UNIFORM CONFIDENCE BOUNDS FOR NONPARAMETRIC REGRESSION

by

Steinar Bjerne
University of Oslo,

Kjell A. Doksum
Department of Statistics
University of California, Berkeley

and

Brian S. Yandell
Statistics Department
University of Wisconsin

Abstract

Let \((X, Y)\) be a bivariate random variable and let \(m(x) = E(Y|X = x)\) be the regression function of \(Y\) on \(X\). Suppose that \(Y_1, \ldots, Y_n\) are independent observations of \(Y\) at \(X = x_1, \ldots, x_n\). We consider nearest neighbor estimates, \(\hat{m}(x)\), and employ well-known inequalities to obtain exact and asymptotic uniform confidence bounds for \(E\hat{m}(x)\) and \(m(x)\) based on \(\hat{m}(x)\). Finally we discuss bias-properties of \(\hat{m}(x)\).

Key words: Nonparametric regression, confidence bands, nearest neighbor.
1. Introduction.

Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be a random sample from a bivariate population with distribution function \(F(x, y)\). We are interested in constructing uniform confidence bounds for the unknown regression function,

\[ m(x) = E(Y|X = x) \]

without making parametric assumptions about either \(m\) or the distributional form of \(F\). We will assume existence of the conditional variance function given by

\[ \sigma^2(x) = \text{var}(Y|X = x) \]

The construction will be based on the \(k\)-nearest neighbor estimator,

\[ \hat{m}(x) = \sum Y_i / k \]

where the summation is taken over the indexes of the \(k\) \(X\)'s that lie closest to \(x\).
2. Preliminaries.

Let $X_{n_i}$ be the $i$th order statistic obtained from $X_1, X_2, \ldots, X_n$ and let $Y_{n_i}$ denote the $i$th induced order statistic of the $Y$-observations. I.e. if $Q_i$ is the anti-rank of $X_i$, $X_{Q_i} = X_{n_i}$, then $Y_{n_i} = Y_{Q_i}$.

Conditional on $X_{n_i} = x_{n_i}$, $i = 1, \ldots, n$, the $Y_{n_i}$ are independent (see e.g. Bhattacharya (1974), Lemma 1) and the distributional assumptions may be written as follows:

$$Y_{n_i} = m(x_{n_i}) + \epsilon_{n_i} \quad i = 1, \ldots, n,$$

where $\epsilon_{n_1}, \ldots, \epsilon_{n_n}$ are independent, (2.1)

$$x_{n_1} \leq \ldots \leq x_{n_n}, \quad E\epsilon_{n_i} = 0 \text{ and } \text{Var}(\epsilon_{n_i}) = \sigma^2(x_{n_i}).$$

For convenience we will, until Section 5, write $Y_i$ for $Y_{n_i}$ and $x_i$ for $x_{n_i}$.

Our estimator for $m(x)$ will be the $k$-nearest neighbor estimator,

$$\hat{m}(x) = \sum_{i \in I_{nk}(x)} \{Y_i/k\},$$

where $I_{nk}(x)$ are the indices of the $k$ values of $x_1, \ldots, x_n$ closest to $x$. We assume that $I_{nk}(x)$ is uniquely determined a.s.

Let

$$J_i = \{x: I_{nk}(x) = \{i+1, \ldots, i+k\}, \quad i = 0, \ldots, n-k \}.$$  (2.3)

Denote

$$\hat{m}_i = \hat{m}(x) \text{ for } x \in J_i, \quad i = 0, \ldots, n-k.$$

Thus

$$\hat{m}_i = (Y_{i+1} + \cdots + Y_{i+k})/k, \quad i = 0, \ldots, n-k.$$  (2.4)
It is seen that $J_0 = (-\infty, \frac{1}{2}(x_1 + x_{k+1}))$ while

$$J_i = \left(\frac{1}{2}(x_i + x_{i+k}), \frac{1}{2}(x_i + x_{i+k})\right), \ i = 1, \ldots, n-k$$

Here, define $x_{n+1} = \infty$, so that $\bigcup_{i=0}^{n-k} J_i = \mathbb{R}$.

Let

$$S_1 = \tilde{m}_i - E\tilde{m}_i = \sum_{j=i+1}^{i+k} W_j, \ i = 0, \ldots, n-k,$$

where

$$W_j = \frac{(Y_j - m(x_j))}{k}, \ j = 1, \ldots, n.$$ 

Define $\tilde{S}_0 = 0$ and

$$\tilde{S}_i = \sum_{j=1}^{i} W_j, \ i = 1, \ldots, n-k.$$ 

Then

$$S_i = \tilde{S}_{i+k} - \tilde{S}_i, \ i = 0, \ldots, n-k.$$ 

**LEMMA.** For each $t > 0$,

$$P(S_i \geq -t; \text{all } i = 0, \ldots, n-k) \geq 1 - 4(kt)^{-2} \left[ \sum_{i=1}^{n-k} \sigma^2(x_i) + \sum_{i=k}^{n} \sigma^2(x_i) \right].$$

**PROOF.** Let

$$A = \{\tilde{S}_{k+i} \geq -\frac{1}{2}t \text{ all } i = 0, \ldots, n-k\}$$

$$B = \{\tilde{S}_i \leq \frac{1}{2}t \text{ all } i = 1, \ldots, n-k\}$$

$$C = \{S_i \geq -t \text{ all } i = 0, \ldots, n-k\}$$

Then $A \cap B \subseteq C$ and by Bonferroni's inequality, we have
\[ P(C) \geq P(A \cap B) \geq 1 - P(A^C) - P(B^C) \]

Kolmogorov's inequality (e.g., Loève, 1963, p. 235), yields

\[ P(A) \geq 1 - 4(kt)^{-2} \sum_{i=k}^{n} \sigma_i^2(x_i) \]

\[ P(B) \geq 1 - 4(kt)^{-2} \sum_{i=k}^{n-k} \sigma_i^2(x_i) \]

and the results follows.
3. Exact uniform confidence bounds for $E \hat{m}(x)$. Consistency.

Since $\inf_x [\hat{m}(x) - E \hat{m}(x)] = \min \{S_i; i = 0, \ldots, n-k\}$, then by the Lemma,

$$P(\inf_x [\hat{m}(x) - E \hat{m}(x)] \geq -t) \geq 1 - 4(kt)^{-2} \left[ \sum_{i=1}^{n-k} \sigma_i^2(x_i) + \sum_{i=k}^{n} \sigma_i(x_i) \right].$$

Thus if we set

$$t_\alpha = \left( \frac{2/\alpha^2 k}{n-k} \sum_{i=1}^{n-k} \sigma_i^2(x_i) + \sum_{i=k}^{n} \sigma_i(x_i) \right)^{1/2}$$

then $\hat{m}(x) + t_\alpha$ is a simultaneous upper confidence boundary for $E \hat{m}(x)$ with confidence coefficient at least $1 - \alpha$.

Similarly, $\hat{m}(x) - t_\alpha$ is a simultaneous lower confidence boundary for $E \hat{m}(x)$ with confidence coefficient at least $1 - \alpha$. Let

$$T^\pm = \sup_x [\hat{m}(x) - E \hat{m}(x)]$$

and $T = \max \{T, T^+\}$ then $P(T \geq t) \leq P(T^+ \geq t) + P(T \leq -t)$.

It follows that, for each $t > 0$,

$$P(\sup_x |\hat{m}(x) - E \hat{m}(x)| \leq t) \geq 1 - 8(kt)^{-2} \left[ \sum_{i=1}^{n-k} \sigma_i^2(x_i) + \sum_{i=k}^{n} \sigma_i(x_i) \right].$$

Thus $\hat{m}(x)$ is a uniformly consistent estimate of $E \hat{m}(x)$ provided

$$k^{-2} \sum_{i=1}^{n-k} \sigma_i^2(x_i) \rightarrow 0 \text{ as } n \rightarrow \infty.$$ If $\sigma^2(x)$ is bounded, this follows if $(n/k^2) \rightarrow 0$ as $n \rightarrow \infty$.

Note that

$$\hat{m}(x) \pm \sqrt{2} t_\alpha$$

is a level $(1 - \alpha)$ simultaneous confidence band for $E \hat{m}(x)$. If we assume that $\sigma^2(x) \equiv \sigma^2$ for all $x$, then

$$t_\alpha = \left( 2 \sqrt{2(n-k)} + 1/\sqrt{\alpha} \right) \sigma$$

and the width of the confidence band is $(4 \sqrt{2} \sqrt{2(n-k)} + 1/\sqrt{\alpha}) \sigma$. If we choose $k = n^{1/4} + \Delta$, $\Delta \leq 1/2$, then the width is
minus a smaller order term. Thus the width tends to zero (and we have consistency) provided $\frac{1}{2} < \Delta < 1$.

To use the band, we need an estimate of $\sigma^2$. A natural estimate is the residual mean square

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2 \quad (3.2)$$
4. Bounds based on asymptotic distribution theory.

We assume \( \sigma^2(x_i) = \sigma^2 < \infty \). Since the \( \tilde{S}_i \), \( i = 1,2,\ldots \) are partial sums of independent identically distributed random variables, we may apply a result for Brownian motion (e.g., Billingsley (1968), p. 72) which yields

\[
P(A) = P\left( \max_{0 \leq i \leq n-k} (-\tilde{S}_{i+k}) < \frac{1}{2} t \right) \approx 2\Phi\left( \frac{\frac{1}{2} t \cdot k / (\sqrt{n} \cdot \sigma)}{\sqrt{n}} \right) - 1.
\]

Similarly we get \( P(B) \approx 2\Phi\left( \frac{\frac{1}{2} t \cdot k / (\sqrt{n} \cdot \sigma)}{\sqrt{n}} \right) - 1 \). As in Section 3 we find that asymptotically a level \((1-\alpha)\) confidence band for \( \hat{E}(x) \) is

\[
\hat{m}(x) + 2\sigma^{-1}(1-\alpha/4)\sigma \cdot n^{k/k}.
\]  

Choosing \( k = n^{1/2+\Delta} \), the width of the band is

\[
4\sigma^{-1}(1-\frac{1}{4}\alpha)n^{-\Delta} \sigma.
\]

The following table compares the widths of the confidence bands (3.1) and (4.1). The widths of the bands are of the form \( 2c \sigma n^{-\Delta} \).

Table 4.1. Values of \( c \) for three frequently used confidence coefficients.

<table>
<thead>
<tr>
<th>Values of ( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band</td>
</tr>
<tr>
<td>Exact (3.1)</td>
</tr>
<tr>
<td>Asymptotic (4.1)</td>
</tr>
</tbody>
</table>
5. Bounds based on nonoverlapping neighborhoods.

The bands in Sections 3 and 4 will be of use only for large data sets. In this section, we develop a band which is much narrower, but it is simultaneous only for a sequence $t_{n1}, \ldots, t_{nl}$ of x-values. The model is

$$Y_{ni} = m(x_{ni}) + \epsilon_{ni}, \quad x_{1n} < \cdots < x_{nn}$$

where $\epsilon_{n1}, \ldots, \epsilon_{nl}$ are independent with $\text{Var}(\epsilon_{ni}) = \sigma^2(x_{ni})$ and

$$\max_j |x_{n,j+k} - x_{nj}| = O(n^{-\Delta/2}), \quad 0 < \Delta < \frac{1}{2}.$$ 

If we choose $t_{n1}, \ldots, t_{nl}$ so that

$$\max_j |t_{n,j+1} - t_{nj}| = O(n^{-\lambda}), \quad 0 < \lambda < \frac{1}{2} - \Delta$$

then for $n$ large enough, there will be no overlap between the $k = n^{1+\Delta}$ nearest x-neighbors to the points $t_{n1}, \ldots, t_{nl}$. Thus if we define

$$T_{ni} = \hat{m}(t_{ni}) - E\hat{m}(t_{ni}), \quad i = 1, \ldots, l$$

then there exists $N$ such that $T_{n1}, \ldots, T_{nl}$ are independent for all $n \geq N$. By Chebychev's inequality and (2.4), for $n \geq N$, $a > 0$,

$$P(\max_i |T_{ni}| \leq a) \geq \frac{1}{\sigma^2(x_{ni})} \left[ 1 - (ak)^{-2} \sum_{j \neq i+1} \sigma^2(x_{nj}) \right].$$

If we assume $\sigma^2(x) \equiv \sigma^2$ then

$$\hat{m}(x) + \frac{\sigma}{\sqrt{k[1 - (1-\alpha)^{-2}]}},$$

is a simultaneous confidence band for $E\hat{m}(x)$ valid for all $x \in \{t_{n1}, \ldots, t_{nl}\}$. The width of this band is of the order $O(n^{-\frac{1}{2}-\Delta}).$

By (2.4) and the Central Limit Theorem,
\[
\lim_{n \to \infty} P(\sqrt{k} T_{nj} \leq t) = \phi(t/\sigma),
\]
where \( k = n^{1 \Delta} \), and \( 1 \) is finite. Thus if we set

\[
M_1 = \max\{ |T_{n1}|, \ldots, |T_{n1}| \}
\]
then

\[
\lim_{n \to \infty} P(\sqrt{k} M_1 \leq t) = \left[ 2\phi(\sqrt{k} t/\sigma) - 1 \right]^1
\]
and

\[
\hat{m}(x) = \phi^{-1}(\frac{1}{\sqrt{2} \log(1 - \alpha)} \sigma/\sqrt{k})
\]
is asymptotically a level \((1 - \alpha)\) simultaneous confidence band for \( E\hat{m}(x) \) valid for \( x \in \{t_{n1}, \ldots, t_{n1}\} \).

This band is of order \( O(n^{-\frac{1}{2}} - \frac{1}{2} \Delta) \). Note that (5.1) and (5.3) are considerably narrower than the bands (3.1) and (4.1).

We next derive approximations to (5.2) and (5.3) valid for large \( n \).

Let \( V_1 = \max_i \{ T_{nj} \}, \ W_1 = \min_i \{ T_{nj} \} \), then \( M_1 = \max\{ V_1, -W_1 \} \) and

\[
P(\sqrt{k} M_1 \leq t) = P(\sqrt{k} V_1 \leq t, -\sqrt{k} W_1 \leq t).
\]

Using this and results on the asymptotic distribution of extreme order statistics (e.g. Galambos, p. 65 and p. 106), we find that if

\[
a_1 = \phi^{-1}(\frac{1}{1 + \frac{1}{2} \log 1 + \log 4}) \quad \text{or} \quad a_1 = (2 \log 1)^{\frac{1}{2}} - \frac{1}{2} (\log \log 1 + \log 4) \quad \text{and} \quad b_1 = 1/(2 \log 1)^{\frac{1}{2}}
\]
and \( b_1 = 1/(2 \log 1)^{\frac{1}{2}} \), then

\[
\lim_{l \to \infty} \lim_{n \to \infty} P(\sqrt{k} M_1 \leq a_1 + b_1 z) = \exp(-2e^{-z})
\]
(5.4)

It follows that for large \( n \), an approximation to (5.3) is
\[ \hat{m}(x) = \sigma \left( a_1 + b_1 z_\alpha \right) / \sqrt{k} \]  
(5.5)

where \( z_\alpha = -\log \left( -\frac{1}{2} \log (1 - \alpha) \right) \).

It would have been more elegant to take the limit in (5.4) as \( l \) and \( n \) simultaneously tend to \( \infty \), say by setting \( l = n^\gamma \), \( 0 < \gamma \leq \frac{1}{2} - \Delta \). With \( \gamma = \frac{1}{2} - \Delta \), this would lead to a band similar to that of Révész (1979).

His estimator differs slightly from \( \hat{m}(x) \) in that his index sets, say \( I_{nk}(x) \), are balanced with \( \frac{1}{2} k \) values of \( x_{n1}, \ldots, x_{nn} \) closest to and less than \( x \), and \( \frac{1}{2} k \) values closest to and greater than \( x \), \( k \) even. Nevertheless, Theorem 1 of Révész (1979) holds for \( \hat{m}(x) \) based on the sets \( I_{nk}(x) \), yielding the confidence band

\[ \hat{m}(x) = \sigma \left( a_s + b_s z_\alpha \right) / \sqrt{k} \]  
(5.6)

with \( s = n^{1/2 - \Delta} \). The width of this band is of order \( O(n^{-1/2 - \Delta} \log n) \)

when \( k = n^{1/2 + \Delta} \). Note that in (5.1), (5.3) and (5.5) as well as (3.1) and (4.1) we have avoided a number of regularity conditions required by Révész.

The band (5.1) can be made asymptotically valid for all \( x \) provided the \( t_{nj} \)'s are chosen dense in the set \( X \) of possible \( x \)'s, that the bias is of smaller order than the widths of the bands (see Section 6), and that \( m(x) \) is uniformly continuous on \( X \).
6. **Bias.**

The bands of the previous sections are for

\[ \tilde{m}(x) = E\tilde{m}(x) = \sum_{i \in I_{nk}(x)} m(x_i)/k. \]

In order to make them valid for \( m(x) \), we need to show that the bias

\[ m(x) - \tilde{m}(x) = \sum_{i \in I_{nk}(x)} [m(x_i) - m(x)]/k \]

is uniformly of smaller order than the width of the bands.

Assume now that \( x \) is in an interval \([a, b]\) and that the regression function satisfies a \( r \)th order Lipschitz condition:

\[ |m(x) - m(y)| \leq c|x - y|^r, \quad x, y \in \mathbb{R}. \] (6.1)

Then,

\[ |\tilde{m}(x) - m(x)| \leq c \cdot \sum_{j \in I_{nk}(x)} |x_j - x|^r/k \]

\[ \leq c \cdot \max \left\{ \left( \frac{x_{i+k+1} - x_i}{2} \right)^r, \left( \frac{x_{i+k+1} - x_{i+k+1}}{2} \right)^r \right\} \]

\[ \leq c \cdot (x_i + x_{i+k+1} - x_i)^r/2^r, \]

when \( x \in J_i = ((x_i + x_{i+k})/2, (x_{i+k} + x_i + k+1)/2], \ i = 0, 1, \ldots, n-k, \)

\( x_0 = a, \ x_{n+1} = b. \)

Thus when \( k = n^{1/2} + \Delta \) and \( \max_j |x_n, j+k-1 - x_n, j| = 0(n^{-\Delta}), \ 0 < \Delta < \frac{1}{2}, \)

where now \( x_i \) and \( x_{n1} \) are used interchangeably, then

\[ \sup_x |\tilde{m}(x) - m(x)| = 0(n^{-r(1/2-\Delta)}) \] (6.2)

and the bias is of smaller order than the width of the bands (3.1) and (4.1) when \( n^{-\Delta} > n^{-r(1/2-\Delta)}, \) i.e. \( \Delta < r/2(r+1). \) Thus, if we are only
willing to assume a first order Lipschitz condition, we must choose \( \Delta \) less than \( \frac{1}{4} \). Second and third order conditions lead to \( \Delta \) less than \( \frac{1}{3} \) and \( \frac{3}{8} \), respectively.

Turning to the bands (5.1) and (5.3), we need \( n^{-\frac{1}{2}} - \frac{1}{2} \Delta > n^{-r(\frac{1}{2} - \Delta)} \), i.e. \( \Delta < \left(\frac{1}{2} r - \frac{1}{4}\right)\left(\frac{1}{2} + r\right)^{-1} \). For \( r = 1, 2 \) and 3, we find \( \Delta \) less than \( \frac{1}{6}, \frac{3}{10} \) and \( \frac{5}{14} \), respectively. Since the bands (5.5) and (5.6) are at least as wide as (5.1) and (5.3), the same restriction is sufficient to make the bias of uniformly smaller order than the widths.

Révész (1979) considers a different restriction of the regressor to the interval \([0, 1]\): Let \( X \) have a density \( f \) such that \( f(x) \geq \lambda \), \( x \in [0, 1] \), for some \( \lambda > 0 \). Let \( k \) be such that

\[
kn^{-2/3} \log n \to 0, \quad k^{-1}(\log n)^3 \to 0 \text{ as } n \to \infty \quad (6.3)
\]

Then (Révész (1979), Lemma 1)

\[
\limsup_{n \to \infty} \frac{1}{k} \sup_{1 \leq i \leq n-k} |X_i - X_{i+k}| \leq 2/\lambda \quad \text{a.s.}
\]

From this and the \( r \)th order Lipschitz condition (6.1), it follows that if \( k = n^{\frac{1}{2}} + \Delta \),

\[
\sup_{x} |\tilde{m}(x) - m(x)| = o(n^{-r(\frac{1}{2} - \Delta)}) \quad \text{a.s.}
\]

This result is valid for \( X \in [a, b] \), not just \([0, 1]\), thus we have shown (6.2) again under a different set of conditions. Note that (6.3) implies that if \( k = n^{\frac{1}{2}} + \Delta \), then \( \Delta < \frac{1}{6} \).

Under the conditions of the above paragraph, and assuming that \( m(x) \) has a uniformly bounded derivative, Révész has shown (Lemma 2) that

\[
\sup_{x} |\tilde{m}(x) - m(x)| = o((k \log n)^{-\frac{1}{2}}) \quad \text{a.s.}
\]
which again implies that the bias is of smaller order than the width of the band.

Rosenblatt (1969) gives asymptotic results for kernel estimators of the regression function. He gets pointwise confidence-intervals of width $n^{-2/5}$ using a bandwidth of order $n^{-1/5}$ which corresponds to $k = n^{4/5}$ or $\Delta = .3$.

Spiegelman and Sacks (1980) obtain window estimators with mean squared error of the order $O(n^{-2/3})$ by imposing a first order Lipschitz condition on $m$. Thus their bias is of order less than $n^{-1/3}$. Their bandwidth is of order $b_n = n^{-1/3}$, corresponding to $k = n^{2/3}$. 
7. An illustration.

To get an idea of the accuracy of the bands, we computed the band (5.6) for data \((x_1, y_1) \cdots (x_{100}, y_{100})\) generated from the model

\[
y_i = \beta_1 \beta_2 x_i^{\beta_3} + e_i
\]

where \(\beta_1 = 5, \beta_2 = -\frac{1}{2}, \beta_3 = 1, e_1, \ldots, e_n\) are i.i.d. \(N(0, 100), x_i = i/25, i = 1, \ldots, 100,\) the significance coefficient is .90, \(\hat{\sigma}\) is computed from (3.2), and \(\Delta = .15.\)

This model has been suggested for agricultural experiments where an amount \(x\) of fertilizer increases yield \(Y\) for low and moderate doses while it decreases yield at high doses.

The result is shown in Figure 1 where the middle curve is the estimate \(\hat{m}(x)\) and the upper and lower curves define the band. The band is fairly accurate with the width being 11.5. Since the band is simultaneous, we can test model assumptions. For instance, since no line fits in the confidence band, regression linear in \(x\) is rejected. Similarly, a parabola does not fit in the band and a quadratic (in \(x\)) regression model is also rejected.

We also computed the widths of the other bands (using \(\sigma = 10\) rather than \(\hat{\sigma}\)). The results are given in the table below using \(k = 20\) and \(l = s = 5.\) Note that the widths of (5.5) and (5.6) are the same.

<table>
<thead>
<tr>
<th>Band</th>
<th>(3.1)</th>
<th>(4.1)</th>
<th>(5.1)</th>
<th>(5.3)</th>
<th>(5.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width</td>
<td>89.6</td>
<td>39.3</td>
<td>31.0</td>
<td>10.3</td>
<td>11.6</td>
</tr>
</tbody>
</table>

Note that since \(e_i\) is normal, (5.3) is exactly a level .90 simultaneous confidence procedure for \(l = 5.\)
Figure 1. The estimate \( \hat{m}(x) \) and 90% simultaneous confidence band for the model (7.1).
Acknowledgement

We wish to thank Professor J. Mac Queen for generously spending his time discussing the topic of this paper.

REFERENCES


Steinar Bjerve  
Matematisk Institutt  
P.O. Box 1053, Blindern  
Oslo 3, Norway  

Kjell A. Doksum  
Department of Statistics  
University of California  
Berkeley, CA 94720  
U.S.A.  

Brian S. Yandell  
Statistics Department  
University of Wisconsin  
1210 West Dayton St.,  
Madison, WI 53706  
U.S.A.