Adaptive Confidence Bands

BY CHRISTOPHER R. GENOVESE[1] AND LARRY WASSERMAN[2]

Department of Statistics
Carnegie Mellon University

December 12, 2006

We show that there do not exist adaptive confidence bands for curve estimation except under very restrictive assumptions. We propose instead to construct adaptive bands that cover a surrogate function $f^\star$ which is close to, but simpler than, f. The surrogate captures the significant features in f. We establish lower bounds on the width for any confidence band for $f^\star$ and construct a procedure that comes within a small constant factor of attaining the lower bound for finite-samples.

KEY WORDS AND PHRASES: Confidence sets, confidence bands, nonparametric regression.

1 Introduction

1.1 Motivation

Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be observations from the nonparametric regression model

$$Y_i = f(x_i) + \sigma\,\epsilon_i \tag{1}$$

where $\epsilon_i \sim N(0,1)$, $x_i \in (0,1)$, and f is assumed to lie in some infinite-dimensional class of functions $\mathcal{H}$. We are interested in constructing confidence bands (L, U) for f. Ideally these bands should satisfy

$$\mathbb{P}_f\{L \le f \le U\} = 1 - \alpha \qquad \text{for all } f \in \mathcal{H} \tag{2}$$

where $L \le f \le U$ means that $L(x) \le f(x) \le U(x)$ for all $x \in \mathcal{X}$, where $\mathcal{X}$ is some subset of $(0,1)$ such as $\mathcal{X} = \{x\}, \mathcal{X} = \{x_1, \ldots, x_n\}$ or $\mathcal{X} = (0,1)$. Throughout this paper, we take $\mathcal{X} = \{x_1, \ldots, x_n\}$ but this particular choice is not crucial in what follows.

Attaining (2) is difficult and hence it is common to settle for pointwise asymptotic coverage:

$$\liminf_{n\to\infty} \mathbb{P}_f\{L \le f \le U\} \ge 1 - \alpha \qquad \text{for all } f \in \mathcal{H}. \tag{3}$$

"Pointwise" refers to the fact that the asymptotic limit is taken for each fixed f rather than uniformly over $f \in \mathcal{H}$. Papers on pointwise asymptotic methods include Claeskens and Van Keilegom (2003), Eubank and Speckman (1993), Härdle and Marron (1991), Hall and Titterington (1988), Härdle and Bowman (1988), Neumann and Polzehl (1998), and Xia (1998).

[1]Research supported by NSF Grant SES 9866147.

[2]Research supported by NIH Grant R01-CA54852-07, NIH grant number MH57881, NSF Grant DMS-98-03433 and NSF Grant DMS-0104016.

Achieving even pointwise asymptotic coverage is nontrivial due to the presence of bias. If $\widehat{f}(x)$ is an estimator with mean $\overline{f}(x)$ and standard deviation $s(x)$ then

$$\frac{\widehat{f}(x) - f(x)}{s(x)} = \frac{\widehat{f}(x) - \overline{f}(x)}{s(x)} + \frac{\text{bias}(x)}{\sqrt{\text{variance(x)}}}.$$

The first term typically satisifes a central limit theorem but the second term does not vanish even asymptotically if the bias and variance are balanced. For discussions on this point, see the papers referenced above as well as Ruppert and Wand (2003) and Sun and Loader (1994).

Pointwise asymptotic bands are not uniform, that is, they do not control

$$\inf_{f \in \mathcal{H}} \mathbb{P}_f\{L \leq f \leq U\}. \tag{4}$$

The sample size $n(f)$ required for the true coverage to approximate the nominal coverage, depends on the unknown function f.

The aim of this paper is to attain uniform coverage over $\mathcal{H}$. We say that $B = (L, U)$ has *uniform coverage* if

$$\inf_{f \in \mathcal{H}} \mathbb{P}_f\{L \leq f \leq U\} \geq 1 - \alpha. \tag{5}$$

Starting in Section 3, we will insist on coverage over $\mathcal{H} = \{\text{all functions}\}$.

The bound in (5) can be achieved trivially using Bonferroni bands. Set $\ell_i = Y_i - c_n\sigma$ and $u_i = Y_i + c_n\sigma$, where $c_n = \Phi^{-1}(1 - \alpha/2n)$ and Φ is the standard Normal CDF. Yet this band is unsatisfactory for several reasons:

1. The width of the band grows with sample size.

2. The band is centered on a poor estimator of the unknown function.

3. The width of the band is independent of the data and hence cannot adapt to the smoothness of the unknown function.

Problems (1) and (2) are easily remedied by using standard smoothing methods. But the results of Low (1997) suggest that (3) is an inevitable consequence of uniform coverage.

The smoother the functions in $\mathcal{H}$, the smaller the width necessary to achieve uniform coverage. Suppose that $\mathcal{F} \subset \mathcal{H}$ contains the "smooth" functions in $\mathcal{H}$ and that $\mathcal{H} - \mathcal{F}$ is nonempty. Uniform coverage over $\mathcal{H}$ requires that the width of fixed-width bands be driven by the "rough" functions in $\mathcal{H} - \mathcal{F}$; the width will thus be large even if $f \in \mathcal{F}$. Ideally, our procedure would adjust automatically to produce narrower bands when the function is smooth ($f \in \mathcal{F}$) and wider bands when the function is rough ($f \notin \mathcal{F}$), but to do that, the width must be determined from the data. Low showed that for density estimation at a single point, fixed-width confidence intervals perform as well as random length intervals; that is, the data do not help reduce the width of the bands for smoother functions. In Section 2, we extend Low's result to nonparametric regression and show that the phenomenon is quite general. Without restrictive assumptions, confidence bands cannot adapt.

These results mean that the width of uniform confidence bands is determined by the greatest roughness we are willing to assume. Because the typical assumptions about $\mathcal{H}$ in the nonparametric regression problem are loosely held and difficult to check, the result is that the confidence band widths are essentially arbitrary. This is not satisfactory in practice.

The contrast with L^2 confidence balls is noteworthy. L^2 confidence sets have been studied by Li (1999), Juditsky and Lambert-Lacroix (2002), Beran and Dümbgen (1998), Genovese and Wasserman (2004), Baraud (2004), Hoffman and Lepski (2003), Cai and Low (2004), and Robins and van der Vaart (2004). Let

$$B = \left\{ f \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^{n}(f_i - \widehat{f}_i)^2 \leq R_n^2 \right\} \tag{6}$$

for some $\widehat{f}$ and suppose that

$$\inf_{f\in\mathbb{R}^n} \mathbb{P}_f\{ f \in B \} \geq 1-\alpha. \tag{7}$$

Then

$$\inf_{f\in\mathbb{R}^n} \mathsf{E}_f(R_n) \geq \frac{C_1}{n^{1/4}}, \quad \text{and} \quad \sup_{f\in\mathbb{R}^n} \mathsf{E}_f(R_n) \geq C_2 \tag{8}$$

where C_1 and C_2 are positive constants. Moreover, there exist confidence sets that achieve the faster $n^{-1/4}$ rate at some points in $\mathbb{R}^n$. Because fixed-radius confidence sets necessarily have radius of size O(1), the supremum in (8) implies such confidence sets must have random radii. We can construct random-radius confidence balls that improve on fixed-radius confidence sets, for example, by obtaining a smaller radius for subsets of smoother functions f. L^2 confidence balls can therefore adapt to the unknown smoothness of f. Unfortunately, confidence balls can be difficult to work with in high dimensions (large n) and tend to constrain many features of interest rather poorly, for which reasons confidence bands are often desired.

It is also interesting to compare the adaptivity results for estimation and inference. Estimators exist (e.g., Donoho et al. 1995) that can adapt to unknown smoothness, achieving near optimal rates of convergence over a broad scale of spaces. But since confidence bands cannot adapt, the minimum width bands that achieve uniform coverage over the same scale of spaces have width $O(1)$, overwhelming the differences among reasonable estimators. We are left knowing that we are close to the true function but being unable to demonstrate it inferentially.

The message we take from the nonadaptivity results in Low (1987) and Section 2 of this paper is that the problem of constructing confidence bands for f over nonparametric classes is simply too difficult under the usual definition of coverage. Instead, we introduce a slightly weaker notion – surrogate coverage – under which it is possible to obtain adaptive bands while allowing sharp inferences about the main features of f.

1.2 Surrogates

Figure 1 shows two situations where a band fails to capture the true function. The top plot shows a conservative failure: the only place where f is not contained in the band is when the bands

are smoother than the truth. The bottom plot shows a liberal failure: the only place where f is not contained in the band is when the bands are less smooth than the truth. The usual notion of coverage treats these failures equally. Yet, in some sense, the second error is more serious than the first since the bands overstate the complexity.

We are thus led to a different approach that treats conservative errors and liberal errors differently. The basic idea is to find a function $f^\star$ that is simpler than f as in Figure 2. We then require that

$$\mathbb{P}_f\{L \leq f \leq U \text{ or } L \leq f^\star \leq U\} \geq 1-\alpha, \quad \text{for all functions } f. \tag{9}$$

More generally, we will define a finite set of surrogates $F^\star \equiv F^*(f) = \{f, f_1^*, \ldots, f_m^*\}$ and require that a surrogate confidence band (L, U) satisfy

$$\inf_f \mathbb{P}_f\{L \leq g \leq U \ \text{ for some } g \in F^\star\} \geq 1-\alpha. \tag{10}$$

We will also consider bands that are adaptive in the following sense: if f lies in some subspace $\mathcal{F}$, then with high probability $\|U-L\|_\infty \leq w(\mathcal{F})$, where $w(\mathcal{F})$ is the best width of a uniformly valid confidence band (under the usual definition of coverage) based on the a priori knowledge that $f \in \mathcal{F}$. Among possible surrogates, a surrogate will be optimal if it admits a valid, adaptive procedure and the set $\{f \in \mathcal{F} : \ F^*(f) = \{f\}\}$ is as large as possible.

1.3 Summary of Results

In Section 2, we show that Low's result on density estimation holds in regression as well. Fixed width bands do as well as random width bands, thus ruling out adaptivity. We show this when $\mathcal{H}$ is the set of all functions and when $\mathcal{H}$ is a ball in a Lipschitz, Sobolev, or Besov space.

Section 3 gives our main results. Theorem 3.2 establishes lower bounds on the width for any valid surrogte confidence band. Let $\mathcal{F}$ be a subspace of dimension d in R^n. The functions that prevent adaptation are those that are close to $\mathcal{F}$ in L^2 but far in L^∞. Loosely speaking, such functions are close to $\mathcal{F}$ except for isolated, spiky features. If $||f - \Pi f||_2 < \epsilon_2$ and $||f - \Pi f||_\infty > \epsilon_\infty$, for tuning constants $\epsilon_2, \epsilon_\infty$, define the surrogate $f^\star$ to be the projection of f onto $\mathcal{F}$, Πf. Otherwise, define $f^\star = f$. We show that if $\mathbb{P}_f\{\|U-L\|_\infty < w\} \geq 1-\gamma$ for all $f \in \mathcal{F}$, then

$$w \geq \max\left(w_{\mathcal{F}}(\alpha,\gamma,\sigma), v(\epsilon_2,\epsilon_\infty,n,d,\alpha,\gamma,\sigma)\right), \tag{11}$$

where $w_{\mathcal{F}}$ is the minimum width for a uniform confidence band knowing a priori that $f \in \mathcal{F}$ and $v(\epsilon_2,\epsilon_\infty,n,d,\alpha,\gamma)$ is described later.

Corollary 3.2 shows that for proper choice of ϵ_2 and ϵ_∞, the v term in the previous equation can be made smaller than $w_{\mathcal{F}}$. Figure 3 represents the functions involved; the gray shaded area are those functions that are replaced by surrogates in the coverage statement, denoted later by $\mathcal{S}(\epsilon_2,\epsilon_\infty)$. These are the functions that are both hard to distinguish from $\mathcal{F}$ (because they are close to it) and hard to cover (because they are "spiky"). The optimal choice of ϵ_2 and ϵ_∞ minimizes the volume of this set while making the right hand side in inequality (11) equal to $w_{\mathcal{F}}$. Put another

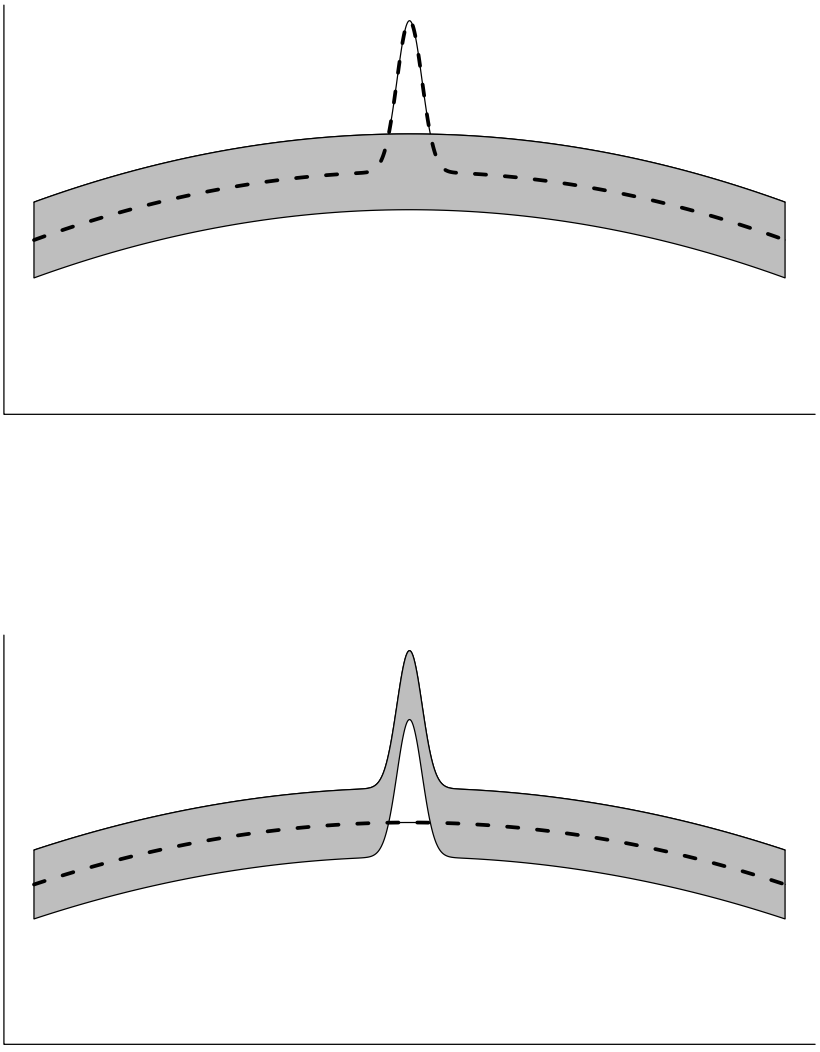

Figure 1: The top plot shows a conservative failure: the only place where f is not contained in the band is when the bands are smoother than the truth. The bottom plot shows a liberal failure: the only place where f is not contained in the band is when the bands are less smooth than the truth. The usual notion of coverage treats these failures equally.

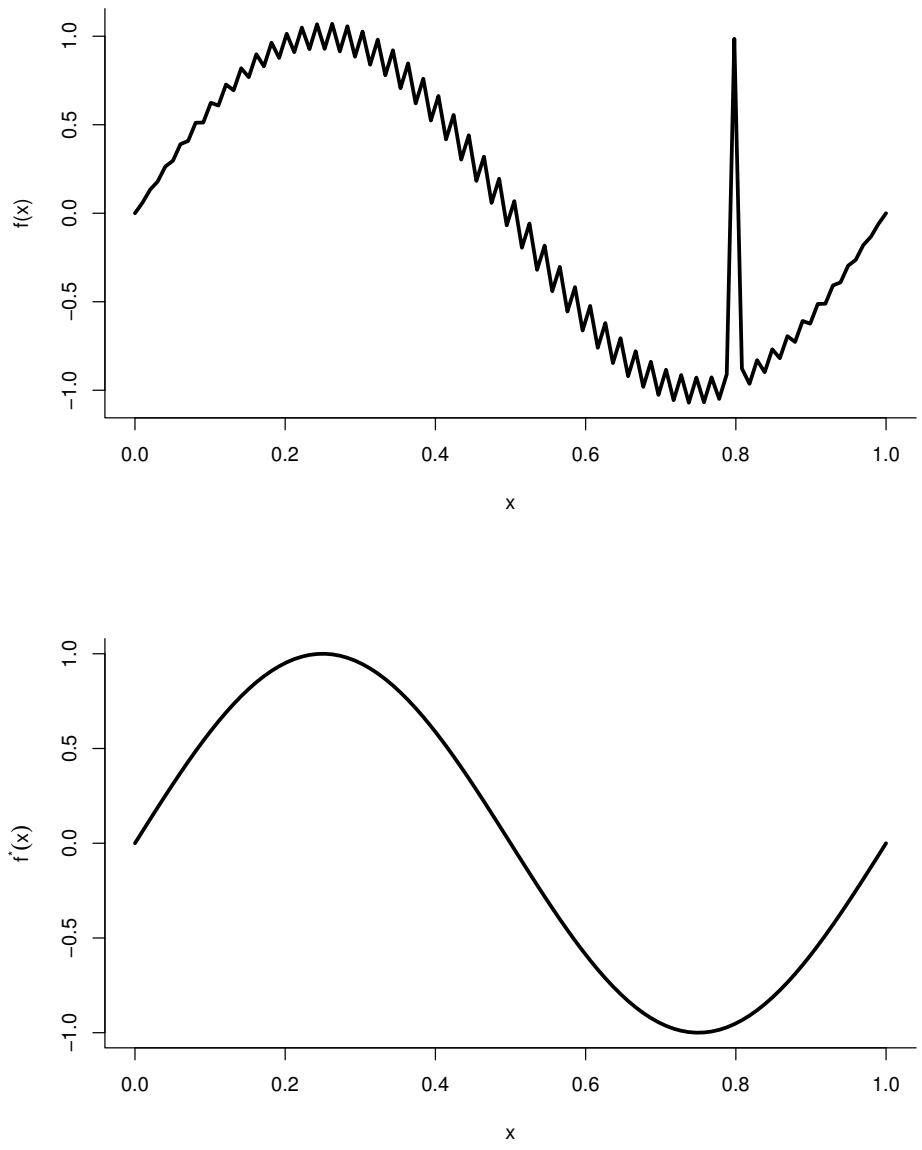

Figure 2: The top plot shows a complicated function f. The bottom shows a surrogate $f^\star$ which is simpler than f but retains the main, estimable features of f. Adaptation is possible if we cover $f^\star$ instead of f.

way, the richest model that permits adaptive confidence bands under the usual notion of coverage is $\mathcal{F} = \mathbb{R}^n - \mathcal{S}(\epsilon_2, \epsilon_\infty)$.

Theorem 3.5 gives a procedure that comes within a factor of 2 of attaining the lower bound for finite-samples. The procedure conducts goodness of fit tests for subspaces and constructs bands centered on the estimator of the lowest dimensional nonrejected subspace. Such a procedure actually reflects common practice. It is not uncommon to fit a model, check the fit, and if the model does not fit then we fit a more complex model. In this sense, we view our results as providing a rigorous basis for common practice. It is known that pretesting followed by inference does not lead to valid inferences for f (Leeb and and Pötscher, 2005). But if we cant accept that sometimes we cover a surrogate $f^\star$ rather than f, then validity is restored.

These results are proved in Section 4.

1.4 Related Work

The idea of estimating the detectable part of f is present, at least implicitly, in other approaches. Davies and Kovac (2001) separate the data into a simple piece plus a noise piece which is similar in spirit to our approach. Another related idea is scale-space inference due to Marron and Chaudhuri (2000) who focus on inference for all smoothed versions of f rather than f itself. Also related is the idea of oversmoothing as described in Terrell (1990) and Scott and Terrell (1985). Terrell argues that "By using the most smoothing that is compatible with the scale of the problem, we tend to eliminate accidental features." The idea of one-sided inference in Donoho (1988) has a similar spirit. Here, one constructs confidence intervals of the form $[L, \infty)$ for functionals such as the number of modes of a density. Bickel and Ritov (2000) make what they call a "radical proposal" to " ... determine how much bias can be tolerated without [interesting] features being obscured." We view our approach as a way of implementing their suggestion. Another related idea is contained in Donoho (1995) who showed that if $\widehat{f}$ is the soft threshold estimator of a function and $f(x) = \sum_j \theta_j \psi_j(x)$ is an expansion in an unconditional basis, then $\mathbb{P}_f\Big\{ \widehat{f} \preceq f \Big\} \geq 1 - \alpha$ where $\widehat{f} = \sum_j \widehat{\theta}_j \psi_j$ and $\widehat{f} \preceq f$ means that $|\widehat{\theta}_j| \leq |\theta_j|$ for all j. Finally, we remind the reader that there is a plethora of work on adaptative estimation; see, for example, Cai and Low (2004) and references therein.

1.5 Notation

If L and U are random functions on $\mathcal{X} = \{x_1, \ldots, x_n\}$ such that $L \leq U$, we define $B = (L, U)$ to be the (random) set of all functions g on $\mathcal{X}$ for which $L \leq g \leq U$. We call B (or equivalently, the pair L, U) a band; the band covers a function f if $f \in B$ (or equivalently, if $L \leq f \leq U$). Define its width to be the random variable

$$W = \|U - L\|_\infty = \max_{1 \leq i \leq n} (U(x_i) - L(x_i)). \tag{12}$$

Because we are constructing bands on $\mathcal{X} = \{x_1, \ldots, x_n\}$, we most often refer to functions

in terms of their evaluations $f = (f(x_1), \ldots, f(x_n)) \in \mathrm{R}^n$. When we need to refer to a space of functions to which f belongs, we use a $\widetilde{\ }$ to denote the function space and no $\widetilde{\ }$ to denote the vector space of evaluations. Thus, if $\widetilde{\mathcal{A}}$ is the space of all functions, then $\mathcal{A} = \mathrm{R}^n$. In both cases, we use the same symbol for the function and let the meaning be clear from context; for example, $f \in \widetilde{\mathcal{A}}$ is the function and $f \in \mathcal{A}$ is the vector $(f(x_1), \ldots, f(x_n))$. Define the following norms on $\mathbb{R}^n$:

$$\begin{aligned} ||f|| &= ||f||_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_i^2} \\ ||f||_\infty &= \max_i |f_i|. \end{aligned}$$

We use $\langle \cdot, \cdot \rangle$ to denote the inner product $\langle f, g \rangle = \frac{1}{n}\sum_{i=1}^n f_i g_i$ corresponding to $\|\cdot\|$.

If $\mathcal{F}$ is a subspace of R^n, we define $\Pi_{\mathcal{F}}$ to be the Euclidean projection onto $\mathcal{F}$, using just Π if the subspace is clear from context. We use

$$e_i = (\underbrace{0, \ldots, 0}_{i-1 \text{ times}}, 1, \underbrace{0, \ldots, 0}_{n-i \text{ times}})^T \tag{13}$$

to denote the standard basis on R^n.

If F_θ is a family of CDFs indexed by θ, we write $F_\theta^{-1}(\alpha)$ to denote the upper-tail α-quantile of F_θ. For the standard normal distribution, however, we use z_α to denote the upper-tail α-quantile, and we denote the CDF and PDF, respectively, by Φ and ϕ.

Throughout the paper we assume that σ is a known constant; in some cases we simply set $\sigma = 1$. But see Remark 3.1 about the unknown σ case.

2 Nonadaptivity of Bands

In this section we construct lower bounds on the width of valid confidence bands analagous to (8) and we show that the lower bound is achieved by fixed-width bands.

Low (1997) considered estimating a density f in the class

$$\mathcal{F}(a, k, M) = \left\{ f : \ f \geq 0, \ \int f = 1, \ f(x_0) \leq a, \ ||f^{(k)}(x)||_\infty \leq M \right\}.$$

He shows that if C_n is a confidence interval for $f(0)$, that is,

$$\inf_{f \in \mathcal{F}(a,k,M)} \mathbb{P}_f\{f(0) \in C_n\} \geq 1 - \alpha,$$

then, for every $\epsilon > 0$, there exists $N = N(\epsilon, M)$ and $c > 0$ such that, for all $n \geq N$,

$$\mathsf{E}_f(\mathrm{length}(C_n)) \geq c\, n^{-k/(2k+1)} \tag{14}$$

for all $f \in \mathcal{F}(a, k, M)$ such that $f(0) > \epsilon$. Moreover, there exists a fixed-width confidence interval C_n and a constant c_1 such that $\mathsf{E}_f(\mathrm{length}(C_n)) \leq c_1 n^{-k/(2k+1)}$ for all $f \in \mathcal{F}(a, k, M)$. Thus,

the data play no role in constructing a rate-optimal band, except in determining the center of the interval.

For example, if we use kernel density estimation, we could construct an optimal bandwidth $h = h(n, k)$ depending only on n and k – but not the data – and construct the interval from that kernel estimator. This makes the interval highly dependent on the minimal amount of smoothness k that is assumed. And it rules out the usual data-dependent bandwidth methods such as cross-validation.

Now return to the regression model

$$Y_i = f_i + \sigma\,\epsilon_i, \quad i = 1, \ldots, n, \tag{15}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent, $\mathrm{Normal}(0, 1)$ random variables, and $f = (f_1, \ldots, f_n) \in \mathbb{R}^n$.

THEOREM 2.1. *Let $B = (L, U)$ be a $1 - \alpha$ confidence band over Θ, where $0 < \alpha < 1/2$ and let $g \in \Theta$. Suppose that Θ contains a finite set of vectors Ω, such that:*

1. *for every distinct pair $f, \nu \in \Omega$, we have $\langle f - g, \nu - g\rangle = 0$ and*
2. *for some $0 < \epsilon < (1/2) - \alpha$,*

$$\max_{f\in\Omega} \frac{e^{n||f-g||^2/\sigma^2}}{|\Omega|} \leq \epsilon^2. \tag{16}$$

Then,

$$\mathsf{E}_g(W) \geq (1 - 2\alpha - 2\epsilon) \min_{f\in\Omega} ||g - f||_\infty. \tag{17}$$

We begin with the case where $\Theta = \mathbb{R}^n$. We will obtain a lower bound on the width of any confidence band and then show that a fixed-width procedure attains that width. The results hinge on finding a least favorable configuration of mean vectors that are as far away from each as possible in L^∞ while staying a fixed distance ϵ in total-variation distance.

THEOREM 2.2. *Let $\mathcal{H} = \mathbb{R}^n$ and fix $0 < \alpha < 1/2$. Let $B = (L, U)$ be a $1 - \alpha$ confidence band over $\mathcal{H}$. Then, for every $0 < \epsilon < (1/2) - \alpha$,*

$$\inf_{f\in\mathbb{R}^n} \mathsf{E}_f(W) \geq (1 - 2\alpha - 2\epsilon)\sigma\sqrt{\log(n\epsilon^2)}. \tag{18}$$

The bound is achieved (up to constants) by the fixed-width Bonferroni bands:

$$\ell_i = Y_i - \sigma z_{\alpha/n},\ u_i = Y_i + \sigma z_{\alpha/n}.$$

THEOREM 2.3 (LIPSHSCHITZ BALLS). *Define $x_i = i/n$ for $1 \leq i \leq n$. Let*

$$\widetilde{\mathcal{H}}(L) \;=\; \left\{f :\ |f(x) - f(y)| \leq L|x - y|, \quad x, y \in [0, 1]\right\}, \tag{19}$$

be a ball in Lipschitz space, and let

$$\mathcal{H}(L) \;=\; \{(f(x_1),\ldots,f(x_n)):\; f\in\widetilde{\mathcal{H}}(L)\} \tag{20}$$

be the vector of evaluations on $\mathcal{X}$ Fix $0<\alpha<1/2$ and let $B=(L,U)$ be a $1-\alpha$ confidence band over $\mathcal{H}(L)$. Then, for every $0<\epsilon<(1/2)-\alpha$,

$$\inf_{f\in\mathcal{H}(L)} \mathsf{E}_f(W) \geq a_n \tag{21}$$

where

$$a_n \;=\; \left(\frac{\log n}{n}\right)^{1/3}\times\left(\frac{L\sigma^2}{2}\right)^{1/3} \times\left(1+\frac{3\log(1+\epsilon^2)}{\log n}+\frac{2\log(L/(2\sigma))}{\log n}-\frac{\log\left(\frac{1}{3}\log n+\log(1+\epsilon^2)+\frac{2}{3}\log(L/(2\sigma))\right)}{\log n}\right).$$

The lower bound is achieved (up to logarithmic factors) by a fixed-width procedure.

THEOREM 2.4 (SOBOLEV BALLS). *Let $\widetilde{\mathcal{H}}(p,c)$ be a Sobolev ball of order p and radius c and let $B=(L,U)$ be a $1-\alpha$ confidence band over $\mathcal{H}(p,c)$. For every $0<\epsilon<(1/2)-\alpha$, for every $\delta>0$, and all large n,*

$$\inf_{F\in\mathcal{H}(p,c-\delta)} \mathsf{E}_F(W) \geq (1-2\alpha-2\epsilon)\left(\frac{c_n}{n^{p/(2p+1)}}\right) \tag{22}$$

for some c_n that increases at most logarithmically. The bound is achieved (up to logarithmic factors) by a fixed-width band procedure.

THEOREM 2.5 (BESOV BALLS). *Let $\widetilde{\mathcal{H}}(p,q,\xi,c)$ be ball of size c in the Besov space $B^{\xi}_{p,q}$ and et $B=(L,U)$ be a $1-\alpha$ confidence band over $\mathcal{H}(p,q,\xi,c)$. For every $0<\epsilon<(1/2)-\alpha$, and every $\delta>0$,*

$$\inf_{f\in\mathcal{H}(p,q,\xi,c-\delta)} \mathsf{E}_f(W) \geq c_n(1-2\alpha-2\epsilon)n^{-1/(1/p-\xi-1/2)}. \tag{23}$$

The bound is achieved (up to logarithmic factors) by a fixed-width procedure.

3 Projection Surrogates

Let $\{\mathcal{F}_T : T\in\mathcal{T}\}$ be a scale of linear subspaces. Let w_T denote the smallest width of any confidence band when it is known that $f\in\mathcal{F}_T$ (defined more precisely below). We would like to define an approporiate surrogate and a procedure that gets as close as possible to the target width w_T when $f\in\mathcal{F}_T$. To clarify the ideas, subsection 3.2 develops our results in the special case where the subspaces are $\{\mathcal{F},\mathrm{R}^n\}$ for a fixed $\mathcal{F}$ of dimension $d<n$. Subsection 3.3 handles the more general case of a sequence of nested subspaces.

3.1 Preliminaries

We begin by defining several quantities that will be used throughout. Let $\tau(\epsilon)$ denote the total variation distance between a $N(0,1)$ and a $N(\epsilon,1)$ distribution. Thus,

$$\tau(\epsilon) = \Phi(\epsilon/2) - \Phi(-\epsilon/2). \tag{24}$$

Then, $\epsilon\phi(\epsilon/2) \leq \tau(\epsilon) \leq \epsilon\phi(0)$ and $\tau(\epsilon) \sim \epsilon\phi(0)$ as $\epsilon \to 0$.

LEMMA 3.1. *If $P = N(f, \sigma^2 I)$ and $Q = N(g, \sigma^2 I)$ are multivariate Normals with $f, g \in \mathbb{R}^n$ then*

$$d_{\mathrm{TV}}(P,Q) = \tau\left(\frac{\sqrt{n}||f-g||}{\sigma}\right). \tag{25}$$

We will need several constants. For $0 < \alpha < 1$ and $0 < \gamma < 1 - 2\alpha$ define

$$\kappa(\alpha,\gamma) = (1/6)\sqrt{2\log(1 + 4(1-\gamma-2\alpha)^2)}. \tag{26}$$

For $0 < \beta < 1 - \xi < 1$ and integer $m \geq 1$ define $Q = Q(m,\beta,\xi)$ to be the solution of

$$\xi = 1 - F_{0,m}(F^{-1}_{Q\sqrt{m},m}(\beta)), \tag{27}$$

where $F_{a,d}$ denotes the CDF of a χ^2 random variable with d degrees of freedom and noncentrality parameter a

LEMMA 3.2. *There is a universal constant $\Lambda(\beta,\xi)$ such that $Q(m,\beta,\xi) \leq \Lambda(\beta,\xi)$ for all $m \geq 1$. For example, $\Lambda(.05,.05) \leq 6.25$. Suppose now that $m = m_n$, $\beta = \beta_n$, and $\xi = \xi_n$ are all functions of n. As long as $-\log\beta_n \leq \log n$ and $-\log\xi_n \leq \sqrt{\log n}$, then $Q(m_n,\beta_n,\xi_n) = O(\sqrt{\log n})$.*

Next, define

$$E(m,\alpha,\gamma) = \max(Q(m,\alpha,\gamma), 2\kappa(\alpha,\gamma)), \tag{28}$$

for $0 < \alpha < 1$ and $0 < \gamma < 1 - 2\alpha$.

Finally, if $\mathcal{F}$ is a subspace of dimension d, define

$$\Omega_{\mathcal{F}} = \max_{1\leq i\leq n} \frac{\|\Pi_{\mathcal{F}} e_i\|}{\|e_i\|}, \tag{29}$$

where e_i is defined in equation (13). Note that $0 \leq \Omega_{\mathcal{F}} \leq 1$. The value of $\Omega_{\mathcal{F}}$ relates to the geometry of $\mathcal{F}$ as a hyperplane embedded in R^n, as seen through the following results.

LEMMA 3.3. *Let $\mathcal{F}$ be a subspace of $\mathbb{R}^n$. Then*

$$\min\left\{\|v\| :\ v \in \mathcal{F}, \|v\|_\infty = \epsilon\right\} = \frac{\epsilon}{\sqrt{n}\Omega_{\mathcal{F}}} \tag{30}$$

$$\max\left\{\|v\|_\infty :\ v \in \mathcal{F}, \|v\| = \epsilon\right\} = \epsilon\sqrt{n}\Omega_{\mathcal{F}}. \tag{31}$$

LEMMA 3.4. *Let $\{\phi_1, \ldots, \phi_J\}$ be orthonormal vectors with respect to $||\cdot||$ in $\mathbb{R}^n$ and let $\mathcal{F}$ be the linear span of these vectors. Then*

$$\Omega_{\mathcal{F}} = \sqrt{\frac{\sum_{j=1}^J \phi_j^2(i)}{n}}. \tag{32}$$

In particular, if $\max_j \max_i \phi_j(i) \leq c$ then

$$\Omega_{\mathcal{F}} \leq c\sqrt{\frac{J}{n}}. \tag{33}$$

LEMMA 3.5. *Let $\{\phi_1, \ldots, \phi_J\}$ be orthonormal functions on $[0,1]$. Define $\mathcal{H}_j$ to be the linear span of $\{\phi_1, \ldots, \phi_j\}$. Let $x_i = i/n$, $i = 1, \ldots, n$ and $\mathcal{F}_j = \{f = (h(x_1), \ldots, h(x_n)) : \ h \in \mathcal{H}_j\}$. Then,*

$$\Omega_{\mathcal{F}} = \sqrt{\frac{\sum_{j=1}^J \phi_j^2(x_i)}{n}} + O(1/n). \tag{34}$$

In particular, if $\max_j \sup_x \phi_j(x) \leq c$ then

$$\Omega_{\mathcal{F}} \leq c\sqrt{\frac{J}{n}} + O(1/n). \tag{35}$$

3.2 Single Subspace

To begin, we start with a single subspace $\mathcal{F}$ of dimension d.

Definition 1 *For given $\epsilon_2, \epsilon_\infty > 0$, define the* surrogate *$f^\star$ of f by*

$$f^\star = \begin{cases} \Pi f & \text{if } ||f - \Pi f||_2 \leq \epsilon_2 \text{ and } ||f - \Pi f||_\infty > \epsilon_\infty \\ f & \text{otherwise.} \end{cases} \tag{36}$$

Define the surrogate set *of f, $F^*(f) = \{f, f^*\}$, which will be a singleton when $f^* = f$. Define the* spoiler set *$\mathcal{S}(\epsilon_2, \epsilon_\infty) = \{f \in \mathrm{R}^n : \ f^\star \neq f\}$ and the* invariant set *$\mathcal{I}(\epsilon_2, \epsilon_\infty) = \{f : \ f^\star = f\}$.*

We give a schematic diagram in Figure 3. The gray area represents $\mathcal{S}(\epsilon_2, \epsilon_\infty)$. These are the functions that preclude adaptivity. Being close to $\mathcal{F}$ in L^2 makes them hard to detect but being far from $\mathcal{F}$ in L^∞ makes them hard to cover. To achieve adaptivity we must settle for sometimes covering $\Pi_{\mathcal{F}} f$.

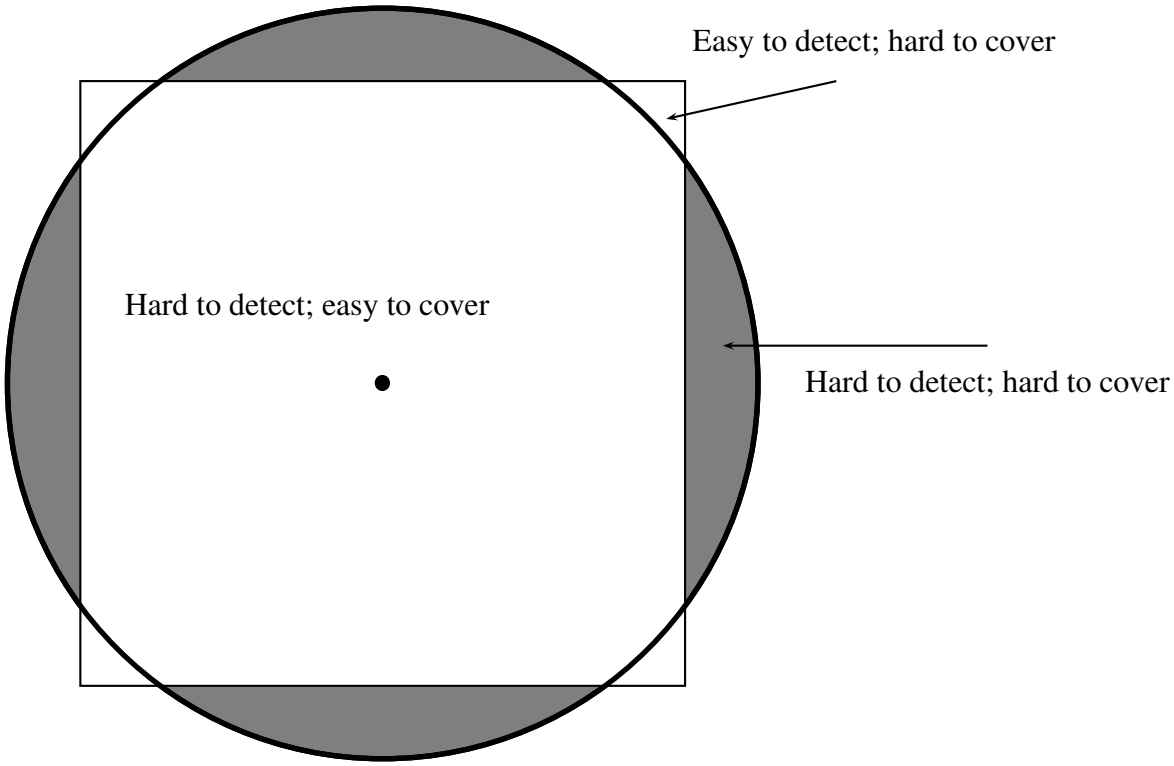

Figure 3: The dot at the center represents the subspace $\mathcal{F}$. The shaded area is the set of spoilers $\mathcal{S}(\epsilon_2, \epsilon_\infty)$ of vectors for which $f^\star \neq f$. If these vectors were not surrogated, adaptation is not possible. The non-shaded area is the invariant set $\mathcal{I}(\epsilon_2, \epsilon_\infty) = \{f : \ f^\star = f\}$.

3.2.1 Lower Bounds

We begin with two lemmas. The first controls the minimum width of a band and the second controls the maximum. The second is of more interest for our purposes; the first lemma is included for completeness. For any $1 \leq p \leq \infty$, $\epsilon > 0$, and $A \subset \mathbb{R}^n$ define

$$M_p(\epsilon, A) = \sup\{d_{\mathrm{TV}}(P_f, P_g) : \ f, g \in A, ||f - g||_p \leq \epsilon\} \tag{37}$$

and

$$m_2(\epsilon, A_0, A_1) = \inf\{d_{\mathrm{TV}}(P_f, P_g) : \ f \in A_0, g \in A_1, \|f - g\|_\infty \geq \epsilon\}. \tag{38}$$

LEMMA 3.6. *Suppose that* $\inf_{f \in A} \mathbb{P}_f\{L \leq f \leq U\} \geq 1 - \alpha$. *Let* $1 \leq p \leq \infty$ *and* $\epsilon > 0$. *For* $f \in A$, *define*

$$\epsilon(f, q) = \sup\{\|f - h\|_q : \ h \in A, \|f - h\|_p \leq \epsilon\},$$

where $1 \leq q \leq \infty$. *Then, for any* $A_0 \subset A$,

$$\inf_{f \in A_0} \mathbb{P}_f\{W > \epsilon(f, \infty)\} \geq 1 - 2\alpha - \sup_{f \in A_0} M_p(\epsilon(f, p), A) \tag{39}$$

where $W = ||U - L||_\infty$. *If every point in* A *is contained in a subset of* A *of* ℓ^p*-diameter* ϵ, *then* $\epsilon(f, p) \equiv \epsilon$, *and*

$$\inf_{f \in A_0} \mathbb{P}_f\{W > \epsilon\} \geq 1 - 2\alpha - M_p(\epsilon, A). \tag{40}$$

LEMMA 3.7. *Suppose that* $\inf_{f\in A}\mathbb{P}_f\{L \le f \le U\} \ge 1-\alpha$. *Suppose that* $A = A_0 \cup A_1$ *(not necessarily disjoint). Let* $\epsilon > 0$ *be such that for each* $f \in A_0$ *there exists* $g \in A_1$ *for which* $\|f-g\|_\infty = \epsilon$. *Then,*

$$\sup_{f\in A_0}\mathbb{P}_f\{W > \epsilon\} \ge 1 - 2\alpha - m_2(\epsilon, A_0, A_1) \tag{41}$$

where $W = ||U-L||_\infty$.

Now we establish the target rate, the smallest width of a band if we knew a priori that $f \in \mathcal{F}$. Define

$$w_{\mathcal{F}} \equiv w_{\mathcal{F}}(\alpha,\gamma,\sigma) = \Omega_{\mathcal{F}}\,\sigma\,\tau^{-1}(1-2\alpha-\gamma). \tag{42}$$

THEOREM 3.1. *Suppose that*

$$\inf_{f\in\mathcal{F}}\mathbb{P}_f\{L \le f \le U\} \ge 1-\alpha. \tag{43}$$

If $\inf_{f\in\mathcal{F}}\mathbb{P}_f\{W \le w\} \ge 1-\gamma$ *then* $w \ge w_{\mathcal{F}}$.

A band that achieves this width, up to logarithmic factors, is $(L,U) = \widehat{f} \pm c$ *where* $\widehat{f} = \Pi Y$ *and* $c = \sigma(\Pi\Pi^T)_{ii} z_{\alpha/2n}$.

Next, we give the main result for this case. Let

$$v_0(\epsilon_2,\epsilon_\infty,n,\alpha,\gamma,\sigma) = \min\Big\{\sqrt{n}\epsilon_2,\ \epsilon_\infty,\ \sigma\tau^{-1}(1-2\alpha-\gamma)\Big\}, \tag{44}$$

$$v_1(\epsilon_2,n,d,\alpha,\gamma,\sigma) = \begin{cases} 0 & \text{if } \epsilon_2 \ge 2\kappa(\alpha,\gamma)(n-d)^{1/4}n^{-1/2} \\ \kappa(\alpha,\gamma)(n-d)^{1/4}n^{-1/2} & \text{if } \epsilon_2 < 2\kappa(\alpha,\gamma)(n-d)^{1/4}n^{-1/2}, \end{cases} \tag{45}$$

and define

$$v(\epsilon_2,\epsilon_\infty,n,d,\alpha,\gamma,\sigma) = \max\Big\{v_0(\epsilon_2,\epsilon_\infty,n,\alpha,\gamma,\sigma), v_1(\epsilon_2,n,d,\alpha,\gamma,\sigma)\Big\}. \tag{46}$$

THEOREM 3.2 (LOWER BOUND FOR SURROGATE CONFIDENCE BAND WIDTH). *Fix* $0<\alpha<1$ *and* $0<\gamma<1-2\alpha$. *Suppose that for bands* $B=(L,U)$

$$\inf_{f\in\mathrm{R}^n}\mathbb{P}_f\{F^*(f)\cap B \ne \emptyset\} \ge 1-\alpha \tag{47}$$

and that

$$\inf_{f\in\mathcal{F}}\mathbb{P}_f\{W \le w\} \ge 1-\gamma. \tag{48}$$

Then,

$$w \ge \underline{w}(\epsilon_2,\epsilon_\infty,n,d,\alpha,\gamma,\sigma) \equiv \max\Big\{w_{\mathcal{F}}(\alpha,\gamma,\sigma),\ v(\epsilon_2,\epsilon_\infty,n,d,\alpha,\gamma,\sigma)\Big\} \tag{49}$$

The inequality (47) ensures that B is a valid surrogate confidence band: for every function, either the function or its surrogate is covered with at least the target probability. The result gives a probabilistic lower bound on the width of the band that is at least as big as the best a priori width for the subspace. As we will see, with proper choice of ϵ_2 and ϵ_∞, the v term can be made small, giving the subspace width $w_{\mathcal{F}}$ for the lower bound.

Next, we address the question of optimality. Consider, for example, the trivial surrogate that maps all functions to 0. We can cover the surrogate using 0 width bands with probability 1, but this would not be too interesting. There is a tradeoff between the width of the bands on low dimensional subspaces and the volume of the spoiler set, the functions that are surrogated. We characterize optimality here as minimizing the volume of the spoiler set $\mathcal{S}(\epsilon_2, \epsilon_\infty)$ while still attaining the target width with high probability when f truly lies in the subspace. In this sense, the surrogate defined above is optimal.

THEOREM 3.3 (OPTIMALITY). *Let $\underline{w}$ denote the right hand side of inequality (49). Then $\underline{w} \geq w_{\mathcal{F}}$, where $w_{\mathcal{F}}$ is defined in (42). Setting*

$$\epsilon_2 = 2\kappa(\alpha, \gamma)(n-d)^{1/4} n^{-1/2}, \quad \epsilon_\infty = w_{\mathcal{F}}$$

minimizes $\mathrm{Volume}(\mathcal{S}(\epsilon_2, \epsilon_\infty))$ *subject to achieving the lower bound on $\underline{w}$.*

3.2.2 Achievability

Having established a lower bound, we need to show that the lower bound is sharp. We do this by constructing a finite-sample procedure that achieves the bound within a factor of 2. Let $F_{a,d}$ denote the CDF of a χ^2 random variable with d degrees of freedom and noncentrality parameter a and let $\chi^2_{\alpha,d} = F^{-1}_{0,d}(1-\alpha)$. Let $T = ||Y - \Pi Y||^2$ and define

$$B = (L, U) = \widehat{f} \pm c\sigma \tag{50}$$

where

$$\widehat{f} = \begin{cases} Y & \text{if } T > \chi^2_{\gamma,n-d} \\ \Pi Y & \text{if } T \leq \chi^2_{\gamma,n-d} \end{cases} \tag{51}$$

and

$$c = \begin{cases} z_{\alpha/(2n)} & \text{if } T > \chi^2_{\gamma,n-d} \\ \omega_{\mathcal{F}} + \epsilon_\infty & \text{if } T \leq \chi^2_{\gamma,n-d}. \end{cases} \tag{52}$$

THEOREM 3.4. *If*

$$\gamma \geq 1 - F_{0,n-d}(F^{-1}_{n\epsilon_2^2,n-d}(\alpha/2)) \tag{53}$$

then

$$\inf_{f \in \mathbb{R}^n} \mathbb{P}_f\{F^\star(f) \cap B \neq \emptyset\} \geq 1 - \alpha \tag{54}$$

and

$$\inf_{f\in\mathcal{F}} \mathbb{P}_f\{W \le w_{\mathcal{F}} + \epsilon_\infty\} \ge 1-\gamma. \tag{55}$$

If $\epsilon_2 \ge E(n-d,\alpha/2,\gamma)(n-d)^{1/4}n^{-1/2}$*, where* $E(m,\alpha,\gamma)$ *is defined in (28), then*

$$\inf_{f\in\mathcal{F}} \mathbb{P}_f\{W \le 2\underline{w}(\epsilon_2,\epsilon_\infty,\alpha,\gamma,n,d)\} \ge 1-\gamma. \tag{56}$$

where $\underline{w}(\epsilon_2,\epsilon_\infty,\alpha,\gamma,n,d)$ *is defined (49). Hence, the procedure adapts to within a constant factor of the lower bound* $\underline{w}$ *given in Theorem 3.2.*

COROLLARY 3.1. *Setting*

$$\epsilon_2 = E(n-d,\alpha/2,\gamma)(n-d)^{1/4}n^{-1/2}, \quad \epsilon_\infty = w_{\mathcal{F}}$$

in the above procedure, minimizes $\mathrm{Volume}(\mathcal{S}(\epsilon_2,\epsilon_\infty))$ *subject to satisfying (56).*

REMARK 3.1. The results can be extended to unknown σ by replacing σ with a nonparametric estimate $\widehat{\sigma}$. However, the results are then asymptotic rather than finite sample. Moreover, a minimal amount of smoothness is required to ensure that $\widehat{\sigma}$ consistently estimates σ; see Genovese and Wasserman (2005). So as not to detract from our main points, we continue to take σ known.

3.2.3 Remarks on Estimation and the Modulus of Continuity

It is interesting to note that the bands defined above cover the true f over a set V that is larger than $\mathcal{F}$. In this section we take a brief look at the properties of V.

Define

$$C(\alpha,a,b) = \sup_{u>0}(au+b)\left(1-\alpha-\frac{1}{4}+\frac{1}{2}\Phi(-u/2)\right), \tag{57}$$

and let $C(\alpha) \equiv C(\alpha,1,0)$. Let $\mathcal{F}^\perp$ be the orthogonal complement of $\mathcal{F}$. Let $B_k^\perp(0,\epsilon)$ be a ℓ^k-ball around 0 in $\mathcal{F}^\perp$ ($k=2,\infty$). For $f\in\mathbb{R}^n$, let $B_k^\perp(f,\epsilon) = f + B_k^\perp(0,\epsilon)$. Define

$$V \equiv V(\epsilon_2,\epsilon_\infty) = \bigcup_{f\in\mathcal{F}}\left(B_2^\perp(f,\epsilon_2)\cap B_\infty^\perp(f,\epsilon_\infty)\right). \tag{58}$$

LEMMA 3.8. *Let* $B=(L,U)$ *be defined as in (50). Then*

$$\inf_{f\in V} \mathbb{P}_f\{L \le f \le U\} \ge 1-\alpha. \tag{59}$$

Let $Tf = f_1$. The next lemma gives the modulus of continuity (Donoho and Liu 1991) of T over V which measures the difficulty of estimation over V. The modulus of continuity of T over a set $\mathcal{A}$ is

$$\omega(u,\mathcal{A}) = \sup\{|Tf-Tg| : \ \|f-g\|_2 \le u; f,g\in\mathcal{A}\}. \tag{60}$$

Donoho and Liu showed that the difficulty of estimation over $\mathcal{A}$ is often characterized by $\omega(1/\sqrt{n},\mathcal{A})$ in the sense that this quantity defines a lower bound on estimation rates.

LEMMA 3.9 (MODULUS OF CONTINUITY). *We have*

$$\omega(u, V) = \left(u\Omega\sqrt{n}\sqrt{\frac{\Omega^2}{1+\Omega^2}} + \min\left(\frac{u\sqrt{n}}{\sqrt{1+\Omega^2}}, \epsilon_2 \wedge (\epsilon_\infty/\sqrt{n}) \right) \right). \tag{61}$$

Note that when $\epsilon_2 = \epsilon_\infty = 0$ and $\Omega \sim \sqrt{d/n}$, we have $\omega(1/\sqrt{n}, \mathcal{A}) \sim \sqrt{d/n}$ as expected. However, when $\epsilon \equiv \epsilon_2 = \epsilon_\infty/\sqrt{n}$ is large we will have that $\omega(1/\sqrt{n}, \mathcal{A}) \sim \sqrt{d/n} + \epsilon/\sqrt{1+d^2/n}$. The extra term $\epsilon/\sqrt{1+d^2/n}$ reflects the "ball-like" behavior of V in addition to the subspace-like behavior of V. The bands need to cover over this extra set to maintain valid coverage and this leads to larger lower bounds than just covering over $\mathcal{F}$.

3.3 Nested Subspaces

Now suppose that we have nested subspaces $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_m \subset \mathcal{F}_{m+1} \equiv \mathbb{R}^n$. Let Π_j denote the projector onto $\mathcal{F}_j$. We define the surrogate as follows.

Definition 2 *For given $\epsilon_2 = (\epsilon_{2,1}, \ldots, \epsilon_{2,m})$ and $\epsilon_\infty = (\epsilon_{\infty,1}, \ldots, \epsilon_{\infty,m})$ define*

$$\mathcal{J}(f) = \{1 \leq j \leq m : \ ||f - \Pi_j f||_2 \leq \epsilon_{2,j} \text{ and } ||f - \Pi_j f||_\infty > \epsilon_{\infty,j} \Big\}. \tag{62}$$

Then define the surrogate set

$$F^\star(f) = \{\Pi_j f : \ j \in \mathcal{J}(f)\} \cup \{f\}. \tag{63}$$

Definition 3 *We say that $B = \{g : \ L \leq g \leq U\} \equiv (L, U)$ has coverage $1-\alpha$ if*

$$\inf_{f \in \mathbb{R}^n} \mathbb{P}_f\{F^\star \cap B \neq \emptyset\} \geq 1 - \alpha. \tag{64}$$

Define, $T_j = ||Y - \Pi_j Y||^2$, $\widehat{f} = \Pi_{\widehat{J}} Y$, where

$$\widehat{J} = \min\{j : \ T_j \leq \chi^2_{\gamma, n-d_j}\}, \tag{65}$$

$$c_j = \begin{cases} \omega_{\mathcal{F}_j}(\alpha_j) + \epsilon_{\infty,j} & \text{if } 1 \leq j \leq m \\ z_{\alpha_{m+1}/n} & \text{if } j = m+1. \end{cases} \tag{66}$$

and finally let $B = (L, U) = \widehat{f} \pm c_{\widehat{J}}\sigma$ where $\sum_j \alpha_j \leq \alpha$.

THEOREM 3.5. *If,*

$$\gamma \geq 1 - \min_j F_{0,n-d_j}(F^{-1}_{n\epsilon^2_{2,j}, n-d_j}(\alpha_j)) \tag{67}$$

then

$$\inf_{f \in \mathbb{R}^n} \mathbb{P}_f\{F^\star \cap B \neq \emptyset\} \geq 1 - \alpha. \tag{68}$$

Let $w_j = w_{\mathcal{F}_j}(\alpha_j) + \epsilon_{\infty,j}$. *If* $w_1 \leq \cdots \leq w_{m+1}$ *then*

$$\inf_{f \in \mathcal{F}_j} \mathbb{P}_f\{W \leq w_j\} \geq 1 - \gamma. \tag{69}$$

If in addition $\epsilon_{2,j} \geq E(n - d_j, \alpha_j, \gamma)(n - d_j)^{1/4} n^{-1/2}$ *and* $\epsilon_{\infty,j} \leq w_{\mathcal{F}_j}$ *then*

$$\inf_{f \in \mathcal{F}_j} \mathbb{P}_f\{W \leq 2\underline{w}(\epsilon_{2,j}, \epsilon_{\infty,j}, \alpha_j, \gamma, n, d_j)\} \geq 1 - \gamma \tag{70}$$

where $\underline{w}(\epsilon_{2,j}, \epsilon_{\infty,j}, \alpha_j, \gamma, n, d_j)$ *is defined (49). Hence, the procedure adapts to within a constant factor of the lower bound* $\underline{w}$ *given in Theorem 3.2.*

COROLLARY 3.2. *Suppose* $\alpha_1 = \cdots = \alpha_{m+1} = \alpha/(m+1)$. *Then* $w_1 \leq \cdots \leq w_{m+1}$ *so (69) holds. Moreover, setting*

$$\epsilon_{2,j} = E(n - d_j, \alpha_j, \gamma)(n - d_j)^{1/4} n^{-1/2} \tag{71}$$

and

$$\epsilon_{\infty,j} = w_{\mathcal{F}_j} \tag{72}$$

in the above procedure, minimizes $\mathrm{Volume}(\mathcal{S}(\epsilon_2, \epsilon_\infty))$ *subject to satisfying (70).*

EXAMPLE 3.1. *Suppose that* $x_i = i/n$ *and let* $B_1 = [0, 1/d], B_2 = (1/d, 2/d], \ldots, B_d = ((d-1)/d, 1]$. *Write* $f = (f(x_i) : i = 1, \ldots, n)$ *and let* $\mathcal{F}$ *denote the subspace of vectors* f *that are constant over each* B_j. *Then* $\Omega_{\mathcal{F}} = \sqrt{d/n}$. *The above procedure then produces a band with width no more that* $O(\sqrt{d/n})$ *with probability at least* $1 - \gamma$.

4 Proofs

In this section, we prove the main results. We omit proofs for a few of the simpler lemmas. Throughout this section, we write $x_n = O^*(b_n)$ to mean that $x_n = O(c_n b_n)$ where c_n increases at most logarithmically with n.

The following lemma is essentially from Section 3.3 of Ingster and Suslina (2003).

LEMMA 4.1. *Let* M *be a probability measure on* $\mathbb{R}^n$ *and let*

$$Q(\cdot) = \int P_f(\cdot) dM(f)$$

where $P_f(\cdot)$ *denotes the measure for a multivariate Normal with mean* $f = (f_1, \ldots, f_n)$ *and covariance* $\sigma^2 I$. *Then*

$$L_1(Q, P_g) \leq \sqrt{\int \int \exp\left\{\frac{n\langle f - g, \nu - g\rangle}{\sigma^2}\right\} dM(f) dM(\nu) - 1}. \tag{73}$$

In particular, if Q is uniform on a finite set Ω, then

$$L_1(Q, P_g) \leq \sqrt{\left(\frac{1}{|\Omega|}\right)^2 \sum_{f,\nu \in \Omega} \exp\left\{\frac{n\langle f-g, \nu-g\rangle}{\sigma^2}\right\} - 1}. \tag{74}$$

PROOF OF LEMMA 4.1. Let p_f denote the density of a multivariate Normal with mean f and covariance $\sigma^2 I$ where I is the identity matrix. Let q be the density of Q:

$$q(y) = \int p_f(y) dM(f).$$

Then,

$$\begin{aligned} \int |p_g(x) - q(x)| dx &= \int \frac{|p_g(x) - q(x)|}{\sqrt{p_g(x)}} \sqrt{p_g(x)} dx \\ &\leq \sqrt{\int \frac{(p_g(x) - q(x))^2}{p_g(x)} dx} = \sqrt{\int \frac{q^2(x)}{p_g(x)} dx - 1}. \end{aligned} \tag{75}$$

Now,

$$\begin{aligned} \int \frac{q^2(x)}{p_g(x)} dx &= \int \left(\frac{q(x)}{p_g(x)}\right)^2 p_g(x) dx = \mathsf{E}_g \left(\frac{q(x)}{p_g(x)}\right)^2 \\ &= \int\int \mathsf{E}_g \left(\frac{p_f(x) p_\nu(x)}{p_g^2(x)}\right) dM(f) dM(\nu) \\ &= \int\int \exp\left\{-\frac{n}{2\sigma^2}(||f-g||^2 + ||\nu-g||^2)\right\} \mathsf{E}_g \left(\exp\left\{\epsilon^T (f+\nu-2g)/\sigma^2\right\}\right) dM(f) dM(\nu) \\ &= \int\int \exp\left\{-\frac{n}{2\sigma^2}(||f-g||^2 + ||\nu-g||^2)\right\} \exp\left\{\sum_{i=1}^n (f_i - g_i + \nu_i - g_i)^2/(2\sigma^2)\right\} dM(f) dM(\nu) \\ &= \int\int \exp\left\{\frac{n\langle f-g, \nu-g\rangle}{\sigma^2}\right\} dM(f) dM(\nu) \end{aligned}$$

and the result follows from (75). □

PROOF OF THEOREM 2.1.. Let $N = |\Omega|$ and let $b = n \max_{f\in\Omega} ||f-g||^2$. Let p_f denote the density of a multivariate Normal with mean f and covariance $\sigma^2 I$ where I is the identity matrix. Define the mixture

$$q(y) = \frac{1}{N} \sum_{f\in\Omega} p_f(y).$$

By Lemma 4.1,

$$\int |p_g(x) - q(x)| dx \leq \sqrt{\left(\frac{1}{N}\right)^2 \sum_{f,\nu\in\Omega} \exp\left\{\frac{n\langle f-g, \nu-g\rangle}{\sigma^2}\right\} - 1}$$

$$= \sqrt{\left(\frac{1}{N}\right)^2 \left[Ne^{b^2/\sigma^2} + N(N-1)\right] - 1}$$
$$\leq \sqrt{e^{b^2/\sigma^2}/N} = \epsilon.$$

Define two events, $A = \{\ell \leq g \leq u\}$ and $B = \{\ell \leq f \leq u, \text{ for some } f \in \Omega\}$. Then, $A \cap B \subset \{w_n \geq a\}$ where

$$a = \min_{f \in \Omega} ||g - f||_\infty.$$

Since $\mathbb{P}_f\{\ell \leq f \leq u\} \geq 1 - \alpha$ for all f, it follows that $\mathbb{P}_f\{B\} \geq 1 - \alpha$ for all $f \in \Omega$. Hence, $Q(B) \geq 1 - \alpha$. So,

$$\begin{aligned}
\mathbb{P}_g\{w_n \geq a\} &\geq \mathbb{P}_g\{A \cap B\} \geq Q(A \cap B) - \epsilon = Q(A) + Q(B) - Q(A \cup B) - \epsilon \\
&\geq Q(A) + Q(B) - 1 - \epsilon \geq Q(A) + (1-\alpha) - 1 - \epsilon \geq \mathbb{P}_g\{A\} + (1-\alpha) - 1 - 2\epsilon \\
&\geq (1-\alpha) + (1-\alpha) - 1 - 2\epsilon = 1 - 2\alpha - 2\epsilon.
\end{aligned}$$

So, $\mathsf{E}_g(w_n) \geq (1 - 2\alpha - 2\epsilon)a$. $\square$

PROOF OF THEOREM 2.2. Let $g \in \mathbb{R}^n$ be arbitrary, let

$$a_n = \sigma\sqrt{\log(n\epsilon^2)}$$

and define

$$\Omega = \left\{g + (a_n, 0, \ldots, 0),\ g + (0, a_n, \ldots, 0),\ \ldots,\ g + (0, 0, \ldots, a_n)\right\}.$$

Then the conditions of Theorem 2.1 are satisfied with $N = n$, and hence

$$\mathsf{E}_g(W) \geq (1 - 2\alpha - 2\epsilon) \min_{f \in \Omega} ||g - f||_\infty = (1 - 2\alpha - 2\epsilon)a_n. \tag{76}$$

This is true for each g and hence (18) follows. The last statement of the theorem follows from standard Gaussian tail inequalities. $\square$

PROOF OF THEOREM 2.3. We construct the appropriate set Ω and apply Theorem 2.1. For simplicity, we build Ω around $g = (0, \ldots, 0)$, the extension to arbitrary g being straightforward. Set $a = a_n$ from the statement of the theorem, and define

$$F(x) = \begin{cases} Lx & 0 \leq x \leq a/L \\ 2a - Lx & a/L \leq x \leq 2a/L. \end{cases}$$

Note that $F \in \mathcal{F}(L)$ and that F minimizes $||F||_2$ among all $F \in \mathcal{F}(L)$ with $||F||_\infty = a$. For simplicity, assume that $2aN/L = 1$ for some integer N. Define $F_1(\cdot) = F(\cdot)$, $F_2(\cdot) = F(\cdot - \delta)$,..., and $F_N(\cdot) = F(\cdot - N\delta)$. Let $\Omega(a) = \{f_1, \ldots, f_N\}$ where $f_j = (F_j(x_1), \ldots, F_j(x_n))$. Now

$$n||f_j||^2 \leq \frac{2na^3}{3L}$$

and so

$$\frac{e^{n||f_j||^2/\sigma^2}}{N} \leq \epsilon^2.$$

Now apply Theorem 2.1.

To prove the last statement, we note that it is well known that if $\widehat{F}$ is a kernel estimator with triangular kernel and bandwidth $h = O(n^{-1/3})$ then

$$\sup_{f \in \Theta} E_F(||\widehat{F} - F||_\infty) \leq C \left(\frac{\log n}{n}\right)^{1/3} \equiv C_n$$

for some $C > 0$. Then $B = (\widehat{F} - \frac{C_n}{\alpha}, \widehat{F} + \frac{C_n}{\alpha})$ (restricted to $x_i = i/n$) is valid by Markov's inequality and has the rate a_n. $\square$

PROOF OUTLINE OF THEOREM 2.4. We will use the fact that an appropriately chosen wavelet basis forms a basis for $\mathcal{F}$. Let

$$J_n \sim \log_2 \left(\frac{n^{1/(2p+1)}}{\log n}\right),$$

$$b_n = \frac{\sigma}{\sqrt{n}} \sqrt{\log(2^{J_n} \epsilon^2)}$$

and

$$F(x) = b_n 2^{J_n/2} \psi(2^{J_n} x)$$

where ψ is a compactly supported mother wavelet. Then $F^{(p)} = b_n 2^{J_n/2} 2^{pJ_n} \psi^{(p)}(2^{J_n} x)$ so that $\int (F^{(p)})^2 < c^2$ for all large n so that $F \in \mathcal{F}$.

Let $f = (F(x_i), \ldots, F(x_n))$. Then,

$$||f||_\infty = b_n 2^{J_n/2} = O^*(n^{-p/(2p+1)})$$

and $\sqrt{n}||f||_2 \sim \sqrt{n} b_n$. Let $f_k = (F(x_1 - k\Delta), \ldots, F(x_n - k\Delta))^T$ where Δ is just large enough so that the F_k's are orthogonal. Hence, $\Delta \approx 1/N$ where $N \sim 2^{J_n}$. Finally, set $\Omega = \{f_1, \ldots, f_N\}$. Then,

$$\frac{e^{n||f||^2/\sigma^2}}{N} = e^{nb_n^2/\sigma^2} 2^{J_n} \leq \epsilon^2$$

for each $f \in \Omega$. The lower bound follows from Theorem 2.1.

A fixed-width procedure that achieves the bound is

$$\ell_i = \widehat{f}_i - c_n z_{\alpha/n}, \; u_i = \widehat{f}_i + c_n z_{\alpha/n}.$$

where $\widehat{f}_i = \widehat{F}(x_i)$,

$$\widehat{F}(x) = \sum_j \widehat{\alpha}_j \phi_j(x) + \sum_{j=1}^{J} \sum_k \widehat{\beta}_{jk} \psi_{jk}(x),$$

$\widehat{\alpha}_j = n^{-1} \sum_i Y_i \phi_j(x_i)$, $\widehat{\beta}_{jk} = n^{-1} \sum_i Y_i \psi_{jk}(x_i)$ and $c_n = \sqrt{\max_x \text{Var}(\widehat{F}(x))}$. $\square$

PROOF OUTLINE OF THEOREM 2.5. Again, we use the fact that an appropriately chosen wavelet basis forms a basis for $\mathcal{F}$. Let

$$J_n \sim \frac{\log_2 \frac{c\sqrt{n}}{\sigma\sqrt{\log 2^J \epsilon^2}}}{\xi + \frac{1}{2} - \frac{1}{p}}.$$

Let

$$a_n = \frac{\sigma}{\sqrt{n}}\sqrt{\log 2^J \epsilon^2}$$

and define $F(x) = a_n 2^{J/2}\psi(x)$, where ψ is a compactly supported mother wavelet. Then, $||f|| = a_n$, $||f||_\infty = a_n 2^{J/2}$, and $||F||_{p,q}^{\xi} \leq c - \delta$ for all large n. Take Ω around g to be non-overlapping translations of F added to g. Then $N \sim 2^J$ and conditions of Theorem 2.1 hold. Moreover,

$$a_n = O^*(n^{-1/(1/p-\xi-1/2)}).$$

The bound is achieved by Markov applied to the soft-thresholded wavelet estimator with universal thersholding. □

PROOF OF LEMMA 3.3. Note that

$$\begin{aligned}
\min\left\{\|v\| :\ v \in \mathcal{F}, \|v\|_\infty = 1\right\} &= \min_{v\in\mathcal{F}} \frac{\|v\|}{\|v\|_\infty} && (77)\\
&= \frac{1}{\max_{v\in\mathcal{F}} \frac{\|v\|_\infty}{\|v\|}}, && (78)\\
&= \frac{1}{\max\left\{\|v\|_\infty :\ v \in \mathcal{F}, \|v\| = 1\right\}}. && (79)
\end{aligned}$$

If v solves one of these problems then ϵv solves the more general version in the statement of the lemma. It now suffices to show just the second equality.

Now, $\Omega_{\mathcal{F}} = \max_i \Omega_i$ where

$$\Omega_i = \frac{\langle e_i, \Pi_{\mathcal{F}} e_i\rangle}{\|e_i\|\,\|\Pi_{\mathcal{F}} e_i\|} = \frac{\|\Pi_{\mathcal{F}} e_i\|}{\|e_i\|}.$$

Maximizing $f_i = e_i^T f$ for $f \in \mathcal{F}$ and $\|f\| \leq 1$ is equivalent to maximizing $n\langle e_i, f\rangle = n\langle \Pi_{\mathcal{F}} e_i, f\rangle$. The maximum subject to the constraint occurs at $f^\star = \Pi e_i/\|\Pi e_i\|$. Hence, the maximum is $e_i^T f^\star = (\Pi e_i)^T f^\star = n\|\Pi e_i\|^2/\|\Pi e_i\| = n\|\Pi e_i\|^2/\|\Pi e_i\|\frac{\|e_i\|}{\|e_i\|} = \sqrt{n}\Omega_i$. Maximizing over i completes the proof. □

PROOF OF LEMMA 3.6. Let $f, g \in A$ be such that $||f - g||_p \leq \epsilon$. Then,

$$\begin{aligned}
\mathbb{P}_g\{L \leq f \leq U\} &= \mathbb{P}_f\{L \leq f \leq U\} + \mathbb{P}_g\{L \leq f \leq U\} - \mathbb{P}_f\{L \leq f \leq U\} && (80)\\
&\geq \mathbb{P}_f\{L \leq f \leq U\} - d_{\mathrm{TV}}(P_f, P_g) && (81)\\
&\geq 1 - \alpha - M_p(||f - g||_p, A) && (82)\\
&\geq 1 - \alpha - M_p(\epsilon(f, p), A). && (83)
\end{aligned}$$

We also have that $\mathbb{P}_g\{L \leq g \leq U\} \geq 1-\alpha$. Hence,

$$\begin{aligned}
\mathbb{P}_g\{L \leq g \leq U, L \leq f \leq U\} &\geq \mathbb{P}_g\{L \leq g \leq U\} + \mathbb{P}_g\{L \leq f \leq U\} - 1 && (84)\\
&\geq 1-\alpha+1-\alpha - M_p(\epsilon(f,p),A) - 1 && (85)\\
&\geq 1-2\alpha - M_p(\epsilon(f,p),A). && (86)
\end{aligned}$$

The event $\{L \leq g \leq U, L \leq f \leq U\}$ implies that $W \geq \|g-f\|_\infty$. Hence,

$$\begin{aligned}
\mathbb{P}_f\{W > ||f-g||_\infty\} &\geq 1-2\alpha - M_p(\epsilon(f,p),A)\\
&\geq 1-2\alpha - M_p(\epsilon(f,p),A)\\
&\geq 1-2\alpha - M_p(\epsilon,A).
\end{aligned}$$

It follows then that

$$\mathbb{P}_f\{W > \epsilon(f,\infty)\} = \inf_g \mathbb{P}_f\{W > ||f-g||_\infty\}. \quad (87)$$

and thus

$$\inf_{f\in A_0} \mathbb{P}_f\{W > \epsilon(f,\infty)\} \geq 1-2\alpha - \sup_{f\in A_0} M_p(\epsilon(f,p),A). \quad (88)$$

This proves the first claim. But $\epsilon(f,\infty) \geq \epsilon(f,p)$ for any $1 \leq p \leq \infty$. The final claim follows immediately. □

PROOF OF LEMMA 3.7. Choose $f \in A_0$. Choose $g \in A_1$ to minimize $d_{\mathrm{TV}}(p_f, p_g)$ such to such that $||f-g||_\infty = \epsilon$. Hence, $d_{\mathrm{TV}}(p_f,p_g) = m_2(\epsilon, A_0, A_1)$. Then,

$$\begin{aligned}
\mathbb{P}_f\{L \leq g \leq U\} &= \mathbb{P}_f\{L \leq f \leq U\} + \mathbb{P}_f\{L \leq f \leq U\} - \mathbb{P}_f\{L \leq f \leq U\} && (89)\\
&\geq \mathbb{P}_f\{L \leq f \leq U\} - d_{\mathrm{TV}}(P_f,P_g) && (90)\\
&\geq 1-\alpha - m_2(\epsilon,A_0,A_1). && (91)
\end{aligned}$$

We also have that $\mathbb{P}_f\{L \leq f \leq U\} \geq 1-\alpha$. Hence,

$$\begin{aligned}
\mathbb{P}_f\{L \leq f \leq U, L \leq g \leq U\} &\geq \mathbb{P}_f\{L \leq f \leq U\} + \mathbb{P}_f\{L \leq g \leq U\} - 1 && (92)\\
&\geq 1-\alpha+1-\alpha - m_2(\epsilon,A_0,A_1) && (93)\\
&\geq 1-2\alpha - m_2(\epsilon,A_0,A_1). && (94)
\end{aligned}$$

The event $\{L \leq f \leq U, L \leq g \leq U\}$ implies that $W \geq \|f-g\|_\infty$. Hence,

$$\mathbb{P}_f\{W > ||f-g||_\infty\} \geq 1-2\alpha - m_2(\epsilon,A_0,A_1). \quad (95)$$

It follows then that

$$\sup_{f\in A_0} \mathbb{P}_f\{W > \epsilon\} \geq 1-2\alpha - m_2(\epsilon,A_0,A_1). \quad (96)$$

□

PROOF OF THEOREM 3.1. First, we compute $m_2(\epsilon, \mathcal{F}, \mathcal{F})$. Note that $d_{TV}(f, 0) = \tau(\sqrt{n}\|f\|)$. Hence, $m_2(\epsilon, \mathcal{F}, \mathcal{F}) = \tau(\sqrt{n}v)$ where $v = \min\{||f|| : \ f \in \mathcal{F}, \ \|f\|_\infty = \epsilon\}$. By Lemma 3.3, $v = \epsilon/(\sqrt{n}\Omega_{\mathcal{F}})$. It follows by Lemma 3.6 that

$$\sup_{f\in\mathcal{F}} \mathbb{P}\{W > w\} \geq 1 - 2\alpha - \tau\left(\frac{w}{\Omega_{\mathcal{F}}}\right). \tag{97}$$

Let $w_* = \Omega\tau^{-1}(1 - 2\alpha - \gamma)$. It follows that if $w < w_*$ then $\inf_{f\in\mathcal{F}} \mathbb{P}\{W \leq w\} < 1 - \gamma$ which is a contradiction.

That the proposed band has correct coverage follows easily. Now, $(\Pi\Pi^T)_i i \leq \Omega_{cF}$ and $z_{\alpha/2n} \leq \sqrt{c \log n}$ for some c and the claim follows. □

PROOF OF THEOREM 3.2. We break the argument up into three parts.

Part I. First, we compute $m_2(\epsilon, \mathcal{F}, \mathcal{F})$. Note that $d_{TV}(f, 0) = \tau(\sqrt{n}\|f\|)$. Hence, $m_2(\epsilon, \mathcal{F}, \mathcal{F}) = \tau(\sqrt{n}v)$ where $v = \min\{||f|| : \ f \in \mathcal{F}, \ \|f\|_\infty = \epsilon\}$. By Lemma 3.3, $v = \epsilon/(\sqrt{n}\Omega_{\mathcal{F}})$. It follows by Lemma 3.6 that

$$\sup_{f\in\mathcal{F}} \mathbb{P}\{W > w\} \geq 1 - 2\alpha - \tau\left(\frac{w}{\Omega_{\mathcal{F}}}\right). \tag{98}$$

Let $w_* = \Omega\tau^{-1}(1 - 2\alpha - \gamma)$. It follows that if $w < w_*$ then $\inf_{f\in\mathcal{F}} \mathbb{P}\{W \leq w\} < 1 - \gamma$ which is a contradiction.

Part II. *Case (a.)* $\epsilon_2 \leq \epsilon_\infty/\sqrt{n}$. First, note that $m_2(w, \mathcal{F}, V) = \tau(\sqrt{n}\frac{w}{\sqrt{n}}) = \tau(w)$ for $w \leq \sqrt{n}\epsilon_2$, because the minimum two-norm for a given infinity-norm is achieved on the coordinate axis. Second, let $A_0 = \mathcal{F}$ and $A_1 = V$ in Lemma 3.6. Then, for $w \leq \sqrt{n}\epsilon_2$,

$$\sup_{f\in\mathcal{F}} \mathbb{P}\{W > w\} \geq 1 - 2\alpha - \tau(w) \tag{99}$$

Let $w_0 = \sqrt{n}\min(n^{-1/2}\tau^{-1}(1 - 2\alpha - \gamma), \epsilon_2)$, then $\sup_{f\in\mathcal{F}} \mathbb{P}\{W > w_0\} \geq \gamma$.

Case (b.) $\epsilon_2 > \epsilon_\infty/\sqrt{n}$. First, note that $m_2(w, \mathcal{F}, V) = \tau(\sqrt{n}\frac{w}{\sqrt{n}}) = \tau(w)$ for $w \leq \epsilon_\infty$. Second, let $A_0 = \mathcal{F}$ and $A_1 = V$ in Lemma 3.6. Then, for $w \leq \epsilon_\infty$,

$$\sup_{f\in\mathcal{F}} \mathbb{P}\{W > w\} \geq 1 - 2\alpha - \tau(w) \tag{100}$$

Let $w_0 = \min(\tau^{-1}(1 - 2\alpha - \gamma), \epsilon_\infty)$, then $\sup_{f\in\mathcal{F}} \mathbb{P}\{W > w_0\} \geq \gamma$.

Part III. The argument here is based on an argument in Baraud (2004). Define a rejection region

$$\mathcal{R} = \{W > w\} \cup \{||\widehat{f} - \Pi\widehat{f}||_2 > W\}. \tag{101}$$

Now, for any $f \in \mathcal{F}$, $f^\star = f$, $||\widehat{f} - \Pi\widehat{f}||_2 \leq ||\widehat{f} - f||_2$ and

$$\begin{aligned}
\mathbb{P}_f\{\mathcal{R}\} &\leq \mathbb{P}_f\{W > w\} + \mathbb{P}_f\Big\{||\widehat{f} - \Pi\widehat{f}||_2 > W\Big\} && (102)\\
&\leq \gamma + \mathbb{P}_f\Big\{||\widehat{f} - \Pi\widehat{f}||_2 > W\Big\} \leq \gamma + \mathbb{P}_f\Big\{||\widehat{f} - \Pi\widehat{f}||_2 > W\Big\} && (103)\\
&\leq \gamma + \mathbb{P}_f\Big\{||f - \widehat{f}||_2 > W\Big\} = \gamma + \mathbb{P}_f\Big\{||f^\star - \widehat{f}||_2 > W\Big\} && (104)\\
&\leq \gamma + \mathbb{P}_f\Big\{||f^\star - \widehat{f}||_\infty > W\Big\} \leq \gamma + \alpha && (105)
\end{aligned}$$

which bounds the type I error of $\mathcal{R}$.

Now let f be such that $||f - \Pi f||_2 \geq \max\{2w, \epsilon_2\}$. Hence, $||f - \Pi f||_2 \geq \epsilon_2$ so that $f^\star = f$. Then,

$$||\widehat{f} - \Pi\widehat{f}||_2 \geq ||f - \Pi\widehat{f}||_2 - ||f - \widehat{f}||_2 \geq 2w - ||f - \widehat{f}||_2. \tag{106}$$

Hence,

$$\begin{aligned}
\mathbb{P}_f\{\mathcal{R}^c\} &= \mathbb{P}_f\Big\{||\widehat{f} - \Pi\widehat{f}||_2 \leq W, W \leq w\Big\} \leq \mathbb{P}_f\Big\{||\widehat{f} - \Pi\widehat{f}||_2 \leq w, W \leq w\Big\} && (107)\\
&\leq \mathbb{P}_f\Big\{||f - \widehat{f}||_2 \geq w, w \geq W\Big\} \leq \mathbb{P}_f\Big\{||f - \widehat{f}||_2 \geq W\Big\} && (108)\\
&= \mathbb{P}_f\Big\{||f^\star - \widehat{f}||_2 \geq W\Big\} \leq \mathbb{P}_f\Big\{||f^\star - \widehat{f}||_\infty \geq W\Big\} && (109)\\
&\leq \alpha. && (110)
\end{aligned}$$

Thus, $\mathcal{R}$ defines a test for $H_0 : f \in \mathcal{F}$ with level $\alpha + \gamma$ whose power on the complement of the sphere of radius $\max\{2w, \epsilon_2\}$ is at least $1 - \alpha$. But, from Baraud (2004), this implies that

$$\max\{w, \epsilon_2/2\} \geq \kappa(\alpha, \gamma)(n - d)^{1/4} n^{-1/2}. \tag{111}$$

□

PROOF OF THEOREM 3.3. The volume is minimized by making ϵ_∞ as large as possible and ϵ_2 as small as possible. To achieve the lower bound on the width requires $\epsilon_\infty \leq w_{\mathcal{F}}$ and $\epsilon_2 \geq 2\kappa(\alpha, \gamma)(n - d)^{1/4} n^{-1/2}$. □

PROOF OF LEMMA 3.2. Q is the solution, with respect to c, to $\xi = 1 - F_{0,m}(r(c))$ where the function $r(c) = F^{-1}_{c\sqrt{m},m}(\beta))$ is monotonically increasing in c. Also, $F_{0,m}(r(0)) = \beta$ and $F_{0,m}(r(\infty)) = 1$ so a solution exists since $0 < \beta < 1 - \xi < 1$. Now we bound Q from above.

To upper bound Q it suffices to find c such that

$$F^{-1}_{c\sqrt{m},m}(\beta) \geq F^{-1}_{0,m}(1 - \xi). \tag{112}$$

From Birgé (2001) we have

$$\begin{aligned}
F^{-1}_{z,d}(u) &\leq z + d + 2\sqrt{(2z + d)\log(1/(1 - u))} + 2\log(1/(1 - u)) && (113)\\
F^{-1}_{z,d}(u) &\geq z + d - 2\sqrt{(2z + d)\log(1/u)}. && (114)
\end{aligned}$$

Hence,

$$\begin{aligned}
F^{-1}_{c\sqrt{m},m}(\beta) &\geq m + c\sqrt{m} - 2\sqrt{(2c\sqrt{m} + m)\log\frac{1}{\beta}} && (115)\\
F^{-1}_{0,m}(1 - \gamma) &\leq m + 2\sqrt{m\log\frac{1}{\gamma}} + 2\log\frac{1}{\gamma}. && (116)
\end{aligned}$$

It suffices to find c that satisfies

$$m + c\sqrt{m} - 2\sqrt{(2c\sqrt{m} + m)\log\frac{1}{\beta}} \geq m + 2\sqrt{m\log\frac{1}{\gamma}} + 2\log\frac{1}{\gamma}, \tag{117}$$

or equivalently,

$$c \geq 2\sqrt{\left(\frac{c}{\sqrt{m}} + 1\right)\log\frac{1}{\beta}} + 2\left(\sqrt{\log\frac{1}{\gamma}} + \log\frac{1}{\gamma}\right). \tag{118}$$

The right hand side of the last inequality is largest when $m = 1$, and equality can be achieved when $m = 1$ at some $\Lambda(\beta, \xi)$ for any β, ξ satisfying the stated conditions. Equality can be achieved then for any m at some $Q(m, \beta, \xi) \leq \Lambda(\beta, \xi)$. This proves the first claim. The second claim follows immediately by inspection. $\square$

PROOF OF THEOREM 3.4. Let $A = \left\{T \leq \chi^2_{\gamma,n-d}\right\}$. Then,

$$\mathbb{P}_f\{f^\star \notin B\} = \mathbb{P}_f\{f^\star \notin B, A\} + \mathbb{P}_f\{f^\star \notin B, A^c\}.$$

We claim that $\mathbb{P}_f\{f^\star \notin B, A\} \leq \alpha/2$ and $\mathbb{P}_f\{f^\star \notin B, A^c\} \leq \alpha/2$. There are four cases.

Case I. $f \in \mathcal{F}$. Then $f = f^\star$ and $\mathbb{P}_f\{f \notin B, A^c\} \leq \mathbb{P}_f\{A^c\} \leq \alpha/2$. $\mathbb{P}_f\{f \notin B, A\} \leq \mathbb{P}_f\{f \notin B\} = \mathbb{P}_{\Pi f}\{\Pi f \notin B\} \leq \mathbb{P}_{\Pi f}\Big\{||\widehat{f} - \Pi f|| > w_{\mathcal{F}}\Big\} \leq \alpha/2$.

Case II. $f \in V - \mathcal{F}$ where $V = \{f : \|f - \Pi f\| \leq \epsilon_2, \|f - \Pi f\|_\infty \leq \epsilon_\epsilon\}$. Again, $f = f^\star$. First, $\mathbb{P}_f\{f \notin B, A^c\} \leq \mathbb{P}_f\Big\{||Y - f||_\infty > z_{\alpha/2n}\Big\} \leq \alpha/2$. Next, we bound $\mathbb{P}_f\{f \notin B, A\}$. Note that $\widehat{f} = \Pi Y \sim N(g, \sigma^2\Pi\Pi^T)$, where $g = \Pi f$. Then $\widehat{f}_i \sim N(g_i, \Omega_i^2)$. Let $B_0 = (L + \epsilon_\infty, U - \epsilon_\infty)$. Then, $\Pi f \in B_0$ implies $f \in B$ and $\mathbb{P}_f\{\notin B, A\} \leq \mathbb{P}_f\{\Pi f \notin B_0\} \leq \alpha/2$.

Case III. $f \notin V$, $||f - \Pi f|| \leq \epsilon_2$ and $||f - \Pi f||_\infty > \epsilon_\infty$. In this case, $f^\star = \Pi f$. Then $\mathbb{P}_f\{f^\star, f \in B^c, A^c\} \leq \mathbb{P}_f\{f \in B^c, A^c\} \leq \alpha/2$. Also, $\mathbb{P}_f\{f^\star, f \in B^c, A\} \leq \mathbb{P}_f\{f^\star \notin B\} = \mathbb{P}_{\Pi f}\{\Pi f \notin B\} \leq \mathbb{P}_{\Pi f}\Big\{||\widehat{f} - \Pi f|| > w_{\mathcal{F}}\Big\} \leq \alpha/2$.

Case IV. $f \notin V$ and $||f - \Pi f|| > \epsilon_2$. In this case, $f^\star = f$. But

$$\mathbb{P}_f\{f \notin B, A\} \leq \mathbb{P}_f\{A\} \leq F_{f-\Pi f,n-d}(\chi^2_{\gamma,n-d}) \leq F_{\epsilon_2,n-d}(\chi^2_{\gamma,n-d}) \leq \alpha/2$$

and

$$\mathbb{P}_f\{f \notin B, A^c\} \leq \mathbb{P}_f\{f \notin B, A^c\} \leq \alpha/2.$$

Thus, $\mathbb{P}_f\{f^\star \notin B\} \leq \alpha$. Equation (55) follows since $\mathbb{P}_f\Big\{T \leq \chi^2_{\gamma,n-d}\Big\} \geq 1 - \gamma$ for all $f \in \mathcal{F}$. $\square$

PROOF OF THEOREM 3.5. Note that $\mathbb{P}_f\{f^\star \cap B = \emptyset\} = \sum_j \mathbb{P}_f\Big\{f^\star \cap B = \emptyset, \widehat{J} = j\Big\}$. We show that $\mathbb{P}_f\Big\{f^\star \cap B = \emptyset, \widehat{J} = j\Big\} \leq \alpha_j$ for each j. There are three cases.

Case I. $||f - \Pi_j f|| > \epsilon_{2,j}$. Then,

$$\begin{aligned}\mathbb{P}_f\Big\{ f^\star \cap B = \emptyset, \widehat{J} = j \Big\} &\leq \mathbb{P}_f\Big\{ \widehat{J} = j \Big\} \leq F_{f-\Pi_j f, n-d_j}(\chi^2_{\gamma,n-d_j}) \\ &\leq F_{\epsilon_{2,j},n-d_j}(\chi^2_{\gamma,n-d_j}) \\ &\leq \alpha_j\end{aligned}$$

due to (67).

Case II. $||f - \Pi_j f|| \leq \epsilon_{2,j}$ and $||f - \Pi_j f||_\infty \leq \epsilon_{\infty,j}$. So,

$$\begin{aligned}\mathbb{P}_f\Big\{ f^\star \cap B = \emptyset, \widehat{J} = j \Big\} &\leq \mathbb{P}_f\Big\{ f \notin B, \widehat{J} = j \Big\} \\ &\leq \mathbb{P}_f\Big\{ ||f - \widehat{f}||_\infty > w_{\mathcal{F}_j} + \epsilon_{\infty,j} \Big\} \\ &\leq \mathbb{P}_f\big\{ ||f - \Pi_j f||_\infty + ||\Pi_j f - \Pi_j Y||_\infty > w_{\mathcal{F}_j} + \epsilon_{\infty,j} \big\} \\ &\leq \mathbb{P}_f\big\{ ||\Pi_j f - \Pi_j Y||_\infty > w_{\mathcal{F}_j} \big\} \\ &= \mathbb{P}_{\Pi_j f}\big\{ ||\Pi_j f - \Pi_j Y||_\infty > w_{\mathcal{F}_j} \big\} \\ &\leq \alpha_j.\end{aligned}$$

Case III. $||f - \Pi_j f|| \leq \epsilon_{2,j}$ and $||f - \Pi_j f||_\infty > \epsilon_{\infty,j}$. Now,

$$\begin{aligned}\mathbb{P}_f\Big\{ f^\star \cap B = \emptyset, \widehat{J} = j \Big\} &\leq \mathbb{P}_f\Big\{ \Pi_j f \notin B, \widehat{J} = j \Big\} \\ &\leq \mathbb{P}_{\Pi_j f}\Big\{ ||\widehat{f} - \Pi_j f|| > w_{\mathcal{F}_j} \Big\} \\ &\leq \alpha_j.\end{aligned}$$

To prove (69), suppose that $f \in \mathcal{F}_j$. Then, $\mathbb{P}_f\Big\{ \widehat{J} > j \Big\} \leq \gamma$. But, as long as $\widehat{J} \leq j$, $W = w_{\widehat{J}}(\alpha_{\widehat{J}}) + \epsilon_{\infty,\widehat{J}} \leq w_{|}(\alpha_j) + \epsilon_{\infty,j}$. The last statement follows since, when $\epsilon_{2,j} \geq Q(n-d_j, \alpha/2, \gamma)(n-d_j)^{1/4} n^{-1/2}$ □

PROOF OF LEMMA 3.9. First note that if B is a ball in $\mathbb{R}^n$ in any norm, then $B - B = 2B$. Second, we have that

$$\begin{aligned}\omega(u) &= \sup\{|Tg| : \ \|g\|_2 \leq u, \ g \in V - V\} &(119)\\ &= \sup\{|Tg| : \ \|g\|_2 \leq u, \ g \in V(2\epsilon_2, 2\epsilon_\infty)\}. &(120)\end{aligned}$$

To see the latter equality, note that if $g, h \in V$, then we can write $g - h = f + \delta_1 - \delta_2$ where $f \in \mathcal{F}$ and δ_i are in $B_k^\perp(0, \epsilon_k)$ for $k = 2, \infty$. Thus, $\delta_1 - \delta_2$ is in $2B_2^\perp(0, \epsilon_2) \cap 2B_\infty^\perp(0, \epsilon_\infty)$.

Set $B^*(f) = B_2^\perp(f, 2\epsilon_2) \cap B_\infty^\perp(f, 2\epsilon_\infty)$. We have that

$$\begin{aligned}\omega(\eta, \mathcal{F}) &= \sup\{f_1 : \ \|f\|_2 \leq \eta, f \in \mathcal{F}\} &(121)\\ \omega(\eta, B^*(0)) &= \sup\{f_1 : \ \|f\|_2 \leq \eta, f \in B^*(0)\}. &(122)\end{aligned}$$

For any $g \in V(2\epsilon_2, 2\epsilon_\infty)$, we can write $g = g_1 + g_2$ where $g_1 \in \mathcal{F}$ and $g_2 \in B^*(0)$ and the two functions are orthogonal. Then,

$$\begin{aligned} w(u,V) &= \sup\left\{T(g):\ g \in V(2\epsilon_2, 2\epsilon_\infty),\ \|g\|_2 \le u\right\} && (123)\\ &= \sup_{0\le c\le u}\left\{T(g_1+g_2):\ \|g_1\|_2 \le \sqrt{u^2-c^2}, \|g_2\|_2 \le c^2,\ g_1\in\mathcal{F}, g_2 \in B^*(0)\right\} && (124)\\ &\le \sup_{0\le c\le u}\left[\sup_{\substack{g_1\in\mathcal{F}\\ \|g_1\|_2\le\sqrt{u^2-c^2}}} T(g_1) + \sup_{\substack{g_2\in B^*(0)\\ \|g_2\|_2\le c}} T(g_2)\right] && (125)\\ &= \sup_{0\le c\le u}\left[\omega(\sqrt{u^2-c^2},\mathcal{F}) + \omega(c, B^*(0))\right]. && (126)\end{aligned}$$

Moreover, equality can be attained for each c by choosing g_1 and g_2 to be the maximizers (or suitably close approximants thereof) of each term in the last equation. Consequently,

$$\omega(u) = \sup_{0\le c\le u} \omega(\sqrt{u^2-c^2},\mathcal{F}) + \omega(c, B^*(0)). \quad (127)$$

To derive $\omega(\eta, B^*(0))$, note that $f = ((\eta\wedge\epsilon_2)\sqrt{n}\wedge\epsilon_\infty, 0, 0, \ldots, 0)$ maximizes f_1 subject to the norm constraint. Hence, $\omega(\eta, B^*(0)) = \min((\eta\wedge\epsilon_2)\sqrt{n}, \epsilon_\infty)$. For $\omega(\eta,\mathcal{F})$, let $e = (1, 0, \ldots, 0) \in \mathbb{R}^n$. Recall that $\Omega_\mathcal{F} = \frac{\langle e, \Pi_\mathcal{F} e\rangle}{\|e\|\,\|\Pi_\mathcal{F} e\|} = \frac{\|\Pi_\mathcal{F} e\|}{\|e\|}$, which is between 0 and 1. Maximizing $e^T f$ for $f \in \mathcal{F}$ and $\|f\|_2 \le \eta$ is equivalent to maximizing $n\langle e, f\rangle = n\langle \Pi_\mathcal{F} e, f\rangle$. The maximum subject to the constraint occurs at $f^\star = \eta\Pi e/\|\Pi e\|$ Hence, $\omega(\eta,\mathcal{F}) = \eta\sqrt{n}\Omega_\mathcal{F}$. Note that η is in terms of the normalized two norm; in the "natural" (root sum of squares) norm, the modulus would be $\omega_\natural(u,\mathcal{F}) = u\Omega_\mathcal{F}$.

It follows that

$$\begin{aligned} \omega(u,V) &= \sup_{0\le c\le u}[\omega(\sqrt{u^2-c^2},\mathcal{F}) + \omega(c, B^*(0))] && (128)\\ &= \sup_{0\le c\le u}\left[\sqrt{n}\Omega_\mathcal{F}\sqrt{u^2-c^2} + \min((c\wedge\epsilon_2)\sqrt{n}, \epsilon_\infty)\right] && (129)\\ &= \sqrt{n}\sup_{0\le c\le u}\left[\Omega_\mathcal{F}\sqrt{u^2-c^2} + \min(c, \epsilon_2\wedge(\epsilon_\infty/\sqrt{n}))\right] && (130)\\ &= \sqrt{n}\left(u\Omega\sqrt{\frac{\Omega^2}{1+\Omega^2}} + \min(\frac{u}{\sqrt{1+\Omega^2}}, \epsilon_2\wedge(\epsilon_\infty/\sqrt{n}))\right) && (131)\\ &= \left(u\sqrt{n}\Omega\sqrt{\frac{\Omega^2}{1+\Omega^2}} + \min\left(\frac{u\sqrt{n}}{\sqrt{1+\Omega^2}},\ \epsilon_2\sqrt{n},\ \epsilon_\infty\right)\right) && (132)\end{aligned}$$

because the supremum over c is maximized at $c = u/(1+\Omega^2)$. In the natural two norm, we have

$$\omega_\natural(u,V) = \left(u\Omega\sqrt{\frac{\Omega^2}{1+\Omega^2}} + \min\left(\frac{u}{\Omega}\sqrt{\frac{\Omega^2}{1+\Omega^2}},\ \epsilon_{2,\natural},\ \epsilon_\infty\right)\right). \quad (133)$$

□

5 Discussion

We have shown that adaptive confidence bands for f are possible if coverage is replaced by surrogate coverage. Of course, there are many other ways one could define a surrogate. Here, we briefly outline a few possibilities.

Wavelet expansions of the form

$$f(x) = \sum_j \alpha_j \phi_j(x) + \sum_j \sum_k \beta_{jk} \psi_{jk}$$

lend themselves quite naturally to the surrogate approach. For example, one can define

$$f^\star(x) = \sum_j \alpha_j \phi_j(x) + \sum_j \sum_k s(\beta_{jk}) \psi_{jk}$$

where $s(x) = \text{sign}(x)(|x| - \lambda)_+$ is the usual soft-thresholding function.

For kernel smoothers and local polynomial smoothers $\widehat{f}_h$ that depends on a bandwidth h, a possible surrogate is $f^\star = \mathsf{E}(\widehat{f}_{h^\star})$ where $h^\star$ is the largest bandwidth h for which $\widehat{f}_h$ passes a goodness of fit test with high probability. In the spirit of Davies and Kovac (2001), one could take the test to be a test for randomness applied to the residuals.

Motivated by ideas in Donoho (1988) we can define another surrogate as follows. Let us switch to the problem of density estimation. Let $X_1, \ldots, X_n \sim F$ for some distribution F. The goal is define an appropriate surrogate band for the density f. Define the smoothness functional $S(F) = \int (f''(x))^2 dx$. To make sure that $S(F)$ is well defined for all F we borrow an idea from Donoho (1988). Let Φ_h denote a Gaussian with standard deviation h and define $S(F) = \lim_{h \to 0} S(F \oplus \Phi_h)$ where $\oplus$ denote convolution. Donoho shows that S is then a well-defined, convex, lower semicontinuous functional.

Let $\widehat{F}_n$ be the empirical distribution function and let $B = B(\widehat{F}, \epsilon_n) = \{F : \ ||F - \widehat{F}_n|| \leq \epsilon_n\}$ where $||\cdot||$ is the Kolmogorov-Smnirnov distance and ϵ_n is the $1 - \beta$ quantile of $||U - U_n||$ where U is the uniform distribution and U_n is the empirical from a sample from U. Thus, B is a nonparametric, $1 - \beta$ confidence ball for F. The simplest $F \in B$ is the distribution that minimize $S(F)$ subject to $F \in B$. We define the surrogate $F^\star$ to be the distribution that minimizes $S(F)$ subject to F belonging to B_F, where B_F is a population version of B. We might then think of $F^\star$ as the simplest distribution that is not empirically dinstinguishable from F. A natural definition of B_F might be $B_F = \{G : \ ||F - G|| \leq \epsilon_n\}$. But this definition only makes sense for fixed radius confidence sets. Another definition is $B_F = \{G : \ \mathbb{P}_F\{G \in B\} \geq 1/2\}$.

To summarize, we define

$$F^\star = \text{argmin}_{F \in B_F} S(F) \tag{134}$$

where

$$B_F = \left\{ G : \ \mathbb{P}_F\Big\{ G \in B(\widehat{F}_n, \epsilon_n) \Big\} \geq 1/2 \right\} \tag{135}$$

and $B(\widehat{F}_n, \epsilon_n) = \{G : \ ||\widehat{F}_n - G|| \leq \epsilon_n\}$. Let

$$\Gamma = \cup \{G^\star : \ G \in B(\widehat{F}_n, \epsilon_n)\}. \tag{136}$$

Then

$$\ell(x) = \inf_{F \in \Gamma} F'(x), \quad u(x) = \sup_{F \in \Gamma} F'(x) \tag{137}$$

defines a valid confidence band for the density of $F^\star$.

Let us also mention average coverage (Wahba 1983; Cummins, Filloon, Nychka 2001). Bands (L, U) have average coverage if $\mathbb{P}_f\{L(\xi) \leq f(\xi) \leq U(\xi)\} \geq 1 - \alpha$ where $\xi \sim \text{Uniform}(0, 1)$. A way to combine average with the surrogate idea is to enforce something stronger than average coverage such as

$$\mathbb{P}_f\Big\{ L(\xi) \leq f(\xi) \leq U(\xi) \text{ and } \widehat{f} \preceq f \Big\} \geq 1 - \alpha$$

where $\widehat{f} = (L + U)/2$ and $\widehat{f} \preceq f$ means that $\widehat{f}$ is simpler than f according to a partial order $\preceq$, for example, $f \preceq g$ if $\int (f'')^2 \leq \int (g'')^2$.

References

Baraud, Y. (2004). Confidence balls in Gaussian regression, *The Annals of Statistics*, 32, 528–551.

Beran, Rudolf and Dümbgen, Lutz. (1998). Modulation of estimators and confidence sets. *The Annals of Statistics*, 26, 1826–1856.

Bickel, P.J. and Ritov, Y. (2000). Non-and semi parametric statistics: compared and contrasted. *J. Statist. Plann. Inference*, 91,

Birgé, L. (2001). An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics.* (M. de Gunst, C. Klaassen and A. van der Vaart, eds.) 113–133, IMS, Beachwood, OH.

Cai, T. and Low, M. (2005). Adaptive Confidence Balls. *The Annals of Statistics*, 34, 202–228.

Cai, T. and Low, Mark, G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.*, 32, 1805–1840.

Chaudhuri, Probal and Marron, J. S. (1999). SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association*, 94, 807–823.

Chaudhuri, Probal and Marron, J. S. (2000). Scale space view of curve estimation. *The Annals of Statistics*, 28, 408–428.

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

Cummins D., Filloon T., Nychka D. (2001). Confidence Intervals for Nonparametric Curve Estimates: Toward More Uniform Pointwise Coverage *Journal of the American Statistical Association*, 96, 233–246.

Donoho, D. (1988). One-Sided Inference about Functionals of a Density. *Annals of Statistics*, 16, 1390–1420.

Donoho, D. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41, 613–627.

Donoho, D. and Liu, R. (1991). Geometrizing Rates of Convergence, II. *The Annals of Statistics*, 19, 633–667.

Eubank, R.L. and Speckman, P.L. (1993). Confidence Bands in Nonparametric Regression. *Journal of the American Statistical Association*, 88, 1287–1301.

Genovese, C. and Wasserman, L. (2005). Nonparametric confidence sets for wavelet regression. *Annals of Statistics*, 33, 698–729.

Hall, P. and Titterington, M. (1988). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis*, 27, 228–254.

Härdle, Wolfgang and Bowman, Adrian W. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of the American Statistical Association*, 83, 102–110.

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

Ingster, Y. and Suslina, I. (2003). *Nonparametric Goodness of Fit Testing Under Gaussian Models.* Springer. New York.

Juditsky, A. and Lambert-Lacroix, S. (2003). Nonparametric confidence set estimation. *Mathematical Methods of Statistics*, 19, 410-428.

Leeb, H. and Pötscher, B.M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21, 21–59.

Li, Ker-Chau. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17, 1001–1008.

Low, Mark G. (1997). On nonparametric confidence intervals. *The Annals of Statistics*, 25, 2547–2554.

Neumann, Michael H. and Polzehl, Jörg. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9, 307–333.

Robins, J. and van der Vaart, Aad. (2006). Adaptive Nonparametric Confidence Sets. *The Annals of Statistics*, 34, 229–253.

Ruppert, D. and Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*, Cambridge University Press. Cambridge.

Terrell, G.R. and Scott, D.W. (1985). Oversmoothed Nonparametric Density Estimates. *Journal of the American Statistical Association*, 80, 209–214.

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

Terrell, G.R. (1990). The Maximal Smoothing Principle in Density Estimation. *Journal of the American Statistical Association*, 85, 470–477.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B, Methodological*, 45, 133–150.

Xia, Y. (1998). Bias-Corrected Confidence Bands in Nonparametric Regression. *Journal of the Royal Statistical Society. Series B*, 60, 797–811.