cemmap
centre for microdata methods and practice

A simple bootstrap method for constructing nonparametric confidence bands for functions

Peter Hall
Joel Horowitz

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP29/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Submitted to the Annals of Statistics

A SIMPLE BOOTSTRAP METHOD FOR CONSTRUCTING NONPARAMETRIC CONFIDENCE BANDS FOR FUNCTIONS

By Peter Hall[*,†,§], and Joel Horowitz[‡,¶]

University of Melbourne[*] and University of California, Davis[†]
and Northwestern University[‡]

Standard approaches to constructing nonparametric confidence bands for functions are frustrated by the impact of bias, which generally is not estimated consistently when using the bootstrap and conventionally smoothed function estimators. To overcome this problem it is common practice to either undersmooth, so as to reduce the impact of bias, or oversmooth, and thereby introduce an explicit or implicit bias estimator. However, these approaches, and others based on nonstandard smoothing methods, complicate the process of inference, for example by requiring the choice of new, unconventional smoothing parameters and, in the case of undersmoothing, producing relatively wide bands. In this paper we suggest a new approach, which exploits to our advantage one of the difficulties that, in the past, has prevented an attractive solution to the problem—the fact that the standard bootstrap bias estimator suffers from relatively high-frequency stochastic error. The high frequency, together with a technique based on quantiles, can be exploited to dampen down the stochastic error term, leading to relatively narrow, simple-to-construct confidence bands.

1. Introduction.

1.1. *Motivation.* There is an extensive literature, summarised in Section 1.4 below, on constructing nonparametric confidence bands for functions. However, this work generally does not suggest practical solutions to the critical problem of choosing tuning parameters, for example smoothing parameters or the nominal coverage level of the confidence band, to ensure a high degree of coverage accuracy or to produce bands that err on the side of conservatism. In this paper we suggest new, simple bootstrap methods for constructing confidence bands using conventional smoothing parameter choices.

In particular, our approach does not require a nonstandard smoothing parameter. The basic algorithm requires only a single application of the bootstrap, although a more refined, double bootstrap technique is also suggested. The greater part of our attention is directed to regression problems, but we also discuss the application of our methods to constructing confidence bands for density functions.

The resulting confidence regions depend on choice of two parameters α and ξ, in the range $0 < \alpha, \xi < 1$, and the methodology results in confidence bands that, asymptotically, cover the regression mean at x with probability at least $1-\alpha$, for at least a proportion $1-\xi$ of values of x. In particular, the bands are pointwise, rather than simultaneous. Pointwise bands are more popular with practitioners, and are the subject of a substantial majority of research on nonparametric confidence bands for functions.

[§]Research supported by ARC and NSF grants.
[¶]Research supported by NSF grant SES-0817552.
Keywords and phrases: Bandwidth, bias, bootstrap, confidence interval, conservative coverage, coverage error, kernel methods, statistical smoothing.

1.2. *Features of our approach, and competing methods.* The "exceptional" $100\,\xi\%$ of points that are not covered are typically close to the locations of peaks and troughs, and so are discernible from a simple estimate of the regression mean. Their location can also be determined using a theoretical analysis—points near peaks and troughs potentially cause difficulties because of bias. See Section 2.6 for theoretical details, and Section 3 for numerical examples.

Our approach accommodates bias by increasing the width of confidence bands. However, the amount by which we increase width is no greater than a constant factor, rather than the polynomial amount (as a function of n) associated with most suggestions for undersmoothing.

Methods based on either under- or oversmoothing are recommended often in the literature. However, there are no empirical techniques, where the data determine the amount of smoothing, that are used even moderately widely in either case. In particular, although theoretical arguments demonstrate clearly the advantages of under- or oversmoothing if appropriate smoothing parameters are chosen, there are no attractive, effective empirical ways of selecting those quantities. Indeed, it is not uncommon to suggest that the issue be avoided altogether, by ignoring the effects of bias. For example, this approach is recommended in textbooks; see Ruppert et al. (2003, p. 133ff), who refer to the resulting bands as "variability bands," and Efron and Tibshirani (1993, pp. 79–80), who suggest plotting many realisations of bootstrapped curve estimators without bias corrections.

In addition to needing unavailable bandwidth choice methods, the drawbacks of undersmoothing include the fact that the confidence bands become both wider and more wiggly as the amount of undersmoothing increases. The increase in wiggliness is so great that, unless sample size is very large, the coverage accuracy does not necessarily improve as the amount of undersmoothing increases. Details are given in Section 3.

Wiggliness can likewise be a problem for bands that result from using oversmoothing to remove bias explicitly. Here the relatively high level of variability from which function derivative estimators suffer means that the confidence bands may again oscillate significantly, and can be difficult to interpret. These results, and those reported in the previous paragraph, are for optimal choices of the amount of under- or oversmoothing. In practice the amount has to be chosen empirically, and that introduces additional noise, which further reduces performance.

1.3. *Intuition.* Our methodology exploits, to our advantage, a difficulty that in the past has hindered a simple solution to the confidence band problem. To explain how, we note first that if nonparametric function estimators are constructed in a conventional manner then their bias is of the same order as their error about the mean, and accommodating the bias has been a major obstacle to achieving good coverage accuracy. Various methods, based on conventional smoothing parameters, can be used to estimate the bias and reduce its impact, but the bias estimators fail to be consistent, not least because the stochastic noise from which they suffer is highly erratic. (In the case of kernel methods, the frequency of the noise is proportional to the inverse of the bandwidth.) However, as we show in this paper, this erratic behaviour is actually advantageous, since if we average over it then we can largely eliminate the negative impact that it has on the bias estimation problem. We do the averaging implicitly, not by computing means but by working with quantiles of the "distribution" of coverage.

1.4. *Literature review.* We shall summarise previous work largely in terms of whether it involved undersmoothing or oversmoothing; the technique suggested in the present paper is almost unique in that it requires neither of these approaches. Härdle and Bowman (1988), Härdle and Marron (1991), Hall (1992a), Eubank and Speckman (1993), Sun and Loader (1994), Härdle et al. (1995) and Xia (1998) suggested methods based on oversmoothing, using either implicit or explicit bias correction. Hall and Titterington (1988) also used explicit bias correction, in the sense that their

bands required a known bound on an appropriate derivative of the target function. Bjerve et al. (1985), Hall (1992b), Hall and Owen (1993), Neumann (1995), Chen (1996), Neumann and Polzehl (1998), Picard and Tribouley (2000), Chen et al. (2003) (in the context of hypothesis testing), Claeskens and Van Keilegom (2003), Härdle et al. (2004) and McMurry and Politis (2008) employed methods that involve undersmoothing. There is also a theoretical literature which addresses the bias issue through consideration of the technical function class from which a regression mean or density came; see e.g. Low (1997) and Genovese and Wasserman (2008). This work sometimes involves confidence balls, rather than bands, and in that respect is connected to research such as that of Eubank and Wang (1994) and Genovese and Wasserman (2005). Wang and Wahba (1995) considered spline and Bayesian methods. The notion of "honest" confidence bands, which have guaranteed coverage for a rich class of functions, was pioneered by Li (1989). Recent contributions include those of Cai and Low (2006), Giné and Nickl (2010) and Hoffmann and Nickl (2011).

2. Methodology.

2.1. *Model.* Suppose we observe data pairs in a sample $\mathcal{Z} = \{(X_i, Y_i),\ 1 \le i \le n\}$, generated by the model

$$Y_i = g(X_i) + \epsilon_i\,, \tag{2.1}$$

where the experimental errors ϵ_i are independent and identically distributed with finite variance and zero mean conditional on X. Our aim is to construct a pointwise confidence band for the true g in a closed, bounded region $\mathcal{R}$. A more elaborate, heteroscedastic model will be discussed in Section 2.4; we omit it here only for the sake of simplicity. We interpret $g(x)$ in the conventional regression manner, as $E(Y \,|\, X = x)$, but our theoretical analysis takes account of the fact that although we condition on the X_is at this point we consider that they originated as random variables, with density f_X.

2.2. *Properties of function estimators and conventional confidence bands.* Let $\hat g$ denote a conventional estimator of g. We assume that $\hat g$ incorporates smoothing parameters computed empirically from the data, using for example cross-validation or a plug-in rule, and that the variance of $\hat g$ can be estimated consistently by $s(\mathcal{X})^2\,\hat\sigma^2$, where $s(\mathcal{X})$ is a known function of the set of design points $\mathcal{X} = \{X_1, \ldots, X_n\}$ and the smoothing parameters, and $\hat\sigma^2$ is an estimator of the variance, σ^2, of the experimental errors ϵ_i, computed from the dataset $\mathcal{Z}$. The case of heteroscedasticity is readily accommodated too; see Section 2.4. We write $\hat g^*$ for the version of $\hat g$ computed using a conventional bootstrap argument. For details of the construction of $\hat g^*$, see step 4 of the algorithm in Section 2.3.

The smoothing parameters used for $\hat g$ would generally be chosen to optimise a measure of accuracy, for example in a weighted L_p metric where $1 \le p < \infty$, and we shall make this assumption implicitly in the discussion below. In particular, it implies that the asymptotic effect of bias, for example as represented by the term $b(x)$ in (2.4) below, is finite and typically nonzero.

An asymptotic, symmetric confidence band for g, constructed naively without considering bias, and with nominal coverage $1 - \alpha$, has the form:

$$\mathcal{B}(\alpha) = \Big\{(x, y) : x \in \mathcal{R}\,,\ \hat g(x) - s(\mathcal{X})(x)\,\hat\sigma\, z_{1-(\alpha/2)} \le y \le \hat g(x) + s(\mathcal{X})(x)\,\hat\sigma\, z_{1-(\alpha/2)}\Big\}\,, \tag{2.2}$$

where $z_\beta = \Phi^{-1}(\beta)$ is the β-level critical point of the standard normal distribution, and Φ is the standard normal distribution function. Unfortunately, the coverage of $\mathcal{B}(\alpha)$ at a point x, given by

$$\pi(x, \alpha) = P\{(x, g(x)) \in \mathcal{B}(\alpha)\}\,, \tag{2.3}$$

is usually incorrect even in an asymptotic sense, and in fact the band typically undercovers, often seriously, in the limit as $n \to \infty$. The reason is that the bias of $\hat{g}$, as an estimator of g, is of the same size as the estimator's stochastic error, and the confidence band allows only for the latter type of error. As a result the limit, as $n \to \infty$, of the coverage of the band is given by

$$\pi_{\lim}(x, \alpha) = \lim_{n \to \infty} \pi(x, \alpha) = \Phi\{z + b(x)\} - \Phi\{-z + b(x)\}\,, \tag{2.4}$$

where $z = z_{1-(\alpha/2)}$ and $b(x)$ describes the asymptotic effect that bias has on coverage. (A formula for $b(x)$ in a general multivariate setting is given in (4.7), and a formula in the univariate case is provided in Section 2.6.) The right-hand side of (2.4) equals $\Phi(z) - \Phi(-z) = 1 - \alpha$ if and only if $b(x) = 0$. For all other values of $b(x)$, $\pi_{\lim}(x, \alpha) < 1 - \alpha$. This explains why the band at (2.2) almost always undercovers unless some sort of bias correction is used.

The band potentially can be recalibrated, using the bootstrap, to correct for coverage errors caused by bias, but now another issue causes difficulty: the standard bootstrap estimator of bias, $E\{\hat{g}^*(x) \,|\, \mathcal{Z}\} - \hat{g}(x)$, is inconsistent, in the sense that the ratio of the estimated bias to its true value does not converge to 1 in probability as $n \to \infty$. This time the problem is caused by the stochastic error of the bias estimator; it is of the same size as the bias itself. The problem can be addressed using an appropriately oversmoothed version of $\hat{g}$ when estimating bias, either explicitly or implicitly, but the degree of oversmoothing has to be determined from the data, and in practice this issue is awkward to resolve. Alternatively, the estimator $\hat{g}$ can be undersmoothed, so that the influence of bias is reduced, but now the amount of undersmoothing has to be determined, and that too is difficult. Moreover, confidence bands computed from an appropriately undersmoothed $\hat{g}$ are an order of magnitude wider than those at (2.2), and so the undersmoothing approach, although more popular than oversmoothing, is unattractive for at least two reasons.

A simpler bootstrap technique, described in detail in the next section, overcomes these problems.

2.3. *The algorithm.*

Step 1. Estimators of g and σ^2. Construct a conventional nonparametric estimator $\hat{g}$ of g. Use a standard empirical method (for example, cross-validation or a plug-in rule), designed to minimise mean L_p error for some p in the range $1 \le p < \infty$, to choose the smoothing parameters on which $\hat{g}$ depends. For example, if the design is univariate then a local linear estimator of $g(x)$ is given by

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^{n} A_i(x)\, Y_i\,, \tag{2.5}$$

where

$$A_i(x) = \frac{S_2(x) - \{(x - X_i)/h\}\, S_1(x)}{S_0(x)\, S_2(x) - S_1(x)^2}\, K_i(x)\,, \tag{2.6}$$

$S_k(x) = n^{-1} \sum_i \{(x - X_i)/h\}^k\, K_i(x)$, $K_i(x) = h^{-1}\, K\{(x - X_i)/h\}$, K is a kernel function and h is a bandwidth.

There is an extensive literature on computing estimators $\hat{\sigma}^2$ of the error variance $\sigma^2 = \mathrm{var}(\epsilon)$; see, for example, Rice (1984), Buckley et al. (1988), Gasser et al. (1986), Müller and Stadtmüller (1987, 1992), Hall et al. (1990), Hall and Marron (1990), Seifert et al. (1993), Neumann (1994), Müller and Zhao (1995), Dette et al. (1998), Fan and Yao (1998), Müller et al. (2003), Munk et al. (2005), Tong and Wang (2005), Brown and Levine (2007), Cai et al. (2009) and Mendez and Lohr (2011). It includes residual-based estimators, which we introduce at (2.8) below, and methods

based on differences and generalised differences. An example of the latter approach, in the case of univariate design, is the following estimator due to Rice (1984):

$$\hat\sigma^2 = \frac{1}{2\,(n-1)} \sum_{i=2}^{n} (Y_{[i]} - Y_{[i-1]})^2\,, \tag{2.7}$$

where $Y_{[i]}$ is the concomitant of $X_{(i)}$ and $X_{(1)} \le \ldots \le X_{(n)}$ is the sequence of order statistics derived from the design variables.

As in Section 2.2, let $s(\mathcal{X})(x)^2\,\hat\sigma^2$ denote an estimator of the variance of $\hat g(x)$, where $s(\mathcal{X})(x)$ depends on the data only through the design points, and $\hat\sigma^2$ estimates error variance, for example being defined as at (2.7) or (2.8). In the local linear example, introduced at (2.5) and (2.6), we take $s(\mathcal{X})(x)^2 = \kappa/\{nh\,\hat f_X(x)\}$, where $\kappa = \int K^2$ and $\hat f_X(x) = (nh_1)^{-1} \sum_{1\le i\le n} K_1\{(x - X_i)/h_1\}$ is a standard kernel density estimator, potentially constructed using a bandwidth h_1 and kernel K_1 different from those used for $\hat g$. There are many effective, empirical ways of choosing h_1, and any of those can be used.

Step 2. Computing residuals. Using the estimator $\hat g$ from step (1), calculate initial residuals $\tilde\epsilon_i = Y_i - \hat g(X_i)$, put $\bar\epsilon = n^{-1} \sum_i \tilde\epsilon_i$, and define the centred residuals by $\hat\epsilon_i = \tilde\epsilon_i - \bar\epsilon$.

A conventional, residual-based estimator of σ^2, alternative to the estimator at (2.7), is

$$\hat\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \hat\epsilon_i^2\,. \tag{2.8}$$

The estimator at (2.7) is root-n consistent for σ^2, whereas the estimator at (2.8) converges at a slower rate unless an undersmoothed estimator of $\hat g$ is used when computing the residuals. This issue is immaterial to the theory in Section 4, although it tends to make the estimator at (2.7) a little more attractive.

Step 3. Computing bootstrap resample. Construct a resample $\mathcal{Z}^* = \{(X_i, Y_i^*),\, 1 \le i \le n\}$, where $Y_i^* = \hat g(X_i) + \epsilon_i^*$ and the ϵ_i^*s are obtained by sampling from $\hat\epsilon_1, \ldots, \hat\epsilon_n$ randomly, with replacement, conditional on $\mathcal{X}$. Note that, since regression is conventionally undertaken conditional on the design sequence, then the X_is are not resampled, only the Y_is.

Step 4. Bootstrap versions of $\hat g$, $\hat\sigma^2$ and $\mathcal{B}(\alpha)$. From the resample drawn in step 3, but using the same smoothing parameter employed to construct $\hat g$, compute the bootstrap version $\hat g^*$ of $\hat g$. (See Section 2.4 for discussion of the smoothing parameter issue.) Let $\hat\sigma^{*2}$ denote the bootstrap version of $\hat\sigma^2$, obtained when the latter is computed from $\mathcal{Z}^*$ rather than $\mathcal{Z}$, and construct the bootstrap version of $\mathcal{B}(\alpha)$, at (2.2):

$$\mathcal{B}^*(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat g^*(x) - s(\mathcal{X})(x)\,\hat\sigma^*\,z_{1-(\alpha/2)} \le y \le \hat g^*(x) + s(\mathcal{X})(x)\,\hat\sigma^*\,z_{1-(\alpha/2)}\Big\}\,. \tag{2.9}$$

Note that $s(\mathcal{X})$ is exactly the same as in (2.2); again this is a consequence of the fact that we are conducting inference conditional on the design points.

If, as in the illustration in step 1, the design is univariate and local linear estimators are employed, then $\hat g^*(x) = n^{-1} \sum_{1\le i\le n} A_i(x)\,Y_i^*$ where $A_i(x)$ is as at (2.6). The bootstrap analogue of the variance formula (2.7) is $\hat\sigma^{*2} = \{2\,(n-1)\}^{-1} \sum_{2\le i\le n} (Y_{[i]}^* - Y_{[i-1]}^*)^2$, where, if the ith largest order statistic $X_{(i)}$ equals X_j, then $Y_{[i]}^* = \hat g(X_j) + \epsilon_j^*$.

Step 5. Estimator of coverage error. The bootstrap estimator $\hat{\pi}(x, \alpha)$ of the probability $\pi(x, \alpha)$ that $\mathcal{B}(\alpha)$ covers $(x, g(x))$ is defined by:

$$\hat{\pi}(x, \alpha) = P\{(x, \hat{g}(x)) \in \mathcal{B}^*(\alpha) \mid \mathcal{X}\}, \tag{2.10}$$

and is computed, by Monte Carlo simulation, in the form

$$\frac{1}{B} \sum_{b=1}^{B} I\{(x, \hat{g}(x)) \in \mathcal{B}_b^*(\alpha)\}, \tag{2.11}$$

where $I(\mathcal{E})$ denotes the indicator function of an event $\mathcal{E}$, and $\mathcal{B}_b^*(\alpha)$ is the bth out of B bootstrap replicates of $\mathcal{B}^*(\alpha)$, where the latter is as at (2.9). The estimator at (2.10) is completely conventional, and in particular, no additional or nonstandard smoothing is needed.

Step 6. Constructing final confidence band. Define $\hat{\beta}(x, \alpha_0)$ to be the solution, in α, of $\hat{\pi}(x, \alpha) = 1 - \alpha_0$, and let $\hat{\alpha}_\xi(\alpha_0)$ denote the ξ-level quantile of points in the set $\{\hat{\beta}(x, \alpha_0) : x \in \mathcal{R}\}$. Specifically:

> take $\mathcal{R}$ to be a subset of $\mathbb{R}^r$, superimpose on $\mathcal{R}$ a regular, r-dimensional, rectangular grid with edge width δ, let $x_1, \ldots, x_N \in \mathcal{R}$ be the grid centres, let $\hat{\alpha}_\xi(\alpha_0, \delta)$ denote the ξ-level empirical quantile of the points $\hat{\alpha}(x_1, \alpha_0), \ldots, \hat{\alpha}(x_N, \alpha_0)$, and, for $\xi \in (0, 1)$, let $\hat{\alpha}_\xi(\alpha_0)$ denote the limit infimum, as $\delta \to 0$, of the sequence $\hat{\alpha}_\xi(\alpha_0, \delta)$. (2.12)

(We use the limit infimum to avoid ambiguity, although under mild conditions the limit exists.) For a value $\xi \in (0, \frac{1}{2}]$, construct the band $\mathcal{B}\{\hat{\alpha}_\xi(\alpha_0)\}$. In practice we have found that taking $1 - \xi = 0.9$ generally gives a slight to moderate degree of conservatism, except for the exceptional points x that comprise asymptotically a fraction ξ of $\mathcal{R}$. Taking $1 - \xi = 0.95$ may be warranted in the case of large samples.

2.4. *Three remarks on the algorithm.*

Remark 1. Calibration. In view of the undercoverage property discussed below (2.4), we expect $\hat{\beta}(x, \alpha_0)$, defined in step 6, to be less than α_0. Equivalently, we anticipate that the nominal coverage of the band has to be increased above $1 - \alpha_0$ in order for the band to cover $(x, g(x))$ with probability at least $1 - \alpha_0$. Conventionally we would employ $\hat{\beta}(x, \alpha_0)$ as the nominal level, but, owing to the large amount of stochastic error in the bootstrap bias estimator that is used implicitly in this technique, it produces confidence bands with poor coverage accuracy. This motivates coverage correction by calibration, along lines suggested by Hall (1986), Beran (1987) and Loh (1987), and resulting in our use of the adjusted nominal level $\hat{\alpha}_\xi(\alpha_0)$, defined in step 6.

Remark 2. Smoothing parameter for $\hat{g}^*$. An important aspect of step 4 is that we use the same empirical smoothing parameters for both $\hat{g}^*$ and $\hat{g}$, even though, in some respects, it might seem appropriate to use a bootstrap version of the smoothing parameters for $\hat{g}$ when estimating $\hat{g}^*$. However, since smoothing parameters should be chosen to effect an optimal tradeoff between bias and stochastic error, and the bias of $\hat{g}$ is not estimated accurately by the conventional bootstrap used in step 3 above, then the bootstrap versions of smoothing parameters, used to construct $\hat{g}^*$, are generally not asymptotically equivalent to their counterparts used for $\hat{g}$. This can cause difficulties. The innate conservatism of our methodology accommodates the slightly nonstandard smoothing parameter choice in step 4. Moreover, by not having to recompute the bandwidth at every bootstrap step we substantially reduce computational labour.

Remark 3. Heteroscedasticity. A heteroscedastic generalisation of the model at (2.1) has the form

$$Y_i = g(X_i) + \sigma(X_i)\,\epsilon_i\,, \tag{2.13}$$

where the ϵ_is have zero mean and unit variance, and $\sigma(x)$ is a nonnegative function that is estimated consistently by $\hat\sigma(x)$, say, computed from the dataset $\mathcal{Z}$ using either parametric or nonparametric methods. In this setting the variance of $\hat g(x)$ generally can be estimated by $s(\mathcal{X})^2\,\hat\sigma(x)^2$, where $s(\mathcal{X})$ is a known function of the design points, and the confidence band at (2.2) should be replaced by

$$\mathcal{B}(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat g(x) - s(\mathcal{X})(x)\,\hat\sigma(x)\,z_{1-(\alpha/2)} \le y \le \hat g(x) + s(\mathcal{X})(x)\,\hat\sigma(x)\,z_{1-(\alpha/2)}\Big\}\,.$$

The model for generating bootstrap data now has the form: $Y_i^* = \hat g(X_i) + \hat\sigma(X_i)\,\epsilon_i^*$, instead of: $Y_i^* = \hat g(X_i) + \epsilon_i^*$ in step 4; and the ϵ_i^*s are resampled conventionally from residual approximations to the ϵ_is.

With these modifications, the algorithm described in steps 1–6 can be implemented as before, and the resulting confidence bands have similar properties. In particular, if we redefine $\mathcal{B}^*(\alpha)$ by

$$\mathcal{B}^*(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat g^*(x) - s(\mathcal{X})(x)\,\hat\sigma^*(x)\,z_{1-(\alpha/2)} \le y \le \hat g^*(x) + s(\mathcal{X})(x)\,\hat\sigma^*(x)\,z_{1-(\alpha/2)}\Big\}$$

(compare (2.9)), and, using this new definition, continue to define $\hat\pi(x,\alpha)$ as at (2.10) (computed as at (2.11)); and if we continue to define $\beta = \hat\beta(x,\alpha_0)$ to be the solution of $\hat\pi(x,\beta) = 1-\alpha_0$, and to define $\hat\alpha_\xi(\alpha_0)$ as in (2.12); then the confidence band $\mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}$ is asymptotically conservative for at least a proportion $1-\xi$ of values $x \in \mathcal{R}$. This approach can be justified intuitively as in Appendix B.1 in the supplementary file, noting that, in the context of the model at (2.13), the expansion at (B.1) in the supplement should be replaced by:

$$E\{\hat g^*(x) \,|\, \mathcal{Z}\} - \hat g(x) = c_1\,g''(x)\,h^2 + (nh)^{-1/2}\,\sigma(x)\,f_X(x)^{-1/2}\,W(x/h) + \text{negligible terms}\,.$$

2.5. *Percentile bootstrap confidence bands.* The methods discussed above are based on the symmetric, asymptotic confidence band $\mathcal{B}(\alpha)$, which in turn is founded on a normal approximation. This approach is attractive because it requires only a single application of the bootstrap for calibration, but it is restrictive in that it dictates a conventional, symmetric "template" for the bands, because the normal model is symmetric. However, particularly if we would prefer the bands to be placed asymmetrically on either side of the estimator $\hat g$ so as to reflect skewness of the distribution of experimental errors, the initial confidence band $\mathcal{B}(\alpha)$, at (2.2), can be constructed using bootstrap methods, and a second iteration of the bootstrap, resulting in a double bootstrap method, can be used to refine coverage accuracy. This allows us to use, for example, equal-tailed intervals (where the amount of probability in either tail is taken to be the same) and so-called "shortest" intervals (where the confidence interval is chosen to be as short as possible, subject to having the desired nominal coverage). Of course, one-sided intervals can be constructed using either a normal approximation or a bootstrap approach, and our method carries over without difficulty to those settings.

The first bootstrap implementation is undertaken using step 4 of the algorithm in Section 2.3, and allows us to define the critical point $\hat z_\beta(x)$ by

$$P\{\hat g^*(x) - \hat g(x) \le s(\mathcal{X})\,\hat z_\beta \,|\, \mathcal{Z}\} = \beta\,, \tag{2.14}$$

for $\beta \in (0,1)$. The confidence band $\mathcal{B}(\alpha)$ is now re-defined as

$$\mathcal{B}(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat g(x) + s(\mathcal{X})(x)\,\hat z_{\alpha/2} \le y \le \hat g(x) + s(\mathcal{X})(x)\,\hat z_{1-(\alpha/2)}\Big\}\,. \tag{2.15}$$

The remainder of the methodology can be implemented in the following six-step algorithm.

(1) Calculate the uncentred bootstrap residuals, $\tilde\epsilon_i^* = Y_i^* - \hat g^*(X_i)$. (2) Centre them to obtain $\hat\epsilon_i^* = \tilde\epsilon_i^* - \bar\epsilon^*$, where $\bar\epsilon^* = n^{-1}\sum_i \tilde\epsilon_i^*$. (3) Draw a double-bootstrap resample, $\mathcal{Z}^{**} = \{(X_i, Y_i^{**}), 1 \le i \le n\}$, where $Y_i^{**} = \hat g^*(X_i) + \epsilon_i^{**}$ and the ϵ_i^{**}s are sampled randomly, with replacement, from the $\hat\epsilon_i^*$s. (4) Construct the bootstrap-world version $\mathcal{B}^*(\alpha)$ of the band $\mathcal{B}(\alpha)$ at (2.15), defined by

$$\mathcal{B}^*(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat g^*(x) + s(\mathcal{X})(x)\,\hat z^*_{\alpha/2} \le y \le \hat g^*(x) + s(\mathcal{X})(x)\,\hat z^*_{1-(\alpha/2)}\Big\}\,,$$

where, reflecting (2.14), $\hat z^*_\beta$ is defined by

$$P\{\hat g^{**}(x) - \hat g^*(x) \le s(\mathcal{X})\,\hat z^*_\beta \,|\, \mathcal{Z}^*\} = \beta\,,$$

and $\mathcal{Z}^*$ is defined as in step 3 of the algorithm in Section 2.3. (5) For this new definition of $\mathcal{B}^*(\alpha)$, define $\hat\pi(x,\alpha)$ as at (2.10). (6) Define $\hat\alpha_\xi(\alpha_0)$ as in (2.12), and take the final confidence band to be $\mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}$, where $\mathcal{B}(\alpha)$ is as at (2.15).

There is also a percentile-t version of this methodology, using our our quantile-based definition of $\hat\alpha_\xi(\alpha_0)$.

2.6. *Values of x that asymptotically are covered with probability at least $1-\alpha_0$.* Define $\|\mathcal{R}\|$ to equal the Lebesgue measure of $\mathcal{R}$, let $\mathcal{S}$ equal the set of $x \in \mathcal{R}$ such that $b(x) = 0$, put $\xi_0 = \|\mathcal{S}\|/\|\mathcal{R}\|$, define $\beta(x,\alpha_0)$ to be the solution, in β, of $\Phi\{z_{1-(\beta/2)} + b(x)\} - \Phi\{-z_{1-(\beta/2)} + b(x)\} = 1-\alpha_0$, and let $\alpha_\xi(\alpha_0)$ denote the $100\,\xi\%$ quantile of values of $\beta(x,\alpha_0)$ for $x \in \mathcal{R}$. Then $\alpha_\xi(\alpha_0)$ is the solution in γ of

$$\Big(\int_{\mathcal{R}} dx\Big)^{-1} \int_{\mathcal{R}} I\{\beta(x,\alpha_0) \le \gamma\}\,dx = \xi\,.$$

As ξ decreases, in order for the identity above to hold the value of γ should decrease. Hence, in accordance with intuition, $\alpha_\xi(\alpha_0)$ decreases as ξ decreases.

It can be proved that $\alpha_\xi(\alpha_0)$ is the limit in probability of $\hat\alpha_\xi(\alpha_0)$. Assume that the design points X_i are univariate and that f_X and g'' are bounded and continuous.

We showed in Section 2.2 that the naive confidence band $\mathcal{B}(\alpha_0)$, defined at (2.2) and having coverage $1-\alpha_0$, strictly undercovers $g(x)$ when evaluated at x, in the asymptotic limit, unless $b(x) = 0$, and that in the latter case the coverage is asymptotically correct, i.e. equals $1-\alpha_0$.

Noting that $\beta(x,\alpha_0)$ is a monotone increasing function of $|b(x)|$, and that $b(x) = -C\,g''(x)\,f_X(x)^{1/2}$ for a positive constant C, we see that if we define $\mathcal{R}(\xi) = \{x \in \mathcal{R} : \beta(x,\alpha_0) > \alpha_\xi(\alpha_0)\}$, and $c(\xi) = \sup\{C\,|g''(x)|\,f_X(x)^{1/2} : x \in \mathcal{R}(\xi)\}$, then the set of exceptional x, for which the confidence band $\mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}$ asymptotically undercovers $(x, g(x))$, is the set $\mathcal{S}_{\rm excep}$ of $x \in \mathcal{R}$ such that $C\,|g''(x)|\,f_X(x)^{1/2} > c(\xi)$. The Lebesgue measure of $\mathcal{S}_{\rm excep}$ equals $\max(0, \xi - \xi_0)\,\|\mathcal{R}\|$. See (2.2) for a definition of $\mathcal{B}(\alpha)$, and step 6 of Section 2.3 for a definition of $\hat\alpha_\xi(\alpha_0)$ and a detailed account of the construction of $\mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}$.

Typically the points in $\mathcal{S}_{\rm excep}$ are close to peaks and troughs, which can be identified from a graph of $\hat g$. In Section 3 we pay particular attention to numerical aspects of this issue.

2.7. *Confidence bands for probability densities.* Analogous methods can be used effectively to construct confidence bands for probability densities. We consider here the version of the single-bootstrap technique introduced in Section 2.3, when it is adapted so as to construct confidence bands for densities of r-variate probability distributions. Specifically, let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote a random sample drawn from a distribution with density f, let h be a bandwidth and K a kernel,

and define the kernel estimator of f by

$$\hat f(x) = \frac{1}{nh^r} \sum_{i=1}^n K\Big(\frac{x - X_i}{h}\Big)\,.$$

This estimator is asymptotically normally distributed with variance $(nh^r)^{-1}\,\kappa\,f(x)$, where $\kappa = \int K^2$, and so a naive, pointwise confidence band for $f(x)$ is given by

$$\mathcal{B}(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat f(x) - \big[(nh^r)^{-1}\,\kappa\,\hat f(x)\big]^{1/2}\,z_{1-(\alpha/2)} \le y \le \hat f(x) + \big[(nh^r)^{-1}\,\kappa\,\hat f(x)\big]^{1/2}\,z_{1-(\alpha/2)}\Big\}\,;$$

compare (2.2).

To correct $\mathcal{B}(\alpha)$ for coverage error, draw a random sample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ from the distribution with density $\hat f_X$, and define $\hat f^*$ to be the corresponding kernel estimator of $\hat f$, based on $\mathcal{X}$ rather than $\mathcal{X}^*$:

$$\hat f^*(x) = \frac{1}{nh^r} \sum_{i=1}^n K\Big(\frac{x - X_i^*}{h}\Big)\,.$$

Importantly, we do not generate $\mathcal{X}^*$ simply by resampling from $\mathcal{X}$. Analogously to (2.9), the bootstrap version of $\mathcal{B}(\alpha)$ is

$$\mathcal{B}^*(\alpha) = \Big\{(x,y) : x \in \mathcal{R}\,,\ \hat f^*(x) - \big[(nh^r)^{-1}\,\kappa\,\hat f^*(x)\big]^{1/2}\,z_{1-(\alpha/2)} \le y \le \hat f^*(x) + \big[(nh^r)^{-1}\,\kappa\,\hat f^*(x)\big]^{1/2}\,z_{1-(\alpha/2)}\Big\}\,.$$

For the reasons given in Remark 2 in Section 2.4 we use the same bandwidth, h, for both $\mathcal{B}(\alpha)$ and $\mathcal{B}^*(\alpha)$.

Our bootstrap estimator $\hat\pi(x,\alpha)$ of the probability $\pi(x,\alpha) = P\{(x, f(x)) \in \mathcal{B}(\alpha)\}$ that $\mathcal{B}(\alpha)$ covers $(x, f(x))$, is given by $\hat\pi(x,\alpha) = P\{(x,\hat g(x)) \in \mathcal{B}^*(\alpha) \,|\, \mathcal{X}\}$. As in step 6 of the algorithm in Section 2.3, for a given desired coverage level $1-\alpha_0$, let $\beta = \hat\beta(x,\alpha_0)$ be the solution of $\hat\pi(x,\beta) = 1-\alpha_0$, and define $\hat\alpha_\xi(\alpha_0)$ as in (2.12). Our final confidence band is $\mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}$. For a proportion of at least $1-\xi$ of the values of $x \in \mathcal{R}$, the limit of the probability that this band covers $f(x)$ is not less than $1-\alpha_0$, and for the remainder of values x the coverage error is close to 0.

In the cases $r = 1$ and 2, which are really the only cases where confidence bands can be depicted, theoretical results analogous to those in Section 4, for regression, can be developed using Hungarian approximations to empirical distribution functions. See, for example, Theorem 3 of Komlós, Major and Tusnády (1976) for the case $r = 1$, and Tusnády (1977) and Massart (1989) for $r \ge 2$. To link this argument to the theoretical development in Appendix B.1 in the supplementary file, we mention that in the univariate case, the analogue of (B.1) in that file is

$$E\{\hat f^*(x) \,|\, \mathcal{Z}\} - \hat f(x) = \frac{1}{2}\,\kappa_2\,f''(x)\,h^2 + (nh)^{-1/2}\,f(x)^{1/2}\,V(x/h) + \text{negligible terms}\,, \tag{2.16}$$

and (B.3) also holds. By way of notation in (2.16) and (B.3), $\kappa_2 = \int u^2\,K(u)\,du$ and, for constants c_1 and c_2, we define $b(x) = -c_1\,f''(x)\,f(x)^{-1/2}$ and $\Delta(x) = -c_2\,V(x)$; and V is a stationary Gaussian process with zero mean and covariance $K'' * K''$.

Alternative to the definition of $\mathcal{B}(\alpha)$ above, a confidence band based on the square-root transform, reflecting the fact that the asymptotic variance of $\hat f$ is proportional to f, could be used. Percentile and percentile-t methods, using our quantile-based method founded on $\hat\alpha_\xi(\alpha_0)$, can also be used.

3. Numerical properties.

3.1. *Parameter settings and comparisons.* In Section 3 we summarise the results of a simulation study addressing the finite-sample performance of methodology described in Section 2. In particular, we report empirical coverage probabilities of nominal 95% confidence intervals for $g(x)$, for different x, different values of $1-\xi$, different choices of g, different error variances σ^2, and different sample sizes n.

For $n = 100$, 200 or 400 we generated data pairs (X_i, Y_i) randomly from the model at (2.1), where the experimental errors ϵ_i were distributed independently as $\mathrm{N}(0, \sigma^2)$ with $\sigma = 1$, 0.5 or 0.2, and the explanatory variables X_i were distributed uniformly on $[-1, 1]$. We worked with the functions g_1, g_2 and g_3, defined by $g_1(x) = x + 5\,\phi(10\,x)$, $g_2(x) = \sin(3\pi x/2)\,/\{1 + 18\,x^2\,(\operatorname{sgn} x + 1)\}$ and $g_3(x) = \sin(\pi x/2)\,/\{1 + 2\,x^2\,(\operatorname{sgn} x + 1)\}$, where ϕ is the standard normal density and $\operatorname{sgn} x = 1$, 0 or -1 according as $x > 0$, $x = 0$ or $x < 0$, respectively. The function g_1 was used by Horowitz and Spokoiny (2001), and also by many subsequent authors; g_2 is the function given by formula (7) of Berry et al. (2002), rescaled here to the interval $[-1, 1]$, and used extensively by Berry et al. (2002) and in subsequent work of other researchers; and g_3 is the version of g_2 obtained by truncating g_2 to the central one third of its support interval, and rescaling so that it is supported on $[-1, 1]$.

The results reported here were obtained using a standard plug-in bandwidth, computed as suggested by Ruppert et al. (1995) but employing the variance estimator at (2.8). The cross-validation bandwidth gives slightly better coverage results for our method, apparently because, on average, it undersmooths a little. However, since computing the plug-in and cross-validation bandwidths involves $O(n)$ and $O(n^2)$ calculations, respectively, then the plug-in method is more attractive in a numerical study that requires 1000 simulations in each setting and sample sizes up to 400. The differences between plug-in and cross-validation were minor in the case of competing methods since, as discussed below, we optimised those methods over the second bandwidth.

In Section 3.2 we report results obtained using our method, undersmoothing without explicit bias correction, and explicit bias correction using an oversmoothed bandwidth to estimate bias. In the latter case we employed the regression version of a bias estimator suggested by Schucany and Sommers (1977). For each parameter setting (that is, each sample size n, each error variance σ^2 and each function g_j), when using undersmoothing we took the bandwidth to be $\gamma\,h$; and when using explicit bias correction we took the bandwidth to be h/λ. The values of γ and λ were chosen to optimise the performance of the two competing methods, and in particular so that those methods had as large as possible a proportion of values $x \in \mathcal{R} = [-0.9, 0.9]$ that were covered with probability at least 0.95. To determine the best γ and λ, for $n = 100$ we varied γ and λ in the ranges $0.1\,(0.1)\,0.9$ and $0.01, 0.02, 0.05, 0.1\,(0.1)\,0.9$, respectively. For $n = 200$ and 400, to reduce computation time we took the respective ranges to be $0.2\,(0.2)\,1.0$ and $0.1\,(0.2)\,0.9$.

This approach favours the two competing methods. It is required because there do not exist, in either case, any alternative approaches that are even moderately widely used. Of course, this situation, which arises because of the sheer difficulty of producing appropriate empirical bandwidths for the competing methods, is one of the motivations for our work. Choosing γ and λ empirically, as would be necessary in practice, would introduce significant extra variability into the competing methodologies, and so would downgrade their performance. Even the approach taken here, which gives competing methods every opportunity to show their advantages, typically produces competing techniques which perform less well than ours.

3.2. *Main results, and discussion.* Graphs of g_1, g_2 and g_3 are shown in Figure 1. The order g_1, g_2, g_3 arranges those functions in terms of decreasing difficulty experienced by each method. In particular, g_1, a single peak on a linear slope, is more challenging than g_2, which represents a deep

trough followed by a moderately high peak, and is more challenging still then g_3, which involves a moderately steep uphill slope followed by a gentle decrease. The extent of difficulty can be deduced from Tables 1–3, which reveal that the proportion of values of x that are covered with probability at least 0.95 increases, for each of the three methods, as we pass from g_1 to g_2 and then to g_3.

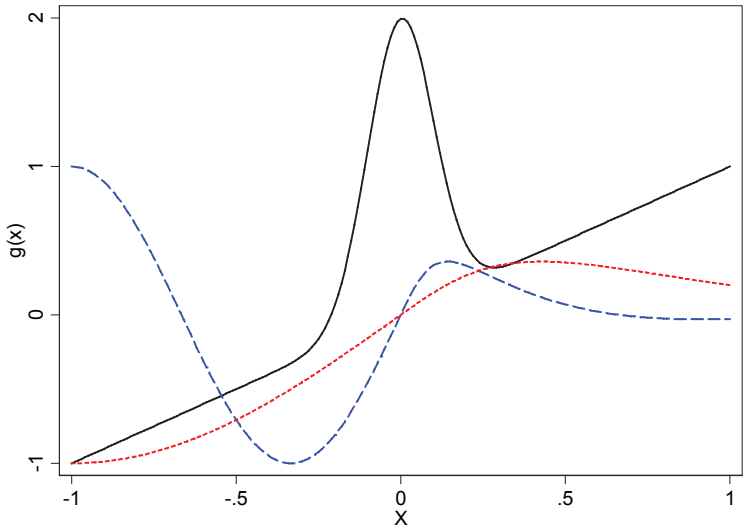


Fig 1: Conditional mean functions. Solid line is $g_1(x)$. Long dashes are $g_2(x)$. Short dashes are $g_3(x)$.

Table 1 treats the case $n = 100$, and shows, in the first column, the values of σ; in the second column, the index j of the function g_j; in the third column, the method; in the fourth column, the value of $1 - \xi$ (for our method), of the optimal γ (for the undersmoothing method), and of the optimal λ (for explicit bias correction); in the fifth column, the proportion of $x \in [-0.9, 0.9]$ for which the confidence band covered $g_j(x)$ with probability not less than 0.95 (referred to below as the "covered proportion"); in the sixth column, the integral average of the absolute values of coverage errors over $x \in [-0.9, 0.9]$; and in the seventh and last column, the average widths of the confidence intervals, i.e. the average widths of the bands constructed on $\mathcal{R}$. See Section 3.1 for definitions of γ and λ, and Section 2 for a definition of ξ.

Tables 2 and 3 provide the same information in the cases $n = 200$ and 400, respectively, although for brevity we give results only for $\sigma = 1$. The numerical values in Tables 1–3 were derived by taking averages over 1000 simulations in each parameter setting. In each instance, for the sake of brevity the tables give results only for three values of $1-\xi$, specifically 0.8, 0.9 and 0.95. When interpreting our results, and comparing them with those of the other methods, the reader should bear in mind that in practice we suggest taking $1 - \xi = 0.9$, whereas the competing methods have a major advantage in that we chose the tuning parameters there to give them the largest possible value of covered proportion.

Panels (a), (b) and (c) of Figure 2 each show three typical confidence bands in the cases of our method, of undersmoothing and of explicit bias correction, respectively, for $g = g_1$, $n = 100$ and $\sigma = 1$. (By "typical" bands we mean bands computed from the dataset for which the integrated squared error (ISE) of the estimator took the median value among 101 different datasets, and from the two datasets for which ISE was closest to but not equal to the median value.) To construct those bands in the case of our method we used $1-\xi = 0.9$. For bands in the other two cases we used the values of γ and λ that maximised covered proportions in the respective parameter settings.

σ	j	Method	$1-\xi$, γ, or λ	Prop. with cov. prob. ≥ 0.95	Av. abs. error of cov. prob.	Av. width
1	1	Ours	0.80	0.685	0.040	1.172
			0.90	0.774	0.041	1.217
			0.95	0.884	0.042	1.397
	2		0.80	0.702	0.025	0.970
			0.90	0.812	0.027	1.146
			0.95	1.000	0.034	1.322
	3		0.80	0.945	0.019	1.009
			0.90	0.995	0.033	1.096
			0.95	1.000	0.042	1.316
	1	Undersmooth	0.70	0.801	0.022	1.105
	2		0.60	0.840	0.018	1.076
	3		0.50	1.000	0.019	0.989
	1	Bias Corr.	0.05	0.737	0.034	0.924
	2		0.05	0.740	0.031	0.834
	3		0.10	0.901	0.015	0.700
0.5	1	Ours	0.80	0.724	0.038	0.949
			0.90	0.812	0.038	1.114
			0.95	0.895	0.039	1.197
	2		0.80	0.823	0.019	0.822
			0.90	0.945	0.027	0.924
			0.95	0.995	0.034	0.993
	3		0.80	0.923	0.018	0.482
			0.90	1.000	0.031	0.562
			0.95	1.000	0.041	0.642
	1	Undersmooth	0.80	0.785	0.024	0.595
	2		0.70	0.856	0.018	0.642
	3		0.70	1.000	0.019	0.452
	1	Bias Corr.	0.40	0.768	0.027	0.533
	2		0.20	0.785	0.019	0.573
	3		0.05	0.906	0.015	0.380
0.2	1	Ours	0.80	0.409	0.019	0.421
			0.90	0.834	0.020	0.497
			0.95	0.930	0.027	0.555
	2		0.80	0.879	0.020	0.366
			0.90	0.950	0.029	0.395
			0.95	0.961	0.036	0.424
	3		0.80	0.945	0.022	0.231
			0.90	1.000	0.033	0.257
			0.95	1.000	0.041	0.293
	1	Undersmooth	0.90	0.801	0.020	0.399
	2		0.80	0.818	0.021	0.282
	3		0.70	0.978	0.020	0.217
	1	Bias Corr.	0.20	0.790	0.022	0.378
	2		0.20	0.796	0.019	0.252
	3		0.90	0.995	0.019	0.190

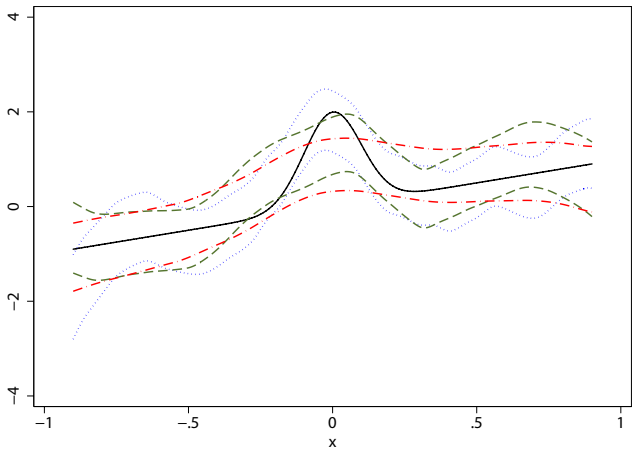
TABLE 1
Simulation results for $n = 100$.

σ	j	Method	$1-\xi$, γ, or λ	Prop. with cov. prob. ≥ 0.95	Av. abs. error of cov. prob.	Av. width
1	1	Ours	0.80	0.745	0.043	0.967
			0.90	0.843	0.042	1.105
			0.95	0.921	0.043	1.243
	2		0.80	0.751	0.023	0.878
			0.90	0.850	0.027	0.920
			0.95	1.000	0.033	0.962
	3		0.80	0.900	0.019	0.734
			0.90	0.995	0.031	0.801
			0.95	1.000	0.041	0.968
	1	Undersmooth	0.40	0.989	0.017	1.266
	2		0.40	1.000	0.020	1.228
	3		0.70	1.000	0.024	0.545
	1	Bias Corr.	0.10	0.762	0.034	0.800
	2		0.20	0.796	0.022	0.777
	3		0.10	0.928	0.018	0.456

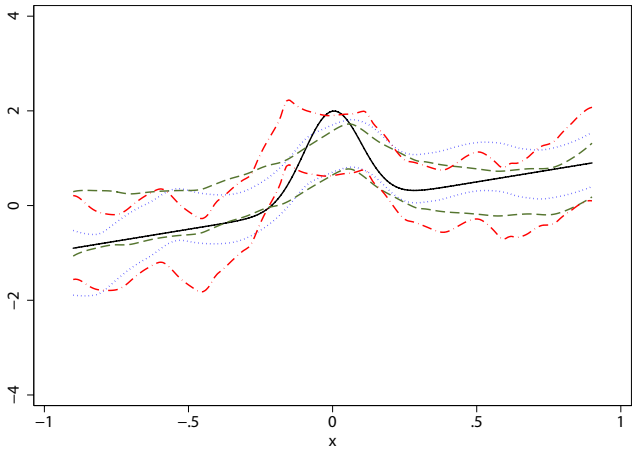
TABLE 2
Simulation results for $n = 200$.

σ	j	Method	$1-\xi$, γ, or λ	Prop. with cov. prob. ≥ 0.95	Av. abs. error of cov. prob.	Av. width
1	1	Ours	0.80	0.746	0.052	0.963
			0.90	0.807	0.048	1.005
			0.95	0.895	0.046	1.005
	2		0.80	0.818	0.022	0.911
			0.90	0.972	0.029	0.953
			0.95	1.000	0.036	0.953
	3		0.80	0.840	0.018	0.907
			0.90	0.995	0.030	0.948
			0.95	1.000	0.041	0.948
	1	Undersmooth	0.30	1.000	0.019	1.208
	2		0.70	1.000	0.024	0.637
	3		0.70	1.000	0.024	0.429
	1	Bias Corr.	0.40	0.801	0.027	0.662
	2		0.30	0.994	0.016	0.533
	3		0.10	0.956	0.019	0.356

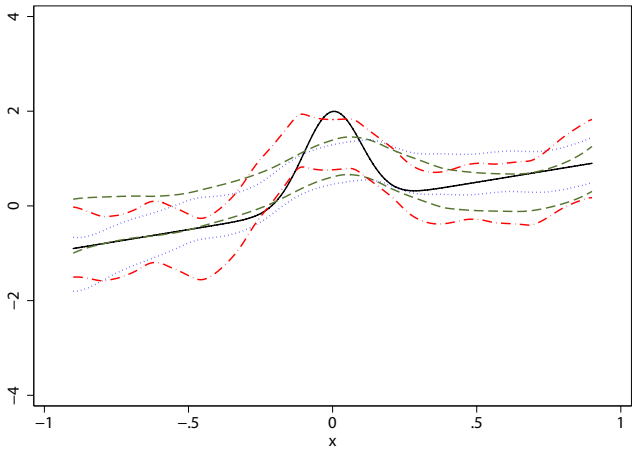
TABLE 3
Simulation results for $n = 400$.



(a) Proposed new method: 0.90 quantile.



(b) Conventional method with undersmoothing: $\gamma = 0.7$



(c) Conventional method with explicit bias correction: $\lambda = 0.05$

Fig 2: Comparison of three methods, each panel showing three confidence bands for interval $[-0.9, 0.9]$ with $n = 100$, $\sigma^2 = 1$, and $g(x) = x + 5\phi(10x)$, $X \sim U[-1, 1]$. Solid line is $g(x)$. Lower and upper limits of the bands indicated by dashes, dots and dash-dots.

The three panels in Figure 3 plot, as functions of x, unsmoothed values of the proportions of times, out of 1000 simulations, that the confidence band covered $(x, g(x))$. Each plot is for the case $n = 100$ and $\sigma = 1$, and panels (a), (b) and (c) in Figure 3 are for $g = g_1$, g_2 and g_3, respectively. The three curves in each panel represent the method suggested in this paper, the undersmoothing method and the explicit bias correction method, respectively. To illustrate coverage levels at endpoints our plots extend right across $[-1, 1]$; they are not restricted to $\mathcal{R} = [-0.9, 0.9]$.

It can be seen from Table 1 that, when $n = 100$, $\sigma^2 = 1$ and $1 - \xi = 0.9$, the proportion of values x for which $g_j(x)$ is covered with probability at least 0.95, when using our method, increases from 0.77 to 0.81 and then to 0.995, for $j = 1$, 2 and 3, respectively. The corresponding values of the "covered proportion" are 0.80, 0.84 and 1.0 for the undersmoothing method, and 0.74, 0.74 and 0.90 in the case of explicit bias correction. In particular, in this respect explicit bias correction is slightly inferior to our approach, and the undersmoothing method is slightly superior, at least in terms of the size of the covered proportion. However, this advantage is of undersmoothing is reversed when $\sigma = 0.5$ or 0.2.

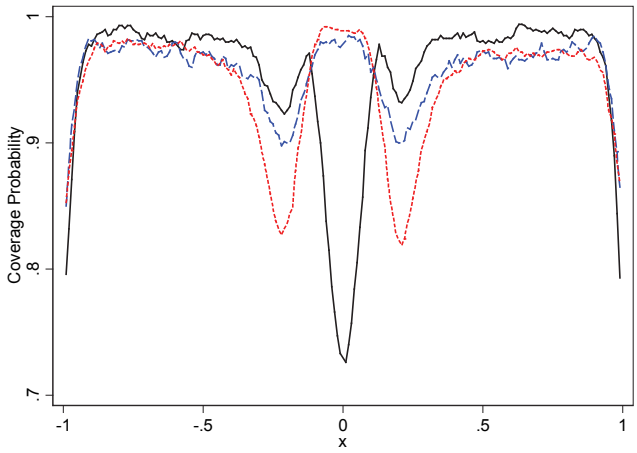
In the case of undersmoothing, the value of the covered proportion can drop sharply if there is stochastic error in choice of the bandwidth fraction, γ. Recall that in our simulation study we determine γ so that undersmoothing performs at its best, although in practice γ would be chosen implicitly using an algorithm based on estimating the second derivative of g_j; this is a noisy procedure at the best of times. To illustrate the difficulty of choosing γ in practice, we mention that, by Table 1, when $n = 100$ the optimal values of γ are 0.7, 0.6 and 0.5 when estimating g_1, g_2 and g_3, respectively, yielding covered proportions 0.801, 0.840 and 1.0, respectively. However, if we were to mistakenly use $\gamma = 0.4$, 0.3 or 0.2 in these respective cases, the covered proportions would drop to 0.558, 0.354 and 0.425, respectively.

Turning to panel (b) in Figure 2, which graphs typical confidence bands computed using the undersmoothing method, we see that the level of undersmoothing needed to achieve a relatively high level of covered proportion has made the band particularly wiggly, and hence very difficult to interpret. In practice this would be quite unsatisfactory. In comparison, the explicit bias corrected band is about as wiggly as the band constructed using our method (compare panels (a) and (c) in Figure 2), and both are easy to interpret.

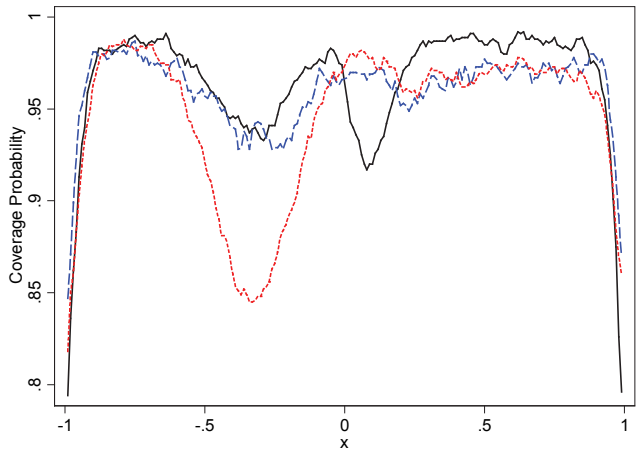
This trend can be seen generally, for different values of σ^2 and different sample sizes: The level of undersmoothing that must be used if the undersmoothing approach is to enjoy good coverage performance, produces bands that are distinctly unattractive because they exhibit a high degree of spatial variability that has nothing to do with actual features of the function g.

We should point out too that, in the case of undersmoothing, the proportion of values $x \in \mathcal{R}$ that are covered with probability at least 0.95 at first increases as the bandwidth decreases, but then starts to decrease. This is a consequence of the fact that the confidence band quickly becomes more erratic as the bandwidth is reduced, even more so than is shown in Figure 2. A similar phenomenon occurs when using explicit bias correction. Here the conservatively covered proportion of $\mathcal{R}$ at first increases as we decrease λ, but then it increases again. The reason is clear: If we were to use a large bandwidth then the bias estimator itself would be too heavily biased, with a consequent decline in coverage performance.

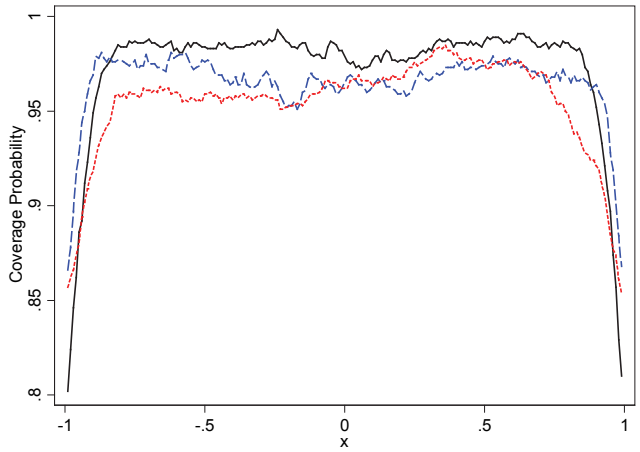
The plots in Figure 3 illustrate clearly the difficulty that each approach has with the bump function g_1 in the interval $(-0.3, 0.3)$, where the gradient of g_1 changes relatively quickly. Our approach undercovers most seriously at $x = 0$, but then again, it is honest about this; since we use $\xi = 0.1$ then our approach concedes from the outset that it can be expected to undercover approximately 10% of points in $\mathcal{R}$, and reflecting this the coverage accuracy improves relatively quickly away from the origin. For example, it is about 0.95 for $x = \pm 0.15$, although it drops briefly down to 0.9 in the near vicinity of ± 0.3. By way of comparison, the undersmoothing and explicit



(a) $g(x) = x + 5\phi(10x)$.



(b) $g(x) = \sin(3\pi x/2)/\{1 + 18x^2[\mathrm{sgn}(x) + 1]\}$.



(c) $g(x) = \sin(\pi x/2)/\{1 + 2x^2[\mathrm{sgn}(x) + 1]\}$

Fig 3: Coverage probabilities of nominal 95% confidence band. Each plot is for the case $n = 100$, $\sigma^2 = 1$ and $X \sim U[-1, 1]$, and panels (a), (b) and (c) are for $g = g_1$, g_2 and g_3, respectively. Solid line: proposed new method. Dashes: conventional method with undersmoothing. Dots: Conventional method with explicit bias correction.

bias correction approaches perform relatively well at $x = 0$, but drop away on either side.

All three methods have less difficulty with the function g_2, although it can be seen that they have more problems near the peak and the trough than anywhere else on $\mathcal{R}$. Finally, each method finds g_3 relatively easy. The same trends are seen also for larger sample sizes and smaller values of σ, although they are less marked in those cases.

The average lengths of confidence bands constructed using different methods vary in ways that are, in many instances, rather predictable. For example, when our method produces bands with larger covered proportion, which it does in most of the cases were considered, the bands themselves tend to be wider, as we would expect. It is of perhaps greater interest to focus on cases where our method has smaller covered proportion, i.e. the case $\sigma = 1.0$ with $n = 100$, 200 and 400. When $n = 100$ our bands are longer by between 7% (in the case of g_2) and 16% (for g_1), despite having lower coverage. However, when $n = 200$ our bands tend to be shorter in two out of three cases (the cases of g_1 and g_2), and when $n = 400$ they are shorter in one out of three cases (the case of g_1). For each method the average lengths of bands decrease relatively slowly as sample size increases.

4. Theoretical properties.

4.1. *Theoretical background.* In the present section we describe theoretical properties of bootstrap methods for estimating the distribution of $\hat{g}$. In Section 4.2 we apply our results to underpin the arguments in Section 2 that motivated our methodology. A proof of Theorem 4.1, below, is given in Appendix B.2 of Hall and Horowitz (2013).

We take $\hat{g}(x)$ to be a local polynomial estimator of $g(x)$, defined by (2.5) and (2.6). The asymptotic variance, Avar, of the local polynomial estimator $\hat{g}$ at x is given by

$$\text{Avar}\{\hat{g}(x)\} = D_1\,\sigma^2\,f_X(x)^{-1}\,(nh_1^r)^{-1}\,, \tag{4.1}$$

where $D_1 > 0$ depends only on the kernel and $\sigma^2 = \text{var}(\epsilon)$. (If $r = k = 1$ then $D_1 = \kappa \equiv \int K^2$.) With this in mind we take the estimator $s(\mathcal{X})(x)^2\,\hat{\sigma}^2$, introduced in Section 2.2, of the variance of $\hat{g}(x)$, to be $D_1\,\hat{\sigma}^2\,\hat{f}_X(x)^{-1}\,(nh^r)^{-1}$, where $\hat{f}_X$ is an estimator of the design density f_X and was introduced in step 1 of the algorithm in Section 2.3.

We assume that:

(4.2)
(a) the data pairs (X_i, Y_i) are generated by the model at (2.1), where the design variables X_i are identically distributed, the experimental errors ϵ_i are identically distributed, and the design variables and errors are totally independent; (b) $\mathcal{R}$ is a closed, nondegenerate rectangular prism in $\mathbb{R}^r$; (c) the estimator $\hat{g}$ is constructed by fitting a local polynomial of degree $2k-1$, where $k \geq 1$; (d) $\hat{f}_X$ is weakly and uniformly consistent, on $\mathcal{R}$, for the common density f_X of the r-variate design variables X_i; (e) g has $2k$ Hölder-continuous derivatives on an open set containing $\mathcal{R}$; (f) f_X is bounded on $\mathbb{R}^r$, and Hölder continuous and bounded away from zero on an open subset of $\mathbb{R}^r$ containing $\mathcal{R}$; (g) the bandwidth, h, used to construct $\hat{g}$, is a function of the data in $\mathcal{Z}$ and, for constants $C_1, C_2 > 0$, satisfies $P\{|h - C_1\,n^{-1/(r+4k)}| > n^{-(1+C_2)/(r+4k)}\} \to 0$, and moreover, for constants $0 < C_3 < C_4 < 1$, $P(n^{-C_4} \leq h \leq n^{-C_3}) = 1 - O(n^{-C})$ for all $C > 0$; (h) the kernel used to construct $\hat{g}$, at (2.5), is a spherically symmetric, compactly supported probability density, and has C_5 uniformly bounded derivatives on $\mathbb{R}^r$, where the positive integer C_5 is sufficiently large and depends on C_2; and (j) the experimental errors satisfy $E(\epsilon) = 0$ and $E|\epsilon|^{C_6} < \infty$, where $C_6 > 2$ is chosen sufficiently large, depending on C_2.

The model specified by (c) is standard in nonparametric regression. The assumptions imposed in (b), on the shape of $\mathcal{R}$, can be generalised substantially and are introduced here for notational simplicity. The restriction to polynomials of odd degree, in (c), is made so as to eliminate the somewhat anomalous behaviour in cases where the degree is even. See Ruppert and Wand (1994) for an account of this issue in multivariate problems. Condition (d) asks only that the design density be estimated uniformly consistently. The assumptions imposed on g and f_X in (e) and (f) are close to minimal when investigating properties of local polynomial estimators of degree $2k-1$. Condition (g) is satisfied by standard bandwidth choice methods, for example those based on cross-validation or plug-in rules. The assertion, in (g), that h be approximately equal to a constant multiple of $n^{-1/(r+2k)}$ reflects the fact that h would usually be chosen to minimise a measure of asymptotic mean L_p error, for $1 \le p < \infty$. Condition (h) can be relaxed significantly if we have in mind a particular method for choosing h. Smooth, compactly supported kernels, such as those required by (h), are commonly used in practice. The moment condition imposed in (j) is less restrictive than, for example, the assumption of normality.

In addition to (4.2) we shall, on occasion, suppose that:

$$\begin{aligned}&\text{the variance estimators } \hat\sigma^2 \text{ and } \hat\sigma^{*2} \text{ satisfy } P(|\hat\sigma - \sigma| > n^{-C_8}) \to 0 \text{ and}\\ &P(|\hat\sigma^* - \hat\sigma| > n^{-C_8}) \to 0 \text{ for some } C_8 > 0.\end{aligned} \tag{4.3}$$

In the case of the estimators $\hat\sigma^2$ defined at (2.7) and (2.8), if (4.2) holds then so too does (4.3).

Let $h_1 = C_1\, n^{-1/(r+4k)}$ be the deterministic approximation to the empirical bandwidth h asserted in (4.2)(g). Under (4.2) the asymptotic bias of a local polynomial estimator $\hat g$ of g, evaluated at x, is equal to $h_1^{2k}\, \nabla\, g(x)$, where ∇ is a linear form in the differential operators $(\partial/\partial x^{(1)})^{j_1} \ldots (\partial/\partial x^{(r)})^{j_r}$, for all choices of $j_1, \ldots, j_r$ such that each j_s is an even, positive integer, $j_1 + \ldots + j_r = 2k$ (the latter being the number of derivatives assumed of g in (4.2)(e)), and $x = (x^{(1)}, \ldots, x^{(r)})$. For example, if $r = k = 1$ then $\nabla = \frac{1}{2}\,\kappa_2\,(d/dx)^2$, where $\kappa_2 = \int u^2\, K(u)\, du$.

Recall that σ^2 is the variance of the experimental error ϵ_i. Let $L = K * K$, denoting the convolution of K with itself, and put $M = L - K$. Let W_1 be a stationary Gaussian process with zero mean and the following covariance function:

$$\operatorname{cov}\{W_1(x_1), W_1(x_2)\} = \sigma^2\, (M * M)(x_1 - x_2)\,. \tag{4.4}$$

Note that, since h_1 depends on n, then so too does the distribution of W_1. Our first result shows that (4.2) is sufficient for a stochastic approximation of local polynomial estimators.

THEOREM 4.1. *If (4.2) holds then, for each n, there exists a zero-mean Gaussian process W, having the distribution of W_1 and defined on the same probability space as the data $\mathcal{Z}$, such that for constants $D_2, C_7 > 0$,*

$$\begin{aligned}P\Big[\sup_{x\in\mathcal{R}} \Big| E\{\hat g^*(x) \,|\, \mathcal{Z}\} - \hat g(x) - \Big\{ h_1^{2k}\, \nabla\, g(x)\\ + D_2\, (n h_1^r)^{-1/2}\, f_X(x)^{-1/2}\, W(x/h_1)\Big\}\Big| > h_1^{2r}\, n^{-C_7}\Big] \to 0\end{aligned} \tag{4.5}$$

as $n \to \infty$. If, in addition to (4.2), we assume that (4.3) holds, then for some $C_7 > 0$,

$$\begin{aligned}P\Big(\sup_{x\in\mathcal{R}} \sup_{z\in\mathbb{R}} \Big| P\Big[\hat g^*(x) - E\{\hat g^*(x) \,|\, \mathcal{Z}\}\\ \le z\, \{D_1\, \hat\sigma^2\, \hat f_X(x)^{-1}\, (n h^r)^{-1}\}^{1/2} \;\Big|\; \mathcal{Z}\Big] - \Phi(z)\Big| > n^{-C_7}\Big) \to 0\end{aligned} \tag{4.6}$$

as $n \to \infty$.

Theorem 4.1 is generically similar to other strong approximations in the literature, although there are two differences that are crucial to our work: the bandwidth in the theorem is a function of the data, and has specific properties, whereas other strong approximations in nonparametric function estimation take the bandwidth to be deterministic; and the theorem treats data obtained using a particular residual-based approach to resampling, and does not treat the originally sampled data.

Result (4.6) asserts that the standard central limit theorem for $\hat{g}^*(x)$ applies uniformly in $x \in \mathcal{R}$. In particular, the standard deviation estimator $\{D_1\,\hat{\sigma}^2 \hat{f}_X(x)^{-1}\,(nh^r)^{-1}\}^{1/2}$, used to standardise $\hat{g}^* - E(\hat{g}^* \,|\, \mathcal{Z})$ on the left-hand side of (4.6), is none other than the conventional empirical form of the asymptotic variance of $\hat{g}$ at (4.1), and was used to construct the confidence bands discussed in Sections 2.2 and 2.3. The only unconventional aspect of (4.6) is that the central limit theorem is asserted to hold uniformly in $x \in \mathcal{R}$, but this is unsurprising, given the moment assumption in (4.2)(j).

4.2. *Theoretical properties of coverage error.* Let $D_3 = D_1^{-1/2}\,\sigma^{-1}$ and $D_4 = D_2\,D_3$, and define

$$b(x) = -D_3\,f_X(x)^{1/2}\,\nabla\,g(x)\,, \quad \Delta(x) = -D_4\,W(x/h_1)\,, \tag{4.7}$$

where W is as in (4.5). To connect these definitions to the theoretical outline in Appendix B.1 in the supplementary file, we note that in the present setting these are the versions of $b(x)$ and $\Delta(x)$ at (B.2) and (B.4), respectively ($D_4\,W$ in (4.7) equals W in (B.4)), and our first result in this section is a detailed version of (B.3):

Corollary 4.1. *If (4.2) and (4.3) hold then, with $z = z_{1-(\alpha/2)}$ and $b(x)$ and $\Delta(x)$ defined as above, we have for some $C_9 > 0$,*

$$P\Big(\sup_{x\in\mathcal{R}} \Big|\hat{\pi}(x,\alpha) - \Big[\Phi\{z + b(x) + \Delta(x)\} - \Phi\{-z + b(x) + \Delta(x)\}\Big]\Big| > n^{-C_9}\Big) \to 0 \tag{4.8}$$

as $n \to \infty$.

Next we give notation that enables us to assert, under specific assumptions, properties of coverage error of confidence bands. See particularly (4.13) in Corollary 4.2, below. Results (4.11) and (4.12) are used to derive (4.13), and are of interest in their own right because they describe large-sample properties of the quantities $\hat{\beta}(x,\alpha_0)$ and $\hat{\alpha}_\xi(\alpha_0)$, respectively, in terms of which our confidence bands are defined; see Section 2.3.

Given a desired coverage level $1-\alpha_0 \in (\frac{1}{2}, 1)$, define $\hat{\beta}(x,\alpha_0)$ and $\hat{\alpha}_\xi(\alpha_0)$ as in step 6 of Section 2.3, and as at (2.12), respectively. Let $b(x)$ and $\Delta(x)$ be as at (4.7), put $d = b + \Delta$, and define $T = T(x,\alpha_0)$ to be the solution of

$$\Phi\{T + d(x)\} - \Phi\{-T + d(x)\} = 1 - \alpha_0\,.$$

Then $T(x,\alpha_0) > 0$, and $A(x,\alpha_0) = 2\,[1 - \Phi\{T(x,\alpha_0)\}] \in (0,1)$. Define $\beta = \beta(x,\alpha_0) > 0$ to be the solution of

$$\Phi\{z_{1-(\beta/2)} + b(x)\} - \Phi\{-z_{1-(\beta/2)} + b(x)\} = 1 - \alpha_0\,, \tag{4.9}$$

and let $\alpha_\xi(\alpha_0)$ be the ξ-level quantile of the values of $\beta(x,\alpha_0)$. Specifically, $\gamma = \alpha_\xi(\alpha_0)$ solves the equation

$$\Big(\int_{\mathcal{R}} dx\Big)^{-1} \int_{\mathcal{R}} I\{\beta(x,\alpha_0) \le \gamma\}\,dx = \xi\,. \tag{4.10}$$

Define $\mathcal{R}_\xi(\alpha_0) = \{x \in \mathcal{R} : I[\beta(x,\alpha_0) > \alpha_\xi(\alpha_0)]\}$. Let the confidence band $\mathcal{B}(\alpha)$ be as at (2.2).

COROLLARY 4.2. *If (4.2) and (4.3) hold, then, for each $C_{10}, C_{11} > 0$, and as $n \to \infty$,*

$$P\Big\{ \sup_{x \in \mathcal{R} \,:\, |\Delta(x)| \le C_{10}} |\hat\beta(x, \alpha_0) - A(x, \alpha_0)| > C_{11} \Big\} \to 0\,, \tag{4.11}$$

$$P\{\hat\alpha_\xi(\alpha_0) \le \alpha_\xi(\alpha_0) + C_{11}\} \to 1\,, \tag{4.12}$$

for each $x \in \mathcal{R}_\xi(\alpha_0)$ the limit infimum of the probability $P[(x, g(x)) \in \mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}]$, as $n \to \infty$, is not less than $1 - \alpha_0$. (4.13)

Property (4.12) implies that the confidence band $\mathcal{B}(\beta)$, computed using $\beta = \hat\alpha_\xi(\alpha_0)$, is no less conservative, in an asymptotic sense, than its counterpart when $\beta = \alpha_\xi(\alpha_0)$. This result, in company with (4.13), underpins our claims about the conservatism of our approach. Result (4.13) asserts that the asymptotic coverage of $(x, g(x))$ by $\mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}$ is, for at most a proportion ξ of values of x, not less than $1 - \alpha_0$. Proofs of Corollaries 4.1 and 4.2 are given in Appendix A, below.

APPENDIX A: OUTLINE PROOFS OF COROLLARIES 4.1 AND 4.2

A.1. Proof of Corollary 4.1. Define

$$\hat d^*(x) = \frac{\hat g(x) - E\{\hat g^*(x) \,|\, \mathcal{Z}\}}{\{D_1\, \hat\sigma^{*2}\, \hat f_X(x)^{-1}\, (nh^r)^{-1}\}^{1/2}}\,, \quad \hat d(x) = \frac{\hat g(x) - E\{\hat g^*(x) \,|\, \mathcal{Z}\}}{\{D_1\, \sigma^2\, f_X(x)^{-1}\, (nh_1^r)^{-1}\}^{1/2}}\,.$$

Recall that, motivated by the variance formula (4.1), we take $s(\mathcal{X})(x)^2\, \hat\sigma^2$, in the definition of the confidence band $\mathcal{B}(\alpha)$ at (2.2), to be $D_1\, \hat\sigma^2\, \hat f_X(x)^{-1}\, (nh^r)^{-1}$. The bootstrap estimator $\hat\pi(x, \alpha)$, defined at (4.10), of the probability $\pi(x, \alpha)$, at (2.3), that the band $\mathcal{B}(\alpha)$ covers the the point $(x, g(x))$, is given by

$$\begin{aligned}
\hat\pi(x, \alpha) &= P\Big\{ \hat g^*(x) - s(\mathcal{X})(x)\, \hat\sigma^*\, z_{1-(\alpha/2)} \le \hat g(x) \le \hat g^*(x) + s(\mathcal{X})(x)\, \hat\sigma^*\, z_{1-(\alpha/2)} \;\Big|\; \mathcal{Z} \Big\} \\
&= P\left[- z_{1-(\alpha/2)} \le \frac{\hat g^*(x) - \hat g(x)}{\{D_1\, \hat\sigma^{*2}\, \hat f_X(x)^{-1}\, (nh^r)^{-1}\}^{1/2}} \le z_{1-(\alpha/2)} \;\middle|\; \mathcal{Z} \right] \\
&= P\left[- z_{1-(\alpha/2)} + \hat d^*(x) \le \frac{\hat g^*(x) - E\{\hat g^*(x) \,|\, \mathcal{Z}\}}{\{D_1\, \hat\sigma^{*2}\, \hat f_X(x)^{-1}\, (nh^r)^{-1}\}^{1/2}} \right. \\
&\qquad\qquad \left. \le z_{1-(\alpha/2)} + \hat d^*(x) \;\middle|\; \mathcal{Z} \right].
\end{aligned} \tag{A.1}$$

If both (4.2) and (4.3) hold then, by (4.5), (4.6), (A.1) and minor additional calculations,

$$P\Big(\sup_{x \in \mathcal{R}} \Big| \hat\pi(x, \alpha) - \Big[\Phi\{ z_{1-(\alpha/2)} + \hat d(x) \} - \Phi\{ - z_{1-(\alpha/2)} + \hat d(x) \} \Big] \Big| > n^{-C_9} \Big) \to 0\,. \tag{A.2}$$

Now, $-\hat d(x) = D_3\, f_X(x)^{1/2}\, \nabla\, g(x) + D_4\, W(x/h_1)$ where $D_3 = D_1^{-1/2}\, \sigma^{-1}$ and $D_4 = D_2\, D_3$, and so (4.8) follows from (A.2).

A.2. Proof of Corollary 4.2. Result (4.11) follows from (4.8). Shortly we shall outline a proof of (4.12); at present we use (4.12) to derive (4.13). To this end, recall that $\gamma = \alpha_\xi(\alpha_0)$ solves equation (4.10) when $z = z_{1-(\beta/2)}$, and $\beta = \beta(x, \alpha_0) > 0$ denotes the solution of equation (4.9). If (4.12) holds then (4.13) will follow if we establish that result when $\hat\alpha_\xi(\alpha_0)$, in the quantity

$P[(x, g(x)) \in \mathcal{B}\{\hat\alpha_\xi(\alpha_0)\}]$ appearing in (4.13), is replaced by $\alpha_\xi(\alpha_0)$. Call this property (P). Now, the definition of $\alpha_\xi(\alpha_0)$, and the following monotonicity property,

$$\Phi(z+b) - \Phi(-z+b) \text{ is a decreasing (respectively, increasing) function of } b \text{ for } b > 0 \text{ (respectively, } b < 0) \text{ and for each } z > 0, \tag{A.3}$$

ensure that

$$\liminf_{n\to\infty} P[(x, g(x)) \in \mathcal{B}\{\alpha_\xi(\alpha_0)\}] \ge 1 - \alpha_0$$

whenever $\beta(x, \alpha_0) \le \alpha_\xi(\alpha_0)$, or equivalently, whenever $x \in \mathcal{R}_\xi(\alpha_0)$. This establishes (P).

Finally we derive (4.12), for which purpose we construct a grid of edge width δ, where δ is small (see (A.4) below), and show that if this grid is used to define $\hat\alpha_\xi(\alpha_0)$ (see (2.12)) then (4.12) holds. Let $x'_1, \ldots, x'_{N_1}$ be the centres of the cells, in a regular rectangular grid in $\mathbb{R}^r$ with edge width δ_1, that are contained within $\mathcal{R}$. (For simplicity we neglect here cells that overlap the boundaries of $\mathcal{R}$; these have negligible impact.) Within each cell that intersects $\mathcal{R}$, construct the smaller cells (referred to below as subcells) of a subgrid with edge width $\delta = m^{-1}\delta_1$, where $m = m(\delta_1) \ge 1$ is an integer and $m \sim \delta_1^{-c}$ for some $c > 0$. Put $N = m^r N_1$; let $x_{j\ell}$, for $j = 1, \ldots, N_1$ and $\ell = 1, \ldots, m^r$, denote the centres of the subcells that are within the cell that has centre x'_j; and let $x_1, \ldots, x_N$ be an enumeration of the values of $x_{j\ell}$, with $x_{11}, \ldots, x_{1m}$ listed first, followed by $x_{21}, \ldots, x_{2m}$, and so on. Recalling the definition of $\hat\alpha_\xi(\alpha_0)$ at (2.12), let $\hat\alpha_\xi(\alpha_0, \delta)$ denote the ξ-level quantile of the sequence $\hat\alpha(x_1, \alpha_0), \ldots, \hat\alpha(x_N, \alpha_0)$.

Let $h_1 = C_1\, n^{-1/(r+4k)}$ represent the asymptotic size of the bandwidth asserted in (4.2)(g), and assume that

$$\delta = O(n^{-B_1}), \quad 1/(r+4k) < B_1 < \infty. \tag{A.4}$$

Then

$$\delta = O(h_1\, n^{-B_2}) \tag{A.5}$$

for some $B_2 > 0$. In particular, δ is an order of magnitude smaller than h_1.

Recall that $A(x, \alpha_0) = 2\,[1 - \Phi\{Z(x, \alpha_0)\}] \in (0, 1)$, where $Z = Z(x, \alpha_0) > 0$ is the solution of

$$\Phi\{Z + b(x) + \Delta(x)\} - \Phi\{-Z + b(x) + \Delta(x)\} = 1 - \alpha_0,$$

and $\Delta(x) = -D_4\, W(x/h_1)$; and that $\beta = \beta(x, \alpha_0) > 0$ solves $\Phi\{\beta + b(x)\} - \Phi\{-\beta + b(x)\} = 1 - \alpha_0$. Define $e(x, \alpha_0) = 2\,[1 - \Phi\{\beta(x, \alpha_0)\}]$. Given a finite set $\mathcal{S}$ of real numbers, let $\mathrm{quant}_\xi(\mathcal{S})$ and $\mathrm{med}(\mathcal{S}) = \mathrm{quant}_{1/2}(\mathcal{S})$ denote, respectively, the ξ-level empirical quantile and the empirical median of the elements of $\mathcal{S}$. Noting (A.3), and the fact that the stationary process W is symmetric (W is a zero-mean Gaussian process the distribution of which does not depend on n), it can be shown that $P\{Z(x, \alpha_0) > \beta(x, \alpha_0)\} = P\{Z(x, \alpha_0) \le \beta(x, \alpha_0)\} = \frac{1}{2}$. Therefore the median value of the random variable $A(x, \alpha_0)$ equals $e(x, \alpha_0)$. Hence, since the lattice subcell centres $x_{j1}, \ldots, x_{jm^r}$ are clustered regularly around x_j, it is unsurprising, and can be proved using (A.5), that the median of $A(x_{j1}, \alpha_0), \ldots, A(x_{jm^r}, \alpha_0)$ is closely approximated by $e(x, \alpha_0)$, and in particular that for some $B_3 > 0$ and all $B_4 > 0$,

$$P\Big\{ \max_{j=1,\ldots,N_1} \Big|\mathrm{med}\{A(x_{j1}, \alpha_0), \ldots, A(x_{jm^r}, \alpha_0)\} - e(x_j, \alpha_0)\Big| > n^{-B_3} \Big\} = O(n^{-B_4}).$$

Therefore, since the ξ-level quantile of the points in the set

$$\bigcup_{j=1}^{N_1} \{A(x_{j1}, \alpha_0), \ldots, A(x_{jm^r}, \alpha_0)\}$$

is bounded below by $\{1 + o_p(1)\}$ multiplied by the ξ-level quantile of the N_1 medians

$$\text{med}\{A(x_{j1}, \alpha_0), \ldots, A(x_{jm^r}, \alpha_0)\}\,, \quad 1 \le j \le N_1\,,$$

then for all $\eta > 0$,

$$P\Big[\text{quant}_{1-\xi}\,\{A(x, \alpha_0) : x \in \mathcal{R}\} \le \text{quant}_{1-\xi}\,\{e(x, \alpha_0) : x \in \mathcal{R}\} + \eta\Big] \to 1\,. \tag{A.6}$$

Since $\text{quant}_{1-\xi}\,\{e(x, \alpha_0) : x \in \mathcal{R}\} = \alpha_\xi(\alpha_0)$ then, by (A.6),

$$P\Big[\text{quant}_{1-\xi}\,\{A(x, \alpha_0) : x \in \mathcal{R}\} \le \alpha_\xi(\alpha_0) + \eta\Big] \to 1\,. \tag{A.7}$$

In view of (4.11),

$$P\Big[\Big|\text{quant}_{1-\xi}\,\{A(x, \alpha_0) : x \in \mathcal{R}\} - \text{quant}_{1-\xi}\,\{\hat\beta(x, \alpha_0) : x \in \mathcal{R}\}\Big| > \eta\Big] \to 0 \tag{A.8}$$

for all $\eta > 0$, and moreover, if δ satisfying (A.4) is chosen sufficiently small,

$$\text{quant}_{1-\xi}\,\{\hat\beta(x, \alpha_0) : x \in \mathcal{R}\} - \hat\alpha_\xi(\alpha_0) \to 0 \tag{A.9}$$

in probability. (This can be deduced from the definition of $\hat\alpha_\xi(\alpha_0)$ at (2.12).) Combining (A.7)–(A.9) we deduce that $P\{\hat\alpha_\xi(\alpha_0) \le \alpha_\xi(\alpha_0) + \eta\} \to 1$ for all $\eta > 0$, which is equivalent to (4.12).

SUPPLEMENTARY MATERIAL

Appendix B: Supplementary material
(doi: 10.1214/13-AOS1137SUPP). The supplementary material in Appendix B.1 outlines theoretical properties underpinning our methodology, while Appendix B.2 contains a proof of Theorem 4.1.

REFERENCES

BERAN, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–468.

BERRY, S.M., CARROLL, R.J. and RUPPERT, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *J. Amer. Statist. Assoc.* **97** 160–169.

BJERVE, S., DOKSUM, K.A. and YANDELL, B.S. (1985). Uniform confidence bounds for regression based on a simple moving average. *Scand. J. Statist.* **12** 159–169.

BROWN, L.D. and LEVINE, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Ann. Statist.* **35** 2219–2232.

BUCKLEY, M.J., EAGLESON, G.K. and SILVERMAN, B.W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75** 189–199.

CAI, T.T., LEVINE, M. and WANG, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivar. Anal.* **100** 126–136.

CAI, T.T. and LOW, M.G. (2006). Adaptive confidence balls. *Ann. Statist.* **34** 202–228.

CHEN, S.X. (1996). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika* **83** 329–341.

CHEN, S.X., HÄRDLE, W. and LI, M. (2003). An empirical likelihood goodness-of-fit test for time series. *J. Roy. Statist. Soc.* Ser. B **65** 663–678.

CLAESKENS, G. and VAN KEILEGOM, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31** 1852–1884.

DETTE, H., MUNK, A. and WAGNER, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. Roy. Statist. Soc.* Ser. B **60** 751–764.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

EUBANK, R. L. AND SPECKMAN, P. L. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88** 1287–1301.

EUBANK, R.L. and WANG, S. (1994). Confidence regions in non-parametric regression. *Scand. J. Statist.* **21** 147–157.
FAN, J. and YAO, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85** 645–660.
GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.
GENOVESE, C. and WASSERMAN, L. (2005). Nonparametric confidence sets for wavelet regression. *Ann. Statist.* **33** 698–729.
GENOVESE, C. and WASSERMAN, L. (2008). Adaptive confidence bands. *Ann. Statist.* **36** 875–905.
GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170.
HALL, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14** 1431–1452.
HALL, P. (1992a). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694.
HALL, P. (1992b). On bootstrap confidence intervals in nonparametric regression. *Ann. Statist.* **20** 695–711.
HALL, P. and HOROWITZ, J. (2013). Supplement to "A simple bootstrap method for constructing nonparametric confidence bands for functions." DOI: 10.1214/13-AOS1137SUPP.
HALL, P., KAY, J.W. and TITTERINGTON, D.M. (1990). Asymptotically optimal difference based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528.
HALL, P. and MARRON, J.S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77** 415–419.
HALL, P. and OWEN, A.B. (1993). Empirical likelihood confidence bands in density estimation. *J. Comput. Graph. Statist.* **2** 273–289.
HALL, P. and TITTERINGTON, D.M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.* **27** 228–254.
HÄRDLE, W. and BOWMAN, A.W. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102–110.
HÄRDLE, W., HUET, S. and JOLIVET, S. (1995). Better bootstrap confidence-intervals for regression curve estimation. *Statistics* **26** 287–306.
HÄRDLE, W., HUET, S., MAMMEN, E. and SPERLICH, S. (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory* **20** 265–300.
HÄRDLE, W. and MARRON, J.S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19** 778–796.
HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409.
HOROWITZ, J.L. and SPOKOINY, V.G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69** 599–531.
KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **34** 33–58.
LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008.
LOH, W.-Y. (1987). Calibrating confidence coefficients. *J. Amer. Statist. Assoc.* **82** 155–162.
LOW, M.G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554.
MCMURRY, T.L. and POLITIS, D.M. (2008). Bootstrap confidence intervals in nonparametric regression with built-in bias correction. *Statist. Probab. Lett.* **78** 2463–2469.
MASSART, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT constructions. *Ann. Probab.* **17** 266–291.
MENDEZ, G. and LOHR, S. (2011). Estimating residual variance in random forest regression. *Comput. Statist. Data Anal.* **55** 2937–2950.
MÜLLER, H.-G. and STADTMÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15** 610–635.
MÜLLER, H.-G. and STADTMÜLLER, U. (1992). On variance estimation with quadratic forms. *J. Statist. Plann. Inference* **35** 213–231.
MÜLLER, H.-G. and ZHAO, P.-L. (1995). On a semiparametric variance function model and a test for heteroscedasticity. *Ann. Statist.* **23** 946–967.
MÜLLER, U.U., SCHICK, A. and WEFELMEYER, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics* **37** 179–188.
MUNK, A., BISSANTZ, N., WAGNER, T. and FRIETAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Statist. Soc.* Ser. B **67** 19–41.
NEUMANN, M.H. (1994). Fully data-driven nonparametric variance estimators. *Scand. J. Statist.* **25** 189–212.
NEUMANN, M.H. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *Ann. Statist.* **23** 1937–1959.
NEUMANN, M.H. and POLZEHL, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *J. Nonparametric Statist.* **9** 307–333.

PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335.
RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.
RUPPERT, D., SHEATHER, S.J. and WAND, M.P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **9** 1257-1270.
RUPPERT, D. and WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370.
RUPPERT, D., WAND, M.P. and CARROLL, R.J. (2003). *Semiparametric Regression*. Cambridge University Press.
SCHUCANY, W.R. and SOMMERS, J.P. (1977). Improvement of kernel type density estimators. *J. Amer. Statist. Assoc.* **72** 420–423.
SEIFERT, B., GASSER, T. and WOLF, A. (1993). Nonparametric-estimation of residual variance revisited. *Biometrika* **80** 373–383.
SUN, J. and LOADER, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Ann. Statist.* **22** 1328–1345.
TONG, T. and WANG, Y. (2005). Estimating residual variance in nonparametric regression using least squares. *Biometrika* **92** 821–830.
TUSNÁDY, G. (1977). A remark on the approximation of the sample DF in the multidimensional case. *Period. Math. Hungar.* **8** 53–55.
WANG, Y.D. and WAHBA, G. (1995). Bootstrap confidence-intervals and for smoothing splines and their comparison to Bayesian confidence-intervals. *J. Statist. Comput. Simul.* **51** 263–279.
XIA, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. Roy. Statist. Soc.* Ser. B **60** 797–811.

Department of Mathematics and Statistics
The University of Melbourne
VIC 3010, Australia
and
Department of Statistics
University of California
Davis, CA 95616, USA.
E-mail: halpstat@ms.unimelb.edu.au

Department of Economics
Northwestern University
2001 Sheridan Road
Evanston, Illinois 60208, USA.
E-mail: joel-horowitz@northwestern.edu