an estimator $\hat{m}(x)$ for $m(x)$ and a bound l_α, such that

$$P\{|\hat{m}(x) - m(x)| \leq l_\alpha \text{ for all } x\} \geq 1 - \alpha. \tag{2}$$

Then, a $100(1 - \alpha)\%$ confidence band can be constructed as

$$\hat{m}(x) \pm l_\alpha.$$

Bickel and Rosenblatt (1973) studied the limiting distribution of the maximum absolute deviation for kernel density estimators. Härdle (1989) used this approach to develop similar results for kernel regression. These confidence bands require the bias of $\hat{m}$ to be negligible relative to its standard error. However, most of the common data-driven bandwidth selectors minimize mean squared error. Consequently, the selected bandwidth always balances the bias and the variance. Hence, it is unwise to hope the bias can be ignored automatically.

There have been two different directions to deal with the bias in the literature. The first direction is to subtract an estimator of the bias from the estimator and use the result as a pivotal quantity for constructing confidence bands. In other words, they try to recenter the naive bands that are often used in this setting. A bias-corrected confidence band was first proposed by Eubank and Speckman (1993). Their confidence band is for the fixed uniform designs. Extensions of the bias-corrected confidence band to random and non uniform designs were developed by Xia (1998). A major disadvantage of a bias-corrected confidence band is that it involves estimating the second derivative of the regression function, which is not stable. When we construct our bias-corrected bands, we are really estimating m by $\hat{m}$-bîas, which should have a bigger variance than that of $\hat{m}$. "Subtracting the estimator of the bias will generally increase variance more than it reduces bias" (Sun and Loader, 1994); we can find similar reports from Härdle and Marron (1991). Fan and Zhang (2000) derived a bias-corrected confidence band but used a smaller bandwidth, $\frac{1}{2}\hat{h}$, in their simulation and application, claiming that "with this small bandwidth, we ignore the bias in the construction of simultaneous confidence bands." However, if the bandwidth is too small, the estimator of the regression function will have a very large variance and be far from the "optimal" estimator. The second direction is to use some techniques to reduce the bias and then ignore the bias instead of estimating the bias. For example, Sun and Loader (1994) suggested using the tricube weight function and fitting local quadratic polynomials. Claeskens and Keilegom (2003) suggested under smoothing to reduce the bias. However, it is well known that the asymptotic distribution of $\sup|\hat{m} - m|$ is very sensitive. Consequently, it is doubtful that the bias can be ignored for general cases after using some techniques to reduce the bias. Even relatively small bias may still have a serious impact on the coverage.

Despite the difficulties involved, the goal of this research is to find a relatively simple and efficient solution to the confidence band problem. Eubank and Speckman (1993) and Xia (1998) attempted to recenter the bands by the correction of the bias. The confidence band takes the form: $\hat{m}(x) - \hat{\text{bias}} \pm l_\alpha$. An alternative approach is to expand the bands to compensate for the bias. The confidence band takes the form: $\hat{m}(x) \pm c * l_\alpha$. We use the second approach to construct our proposed confidence band.

This article is organized as follows. We introduce our adjusted confidence band in Sec. 2. The simulation study is reported in Sec. 3. Finally, we give a summary in Sec. 4.

2. Adjusted Confidence Bands

The confidence bands using the approach proposed by Bickel and Rosenblatt (1973) are widely accepted because these confidence bands have a limiting distribution that does not involve the unknown mean function in theory. Eubank and Speckman (1993) and Xia (1998) estimated the bias and recentered the confidence bands to make them more possible to be used in practice. However, from their article we can see that the convergence rate is very slow. One difficulty is that, in practice, we have finite samples, such as $n = 200$. (This is considered as a large sample size in general regression problems.) Since the convergence rate is very slow, this sample size is not large enough to approximate the population quantities by sample estimators. Even with some modifications of the estimators of the bounds l_α (as we can see in Eubank and Speckman, 1993, and Xia, 1998, they modified the asymptotic variances of the estimator of $m(x)$ for the finite sample), the coverage is still not satisfied. We did simulations using the methods proposed by Xia (1998) with bias correction and also without bias correction. The simulation results are available in Tables 1 and 2. (Similar results can be also found from Claeskens and Keilegom, 2003.) Xia's method does not work well when using function $m(x) = x * (1 - x)$ and the fixed uniform design. The coverage is not improved at all. From Table 1, when $\sigma = 0.05$, with nominal coverage = 90%, coverage for $n = 200$ is only 0.766. This is even lower than the coverage 0.814 (see Table 2) using the same setting but without the bias correction term.

Table 1
With the bias correction term (we use function $m(x) = x * (1 - x)$ and the fixed uniform design in this simulation)

σ	Nominal coverage (%)	Method	Coverage for the following n: 50	100	200
0.05	90	xia	0.804	0.830	0.766
	95	xia	0.874	0.888	0.870
0.1	90	xia	0.774	0.782	0.730
	95	xia	0.898	0.880	0.846

Table 2
Without the bias correction term (we use function $m(x) = x * (1 - x)$ and the fixed uniform design in this simulation)

σ	Nominal coverage (%)	Method	Coverage for the following n: 50	100	200
0.05	90	No bias	0.760	0.790	0.814
	95	No bias	0.848	0.892	0.878
0.1	90	No bias	0.774	0.782	0.818
	95	No bias	0.858	0.864	0.898

Besides recentering the bands by the correction of the bias, another approach is to expand the bands to account for the bias. So, the confidence bands take the form

$$\hat{m}(x) \pm c * l_\alpha(x),$$

where

$$\begin{aligned}
l_\alpha(x) &= \frac{\hat{\sigma}(x)\widehat{V}}{\left(\sum_{t=1}^{n} w_t(x)\right)^{1/4}} \left\{ \sqrt{-2\log(\hat{h})} + \frac{1}{\sqrt{-2\log(\hat{h})}}(A - X_\alpha) \right\}, \\
A &= \log\left\{ \frac{1}{2\pi} \left(\int K'(u)^2 du \Big/ \int K(u)^2 du \right)^{1/2} \right\}, \\
X_\alpha &= \log\left\{ \frac{-\log(1-\alpha)}{2} \right\}, \\
w_t(x) &= K\left(\frac{x_t - x}{h}\right)\left(s_2 - \frac{x_t - x}{h} s_1\right), \\
s_l &= \sum_{t=1}^{n} K\left(\frac{x_t - x}{h}\right)\left(\frac{x_t - x}{h}\right)^l \quad l = 1, 2, \\
\hat{m}(x) &= \sum_{t=1}^{n} w_t(x) y_t \Big/ \sum_{t=1}^{n} w_t(x) \\
V &= \sqrt{\int K(u)^2 du},
\end{aligned} \tag{3}$$

$K(\cdot)$ is a kernel function, $\hat{h}$ is the selected bandwidth, $\hat{\sigma}(x)$ is the consistent estimator of $\sigma(x)$, and

$$c = \inf_{b \in R} \left\{ P(|\hat{m}(x) - m(x)| \le b * l_\alpha \text{ for all } x) \ge 1 - \alpha \right\}. \tag{4}$$

We call c an adjustment value. The advantage of our suggested confidence band is that it is simple to use and can be applied to random designs.

Proposition 2.1. *The adjustment value is a monotone decreasing function of sample size n and the limit of the adjustment value is* 1.

Proof. From (3), the bound l_α has two parts. The first part, $\hat{\sigma}(x)\widehat{V}/\left(\sum_{t=1}^{n} w_t(x)\right)^{1/4}$, is the estimated standard deviation of $\hat{m}$. The second part, $\left\{\sqrt{-2\log(\hat{h})} + \frac{1}{\sqrt{-2\log(\hat{h})}}\right\}(A - X_\alpha)$, is a decreasing function of the bandwidth. As the sample size increases, the selected bandwidth decreases. The bias and the first part of the bound decay at the same rate: namely, $n^{-\frac{2}{5}}$. The second part of the bound increases as the sample size increases. Hence, the bias decays faster than the bound. We only need a smaller c to compensate for the bias when we use a bigger sample size. In other words, the adjustment value c is a monotone decreasing function of the sample size n. It is also clear from (3) that $l_{\alpha 1}/l_{\alpha 2} \to 1$ as $n \to \infty$ for any fixed $0 < \alpha_1, \alpha_2 < 1$. The effect of α on the width of the band l_α is of second order. Consequently, using the definition from (4), the limit of the adjustment value is 1.

The remaining problem is to find an appropriate estimator of c. We did pilot simulation to find some properties of the adjustment value c. The lower bound of c is obviously 1. By the fact that the absolute value of the bias is half of the standard deviation when we use the optimal bandwidth (see Prewitt and Lohr, 2006), we can use $\frac{1}{2}l_\alpha$ to compensate for the bias conservatively so that $\frac{3}{2}l_\alpha$ is a possible upper bound for our confidence band. We did simulation for the fixed uniform design using function $m(x) = \exp\{-16 * (x - 0.5)^2\}$ and $c = 1.5$. The simulation result is found in Table 3. We can see that the coverage is always greater than the nominal level. This suggests that 1.5 is a possible upper bound.

Given true regression function $m(x)$, sample size n, and σ, using the definition of (4), we can find c. In practice, the true regression function is generally not available. So, we define c^* as the random variable which is corresponding to the estimator of the regression function $\hat{m}(x)$, sample size n, and $\hat{\sigma}$ from the given data. Based on our experience, c^* is very close to c. The basic idea is that we can generate random samples using the estimator of the regression function, sample size and $\hat{\sigma}$. Then, set a grid for c from 1 to 1.5 and do simulations for each value. Using definition (4), we can find the estimator of c^*, i.e., the value for which the simulation coverage is closest and greater than or equal to the nominal coverage. Unfortunately, this

Table 3

The fixed uniform design, function $m(x) = \exp\{-16 * (x - 0.5)^2\}$ (The confidence band takes the form $\hat{m}(x) \pm c * l_\alpha(x)$. We can see that the coverage is always greater than the nominal level. This suggests that 1.5 is a possible upper bound)

σ	Nominal coverage (%)	Method	Coverage for the following n: 50	100	200
0.05	90	$c = 1.5$	0.974	0.986	0.994
	95	$c = 1.5$	0.992	0.996	0.996
0.1	90	$c = 1.5$	0.988	0.968	0.992
	95	$c = 1.5$	0.990	0.998	0.998

process is not efficient at all. Hence, we develop a process to estimate the value of c by incorporating logistic regression in an innovative way.

We describe the process, process 1, as follows. Given a data set, we estimate $m(x)$ and σ, say $\hat{m}, \hat{\sigma}$. We then generate 1,000 samples (for each sample, we have n independent pairs of observations) from $\hat{m}, \hat{\sigma}$, and the x's. Between the lower bound 1 and upper bound 1.5 of c, we set d_i $(i = 1, 2, \ldots, 1{,}000)$ as equally spaced points. For different samples, we try different d_t's. If the confidence band contains the estimator of the regression function, set $w_i = 1$ and otherwise take w_i to be zero. The process will generate a sequence of response w_i $(i = 1, 2, \ldots, 1{,}000)$ which are binary variables with a Bernoulli distribution. We use the logistic regression model to fit the data, $\{(d_i, w_i)\}$ $i = 1, 2, \ldots, 1{,}000$ and find the value corresponding to the probability $100(1 - \alpha)\%$ using $\pi(x) = e^{\beta_0 + \beta_{1x}}/(1 + e^{\beta_0 + \beta_1 x})$. This value is our estimator of c.

3. Simulation

In this section we use simulations to study the performance of the method. We use the following regression function that has been used in Eubank and Speckman (1993) and Xia (1998),

$$m(x) = \exp\{-16 * (x - 0.5)^2\},$$

where $0 \leq x \leq 1$. We use the random normal design, $x_i \sim N(0.5, 1)$, $\{x_i\}$ is a sequence of independent, identically distributed, random variables with $E(x_i) = 0.5$ and $E((x_i - 0.5)^2) = 1$. We tried two sample sizes $n = 100, 200$. σ is chosen to be 0.05 and 0.1. α is set to be 0.1 and 0.05. The confidence bands are investigated for the different settings combining sample size, σ and α. For all the data analyzed below, we use the Epanechnikov kernel for the local linear smoother of $m(x)$. Cross-validation method is used to select the bandwidth and GSJS method (proposed by Gasser et al., 1986, and thereafter referred to as the GSJS estimator) is used to estimate the σ. We conduct 5,000 replications at each combination of function m, error noise σ, and sample size n, using a different seed for every case. The grid of 100 evenly spaced points in the interval [0, 1] is used.

The preferred simulation should generate 5,000 samples and for each sample we use process 1 to estimate c and then use this $\hat{c}$ to construct the confidence band and have a success or failure result. However, it takes two hours to estimate c using process 1. Consequently, we need 10,000 hours. Practically speaking, it is impossible

Table 4
The estimation of c (by logistic regression model)

σ	Nominal coverage (%)	Coverage for the following n: 100	200
0.05	90	1.339188	1.286716
	95	1.385901	1.286741
0.1	90	1.283638	1.258963
	95	1.305620	1.247548

Table 5
Nonuniform random design, function $m(x) = \exp\{-16 * (x - 0.5)^2\}$, using $\hat{c}$ from Table 4

σ	Nominal coverage (%)	Coverage for the following n: 100	200
0.05	90	0.8866	0.9182
	95	0.9520	0.9556
0.1	90	0.8980	0.8992
	95	0.9530	0.9514

to carry out this simulation. Notice that when the sample size is bigger than 100, $\hat{m}(x)$ and $\hat{\sigma}$ are very close to the true function and σ. The c value for estimated settings using $\hat{m}(x)$ and $\hat{\sigma}$ (we defined it c^* previously) should be very close to the c value for the true settings. Hence, we did following simulations to investigate this method. We generated 1,000 samples for each experimental setting and estimated c using process 1. We call this estimator $\hat{c}$. The results were shown in Table 4. Then we did 5,000 simulations for each experimental setting using corresponding $\hat{c}$ value. The results were shown in Table 5. The empirical coverage are very close to nominal level. Hence, we believe logistic regression model is appropriate to be used here. The simulation we did shows that in practice $\hat{c}^*$ from the logistic regression model should be very close to c^*, which suggest $\hat{c}^*$ could be a good estimator of c when c^* is very close to c.

4. Concluding Remarks

In this article, we have proposed a new method to construct simultaneous confidence bands for the random designs in nonparametric regression. The local linear smoother is used to estimate the regression curve. Instead of subtracting the estimator of the bias, we expand the confidence bands by multiplying by an adjustment value c to account for the bias of the smoother. The confidence bands take the form $\hat{m}(x) \pm c * l_\alpha(x)$. Logistic regression is used in an innovative way to estimate the bias adjustment c. The resulting procedure for constructing confidence bands is entirely data-driven. The advantage of our method is that it is simple to use and can be applied to all the cases, such as the uniform design, nonuniform random design, etc. Simulations show that the method we proposed provides a promising way to construct confidence bands in nonparametric regression.

References

Bickel, P. J., Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates (Corr: V3 p1370). *The Annals of Statistics* 1:1071–1095.

Claeskens, G., Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics* 31(6):1852–1884.

Eubank, R. L., Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association* 88:1287–1301.

Fan, J., Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics* 27(4):715–731.

Gasser, T. Sroka, L., Jennen-Steinmetz, C. (1986). Residual vairance and residual pattern in nonlinear regression. *Biometrika* 73:625–633.

Härdle, W. (1989). Asymptotic maximal deviation of *M*-smoothers. *Journal of Multivariate Analysis* 29:163–179.

Härdle, W., Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics* 19:778–796.

Prewitt, K., Lohr, S. (2006). Bandwidth selection in local polynomial regression using eigenvalues. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 68(1):135–154.

Sun, J., Loader, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics* 22:1328–1345.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society, Series B, Methodological* 60:797–811.