Adaptive Confidence Bands for Nonparametric Regression Functions

T. Tony Cai*, Mark Low and Zongming Ma
University of Pennsylvania

Abstract

A new formulation for the construction of adaptive confidence bands in nonparametric function estimation problems is proposed. Confidence bands are constructed which have size that adapts to the smoothness of the function while guaranteeing that both the relative excess mass of the function lying outside the band and the measure of the set of points where the function lies outside the band are small. It is shown that the bands adapt over a maximum range of Lipschitz classes. The adaptive confidence band can be easily implemented in standard statistical software with wavelet support. Numerical performance of the procedure is investigated using both simulated and real datasets. The numerical results agree well with the theoretical analysis. The procedure can be easily modified and used for other nonparametric function estimation models.

Keywords: Adaptive confidence band, average coverage, coverage probability, excess mass, lower bounds, noncovered points, nonparametric regression, wavelets, white noise model.

AMS 2000 subject classifications: Primary 62G07; secondary 60F05

1 Introduction

Adaptive inference has been a major focus in nonparametric function estimation. Within this area there has been considerable success constructing procedures for estimating a regression function or density which adapt to the smoothness properties of the unknown function. A particularly successful example is wavelet thresholding but there are a wide variety of estimation procedures with proven optimality properties.

Unfortunately the development of a satisfactory theory for adaptive confidence bands has proved to be more difficult. Ideally, an adaptive confidence band should have its size automatically adjusted to the smoothness of the underlying function, while maintaining a prespecified coverage probability. However as we shall show such a goal is impossible even for Lipschitz function classes and hence a new framework for investigating adaptive confidence bands is needed. The primary goal of the present paper is to provide such a framework along with a new confidence band procedure that not only has good numerical performance but also achieves adaptivity in this new framework.

Consider the nonparametric regression model

$$y_i = f(t_i) + \sigma \epsilon_i, \quad i = 1, ..., n \tag{1}$$

where $t_i = \frac{i}{n}$ and $\epsilon_i \overset{iid}{\sim} N(0,1)$. The goal is to construct a confidence band for f on the interval $[0,1]$. A confidence band CB can be represented by two random functions, the lower limit $L(\cdot)$ and the upper limit $U(\cdot)$ where $L(t)$ and $U(t)$ are two functions based on the observations $\{y_1, ..., y_n\}$ such that $L(t) \leq U(t)$ for all $0 \leq t \leq 1$. We shall write $CB = [L(t), U(t)]$.

For a fixed collection of functions $\mathcal{F}$, write $\mathcal{B}_\alpha(\mathcal{F})$ for the collection of all confidence bands which have guaranteed coverage probability of at least $1 - \alpha$ over $\mathcal{F}$, i.e.,

$$\mathcal{B}_\alpha(\mathcal{F}) = \left\{ CB = [L(t), U(t)] : \inf_{f \in \mathcal{F}} P_f \left(f(t) \in [L(t), U(t)], \ \forall \, 0 \leq t \leq 1 \right) \geq 1 - \alpha \right\}. \tag{2}$$

Useful bands for the unknown function should then be chosen from this collection so that the size of the resulting band is "small" while guaranteeing coverage. Two natural measures of the size of the band are given by the average width $\int_0^1 (U(t) - L(t))dt$ and the maximum width $\max_t (U(t) - L(t))$.

Given that the size of the confidence band is allowed to be random it is helpful to evaluate the expected width of the band which typically may also depend on the function f. For a

confidence band $CB = [L(t), U(t)]$, write

$$w(CB, f) = E_f \int_0^1 (U(t) - L(t)) dt$$

for the expected average width for a particular $f \in \mathcal{F}$. In this setting an adaptive band should have values of $w(CB, f)$ which adjust to the unknown function f in the sense that it is small when a function f is easier to estimate. However, before explaining why this goal is not typically possible, it is helpful to first introduce

$$w(CB, \mathcal{F}) = \sup_{f \in \mathcal{F}} w(CB, f) = \sup_{f \in \mathcal{F}} E_f \int_0^1 (U(t) - L(t)) dt,$$

the maximum expected average width where the maximum is taken over all $f \in \mathcal{F}$. In addition, the minimax expected average width $W_\alpha(\mathcal{F})$ of confidence bands which have guaranteed coverage probability at least $1 - \alpha$ over $\mathcal{F}$ is denoted by

$$W_\alpha(\mathcal{F}) = \inf_{CB \in \mathcal{B}_\alpha(\mathcal{F})} w(CB, \mathcal{F}) = \inf_{CB \in \mathcal{B}_\alpha(\mathcal{F})} \sup_{f \in \mathcal{F}} w(CB, f).$$

For example consider the Lipschitz classes

$$\Lambda(\beta, M) = \{f : |f(y) - f(x)| \leq M|y - x|^\beta \text{ for } x, y \in [0, 1]\}$$

for $0 < \beta \leq 1$, and for $\beta > 1$

$$\Lambda(\beta, M) = \{f : |f^{(\lfloor \beta \rfloor)}(x) - f^{(\lfloor \beta \rfloor)}(y)| \leq M|x - y|^{\beta'} \text{ for } x, y \in [0, 1]\},$$

where $\lfloor \beta \rfloor$ is the largest integer less than β and $\beta' = \beta - \lfloor \beta \rfloor$. These are among the most commonly considered parameter spaces in the nonparametric function estimation literature. The minimax theory for such parameter spaces can be developed relatively easily and as shown later the minimax expected average width is given by

$$W_\alpha(\Lambda(\beta, M)) \asymp M^{\frac{1}{2\beta+1}} \left(\frac{\log n}{n} \right)^{\frac{\beta}{1+2\beta}}$$

and can be attained by a fixed width confidence band centered on a linear estimator. However as is typical the confidence band centered on a linear procedure that attains this bound for a given Lipschitz class behaves poorly for other classes. It either has poor coverage or the expected average width of the band is unnecessarily large. Such a band is clearly not adaptive to the smoothness property of the function. This therefore leads naturally to the question of whether it is possible to construct a confidence band that performs well simultaneously over a collection of Lipschitz classes.

3

1.1 Impossibility of Adaptation over Lipschitz Classes

An adaptive confidence band over a collection of parameter spaces $\mathcal{C} = \{\mathcal{F}_i : i \in I\}$ where I is an index set should guarantee a given coverage probability over $\mathcal{C}$ while simultaneously minimizing the maximum expected average width over each of the parameter spaces $\mathcal{F}_i$. Hence a confidence band $CB \in \mathcal{B}_\alpha(\cup_{i\in I}\mathcal{F}_i)$ is called adaptive over $\{\mathcal{F}_i : i \in I\}$ if for all $i \in I$,

$$w(CB, \mathcal{F}_i) \leq C_i W_\alpha(\mathcal{F}_i)$$

where C_i are constants not depending on n, and we say that adaptation is possible over the collection $\{\mathcal{F}_i : i \in I\}$ if such a procedure exists.

Unfortunately this adaptation goal is not typically attainable. For example it is not possible to adapt over even two Lipschitz classes $\Lambda(\beta_0, M_0)$ and $\Lambda(\beta_1, M_1)$ with $\beta_0 < \beta_1$. That is, for all $CB \in \mathcal{B}_\alpha(\Lambda(\beta_0, M_0) \cup \Lambda(\beta_1, M_1))$ there is a constant $d > 0$ such that

$$w(CB, \Lambda(\beta_1, M_1)) \geq dn^{-\frac{\beta_0}{2\beta_0+1}} \gg W_\alpha(\Lambda(\beta_1, M_1)). \tag{3}$$

This result is an immediate consequence of the minimax lower bound given in Theorem 2 in Section 4, which provides even stronger negative statements. These results show that there is essentially no room for improvement in terms of rate of convergence. The expected average width (up to log terms) is essentially the same for every function and hence the size must be essentially of the same order as in the worst case no matter the true function. In marked contrast to estimating the unknown function under integrated mean squared error, the construction of adaptive bands in this context is thus impossible from the classical view of covering the entire function.

This impossibility of constructing adaptive confidence bands in such settings is now well known and has led to alternative formulations of the adaptation problem. In the literature, there are at least two different approaches toward this goal. One approach is to impose additional structural assumptions. This reduces the parameter space and makes the coverage requirement (2) easier to satisfy. For example, Hengartner and Stark (1995), Dümbgen (1998), among many others, considered shape constraints such as monotonicity or convexity, and showed that adaptation is achievable under such constraints. Recently, Giné and Nickl (2010) considered a self-similarity-type constraint which also leads to adaptation. Moreover, their results also implied that functions not satisfying such constraint are nowhere dense in

Lipschitz classes. See also Hoffman and Nickl (2011) and Bull (2011b). The other approach toward adaptation is to relax the notion of coverage. In particular Genovese and Wasserman (2008) suggested the notion of surrogate coverage, which requires the band to cover either the function f or a smoother surrogate with probability $1 - \alpha$. Under this new notion of coverage, the authors showed that a particular type of adaptation can be achieved. Wahba (1983) proposed the notion of average coverage. Instead of covering the entire function with probability $1 - \alpha$, the average coverage criterion requires the confidence band to cover on average $100 \times (1 - \alpha)\%$ of the points. See also Nychka (1988). However for average coverage an adaptation theory has not yet been developed.

1.2 New Formulation

The focus of the present paper is to introduce two different but related relaxations of the classical notion usually required of a confidence band namely that of covering the function at all points. The goal is still to cover the true function rather than some surrogate function and we do not wish to impose order constraints on the function or to restrict attention only to special self-similar-type functions within a smoothness class. Instead we shall, as in the case for average coverage, give up guaranteeing coverage at all points with the goal of allowing more adaptive confidence bands where the size of the band reflects the underlying difficulty in recovering the particular unknown function.

More specifically the first relaxation focuses on the measure of the set of points where coverage does not occur whereas the second focuses on the excess mass of the function lying outside of the confidence band. Hence for the first relaxation, the goal is to construct a confidence band with bandwidth automatically adjusting to the smoothness of the underlying function, while maintaining coverage of the function at "most" of the points in $[0, 1]$. This point of view is related to that of guaranteeing average coverage as described earlier. Under the second relaxation, the goal is to have confidence bands that, with a pre-specified probability, limit the amount of excess mass of the true function outside of the confidence band. The goal is to guarantee that the excess mass compared to the size of the band, is negligible.

Set of Noncovered Points

For a confidence band $CB = [L(t), U(t)]$, define the set of noncovered points by

$$N(CB, f) = \{t \in [0, 1] : f(t) \notin [L(t), U(t)]\}.$$

Note that $N(CB, f)$ is a random subset of $[0, 1]$ since CB is random. It is natural to require that this random set $N(CB, f)$ be "small" for a good confidence band procedure CB. That is, one would like CB to cover the function f over "most" of the points in $[0, 1]$ with probability at least $1 - \alpha$.

In this paper "most" will refer to a set of points with measure that goes to zero as the sample size increases. More specifically, the coverage probability condition (2) is relaxed to

$$\inf_{f \in \mathcal{F}} P_f(\mu(N(CB, f)) \leq \xi_n) \geq 1 - \alpha \tag{4}$$

for some sequence of positive numbers ξ_n such that $\xi_n \to 0$ as $n \to \infty$.

Under this relaxation the goal of an adaptive band can then be formulated for the Lipschitz classes. Subject to guaranteeing covering the function at most points, the aim is to minimize the expected average width simultaneously for an entire collection of Lipschitz classes, a goal that is ruled out by (3) for usual confidence bands.

Relative Excess Mass

In addition to wanting a confidence band to cover the true function at most points it is also natural to want the total mass of the function that lies outside the band to be small. For a confidence band $CB = [L(t), U(t)]$ and a function f, define the excess mass function by

$$e_f(t) = [f(t) - U(t)]_+ + [L(t) - f(t)]_+. \tag{5}$$

Then the integrated excess mass of the function f with respect to CB is $\int_0^1 e_f(t)dt$. In other words, $\int_0^1 e_f(t)dt$ is the total amount of mass of f that lies outside of the band CB. We then measure the performance of CB by its relative excess mass

$$RE(CB, f) = \frac{\int_0^1 e_f(t)dt}{\int_0^1 [U(t) - L(t)]dt}.$$

For a good confidence band procedure, with probability at least $1 - \alpha$, the area of the true function lies outside of the band should be "small" compared to the area of the band itself, i.e.,

its relative excess mass should be small. More precisely, we relax the coverage requirement (2) to

$$\inf_{f \in \mathcal{F}} P_f(RE(CB, f) \leq \xi'_n) \geq 1 - \alpha \tag{6}$$

for some sequence of positive numbers ξ'_n such that $\xi'_n \to 0$ as $n \to \infty$.

1.3   Adaptive Procedure

One focus of the present paper is to develop an adaptive confidence band which controls both the measure of the set of noncovered points and the relative excess mass and for which both go to zero asymptotically. Such a goal is possible for particular ranges of Lipschitz classes. However before we discuss in detail our adaptive band it is important to first discuss limits on the possible range of adaptation as this range will enter naturally into our adaptive band. Note that a band that is adaptive over two Lipschitz classes $\Lambda(\beta_0, M_0)$ and $\Lambda(\beta_1, M_1)$ should satisfy either (4) if attention is focused on the collection of points where coverage does not occur or (6) if attention is focused on excess mass where in both cases $\mathcal{F} = \Lambda(\beta_0, M_0) \cup \Lambda(\beta_1, M_1)$. For the band to be adaptive the maximum expected average width should be $(\log n / n)^{\beta_i / (1 + 2\beta_i)}$ over $\Lambda(\beta_i, M_i)$ for $i = 0$ and $i = 1$.

Unfortunately lower bound results given in Section 4 show that this goal cannot be achieved if $\beta_1 > 2\beta_0 > 0$. In fact in such a case if the maximum expected average width over $\Lambda(\beta_1, M_1)$ is of order $(\log n / n)^{\beta_1 / (1 + 2\beta_1)}$, then

$$\sup_{f \in \Lambda(\beta_0, M_0)} P_f \left(\mu(N(CB, f)) \geq \frac{1}{2} \right) \geq \frac{1}{2}$$

and

$$\sup_{f \in \Lambda(\beta_0, M_0)} P_f \left(RE(CB, f) \geq r \right) \geq \frac{1}{2}$$

for any given $r > 0$, when n is sufficiently large. That is, there is better than 50% of chance that the confidence band misses some function in $\Lambda(\beta_0, M_0)$ over more than half of the interval $[0, 1]$ and there is better than 50% of chance that some function in $\Lambda(\beta_1, M_1)$ has excess mass much more than the area of the band.

This shows that adaptation is not possible over Lipschitz classes $\Lambda(\beta, M)$ for $\beta \in [\beta_0, \beta_1]$ with $\beta_1 > 2\beta_0 > 0$ even under either of the more relaxed conditions. These extremely negative

results however do not apply when $\beta_1 < 2\beta_0$ and so our focus is on constructing confidence bands that are adaptive over the collection of Lipschitz classes $\Lambda(\beta, M)$ for $\beta \in [\beta_0, 2\beta_0]$ for a prespecified minimum smoothness value $\beta_0 > 0$.

One major goal of the present paper is to show that it is indeed possible to adapt over the range $[\beta_0, 2\beta_0]$ under both the set of noncovered points criterion (4) and the relative excess mass criterion (6). Given the minimum smoothness β_0 and the maximum Lipschitz constant M_0, we construct a data-driven confidence band using wavelet techniques. The proposed band centers on a wavelet projection estimator of the regression function where the projection level is determined by the results of testing multiple hypotheses. The null hypotheses are naturally constructed from the Hölder conditions on the wavelet coefficients of Lipschitz functions, while the alternative hypotheses are carefully designed to control both the set of noncovered points and excess mass. After determining the projection level and hence the center, we specify the width of the band by controlling the stochastic error and the bias of such projection estimators separately. The resulting band is a *uniform band* where the width of the band $U(t) - L(t) = \hat{w}_n$ does not depend on t. It is shown to meet both criteria (4) and (6) simultaneously over all Lipschitz classes $\Lambda(\beta, M)$ where $\beta \in [\beta_0, 2\beta_0]$ and $M \in [1, M_0]$. In addition, the adaptive confidence band is shown to have desirable average coverage probability.

The proposed confidence band procedure can be implemented efficiently in standard statistical software with wavelet support. Numerical performance of the procedure is investigated using both simulated examples and a call center dataset. For simulated examples, the performance of the band agrees well with the asymptotic theory even when the sample size is not large. For the call center data, the procedure leads to a smooth and interpretable band and confirms the significance of a peak of call arrival.

1.4   Organization of the Paper

The rest of the paper is organized as follows. Section 2 presents the detailed construction of an adaptive confidence band using wavelet techniques. Section 3 analyzes the theoretical properties of the confidence band, and investigates its numerical performance by simulations and real data analysis. A call center dataset is analyzed to illustrate the procedure. Section

4 formally states the limits on the range of adaptation over the Lipschitz classes $\Lambda(\beta, M)$ under both the set of noncovered points criterion (4) and the excess mass criterion (6) by establishing lower bounds under both criteria. The lower bounds together with the upper bounds obtained in Section 2 show that the proposed confidence band is optimally adaptive under both criteria. Further discussions on the connections of our results and those of related problems are given in Section 5. The main results are proved in Section 6. Additional technical details are provided in a supplement to this paper.

2   Construction of Adaptive Confidence Band

Before providing the detailed construction of the adaptive confidence band it is useful to restate a precise formulation of our goal in the construction of adaptive confidence bands over Lipschitz classes. For a prespecified minimum smoothness parameter β_0, the collection of function spaces that we aim to adapt over is

$$\mathcal{A}(\beta_0, M_0) = \{\Lambda(\beta, M) : \beta \in [\beta_0, 2\beta_0], M \in [1, M_0]\}, \tag{7}$$

where $M_0 > 1$ is also given. In addition, we require $\beta_0 > \frac{1}{4}$.

For a prespecified confidence level $1 - \alpha$, the goal is to construct a single confidence band $CB = [L(t), U(t)]$ which simultaneously satisfies the following three requirements.

(a) (Average width condition) There exist a constant C, such that for any $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$,

$$\sup_{f \in \Lambda(\beta, M)} \mathsf{E}_f \int_0^1 [U(t) - L(t)]dt \leq CM^{\frac{1}{2\beta+1}} \left(\frac{\sigma^2 \log n}{n} \right)^{\frac{\beta}{2\beta+1}}. \tag{8}$$

(b) (Noncovered points condition) There exist a sequence of positive numbers $\xi_n = \xi_n(\beta_0, M_0)$ with $\xi_n \to 0$ as $n \to \infty$, such that for each $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$,

$$\lim_{n \to \infty} \inf_{f \in \Lambda(\beta, M)} P_f \left(\mu(N(CB, f)) \leq \xi_n \right) \geq 1 - \alpha. \tag{9}$$

(c) (Excess mass condition) There exist a sequence of positive numbers $\xi_n' = \xi_n'(\beta_0, M_0)$ with $\xi_n' \to 0$ as $n \to \infty$, such that for each $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$,

$$\lim_{n \to \infty} \inf_{f \in \Lambda(\beta, M)} P_f \left(RE(CB, f) \leq \xi_n' \right) \geq 1 - \alpha. \tag{10}$$

If a confidence band satisfies all three conditions, then its size contracts at an optimal rate with respect to the smoothness parameters β and M. In addition, with asymptotic probability at least $1 - \alpha$, it covers the function on most points in $[0, 1]$ and the excess mass of the function is negligible compared to the band size.

In this section such an adaptive confidence band is constructed based on the observed data $\{y_i : 1 \leq i \leq n\}$. The band is a uniform band with width that depends on the data. The detailed construction depends on an estimate of the underlying function which is taken to be the center of the band along with a specification of the data dependent width. The center is given by a wavelet estimate of the function. It is thus helpful to first introduce a few useful facts about the wavelet coefficients of Lipschitz functions. Then, we investigate the bias and variance properties of projection estimators, which leads to a rate optimal oracle band. Motivated by this oracle procedure, we introduce a hypothesis testing scheme for selecting the projection level based on data, which results in a data-driven choice for both the center and the width of the band.

2.1 Wavelet Preliminaries

We first characterize Lipschitz functions via their wavelet coefficients. Let $\{\psi_{lk} : l \geq 0, k = 1, \ldots, 2^l\}$ form a wavelet basis on $[0, 1]$ with the mother wavelet $\psi \in C^s$ for some integer $s > 2\beta_0$. In addition, we assume that ψ is compactly supported with support length S. For any $f \in \Lambda(\beta, M)$ let $\theta[f] = (\theta_{lk}) = (\langle f, \psi_{lk} \rangle)$ be its wavelet coefficients. Then, see for example Lemma 7.3 in Johnstone (2012)

$$\max_k |\theta_{lk}(f)| \leq c_\psi M \, 2^{-(\beta + \frac{1}{2})l}, \quad \text{for all } l, \tag{11}$$

where c_ψ is a constant depending only on the wavelet basis and β_0. For instance, we could let $c_\psi = \max\{1, c_{\beta_0} \int [|x|^{2\beta_0} \vee 1] |\psi(x)| dx\}$, where $c_{\beta_0} = \prod_{j=1}^{[2\beta_0]} (2\beta_0 + 1 - j)$ if $2\beta_0 > 1$ and 1 otherwise. Thus, c_ψ can be evaluated numerically given ψ and β_0.

For convenience, we assume the sample size $n = 2^J$ for some integer $J > 0$. With the same wavelet basis, the observed data $\{y_i : 1 \leq i \leq n\}$ can be transformed into empirical wavelet coefficients

$$\{\hat{\theta}_{lk} : 1 \leq k \leq 2^l, 1 \leq l < J\}. \tag{12}$$

Let ϕ be the father wavelet of the wavelet basis, then

$$\mathsf{E}\hat{\theta}_{lk} = \bar{\theta}_{lk} \equiv \langle f_n, \psi_{lk} \rangle, \quad \text{with} \quad f_n(t) = \sum_{k=1}^{n} n^{-1/2} f(\frac{k}{n}) \phi_{Jk}(t). \tag{13}$$

If $f \in \Lambda(\beta, M)$, by making c_ψ in (11) sufficiently large, we also have

$$\max_k |\bar{\theta}_{lk}(f)| \le c_\psi M \, 2^{-(\beta + \frac{1}{2})l}, \quad \text{for all } l < J. \tag{14}$$

For a proof, see the supplement.

2.2   A Confidence Band For A Given Lipschitz Class

Our confidence band uses a projection estimator as its center. In this part, we investigate the bias and variance properties of projection estimators, which leads to a minimax rate optimal band for a given Lipschitz class.

For any resolution level $j < J$, the projection estimator of f at level j is

$$\hat{f}_j(t) = \sum_{l=0}^{j} \sum_{k=1}^{2^l} \hat{\theta}_{lk} \psi_{lk}(t) \tag{15}$$

where the empirical wavelet coefficients $\hat{\theta}_{lk}$ are given in (12). Let $f_j(t) = \mathsf{E}\hat{f}_j(t)$. Then a band can be formed by taking its center as $\hat{f}_j$ and setting the width of the band to be twice the sup-norm of the difference

$$f(t) - \hat{f}_j(t) = \left(f_j(t) - \hat{f}_j(t) \right) + \left(f(t) - f_j(t) \right). \tag{16}$$

Here, the first term is stochastic error and the second term is bias. In what follows, we bound the two terms on the right side respectively.

Bounding Stochastic Error

We use a result in Bull (2011a) to bound the stochastic error, which builds on the extreme value theory for cyclostationary Gaussian processes (Piterbarg and Seleznjev, 1994; Husler, 1999). It provides an extension of Theorem 2 of Giné and Nickl (2010), both of which improve earlier works of Smirnov (1950) and Bickel and Rosenblatt (1973). To this end, we need the following assumption on the mother wavelet ψ of the wavelet basis $\{\psi_{lk}\}$.

Assumption (W). The mother wavelet ψ of the wavelet basis $\{\psi_{lk}\}$ is compactly supported, and for $\sigma_\psi^2(t) = \sum_{k \in \mathbb{Z}} \psi(t-k)^2$, its maximum is attained at a unique point t_0 on $[0,1)$ with $(\sigma_\psi^2)''(t_0) < 0$.

Giné and Nickl (2010) and Giné et al. (2011) verified that the unique maximum assumption on $\sigma_\psi^2(t)$ is satisfied by spline bases, and Bull (2011a) showed numerically that it is also satisfied by the Daubechies and Symmlet classes. Thus, assumption (W) is satisfied by the Daubechies and Symmlet bases, whose mother wavelets are compactly supported. Under this assumption, let

$$\bar{\sigma}_\psi^2 = \sigma_\psi^2(t_0) = \max_{t \in [0,1)} \sigma_\psi^2(t) \qquad \text{and} \qquad v_\psi = -\frac{\sum_{k \in \mathbb{Z}} \psi'(t_0 - k)^2}{\bar{\sigma}_\psi \sigma_\psi''(t_0)}. \tag{17}$$

For any positive integer j, further define

$$a_j = \sqrt{2 \log 2}\,(j+1)^{\frac{1}{2}}, \tag{18}$$

$$b_j = a_j - \frac{\log(\pi \log 2) + \log(j+1) - \frac{1}{2}\log(1+v_\psi)}{2\,a_j}, \tag{19}$$

$$c_j = \frac{\sigma}{\sqrt{n}}\bar{\sigma}_\psi 2^{\frac{j+1}{2}}. \tag{20}$$

Proposition 1. *Let* $j_n \to \infty$, $\alpha_0 \in (0,1)$, *and* $\Gamma_n = [\alpha_n, \alpha_0]$, *where* $\alpha_n \in (0, \alpha_0)$ *and* $\alpha_n^{-1} = o(e^{Cj_n})$ *for any* $C > 0$. *If Assumption (W) is satisfied, then as* $n \to \infty$, *for* $x_\alpha = -\log(-\log(1-\alpha))$, $\mu_{j_n} = c_{j_n} b_{j_n}$ *and* $\sigma_{j_n} = c_{j_n}/a_{j_n}$,

$$\sup_{\alpha \in \Gamma_n} \left| \frac{1}{\alpha} P\left(\|\hat{f}_{j_n} - f_{j_n}\|_\infty > \mu_{j_n} + \sigma_{j_n} x_\alpha \right) - 1 \right| \to 0.$$

Remark 1. The quantity $\hat{f}_{j_n}(t) - f_{j_n}(t) = \sum_{l \leq j_n} \sum_k (\hat{\theta}_{lk} - \bar{\theta}_{lk})\psi_{lk}(t)$ does not depend on the underlying function f. Therefore, the convergence is uniform for all the function f that we are interested in. Following the lines of the proof in Bull (2011), one sees that the convergence is also uniform for all sequences $\{j'_n\}$ such that $j'_n \geq j_n$ for all n.

By Proposition 1, with proper centering and scaling determined by the projection level j_n and the wavelet basis, the stochastic error $\|f_{j_n} - \hat{f}_{j_n}\|_\infty$ converges weakly to a Gumbel distribution. When $j_n \to \infty$, $\sigma_{j_n} \asymp \mu_{j_n}/j_n \ll \mu_{j_n}$. Thus, with asymptotic probability $1 - \alpha$,

$$\|f_{j_n} - \hat{f}_{j_n}\|_\infty \asymp \mu_{j_n} \asymp \sigma n^{-\frac{1}{2}} 2^{\frac{j}{2}} j^{\frac{1}{2}}. \tag{21}$$

Bounding Bias

The bias term $\|f - f_j\|_\infty$ can be bounded by $\|f - f_j\|_\infty \leq \|f_n - f_j\|_\infty + \|f - f_n\|_\infty$, with f_n given by (13). For an analysis of the first term $\|f_n - f_j\|_\infty$ define

$$\tau_\psi = \sup_{l \geq 0} \sup_{t \in [0,1]} 2^{-\frac{l}{2}} \sum_{k \in \mathbb{Z}} |\psi_{lk}(t)|. \tag{22}$$

Since ψ has compact support with support length S,

$$\tau_\psi = \sup_{l \geq 0} \sup_{t \in [0,1]} 2^{-\frac{l}{2}} \sum_{k \in \mathbb{Z}} |2^{\frac{l}{2}} \psi(2^l t - k)| \leq S \max_{t \in \mathbb{R}} |\psi(t)| = O(1).$$

In practice, for any particular wavelet basis with compact support, τ_ψ can be evaluated numerically. For any $f \in \Lambda(\beta, M)$, (14) and (22) lead to

$$\|f_n - f_j\|_\infty = \left\| \sum_{l=j+1}^{J-1} \sum_{k=1}^{2^l} \bar{\theta}_{lk} \psi_{lk} \right\|_\infty \leq \tau_\psi \sum_{l=j+1}^{J-1} 2^{\frac{l}{2}} \max_{1 \leq k \leq 2^l} |\bar{\theta}_{lk}| \leq \tau_\psi c_\psi M \sum_{l=j+1}^{J-1} 2^{-\beta l}.$$

A similar analysis on the second term yields $\|f - f_n\|_\infty \leq \tau_\psi \sum_{l \geq J} 2^{\frac{l}{2}} \max_{1 \leq k \leq 2^l} |\theta_{lk}| \leq \tau_\psi c_\psi M \sum_{l \geq J} 2^{-\beta l}$.

Putting these two bounds together results in a further bound

$$\|f - f_j\|_\infty \leq \tau_\psi c_\psi M \sum_{l > j} 2^{-\beta l} = \frac{\tau_\psi c_\psi}{1 - 2^{-\beta}} M 2^{-\beta(j+1)} \tag{23}$$

and thus, $\|f - f_j\|_\infty \asymp M 2^{-\beta j}$.

Bias-Variance Tradeoff and an Oracle Band

Suppose that the band is centered at projection estimators $\hat{f}_{j_n}$ where $j_n \to \infty$ as $n \to \infty$. Then by Proposition 1 and (23) an asymptotic $1 - \alpha$ confidence band over $\Lambda(\beta, M)$ is given by

$$[\hat{f}_{j_n} - w_n, \, \hat{f}_{j_n} + w_n], \tag{24}$$

where the half-width

$$w_n = w_n(\beta, M) = (\mu_{j_n} + \sigma_{j_n} x_\alpha) + \frac{\tau_\psi c_\psi}{1 - 2^{-\beta}} M 2^{-\beta(j_n+1)}. \tag{25}$$

This band is constructed with the knowledge of both M and β. To minimize the half-width, and hence achieve the smallest average width among all bands of the form (24), a level

13

$j^{\mathrm{cb}} = j_n^{\mathrm{cb}}(\beta, R)$ is chosen which balances the two terms in the half-width expression, where the superscript "cb" stands for confidence bands. In other words, we require $\sigma n^{-\frac{1}{2}} 2^{\frac{j^{\mathrm{cb}}}{2}} (j^{\mathrm{cb}})^{\frac{1}{2}} \asymp M 2^{-\beta j^{\mathrm{cb}}}$. Note that this necessarily requires $j^{\mathrm{cb}} \asymp \log n$. Thus, we could require more precisely that j^{cb} is the solution of

$$2^{-\beta} \sigma \sqrt{\frac{\log n}{n}} < c_\psi M 2^{-(\beta + \frac{1}{2})j} \le \sigma \sqrt{\frac{2 \log n}{n}}, \tag{26}$$

which leads to

$$2^{j^{\mathrm{cb}}} \asymp \left(\frac{M}{\sigma}\right)^{\frac{2}{2\beta+1}} \left(\frac{n}{\log n}\right)^{\frac{1}{2\beta+1}}. \tag{27}$$

This leads to the optimal bias-variance tradeoff up to a constant multiplier. When $j_n = j_n^{\mathrm{cb}}$, the average width of the band in (24) is $w_n = O\left(M^{\frac{1}{2\beta+1}}(\sigma^2 \log n/n)^{\frac{\beta}{2\beta+1}}\right)$. By Theorem 2, such a band achieves over the class $\Lambda(\beta, M)$ the minimax rate for the average width of the band.

The band however involves the knowledge of β and M in finding the right level j^{cb} in (26) and in specifying w_n. Though this band is not adaptive, the above discussion suggests that we can obtain a data-driven adaptive confidence band by estimating the level j^{cb} and the half-width w_n based on data.

2.3   A Data-Driven Confidence Band

We are now in the position to construct a data-driven adaptive confidence band. To this end, we first give a scheme for selecting an appropriate projection level based on repeated hypothesis testing. After selecting such a level, we use an upper bound on the stochastic error of this estimator along with an estimate of its bias to choose the width of the band.

Data-Based Selection of Projection Level

We first define two levels $j^{\min} = j_n^{\min}$ and $j^{\max} = j_n^{\max}$ as the largest integers such that

$$2^{j^{\min}} \le \left\lceil \left(\frac{n}{\sigma^2 \log n}\right)^{\frac{1}{4\beta_0+1}} \right\rceil, \qquad 2^{j^{\max}} \le \left\lceil \left(\frac{c_\psi^2 M_0^2 \, n}{\sigma^2 \log n}\right)^{\frac{1}{2\beta_0+1}} \right\rceil. \tag{28}$$

Note that $j^{\min}$ and $j^{\max}$ are near optimal projection levels for the $\Lambda(2\beta_0, 1)$ and the $\Lambda(\beta_0, M_0)$ classes. For any other $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$, the corresponding projection level should be

14

sandwiched by these two extremes. Thus, we focus on those levels in the set

$$\mathcal{J} = [j^{\min}, j^{\max}] \cap \mathbb{N}. \tag{29}$$

Our goal is to construct an adaptive confidence band over the collection $\mathcal{A}(\beta_0, M_0)$ and so we are interested in confidence bands for functions f where $f \in \Lambda(\beta, M)$ for some $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$. In fact in most cases $f \in \Lambda(\beta, M)$ for an entire collection of $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$. Since bands corresponding to projection levels with smaller j values are narrower, our ideal projection level is the smallest j which satisfies (26) for some M and β where $f \in \Lambda(\beta, M)$ with $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$.

The actual selection proceeds as follows. We progressively search for the projection level in $\mathcal{J}$, starting at $j^{\min}$. Suppose we are now investigating some $j \in \mathcal{J}$. Then there exists some class $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$ such that the level j satisfies (26). In other words, j is optimal for $\Lambda(\beta, M)$. If the underlying function $f \in \Lambda(\beta, M)$, (14) implies that for all $j \le l < J$,

$$\max_k |\bar{\theta}_{lk}| \le c_\psi M \, 2^{-(\beta + \frac{1}{2})l} \le \left(c_\psi M \, 2^{-(\beta + \frac{1}{2})j} \right)^{\frac{l}{j}} \le \left(\frac{2\sigma^2 \log n}{n} \right)^{\frac{l}{2j}} \equiv c_{jl}. \tag{30}$$

Here, the second inequality holds because $c_\psi M \ge 1$, and the last inequality comes from (26). Thus, if we test the null hypotheses

$$H_{0,jl} : \max_{1 \le k \le 2^l} |\bar{\theta}_{lk}| \le c_{jl}, \tag{31}$$

for all $j \le l < J$, we should not reject any of them. If any of $H_{0,jl}, \, j \le l \le J$ is rejected, we move on to investigate the level $j + 1$ until $j = j^{\max}$. Otherwise, we select the current level j as our estimated projection level $\hat{j}^{\mathrm{cb}}$. If the current level is $j^{\max}$, we let $\hat{j}^{\mathrm{cb}} = j^{\max}$ directly.

We now spell out the details about how to test $H_{0,jl}$. For testing $H_{0,jl}$ for $j \le l < J$ we consider three test statistics

$$T_{0,jl} = \max_k |\hat{\theta}_{lk}|, \quad T_{1,jl} = \sum_{k=1}^{2^l} |\hat{\theta}_{lk}| I_{\{|\hat{\theta}_{lk}| > \tau_{jl}\}}, \quad T_{2,jl} = \sum_{k=1}^{2^l} |\hat{\theta}_{lk}| I_{\{|\hat{\theta}_{lk}| > \sigma_n\}} \tag{32}$$

where $\sigma_n = \sigma n^{-1/2}$ and $\tau_{jl} = c_{jl} + \sigma_n (l/2)^{1/2}$. To define the rejection rule, for any $a, t > 0$, let

$$\mu(a; t) = \phi(t + a) + \phi(t - a) + a[\Phi(t + a) - \Phi(t - a)], \tag{33}$$

15

where ϕ and Φ are the density and distribution functions of the standard normal distribution. Define events

$$R_{0,jl} = \left\{ T_{0,jl} > \sigma_n(\sqrt{3} + \sqrt{2})\sqrt{\log n} \right\},$$

$$R_{1,jl} = \left\{ \frac{T_{1,jl}}{\sigma_n} > 2^l \mu \left(\frac{c_{jl}}{\sigma_n}; \frac{\tau_{jl}}{\sigma_n} \right) + \left(\frac{2^l \log n}{4} \right)^{1/2} \left(\frac{c_{jl}}{\sigma_n} + \left(\frac{5 \log n}{2} \right)^{1/2} \right) \right\}, \qquad (34)$$

$$R_{2,jl} = \left\{ \frac{T_{2,jl}}{\sigma_n} > 2^l \mu \left(\frac{c_{jl}}{\sigma_n}; 1 \right) + \left(\left(1 + \frac{c_{jl}^2}{\sigma_n^2} \right) 2^l \log n \right)^{1/2} \right\}.$$

Finally, we test $H_{0,jl}$ according to the following rejection rule:

$$\text{We reject } H_{0,jl} \text{ on the event} \begin{cases} R_{0,jl} \cup R_{1,jl}, & \text{if } c_{jl} > \sigma_n(\log n)^{-1/2}, \\[2mm] R_{0,jl} \cup R_{2,jl}, & \text{otherwise.} \end{cases} \qquad (35)$$

Thus our estimated projection level is given by

$$\hat{j}^{\mathrm{cb}} = \min \left\{ j \in \mathcal{J} : H_{0,jl} \text{ is not rejected by (35) for } j \leq l < J \right\} \qquad (36)$$

with the convention that $H_{0,j^{\max}l}, j^{\max} \leq l < J$ are never rejected. We center our band at

$$\hat{f}_{\hat{j}^{\mathrm{cb}}}(t) = \sum_{l \leq \hat{j}^{\mathrm{cb}}} \sum_k \hat{\theta}_{lk} \psi_{lk}(t). \qquad (37)$$

Construction of the Band

We now specify the width of the band and to this end, we essentially need to provide estimators for the quantity on the righthand side of (25). The quantity is the sum of two terms, with the first term bounding the stochastic error and the second bounding the bias. In what follows, we deal with the two terms separately. For the first term, in order to accommodate the uncertainty of $\hat{j}^{\mathrm{cb}}$, we replace x_α by

$$x_{\alpha_n} = -\log\left(-\log\left(1 - \alpha_n\right)\right), \quad \text{with} \quad \alpha_n = \frac{\alpha}{|\mathcal{J}|}. \qquad (38)$$

Here, $|\mathcal{J}| = j^{\max} - j^{\min} + 1$ gives the cardinality of the set $\mathcal{J}$. Moreover, we replace all j_n by $\hat{j}^{\mathrm{cb}}$ and obtain the bound for this term as

$$\hat{w}_n^s = \mu_{\hat{j}^{\mathrm{cb}}} + \sigma_{\hat{j}^{\mathrm{cb}}} x_{\alpha_n}. \qquad (39)$$

For the second term in (25) note that it equals $\tau_\psi \sum_{l>j} c_\psi M 2^{-\beta l}$ and since we no longer know (β, M), it cannot be evaluated directly. However, (30) suggests that the summands can be bounded above by c_{jl}'s for all $l < J$. In addition, we multiply the partial sum from $\hat{j}^{\mathrm{cb}}$ to $J-1$ by a factor of 1.01 to cover the sum over those levels beyond $J-1$. This leads to the replacement of the second term by

$$\hat{w}_n^b = 1.01 \cdot \tau_\psi \sum_{l=\hat{j}^{\mathrm{cb}}+1}^{J-1} 2^{l/2} c_{\hat{j}^{\mathrm{cb}} l}. \tag{40}$$

Since τ_ψ can be evaluated numerically, $\hat{w}_n^b$ can be computed given $\hat{j}^{\mathrm{cb}}$.

Finally, the data-based confidence band is

$$[\hat{f}_{\hat{j}^{\mathrm{cb}}} - \hat{w}_n, \ \hat{f}_{\hat{j}^{\mathrm{cb}}} + \hat{w}_n], \qquad \text{with} \quad \hat{w}_n = \hat{w}_n^s + \hat{w}_n^b. \tag{41}$$

Here $\hat{j}^{\mathrm{cb}}$, $\hat{w}_n^s$ and $\hat{w}_n^b$ are given by (36), (39) and (40), respectively.

3 Performance of Confidence Band

Both the center and width of the confidence band given in Section 2 adjust to the underlying smoothness of the unknown function. We now look at the properties of the band providing theoretical properties, some simulation results as well as an application to some call center data. In the application to call center data we also indicate how the theory developed for Normal errors can be naturally extended to other settings.

3.1 Theoretical Properties

We now state theoretical properties of the confidence band (41). In particular, Theorem 1 below establishes that this band satisfies the requirements (8)–(10).

Theorem 1. *Suppose the wavelet basis satisfies Assumption (W). For any fixed $\beta_0 > \frac{1}{4}$ and $M_0 > 1$ and $\alpha \in (0,1)$, the confidence band (41) satisfies the area condition (8), the noncovered points condition (9) and the excess mass condition (10) simultaneously over the collection of function spaces $\mathcal{A}(\beta_0, M_0)$. Moreover, we could let $\xi_n = C(\log n)^{-\frac{\beta_0}{2(4\beta_0+1)}}$ in (9) and $\xi'_n = C(\beta_0, M_0)(\log n)^{-\frac{\beta_0}{4\beta_0+1}}$ in (10).*

The lower bound results given in Section 4 show that the confidence band (41) is optimally adaptive under both the set of noncovered points criterion (9) and the excess mass criterion (10).

Theorem 1 also implies the following adaptation result on average coverage. For a confidence band $CB = [L(t), U(t)]$ average coverage can be defined by

$$AC(CB, f) = \int_0^1 P_f(L(t) \leq f(t) \leq U(t))dt. \tag{42}$$

Note that

$$AC(CB, f) = E_f(1 - \mu(N(CB, f))) \tag{43}$$

Hence if $P_f(\mu(N(CB, f)) \leq \xi) \geq 1 - \alpha$, it follows that

$$AC(CB, f) = E_f(1 - \mu(N(CB, f))) \geq (1 - \epsilon)(1 - \alpha).$$

It is then easy to check that the adaptive confidence band has average coverage probability.

Corollary 1. *Under the assumptions of Theorem 1, the confidence band* (41) *satisfies*

$$\lim_{n \to \infty} \inf_{f \in \Lambda(\beta, M)} AC(CB, f) \geq 1 - \alpha$$

for all $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, R_0)$.

This together with the lower bound on the minimum average width given in Corollary 2 in Section 4 show that the confidence band (41) is also optimally adaptive under the average coverage criterion.

Remark 2. Bull (2011b) constructed a confidence band that was shown to achieve adaptive coverage over a collection of subsets of Lipschitz functions that also satisfy a self-similarity condition (Eq.(2.1) in Bull (2011b)). Moreover, it was shown that such subsets are in some sense large within the corresponding Lipschitz class $\Lambda(\beta, M)$ (Proposition 2.3 in Bull (2011b)). It is easy to see that the center of our band contains at least as many resolution levels as the center of the confidence band given in Bull (2011b). Using this fact, and following the lines of the proof to Theorem 3.3 in Bull (2011b), it can be shown that the confidence band proposed in the present paper also satisfies the conclusion of Theorem 3.3 in Bull (2011b) and hence has true adaptive coverage over self-similar subsets of $\Lambda(\beta, M)$ classes. That is,

our confidence band satisfies (9) and (10) with $\xi_n = \xi'_n = 0$ over the collections of self-similar Lipschitz functions. It is worth noting that the rate $O((\log n/n)^{\frac{\beta}{2\beta+1}})$ in (8) is the tightest possible for achieving true coverage on these subsets. See Theorem 3.4 of Bull (2011b).

3.2 Simulation Studies

The proposed adaptive confidence band is easily implementable. We report here the application of the proposed confidence band procedure to four test functions. The four functions are

Case 1 $\quad f(t) \propto B_{10,5}(t) + B_{7,7}(t) + B_{5,10}(t),$

Case 2 $\quad f(t) \propto 3B_{30,17}(t) + 2B_{3,11}(t),$

Case 3 $\quad f(t) \propto 7B_{15,30}(t) + 2\sin(32\pi t - \frac{2\pi}{3}) - 3\cos(16\pi t) - \cos(64\pi t),$

Case 4 $\quad f(t) \propto (t - \frac{1}{3})I(\frac{1}{3} \leq t \leq \frac{1}{2}) + (\frac{2}{3} - t)I(\frac{1}{2} \leq t \leq \frac{2}{3}),$

where $B_{a,b}(t)$ stands for the density function of a Beta(a, b) distribution. In all cases, we rescale the function such that $\int_0^1 f^2 = 1$. The test functions are plotted in Figure 1 as the black solid curves. As can be seen from the plots, the first three cases are smooth functions with decreasing level of smoothness. Case 4 has discontinuity in its first order derivative, and is included here as an attempt to defeat the procedure. Except Case 3, the other three cases have been previously used in Wahba (1983).

In each repetition, the data is generated from one test function according to model (1) with $n = 512$ and $\sigma = 0.25$. We then apply the band procedure with $1 - \alpha = 0.95$, $\beta_0 = 3$ and $M_0 = 100$ using a Symmlet 8 basis. Figure 1 shows for each case a typical realization of the observed data and the resulting band.

Table 1 summarizes the simulation results from 1000 repetitions. The first column (Non-coverage) reports the 95th percentile of the proportion of the points not covered by the band, and the second column (Relative excess) gives the 95th percentile of relative excess mass. If the band maintains the traditional coverage, then these two quantities should both be zeros. The third column reports the average size of the band. As reference, the last column gives the average ℓ_∞ distance from the band center to the true function.

From Table 1, we see that for the first two cases, our procedure seems to maintain tradi-
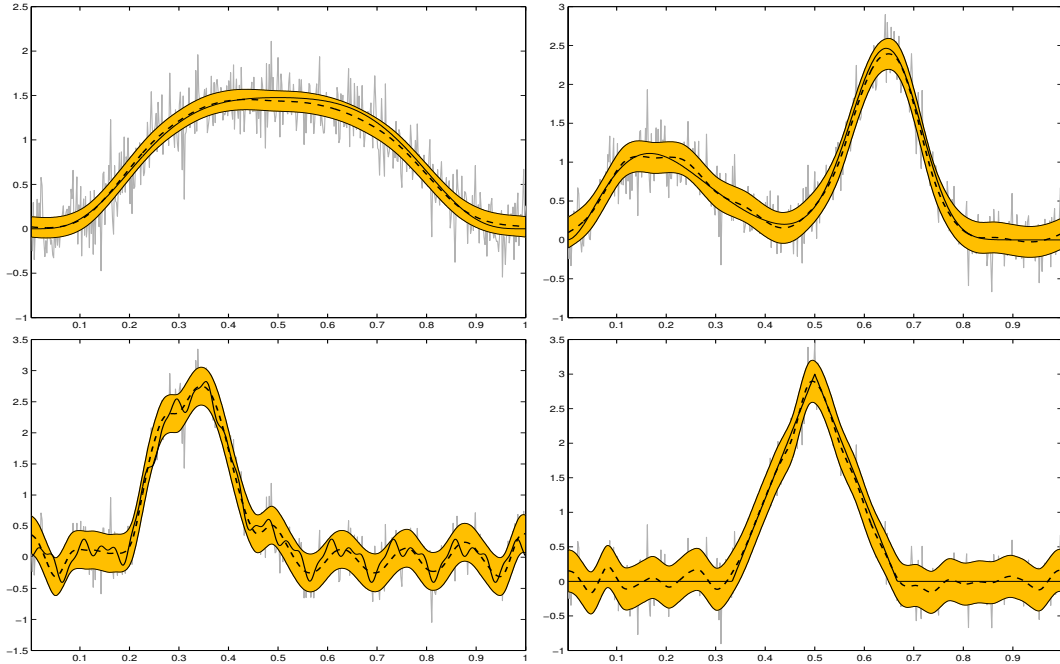
Figure 1: One realization of the observed data and the resulting band. Case 1: top left; Case 2: top right; Case 3: bottom left; Case 4: bottom right. Black solid: the true function. Gray: observed data. Orange: confidence band. Black dashed: band center.

	Non-coverage	Relative excess	Average size	Average ℓ_∞ loss
Case 1	0	0	0.2895	0.0702
Case 2	0	0	0.4841	0.1149
Case 3	$< 10^{-4}$	0.0039	0.7270	0.3051
Case 4	0.0005	0.0078	0.4935	0.2246

Table 1: Simulation results for confidence bands from 1000 repetitions: $\beta_0 = 3$, $M_0 = 100$.

tional coverage for both functions, while the band size adapts automatically to the smoothness of the function. For Case 3, we do not achieve traditional coverage, but both measure of non-covered points and relative excess mass are well controlled. Moreover, the average band size is larger than the first two cases as we expected. In the last case, though the function violates our assumptions, the measure of non-covered points and relative excess mass are still under control. However, the earlier theoretical results do not apply to Case 4, and the performance of the band could be worse for larger n. Last but not least, in each case, the average size of

	$\beta_0 = 2, M_0 = 100$	$\beta_0 = 2, M_0 = 200$	$\beta_0 = 3, M_0 = 200$
Case 1	0.3049	0.3126	0.2923
Case 2	0.5191	0.5282	0.4911
Case 3	0.7732	0.8170	0.7941
Case 4	0.5201	0.5295	0.5027

Table 2: Average sizes for confidence bands from different choices of (β_0, M_0).

the band is always within five times the average ℓ_∞ loss of the band center as an estimator of the function f.

The construction of the adaptive confidence band requires a choice for the parameters β_0 and M_0. To investigate the sensitivity of the proposed band to the choice of β_0 and M_0, we repeated the simulations reported above with three additional combinations of (β_0, M_0): $(2, 100)$, $(2, 200)$ and $(3, 200)$. All the other parameters remain the same. The resulting average sizes of the bands are reported in Table 2. These results indicate that the proposed band is not very sensitive to the choices of β_0 and M_0 values in terms of the average size. Other measures of performance also remain similar. Note that all the functions here are scaled to have unit L_2 norm. Thus, in practice, if no domain knowledge is available, we recommend setting β_0 to be either 2 or 3, and M_0 to be either 100 or 200 times a reasonable estimator of the L_2 norm of the underlying function, such as that in Cai and Low (2006b).

Confidence bands satisfying the three requirements (8), (9) and (10) can in theory also be constructed based on adaptive minimax L_2 confidence balls such as those given in Juditsky and Lambert-Lacroix (2003), Cai and Low (2006a), and Robins and van der Vaart (2006). See also Hoffman and Lepski (2002). Let $(\hat{f}_n, s_n(\alpha))$ denote an adaptive confidence ball with coverage $1 - \alpha$, where $\hat{f}_n$ is the center and $s_n(\alpha)$ is the radius. One could transform it into a confidence band

$$[\hat{f}_n - C_n(\alpha)\, s_n(\alpha/2),\ \hat{f}_n + C_n(\alpha)\, s_n(\alpha/2)].$$

Section 5.8 of Wasserman (2006) suggests that one can set $C_n(\alpha) = \sqrt{2/\alpha}$ to achieve average coverage. Since the requirements (9)–(10) are stronger than average coverage, the actual $C_n(\alpha)$ needed here has to be larger than $\sqrt{2/\alpha}$.

Table 3 summarizes the average value of $2 \times s_n(\alpha/2)$ with $\alpha = 0.05$ for confidence balls

	Cai and Low	Robins and van der Vaart
Case 1	0.7576	1.6425
Case 2	0.7948	1.6374
Case 3	0.8577	1.6643
Case 4	0.7906	1.6477

Table 3: Average values of $2 \times s_n(\alpha/2)$ for confidence balls from 1000 repetitions: $\alpha = 0.05$.

proposed by Cai and Low (2006a) and Robins and van der Vaart (2006) for the four test functions with $\beta_0 = 3$ and $M_0 = 100$[1]. For the method in Robins and van der Vaart (2006), we use the block thresholding estimator used in Cai and Low (2006a) as the center.

From Table 3, we find that the radii of the confidence balls seem to be less adaptive to the smoothness of the underlying signals compared to sizes of the proposed band in Table 1. By the above discussion, to make fair comparison to the third column of Table 1 in terms of magnitude, one needs to further multiply each number in Table 3 by a factor $C_n(\alpha) > \sqrt{2/\alpha} \doteq 6.3246$ when $\alpha = 0.05$. Thus, one can conclude that confidence bands obtained from transforming these confidence bands have much larger sizes than the one proposed in the current paper on these simulation examples.

In summary, the simulation results show the practicality of the proposed confidence band procedure, and seem to agree well with the earlier theoretical analysis. In addition, the resulting bands do not seem to be sensitive to the choices of β_0 and M_0 and perform favorably to confidence bands obtained by transforming confidence intervals.

3.3 Call Center Data

We now illustrate our confidence band on a real data example. The dataset consists of the arrival time of regular service calls to the call center of an Israeli bank from August to October in 1999 (Brown, et al., 2005). We assume that the arrival rate follows an inhomogeneous Poisson with mean $\mu(t)$. Our goal is to provide a confidence band for this mean function.

---

[1]The radius parameters used in both Cai and Low (2006a) and Robins and van der Vaart (2006) are for the sequence domain. In view of (11), we use $c_\psi M_0$ as the radius of the parameter spaces in the sequence domain when using the formulas in both papers.
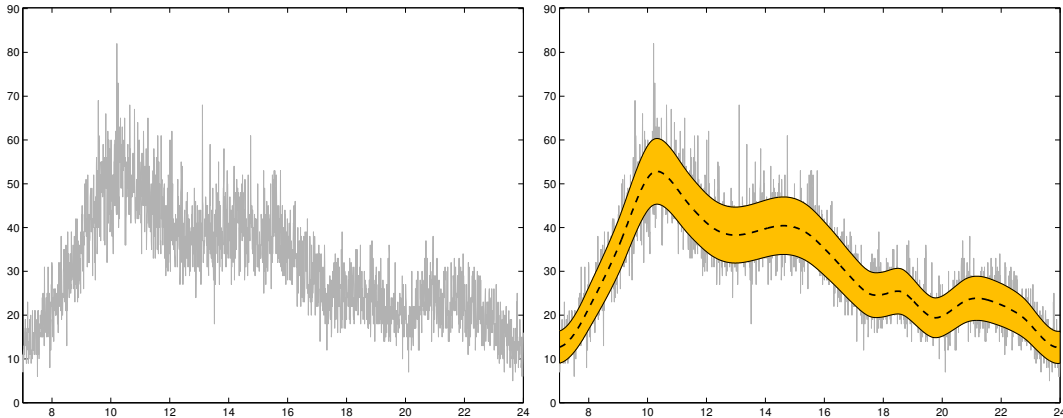
Figure 2: 95% confidence band for $\mu(t)$: original data (left panel); confidence band (right panel, orange) with band center in dashed line.

We first divide the daily operating time (7 AM – midnight) to $n = 2048$ equally spaced intervals. Let $N_i \sim \text{Poisson}(\mu(t_i))$ be the number of calls arriving in the ith interval. Then the transformed data

$$y_i = \sqrt{N_i + \tfrac{1}{4}}$$

approximately follows model (1) with $f(t) = \sqrt{\mu(t)}$. Then we compute the data-based band (41) with $1 - \alpha = 95\%$ using the y_i's and finally transform everything back by a square transform. For details about this root-unroot procedure, see Brown, et al. (2010). When computing (41), we used Symmlet 8 basis, $\beta_0 = 3$ and $M_0 = 100$. In addition, the noise standard deviation is estimated by $\hat{\sigma} = 1.4826 \times MAD(\hat{\theta}_{J-1,k})$ as suggested by Donoho and Johnstone (1994), where $\{\hat{\theta}_{J-1,k}\}$ are the empirical wavelet coefficient at the $J - 1$ level.

Figure 2 plots the confidence band for the mean function $\mu(t)$ of the inhomogeneous Poisson process used to model call arrival. From the plot, there is a clear peak of call arrival at around 10 AM, which was previously noted in Brown, et al. (2005). On the other hand, the second peak at around 3 PM is not as significant.

4 Lower Bounds For Miscoverage

This section examines the intrinsic difficulty of constructing confidence bands that are adaptive. First we give some bounds that explain why it is not possible to create adaptive bands

over any two Lipschitz classes that cover the entire function. This explains why we focus on adaptation while controlling excess mass or the measure of the points where the function is not covered. We then turn attention to bands which allow for some points where the function is not covered. Bounds given here show why adaptation must be limited to the range of Lipschitz classes considered in this paper.

4.1   Bounds For Bands Covering The Entire Function

Hall and Titteringon (1988) gave lower bounds for the maximum width of uniform confidence bands in the context of a function assumed to have a given number of derivatives. Recall that for uniform band, we write $U(t) - L(t) = \hat{w}_n$. In the case of the Lipschitz classes considered in the present paper their bound can be written as

$$\sup_{f \in \Lambda(\beta, M)} P_f \left(\hat{w}_n \geq \eta M^{\frac{1}{2\beta+1}} \left(\frac{\log n}{n} \right)^{\frac{\beta}{1+2\beta}} \right) \geq 1 - \alpha. \tag{44}$$

where $\eta > 0$ is a fixed constant not depending on β or M. Even though this lower bound is useful for evaluating the performance of the largest maximum width of a uniform confidence band for a given parameter space it is not sufficient for the goals of the present paper since a bound is needed for each f and not just for the supremum.

Our first collection of lower bounds concern bands that have guaranteed coverage for the entire function over a particular $\Lambda(\beta, M)$ class.

Theorem 2. *Suppose that the confidence band $CB = [L(t), U(t)] \in \mathcal{B}_\alpha(\Lambda(\beta, M))$ has a guaranteed coverage probability of $1 - \alpha$ over $\Lambda(\beta, M)$. Then there is a $C_1 > 0$ such that for all $\epsilon > 0$ there is an N such that for all $n > N$*

$$\sup_{f \in \Lambda(\beta, M)} P_f \left(\int_0^1 (U(t) - L(t)) dt \geq C_1 M^{\frac{1}{2\beta+1}} \left(\frac{\log n}{n} \right)^{\frac{\beta}{1+2\beta}} \right) \geq 1 - \alpha - \epsilon \tag{45}$$

and hence there is a $C_2 > 0$ such that for $n \geq N$,

$$\sup_{f \in \Lambda(\beta, M)} E_f \int_0^1 (U(t) - L(t)) dt \geq C_2 M^{\frac{1}{2\beta+1}} \left(\frac{\log n}{n} \right)^{\frac{\beta}{1+2\beta}}. \tag{46}$$

For each $f \in \Lambda(\beta, M')$ with $M' < M$, there is a $C > 0$ and $a > 0$ such that for all n,

$$P_f \left(\int_0^1 (U(t) - L(t)) dt \geq C n^{-\frac{\beta}{1+2\beta}} \right) \geq a. \tag{47}$$

24

Finally for a uniform band $[L(t), U(t)]$ with $U(t) - L(t) = \hat{w}_n$ then for each $f \in \Lambda(\beta, M')$ with $M' < M$, there is a $C > 0$ and $a > 0$ such that for all n

$$P_f\left(\hat{w}_n \geq C\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+2\beta}}\right) > a. \tag{48}$$

The bounds given in this Theorem, particularly those of (47) and (48) show that, for confidence bands that have honest coverage of the entire function, it is not possible to adapt over any pair of Lipschitz classes $\Lambda(\beta_1, M_1)$ and $\Lambda(\beta_2, M_2)$ whenever $\beta_1 \neq \beta_2$. It is for this reason that we have allowed the band to have points where the function is not covered.

4.2 Lower Bounds For Set of Noncovered Points and Excess Mass

As mentioned the lower bounds given in the previous section rule out the construction of adaptive confidence bands that have coverage for the entire band. This does not rule out adaptation of bands in the sense of covering the function at most points.

We shall now establish lower bounds for confidence bands under both the set of noncovered points criterion (4) and the excess mass criterion (6). These lower bounds yield directly the limits on the range of Lipschitz classes over which adaptation is possible under either criterion.

Theorem 3. *Suppose that a confidence band either satisfies*

$$\inf_{g \in \Lambda(\beta, M)} P_g(\mu(N(CB, g)) < \frac{1}{2} - \epsilon) \geq 1 - \alpha \tag{49}$$

where $\epsilon > 0$ and $\alpha < \frac{1}{2}$ or satisfies for some $r > 0$

$$\inf_{g \in \Lambda(\beta, M)} P_g(RE(CB, g) \leq r) \geq 1 - \alpha. \tag{50}$$

Then for all $h \in \Lambda(\beta, M')$ with $M' < M$, there is a $c > 0$ (which may depend on h) such that

$$w(CB, h) \geq cn^{-\frac{2\beta}{1+4\beta}}. \tag{51}$$

Remark 3. It is useful to compare Theorem 2 and Theorem 3. Theorem 2 rules out adaptation over any pair of Lipschitz classes whereas Theorem 3 rules out adaptation over any pair of Lipschitz classes $\Lambda(\beta_0, M_0)$ and $\Lambda(\beta_1, M_1)$ whenever $\beta_1 > 2\beta_0$. More precisely suppose that $\beta_1 > 2\beta_0$ and that

$$w(CB, \Lambda(\beta_1, M_1)) \leq C\left(\frac{\log n}{n}\right)^{\frac{\beta_1}{2\beta_1+1}} \tag{52}$$

and so the confidence band has width achieving the minimax bound for the class $\Lambda(\beta_1, M_1)$. Theorem 3 then shows that there is some function $f \in \Lambda(\beta_0, M_0)$ such that for sufficiently large n

$$P_f\left(\mu(N(CB, f)) \geq \frac{1}{2}\right) \geq \frac{1}{2}.$$

That is, there is better than 50% of chance that the confidence band misses the function over more than half of the interval $[0, 1]$. Moreover there also exists $f \in \Lambda(\beta_0, M_0)$, such that

$$P_f\left(RE(CB, h) \geq 1\right) \geq \frac{1}{2}.$$

That is, with probability at least more than 50%, the integrated excess mass is at least as large as the area of the band. Therefore, adaptation is still impossible over Lipschitz classes $\Lambda(\beta, M)$ for $\beta \in [\beta_0, \ \beta_1]$ with $\beta_1 > 2\beta_0 > 0$ even under the more relaxed coverage constraint (4) or under the excess mass constraint (6).

Remark 4. It is also possible to give bounds on the average coverage probability defined in (42). Note once more that $AC(CB, f) = E_f(1 - \mu(N(CB, f)))$. Hence if $AC(CB, f) \geq 1 - \alpha$ it follows that

$$1 - \alpha \leq E_f(1 - \mu(N(CB, f))) \leq (1 - c)P_f(\mu(N(CB, f)) > c) + P_f(\mu(N(CB, f)) \leq c).$$

Hence

$$P_f(\mu(N(CB, f)) > c) \leq \frac{\alpha}{c}$$

or alternatively

$$P_f(\mu(N(CB, f)) \leq c) \geq 1 - \frac{\alpha}{c}.$$

The following corollary gives a bound on the minimum average width of such a confidence band.

Corollary 2. *Suppose that the confidence band CB has average coverage probability of at least $1 - \alpha$ over $\Lambda(\beta_0, M_0)$. Then for all $g \in \Lambda(\beta_0, M)$ with $M < M_0$ it follows that there is a $c > 0$ such that*

$$w(CB, g) \geq cn^{-\frac{2\beta_0}{1+4\beta_0}}. \tag{53}$$

Since $n^{-\frac{2\beta_0}{1+4\beta_0}} \gg \left(\frac{\log n}{n}\right)^{\frac{\beta_1}{1+2\beta_1}}$ whenever $\beta_1 > 2\beta_0$, this corollary shows that even under this criteria adaptation over Lipschitz classes is still ruled out whenever $\beta_1 > 2\beta_0$.

Theorem 3 shows that for an assumed minimum smoothness parameter $\beta_0 > 0$, the best one can hope for is to construct confidence bands that are adaptive over the Lipschitz classes $\Lambda(\beta, M)$ for $\beta \in [\beta_0, \ 2\beta_0]$, under either the set of noncovered points criterion (4) and the excess mass criterion (6). We should note that such limitation also occurs in the construction of adaptive confidence balls. See, for example, Cai and Low (2006a) and Robins and van der Vaart (2006).

5 Conclusion and Discussion

One of the primary goals of the present paper is to introduce a concrete confidence band which not only fits our new theoretical framework but also works well for relatively small to moderate sample sizes. As mentioned in Section 3.2, confidence bands satisfying the three requirements (8), (9) and (10) can in theory also be constructed based on adaptive minimax L_2 confidence balls such as those given in Juditsky and Lambert-Lacroix (2003), Cai and Low (2006a), and Robins and van der Vaart (2006). However bands constructed that way appear to be more of theoretical interest rather than practical use. In particular the procedure in Juditsky and Lambert-Lacroix (2003) involves an unspecified tuning parameter and is not readily implementable. The simulation study in Section 3.2 also demonstrates the favorable performance of our proposed procedure over the bands transformed from the L_2 confidence balls given in Cai and Low (2006a) and Robins and van der Vaart (2006).

It is worth noting that the band procedure proposed in Section 2 does not depend on the sequences ξ_n and ξ_n' used in (9) and (10) and it also maintains true coverage over self-similar Lipschitz functions. Finding the optimal rates of convergence for these two sequences subject to condition (8) is an interesting and open theoretical problem that is beyond the scope of the present paper.

The confidence band that was developed in this paper was for periodic regression functions based on a nonparametric regression with Gaussian noise model. However as illustrated by the call center data example existing techniques in the literature can help to transform more complex data sets to settings where our procedure is still appropriate. This will for example include non-periodic functions as well as other data generating distributions.

In settings where the underlying function is not periodic boundary corrected wavelet bases

developed by Cohen, et al. (1993) can replace the periodized wavelet bases considered in the present paper. Cohen, et al. (1993) constructed boundary corrected orthonormal wavelets on $[0, 1]$ with 2^j wavelet functions at resolution level j. The wavelets have the same vanishing moments property as the wavelets on the whole line. The boundary correction affects only a fixed number of wavelet coefficients at each resolution level and the corresponding discrete wavelet transform introduces correlations to these wavelet coefficients. See Cohen, et al (1993) for more on boundary corrected wavelet bases. The required modification for the adaptive confidence band procedure is minor.

In the present paper we have focused on nonparametric regression with Gaussian noise. The method can be extended to a number of other nonparametric models. For nonparametric regression with an unknown noise distribution that is possibly heavy-tailed, the local median transformation introduced in Brown, Cai, and Zhou (2008) and Cai and Zhou (2009) can be used to transform the problem into a standard nonparametric regression with Gaussian noise. A key step is a local median transformation, where the original observations are first divided into small groups with the same number of observations in each group and then the medians of the data in these groups are taken as a new data set. The central idea is that the new data set can be well approximated by Gaussian random variables for a wide collection of noise distributions. After the local median transformation, the confidence band introduced in the present paper which is designed for Gaussian noise can then be applied to the new data set. All the claims still hold with minor changes to the proofs.

Similar ideas can also be used to construct confidence bands for nonparametric density estimation by using the root-unroot transformation introduced in Brown, et al (2010). In addition, the confidence band procedure introduced in this paper can be generalized for nonparametric regression in exponential families such as nonparametric Poisson regression and binomial regression by using the mean-matching variance stabilizing transformation. See Brown, Cai and Zhou (2010) and Cai and Zhou (2010).

In the present paper we have focused on the construction of uniform bands. An interesting topic for future investigation is the construction of variable width bands that also achieve spatial adaptivity.

6   Proofs

In this section we provide the proofs of all the main results. In section 6.1 we provide a proof of Theorem 1 which gives performance guarantees on the performance of the adaptive confidence procedure described in Section 2. In Section 6.2 we turn to the proof of Theorem 3 which gives lower bounds for confidence bands which cover most points or have small excess mass. The proof of Theorem 2 which gives lower bounds for confidence bands with guaranteed coverage of the entire function is given in the supplement.

6.1   Proof of Upper Bounds

We first introduce a couple of propositions describing the performance of the tests used to construct the projection estimator which is used as the center of the confidence band. We then turn to a proof of Theorem 1.

6.1.1   Testing Propositions

We now prove Theorem 1 and Corollary 1. To this end, we first introduce two propositions which give non-asymptotic bounds for the probabilities of type I error and powers of the test (35). The proofs of both propositions are given in the supplement.

Recall that $\sigma_n = \sigma n^{-1/2}$. In what follows, for $i = 0, 1$, $f \in H_{i,jl}$ means that the wavelet coefficients of both f and f_n at the lth resolution level satisfy the statement in $H_{i,jl}$, respectively. The first proposition deals with excess mass type alternatives.

Proposition 2. *Let j satisfy (26) for some $\beta \in [\beta_0, 2\beta_0]$ and $M \in [1, M_0]$. Consider testing $H_{0,jl}$ (31) against*

$$H_{1,jl} : \sum_{k=1}^{2^l} \left(|\bar{\theta}_{lk}| - c_{jl} \right)_+ > e_{jl}, \tag{54}$$

where for a sufficiently large constant C,

$$e_{jl} = \begin{cases} C\,\sigma_n 2^l (\log n)^{-\frac{1}{2}}, & \text{if } 2^{-(\beta+\frac{1}{2})(l-j)} \in [(\log n)^{-1}, 1], \\ C\,2^l c_{jl}, & \text{if } 2^{-(\beta+\frac{1}{2})(l-j)} \in [2^{-\frac{l}{4}} l^{-\frac{1}{4}}, (\log n)^{-1}), \\ C\,\sigma_n 2^{\frac{3l}{4}} l^{\frac{1}{4}}, & \text{if } 2^{-(\beta+\frac{1}{2})(l-j)} \in (0, 2^{-\frac{l}{4}} l^{-\frac{1}{4}}). \end{cases} \tag{55}$$

Let $\phi_{jl} \in \{0,1\}$ be the test specified by (35) with $\phi_{jl} = 1$ for rejection. Then there exists another absolute constant C', s.t.

$$\sup_{j \le l \le j^{\mathrm{t}}} \sup_{f \in H_{0,lj}} \mathsf{E}_f \phi_{jl} \le C' n^{-\frac{1}{2}}, \tag{56}$$

$$\inf_{j \le l \le j^{\mathrm{t}}} \inf_{f \in H_{1,lj}} \mathsf{E}_f \phi_{jl} \ge 1 - C' 2^{-\frac{l}{2}}. \tag{57}$$

The next proposition deals with noncovered points type alternatives. For any set A, we use $|A|$ to denote its cardinality. Moreover, let

$$j^{\mathrm{t}} = 2j^{\min} + \frac{1}{4\beta_0 + 1} \log_2 \log n + \frac{4}{4\beta_0 + 1} \left[\log_2 M_0 - \log_2(1 - 2^{-\beta_0}) \right]. \tag{58}$$

Note that when $l > j^{\mathrm{t}}$, $c_\psi M_0 2^{-(\beta_0 + \frac{1}{2})l} \le C\sigma_n 2^{-l/4} l^{1/4}$. Moreover, since $\beta_0 > 1/4$, we have $j^{\mathrm{t}} < J$, at least for sufficiently large n.

Proposition 3. *Let j satisfy (26) for some $\beta \in [\beta_0, 2\beta_0]$ and $M \in [1, M_0]$. Consider testing $H_{0,jl}$ (31) against*

$$H'_{1,jl} : 2^{-l} |\{\bar{\theta}_{lk} : |\bar{\theta}_{lk}| > \tilde{c}_{jl}\}| > \kappa_{jl}, \tag{59}$$

where $\tilde{c}_{jl} = (\gamma_{jl} + 1)(c_{jl} \vee \sigma n^{-\frac{1}{2}} 2^{-\frac{l}{4}} l^{\frac{1}{4}})$ with

$$\gamma_{jl} = \begin{cases} (\log n)^{-\frac{1}{4}}, & \text{if } 2^{-(\beta + \frac{1}{2})(l-j)} \in [(\log n)^{-\frac{1}{2}}, 1], \\ 2^{\frac{1}{2}\beta_0(l-j)}, & \text{if } 2^{-(\beta + \frac{1}{2})(l-j)} \in [2^{-\frac{l}{4}} l^{-\frac{1}{4}}, (\log n)^{-\frac{1}{2}}), \\ (\log n)^{\frac{\beta_0}{2(4\beta_0+1)}} 2^{\frac{1}{8}(j^{\mathrm{t}} - l)}, & \text{if } 2^{-(\beta + \frac{1}{2})(l-j)} \in (0, 2^{-\frac{l}{4}} l^{-\frac{1}{4}}), \end{cases} \tag{60}$$

and for a sufficiently large constant C

$$\kappa_{jl} = \begin{cases} C\sigma_n c_{jl}^{-1} (\log n)^{-\frac{1}{4}}, & \text{if } 2^{-(\beta + \frac{1}{2})(l-j)} \in [(\log n)^{-\frac{1}{2}}, 1], \\ C\gamma_{jl}^{-1} (1 \wedge \sigma_n c_{jl}^{-1} (\log n)^{-\frac{1}{2}}), & \text{otherwise.} \end{cases}$$

Let $\phi_{jl} \in \{0,1\}$ be the test specified by (35) with $\phi_{jl} = 1$ for rejection. Then there exists another absolute constant C', s.t.

$$\sup_{j \le l \le j^{\mathrm{t}}} \sup_{f \in H_{0,lj}} \mathsf{E}_f \phi_{jl} \le C' n^{-\frac{1}{2}}, \tag{61}$$

$$\inf_{j \le l \le j^{\mathrm{t}}} \inf_{f \in H'_{1,lj}} \mathsf{E}_f \phi_{jl} \ge 1 - C' 2^{-\frac{l}{2}}. \tag{62}$$

6.1.2 Proof of Theorem 1

We divide the proof into three parts. First, we verify the average area condition (8). Then, we prove that the excess mass condition (10) is satisfied. Finally, we come back to verify the noncovered points condition (9), which uses some intermediate results in the proof of (10).

1°. We first verify the area condition (8). Fix a function class $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$. For this class, let $j^{cb} = j_n^{cb}$ satisfy (26) and hence (27).

Note that $\hat{j}^{cb} \in \mathcal{J}$ and that $j^{\min}, j^{\max} \asymp \log n$. By (41), (18), (19) and (20), the width, and hence the area, of the band (41) is of order $O\big(\sigma 2^{\hat{j}^{cb}/2}(\log n/n)^{1/2}\big)$. Thus, to verify (8), it suffices to show that uniformly over $\Lambda(\beta, M)$, $\hat{j}^{cb} \leq j^{cb} + l_\psi$ with sufficiently high probability, where l_ψ is a positive integer depending only on the wavelet basis.

Note that $f \in \Lambda(\beta, M)$ implies $f \in H_{0,jl}$ for all pairs (j, l) where $j \leq j^{cb} + l_\psi$ and $l \geq j$. Since $j^{\min}, j^{\max}, J \asymp \log n$, there are at most $O((\log n)^2)$ hypotheses testing when we obtain $\hat{j}^{cb}$. Thus, (56) and (61), together with the union bound, ensure that with probability at least $1 - C'n^{-\frac{1}{2}}(\log n)^2$, we have $\hat{j}^{cb} \leq j^{cb} + l_\psi$. So, for any n,

$$\sup_{f \in \Lambda(\beta, M)} P_f(\hat{j}^{cb} \leq j^{cb} + l_\psi) \geq 1 - C'n^{-\frac{1}{2}}(\log n)^2.$$

Therefore, we have

$$\sup_{f \in \Lambda(\beta, M)} \mathsf{E}_f \int_0^1 [U(t) - L(t)]dt$$

$$\leq C\sigma \left(\frac{\log n}{n}\right)^{\frac{1}{2}} \sup_{f \in \Lambda(\beta, M)} \left[\mathsf{E}_f\left(2^{\hat{j}^{cb}/2}I_{\{\hat{j}^{cb} \leq j^{cb} + l_\psi\}}\right) + \mathsf{E}_f\left(2^{\hat{j}^{cb}/2}I_{\{\hat{j}^{cb} > j^{cb} + l_\psi\}}\right)\right]$$

$$\leq C\sigma \left(\frac{\log n}{n}\right)^{\frac{1}{2}} \sup_{f \in \Lambda(\beta, M)} \left[\mathsf{E}_f\left(2^{(j^{cb}+l_\psi)/2}I_{\{\hat{j}^{cb} \leq j^{cb} + l_\psi\}}\right) + \mathsf{E}_f\left(2^{j^{\max}/2}I_{\{\hat{j}^{cb} > j^{cb} + l_\psi\}}\right)\right]$$

$$\leq C\sigma \left(\frac{\log n}{n}\right)^{\frac{1}{2}} \left[2^{(j^{cb}+l_\psi)/2} + 2^{j^{\max}/2}C'n^{-\frac{1}{2}}(\log n)^2\right]$$

$$\leq C_\psi\sigma \left(\frac{\log n}{n}\right)^{\frac{1}{2}} \left(\frac{M^2 n}{\sigma^2 \log n}\right)^{\frac{1}{4\beta+2}} = C_\psi M^{\frac{1}{2\beta+1}} \left(\frac{\sigma^2 \log n}{n}\right)^{\frac{\beta}{2\beta+1}}.$$

Here, the last inequality holds because $2^{j^{\max}/2}n^{-1/2}(\log n)^2 \leq 1 \leq 2^{j^{cb}/2}$ when $n \geq n_0(\beta_0, M_0, \sigma)$. This completes the verification of the area condition.

2°. In the second step, we verify the excess mass condition (10). Recall $\hat{w}_n^s$ in (39) and

$\hat{w}_n^b$ in (40), where for $f_{\hat{j}^{\mathrm{cb}}} = \sum_{l \leq \hat{j}^{\mathrm{cb}}} \sum_k \bar{\theta}_{lk} \psi_{lk}$, $\hat{w}_n^s$ is intended to bound the stochastic error $\|\hat{f}_{\hat{j}^{\mathrm{cb}}} - f_{\hat{j}^{\mathrm{cb}}}\|_\infty$ and $\hat{w}_n^b$ to bound the bias $\|f - f_{\hat{j}^{\mathrm{cb}}}\|_\infty$.

Let $\bar{e}_f = (f - U)_+$ be the excess mass exceeding the upper limit, and $\underline{e}_f = (L - f)_+$ the excess mass exceeding the lower limit. Note that $f - U = (f - f_{\hat{j}^{\mathrm{cb}}} - \hat{w}_n^b) + (f_{\hat{j}^{\mathrm{cb}}} - \hat{f}_{\hat{j}^{\mathrm{cb}}} - \hat{w}_n^s)$. Since $(a + b)_+ \leq a_+ + b_+$, this leads to

$$\bar{e}_f \leq (f - f_{\hat{j}^{\mathrm{cb}}} - \hat{w}_n^b)_+ + (f_{\hat{j}^{\mathrm{cb}}} - \hat{f}_{\hat{j}^{\mathrm{cb}}} - \hat{w}_n^s)_+ \equiv \bar{e}_f^b + \bar{e}_f^s. \tag{63}$$

As before, the superscript s stands for stochastic error and b for bias. In what follows, we bound $\bar{e}_f$ by controlling $\bar{e}_f^b$ and $\bar{e}_f^s$ separately. A completely analogous argument will lead to the same bound for $\underline{e}_f$.

Now fix a class $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$, and pick any $f \in \Lambda(\beta, M)$. Define the event

$$E_f = \{\|f_{\hat{j}^{\mathrm{cb}}} - \hat{f}_{\hat{j}^{\mathrm{cb}}}\|_\infty \leq \hat{w}_n^s, \text{ and all } H_{0,jl} \text{ vs. } H_{1,jl} \text{ are tested correctly}\}. \tag{64}$$

On this event, we have $\bar{e}_f^s = 0$. From now on, we focus on controlling $\bar{e}_f^b$.

We start with a simple case. If f also belongs to $\Lambda(2\beta_0, 1)$, then $\hat{j}^{\mathrm{cb}} \geq j^{\min}$. Thus, for large n, $\hat{w}_n^b$ is no less than the rightmost side of (23), and so $\bar{e}_f^b = 0$.

When $f \notin \Lambda(2\beta_0, 1)$, let $\hat{j}^{\mathrm{cb}}$ satisfy (26) for some $\hat{\beta} \in [\beta_0, 2\beta_0]$ and $\hat{M} \in [1, M_0]$. Moreover, let $c_{\hat{j}^{\mathrm{cb}}l}$ be defined as in (30) with j replaced by $\hat{j}^{\mathrm{cb}}$ for $\hat{j}^{\mathrm{cb}} \leq l < J$, and 0 otherwise. By (22) and (40), we obtain

$$\hat{w}_n^b \geq \sum_{l > \hat{j}^{\mathrm{cb}}} \sum_k c_{\hat{j}^{\mathrm{cb}}l} |\psi_{lk}(t)|.$$

Recall θ_{lk} and $\bar{\theta}_{lk}$ defined in Section 2.1. With slight abuse of notation, we define $\bar{\theta}_{lk} = \theta_{lk}$ for all $l \geq J$. Thus, $|f(t) - f_{\hat{j}^{\mathrm{cb}}}(t)| \leq \sum_{l > \hat{j}^{\mathrm{cb}}} \sum_k |\bar{\theta}_{lk}| |\psi_{lk}(t)|$. Hence,

$$\int_0^1 \bar{e}_f^b(t) dt \leq \int_0^1 \left[\sum_{l > \hat{j}^{\mathrm{cb}}}^\infty \sum_k |\bar{\theta}_{lk}| |\psi_{lk}(t)| - \sum_{l > \hat{j}^{\mathrm{cb}}}^\infty \sum_k c_{\hat{j}^{\mathrm{cb}}l} |\psi_{lk}(t)| \right]_+ dt.$$

We apply the inequality $(a + b)_+ \leq a_+ + b_+$ repeatedly to further bound the right side by

$$\sum_{l > \hat{j}^{\mathrm{cb}}} \sum_k \int_0^1 (|\bar{\theta}_{lk}| |\psi_{lk}| - \sum_k c_{\hat{j}^{\mathrm{cb}}l} |\psi_{lk}|)_+ \leq \sum_{l > \hat{j}^{\mathrm{cb}}} \sum_k (|\bar{\theta}_{lk}| - c_{\hat{j}^{\mathrm{cb}}l})_+ \int_0^1 |\psi_{lk}|.$$

Further note that $\int_0^1 |\psi_{lk}(t)| dt \leq 2^{-l/2} \|\psi\|_1$ where ψ is the mother wavelet and $\|\psi\|_1 = \int_{\mathbb{R}} |\psi(t)| dt$. The last two displays thus lead to

$$\int_0^1 \bar{e}_f^b(t) dt \leq \|\psi\|_1 \sum_{l > \hat{j}^{\mathrm{cb}}} 2^{-\frac{l}{2}} \sum_k (|\bar{\theta}_{lk}| - c_{\hat{j}^{\mathrm{cb}}l})_+. \tag{65}$$

32

To further bound the right side of (65), we divide the resolution levels above $\hat{j}^{\mathrm{cb}}$ into three parts as $\{l : l > \hat{j}^{\mathrm{cb}}\} = \mathcal{J}_1 \cup \mathcal{J}_2 \cup \mathcal{J}_3$, where

$$
\begin{aligned}
\mathcal{J}_1 &= \{l : 2^{-(\hat{\beta}+\frac{1}{2})(l-\hat{j}^{\mathrm{cb}})} \in [(\log n)^{-1}, 1)\}, \\
\mathcal{J}_2 &= \{l : 2^{-(\hat{\beta}+\frac{1}{2})(l-\hat{j}^{\mathrm{cb}})} \in [2^{-\frac{l}{4}}l^{-\frac{1}{4}}, (\log n)^{-1})\}, \\
\mathcal{J}_3 &= \{l : 2^{-(\hat{\beta}+\frac{1}{2})(l-\hat{j}^{\mathrm{cb}})} \in (0, 2^{-\frac{l}{4}}l^{-\frac{1}{4}})\}.
\end{aligned}
\tag{66}
$$

In what follows, we bound the sum in (65) over each $\mathcal{J}_i$ separately. For notational convenience, we let

$$
\hat{\omega}_n = 2^{\hat{j}^{\mathrm{cb}}/2}\left(\frac{\sigma^2 \log n}{n}\right)^{1/2}.
\tag{67}
$$

For $\mathcal{J}_1$, on E_f, Proposition 2 leads to

$$
I \equiv \sum_{\mathcal{J}_1} 2^{-\frac{l}{2}} \sum_k (|\bar{\theta}_{lk}| - c_{\hat{j}^{\mathrm{cb}}l})_+ \leq C\sigma_n(\log n)^{-\frac{1}{2}} \sum_{\mathcal{J}_1} 2^{\frac{l}{2}}.
$$

The definition of $\mathcal{J}_1$ implies that $l - \hat{j}^{\mathrm{cb}} \leq 2(2\hat{\beta}+1)^{-1} \log_2 \log n$, and so for all $l \in \mathcal{J}_1$, $2^{l/2} \leq 2^{\hat{j}^{\mathrm{cb}}/2}(\log n)^{1/(2\hat{\beta}+1)}$. Since $\{2^{l/2}\}$ is a geometric increasing sequence, the last display implies

$$
I \leq C 2^{\frac{\hat{j}^{\mathrm{cb}}}{2}} \sigma_n(\log n)^{\frac{1}{2\hat{\beta}+1}-\frac{1}{2}} \leq C\hat{\omega}_n(\log n)^{-\frac{2\hat{\beta}}{2\hat{\beta}+1}}.
\tag{68}
$$

For $\mathcal{J}_2$, on E_f, Proposition 2 leads to

$$
II \equiv \sum_{\mathcal{J}_2} 2^{-\frac{l}{2}} \sum_k (|\bar{\theta}_{lk}| - c_{\hat{j}^{\mathrm{cb}}l})_+ \leq C \sum_{\mathcal{J}_2} 2^{\frac{l}{2}} c_{\hat{j}^{\mathrm{cb}}l} \leq C\hat{\omega}_n(\log n)^{-\frac{4\beta_0}{4\beta_0+1}}.
\tag{69}
$$

Here, the second inequality holds because the summands in the middle term is geometrically decreasing which is implied by $\hat{\beta} \geq \beta_0 > 1/4$.

Turn to $\mathcal{J}_3$. Since $f \in \Lambda(\beta, M)$, Proposition 2 and (11) imply that

$$
2^{-\frac{l}{2}} \sum_{k=1}^{2^l} (|\bar{\theta}_{lk}| - c_{\hat{j}^{\mathrm{cb}}l})_+ \leq \begin{cases} C\sigma_n 2^{\frac{l}{4}}l^{\frac{1}{4}}, & \text{for all } l \leq j^{\mathrm{t}}, \\ c_\psi M\, 2^{-\beta l}, & \text{for all } l \in \mathcal{J}_3. \end{cases}
$$

Consider the critical level

$$
l_1 = \frac{4}{4\beta+1}\log_2\left(\frac{M}{1-2^{-\beta}}\right) + \frac{1}{4\beta+1}\log_2 \log n + \frac{2}{4\beta+1}\log_2\left(\frac{n}{\sigma^2 \log n}\right).
\tag{70}
$$

Since $\beta \geq \beta_0 > 1/4$ and $M \leq M_0$, we obtain $l_1 \leq j^{\mathrm{t}} < J$. Thus, we have

$$
III \equiv \sum_{\mathcal{J}_3} 2^{-\frac{l}{2}} \sum_k (|\theta_{lk}| - c_{\hat{j}^{\mathrm{cb}}l})_+ \leq \sum_{\mathcal{J}_3 \ni l \leq l_1} C\sigma_n 2^{\frac{l}{4}}l^{\frac{1}{4}} + \sum_{l > l_1} c_\psi M\, 2^{-\beta l}.
$$

Further note that $M\,2^{-\beta l_1}/(1-2^{-\beta}) \asymp \sigma_n 2^{\frac{l_1}{4}} l_1^{\frac{1}{4}} \leq C(\log n)^{-\frac{\beta}{4\beta+1}} M^{\frac{1}{4\beta+1}} (\sigma^2 \log n/n)^{\frac{2\beta}{4\beta+1}}$. We thus bound the right side of the last display to obtain

$$III \leq C\,\sigma_n 2^{\frac{l_1}{4}} l_1^{\frac{1}{4}} + \frac{c_\psi M\,2^{-\beta l_1}}{1-2^{-\beta}} \leq C(\log n)^{-\frac{\beta}{4\beta+1}} M^{\frac{1}{4\beta+1}} \left(\frac{\sigma^2 \log n}{n}\right)^{\frac{2\beta}{4\beta+1}}. \qquad (71)$$

We now assemble (68), (69) and (71) to bound the right side of (65). Note that $\hat{\beta} \leq 2\beta_0$, and so $2\hat{\beta}/(2\hat{\beta}+1) < 4\beta_0/(4\beta_0+1)$. Moreover, $\beta \geq \beta_0$ leads to $(\sigma^2 \log n/n)^{\frac{2\beta}{4\beta+1}} \leq (\sigma^2 \log n/n)^{\frac{2\beta_0}{4\beta_0+1}} \leq C2^{j^{\min}/2}(\sigma^2 \log n/n)^{\frac{1}{2}} \leq C\hat{\omega}_n$. Therefore, we obtain

$$\int_0^1 \bar{e}_f^b(t)dt \leq I + II + III \leq C_\psi[(\log n)^{-\frac{2\hat{\beta}}{2\hat{\beta}+1}} + (\log n)^{-\frac{\beta}{4\beta+1}} M^{\frac{1}{4\beta+1}}]\,\hat{\omega}_n.$$

On the other hand, the area of the band $\int_0^1[U(t) - L(t)]dt \geq \hat{w}_n^b \geq C\tau_\psi \hat{\omega}_n$. On the event E_f, $\bar{e}_f = \bar{e}_f^b$, and so

$$\frac{\int_0^1 \bar{e}_f(t)dt}{\int_0^1[U(t)-L(t)]dt} \leq C_\psi[(\log n)^{-\frac{2\hat{\beta}}{2\hat{\beta}+1}} + (\log n)^{-\frac{\beta}{4\beta+1}} M^{\frac{1}{4\beta+1}}] \leq C(\beta_0, M_0)(\log n)^{-\frac{\beta_0}{4\beta_0+1}},$$

where $C(\beta_0, M_0) = C_\psi M_0^{\frac{1}{4\beta_0+1}}$. By symmetry, the same result holds for $\underline{e}_f$. So on E_f,

$$RE(CB, f) \leq C(\beta_0, M_0)(\log n)^{-\frac{\beta_0}{4\beta_0+1}}, \quad \text{for} \quad C(\beta_0, M_0) = C_\psi M_0^{\frac{1}{4\beta_0+1}}. \qquad (72)$$

To complete the verification of (10), we evaluate the probability of E_f. By (64), we have

$$P_f(E_f^c) \leq P_f(\|\hat{f}_{\hat{j}^{\text{cb}}} - f_{\hat{j}^{\text{cb}}}\|_\infty > \hat{w}_n^s) + P_f(\text{some } H_{0,jl} \text{ vs. } H_{1,jl} \text{ was not tested correctly}).$$

By Proposition 1 and the remark after it, we apply a union bound to obtain

$$\sup_{\Lambda(\beta, M)} P_f(\|\hat{f}_{\hat{j}^{\text{cb}}} - f_{\hat{j}^{\text{cb}}}\|_\infty > \hat{w}_n^s) \leq \sum_{j \in \mathcal{J}} \sup_{f \in \Lambda(\beta, M)} P_f(\|\hat{f}_j - f_j\|_\infty > \hat{w}_n^s) \leq |\mathcal{J}| \frac{\alpha(1 + o(1))}{|\mathcal{J}|} = \alpha + o(1).$$

In addition, the total number of $H_{0,jl}$ vs. $H_{1,jl}$ tested are of order $O((\log n)^2)$. Thus, Proposition 2, together with the union bound, implies that

$$\sup_{\Lambda(\beta, M)} P_f(\text{some } H_{0,jl} \text{ vs. } H_{1,jl} \text{ was not tested correctly}) \leq C(\log n)^2(n^{-\frac{1}{2}} + 2^{-\frac{j^{\min}}{2}}) = o(1).$$

The last three displays together imply that

$$\lim_{n \to \infty} \sup_{\Lambda(\beta, M)} P_f(E_f) \geq 1 - \alpha. \qquad (73)$$

34

Together with (72), this completes the verification of (10).

3°. Finally, we turn to the verification of (9). The proof strategy is similar to the previous case. Fix any class $\Lambda(\beta, M) \in \mathcal{A}(\beta_0, M_0)$. Pick any $f \in \Lambda(\beta, M)$. For $\hat{j}^{\mathrm{cb}}$, let (26) be satisfied with $\hat{\beta}$ and $\hat{M}$. For any $l \geq \hat{j}^{\mathrm{cb}}$, let $\tilde{c}_{\hat{j}^{\mathrm{cb}}l}$ be defined in Proposition 3 with β and j replaced by $\hat{\beta}$ and $\hat{j}^{\mathrm{cb}}$ if $j \leq l \leq j^{\mathrm{t}}$. When $l > j^{\mathrm{t}}$, let $\tilde{c}_{\hat{j}^{\mathrm{cb}}l} = c_\psi M_0 2^{-(\beta_0 + \frac{1}{2})l}$. Define the event

$$E'_f = \{\|f_{\hat{j}^{\mathrm{cb}}} - \hat{f}_{\hat{j}^{\mathrm{cb}}}\|_\infty \leq \hat{w}^s_n, \text{ and all } H_{0,jl} \text{ vs. } H'_{1,jl} \text{ are tested correctly}\}. \tag{74}$$

Let the set of uncovered points be $N(CB, f)$. Note that if

$$\hat{w}^b_n \geq \tau_\psi \sum_{l > \hat{j}^{\mathrm{cb}}}^\infty 2^{l/2} \tilde{c}_{\hat{j}^{\mathrm{cb}}l}, \tag{75}$$

then on E'_f,

$$\mu(N(CB, f)) \leq \sum_{l > \hat{j}^{\mathrm{cb}}}^{j^{\mathrm{t}}} \kappa_{\hat{j}^{\mathrm{cb}}l}. \tag{76}$$

In what follows, we focus on verifying (75) and further bounding the right side of (76). To this end, we decompose $\{l : l > \hat{j}^{\mathrm{cb}}\} = \mathcal{J}_1 \cup \mathcal{J}_2 \cup \mathcal{J}_3$, where the J_i's are given by (66).

For $\mathcal{J}_1$, we further divide it into two disjoint subsets $\mathcal{J}_1 = \mathcal{J}_{10} \cup \mathcal{J}_{11}$, where

$$\mathcal{J}_{10} = \{l : 2^{-(\hat{\beta} + \frac{1}{2})(l - \hat{j}^{\mathrm{cb}})} \in [(\log n)^{-\frac{1}{2}}, 1)\},$$
$$\mathcal{J}_{11} = \{l : 2^{-(\hat{\beta} + \frac{1}{2})(l - \hat{j}^{\mathrm{cb}})} \in [(\log n)^{-1}, (\log n)^{-\frac{1}{2}})\}.$$

On $\mathcal{J}_{10}$, we have $\gamma_{\hat{j}^{\mathrm{cb}}l} = (\log n)^{-1/4}$, and so

$$\sum_{\mathcal{J}_{10}} 2^{\frac{l}{2}} \tilde{c}_{\hat{j}^{\mathrm{cb}}l} = (1 + (\log n)^{-\frac{1}{4}}) \sum_{\mathcal{J}_{10}} 2^{\frac{l}{2}} c_{\hat{j}^{\mathrm{cb}}l}.$$

On $\mathcal{J}_{11}$, Proposition 3 gives $\gamma_{\hat{j}^{\mathrm{cb}}l} = 2^{\frac{1}{2}\beta_0(l - \hat{j}^{\mathrm{cb}})}$, which leads to

$$\sum_{\mathcal{J}_{11}} 2^{\frac{l}{2}} \gamma_{\hat{j}^{\mathrm{cb}}l} c_{\hat{j}^{\mathrm{cb}}l} \leq C\hat{\omega}_n \sum_{\mathcal{J}_{11}} 2^{-(\hat{\beta} - \frac{1}{2}\beta_0)(l - \hat{j}^{\mathrm{cb}})} \leq C\hat{\omega}_n \sum_{\mathcal{J}_{11}} 2^{-\frac{1}{2}\hat{\beta}(l - \hat{j}^{\mathrm{cb}})} \leq C\hat{\omega}_n (\log n)^{-\frac{\hat{\beta}}{4\hat{\beta}+2}}.$$

Here, the second last inequality holds because $\hat{\beta} \geq \beta_0$. Putting together both parts and noting that the rightmost side of the last display achieves its maximum when $\hat{\beta} = \beta_0$, we have

$$\sum_{\mathcal{J}_1} 2^{\frac{l}{2}} \tilde{c}_{\hat{j}^{\mathrm{cb}}l} \leq \sum_{\mathcal{J}_1} 2^{\frac{l}{2}} c_{\hat{j}^{\mathrm{cb}}l} + C\hat{\omega}_n (\log n)^{-\frac{\beta_0}{4\beta_0+2}}. \tag{77}$$

Turn to $\mathcal{J}_2$. Similar to the case of $\mathcal{J}_{11}$, we have $\sum_{\mathcal{J}_2} 2^{\frac{l}{2}} \gamma_{\hat{j}^{cb}l} c_{\hat{j}^{cb}l} \leq C\hat{\omega}_n \sum_{\mathcal{J}_2} 2^{-\frac{1}{2}\hat{\beta}(l-\hat{j}^{cb})}$, and hence

$$\sum_{J_2} 2^{\frac{l}{2}} \tilde{c}_{\hat{j}^{cb}l} \leq \sum_{J_2} 2^{\frac{l}{2}} c_{\hat{j}^{cb}l} + C\hat{\omega}_n (\log n)^{-\frac{\hat{\beta}}{2\hat{\beta}+1}} \leq \sum_{J_2} 2^{\frac{l}{2}} c_{\hat{j}^{cb}l} + C\hat{\omega}_n (\log n)^{-\frac{\beta_0}{2\beta_0+1}}. \tag{78}$$

For $\mathcal{J}_3$, we further decompose it into $\mathcal{J}_3 = \mathcal{J}_{30} \cup \mathcal{J}_{31}$, where $\mathcal{J}_{30} = \{l \in \mathcal{J}_3 : l \leq j^{t}\}$, and $\mathcal{J}_{31} = \{l \in \mathcal{J}_3 : l > j^{t}\}$. For each $l \in \mathcal{J}_{30}$, we have

$$2^{\frac{l}{2}} \tilde{c}_{\hat{j}^{cb}l} \leq C\sigma_n (\log n)^{\frac{\beta_0}{2(4\beta_0+1)}} 2^{\frac{1}{8}(j^{t}+l)} l^{\frac{1}{4}} \leq C\sigma_n (\log n)^{\frac{6\beta_0+1}{4(4\beta_0+1)}} 2^{\frac{1}{8}(j^{t}+l)}.$$

Note that the right side is geometrically increasing in l, and so

$$\sum_{\mathcal{J}_{30}} 2^{\frac{l}{2}} \tilde{c}_{\hat{j}^{cb}l} \leq C\sigma_n (\log n)^{\frac{6\beta_0+1}{4(4\beta_0+1)}} 2^{\frac{1}{4}j^{t}} \leq CM_0^{\frac{1}{4\beta_0+1}} (\log n)^{-\frac{\beta_0}{2(4\beta_0+1)}} \hat{\omega}_n,$$

where the last inequality relies on the fact that $(\sigma^2 \log n/n)^{2\beta_0/(4\beta_0+1)} \leq C\hat{\omega}_n$.

On $\mathcal{J}_{31}$, $\tilde{c}_{\hat{j}^{cb}l} = c_\psi M_0 2^{-(\beta_0+\frac{1}{2})l}$, and so

$$\sum_{\mathcal{J}_{31}} 2^{\frac{l}{2}} \tilde{c}_{\hat{j}^{cb}l} \leq \frac{M_0}{1-2^{-\beta_0}} 2^{-\beta_0 j^{t}} \leq CM_0^{\frac{1}{4\beta_0+1}} (\log n)^{-\frac{\beta_0}{4\beta_0+1}} \hat{\omega}_n,$$

where the last inequality also relies on $(\sigma^2 \log n/n)^{2\beta_0/(4\beta_0+1)} \leq C\hat{\omega}_n$. The last two displays jointly imply

$$\sum_{\mathcal{J}_3} 2^{\frac{l}{2}} \tilde{c}_{\hat{j}^{cb}l} \leq CM_0^{\frac{1}{4\beta_0+1}} (\log n)^{-\frac{\beta_0}{2(4\beta_0+1)}} \hat{\omega}_n. \tag{79}$$

Putting (77), (78) and (79) together, we obtain that for $n \geq n_0(\beta_0, M_0, \sigma)$,

$$\tau_\psi \sum_{l>\hat{j}^{cb}} 2^{\frac{l}{2}} \tilde{c}_{\hat{j}^{cb}l} \leq \tau_\psi [1 + CM_0^{\frac{1}{4\beta_0+1}} (\log n)^{-\frac{\beta_0}{2(4\beta_0+1)}}] \sum_{l>\hat{j}^{cb}}^{j^{t}} 2^{\frac{l}{2}} c_{\hat{j}^{cb}l} \leq \hat{w}_n^b,$$

i.e., (75) is satisfied.

Given (75), we now bound the right side of (76) on the event E_f'. To this end, note that $\{l : \hat{j}^{cb} < l \leq j^{t}\} = \mathcal{J}_{10} \cup \mathcal{J}_{11} \cup \mathcal{J}_2 \cup \mathcal{J}_{30}$. So we compute the sum over these four sets separately. For $\mathcal{J}_{10}$, we have $\kappa_{\hat{j}^{cb}l} \leq C\sigma_n c_{\hat{j}^{cb}l}^{-1} (\log n)^{-\frac{1}{4}} \leq C(\log n)^{-3/4} 2^{(\hat{\beta}+\frac{1}{2})(l-\hat{j}^{cb})}$, and so

$$\sum_{\mathcal{J}_{10}} \kappa_{\hat{j}^{cb}l} \leq C(\log n)^{-\frac{3}{4}} \sum_{\mathcal{J}_{10}} 2^{(\hat{\beta}+\frac{1}{2})(l-\hat{j}^{cb})} \leq C(\log n)^{-\frac{1}{4}}. \tag{80}$$

On $\mathcal{J}_{11}$, $\kappa_{\hat{j}^{cb}l} \leq C\sigma_n (c_{\hat{j}^{cb}l} \gamma_{\hat{j}^{cb}l})^{-1} (\log n)^{-\frac{1}{2}} \leq C2^{-\frac{1}{2}\beta_0(l-\hat{j}^{cb})}$, which leads to

$$\sum_{\mathcal{J}_{11}} \kappa_{\hat{j}^{cb}l} \leq C \sum_{\mathcal{J}_{11}} 2^{-\frac{1}{2}\beta_0(l-\hat{j}^{cb})} \leq C(\log n)^{-\frac{\beta_0}{4\hat{\beta}+2}} \leq C(\log n)^{-\frac{\beta_0}{8\beta_0+2}}. \tag{81}$$

36

The last inequality holds as $\hat{\beta} \leq 2\beta_0$. On $\mathcal{J}_2$, $\kappa_{\hat{j}^{\mathrm{cb}}l} = C\gamma^{-1}_{\hat{j}^{\mathrm{cb}}l}$, and so

$$\sum_{\mathcal{J}_2} \kappa_{\hat{j}^{\mathrm{cb}}l} \leq C \sum_{\mathcal{J}_2} 2^{-\frac{1}{2}\beta_0(l-\hat{j}^{\mathrm{cb}})} \leq C(\log n)^{-\frac{\beta_0}{2\hat{\beta}+1}} \leq C(\log n)^{-\frac{\beta_0}{4\beta_0+1}}. \tag{82}$$

Last but not least, on $\mathcal{J}_{30}$, we have $\kappa_{\hat{j}^{\mathrm{cb}}l} \leq C\gamma^{-1}_{\hat{j}^{\mathrm{cb}}l} \leq C(\log n)^{-\frac{\beta_0}{2(4\beta_0+1)}} 2^{-\frac{1}{8}(j^{\mathrm{t}}-l)}$, and so

$$\sum_{l \in \mathcal{J}_{30}} \kappa_{\hat{j}^{\mathrm{cb}},l} \leq C(\log n)^{-\frac{\beta_0}{2(4\beta_0+1)}} \sum_{l \in \mathcal{J}_{30}} 2^{-\frac{1}{8}(j^{\mathrm{t}}-l)} \leq C(\log n)^{-\frac{\beta_0}{2(4\beta_0+1)}}. \tag{83}$$

Assembling (80), (81), (82) and (83), we further bound the right side of (76) to obtain

$$\mu(N(CB, f)) \leq C(\log n)^{-\frac{\beta_0}{2(4\beta_0+1)}}. \tag{84}$$

In addition, a similar argument to that leading to (73) leads to

$$\varliminf_{n \to \infty} \sup_{\Lambda(\beta, M)} P_f(E'_f) \geq 1 - \alpha. \tag{85}$$

Together with (84), this completes the verification of (9) and hence completes the proof of Theorem 1. ∎

6.2 Proof of Theorem 3

Before we turn to the proof of the lower bounds given in Theorem 3 we introduce a lemma which gives a bound on the chi-square distance between a Normal random vector and particular mixtures of such vectors. The proof of this lemma is given in the supplement. Let n be a positive integer and let $\{J_1, J_2, ..., J_m\}$ be a partition of the index set $\{1, 2, ..., n\}$ with $|J_i| = k_i$ and $\sum_{i=1}^{m} k_i = n$. Let $B_1, ..., B_m$ be independent and identically distributed Rademacher variables with $P(B_1 = -1) = P(B_1 = 1) = \frac{1}{2}$. For a fixed vector $\gamma = (\gamma_1, ..., \gamma_n) \in \mathbb{R}^n$, define the random vector $\theta \in \mathbb{R}^n$ by $\theta_{J_i} = B_i\gamma_{J_i}$ for $i = 1, 2, ...k$. Let $y|\theta \sim N_n(\theta, \sigma^2 I)$. Denote the marginal distribution of y and its density function by P_1 and h_1, respectively.

For a vector $\xi \in \mathbb{R}^d$, denote the density of a d-variate normal distribution $N_d(\xi, \sigma^2 I)$ by ϕ_ξ and set $\psi_\xi = \frac{1}{2}\phi_{-\xi} + \frac{1}{2}\phi_\xi$. Then it is easy to see that the marginal density h_1 of y is given by $h_1(y) = \prod_{i=1}^{m} \psi_{\gamma_{J_i}}(y_{J_i})$. Denote by P_0 and h_0 respectively the joint distribution and joint density of the normal distribution $N_n(0, \sigma^2 I)$.

Lemma 1. *The chi-squared distance between P_0 and P_1, $\chi^2(P_0, P_1)$, satisfies*

$$\chi^2(P_0, P_1) \equiv \mathsf{E}_{h_0}\left(\frac{h_1(y)}{h_0(y)} - 1\right)^2 \leq \exp\left(\frac{1}{2\sigma^4}\sum_{i=1}^{m}\|\gamma_{J_i}\|_2^4\right) - 1.$$

If $\sum_{i=1}^{m}\|\gamma_{J_i}\|_2^4 \leq 2\sigma^4 \log(1 + \epsilon_0^2)$ and A is any event such that $P_0(A) \geq \alpha$, then

$$P_1(A) \geq \alpha - \frac{1}{2}\epsilon_0. \tag{86}$$

Now without loss of generality we shall assume the noise level $\sigma = 1$. Let g be an infinitely differentiable function supported on $[0, 1]$ with $g(t) > 0$ for $t \in (0, 1)$ and $\int_0^1 g^2(t)dt = 1$. For instance, one can set

$$g(t) = \begin{cases} c_g\left(\exp\left(-\frac{1}{t}e^{-\frac{1}{1-t}}\right) + \exp\left(-\frac{1}{1-t}e^{-\frac{1}{t}}\right) - 1\right), & t \in [0, 1], \\ 0, & \text{otherwise.} \end{cases} \tag{87}$$

Here, the normalizing constant $c_g \doteq 0.346$. Suppose $h \in \Lambda(\beta, M')$ with $M' < M$, Let m be a positive integer and let $B_1, ..., B_m$ be iid Rademacher variable with $P(B_1 = -1) = P(B_1 = 1) = \frac{1}{2}$. Define the random function f by

$$f(t) = h(t) + \sum_{i=1}^{m} B_i c_0 m^{-\beta} g(m(t - x_i)) \tag{88}$$

where $x_i = \frac{i-1}{m}$ and $c_0 > 0$ is a constant. It is easy to verify that, when the constant c_0 is chosen sufficiently small, all realizations of f are in $\Lambda(\beta, M)$. Set $m = \lceil n^{\frac{2}{4\beta+1}} \rceil$. Without loss of generality we shall assume that n is divisible by m and let $k_n = n/m$. Note that the Riemann sum

$$A_{k_n} \equiv \frac{1}{k_n}\sum_{j=1}^{k_n} g^2(\frac{j}{k_n}) \to \int_0^1 g^2(t)dt = 1$$

and so for all sufficiently large n, $A_{k_n} \leq \sqrt{2}$.

Now consider the nonparametric regression model (1) with the mean function f given in (88). Denote the joint marginal distribution of the $y_1, ..., y_n$ by P_1. If the B_i Rademacher variables are instead set equal to zero denote the joint distribution of the $y_1, ..., y_n$ by P_0. It follows from Lemma 1 that the chi-squared distance between P_0 and P_1 satisfies

$$\begin{aligned} \chi^2(P_0, P_1) &\leq \exp\left\{\frac{1}{2}mc_0^2 m^{-4\beta}\left(\sum_{j=1}^{k_n} g^2(\frac{j}{k_n})\right)^2\right\} - 1 = \exp\left\{\frac{1}{2}mc_0^2 m^{-4\beta}k_n^2 A_{k_n}^2\right\} - 1 \\ &\leq e^{c_0^2} - 1. \end{aligned}$$

Set $b_n = bn^{\frac{-2\beta}{4\beta+1}}$ and suppose that $w(CB, h) \leq bn^{\frac{-2\beta}{4\beta+1}}$ it follows that

$$P_0 \left(\int_0^1 (U(t) - L(t))dt \leq \gamma_1^{-1}b_n \right) \geq 1 - \gamma_1.$$

For a given constant $0 < \gamma_2 < 1$, define the set $S_1 = \{t \in [0,1] : U(t) - L(t) \leq (\gamma_1\gamma_2)^{-1}b_n\}$. Then it follows that

$$P_0\left(\mu(S_1) \geq 1 - \gamma_2\right) \geq P_0 \left(\int_0^1 (U(t) - L(t))dt \leq \gamma_1^{-1}b_n \right) \geq 1 - \gamma_1.$$

where $\mu(\cdot)$ is the Lebesgue measure. Define the set $S_2 = \{t \in [0,1] : h(t) \in [L(t), U(t)]\}$.

Suppose that

$$P_0 \left(\mu(S_2) \geq \frac{1}{2} + \epsilon \right) \geq 1 - \alpha.$$

Now set $A = S_1 \cap S_2$. If $\mu(S_1) \geq 1 - \gamma_2$ and $\mu(S_2) \geq \frac{1}{2} + \epsilon$, then $\mu(A) \geq \frac{1}{2} + \epsilon - \gamma_2$. Hence

$$P_0(\mu(A) \geq \frac{1}{2} + \epsilon - \gamma_2) \geq 1 - \alpha - \gamma_1.$$

It now follows from Lemma 1 that

$$P_1(\mu(A) \geq \frac{1}{2} + \epsilon - \gamma_2) \geq 1 - \alpha - \gamma_1 - \frac{1}{2}\epsilon_0.$$

Note that for any function $f(\cdot)$ of the form (88) with $m = \lfloor n^{\frac{2}{4\beta+1}} \rfloor$ and $B_i \in \{-1, 1\}$, for any $c > 0$ there is a $b > 0$ and $d > 0$ both depending on c_0 and $\gamma_1\gamma_2$ such that if $d_n = dn^{-\frac{2\beta}{4\beta+1}}$,

$$\mu(S_3) \geq 1 - c,$$

where $S_3 = \{t : |f(t) - h(t)| \geq d_n\}$.

Note also that for $t \in A$, $h(t) \in [L(t), U(t)]$ and $U(t) - L(t) \leq (\gamma_1\gamma_2)^{-1}b_n$, so $|L(t) - h(t)| \leq (\gamma_1\gamma_2)^{-1}b_n$ and $|U(t) - h(t)| \leq (\gamma_1\gamma_2)^{-1}b_n$. So under P_1, for any f of the form (88), the set of noncovered points $N(CB, f)$ satisfies under the event $\mu(A) \geq \frac{1}{2} + \epsilon - \gamma_2$

$$\mu(N(CB, f)) \geq \mu(A \cap S_3) \geq \frac{1}{2} + \epsilon - \gamma_2 - c.$$

Hence,

$$P_1 \left(\mu(N(CB, f)) \geq \frac{1}{2} + \epsilon - \gamma_2 - c \right) \geq 1 - \alpha - \gamma_1 - \frac{1}{2}\epsilon_0.$$

By taking $\gamma_2 + c \leq \epsilon$ and selecting γ_1 and ϵ_0 such that $1 - \alpha - \gamma_1 - \frac{1}{2}\epsilon_0 > \alpha$ yields

$$P_1 \left(\mu(N(CB, f)) \geq \frac{1}{2} \right) > \alpha.$$

39

Hence there is an f for which

$$P_f\left(\mu(N(CB, f)) \le \frac{1}{2}\right) < 1 - \alpha.$$

It thus follows that if a confidence band satisfies (49) then (51) must hold.

We shall now show that if (50) holds then (51) must also hold. Once again set $b_n = bn^{\frac{-2\beta}{4\beta+1}}$ and suppose that $w(CB, h) \le bn^{\frac{-2\beta}{4\beta+1}}$. Defining S_1 as before note that

$$P_0\left(\int_0^1 (U(t) - L(t))dt \le \gamma_1^{-1}b_n, \text{ and } \mu(S_1) \ge 1 - \gamma_2\right) \ge 1 - \gamma_1.$$

Then for some $0 < \gamma_3 < 1$, define the set $S_2' = \{t \in [0, 1] : U(t) - h(t) \ge -\gamma_3^{-1}b_n \text{ and } L(t) - h(t) \le \gamma_3^{-1}b_n\}$. On $(S_2')^c$ at any point t the true function is at least $\frac{b_n}{\gamma_3}$ away from the band. Hence the absolute excess is at least $\frac{b_n}{\gamma_3}(1 - \mu(S_2'))$. Set $A' = S_1 \cap S_2'$. Then the previous display and the discussion afterwards implies

$$P_0\left(\int_0^1 (U(t) - L(t))dt \le \gamma_1^{-1}b_n, \text{ and } \mu(A') \ge 1 - \gamma_2 - \frac{\gamma_3}{\gamma_1}r\right) \ge 1 - \alpha - \gamma_1.$$

Moreover, Lemma 1 further implies

$$P_1\left(\int_0^1 (U(t) - L(t))dt \le \gamma_1^{-1}b_n, \text{ and } \mu(A') \ge 1 - \gamma_2 - \frac{\gamma_3}{\gamma_1}r\right) \ge 1 - \alpha - \gamma_1 - \epsilon_0.$$

For any function $f(\cdot)$ of the form (88) for any $c > 0$ there is a $b > 0$ and c_3 both depending on c_0, $\gamma_1\gamma_2$ and γ_3, and $d = (1 + c_3)((\gamma_1\gamma_2)^{-1} + \gamma_3^{-1})b \le c_0$ such that if $d_n = dn^{-\frac{2\beta}{4\beta+1}}$, we have for the set S_3

$$\mu(S_3) \ge 1 - c.$$

On the set A', we have $|U(t) - h(t)| \le ((\gamma_1\gamma_2)^{-1} + \gamma_3^{-1})b_n$ and $|L(t) - h(t)| \le ((\gamma_1\gamma_2)^{-1} + \gamma_3^{-1})b_n$. So, we have

$$\int_0^1 e_f(t)dt \ge \int_{A' \cap S_3} \left(d_n - ((\gamma_1\gamma_2)^{-1} + \gamma_3^{-1})b_n\right)dt \ge c_3((\gamma_1\gamma_2)^{-1} + \gamma_3^{-1})b_n \, \mu(A' \cap S_3).$$

So, on the event $\{\int_0^1 (U(t) - L(t))dt \le \gamma_1^{-1}b_n, \text{ and } \mu(A') \ge 1 - \gamma_2 - \frac{\gamma_3}{\gamma_1}r\}$, the last display leads to

$$RE(CB, f) \ge c_3(\gamma_2^{-1} + \gamma_1\gamma_3^{-1})(1 - \gamma_2 - \frac{\gamma_3}{\gamma_1}r - c).$$

Therefore,

$$P_1\left(RE(CB, h) \ge c_3(\gamma_2^{-1} + \gamma_1\gamma_3^{-1})(1 - \gamma_2 - \frac{\gamma_3}{\gamma_1}r - c)\right) \ge 1 - \alpha - \gamma_1 - \epsilon_0 > \alpha \quad \blacksquare$$

References

[1] Bickel, P. J. and Rosenblatt M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1**, 1071-1095.

[2] Brown, L. D., Cai, T., Zhang, R., Zhao, L. and Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields* **146**, 401-433.

[3] Brown, L. D., Cai, T. and Zhou, H. (2008). Robust nonparametric estimation via wavelet median regression. *Ann. Statist.* **36**, 2055-2084.

[4] Brown, L. D., Cai, T. and Zhou, H. (2010). Nonparametric regression in exponential families. *Ann. Statist.* **38**, 2005-2046.

[5] Brown, L. D., Fu, X. and Zhao, L. (2010). Confidence intervals for nonparametric regression. *J. Nonparametric Statist.,* **23**, 149-163.

[6] Brown, L.D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L.H. (2005). Statistical analysis of a telephone call center: a queueing science perspective. *J. Amer. Statist. Assoc.,* **100**, 36-50.

[7] Bull, A. D. (2011a). A Smirnov-Bickel-Rosenblatt theorem for compactly-supported wavelets. ArXiv preprint, arXiv:1110.4961.

[8] Bull, A. D. (2011b). Honest adaptive confidence bands and self-similar functions. ArXiv preprint, arXiv:1110.4985.

[9] Cai, T. and Low, M. G. (2006a). Adaptive confidence balls. *Ann. Statist.* **34**, 202-228.

[10] Cai, T. and Low, M. G. (2006b). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34**, 2298-2325.

[11] Cai, T. and Zhou, H. (2009). Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.* **37**, 3204-3235.

[12] Cai, T. and Zhou, H. (2010). Nonparametric regression in natural exponential families. In *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, IMS Collections Vol. 6, 199-215.

[13] Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993). Multiresolution analysis, wavelets, and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris* (A). **316**, 417–421.

[14] Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

[15] Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist.* **26**, 288-314.

[16] Genovese, C. and Wasserman, L. (2008). Adaptive confidence bands. *Ann. Statist.* **36**, 875-905.

[17] Giné, E., Güntürk, C. S. and Madych, W. R. (2011). On the Periodized Square of L^2 Cardinal Splines. *Experimental Mathematics* **20**, 177-188.

[18] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38**, 1122-1170.

[19] Hall, P. and Titterington, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.* **27**, 228-254.

[20] Hengartner, N. W. and Stark, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23**, 525-550.

[21] Hoffman M. and Lepski, O. (2002). Random rates in anisotropic regression. *Ann. Statist.* **30**, 325-396.

[22] Hoffman, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39**, 2383-2409.

[23] Hüsler, J. (1999). Extremes of Gaussian processes, on results of Piterbarg and Seleznjev. *Statist. Prob. Lett.* **44**, 251-258.

[24] Johnstone, I. (2012). *Gaussian estimation: Sequence and wavelet models.* Book draft.

[25] Juditsky, A. and Lambert-Lacroix, S. (2003). Nonparametric confidence set estimation. *Math. Methods Statist.* **12**, 410-428.

[26] Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.* **83**, 1134-1143.

[27] Piterbarg V. and Seleznjev O. (1994). Linear interpolation of random processes and extremes of a sequence of Gaussian nonstationary processes. Technical report 446, Department of Statistics, University of North Carolina, Chapel Hill, NC.

[28] Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34**, 229-253.

[29] Smirnov, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables. *Doklady Akad. Nauk SSSR (N.S.)* **74**, 189-191.

[30] Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc. B* **45**, 133-150.

[31] Wasserman, L. (2006). *All of Nonparametric Statistics.* Springer.