Confidence Intervals for Nonparametric Regression

Lawrence D. Brown, Xin Fu and Linda H. Zhao

Department of Statistics

University of Pennsylvania

Abstract

In non-parametric function estimation, providing a confidence interval with the right coverage is a challenging problem. This is especially the case when the underlying function has a wide range of unknown degrees of smoothness. Here we propose two methods of constructing an average coverage confidence interval built from block shrinkage estimation methods. One is based on the James-Stein shrinkage estimator; the other begins with a Bayesian perspective and is based on a modification of the harmonic prior estimator. Simulation shows that these confidence intervals have average coverage close to or above the nominal coverage even when the underlying function is rough and/or the signal to noise ratio is small. Both of the confidence intervals perform consistently well across all the investigated test functions even through these functions have very different shapes and smoothness.

Keywords: Blockwise estimators; Confidence interval; Harmonic prior; James-Stein; Level of average coverage; Nonparametric estimation.

AMS 2000 Subject Classification: Primary: 62G15; Secondary: 62F15, 62F99.

1 Introduction

Confidence intervals play a central inferential role in non-parametric function estimation. Many attempts have been made to construct confidence intervals based on different estimation methods, for example, Wahba (1983) and Nychka (1988) on smoothing spline, Eubank and Speckman (1993), Hardle and Marron (1991) on kernel estimation, Xia (1998) on local polynomial estimation with corrected bias, Hall (1986,1988,1993) and Efron (1987) on bootstrap confidence intervals. More recently, Barber, Nason and Silverman (2002) proposed a method of constructing confidence intervals for a wavelet thresholding method. Mao and Zhao (2003) studied confidence intervals based on free-knot polynomial splines.

However, constructing a confidence interval with right coverage and optimal average length remains a challenging problem. This is especially the case when the underlying function has a wide range of unknown degrees of smoothness. We propose two methods of constructing confidence intervals. Both of them are based on block-wise shrinkage estimators. Such estimators have received a lot of attention recently in the non-parametric function estimation context. See for example Donoho and Johnstone (1994, 1995a and 1995b), Kerkyacharian, Picard and Tribouley (1996), Cavalier,Golubev, Picard and Tsybakov (2002) and Cai, Low and Zhao (2000). It has been shown by Cavalier, et al., (2002) and Cai, Low and Zhao (2000) that the estimator of the underlying function can be sharply adaptive over a wide range of (Sobolev) smoothness classes.

We consider the nonparametric function estimation model,

$$Y_i = f(x_i) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \tau^2), x_i = i/n, \tag{1.1}$$

where $f(x)$ is supported on $[0, 1]$ and n is the number of data points. Our goal is to provide a $1 - \alpha$ level average coverage confidence interval so that

$$E\left(\frac{1}{n}\sum_{i=1}^{n} I\{f(x_i) \in CI(x_i)\}\right) = 1 - \alpha. \tag{1.2}$$

In words, we aim to construct a confidence interval so that on average $1 - \alpha$ percent of the function will be covered. This criteria has been one popular standard for nonparametric confidence interval construction; see for example, Wahba (1983).

The confidence interval problem is challenging mainly because it is difficult to directly measure the bias of the function estimate. The first confidence interval method we propose, which is based on the blockwise James-Stein shrinkage estimator, is a construction that takes account of combined effect of bias and variance. Basically instead of estimating variance and bias separately, we estimate mean square error loss and use it to construct the confidence interval. The other method involves a Bayesian point of view. We use modified harmonic priors on each blocks. Then we calculate the posterior variance of the function estimate and provide approximate $1-\alpha$ Bayesian credible intervals.

The performance of our confidence intervals will be demonstrated through numerical methods – comparisons with two other popular methods, confidence intervals based on the spline estimate of Wahba (1983) and Nychka (1988) and a variance band based on local polynomial estimators. We choose two different groups of test functions. One group contains test functions similar to those in the traditional confidence intervals literature. These functions are shown in panels (a)-(e) of Figure 3 in Section 4. The other group contains some of the popular test functions in the nonparametric function estimation literature, see panels (f)-(h) of Figure 3. The simulation studies in Section 4 have shown that both of our methods provide average coverage confidence intervals with coverage very close to nominal coverage for all the test functions and for a range of levels of signal to noise ratio.

The rest of the paper is organized as follows. In Section 2, shrinkage estimators through orthonormal transformation are discussed. Detailed confidence interval constructions are presented in Section 3. Comparison with other methods along with discussions is contained in Section 4. An application to call center data is also provided in Section 4.

2 Block-wise Shrinkage Estimators

2.1 Orthonormal Transformation

Suppose we have an orthonormal basis $\{\varphi_j(x_i) : j = 1, 2, \cdots, n\}$, i.e., $\frac{1}{n}\sum_{i=1}^{n}\varphi_j(x_i)\varphi_k(x_i) = \delta_{jk}$. This type of basis arises naturally when using either discrete wavelet transforms or discrete Fourier transforms.

Then we can write

$$f(x_i) = \sum_{j=1}^{n} \xi_j \varphi_j(x_i). \tag{2.3}$$

where $\xi_j, j = 1, \cdots, n$ are the coefficients. Similarly we can transform $\{Y_i\}$ accordingly. That is

$$W_j = \frac{1}{n} \sum_i Y_i \varphi_j(x_i).$$

Because $\{\varphi_j(x_i) : j = 1, 2, \cdots, n\}$ is orthonormal, we can easily see that

$$W_j \overset{ind}{\sim} N(\xi_j, \sigma^2 = \tau^2/n).$$

Hence, after the transformation, the non-parametric function estimation problem becomes a multivariate normal mean problem, i.e., we observe $W_j \overset{ind}{\sim} N(\xi_j, \sigma^2 = \tau^2/n)$ and we try to estimate ξ_j.

2.2 Blocking Scheme

A Blockwise shrinkage technique is used in estimating the multivariate normal means. Basically, we partition all the $\{\xi_j\}$ into $K(n)$ blocks, $\{B_1, B_2, \cdots, B_k, \cdots, B_{K(n)}\}$, where

$$B_k = \{l_k + 1, l_k + 2, \cdots, l_{k+1}\}.$$

The size of block B_k is $m_k = l_{k+1} - l_k$. We choose our block size to be $m_k = \lfloor b(n)^k \rfloor$, the integer part of the $b(n)^k$ and $k = 1, 2, \cdots, K$. Cavalier, et al., (2002) and Cai, Low and Zhao (2000) have shown that if blocks are asymptotically of size b^k with $b \to 1$ as $n \to \infty$, then the estimators for unknown function under integrated squared error loss are sharply adaptive over a wide range of (Sobolev) smoothness classes. We have used $b = 1/(1 + log(n))$ as suggested in Car, Low and Zhao (2000) in our simulations as well as in the real data set.

2.3 Two Shrinkage Methods

Let w_k denote the vector of observations within block B_k, that is $w_{k_i} = w_{l_k+i}, i = 1, \cdots, m_k$. We apply the shrinkage method within each block. Two shrinkage methods are considered. The first method involves the James-Stein plus estimator; namely for block B_k

$$\hat{\xi}_{JS^+{}_k} = \left(1 - \frac{c\sigma^2(m_k - 2)}{||w_k||^2}\right)_+ w_k,$$

here again m_k denotes the block size and c is a constant typically between 0 and 2. Larger c implies more shrinkage. In the simulation study below we choose $c = 1.5$. See Cai (1999) for motivation of such a choice. We refer to this blockwise shrinkage method as the blockwise James-Stein shrinkage method.

The second method is referred to as the blockwise harmonic plus shrinkage method. This method is based on the generalized Bayes estimator using the harmonic plus prior, which we will briefly refer to as the harmonic plus estimator. The harmonic plus prior H^+ is of the form $H + \beta\delta_{\{0\}}$, where $\delta_{\{0\}}$ denotes the point mass at 0. Here H denotes the harmonic prior with its density function $h(\xi) = 1/||\xi||^{m_k-2}$. Notice that $h(\xi)$ is not a proper density function because H is an improper prior; that is $\int h(\xi)d\xi = \infty$. The Harmonic prior is well known to have very desirable properties for the multivariate normal mean problem, see Stein (1981) for details. In particular the harmonic estimator is minimax. Moreover, as a direct implication of Brown (1971), the harmonic estimator is also admissible.

However, when ξ is close to 0, the harmonic prior has much less shrinkage effect compared with that of the James-Stein plus method. Hence the harmonic plus prior is constructed so that the corresponding Bayes estimator has a similar risk function to that of the James-Stein plus estimator with $c = 1$. $\beta\delta_{\{0\}}$ is added so that the prior will have more shrinkage effect when ξ is close to 0. β can of course be thought as the prior probability that $\xi = 0$. It is obvious that the larger β is, the more shrinkage the harmonic plus estimator has. We will discuss later in detail how to choose β. For comparison of the risk functions for the James-Stein plus with $c = 1$, harmonic and harmonic plus estimators, see Figure 1. Detailed study of this type of estimators can be found in Brown and Xu (2005).

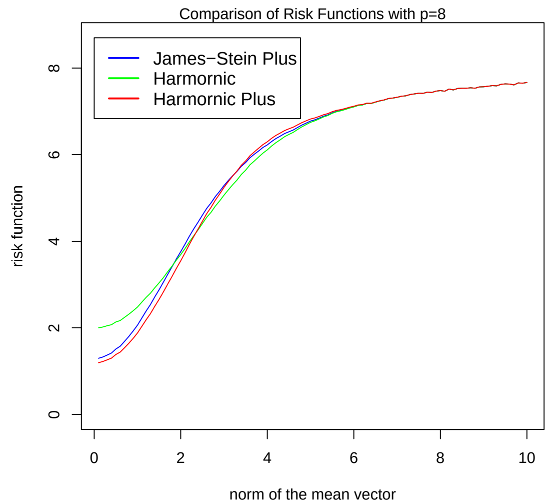


Figure 1: Comparison of Risk Function under L^2 loss for James-Stein plus, Harmonic and Harmonic plus Estimators. Risk function is computed using Monte Carlo simulation with 100,000 repetitions.

For the blockwise harmonic plus estimator, we put independent versions of the harmonic plus prior H^+ on each block to obtain the estimator for ξ.

We will now give the analytical form of the Bayes estimator of the harmonic plus prior. To simplify the notation, we will use m instead of m_k for the block size of B_k, and omit the subscripts on w_k, etc.

Let $g^*_G(x)$ denote the marginal density function of x under prior G and $\hat{\xi}_G$ the Bayes estimate under prior G. Then by Brown(1971) we can write

$$\hat{\xi}_G = (1 - \rho_G(||w||))w, \tag{2.4}$$

where $\rho_G(||w||) = -\sigma^2 \frac{\frac{\partial}{\partial ||w||}(log(g^*_G(w)))}{||w||}$. Brown and Xu (2005) give the following expressions for g^*_H.

1) When $m \geq 4$ even,

$$g^*_H(w) = \left(\frac{1}{\sigma}\right)^{m-2} \left(\frac{1}{\sqrt{2\pi}}\right)^m ||y||^{2-m} \left(1 - P_{||y||^2/2}\left(\frac{m}{2} - 2\right)\right) \tag{2.5}$$

where $y = w/\sigma$ and $P_{||w||^2/2}(\cdot)$ is the cdf of the Poisson distribution with rate equals to $||w||^2/2$.

2) When $m \geq 3$ odd,

$$g_H^*(w) = \left(\frac{1}{\sigma}\right)^{m-2}\left(\frac{1}{\sqrt{2\pi}}\right)^m ||y||^{2-m}\left(\Psi(||y||) - e^{-||y||^2/2}W_p(||y||)\right) \tag{2.6}$$

where $y = w/\sigma$, $\Psi(s) = \sqrt{2\pi}(\Phi(s) - 1/2)$ and $W_m(s) = \sum_{j=0}^{(m-1)/2-1} \frac{2^j j!}{(2j+1)!} s^{2j+1}$. Here $\Phi(\cdot)$ is the cdf of standard normal distribution.

It is easy to see that for harmonic plus, we have

$$g_{H^+}^*(w) = g_H^*(w) + \beta f_{\xi=0}(w), \tag{2.7}$$

where $f_\xi(w) = \frac{1}{\sqrt{2\pi\sigma^{2m}}} exp(-||w - \xi||^2/2\sigma^2)$ is the density of w given ξ. Therefore, using equation (2.5), (2.6) and (2.7), one can find that

1) When $m \geq 4$ even,

$$\begin{aligned}\rho_{H^+}(||w||) &\equiv -\sigma^2 \cdot \frac{\frac{\partial}{\partial ||w||}(\log(g_H^*))}{||w||} \\ &= \frac{\sigma^2(m-2)(1 - P_{||y||^2/2}(m/2-1)) + \beta||y||^m e^{-||y||^2/2}}{\sigma^2||y||^2(1 - P_{||y||^2/2}(m/2-2)) + \beta||y||^m e^{-||y||^2/2}},\end{aligned} \tag{2.8}$$

where $y = w/\sigma$ and $P_{||w||^2/2}(\cdot)$ is the cdf for the Poisson distribution with rate equal to $||w||^2/2$.

2) When $m \geq 3$ odd,

$$\begin{aligned}\rho_{H^+}(||w||) &\equiv -\sigma^2 \cdot \frac{\frac{\partial}{\partial ||w||}(\log(g_H^*))}{||w||} \\ &= \frac{\sigma^2(m-2)(\Psi(||y||) - e^{-||y||^2/2}W_{m+2}(||y||)) + \beta||y||^m e^{-||y||^2/2}}{\sigma^2||y||^2(\Psi(||y||) - e^{-||y||^2/2}W_m(||y||)) + \beta||y||^m e^{-||y||^2/2}},\end{aligned} \tag{2.9}$$

where $y = w/\sigma$, $\Psi(s) = \sqrt{2\pi}(\Phi(s) - 1/2)$ and $W_m(s) = \sum_{j=0}^{(m-1)/2-1} \frac{2^j j!}{(2j+1)!} s^{2j+1}$.

In terms of choice of β we slightly modify the basic idea in Brown and Xu (2005). They proposed that one choose β so that the posterior probability of $\xi = 0$ given $||y||^2 = m - 2$ is 1/2, i.e., $P(\xi = 0| \parallel y \parallel^2 = m - 2) = 1/2$. In other words, β can be chosen so that the marginal density coming from the harmonic prior equals to the marginal density coming from the point mass prior when $||y||^2 = (m - 2)$. The risk performance of the Bayes estimator with the chosen β can be found in Figure 1.

Because we judge it more desirable to introduce additional shrinkage, as in Cai, Low and Zhao (2000) we choose β so that $P(\xi = 0|||y||^2 = c(m - 2)) = 1/2$, where $c = 1.5$.

Thus, when β satisfies the condition $P(\xi = 0| \parallel y \parallel^2 = c(m - 2)) = 1/2$, it is given by

1) when $m \geq 4$ even,

$$\beta = \sigma^2 e^{(c(m-2))/2} (c(m-2))^{(2-m)/2} \left(1 - P_{(c(m-2))/2}\left(\frac{m}{2} - 2\right)\right).$$

2) when $m \geq 3$ odd,

$$\beta = \sigma^2 e^{(c(m-2))/2} (c(m-2))^{(2-m)/2} \left(\Psi(\sqrt{c(m-2)}) - e^{-(c(m-2))/2} W_m(\sqrt{c(m-2)})\right).$$

See Figure 2 for comparison of risk functions.

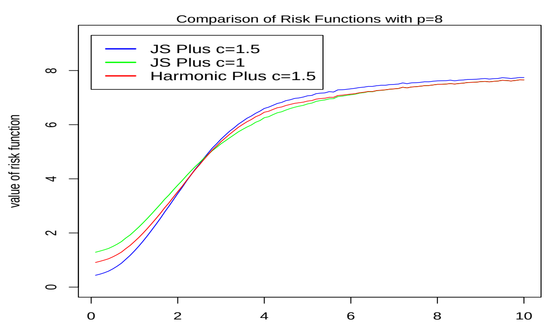


Figure 2: Comparison of Risk Function for JS Plus with c=1.5, c=1 and and Harmonic plus Estimators with c=1.5. Risk function is computed using Monte Carlo simulation with 100,000 repetitions.

3 Construction of Confidence Interval

3.1 Confidence Intervals using the Harmonic Plus Estimators

Here we introduce our method of constructing confidence intervals. First, we describe the construction that is based on the harmonic plus estimator.

We produce the average coverage band by constructing $1-\alpha$ level confidence intervals for

$$f(x_i) = \sum_j \xi_j \varphi_j(x_i), \forall i.$$

Since the Harmonic plus estimator is Bayesian, it might be thought of as desirable to use the HPD for the $f(x_i)$. HPD regions for the harmonic or harmonic plus are hard to compute and have undesirable geometric form. But it is reasonable and desirable to assume approximate normality for the posterior distribution. Such an approximation is also suggested in Berger (1980). In order to approximate the posterior we need only to know the posterior mean and variance. The mean is given via (2.5) and (2.6). We thus only need to find the posterior variance of $f(x_i)$.

Since the prior distributions for different blocks are independent, the posterior distribution of $\xi|w$ are independent across different blocks. Hence, the posterior variance of $f(x)$ given w can be written as

$$var(f(x_i)|w) = var(\sum_j \xi_j \varphi_j(x_i)|w) = \sum_k var(\sum_{j \in B_k} \xi_j \varphi_j(x_i)|w_{B_k}),$$

where B_k denotes kth block and $w_{B_k} = \{w_i : i \in B_k\}$.

To find out the posterior variance $var(\sum_{j \in B_k} \xi_j \varphi_j(x_i)|w_{B_k})$, let us first rewrite it as

$$var(a_{B_k} \cdot \xi_{B_k}|w_{B_k}),$$

where

$$a_{B_k} = \{\varphi_j(x_i) : i \in B_k\}, \;\; \xi_{B_k} = \{\xi_i : i \in B_k\}.$$

To simplify the notation, we will suppress subscripts in the following lemma:

Lemma 3.1 *Let $W \sim N_m(\xi, \sigma^2 I_m)$ and let ξ have a spherically symmetric prior G, and a be any $m-$dimensional vector. Then*

$$var(a \cdot \xi | w) = \sigma^2 ||a|| \left(1 - \rho'(||w||) \frac{a \cdot w}{||a|| \, ||w||} - \rho(||w||)\right), \tag{3.10}$$

where

$$\rho(||w||) = -\sigma^2 \cdot \frac{\frac{\partial}{\partial ||w||}(\log(g_G^*))}{||w||}.$$

Proof: Let $d = a / \parallel a \parallel$. There exists an orthonormal matrix Q such that $d'Q' = e_1 = (1, 0, 0, \cdots, 0)$. Therefore,

$$var(a \cdot \xi | w) = \parallel a \parallel var(d'Q'Q\xi | w) = ||a|| e_1 var(Q\xi | w) e_1',$$

which is just the posterior variance for the first component of $Q\xi$. Since Q is orthonormal, $Qw \sim N(Q\xi, \sigma^2 I_m)$. And $(Qw)_1 = d'w$. By Brown (1971), we know

$$var((Q\xi)_1 | w) = \sigma^2 \left(1 - \rho'(||Qw||) \frac{(Qw)_1}{||Qw||} - \rho(||Qw||)\right),$$

where $\rho'(t) = \frac{\partial \rho(t)}{\partial t}$. After plugging in $d = a/||a||$ and noticing that $||Qw|| = ||w||$, we obtained the desired result. ∎

Then the $1 - \alpha$ level confidence interval for the harmonic plus estimator is of the form

$$\hat{f}(x_i) \pm z_{1-\alpha/2} \sigma \sqrt{\sum_k ||a_{B_k}||^2 \left(1 - \rho'_{H^+{}_k} \frac{a_{B_k} \cdot w_{B_k}}{||a_{B_k}|| \, ||w_{B_k}||} - \rho_{H^+{}_k}\right)}, \tag{3.11}$$

where $z_{1-\alpha/2}$ denotes the $1 - \alpha/2$th quantile of a standard normal distribution, $\rho'_{H^+{}_k(t)} = \frac{\partial \rho_{H^+{}_k}(t)}{dt}$ and the detailed formula is given in appendix.

3.2 Confidence Interval using James-Stein Plus Estimators

To produce an average coverage confidence interval based on the James-Stein plus estimator, we again try to construct a $1-\alpha$ confidence interval for each $f(x_i)$. However, since the James-

Stein plus estimator is biased, instead of adding of the variances across all the blocks, we add an estimate of the mean square errors across all the blocks to compensate for bias. That is, we approximate

$$\begin{aligned} E(\hat{f}(x_i) - f(x_i))^2 &\equiv E\left(\sum_{i=1}^{n}(a_i\hat{\xi}_i - a_i\xi_i)^2\right) \\ &\approx \sum_{k=1}^{K} E(\sum_{i\in B_k}(a_i\hat{\xi}_i - a_i\xi_i)^2 \end{aligned}$$

Our confidence interval will be of the form:

$$\hat{f}(x_i) \pm z_{1-\alpha/2}\sqrt{\sum_{k=1}^{K}\hat{E}(a_{B_k}\hat{\xi_{B_k}} - a_{B_k}\xi_{B_k})^2}. \tag{3.12}$$

In order to estimate $E(a_{B_k}\hat{\xi_{B_k}} - a_{B_k}\xi_{B_k})^2$ for each block k, we can rotate coordinates as in the proof of lemma 3.1. If we let $\zeta = Q\xi$ then we can rewrite $E(a'\hat{\xi} - a'\xi)^2$ as

$$||a||d'Q'E(Q\hat{\xi} - Q\xi)^2Qd = ||a||d'Q'E(\widehat{Q\xi} - Q\xi)^2Qd = ||a||E(\hat{\zeta}_1 - \zeta_1)^2,$$

where again $d = \frac{a}{||a||}$, and $d'Q' = e_1 = (1, 0, 0, \cdots, 0)$.

To estimate $E(\hat{\zeta}_1 - \zeta_1)^2$, let us look at the first order expansion of $\hat{\zeta}_1 - \zeta_1$. Consider an estimator of the form $\hat{\zeta}_1 = (1 - \rho(||z||))z_1$, where $z_i \overset{ind}{\sim} N(\zeta_i, \sigma^2)$. Then the first order expansion of $\hat{\zeta}_1 - \zeta_1$ is

$$\begin{aligned} \hat{\zeta}_1 - \zeta_1 &= z_1 - \zeta_1 - \rho(||z||)z_1 \\ &\approx z_1 - \zeta_1 - \rho(||\zeta||)\zeta_1 - \sum_{j=1}^{m}\frac{\partial}{\partial z_j}(\rho(||z||)z_1)\bigg|_{z=\zeta}(z_j - \zeta_j) \\ &= z_1 - \zeta_1 - \rho(||\zeta||)\zeta_1 - (z_1 - \zeta_1)\left[\rho'(||\zeta||)\frac{\zeta_1^2}{||\zeta||} + \rho(||\zeta||)\right] \\ &- \sum_{j=2}^{m}(z_j - \zeta_j)\rho'(||\zeta||)\frac{\zeta_1\zeta_j}{||\zeta||} \end{aligned} \tag{3.13}$$

Thus,

$$\begin{aligned} &E(\hat{\zeta}_1 - \zeta_1)^2 \\ \approx\ &\sigma^2[(1 - \rho(||\zeta||) - ||\zeta||\rho'(||\zeta||)\Upsilon_1^2(\zeta))^2 + \Upsilon_1^2(1 - \Upsilon_1^2)(||\zeta||\rho'(||\zeta||))^2] + \rho^2\Upsilon_1^2||\zeta||^2, \end{aligned}$$

where $\Upsilon_1^2 = \frac{\zeta_1^2}{||\zeta||^2}$. Here, both $||\zeta||^2$ and Υ_1^2 are unknown and need to be estimated. The estimate we use here are $(||z||^2 - m)_+$ for $||\zeta||^2$ and $z_1/||z||$ for Υ_1.
Substitute Qw_{B_k} into z to get

$$
\begin{aligned}
& E(a_{B_k}\hat{\xi_{B_k}} - a_{B_k}\xi_{B_k})^2 \approx \hat{E}(a_{B_k}\hat{\xi_{B_k}} - a_{B_k}\xi_{B_k})^2 \\
= \; & \sigma^2 \left[\left(1 - \rho_{H^+} - \sqrt{(||w_{B_k}||^2 - m_k)_+} * \rho_{JS^+} \frac{(a_{B_k} \cdot w_{B_k})^2}{||w_{B_k}||^2} \right)^2 \right. \\
+ \; & \left. \frac{(a_{B_k} \cdot w_{B_k})^2}{||w_{B_k}||^2} (1 - \frac{(a_{B_k} \cdot w_{B_k})^2}{||w_{B_k}||^2})(||w_{B_k}||^2 - m_k)_+ (\rho'_{JS^+})^2 \right] \\
+ \; & (\rho_{JS^+})^2 \frac{(a_{B_k} \cdot w_{B_k})^2}{||w_{B_k}||^2} (||w_{B_k}||^2 - m_k)_+, \qquad (3.14)
\end{aligned}
$$

This expression along with (3.12) describes the relevant intervals.

3.3 Estimating Variance

So far, we have described two constructions for non-parametric function estimator based on blockwise shrinkage methods, assuming that we know the variance τ. In applications, τ^2 is not usually known, however it can be estimated quite accurately and satisfactorily by using the difference estimator proposed by Rice (1984). That is

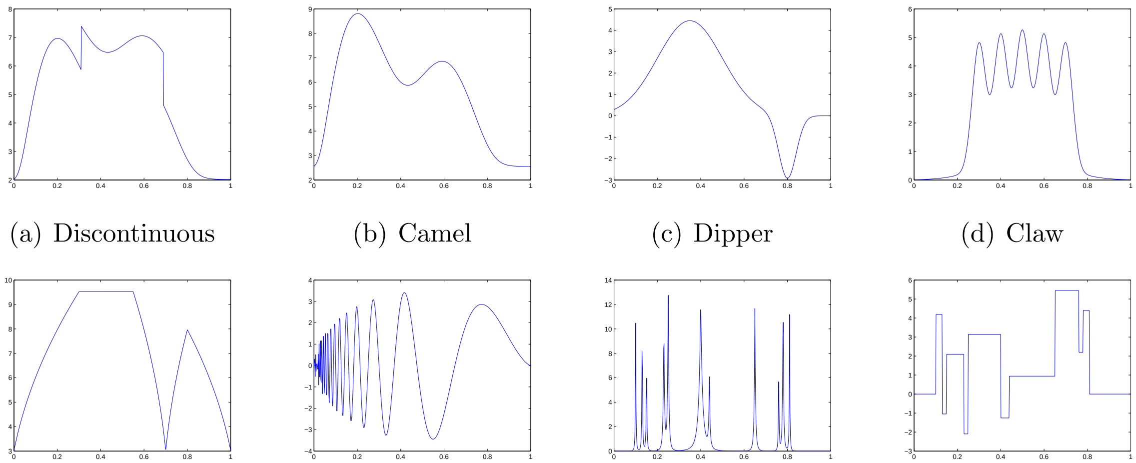
$$\hat{\tau}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2. \qquad (3.15)$$

4 Empirical Results

4.1 Simulation Results

We use 8 functions to test the performance of the confidence intervals constructed based on the James-Stein plus and the harmonic plus estimator. These functions can be divided into two groups. One group contains test functions similar to those in the traditional confidence intervals literature. These functions are shown in panels (a)-(e) of Figure 3. The other group contains some of the popular test functions in the nonparametric function estimation literature; see panels (f)-(h) of Figure 3.

For each function, we use 4 different values of signal to noise ratio (STNR), 1, 4, 16,



(a) Discontinuous (b) Camel (c) Dipper (d) Claw

(e) Corner (f) Doppler (g) Bumps (h) Blocks

Figure 3: Plots of Test Functions for the simulations

64. Here signal to noise ratio is defined as the ratio of variance of the true function to the variance of the noise τ^2 ,

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(f(x_i)-\overline{f})^2}{\tau^2}.$$

It is easy to see that the larger the signal to noise ratio is, the easier the estimation problem is. Therefore, given the same underlying function, the average length for the confidence interval should be narrower with larger signal to noise ratio. See Figure 4 for illustration of the discontinuous function with the various signal to noise ratios.

In the simulation we used the discrete wavelet transform with wavelet sym8. This is a Symmlet with 8 parameters, 7 vanishing moments and support length 15. Detailed discussion about this wavelet can be found in Daubechies (1993). Periodic extension of the data is used in the implementation.

Table 1 contains the main information about the coverage and average width of different confidence intervals. The average coverage and the associated average width are obtained by averaging 1000 independent simulations. The sample size is 512. We compare our confidence interval with two other methods. The first one is based on the smoothing spline estimator, proposed and studied by Wahba (1983) and Nychka (1988). They view the smoothing spline estimator as a Bayes estimator with dependent normal priors and construct Bayesian poste-

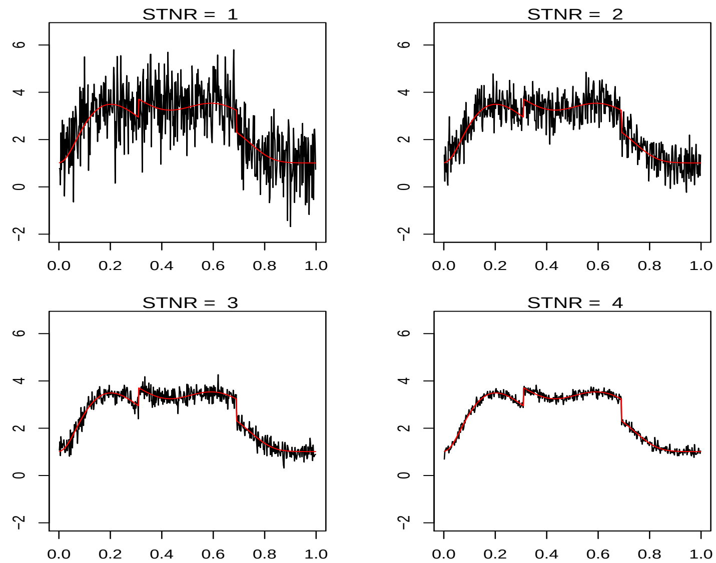


Figure 4: Discontinuous Function with Different STNR.

rior confidence intervals. The penalty constant is selected by leave-one-out cross-validation. The resultant confidence intervals can thus be viewed as empirical Bayes HPD regions. The other method is based on the local linear estimator. The variance given the bandwidth is calculated for the estimator and the confidence interval is based on its variance only, as if the estimator were unbiased. For bandwidth selection, we use the direct plug-in method proposed by Ruppert, Sheather and Wand (1995). Since the confidence interval for local linear estimator is just a variance band with no bias correction, we can expect that its average length is far narrower than that of other confidence intervals, but that its coverage may fall below the nominal value.

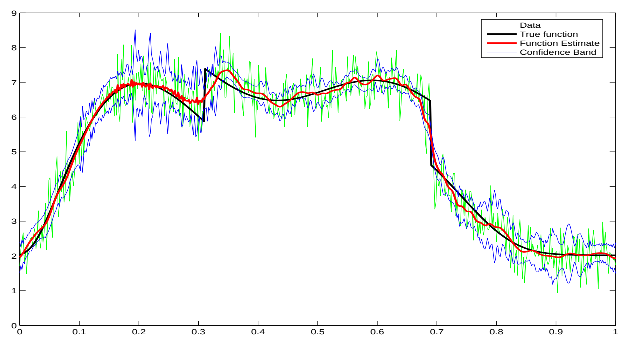
For a clean view of Table 1, we decide not to include the simulation standard errors. Moreover, the standard errors are all small. For average coverage, the standard errors are all less than 0.4 and most of them are less than 0.1. For the average length, the standard errors are all less than 0.01. All the standard errors are decreasing when the signal to noise ratio becomes larger, with the underlying function and construction method fixed.

From Table 1, we can see that the coverage of the James-Stein plus method and the harmonic plus method are consistently around 95% with acceptable average length across all the test functions and all levels of signal to noise ratio. For the local linear estimator,

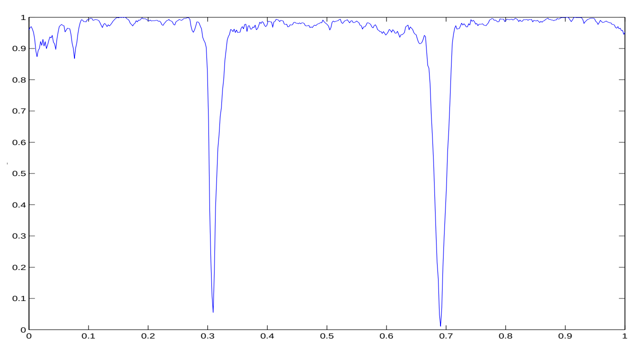
		JS plus		Harmonic Plus		Spline		Local Linear	
f	STNR	coverage	width	coverage	width	coverage	width	coverage	width
discon	1	96.2	2.66	98.1	2.70	79.7	0.94	85.6	1.18
	4	94.7	1.43	96.8	1.41	86.6	0.68	83.1	0.61
	16	93.9	0.82	95.5	0.78	91.6	0.50	83.0	0.32
	64	94.2	0.48	95.9	0.45	94.7	0.37	84.9	0.22
camel	1	97.1	2.58	98.4	2.66	84.5	0.79	91.4	0.95
	4	97.1	1.30	98.4	1.33	87.3	0.48	90.9	0.58
	16	96.4	0.66	98.2	0.67	89.9	0.29	91.3	0.30
	64	96.3	0.33	97.8	0.34	92.8	0.18	91.0	0.18
dipper	1	96.7	2.60	98.4	2.66	87.1	1.03	89.2	0.96
	4	96.6	1.31	98.3	1.35	91.0	0.48	90.9	0.57
	16	96.0	0.68	98.0	0.68	93.2	0.29	90.8	0.34
	64	95.7	0.35	97.6	0.35	94.2	0.18	90.2	0.19
claws	1	96.3	2.84	97.9	2.81	92.9	1.62	89.0	1.80
	4	96.5	1.42	97.9	1.42	94.6	0.93	89.5	1.07
	16	96.4	0.73	97.7	0.72	95.2	0.52	90.8	0.60
	64	95.8	0.37	97.3	0.37	95.8	0.29	92.8	0.39
corner	1	96.4	2.64	98.3	2.69	89.8	1.16	90.2	1.25
	4	95.8	1.38	97.6	1.37	91.6	0.71	90.7	0.58
	16	94.9	0.73	97.0	0.72	93.4	0.44	90.9	0.38
	64	95.0	0.38	96.6	0.38	94.7	0.28	91.2	0.21
doppler	1	94.1	3.05	96.0	2.79	89.0	2.31	79.4	0.63
	4	94.5	1.65	95.6	1.62	91.7	1.66	77.1	0.51
	16	95.3	0.93	96.8	0.95	94.8	1.24	72.1	0.32
	64	96.6	0.62	97.9	0.64	97.8	1.05	69.6	0.16
bumps	1	93.1	4.26	94.3	3.91	94.1	3.94	76.1	1.06
	4	92.6	2.84	94.4	2.67	96.7	3.12	73.2	0.74
	16	92.7	2.19	94.7	2.13	98.4	2.85	63.6	0.40
	64	92.5	1.97	94.5	1.93	99.5	3.04	59.7	0.26
blocks	1	91.7	3.48	94.0	3.26	89.4	2.54	67.3	1.45
	4	91.0	2.07	92.5	1.94	90.7	1.71	66.8	0.89
	16	91.4	1.40	93.2	1.31	94.4	1.37	68.8	0.59
	64	92.3	1.10	94.1	1.04	96.1	1.21	69.7	0.44

Table 1: Average Coverage and Length of Average Coverage Confidence Intervals. The nominal coverage is 95%. Sample size $n = 512$, 1000 repetitions.

the variance band is noticeably too narrow. This is reflected by the fact that the average coverage is far below the nominal coverage. This is especially the case for functions that are rougher, such as the bumps and blocks functions. It is because when the underlying function is less smooth, the local linear estimator has larger bias and this bias is not corrected by the variance band. The smoothing spline Bayesian confidence intervals in general performs better than local linear, but it tends to considerably under-cover the true function values (coverage is below 90%) when the signal to noise ratio is small.



(a) Function Estimate and its Confidence Interval



(b) Pointwise Coverage

Figure 5: Confidence Intervals and Average Coverage of Discontinuous Function, STNR=16:1

The underlying model for Figure 5 is the discontinuous function with signal to noise ratio 16 to 1. The method for constructing confidence intervals is based on the James-Stein plus estimator. The top graph shows a typical picture for the underlying function, data and confidence intervals from the model. The bottom graph shows the coverage at each value of x_i as estimated by the Monte Carlo simulation. It is clear that the coverage around the two jump points is very low, and the smooth parts of the function in general have higher coverage than 95%. The average coverage for the simulation of Figure 5 (b) is 93.9%.

A feature of our bands is that the width varies (across values of x) in response to locally perceived smoothness of underlying function. This is reasonable. It is much more difficult to cover parts of the function that are rough. On the other hand, the average coverage criterion we have chosen gives up to some extent the most difficult points and aims to control only the average coverage. For this reason, we expect non-homogenetic coverage across different parts of the function. That is exactly what is shown in Figure 5.

Figure 5 is a typical graph in the way it shows this feature of the average coverage criteria. Even though we construct a pointwise confidence interval, the coverage around the two jump points is very low. This is typically the case for construction of confidence intervals because of the large magnitude of bias around jump points.

4.2 Application to Call Center Data

In this section, we apply our method to construct a confidence interval for the arrival rate of call center data that is studied in Brown, et al., (2005). The source of the data is a small call center for one of Israel's banks.

In this paper, we will focus on the arrival rate of telephone calls requesting basic service. It is assumed that the number of calls arriving has an inhomogeneous Poisson process with mean μ. The task is to provide the nonparametric function estimator and confidence interval for this rate using our method.

The first step is to divide up the duration of the working day into relatively short blocks of time, in our case evenly-spaced 2 minute intervals. Denote $N_i \sim Possion(\mu)$ to be the number of calls within time interval i. Then we transform N_i according to

$$Y_i = \sqrt{N_i + 1/4}$$

so that Y_i is approximately normal with mean $\sqrt{\mu}$ and almost constant variance. Then after using our James-Stein procedure to produce a nonparametric estimator and confidence interval, we transform everything back by taking the square. (Notice that the unroot step does not involve the constant $1/4$. Detailed discussions of the this root-unroot procedure can be found in Brown, et al., (2010).) This reference validates the accuracy and utility of the root-unroot procedure and also discusses the relation of this type of Poisson process analysis to the familiar formulation for nonparametric density estimation.

The number of arriving calls with the nonparametric estimators for arrival rate and associated confidence interval is shown by Figure 6. Note the peak in arrival rate shortly after 10 AM and again at around 3 PM. The confidence interval makes fairly clear that these should be considered as separate modes with a dip in between. This generally bimodal pattern of call arrivals was noted and commented on further in Brown, et. al. (2005). Of course, since the band in Figure 6 is an average coverage band rather than a simultaneous coverage confidence interval it is not well formulated to provide a significance test of such an assertion. Dumbgen and Walther (2008) and Rufibach and Walther (2009) describe a method that could be adapted to investigate the statistical significance of these modes. In addition, Chaudhuri and Marron (1999) discuss a different, useful perspective on the search for modes in such data. Our curve estimate also shows a brief, local dip at about 3 pm. The confidence interval suggests that this dip might be statistically significant, but does not provide clear or convincing evidence of this.

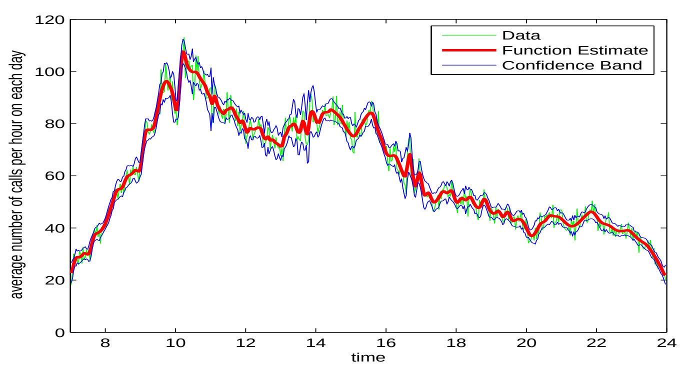


Figure 6: Confidence Interval for Arrival Rate of Call Center Data.

5 Conclusion

We have considered two methods for producing non-parametric confidence intervals based on blockwise shrinkage estimators. According to the simulation results, both of the confidence intervals perform consistently well over functions with very different degree of smoothness and various signal to noise ratios. This shows the method's ability to adapt to unknown smoothness. Moreover, the computation cost is very low due to the convenience of wavelet methods and the availability of the analytical forms of the shrinkage estimators as well as the additional feature that they do not have tuning parameters.

Acknowledgements

The authors would like to thank Tony Cai for many useful suggestions.

A Appendix

1), When $m \geq 4$ even,

$$\begin{aligned}
\rho_H'(||w||) &= \sigma^3 \Bigg[\sigma^2(m-2)||y||^3 \left(1 - P_{||y||^2/2}\left(\frac{m}{2}-2\right)\right) p_{||y||^2/2}\left(\frac{m}{2}-1\right) \\
&+ (m-2)\beta||y||^m p_{||y||^2/2}\left(\frac{m}{2}-1\right) e^{-||y||^2/2}(||y||-1) \\
&+ \beta||y||^{m+1}\left(1 - P_{||y||^2/2}\left(\frac{m}{2}-2\right)\right) e^{-||y||^2/2}(m-2-||y||^2) \\
&- 2\sigma^2(m-2)||y||\left(1 - P_{||y||^2/2}\left(\frac{m}{2}-2\right)\right)\left(1 - P_{||y||^2/2}\left(\frac{m}{2}-1\right)\right) \\
&- \sigma^2(m-2)||y||^2 p_{||y||^2/2}\left(\frac{m}{2}-2\right)\left(1 - P_{||y||^2/2}\left(\frac{m}{2}-1\right)\right) \\
&- (m-2)\beta||y||^{m-1}(m-||y||^2)\left(1 - P_{||y||^2/2}\left(\frac{m}{2}-1\right)\right) e^{-||y||^2/2}\Bigg] \\
&/ \; (\sigma^2||y||^2(1 - P_{||y||^2/2}(m/2-2)) + \beta||y||^m e^{-||y||^2/2})^2,
\end{aligned}$$

where $y = w/\sigma, p_\mu(\cdot)$ is the Poisson probability function with mean equal to μ.

2), When $m \geq 3$ odd,

$$\begin{aligned}
\rho_H'(||w||) &= \sigma^3 e^{-||y||^2/2}\Big[\beta||y||^m e^{-||y||^2/2}(1 + ||y||w_{m+2}(||y||)) \\
&+ \sigma^2(m-2)||y||^2(1 + ||y||w_{m+2}(||y||))(\Psi(||y||) - e^{-||y||^2/2}W_m(||y||)) \\
&+ \beta||y||^{m+1}(\Psi(||y||) - e^{-||y||^2/2}W_m(||y||))(m-2-||y||^2)
\end{aligned}$$

$$
\begin{aligned}
- &\quad 2e^{||y||^2/2}\sigma^2(m-2)||y||(\Psi(||y||) - e^{-||y||^2/2}W_m(||y||)) \\
&\quad (\Psi(||y||) - e^{-||y||^2/2}W_{m+2}(||y||)) \\
- &\quad \sigma^2(m-2)||y||^2(1+||y||w_m(||y||))(\Psi(||y||) - e^{-||y||^2/2}W_{m+2}(||y||)) \\
- &\quad ||y||^{m+2}\beta e^{-||y||^2/2}(1+||y||w_m(||y||)) \\
- &\quad (m-2)\beta||y||^{m-1}(m-||y||^2)(\Psi(||y||) - e^{-||y||^2/2}W_{m+2}(||y||))\Big] \\
/ &\quad (\sigma^2||y||^2(\Psi(||y||) - e^{-||y||^2/2}W_m(||y||)) + \beta||y||^m e^{-||y||^2/2})^2,
\end{aligned}
$$

where $y = w/\sigma$, $\Psi(s) = \sqrt{2\pi}(\Phi(s) - 1/2)$. $w_m(s)$ is the last term in $W_m(s)$ i.e., $w_m(s) = \frac{2^{(m-1)/2-1}((m-1)/2-1)!}{(m-2)!}s^{m-2}$. Here $\phi(\cdot)$ is the density function of standard normal distribution.

References

[1] Barber, S., Nason, G.P. and Silverman, B.W. (2002), "Posterior Probability Intervals for Wavelet Thresholding." *Journal of the Royal Statisitcal Society, SerB*, **64** 189-205.

[2] Berger, J. (1980), "A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Nornal Mean." *Ann. Statist.* **8** 716-761.

[3] Brown, L.D. (1971), "Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems." *The Annals of Mathematical Statistics* **42** 855-903.

[4] Brown, L.D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005), "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." *J. Amer. Statist. Assoc.* **100** 36-50.

[5] Brown, L.D., Cai, T.T., Zhang, R., Zhao, L., and Zhou, H.H. (2010), "The Root-Unroot Algorighm for Density Estimation as Implemented via Wavelet Blcok Thresholding." *Probability Theory and Related Fields.* to appear.

[6] Brown, L.D. and Xu, J. (2005), "A Closed Form for The Harmonic-Prior Bayes Estimator with Associated Confidence Sets for the Means of a Multivariate Normal Distribution", preprint.

[7] Cai, T.T. (1999), "Adaptive Wavelet Estimation: a Block Thresholding and Oracle Inequality Approach." *Ann. Statist.* **27** 898-924.

[8] Cai, T.T., Low, M.G. and Zhao, L.H. (2000), "Sharp Adaptive Estimation by a Blockwise Method." to appear in *J. of Nonparametric Statistics.*

[9] Chaudhuri, P. and Marron, J.S. (1999), "SiZer for Exploration of Structures in Curves." *J. Amer. Statist. Assoc.* **94** 807-823.

[10] Cavalier, L., Golubev, G.K., Picard, D. and Tsybakov, A.B. (2002), "Oralcle Inequalities for Inverse Problems." *Ann. Statist.* **30** 843-874.

[11] Daubechies, I. (1993), "Orthonormal Bases of Compactly Supported Wavelets II: Variations on a Theme." *SIAM J.math. Anal.* **24**, 499-519.

[12] Donoho, D.L. and Johnstone, I.M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage." *Biometrika* **81** 425-455.

[13] Donoho, D.L. and Johnstone, I.M. (1995a), "Wavelet Shrinkage:Asymptopia." *Journal of the Royal Statistical Society, SerB* **57** 301-369.

[14] Donoho, D.L. and Johnstone, I.M. (1995b), "Adapting to Unknown Smoothness via Wavelet Shrinkage", *J. Amer. Statist. Assoc.* **90** 1200-1224.

[15] Dumbgen, L. and Walther, G. (2008), "Multiscale Inference about a Density", *Ann. Statist.* **36** 1758-1785.

[16] Efron, B. (1987), "Better Bootstrap Confidence Intervals (with discussion)." *J. Amer. Statist. Assoc.* **82** 171-200.

[17] Eubank, R.L. and Speckman P.L. (1993), "Confidence Bands in Nonparametric Regression." *J. Amer. Statist. Assoc.* **88** 1287-1301.

[18] Hall, P. (1986), "On the Bootstrap and Confidence Intervals." *Ann. Statist.* **14** 1431-1452.

[19] Hall, P. (1988), "Theoretical Comparison of Bootstrap Confidence Intervals (with discussion)." *Biometrika* **75** 661-671.

[20] Hall, P. (1993), "On Edgeworth Expansion and Bootstrap Confidence Bands in Nonparametric Curve Estimation." *Journal of the Royal Statisitcal Society, SerB* **55** 291-304.

[21] Hardle, W. and Marron, J.S. (1991), "Bootstrap Simultaneous Error Bars for Nonparametric Regression." *Ann. Statist.* **19** 778-796.

[22] Kerkyacharian, G., Picard, D. and Tribouley, K. (1996), "Adaptive Density Estimation." *Bernoulli* **2** 229-247.

[23] Mao, W. and Zhao L.H. (2003), "Free-Knot Polynomial Splines with Confidence Intervals." *Journal of the Royal Statisitcal Society, SerB* **65** 901-919.

[24] Nychka, D. (1988), "Bayesian Confidence Intervals for Smoothing Splines." *J. Amer. Statist. Assoc.* **83** 1134-1143.

[25] Rice, J. (1984), "Bandwidth Choice for Nonparametric Kernel Regression." *Ann. Statist.* **12** 1215-1230.

[26] Rufibach, K. and Walther, G. (2009), "The Block Criterion for Multiscale Inference About a Density, With Applications to Other Multiscale Problems. " *Journal of Computational and Graphical Statistics*, posted online.

[27] Ruppert, D., Sheather, S.J. and Wand, M.P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression." *J. Amer. Statist. Assoc.* **90** 1257-1270.

[28] Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution." *Ann. Statist.* **9** 1135-1151.

[29] Wahba, G. (1983), "Bayesian 'Confidene Intervals' for the Cross-Validated Smoothing Spline." *Journal of the Royal Statistical Society, SerB* **45**, 133-150.

[30] Xia, Y. (1998), "Bias-Corrected Confidence Bands in Nonparametric Regression." *Journal of the Royal Statistical Society. SerB* **60**, 797-811.