Yale University

EliScholar – A Digital Platform for Scholarly Publishing at Yale

Cowles Foundation Discussion Papers | Cowles Foundation

5-1-2016

Optimal Inference in a Class of Regression Models

Timothy B. Armstrong

Michal Kolesár

Follow this and additional works at: https://elischolar.library.yale.edu/cowles-discussion-paper-series

 Part of the Economics Commons

Recommended Citation

Armstrong, Timothy B. and Kolesár, Michal, "Optimal Inference in a Class of Regression Models" (2016). *Cowles Foundation Discussion Papers*. 2494.
https://elischolar.library.yale.edu/cowles-discussion-paper-series/2494

This Discussion Paper is brought to you for free and open access by the Cowles Foundation at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Cowles Foundation Discussion Papers by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

OPTIMAL INFERENCE IN A CLASS OF REGRESSION MODELS

By

Timothy B. Armstrong and Michal Kolesár

May 2016

COWLES FOUNDATION DISCUSSION PAPER NO. 2043



COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Optimal inference in a class of regression models*

Timothy B. Armstrong† Yale University

Michal Kolesár‡ Princeton University

May 26, 2016

Abstract

We consider the problem of constructing confidence intervals (CIs) for a linear functional of a regression function, such as its value at a point, the regression discontinuity parameter, or a regression coefficient in a linear or partly linear regression. Our main assumption is that the regression function is known to lie in a convex function class, which covers most smoothness and/or shape assumptions used in econometrics. We derive finite-sample optimal CIs and sharp efficiency bounds under normal errors with known variance. We show that these results translate to uniform (over the function class) asymptotic results when the error distribution is not known. When the function class is centrosymmetric, these efficiency bounds imply that minimax CIs are close to efficient at smooth regression functions. This implies, in particular, that it is impossible to form CIs that are tighter using data-dependent tuning parameters, and maintain coverage over the whole function class. We specialize our results to inference in a linear regression, and inference on the regression discontinuity parameter, and illustrate them in simulations and an empirical application.

*We thank Isaiah Andrews, Matias Cattaneo, Gary Chamberlain, Denis Chetverikov, Ulrich Müller, and Azeem Shaikh for useful discussions, and numerous seminar and conference participants for helpful comments and suggestions. All remaining errors are our own.

†email: timothy.armstrong@yale.edu

‡email: mkolesar@princeton.edu

1 Introduction

In this paper, we study the problem of constructing one- and two-sided confidence intervals (CIs) for a linear functional Lf of an unknown regression function f in a broad class of regression models with fixed regressors, in which f is known to belong to some convex function class $\mathcal{F}$. The linear functional may correspond to the value of f at a point, the regression discontinuity parameter, an average treatment effect under unconfoundedness, or a regression coefficient in a linear or partly linear regression. The class $\mathcal{F}$ may contain smoothness restrictions (e.g. a bound on the second derivative, or assuming f is linear as in a linear regression), and/or shape restrictions (such as monotonicity, or sign restrictions on regression coefficients in a linear regression). Often in applications, the function class will be indexed by a smoothness parameter C. This is the case, for instance, when $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$, the class of Lipschitz continuous functions with Lipschitz constant C.

We further assume that the regression errors are normal, with known variance, which allows us to derive finite-sample optimal CIs and sharp finite-sample efficiency bounds. We show that these finite-sample results translate to uniform asymptotic results when the error distribution is unknown under high-level regularity conditions, and derive sufficient low-level conditions in an application to regression discontinuity. This finite-sample approach allows us to use the same framework and methods to cover problems that are often seen as outside of the scope of nonparametric methods, such as discrete regressors in regression discontinuity (Lee and Card, 2008) and linear regression with restrictions on the sign and magnitude of coefficients (Andrews, 2001). In our setup, one need not worry about whether regressors can be considered continuous or discrete, or whether the constraints on f are "parametric" or "nonparametric."

Our main contribution is to derive sharp efficiency bounds that have implications for data-driven approaches to model and bandwidth selection in both "parametric" and "nonparametric" settings. First, for a given quantile β, we characterize one-sided CIs that minimize the maximum β quantile of excess length over a convex class $\mathcal{G}$. The optimal CI $[\hat{c}, \infty)$ takes a simple form. The lower limit $\hat{c}$ is obtained by taking an estimator $\hat{L}$ that trades off bias and variance in a certain optimal sense and is linear in the outcome vector, and subtracting (1) the standard deviation of $\hat{L}$ times the usual critical value based on a normal distribution and (2) a bias correction to guarantee proper coverage. This bias correction, in contrast to bias corrections often used in practice, is based on the maximum bias of $\hat{L}$ over $\mathcal{F}$, and is therefore non-random.

When $\mathcal{G} = \mathcal{F}$, this procedure yields minimax one-sided CIs. Setting $\mathcal{G} \subset \mathcal{F}$ to a class

of smoother functions is equivalent to "directing power" or attempting to "adapt" to these smoother functions while maintaining coverage over $\mathcal{F}$. The procedure gives a sharp bound on the scope for adaptation for one-sided CIs. We show that when the class $\mathcal{F}$ is centrosymmetric (i.e. $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$), the scope for directing power is severely limited: CIs that are minimax for β quantile of excess length also optimize excess length over a class $\mathcal{G}$ of functions that are sufficiently smooth (such as the singleton class comprising just the zero function, or the class of constant functions if $\mathcal{F}$ places bounds on derivatives) but at a different quantile. Furthermore, a CI that is minimax for a given quantile is also highly efficient at smooth functions for the same quantile. For instance, a CI for the conditional mean at a point that is minimax over the Lipschitz class $\mathcal{F}_{\text{Lip}}(C)$ is asymptotically 95.2% efficient at a constant function relative to a CI that directs all power at this function. For function classes smoother than $\mathcal{F}_{\text{Lip}}(C)$, the efficiency is even higher.

Our second main result is to derive a confidence set that minimizes its expected length at a single function g. We compare the performance of this confidence set to the optimal fixed-length CI derived in Donoho (1994) (i.e. confidence intervals of the form $\hat{L} \pm \chi$, where $\hat{L}$ is an affine estimator and the non-random half-length χ, which depends only on the regressors, is chosen to satisfy the coverage requirement). We find that, similarly to minimax one-sided CIs, when $\mathcal{F}$ is centrosymmetric, confidence sets that optimize expected length at a function g that is sufficiently smooth are not shorter than fixed-length CI by more than a few percentage points. For instance, the fixed-length CI for a conditional mean at a point when f is constrained to be in $\mathcal{F}_{\text{Lip}}(C)$ is asymptotically 95.6% efficient at any constant function relative to a confidence set that optimizes its excess length at this function.

An important practical implication of these results is that it is not possible to avoid having to explicitly specify the smoothness constant C: procedures that use data-driven rules to determine the smoothness of f (such as using data-driven bandwidths or variable selection) must either fail to improve upon the minimax CIs or fixed-length CIs (that effectively assume the worst case smoothness), or else fail to maintain coverage over the whole parameter space.

In order to avoid having to specify the smoothness constant, one has to strengthen the assumptions on f. For instance, one can impose shape restrictions that break the centrosymmetry, as in Cai, Low, and Xia (2013) or Armstrong (2015), or self-similarity assumptions that break the convexity, as in Giné and Nickl (2010) or Chernozhukov, Chetverikov, and Kato (2014).

Alternatively, one can consider intervals that satisfy weaker notions of coverage than the traditional definition of a confidence interval, such as average coverage (see Cai, Low, and

Ma, 2014 and Hall and Horowitz, 2013 for recent examples).

We apply these results to two popular models. First, we consider the problem of inference in linear regression with restricted parameter space.

The general results give bounds for the scope for directing power at "smooth" alternatives where certain parameters are zero while maintaining coverage over a convex parameter space. Since directing power at such alternatives is often the goal of model or variable selection, unless one imposes non-convex or asymmetric restrictions on the parameter space, the scope for model or variable selection, such as using CIs considered in Andrews and Guggenberger (2009a) and McCloskey (2012), is severely limited. We also discuss sparsity as a non-convex constraint and point out that, while it is possible to adapt to the indices of non-zero coefficients, our results bound the scope for adapting to the number of non-zero coefficients.

Second, we consider inference about the regression discontinuity parameter. We illustrate our results an empirical application from Lee (2008), and show that the resulting CIs are informative and simple to construct. We also consider one-sided CIs and two-sided fixed-length CIs based on local linear estimators, with bandwidths chosen to optimize their maximum excess length and half-length, respectively.

Local linear estimators have been popular in empirical practice for regression discontinuity due to asymptotic relative efficiency results of Cheng, Fan, and Marron (1997) for minimax estimation with squared error loss. Using the same function classes as Cheng, Fan, and Marron (1997), we consider finite-sample efficiency for CIs, and we compute efficiency at smooth functions as well as minimax efficiency. We show that in the Lee (2008) application, CIs based on local linear estimators with triangular kernel are highly efficient relative to the optimal CIs discussed above. We also illustrate through a Monte Carlo study that popular data-driven bandwidth selectors used in the regression discontinuity setting lead to undercoverage, even when one uses these bandwidth selectors as a starting point for bias correction or undersmoothing (see Appendix A).

Our results and setup build on a large statistics literature on optimal estimation and inference in the nonparametric regression model. This literature has mostly been concerned with constructing an optimal estimator of Lf (see, e.g., Stone (1980), Ibragimov and Khas'minskii (1985), Fan (1993), Donoho (1994), Cheng, Fan, and Marron (1997) and references therein), and it is often cited in econometrics to formalize claims about optimal kernels and rates of convergence.[1] Our results are closely related to those in Low (1997) and Cai and Low

[1]For example, in their survey of nonparametric methods in econometrics, Ichimura and Todd (2007) cite

(2004a), who consider confidence sets that take the form of a two-sided CI, and, subject to coverage over $\mathcal{F}$, derive bounds on the maximum expected length over $\mathcal{G}$, and the results in Cai, Low, and Xia (2013), who obtain bounds on the expected length of two-sided CIs at a single function g. The bounds of Low (1997) and Cai and Low (2004a) imply that when $\mathcal{F}$ is constrained only by bounds on a derivative, the expected length of any CI that maintains coverage must shrink at the minimax rate for any any f in the interior of $\mathcal{F}$. We extend and sharpen these findings by showing that, for smooth f, this remains true whenever $\mathcal{F}$ is centrosymmetric, even if we don't require the confidence set to take the form of an interval, and, moreover, not only is the rate the same as the minimax rate, the constant must be close to that for fixed-length CIs.

Many procedures popular in practice avoid having to specify C by dropping the requirement that the CI be valid uniformly over $\mathcal{F}$, and only require it to be valid pointwise for each $f \in \mathcal{F}$. For example, under the assumption that f has at least one derivative, one can construct a 95% CI for a conditional mean at a boundary point $x = 0$ by using a kernel estimator with bandwidth that shrinks at a rate slightly faster than $n^{-1/3}$ (i.e. undersmooth relative to the mean-square error optimal bandwidth), and adding and subtracting 1.96 times the standard deviation of this estimator. Even though the estimator is biased in finite samples, so that the CI will undercover in finite samples, for any given f with at least one derivative, the bias is of lower order than the variance, so that asymptotically, it will be negligible and the CI will have pointwise asymptotic coverage equal to 95% under regularity conditions. However, it is clear that, in any given sample, one can make the bias arbitrarily large, and hence the finite-sample coverage arbitrarily close to zero, by setting $f(x) = Cx$ with C sufficiently large. Indeed, any bandwidth choice—including one that makes an "asymptotic promise" to undersmooth—implies a maximum value of C beyond which the coverage of a nominal 95% CI in any given sample will fail to be close (say within 5%) to the nominal coverage. Thus, even if one is willing to accept a given amount of undercoverage, a CI based on undersmoothing involves an implicit choice of C.

One way to address this problem is to allow for more flexible bandwidth sequences as in Calonico, Cattaneo, and Titiunik (2014) and Calonico, Cattaneo, and Farrell (2015). Another approach is to try to estimate an upper bound on the possible bias, as in Schennach (2015). However, our results imply that, in order to achieve good coverage over a range of functions f in a given sample, one cannot avoid having to specify an explicit bound on the smoothness of f. Once this is done, there is very little scope for improving upon a CI that

optimal rates of convergence for nonparametric estimation given in Stone (1980).

uses this a priori smoothness bound to choose the optimal bandwidth and to bound the bias.

Similar problems with CIs that are valid pointwise, but not uniformly, have been pointed out in several parametric and semiparametric models popular in econometrics, including instrumental variables models (Bound, Jaeger, and Baker, 1995; Staiger and Stock, 1997) and moment inequalities (Andrews and Guggenberger, 2009b). They are also central to inference after model selection (Leeb and Pötscher, 2005), as we discuss in detail in the application to linear regression in Section 4. As this literature points out, in any given sample, there will be part of the parameter space where pointwise CIs will severely undercover. In nonparametric settings, however, the problem can be much worse in the sense that the problematic part of the parameter space may be much larger. Brown, Low, and Zhao (1997) give examples of nonparametric estimation problems where every point of the parameter space can be a point of superefficiency, in contrast to parametric estimation problems, where the set of superefficiency has Lebesgue measure zero (see also Chapter 1.2.4 in Tsybakov, 2009). As Robins and van der Vaart (2006) point out, dropping uniformity "appears to contradict the very definition of a confidence set"—to construct a CI for Lf, one must specify some parameter space $\mathcal{F}$ such that the CI covers Lf with the prespecified probability for all f in the parameter space.

Pointwise-in-f asymptotics may lead to other inconsistencies, such as assuming that the covariates are continuous even if they are clearly discrete in the given sample. This has lead to considerable confusion in the regression discontinuity literature, in which very different modeling approaches have been proposed when covariates are discrete instead of continuous (see Lee and Card, 2008). In contrast, in this paper we take a finite-sample approach, and only use asymptotics to relax the normality assumption. When covariates are continuous, additional simplifications obtain: certain sums are approximated by integrals that do not depend on the design points, and the optimal procedures correspond asymptotically to kernel estimators with different bandwidths. However, one need not use these simplifications in forming estimates and CIs: the finite sample approach still leads to CIs that are easily computable and relatively simple, as we illustrate in our regression discontinuity application in Section 5. Thus, one can take the same approach whether the covariates are discrete or continuous and not worry about how to best model them.

The remainder of this paper is organized as follows. In Section 2, we illustrate our results in a simple example. Section 3 introduces the general setup and states the main results. Section 4 applies these results to linear regression. Section 5 considers an application to regression discontinuity. Section 6 concludes. Proofs, long derivations, and additional results

are collected in appendices. Appendix A conducts a Monte Carlo study to illustrate the main results. Appendix B contains proofs for the main results in Section 3, and Appendix C additional details for constructing two-sided CIs studied in that section. The Supplemental Materials contain further appendices. Supplemental Appendix D contains derivations for Sections 4 and 5. Supplemental Appendices E, F, and G contain asymptotic results. Supplemental Appendix H contains additional figures for the application in Section 5.

2 Simple example

To illustrate the main theoretical results, consider the nonparametric regression $y_i = f(x_i) + u_i$, where $i = 1, \dots, n$, the regressors $x_i \in \mathbb{R}$ are treated as fixed and the errors u_i are i.i.d. standard normal. We assume that f lies in a class of Lipschitz continuous functions with constant C,

$$\mathcal{F}_{\text{Lip}}(C) = \{f \colon |f(x_1) - f(x_2)| \leq C|x_1 - x_2|\}. \tag{1}$$

We are interested in inference on the the value of the regression function f at a point, which we can normalize to zero.

Consider first the problem of constructing one-sided confidence intervals (CIs). In particular, consider the problem of constructing CIs $[\hat{c}, \infty)$ that minimize the maximum βth quantile of excess length, $\sup_{f \in \mathcal{F}} q_{f,\beta}(f(0) - \hat{c})$, where $q_{f,\beta}$ denotes the βth quantile of the excess length $f(0) - \hat{c}$. Such CIs can be obtained by inverting tests of the null hypothesis $H_0 \colon f(0) \leq L_0$ that maximize their minimum power under the alternative $H_1 \colon f(0) \geq L_0 + 2b$, where the half-distance b to the alternative is calibrated so that the minimum power of the minimax test is given by β (see Section 3.3 for derivation).

To construct the minimax test, note that if we set $\mu = (f(x_1), \dots, f(x_n))'$, and $Y = (y_1, \dots, y_n)'$, we can view the testing problem as an n-variate normal mean problem $Y \sim N(\mu, I_n)$. The vector of means μ is constrained take values in the convex sets $M_0 = \{(f(x_1), \dots, f(x_n))' \colon f \in \mathcal{F}, f(0) \leq L_0\}$ under the null, and $M_1 = \{(f(x_1), \dots, f(x_n))' \colon f \in \mathcal{F}, f(0) \geq L_0 + 2b\}$ under the alternative. To solve this problem, let's first consider a two-point testing problem with the null and alternative given by some $\mu_0 \in M_0$ and $\mu_1 \in M_1$. By the Neyman-Pearson lemma, the optimal test of μ_0 vs μ_1 is the likelihood ratio test, which rejects for large values of $(\mu_1 - \mu_0)'Y$, and has power $\Phi\left(\|\mu_1 - \mu_0\| - z_{1-\alpha}\right)$ at μ_1. Since this testing problem is easier than testing M_0 against M_1, minimizing this power over $\mu_0 \in M_0$ and $\mu_1 \in M_1$ must give an upper bound for the minimum power of the minimax test. Let us conjecture that the solution to the minimax testing problem is given by the

solution to the two point testing problem with μ_0 and μ_1 given by the minimizers μ_0^* and μ_1^* of $\Phi(\|\mu_1 - \mu_0\| - z_{1-\alpha})$. To verify this conjecture, we need to show that the power of the resulting test is minimized over M_1 at μ_1^*, and it controls size over M_0 (see Theorem 8.1.1 in Lehmann and Romano 2005), in which case μ_0^* and μ_1^* are called "least favorable." In Lemma B.2, which follows directly from Section 2.4.3 in Ingster and Suslina (2003), we show that for convex M_0 and M_1, this is indeed the case.

Since the power $\Phi(\|\mu_1 - \mu_0\| - z_{1-\alpha})$ is increasing in the distance between μ_1 and μ_0, the least favorable functions correspond to the points $\mu_0^* = (f^*(x_1), \ldots, f^*(x_n))'$ and $\mu_1^* = (g^*(x_1), \ldots, g^*(x_n))'$ that minimize the Euclidean distance between the sets M_0 and M_1,

$$(f^*, g^*) = \underset{f,g \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \qquad \text{subject to } f(0) \leq L_0,\ g(0) \geq L_0 + 2b. \tag{2}$$

To satisfy the constraints, the solution must satisfy $g^*(x) \geq L_0 + 2b - C|x|$ and $f^*(x) \leq L_0 + C|x|$ for all x. Therefore, the difference between the two functions is bounded by $|g^*(x) - f^*(x)| \geq 2\max\{b - C|x|, 0\}$. Since we can make the bound sharp by setting

$$g^*(x) = L_0 + b + \max\{b - C|x|, 0\}, \qquad f^*(x) = L_0 + b - \max\{b - C|x|, 0\},$$

these functions must solve (2). The first panel of Figure 1 shows the least favorable functions. Intuitively, to make H_0 and H_1 hardest to distinguish, the null and alternative functions f^* and g^* converge to each other as fast as possible under the Lipschitz constraint and the null and alternative constraints $f^*(0) \leq L_0$ and $g^*(0) \geq L_0 + 2b$.

The likelihood ratio test that corresponds to the two-point test based on the least favorable means rejects for large values of $Y'(\mu_1^* - \mu_0^*)$, with critical value given by the distribution of Y under μ_0^*. By working out this critical value and rearranging the resulting expression, we obtain the minimax test that rejects whenever

$$\hat{L}_h - L_0 - \text{bias}_{f^*}(\hat{L}_h) \geq \text{var}(\hat{L}_h)^{1/2} z_{1-\alpha}. \tag{3}$$

Here $\hat{L}_h$ is a Nadaraya-Watson kernel estimator based on the triangular kernel $k(u) = \max\{0, 1 - |x|\}$ and bandwidth $h = b/C$,

$$\hat{L}_h = \frac{\sum_{i=1}^n (g^*(x_i) - f^*(x_i)) y_i}{\sum_{i=1}^n (g^*(x_i) - f^*(x_i))} = \frac{\sum_{i=1}^n k(x_i/h) y_i}{\sum_{i=1}^n k(x_i/h)},$$

$\text{var}(\hat{L}_h) = \frac{\sum_{i=1}^n k(x_i/h)^2}{(\sum_{i=1}^n k(x_i/h))^2}$ is the variance of $\hat{L}_h$, $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution, and $\text{bias}_{f^*}(\hat{L}_h) = b\left(1 - \frac{\sum_{i=1}^n k(x_i/h)^2}{\sum_{i=1}^n k(x_i/h)}\right)$ is the bias of the estimator $\hat{L}_h$ under f^*. The estimator $\hat{L}_h$ is normally distributed with variance that does not depend on the true function f. Its bias, however, does depend on f. To control size under H_0 in finite samples, it is necessary to subtract the largest possible bias of $\hat{L}_h$ under the null, which obtains at f^* (we show in the next section that this is in fact the largest bias over all of $\mathcal{F}_{\text{Lip}}(C)$). Since the rejection probability of the test is decreasing in the bias, its minimum power occurs when the bias is minimal under H_1, which occurs at g^*, and is given by

$$\beta = \Phi\Big(2b\sqrt{\textstyle\sum_{i=1}^n k(x_i/h)^2} - z_{1-\alpha}\Big). \tag{4}$$

Since the estimator, its variance, and the non-random bias correction are all independent of the particular null L_0, the CI based on inverting these tests as H_0 varies over $\mathbb{R}$ is given by

$$[\hat{c}_{\alpha,b}, \infty), \quad \text{where} \quad \hat{c}_{\alpha,b} = \hat{L}_h - \text{bias}_{f^*}(\hat{L}_h) - \text{sd}(\hat{L}_h) z_{1-\alpha}. \tag{5}$$

This CI minimizes the βth quantile maximum excess length with β given by the minimax power of the tests (4). Equivalently, given a quantile β that we wish to optimize, set the half-distance to the alternative b_β as the solution to $b_\beta = (z_\beta + z_{1-\alpha})/\sqrt{4\sum_{i=1}^n k(x_i/(C/b_\beta))^2}$.

This solution has four important features. First, it is simple to construct. Second, different choices of the constants C and b (or β) affect the optimal bandwidth, but not the kernel—the triangular kernel is therefore minimax optimal for the Lipschitz class (see Armstrong and Kolesár (2016) and references therein for general results on optimal kernels in these settings). Third, the least favorable functions, g^* and f^*, correspond to scaled versions of this optimal kernel—the least favorable functions and the kernel have the same shape. Fourth, the bias correction is non-random, depends on the worst-case bias of $\hat{L}_h$ (rather than an estimate of its bias), and doesn't disappear asymptotically. In particular, suppose that for some d, $\frac{1}{nh}\sum_{i=1}^h k(x_i/h)^2 \to d\int k(u)^2 du = \frac{2}{3}d$ and $\frac{1}{nh}\sum_{i=1}^h k(x_i/h) \to d$ as $n \to 0$, $nh \to \infty$ and $h \to 0$ (under random sampling of the regressors x_i, this holds with d corresponding to the density of x_i at 0). Let h_β denote the bandwidth that is optimal for the β quantile. Then the worst case bias of $\hat{L}_{h_\beta}$ equals $Ch_\beta/3(1+o(1))$, while its variance equals $\frac{2}{3nh_\beta f_x(0)}(1+o(1))$, with the optimal bandwidth given by

$$h_\beta = \frac{1}{2}\left(\frac{3}{C^2 n f_x(0)}\right)^{1/3}(z_\beta + z_{1-\alpha})^{2/3}(1+o(1)), \tag{6}$$

so that the squared bias and variance are of the same order, $O(n^{-2/3})$. Consequently, no CI that "undersmooths" in the sense that it is based on an estimator whose bias is of lower order than its variance can be minimax optimal asymptotically or in finite samples.

An apparent disadvantage of this CI is that it requires the researcher to choose a smoothing constant C. Addressing this issue leads to "adaptive" CIs. Adaptive CIs achieve good excess length properties for a range of parameter spaces $\mathcal{F}_{\text{Lip}}(C_j)$, $C_1 < \cdots < C_J$, while maintaining coverage over their union, which is given by $\mathcal{F}_{\text{Lip}}(C_J)$, where C_J is a conservative upper bound on the possible smoothness of f. In contrast, a minimax CI only considers worst-case excess length over $\mathcal{F}_{\text{Lip}}(C_J)$. To derive an upper bound on the scope for adaptivity, consider the problem of finding a CI that optimizes excess length over $\mathcal{F}_{\text{Lip}}(0)$ (the space of constant functions), while maintaining coverage over $\mathcal{F}_{\text{Lip}}(C)$ for some $C > 0$.

To derive the form of such CI, consider the one-sided testing problem $H_0\colon f(0) \leq L_0$ and $f \in \mathcal{F}_{\text{Lip}}(C)$ against the one-sided alternative $H_1\colon f(0) \geq L_0 + b$ and $f \in \mathcal{F}_{\text{Lip}}(0)$ (so that now the half-distance to the alternative is given by $b/2$ rather than b). This is equivalent to a multivariate normal mean problem $Y \sim N(\mu, I_n)$, with $\mu \in M_0$ under the null as before, and $\mu \in \tilde{M}_1 = \{(L, \ldots, L)\colon L \geq L_0 + b\}$. Since the null and alternative are convex, by the same arguments as before, the least favorable functions minimize the Euclidean distance between the two sets. The minimizing functions are given by $\tilde{g}^*(x) = L_0 + b$, and $\tilde{f}^* = f^*$ (same function as before). The second panel of Figure 1 plots this solution. Since $\tilde{g}^* - \tilde{f}^* = (g^* - f^*)/2$, the resulting test is again given by (3), and the CI is also the same as before—the only difference is that we moved the half-distance to the alternative from b to $b/2$. Hence, the minimax CI that optimizes a given quantile of excess length over $\mathcal{F}_{\text{Lip}}(C)$ also optimizes its excess length over the space of constant functions, but at a different quantile. By calculating the power of the minimax test at constant alternatives, it can be seen that the scope for improvement is still small if one compares excess length at the same quantile: in Section 3.3, we show that, for this smoothness class, the CI that minimaxes excess length at a given quantile is at least 95.2% optimal asymptotically for constant functions at the same quantile. For function classes smoother than $\mathcal{F}_{\text{Lip}}(C)$, the efficiency is even higher.

Therefore, it is not possible to "adapt" to cases in which the regression function is smoother than the least favorable function.

A two-sided CI based on $\hat{L}_h$ could be formed by adding and subtracting $\text{bias}_{f^*}(\hat{L}_h) + \text{sd}(\hat{L}_h)z_{1-\alpha/2}$, thereby accounting for possible bias on either side. However, this is conservative, since the bias cannot be in both directions at once. Since $(\hat{L}_h - Lf)/\text{sd}(\hat{L}_h)$ follows a normal distribution with variance one and bias ranging from $-\text{bias}_{f^*}(\hat{L}_h)/\text{sd}(\hat{L}_h)$ to

$\text{bias}_{f^*}(\hat{L}_h)/\,\text{sd}(\hat{L}_h)$, a nonconservative CI takes the form $\hat{L}_h \pm \text{sd}(\hat{L}_h)\,\text{cv}_\alpha(\text{bias}_{f^*}(\hat{L}_h)/\,\text{sd}(\hat{L}_h))$, where $\text{cv}_\alpha(t)$ is the $1-\alpha$ quantile of the absolute value of a $N(t,1)$ distribution. This corresponds to a fixed-length CI, as defined in Donoho (1994). The optimal choice of h for a fixed-length CI simply minimizes $\text{sd}(\hat{L}_h)\,\text{cv}_\alpha(\text{bias}_{f^*}(\hat{L}_h)/\,\text{sd}(\hat{L}_h))$ (since the length of the CI is nonrandom, minimizing it does not invalidate the CI). It follows from results in Donoho (1994) that the fixed-length CI centered at the optimal $\hat{L}_h$ is in fact optimal among all fixed-length CIs centered at affine functions of the y_is, and is close to optimal among fixed-length CIs centered at any estimate.

The restriction to fixed-length CIs rules out adaptivity: the length of the CI must always reflect the worst possible bias of the estimator. In Section 3.4 we derive a sharp efficiency bound that shows that, similar to the one-sided case, these CIs are nonetheless highly efficient relative to variable-length CIs that optimize their length at smooth functions.

The key to these non-adaptivity results is that the class $\mathcal{F}$ is centrosymmetric (i.e. $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$) and convex. The centrosymmetry implies that the least favorable functions in the minimax problem (2) are, up to constants, negatives of one another, and the convexity is necessary for Lemma B.2 to apply. For adaptivity to be possible, we need shape restrictions like monotonicity, or non-convexity of $\mathcal{F}$. In the next section, we give general statements of these results.

3 General characterization of optimal procedures

We consider the following setup and notation, much of which follows Donoho (1994). We observe data Y of the form

$$Y = Kf + \sigma\varepsilon \tag{7}$$

where f is known to lie in a convex subset $\mathcal{F}$ of a vector space, and $K : \mathcal{F} \to \mathcal{Y}$ is a linear operator between $\mathcal{F}$ and a Hilbert space $\mathcal{Y}$. We use $\langle \cdot, \cdot \rangle$ to denote the inner product on $\mathcal{Y}$ and $\| \cdot \|$ to denote the norm. The error term ε is standard Gaussian with respect to this inner product: for any $g \in \mathcal{Y}$, $\langle \varepsilon, g \rangle$ is normal with $E\langle \varepsilon, g \rangle = 0$ and $\text{var}\,(\langle \varepsilon, g \rangle) = \|g\|^2$. We are interested in constructing a confidence set for a linear functional Lf.

3.1 Special cases

The general setup (7) covers a number of important models as special cases. First, it can be used to study Gaussian nonparametric regression with fixed design, in which we observe

$\{x_i, y_i\}_{i=1}^n$ with x_i a deterministic vector, and

$$y_i = f(x_i) + u_i, \quad u_i \sim N(0, \sigma^2(x_i)) \text{ independent across } i, \tag{8}$$

where $\sigma^2(x)$ is known. Here $Y = (y_1/\sigma(x_1), \ldots, y_n/\sigma(x_n))'$, $\mathcal{Y} = \mathbb{R}^n$, $Kf = (f(x_1)/\sigma(x_1), \ldots, f(x_n)/\sigma(x_n))'$ and with $\langle x, y\rangle$ given by the Euclidean inner product $x'y$. Depending on the definition of the linear functional L, this model covers several important situations encountered in applied econometrics, including: inference at a point, regression discontinuity (see Section 5), and average treatment effects under unconfoundedness (with $Lf = \frac{1}{n}\sum_{i=1}^n (f(w_i, 1) - f(w_i, 0))$ where $x_i = (w_i, d_i')'$, d_i is a treatment indicator and w_i are controls). The finite sample results in this model will often lead to analogous uniform (over $\mathcal{F}$) asymptotic results in the more realistic setting in which the distribution of u_i is not known (see Section 3.6).

Second, the setup (7) can be used to study the linear regression model with restricted parameter space. For simplicity, we consider the case with homoskedastic errors

$$Y = X\theta + \sigma\varepsilon, \quad \varepsilon \sim N(0, I_n), \tag{9}$$

where X is a fixed $n \times k$ design matrix and σ is known. This fits into our framework with $f = \theta$, X playing the role of K, taking $\theta \in \mathbb{R}^k$ to $X\theta \in \mathbb{R}^n$, and $\mathcal{Y} = \mathbb{R}^n$ with the Euclidean inner product $\langle x, y\rangle = x'y$. We are interested in a linear functional $L\theta = \ell'\theta$ where $\ell \in \mathbb{R}^k$. We consider this model in Section 4. While we focus on homoskedastic linear regression for exposition, the results extend to the multivariate normal location model $\hat{\theta} \sim N(\theta, \Sigma_\theta)$, which obtains as a limiting experiment of regular parametric models. Thus, the finite sample results for OLS could be extended to local asymptotic results for other regular parametric models, with the constraint sets $\mathcal{F}$ and $\mathcal{G}$ (defined below) shrinking at a $\sqrt{n}$ rate.

In addition to the regression models (8) and (9), the setup (7) includes other nonparametric and semiparametric regression models such as the partly linear model (where f takes the form $g(w_1) + \gamma' w_2$, and we are interested in a linear functional of g or γ). It also includes the Gaussian white noise model, which can be obtained as a limiting model for nonparametric density estimation (see Nussbaum, 1996) as well as nonparametric regression (see Brown and Low, 1996). We refer the reader to Donoho (1994, Section 9) for details of these and other models that fit into the general setup (7).

3.2 Performance criteria and a class of estimators

Let us now define the performance criteria that we use to evaluate confidence sets for Lf. Following the usual definition, a set $\mathcal{C} = \mathcal{C}(Y)$ is a $100 \cdot (1-\alpha)\%$ confidence set for Lf if

$$\inf_{f \in \mathcal{F}} P_f \left(Lf \in \mathcal{C} \right) \geq 1 - \alpha. \tag{10}$$

We denote the collection of all confidence sets $\mathcal{C}$ that satisfy (10) by $\mathcal{I}_\alpha$. Among confidence sets in this collection, we can compare their performance at a particular $f \in \mathcal{F}$ using expected length,

$$\Lambda_f(\mathcal{C}) = E_f \lambda(\mathcal{C}(Y)),$$

where λ is Lebesgue measure.

Allowing confidence sets to have arbitrary form can lead to sets $\mathcal{C}$ that are complicated and difficult to interpret or even compute. One way of avoiding this is to restrict attention to sets in $\mathcal{I}_\alpha$ that take the form of a fixed-length confidence interval (CI). A fixed-length CI takes the form $[\hat{L} - \chi, \hat{L} + \chi]$ for some estimate $\hat{L}$ and some nonrandom χ (for instance, in the regression model (8), χ may depend on the regressors x_i and $\sigma^2(x_i)$, but not on y_i). For an estimator $\hat{L}$, let

$$\chi_\alpha(\hat{L}) = \min \left\{ \chi \colon \inf_{f \in \mathcal{F}} P_f\big(|\hat{L} - Lf| \leq \chi \big) \geq 1 - \alpha \right\}$$

denote the half-length of the shortest fixed-length $100 \cdot (1-\alpha)\%$ CI centered at $\hat{L}$.

The restriction to fixed-length CIs simplifies their comparison: for any $f \in \mathcal{F}$, the expected length equals $2\chi_\alpha(\hat{L})$, so among fixed-length CIs, one simply prefers those with smaller half-length. On the other hand, one may worry that fixed-length CIs may be costly since the length cannot "adapt" to reflect greater precision for different functions $f \in \mathcal{F}$. To address this concern, in Section 3.4, we compare the length of fixed-length CIs to sharp bounds on the optimal expected length $\inf_{\mathcal{C} \in \mathcal{I}_\alpha} \Lambda_f(\mathcal{C})$.

If $\mathcal{C}$ is restricted to take the form of a one-sided confidence interval (CI) $[\hat{c}, \infty)$, we cannot use expected length as a criterion. We can, however, compare performance at a particular parameter f using the βth quantile of excess length,

$$q_{f,\beta}(Lf - \hat{c}),$$

where $q_{f,\beta}(Lf - \hat{c})$ denotes the βth quantile of $Lf - \hat{c}$, the excess length, under f. To measure

performance globally over some set $\mathcal{G}$, we use the maximum βth quantile of the excess length,

$$q_\beta(\hat{c}, \mathcal{G}) = \sup_{g \in \mathcal{G}} q_{g,\beta}(Lg - \hat{c}). \tag{11}$$

If $\mathcal{G} = \mathcal{F}$, minimizing $q_\beta(\hat{c}, \mathcal{F})$ over one-sided CIs that satisfy (10) gives minimax excess length. If $\mathcal{G} \subset \mathcal{F}$ is a class of smoother functions, minimizing $q_\beta(\hat{c}, \mathcal{G})$ yields CIs that direct power: they achieve good performance when f is smooth, while maintaining coverage over all of $\mathcal{F}$. A CI that achieves good performance over multiple classes $\mathcal{G}$ is said to be "adaptive" over these classes. In Section 3.3, we give sharp bounds on (11) for a single class $\mathcal{G}$, which gives a benchmark for adapting over multiple classes (cf. Cai and Low, 2004a).

We will also relate the optimal decision rules for constructing CIs to the rules for constructing estimators that minimize the maximum mean squared error (MSE) over $\mathcal{F}$. For an estimator $\hat{L}$, the maximum mean squared error over $\mathcal{F}$ is defined as

$$R(\hat{L}) = \sup_{f \in \mathcal{F}} E_f(\hat{L} - Lf)^2.$$

The main tool in deriving decision rules that are optimal or close to optimal for these performance criteria will be the ordered modulus of continuity between $\mathcal{F}$ and $\mathcal{G}$, defined by Cai and Low (2004a)

$$\omega(\delta; \mathcal{F}, \mathcal{G}) = \sup \{Lg - Lf \colon \|K(g - f)\| \leq \delta, f \in \mathcal{F}, g \in \mathcal{G}\}$$

for any sets $\mathcal{F}$ and $\mathcal{G}$ with a non-empty intersection (so that the set over which the supremum is taken is non-empty). When $\mathcal{G} = \mathcal{F}$, $\omega(\delta; \mathcal{F}, \mathcal{F})$ is the (single-class) modulus of continuity over $\mathcal{F}$ (Donoho and Liu, 1991), and we will denote it by $\omega(\delta; \mathcal{F})$. The ordered modulus $\omega(\cdot; \mathcal{F}, \mathcal{G})$ is concave, which implies that the superdifferential at δ (the set of slopes of tangent lines at $(\delta, \omega(\delta; \mathcal{F}, \mathcal{G}))$ is nonempty for any $\delta > 0$. Throughout the paper, we let $\omega'(\delta; \mathcal{F}, \mathcal{G})$ denote an (arbitrary unless otherwise stated) element in this set. Typically, $\omega(\cdot; \mathcal{F}, \mathcal{G})$ is differentiable, in which case $\omega'(\delta; \mathcal{F}, \mathcal{G})$ is defined uniquely as the derivative at δ. We use $g^*_{\delta,\mathcal{F},\mathcal{G}}$ and $f^*_{\delta,\mathcal{F},\mathcal{G}}$ to denote a solution to the ordered modulus problem (assuming it exists), and $f^*_{M,\delta,\mathcal{F},\mathcal{G}} = (f^*_{\delta,\mathcal{F},\mathcal{G}} + g^*_{\delta,\mathcal{F},\mathcal{G}})/2$ to denote the midpoint.

We will show that optimal decision rules will in general depend on the data Y through an estimator of the form

$$\hat{L}_{\delta,\mathcal{F},\mathcal{G}} = Lf^*_{M,\delta,\mathcal{F},\mathcal{G}} + \frac{\omega'(\delta; \mathcal{F}, \mathcal{G})}{\delta} \left\langle K(g^*_{\delta,\mathcal{F},\mathcal{G}} - f^*_{\delta,\mathcal{F},\mathcal{G}}), Y - Kf^*_{M,\delta,\mathcal{F},\mathcal{G}} \right\rangle, \tag{12}$$

with δ and $\mathcal{G}$ depending on the optimality criterion. When $\mathcal{F} = \mathcal{G}$, we denote the estimator $\hat{L}_{\delta,\mathcal{F},\mathcal{F}}$ by $\hat{L}_{\delta,\mathcal{F}}$. When the sets $\mathcal{F}$ and $\mathcal{G}$ are clear from the context, we use $\omega(\delta)$, $\hat{L}_\delta$, f^*_δ, g^*_δ and $f^*_{M,\delta}$ in place of $\omega(\delta;\mathcal{F},\mathcal{G})$, $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$, $f^*_{\delta,\mathcal{F},\mathcal{G}}$, $g^*_{\delta,\mathcal{F},\mathcal{G}}$ and $f^*_{M,\delta,\mathcal{F},\mathcal{G}}$ to avoid notational clutter.

Let $\overline{\text{bias}}_\mathcal{G}(\hat{L}) = \sup_{f\in\mathcal{G}} E_f(\hat{L}-Lf)$ and $\underline{\text{bias}}_\mathcal{G}(\hat{L}) = \inf_{f\in\mathcal{G}} E_f(\hat{L}-Lf)$ denote the maximum and minimum bias of an estimator $\hat{L}$ over the set $\mathcal{G}$. As we show in Lemma B.1 in the Appendix, a useful property of $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ is that its maximum bias over $\mathcal{F}$ and minimum bias over $\mathcal{G}$ are attained at f^*_δ and g^*_δ, respectively, and are given by

$$\overline{\text{bias}}_\mathcal{F}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}}) = -\underline{\text{bias}}_\mathcal{G}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}}) = \frac{1}{2}\left(\omega(\delta;\mathcal{F},\mathcal{G}) - \delta\omega'(\delta;\mathcal{F},\mathcal{G})\right). \tag{13}$$

As remarked by Cai and Low (2004b), no estimator can simultaneously achieve lower maximum bias over $\mathcal{F}$, higher minimum bias over $\mathcal{G}$, and lower variance (which for $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ doesn't depend on f) than the estimators in the class $\{\hat{L}_{\delta,\mathcal{F},\mathcal{G}}\}_{\delta>0}$. Estimators (12) can thus be used to optimally trade off various levels of bias and variance.

Let us briefly discuss two symmetry properties that lead to simplifications when satisfied by $\mathcal{F}$. The first we call translation invariance.

Definition 1 (Translation Invariance)**.** *The function class $\mathcal{F}$ is translation invariant if there exists a function $\iota \in \mathcal{F}$ such that $L\iota = 1$ and $f + c\iota \in \mathcal{F}$ for all $c \in \mathbb{R}$ and $f \in \mathcal{F}$.*

Translation invariance will hold in most cases where the parameter of interest Lf is unrestricted. For example, if $Lf = f(0)$, it will hold with $\iota(x) = 1$ if $\mathcal{F}$ places monotonicity restrictions and/or restrictions on the derivatives of f, but not if $\mathcal{F}$ places a bound on the function itself. Under translation invariance, by Lemma B.3 in the Appendix, the modulus is differentiable and $\omega'(\delta;\mathcal{F},\mathcal{G}) = \delta/\langle K\iota, K(g^*_\delta - f^*_\delta)\rangle$, which gives

$$\hat{L}_{\delta,\mathcal{F},\mathcal{G}} = Lf^*_{M,\delta} + \frac{\langle K(g^*_\delta - f^*_\delta), Y - Kf^*_{M,\delta}\rangle}{\langle K(g^*_\delta - f^*_\delta), K\iota\rangle}.$$

The second property we consider is centrosymmetry.

Definition 2 (Centrosymmetry)**.** *The function class $\mathcal{F}$ is centrosymmetric if $f \in \mathcal{F} \implies -f \in \mathcal{F}$.*

Under centrosymmetry, the functions that solve the single-class modulus problem can be seen to satisfy $g^*_\delta = -f^*_\delta$, and the modulus is given by

$$\omega(\delta;\mathcal{F}) = \sup\left\{2Lf \colon \|Kf\| \le \delta/2, f \in \mathcal{F}\right\}. \tag{14}$$

Since $f^*_\delta = -g^*_\delta$, $f^*_{M,\delta}$ is the zero function and $\hat{L}_{\delta,\mathcal{F}}$ is linear:

$$\hat{L}_{\delta,\mathcal{F}} = \frac{2\omega'(\delta;\mathcal{F})}{\delta}\langle Kg^*_\delta, Y\rangle = \frac{\langle Kg^*_\delta, Y\rangle}{\langle Kg^*_\delta, K\iota\rangle}, \tag{15}$$

where the last equality holds when $\mathcal{F}$ is translation invariant as well as centrosymmetric.

Centrosymmetry and translation invariance are not needed for most of the results in this paper. However, centrosymmetry will play central role in bounding the gains from directing power at smooth functions, as we show in Section 3.3 for one-sided CIs and in Section 3.4 for two-sided CIs.

3.3 Optimal one-sided CIs

Given β, a one-sided CI that minimizes (11) among all one-sided CIs with level $1-\alpha$ is based on $\hat{L}_{\delta_\beta;\mathcal{F},\mathcal{G}}$ where $\delta_\beta = \sigma(z_\beta + z_{1-\alpha})$ and z_q denotes the qth quantile of a standard normal distribution. The CI takes a simple form, which is given in the following theorem. Proofs of the results in this section are given in Appendix B.

Theorem 3.1. *Let $\mathcal{F}$ and $\mathcal{G}$ be convex with $\mathcal{G} \subseteq \mathcal{F}$, and suppose that f^*_δ and g^*_δ achieve the ordered modulus at δ with $\|K(f^*_\delta - g^*_\delta)\| = \delta$. Let*

$$\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}} = \hat{L}_{\delta,\mathcal{F},\mathcal{G}} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}}) - z_{1-\alpha}\sigma\omega'(\delta;\mathcal{F},\mathcal{G}).$$

Then, for $\beta = \Phi(\delta/\sigma - z_{1-\alpha})$, $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}$ minimizes $q_\beta(\hat{c},\mathcal{G})$ among all one-sided $1-\alpha$ CIs, where Φ denotes the standard normal cdf. The minimum coverage is taken at f^*_δ and equals $1-\alpha$. All quantiles of excess length are maximized at g^*_δ. The worst case βth quantile of excess length is $q_\beta(\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}},\mathcal{G}) = \omega(\delta;\mathcal{F},\mathcal{G})$.

The assumption that the modulus is achieved with $\|K(f^*_\delta - g^*_\delta)\| = \delta$ rules out degenerate cases: if $\|K(f^*_\delta - g^*_\delta)\| < \delta$, then relaxing this constraint does not increase the modulus, which means that $\omega'(\delta;\mathcal{F},\mathcal{G}) = 0$ and the optimal CI does not depend on the data.

The estimator $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ is normally distributed with bias that depends on f, and variance $\sigma^2\omega'(\delta;\mathcal{F},\mathcal{G})^2$, which is independent of f. The CI in Theorem 3.1 uses the fact that the maximum bias over $\mathcal{F}$ and minimum bias over $\mathcal{G}$ are taken at f^*_δ and g^*_δ. Since the coverage of a one-sided CI decreases with the bias of the estimator that it is based on, to ensure proper coverage, we need to subtract $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}})$, the maximum bias under $\mathcal{F}$, from $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$, and then subtract the $1-\alpha$ quantile of the of a mean zero normal variable with the same

variance as $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$. On the other hand, all quantiles of excess length decrease with the bias: they are greatest when the bias is minimal, which gives the second part of the theorem.

One can't do better than using $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}$ because the test that rejects L_0 when $L_0 \notin [\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}, \infty)$ is minimax for $H_0 : Lf \leq L_0$ and $f \in \mathcal{F}$ against $H_1 : Lf \geq L_0 + \omega(\delta; \mathcal{F}, \mathcal{G})$ and $f \in \mathcal{G}$, where $L_0 = Lf^*_\delta$. If both $\mathcal{F}$ and $\mathcal{G}$ are translation invariant, $f^*_\delta + c\iota$ and $g^*_\delta + c\iota$ achieve the ordered modulus for any $c \in \mathbb{R}$, so that, varying c, this test can be seen to be minimax for any L_0. Thus, under translation invariance, the CI in Theorem 3.1 inverts minimax one sided tests with distance to the null given by $\omega(\delta)$. These results for minimax tests can be derived from an application of a result characterizing minimax tests as Neyman-Pearson tests for mixtures over least favorable distributions over the null and alternative (Theorem 8.1.1 in Lehmann and Romano, 2005), where the least favorable null and alternative are given by point masses at f^*_δ and g^*_δ (see Lemma B.2 in the Appendix and Section 2.4.3 in Ingster and Suslina, 2003).

Given the model (7), implementing the CI from Theorem 3.1 requires the researcher to choose a quantile β to optimize, and to choose the set $\mathcal{G}$. There are two natural choices for β. If the objective is to optimize the performance of the CI "on average", then optimizing the median excess length ($\beta = 0.5$) is a natural choice. Since for any CI $[\hat{c}, \infty)$ that is an affine function of the data Y, the median and expected excess lengths coincide, and since $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}$ is affine in the data, setting $\beta = 0$ also has the advantage that it minimizes the expected excess length among CIs that are affine. Alternatively, if the CI is being computed as part of a power analysis, then setting $\beta = 0.8$ is natural, as under translation invariance, it translates directly to statements about 80% power, a standard benchmark in such analyses (Cohen, 1988).

For the set $\mathcal{G}$, there are two leading choices. First, setting $\mathcal{G} = \mathcal{F}$ yields minimax CIs:

Corollary 3.1 (One-sided minimax CIs)**.** *Let $\mathcal{F}$ be convex, and suppose that f^*_δ and g^*_δ achieve the single-class modulus at δ with $\|K(f^*_\delta - g^*_\delta)\| = \delta$. Let*

$$\hat{c}_{\alpha,\delta,\mathcal{F}} = \hat{L}_{\delta,\mathcal{F}} - \frac{1}{2}\left(\omega(\delta;\mathcal{F}) - \delta\omega'(\delta;\mathcal{F})\right) - z_{1-\alpha}\sigma\omega'(\delta;\mathcal{F}).$$

Then, for $\beta = \Phi(\delta/\sigma - z_{1-\alpha})$, $\hat{c}_{\alpha,\delta,\mathcal{F}}$ minimizes the maximum βth quantile of excess length among all $1-\alpha$ CIs for Lf. The minimax excess length is given by $\omega(\delta;\mathcal{F})$.

The minimax criterion may be considered overly pessimistic: it focuses on controlling the excess length under the least favorable function. This leads to the second possible choice for $\mathcal{G}$: set it to a smaller convex class of smoother functions $\mathcal{G} \subset \mathcal{F}$. The resulting CIs will then

achieve the best possible performance when f is smooth, while maintaining coverage over all of $\mathcal{F}$.

It is instructive to consider the case in which $\mathcal{F}$ is centrosymmetric, and the solution to the ordered modulus problem satisfies

$$f - g^*_{\delta,\mathcal{F},\mathcal{G}} \in \mathcal{F} \quad \text{for all } f \in \mathcal{F}. \tag{16}$$

This will be satisfied if $g^*_{\delta,\mathcal{F},\mathcal{G}}$ is "smooth" enough. If $\mathcal{F}$ is translation invariant, then (16) holds for $\mathcal{G} = \text{span}(\iota)$. If $\mathcal{F}$ places a bound on the pth derivative of f (e.g. $\mathcal{F}$ is a Hölder class) it holds if all $g \in \mathcal{G}$ are polynomials of order $p-1$ or lower: the pth derivative of any $g \in \mathcal{G}$ is always zero, so that if f satisfies the particular bound, so does $f - g$.

Under condition (16), if $f^*_{\delta,\mathcal{F},\mathcal{G}}$ and $g^*_{\delta,\mathcal{F},\mathcal{G}}$ solve the modulus problem $\omega(\delta, \mathcal{F}, \mathcal{G})$, then $f^*_{\delta,\mathcal{F},\mathcal{G}} - g^*_{\delta,\mathcal{F},\mathcal{G}}$ and 0 (the zero function) solve $\omega(\delta; \mathcal{F}, \{0\})$ and vice versa (note that, under centrosymmetry, Equation (16) holds for $g^*_{\delta,\mathcal{F},\mathcal{G}}$ iff. it holds for $-g^*_{\delta,\mathcal{F},\mathcal{G}}$), so that

$$\omega(\delta; \mathcal{F}, \mathcal{G}) = \omega(\delta; \mathcal{F}, \{0\}) = \sup\{-Lf \colon \|Kf\| \leq \delta, f \in \mathcal{F}\} = \frac{1}{2}\omega(2\delta; \mathcal{F}), \tag{17}$$

where the last equality obtains because under centrosymmetry, maximizing $-Lf = L(-f)$ and maximizing Lf are equivalent, so that the maximization problem is equivalent to (14). Furthermore, $g^*_{\delta,\mathcal{F},\mathcal{G}} - f^*_{\delta,\mathcal{F},\mathcal{G}} = \frac{1}{2}(g^*_{2\delta,\mathcal{F}} - f^*_{2\delta,\mathcal{F}})$, so that

$$\begin{aligned} \hat{L}_{\delta,\mathcal{F},\mathcal{G}} &= \hat{L}_{2\delta,\mathcal{F}} + Lf^*_{M,\delta,\mathcal{F},\mathcal{G}} - \frac{\omega'(2\delta;\mathcal{F})}{2\delta}\left\langle K(g^*_{2\delta,\mathcal{F}} - f^*_{2\delta,\mathcal{F}}), Kf^*_{M,\delta,\mathcal{F},\mathcal{G}}\right\rangle \\ &= \hat{L}_{2\delta,\mathcal{F}} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{2\delta,\mathcal{F}})/2, \end{aligned} \tag{18}$$

where the second line follows since $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}}) = \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{2\delta,\mathcal{F}})/2$ by (17). Since $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ and $\hat{L}_{2\delta,\mathcal{F}}$ are equal up to a constant, $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}} = \hat{c}_{\alpha,\delta,\mathcal{F},\{0\}} = \hat{c}_{\alpha,2\delta,\mathcal{F}}$. Thus, when (16) holds, optimizing excess length over $\mathcal{G}$ is equivalent to optimizing excess length at $\{0\}$, and it leads to the same class of CIs as the minimax criterion—the only difference is that the excess length is calibrated differently:

Corollary 3.2. *Let $\delta_\beta = \sigma(z_\beta + z_{1-\alpha})$. Let $\mathcal{F}$ be centrosymmetric, and let $\mathcal{G} \subseteq \mathcal{F}$ be any convex set such that the solution to the ordered modulus problem exists and satisfies* (16) *with $\|K(f^*_{\delta_\beta} - g^*_{\delta_\beta})\| = \delta_\beta$. Then the one-sided CI $\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}$ that is minimax for the βth quantile also optimizes $q_{\tilde{\beta}}(\hat{c}; \mathcal{G})$, where $\tilde{\beta} = \Phi((z_\beta - z_{1-\alpha})/2)$. In particular, $\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}$ optimizes $q_{\tilde{\beta}}(\hat{c}; \{0\})$.*

Moreover, the efficiency of $\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}$ for the βth quantile of maximum excess length over $\mathcal{G}$

is given by

$$\frac{\inf_{\hat{c}:\,[\hat{c},\infty)\in\mathcal{I}_\alpha} q_\beta(\hat{c},\mathcal{G})}{q_\beta(\hat{c}_{\alpha,\delta_\beta,\mathcal{F}},\mathcal{G})} = \frac{\omega(\delta_\beta;\mathcal{F},\mathcal{G})}{q_\beta(\hat{c}_{\alpha,\delta_\beta,\mathcal{F}},\mathcal{G})} = \frac{\omega(2\delta_\beta;\mathcal{F})}{\omega(\delta_\beta;\mathcal{F})+\delta_\beta\omega'(\delta_\beta;\mathcal{F})}. \tag{19}$$

The second part of the Corollary follows since by (18), $\underline{\text{bias}}_{\mathcal{G}}(\hat{L}_{\delta,\mathcal{F}}) = 0$, which implies $q_\beta(\hat{c}_{\alpha,\delta_\beta,\mathcal{F}},\mathcal{G}) = (\omega(\delta_\beta;\mathcal{F})+\delta_\beta\omega'(\delta_\beta;\mathcal{F}))/2$.

The first part of Corollary 3.2 states that minimax CIs that optimize a particular quantile β will also minimize the maximum excess length over $\mathcal{G}$ at a different quantile $\tilde{\beta}$. For instance, a CI that is minimax for median excess length among 95% CIs also optimizes $\Phi(-z_{0.95}/2) \approx 0.205$ quantile under the zero function. Vice versa, the CI that optimizes median excess length under the zero function is minimax for the $\Phi(2z_{0.5}+z_{0.95}) = 0.95$ quantile.

The second part of Corollary 3.2 gives the exact cost of optimizing the "wrong" quantile $\tilde{\beta}$. Since the one-class modulus is concave, $\delta\omega'(\delta) \leq \omega(\delta)$, and we can lower bound the efficiency of $\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}$ given in (19) by $\omega(2\delta_\beta)/(2\omega(\delta_\beta)) \geq 1/2$. Typically, however, the efficiency is much higher. In particular, in the regression model (8), the one-class modulus often satisfies

$$\omega(\delta;\mathcal{F}) = n^{-r/2}A\delta^r(1+o(1)) \tag{20}$$

as $n \to \infty$ for some constant A, where $r/2$ is the rate of convergence of the minimax root MSE. We show that this is the case under regularity conditions in the regression discontinuity application in Lemma G.6 (see Donoho and Low, 1992, for other cases where (20) holds). In this case, (19) evaluates to $\frac{2^r}{1+r}(1+o(1))$, so that the asymptotic efficiency depends depends only on r. Figure 2 plots the asymptotic efficiency as a function of r.

Suppose $\mathcal{F}$ is smooth enough so that the rate of convergence satisfies $r \geq 1/2$ (as is the case for inference at a point when functions in $\mathcal{F}$ have at least one directional derivative). Then the asymptotic efficiency of minimax CIs relative to CIs that optimize their excess length for the zero function is at least $2^{1/2}/(1+1/2) = 94.3\%$ when indeed $f = 0$. Since adapting to the zero function is at least as hard as adapting to any set $\mathcal{G}$ that includes it, this implies that if $\mathcal{F}$ is convex and centrosymmetric, "directing power" yields very little gain in excess length no matter how optimistic one is about where to direct it.

This result places a severe bound on the scope for adaptivity in settings in which $\mathcal{F}$ is convex and centrosymmetric: any CI that performs better than the minimax CI by more than the ratio in (19) must fail to control coverage at some $f \in \mathcal{F}$. Adaptation is only

possible when centrosymmetry fails (typically by placing shape restrictions on f, such as monotonicity), or convexity fails (by say placing sparsity assumptions on the coefficients in a series expansion of f).

3.4 Two-sided CIs and minimax MSE estimators

Finding optimal rules for two-sided confidence intervals, or for estimation criteria such as mean squared error, is more complicated. However, it is known that estimators in the class $\hat{L}_{\delta,\mathcal{F}}$ and the associated fixed-length CIs are minimax optimal when one restricts attention to affine estimators (i.e. estimators of the form $\hat{L} = a + \langle b, Y \rangle$ for constants $a \in \mathbb{R}$ and $b \in \mathcal{Y}$) if δ is chosen optimally. These results are due to Donoho (1994), and we state them below for convenience. We then give a solution to the problem of constructing confidence sets that optimize expected length $\Lambda_f(\mathcal{C})$ at a single function f, and use this result to bound the efficiency of fixed-length affine CIs among all confidence sets.

To describe the Donoho (1994) results, first consider the normal model $Z \sim N(\mu, 1)$ where $\mu \in [-\tau, \tau]$. The minimax affine mean squared error for this problem is

$$\rho_A(\tau) = \min_{\delta(Y) \text{ affine}} \max_{\mu \in [-\tau,\tau]} E_\mu(\delta(Y) - \mu)^2.$$

The solution is achieved by shrinking Y toward 0, namely $\delta(Y) = c_\rho(\tau)Y$, with $c_\rho(\tau) = \tau^2/(1+\tau^2)$, which gives $\rho_A(\tau) = \tau^2/(1+\tau^2)$. The length of the smallest fixed-length affine $100 \cdot (1-\alpha)\%$ confidence interval is

$$\chi_{A,\alpha}(\tau) = \min \left\{ \chi \colon \text{there exists } \delta(Y) \text{ affine s.t. } \inf_{\mu \in [-\tau,\tau]} P_\mu(|\delta(Y) - \mu| \leq \chi) \geq 1 - \alpha \right\}.$$

The solution is achieved at some $\delta(Y) = c_\chi(\tau)Y$, and it is characterized in Drees (1999). We give the details in Appendix C for convenience.

By a sufficiency argument, the minimax MSE affine estimator in the one-dimensional submodel $\{g\lambda + f(1-\lambda) \colon \lambda \in [0,1]\}$ is characterized by a scaling of $\rho_A(\tau)$ for an appropriate choice of τ, and similarly for $\chi_{A,\alpha}$. Donoho (1994) then uses the modulus of continuity to find the least favorable submodel such that minimax affine estimators and fixed-length CIs in the submodel are also minimax in the full model. This leads to the following result:

Theorem 3.2 (Donoho 1994)**.** *Suppose that δ_ρ is a solution to*

$$c_\rho(\delta/(2\sigma)) = \delta\omega'(\delta)/\omega(\delta),$$

and that $f^*_{\delta_\rho}, g^*_{\delta_\rho}$ achieve the one-class modulus $\omega(\cdot;\mathcal{F})$ at δ_ρ. Then the MSE minimax affine estimator is of Lf is $\hat{L}_{\delta_\rho,\mathcal{F}}$, and its maximum root MSE is given by

$$R(\hat{L}_{\delta_\rho,\mathcal{F}})^{1/2} = \frac{\omega(\delta_\rho)}{\delta_\rho}\sqrt{\rho_A\left(\frac{\delta_\rho}{2\sigma}\right)}\sigma.$$

Similarly, suppose that δ_χ is a solution to

$$c_\chi(\delta/(2\sigma)) = \delta\omega'(\delta)/\omega(\delta),$$

and that $f^*_{\delta_\chi}, g^*_{\delta_\chi}$ achieve the one-class modulus $\omega(\cdot;\mathcal{F})$ at δ_χ. Then the shortest fixed-length affine CI is given by

$$\hat{L}_{\delta_\chi,\mathcal{F}} \pm \frac{\omega(\delta_\chi)}{\delta_\chi}\chi_{A,\alpha}\left(\frac{\delta_\chi}{2\sigma}\right)\sigma.$$

Theorem 3.2 gives the optimal δ for a particular performance criterion in terms of the shrinkage coefficient in the one dimensional bounded normal means problem ($c_\rho(\cdot)$ or $c_\chi(\cdot)$). Often (at least asymptotically), $\hat{L}_{\delta,\mathcal{F}}$ takes the form of a kernel estimator with bandwidth determined by δ; this allows for comparisons of optimal bandwidths for different performance criteria. We perform such comparisons in a companion paper (Armstrong and Kolesár, 2016).

Donoho (1994) also bounds the penalty for restricting attention to affine procedures, using a formula based on the modulus of continuity. Since the bounds turn out to be very tight in many situations, the cost of restricting attention to affine procedures is typically not too large. We refer the reader to Donoho (1994), Drees (1999) and references therein for details.

On the other hand, just as with minimax one-sided CIs, one may worry that since the length of fixed-length CIs is driven by the least favorable functions, restricting attention to fixed-length CIs may be very costly when the true f is smoother. The next result characterizes the confidence sets that optimizes expected length at a single function g, and thus gives bounds for the possible performance gains.

Theorem 3.3. *Let $g \in \mathcal{F}$, and assume that a minimizer f_{L_0} of $\|K(g-f)\|$ subject to $Lf = L_0$ and $f \in \mathcal{F}$ exists for all $L_0 \in \mathbb{R}$, and let $\delta_{L_0} = \|K(g - f_{L_0})\|$. Then the confidence set $\mathcal{C}_g(Y)$ that minimizes $E_g\lambda(\mathcal{C})$ subject to $1-\alpha$ coverage on $\mathcal{F}$ inverts the family of tests ϕ_{L_0} that reject for large values of $\langle K(g-f_{L_0}), Y\rangle$ with critical value given by the $1-\alpha$ quantile under f_{L_0}. The expected length of this confidence set is given by*

$$E_g[\lambda(\mathcal{C}_g(Y))] = (1-\alpha)E\left[(\omega(\sigma(z_{1-\alpha}-Z);\mathcal{F},\{g\}) + \omega(\sigma(z_{1-\alpha}-Z);\{g\},\mathcal{F})) \mid Z \le z_{1-\alpha}\right],$$

where Z is a standard normal random variable.

This result gives the exact solution to the problem of "adaptation to a function" posed by Cai, Low, and Xia (2013), who obtain bounds for this problem in the case where $\mathcal{C}$ is required to be an interval. It follows from the observation in Pratt (1961) that minimum expected length CIs are obtained by inverting a family of uniformly most powerful tests of $H_0\colon Lf = L_0$ and $f \in \mathcal{F}$ against $H_1\colon f = g$. The least favorable null f_{L_0} for such a test is given by a minimizing $\|K(g-f)\|$ subject to $Lf = L_0$. Equivalently, we can obtain it as a solution to the ordered modulus problem $\omega(\delta_{L_0}; \mathcal{F}, \{g\})$ (if $L_0 \leq Lg$), or $\omega(\delta_{L_0}; \{g\}, \mathcal{F})$ (if $L_0 \geq Lg$). The expression for the expected length of $\mathcal{C}_g(Y)$ follows by computing the power of these tests. The assumption that a minimizer of $\|K(g-f)\|$ subject to $Lf = L_0$ and $f \in \mathcal{F}$ exists for all $L_0 \in \mathbb{R}$ means that Lf is unbounded over $\mathcal{F}$. This assumption is made to simplify the statement; a truncated version of the same formula holds when $\mathcal{F}$ places a bound on Lf.

Directing power at a single function is seldom desirable in practice. Theorem 3.3 is very useful, however, in bounding the efficiency of other procedures, such as fixed-length CIs from Theorem 3.2. In particular, suppose $f - g \in \mathcal{F}$ for all f (so that (16) holds with $\mathcal{G} = \{g\}$) and that $\mathcal{F}$ is centrosymmetric. Then, by arguments in Section 3.3, $\omega(\delta; \mathcal{F}, \{g\}) = \omega(\delta; \{g\}, \mathcal{F}) = \frac{1}{2}\omega(2\delta; \mathcal{F})$, which yields:

Corollary 3.3. *Consider the setup in Theorem 3.3 with the additional assumption that $\mathcal{F}$ is centrosymmetric and g satisfies $f - g \in \mathcal{F}$ for all f. Then the efficiency of the fixed-length CI around $\hat{L}_{\delta_\chi, \mathcal{F}}$ at g relative to all confidence sets is*

$$\frac{\inf_{\mathcal{C} \in \mathcal{I}_\alpha} \Lambda(\mathcal{C}(Y), \{g\})}{\frac{\omega(\delta_\chi) 2\sigma}{\delta_\chi} \chi_{A,\alpha}\left(\frac{\delta_\chi}{2\sigma}\right)} = \frac{(1-\alpha) E\left[\omega(2\sigma(z_{1-\alpha} - Z); \mathcal{F}) \mid Z \leq z_{1-\alpha}\right]}{\frac{\omega(\delta_\chi) 2\sigma}{\delta_\chi} \chi_{A,\alpha}\left(\frac{\delta_\chi}{2\sigma}\right)}. \tag{21}$$

The assumption of Corollary 3.3 will be satisfied for smooth functions g, including the zero function. This efficiency ratio can easily be computed in particular applications, and we do in Section 5.2 in an application to regression discontinuity. However, it is insightful to consider the asymptotic efficiency implied by (21) when the one-class modulus satisfies (20). In this case Theorem 3.2 implies that $\delta_\chi = 2\sigma c_\chi^{-1}(r) + o(1)$, so that the length of the fixed-length CI around $\hat{L}_{\delta_\chi}$ is given by

$$n^{-r/2} A 2^r \sigma^r \frac{\chi_{A,\alpha}(c_\chi^{-1}(r))}{(c_\chi^{-1}(r))^{1-r}}(1 + o(1)),$$

and we get

$$\frac{\inf_{\mathcal{C}\in\mathcal{I}_\alpha}\Lambda(\mathcal{C}(Y),\{g\})}{\frac{\omega(\delta)2\sigma}{\delta}\chi_{A,\alpha}\left(\frac{\delta}{2\sigma}\right)}=\frac{(1-\alpha)E[(z_{1-\alpha}-Z)^r\mid Z\le z_{1-\alpha}]}{(c_\chi^{-1}(r))^{r-1}\chi_{A,\alpha}(c_\chi^{-1}(r))}(1+o(1)) \tag{22}$$

(here, we use properties of the modulus and $\chi_{A,\alpha}$ to obtain the above display from the pointwise-in-δ convergence in (20); see Lemma F.2 in the Supplemental Materials). This asymptotic efficiency is plotted in Figure 2 as a function of r for $\alpha = 0.05$. When $r = 4/5$ (as in the regression discontinuity application in Section 5), for instance, the asymptotic efficiency is 95.7%. When $r = 1$ (parametric rate of convergence), the asymptotic efficiency equals $((1-\alpha)z_{1-\alpha}+\phi(z_{1-\alpha}))/z_{1-\alpha/2}$, as in the normal mean example in Pratt (1961, Section 5), where ϕ is standard normal density. For $\alpha = 0.05$, this yields 84.99%.

Just like with minimax one-sided CIs, this result places a severe bound on the scope for improvement over fixed-length CIs when $\mathcal{F}$ is centrosymmetric. It strengthens the finding in Low (1997) and Cai and Low (2004a), who derive bounds on the expected length of random length $1-\alpha$ CIs (i.e. CIs in the set $\mathcal{I}_\alpha$). Their bounds imply that when $\mathcal{F}$ is constrained only by bounds on a derivative, the expected length of any CI in $\mathcal{I}_\alpha$ must shrink at the minimax rate $n^{-r/2}$ for *any* any f in the interior of $\mathcal{F}$. Equation (22) shows that for smooth functions f, this remains true whenever $\mathcal{F}$ is centrosymmetric, even if we don't require $\mathcal{C}$ to take the form of an interval, and, moreover, not only is the rate the same as the minimax rate, the constant must be close to that for fixed-length CIs.

On the other hand, when $\mathcal{F}$ is not centrosymmetric, it is possible to improve upon the fixed-length CIs, and Cai and Low (2004a) give a general procedure that is rate-adaptive.

3.5 Confidence Intervals Based on Suboptimal Estimators

The confidence intervals discussed in Sections 3.3 and 3.4 are based on the worst case bias of $\hat{L}_{\delta,\mathcal{F}}$ for δ chosen optimally. More generally, for any affine estimator $\hat{L}$, the set of possible distributions of $\hat{L} - Lf$ as f ranges over $\mathcal{F}$ is characterized by the set of possible biases of $\hat{L}$, and a CI can be constructed based on the maximum and minimum bias. For $\hat{L}_{\delta;\mathcal{F}}$, the maximum and minimum bias are attained g_δ^* and f_δ^* no matter how δ is chosen (see Lemma 4 in Donoho, 1994, and Lemma B.1 in the Appendix); this allows a further simplification.

To describe the results, let $\mathrm{cv}_\alpha(b)$ be the shortest half-length of a $1-\alpha$ CI for some parameter that is centered around a normally distributed estimator with variance one and maximum absolute bias equal to b. In other words, $\mathrm{cv}_\alpha(b)$ solves $P(|Z+b|\le \mathrm{cv}) = \Phi(\mathrm{cv}-b)-\Phi(-\mathrm{cv}-b) = 1-\alpha$, where $Z\sim N(0,1)$. We tabulate these critical values in Table 1.

Theorem 3.4. *Let $\hat{L} = a + \langle w, Y \rangle$ be an affine estimator such that $\overline{\text{bias}}_{\mathcal{F}}(\hat{L})$ and $\underline{\text{bias}}_{\mathcal{F}}(\hat{L})$ are finite. Let $b = \max\{|\overline{\text{bias}}_{\mathcal{F}}(\hat{L})|, |\underline{\text{bias}}_{\mathcal{F}}(\hat{L})|\}$. Then: (i) $[\hat{L} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}) - \|w\| z_{1-\alpha}\sigma, \infty)$ is a valid CI, and it has maximum excess length*

$$q_\beta(\hat{L}; \mathcal{F}) = \sigma\|w\|(z_\beta + z_{1-\alpha}) + \overline{\text{bias}}_{\mathcal{F}}(\hat{L}) - \underline{\text{bias}}_{\mathcal{F}}(\hat{L}).$$

(ii) $\hat{L} \pm \text{cv}_\alpha(b)\sigma\|w\|$ is the shortest fixed-length $1-\alpha$ CI centered at $\hat{L}$.

For $\hat{L}_{\delta,\mathcal{F}}$, this holds with

$$\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{\delta,\mathcal{F}}) = -\underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{\delta,\mathcal{F}}) = \frac{1}{2}(\omega(\delta) - \delta\omega'(\delta)) \text{ and } \|w\| = \omega'(\delta). \tag{23}$$

Theorem 3.4 can be used along with Theorem 3.2 to bound the efficiency loss from basing a confidence interval on a suboptimal estimator, or from basing a confidence interval on an estimator that is optimal for mean squared error, rather than CI length. We do this in Section 5 for a regression discontinuity application. In Armstrong and Kolesár (2016), we consider asymptotic implications of this result.

3.6 Unknown Error Distribution

Throughout this section, we have assumed that the error term ε is normal with known variance. When the error distribution is unknown, one can form estimates and CIs based on an estimate or guess for the variance function. If the variance function used in forming the estimate is misspecified, one can use a robust estimate of the variance of the estimate along with the approach in Section 3.5 in forming the CI. In Supplemental Appendix E we consider a version of the nonparametric fixed-design regression model with non-normal errors and show that, under regularity conditions, this leads to CIs that are valid in a uniform asymptotic sense, with the efficiency bounds carrying over to this setup in an asymptotic sense as well. These results show that optimal CIs are based on asymptotically normal estimates in a broad class of settings with non-normal errors.

4 Linear regression

This section considers the linear regression model (9). The results in Section 3 apply to the problem of optimizing performance over $\theta \in \mathcal{G}$ subject to a coverage requirement over $\theta \in \mathcal{F}$, where $\mathcal{F}$ and $\mathcal{G}$ are convex sets. Many constraints used in parametric models in econometrics

lead to convex parameter sets, including restrictions on the sign and magnitude of particular coefficients (see Andrews, 2001, and references therein).

Consider a linear functional $L\beta = \ell'\beta$, where ℓ is a $k \times 1$ column vector. The ordered modulus problem for $\omega(\delta; \mathcal{F}, \mathcal{G})$ is

$$\sup_{\beta} \ell'(\gamma - \theta) \quad \text{s.t.} \quad \|X(\gamma - \theta)\| \leq \delta,\ \gamma \in \mathcal{G}, \theta \in \mathcal{F}, \tag{24}$$

which is a finite dimensional convex optimization problem. For translation invariance, we can take $\iota = \iota_\ell = \ell/\|\ell\|^2$. In the remainder of this section, we discuss the form of optimal procedures in some special cases (in Section 4.1), as well as implications of the results in Section 3 for variable selection (in Section 4.2).

4.1 Examples

We solve (24) in some examples. First, we show that the problem reduces to inference based on the ordinary least squares (OLS) estimate when the parameter space is unconstrained. Next, we note that elliptical constraints lead to inference based on ridge regression estimates. Finally, we consider the bivariate case and analyze how restrictions on the coefficient of one variable affect inference on the other variable.

4.1.1 Unconstrained Parameter Space

In the unconstrained case $\mathcal{F} = \mathcal{G} = \mathbb{R}^k$ the modulus problem (24) reduces to $2 \max_\theta \ell'\theta$ s.t. $\|X\theta\| \leq \delta/2$. Simple calculations involving the Lagrangian leads to the solution $\hat{L}_\delta = \ell'(X'X)^{-1}X'Y$, (see Supplemental Appendix D.1 for details). Thus, $\hat{L}_\delta$ is given by applying the linear transformation L to the OLS estimator $(X'X)^{-1}X'Y$, regardless of δ. The worst-case bias is zero, and the fact that the estimator minimizes variance subject to this bound on the bias reduces to the Gauss-Markov theorem. Since the parameter space is unconstrained, we can take ι to be any element with $\ell'\iota = 1$. By centrosymmetry, the one-sided confidence set that minimizes any quantile of excess length uniformly over the span of ι is based on $\hat{L}_\delta$ for δ chosen appropriately. Since $\hat{L}_\delta$ does not depend on δ and the span of ι gives the entire parameter space by varying the definition of ι, we obtain the classic result that the uniformly most powerful test of $H_0 : \ell'\theta \leq L_0$ is the one-sided z-test based on the OLS estimate.

4.1.2 Elliptical Constraints

Suppose that $\mathcal{F} = \mathcal{G} = \{\theta \colon \|M\theta\| \leq C\}$ for some $k \times k$ matrix M. The form of the class of optimal estimators can again be derived by solving the Lagrangian; it is given by

$$\hat{L}_\delta = \ell'(X'X + \tilde{\lambda}_\delta M'M)^{-1}X'Y$$

where $\tilde{\lambda}_\delta$ is given by the ratio of Lagrange multipliers (see Supplemental Appendix D.2 for details). Note that $\hat{L}_\delta$ is obtained by applying the transformation L to the ridge regression estimator $(X'X+\tilde{\lambda}_\delta M'M)^{-1}X'Y$, with the regularization parameter $\tilde{\lambda}_\delta$ depending on δ. The minimaxity of this class of estimators for mean squared error has been noted by Li (1982). The results in Section 3 show that minimax one-sided CIs take this form as well. In addition, since the class $\mathcal{F}$ is centrosymmetric, one-sided CIs that optimize performance at $\theta = 0$ also take this form, and Corollaries 3.2 and 3.3 give bounds on the scope for "adapting" to $\theta = 0$ while maintaining correct coverage over the elliptical class.

4.1.3 Sign Restrictions in the Two Parameter Case

Consider the case where $k = 2$, and we are interested in inference on $L\theta = \theta_1$ with θ_1 unconstrained and θ_2 restricted to be positive: $\mathcal{F} = \mathbb{R} \times [0, \infty)$. For the minimax criterion ($\mathcal{G} = \mathcal{F}$), the modulus problem is

$$\sup_{\theta,\gamma} \gamma_1 - \theta_1 \text{ s.t. } (\gamma - \theta)'X'X(\gamma - \theta) \leq \delta^2, \gamma_2 \geq 0, \theta_2 \geq 0.$$

For any θ and γ that solve this problem without the second constraint, we can add $(c, c)'$ to both θ and γ for a large constant c and obtain the same value without the second constraint binding. Thus, for $\mathcal{G} = \mathcal{F}$, the constraint on θ_2 does not affect the optimal procedure.

Suppose that we wish to optimize performance over the set $\mathbb{R} \times \{\tilde{\gamma}_2\}$ for some fixed $\tilde{\gamma}_2 > 0$. Let us normalize the parameter θ so that the diagonal elements of $X'X$ are 1, and let ρ be the off-diagonal element of $X'X$ (this reparameterizes θ as $\text{diag}(X'X)\theta$). The modulus problem is

$$\sup_{\theta,\gamma_1} \gamma_1 - \theta_1 \text{ s.t. } (\gamma_1 - \theta_1)^2 + 2\rho(\gamma_1 - \theta_1)(\tilde{\gamma}_2 - \theta_2) + (\tilde{\gamma}_2 - \theta_2)^2 \leq \delta^2, \theta_2 \geq 0.$$

The constraint $\theta_2 \geq 0$ will bind iff. dropping the constraint leads to a negative value of θ. Dropping this constraint, the first order conditions for θ_2 give $-2\lambda(\rho(\gamma_1-\theta_1)+(\tilde{\gamma}_2-\theta_2)) = 0$

so that $\tilde{\gamma}_2 - \theta_2 = -\rho(\gamma_1 - \theta_1)$. Thus, the unconstrained θ_2 is given by $\theta_2 = \tilde{\gamma}_2 + \rho\omega(\delta)$. If $\rho > 0$, the constraint will never bind, and the test will be the same as in the unconstrained problem. If $\rho < 0$, the constraint will always bind when $\tilde{\gamma}_2 = 0$, and the range of $\tilde{\gamma}_2$ on which the constraint binds is given by $[0, |\rho|\omega(\delta))$.

To get some intuition for this, note that, for $\hat{\theta}_{OLS}$, the covariance between the estimates of the two parameters is positive iff. ρ is negative. Thus, if $\rho < 0$, one can decrease the variance of the OLS estimate $\hat{\theta}_{OLS,1}$ by subtracting some fraction of $\hat{\theta}_{OLS,2}$. If we maintain the restriction $\theta_2 \geq 0$ under the null, then this can only introduce downward bias, so we do not need to adjust the critical value when constructing a lower CI. This strategy works for "directing power" against $\tilde{\gamma}_2$ so long as $\tilde{\gamma}_2$ is not too large, so that the negative bias does not decrease power too much under the alternative. Another source of intuition is the formula for omitted variables bias. If $\rho < 0$ (the regressors are negatively correlated), then, under the maintained hypothesis $\theta_2 \geq 0$, ignoring the second regressor leads to downward bias, so it is possible to form a lower CI based on the OLS estimate in the regression with the second regressor omitted, or by using some combination of the OLS estimates of θ_1 with and without the second regressor.

4.2 Implications for Variable Selection

The results of Section 3 can be used to address the question: under what conditions does variable selection or shrinkage make sense for confidence interval construction? Inference after model selection has been a topic of interest in the recent econometrics and statistics literature (see, among others, Andrews and Guggenberger, 2009a; Belloni, Chernozhukov, and Hansen, 2014; Leeb and Pötscher, 2005; McCloskey, 2012; van de Geer, Bühlmann, Ritov, and Dezeure, 2014; Zhang and Zhang, 2014).

If the parameter space is completely unrestricted under the null ($\mathcal{F} = \mathbb{R}^k$), then, as discussed in Section 4.1.1, the one-sided test based on the unrestricted OLS estimator is uniformly most powerful. This is an extremely powerful result regarding the use of anything other than the one-sided z-test based on the unrestricted OLS estimator: even if one only cares about power in the case where all parameters are zero except for the parameter of interest, the optimal test still uses the unrestricted OLS estimator.

To get around this negative result, one must restrict the parameter space under the null. Consider the case where L is a single element of the parameter vector: $L\theta = \theta_1$. Consider inference on θ_1 with the remaining parameters $\theta_{-1} = (\theta_2, \ldots, \theta_k)' \in \mathbb{R}^{k-1}$ constrained to some set $\mathcal{F}_{-1} \subseteq \mathbb{R}^{k-1}$. This fits into our framework with $\mathcal{F} = \mathbb{R} \times \mathcal{F}_{-1}$. If $\mathcal{F}_{-1}$ places

nontrivial restrictions on the remaining parameters, optimal one-sided tests will, in general, not be based on the unrestricted OLS estimator.

Suppose that we suspect that the remaining coefficients θ_{-1} are zero, and want to optimize the performance of a confidence interval for this parameter value while maintaining size control over $\mathcal{F}_{-1}$. If $\mathcal{F}_{-1}$ is centrosymmetric, then it follows from Corollary 3.2 that the minimax one-sided CI for β quantile excess length also optimizes $\tilde{\beta}$ quantile excess length at $\theta_{-1} = 0$, where $\tilde{\beta} = \Phi((z_\beta - z_{1-\alpha})/2)$. Furthermore, Corollary 3.2 gives the relative efficiency for the minimax one-sided CI for optimizing excess length at $\theta_{-1} = 0$. For two-sided CIs, Corollary 3.3 gives the potential improvement from optimizing expected length at a value of $(\theta_1, \theta_{-1}')'$ with $\theta_{-1} = 0$ relative to fixed-length affine CIs. Note that the same argument holds for optimizing performance at some parameter value $\tilde{\theta}$ if the parameter space is centrosymmetric about $\tilde{\theta}$. For example, if one defines the parameter space by choosing a plausible parameter value and placing symmetric bounds around it, Corollaries 3.2 and 3.3 give bounds on the scope for directing power at this parameter.

These results severely limit the scope for variable selection or other procedures that attempt to "adapt" to particular parameter values when $\mathcal{F}_{-1}$ is convex and centrosymmetric. To get around this, one must consider situations where parameters are restricted to a non-convex or asymmetric parameter space under the null. Sparsity is one example of a non-convex restriction under which variable selection has been used fruitfully (see Belloni, Chernozhukov, and Hansen, 2014, for an example). If $\mathcal{F}_{-1}$ is the set of s-sparse vectors $\{\theta_{-1} : \#\{j : \theta_{-1,j} \neq 0\} \leq s\}$ one can use pre-testing to find the indices of the non-zero coefficients while controlling size. However, Corollaries 3.2 and 3.3 are relevant here as well. While we do not need to know the location of the non-zero coefficients, we must impose sparsity when defining size. Furthermore, Corollaries 3.2 and 3.3 can be used to bound the scope for adapting to the level of sparsity.

Suppose that we wish to impose only s-sparsity under the null, while optimizing performance when the parameter vector is p-sparse, where $p < s < k$. Using βth quantile excess length of one-sided CIs as the performance criterion, this amounts to optimizing $q_\beta(\hat{c}; \{\theta_{-1} : \#\{j : \theta_{-1,j} \neq 0\} \leq p\})$ subject to $1 - \alpha$ coverage of $[\hat{c}, \infty)$ over $\{\theta_{-1} : \#\{j : \theta_{-1,j} \neq 0\} \leq s\}$. Since the sets involved in this problem are non-convex, the results in this paper do not apply immediately. However, relaxing the problem by assuming that we know the indices of the nonzero components under the null and alternative can only make the problem easier: the convex problem of optimizing $q_\beta(\hat{c}; \{\theta_{-1} : \theta_{-1,j} = 0 \text{ for } j > p\})$ subject to coverage over $\{\theta_{-1} : \theta_{-1,j} = 0 \text{ for } j > s\}$ provides a lower bound. By Corol-

laries 3.2 and 3.3, one cannot do much better at p-sparse parameters subject to coverage over s-sparse parameters than than the minimax CI over s-sparse parameters with the non-zero components known. Thus, for confidence interval construction, the scope for adapting between different levels of sparsity is severely limited. The same arguments go through if one considers approximately sparse sets of the form $\{\theta_{-1} : \#\{j : |\theta_{-1,j}| > c\} \leq s\}$, or if one considers the set of regression functions with a bound on the approximation error of sparse linear functions. See Cai and Guo (2015) for recent work on adaptation to sparsity in high-dimensional regression.

5 Regression discontinuity

In a (sharp) regression discontinuity (RD) design, we are interested in estimating a jump in the regression function in the model (8) at a known threshold, which we normalize to 0, so that throughout this section, we set

$$Lf = \lim_{x \downarrow 0} f(x) - \lim_{x \uparrow 0} f(x).$$

The threshold determines participation in a binary treatment: units with $x_i > 0$ are treated; units with $x_i < 0$ are controls (we assume that $x_i \neq 0$ for all i). If the regression functions of potential outcomes are continuous at zero, then Lf measures the average effect of the treatment for units with covariate values equal to the threshold.

Let $f_+(x) = f(x)I(x > 0)$ and $f_-(x) = -f(x)I(x < 0)$ so that we can write $f = f_+ - f_-$. Also let $f_+(0) = \lim_{x \downarrow 0} f_+(x)$ and $f_-(0) = \lim_{x \uparrow 0} f_-(x)$, so that $Lf = f_+(0) + f_-(0)$. We will assume that f lies in the class of functions

$$\mathcal{F}_{RDT,p}(C) = \{f_+ - f_- \colon f_+ \in \mathcal{F}_{T,p}(C; \mathbb{R}_+),\ f_- \in \mathcal{F}_{T,p}(C; \mathbb{R}_-)\},$$

where $\mathcal{F}_{T,p}(C; \mathcal{X})$ consists of functions f such that the approximation error from pth order Taylor expansion of $f(x)$ about 0 is bounded by $C|x|^p$, uniformly over $\mathcal{X}$:

$$\mathcal{F}_{T,p}(C; \mathcal{X}) = \left\{ f \colon \left| f(x) - \textstyle\sum_{j=0}^{p-1} \frac{f^{(j)}(0)}{j!} x^j \right| \leq C|x|^p \text{ all } x \in \mathcal{X} \right\}.$$

The class $\mathcal{F}_{T,p}(C; \mathcal{X})$ formalizes the idea that the pth derivative of f at zero should be bounded by $p!C$.

Minimax estimation using this class of functions goes back at least to Legostaeva and

Shiryaev (1971). Sacks and Ylvisaker (1978) and Cheng, Fan, and Marron (1997) considered minimax MSE estimation of $f(0)$ in this class of functions when 0 is a boundary point. Their results formally justify using local polynomial regression to estimate the RD parameter. When the degree of smoothness p is not known, Sun (2005) proposes an adaptive version of the local polynomial estimator that achieves the optimal rate of convergence up to a logarithmic factor. In contrast, since the class $\mathcal{F}_{RDT,p}(C)$ is symmetric, Corollaries 3.2 and 3.3 imply that it is not possible to construct confidence intervals that shrink at the optimal rate without knowing p. The researcher will therefore need to specify both p and C to construct confidence intervals.

To illustrate the theoretical results in this section, we use the dataset from Lee (2008). The dataset consists of 6,558 observations that correspond to elections to the US House of Representatives between 1946 and 1998. The running variable $x_i \in [-100, 100]$ is the Democratic margin of victory (in percentages) in a given election i. The outcome variable $y_i \in [0, 100]$ is the Democratic vote share (in percentages) in the next election. Given the inherent uncertainty in final vote counts, the party that wins is essentially randomized in elections that are decided by a narrow margin, so that Lf measures the incumbency advantage for Democrats for elections decided by a narrow margin—the impact of being the current incumbent party in a congressional district on the probability of winning the next election.

To implement the optimal procedures in the Lee application, we will need to use an estimated version of $\sigma(x)^2$, as the true variance function is unknown. We assume that the variance is homoscedastic on either side of the cutoff and use the estimates $\hat{\sigma}^2_+(x) = 14.5^2$ and $\hat{\sigma}^2_-(x) = 12.5^2$, which are based on residuals form a local linear regression with bandwidth selected using the Imbens and Kalyanaraman (2012, IK hereafter) selector. In Section 5.5, we show that the resulting confidence intervals will be asymptotically valid and optimal so long as $\hat{\sigma}_+(0)$ and $\hat{\sigma}_-(0)$ converge to $\sigma_+(0)$ and $\sigma_-(0)$ uniformly over $\mathcal{F}$, even if the true variance function is not constant.

We use variance estimates based on the IK bandwidth for simplicity and for comparison with the previous literature. While the optimality-within-a-class results of IK for estimating the regression discontinuity parameter do not apply in the uniform sense considered in this paper, the tuning parameters they use guarantee uniform convergence of the variance estimate based on this bandwidth when the regression and variance functions are restricted to an appropriate class. In Section 5.5, we show that the particulars of the variance estimate do not matter for first order asymptotics (so long as it is uniformly consistent). On the

other hand, the Edgeworth expansions in Calonico, Cattaneo, and Farrell (2015) suggest that alternative variance estimators may be preferred.

5.1 Least favorable functions

To construct optimal estimators and confidence sets, we first need to find functions g_δ^* and f_δ^* that solve the modulus problem. Since the class $\mathcal{F}_{RDT,p}(C)$ is symmetric, $f_\delta^* = -g_\delta^*$, and the (single-class) modulus of continuity $\omega(\delta; \mathcal{F}_{RDT,p}(C))$ is given by the value of the problem

$$\sup_{f_+ - f_- \in \mathcal{F}_{RDT,p}(C)} 2(f_+(0) + f_-(0)) \quad \text{st} \quad \sum_{i=1}^n \frac{f_-(x_i)^2}{\sigma^2(x_i)} + \sum_{i=1}^n \frac{f_+(x_i)^2}{\sigma^2(x_i)} \leq \delta^2/4. \tag{25}$$

Let $g_{\delta,C}^*$ denote the (unique up to the values at the x_is) solution to this problem. The solution $g_{\delta,C}^*$ can be obtained using a simple generalization of Theorem 1 of Sacks and Ylvisaker (1978); it is characterized by a system of $2p$ equations in $2p$ unknowns. We provide details in Supplemental Appendix D.3.

Using the fact that the class $\mathcal{F}_{RDT,p}(C)$ is translation invariant (we can take $\iota(x) = c_0 + 1(x > 0)$ for any c_0) and $\sum_{i=1}^n \frac{g_{+,\delta,C}^*(x_i)}{\sigma^2(x_i)} = \sum_{i=1}^n \frac{g_{-,\delta,C}^*(x_i)}{\sigma^2(x_i)}$ (this can be seen by noting that the bias at any constant function must be zero—otherwise the bias could be made arbitrarily large by increasing the constant; see Supplemental Appendix D.3 for details), the class of estimators $\hat{L}_\delta$ can be written as

$$\hat{L}_\delta = \hat{L}_{\delta,\mathcal{F}_{RDT,p}(C)} = \frac{\sum_{i=1}^n g_{+,\delta,C}^*(x_i) y_i/\sigma^2(x_i)}{\sum_{i=1}^n g_{+,\delta,C}^*(x_i)/\sigma^2(x_i)} - \frac{\sum_{i=1}^n g_{-,\delta,C}^*(x_i) y_i/\sigma^2(x_i)}{\sum_{i=1}^n g_{-,\delta,C}^*(x_i)/\sigma^2(x_i)}. \tag{26}$$

To illustrate these results using the Lee data, we fix $p = 2$. Figure 3 plots the least favorable function $g_{\delta,C}^*$ for this data, with δ calibrated to be optimal for one-sided CIs that minimax the excess length at the $\beta = 0.8$ quantile (so that $\delta = z_{0.95} + z_{0.8} = 2.49$), and several choices of C. It is clear from the figure that the smoothness parameter C effectively rescales the least favorable function while preserving its shape. Indeed, we show in Armstrong and Kolesár (2016) that $\hat{L}_\delta$ is asymptotically equivalent to a local linear estimator with bandwidth that depends on C and δ, and kernel for which the equivalent kernel (as defined in Fan and Gijbels, 1996, p. 72) is given by $k(u) = (3.95 - 9.11u + 4.88u^2)_+ - (3.95 - 9.11u - 4.88u^2)_-$ where $(t)_+ = \max\{t, 0\}$ and $(t)_- = -\min\{t, 0\}$.

It is also clear from Figure 3 that the least favorable function is not smooth away from the cutoff—indeed the Taylor class doesn't impose smoothness of f away from cutoff, which

may be too conservative in many applications. If one bounds the second derivative globally by $2C$, the least favorable function, derived by Gao (2016) in an asymptotic setting, has a more smooth appearance. As we show in Armstrong and Kolesár (2016), imposing a global bound on the second derivative tightens optimal CIs by about 10% in large samples (see also Appendix A for a Monte Carlo study of CIs under global smoothness).

Let us briefly discuss the interpretation of the smoothness constant C in this application. By definition of the class $\mathcal{F}_{RDT,2}(C)$, C determines how large the approximation error can be if we approximate the regression functions f_+ and f_- on either side of the cutoff by a linear Taylor approximation at the cutoff: the approximation error is no greater than Cx^2. Thus, if $C = 0.05$, and we are predicting the vote share in the next election when the margin of victory is, say, $x = 10\%$, the linear approximation and the true conditional expectation differ by at most 5%, and they differ by no more than 20% when $x = 20\%$. Suppose that the conditional variance is homoscedastic and equal to the IK estimate of 14.5%. Then $C = 0.05$ implies that the prediction MSE at can be reduced by at most $5^2/(14.5^2 + 5^2) = 10.6\%$ at $x = 10\%$, and at most by $20^2/(14.5^2 + 20^2) = 65.5\%$ at $x = 20\%$ when we use the true regression function rather than the linear approximation. To the extent that researchers agree that the vote share in the next election varies smoothly enough with the margin of victory in the current election to make such large reductions in MSE unlikely, $C = 0.05$ is quite a conservative choice.

5.2 Bounds on adaptation

Since the class $\mathcal{F}_{RDT,p}(C)$ is centrosymmetric, Corollaries 3.2 and 3.3 apply to bound the scope for adaptation to C. To the extent that the bounds are tight (which, as we will see below, is indeed the case), the a priori choice of C for confidence interval construction cannot be avoided if one is only willing to place f in the smoothness class $\mathcal{F}_{RDT,p}(C)$ for some C.

Let $\mathcal{G}_p = \left\{(\sum_{j=0}^{p-1} a_j x^j)1(x < 0) + (\sum_{j=0}^{p-1} b_j x^j)1(x > 0) \colon a_1, a_2, a_3, a_4 \in \mathbb{R}\right\}$ denote the class of piecewise polynomial functions. Since for any $f \in \mathcal{F}_{RDT,p}(C), g \in \mathcal{G}_p$, $f - g \in \mathcal{F}_{RDT,p}(C)$, it follows from Corollary 3.2 that the efficiency of minimax CIs relative to CIs that direct power at any subset of $\mathcal{G}_0 \subseteq \mathcal{G}_p$ is given by

$$\frac{\inf_{\hat{c}\colon [\hat{c},\infty)\in\mathcal{I}_\alpha} q_\beta(\hat{c}, \mathcal{G}_0)}{q_\beta(\hat{c}_{\alpha,\delta_\beta,\mathcal{F}_{RDT,p}(C)}, \mathcal{G}_0)} = \frac{\omega(2\delta; \mathcal{F}_{RDT,p}(C))}{\delta^2/(2\sum_{i=1}^n g^*_{+,\delta,C}(x_i)/\sigma^2(x_i)) + \omega(\delta; \mathcal{F}_{RDT,p}(C))},$$

with $\delta = z_{1-\alpha} + z_\beta$. In the Lee dataset with $p = 2$, the relative efficiency of CIs that minimax the 0.8 quantile is between 96% and 99.6% in for $C \in [0.00002, 0.1]$. The relative

efficiency of CIs that minimax the median is between 96% and 99.4%. Since the optimal rate of convergence is $r = 4/5$, this is very close to the asymptotic prediction $2^r/(1+r) = 96.7\%$.

For the fixed-length CIs, the efficiency at any $g \in \mathcal{G}_p$ is given by Corollary 3.3. In the Lee example with $p = 2$ and $C \in [0.0001, 0.1]$, which corresponds to a very wide range of smoothness classes, the efficiency varies between 95.4% and 95.9%. For very small C, it drops down to 91.3%. Unless C is extremely small, this matches the asymptotic efficiency of 95.7% implied by Equation (22) almost exactly.

5.3 Optimal inference procedures

To construct procedures that are optimal for a given performance criterion, we need to calibrate δ optimally. For one-sided CIs that minimax the excess length at the β quantile, the optimal δ is given by $\delta = z_{1-\alpha} + z_\beta$. The one-sided CI is then given by Corollary 3.1. The optimal δ for constructing fixed-length CIs and minimax MSE estimators is given in Theorem 3.2. We give implementation details in Supplemental Appendix D.3.

To illustrate the sensitivity of the results to the choice of C, Figure 4 plots these estimators and confidence intervals for the Lee data for $C \in [0.00002, 0.1]$. To understand the effect of C on the optimal amount of smoothing, we use the following definition of effective sample size. Let $\hat{L} = \sum_{i=1}^n w_+(x_i)y_i - \sum_{i=1}^n w_-(x_i)y_i$ be a linear estimator, where the weights w_+ and w_- satisfy $w_+(x) = 0$ if $x < 0$ and $w_-(x) = 0$ if $x > 0$. Then define the effective sample size by the variance measure of $\hat{L}$ (Klemelä, 2014)

$$n_e = \frac{1}{\sum_{i=1}^n w_+^2(x_i)} + \frac{1}{\sum_{i=1}^n w_-^2(x_i)}. \tag{27}$$

The logic behind this definition is that under homoscedasticity, the variance is given by $\text{var}(\sum_{i=1}^n w_+(x_i)y_i) = \sigma^2 \sum_{i=1}^n w_+^2(x_i)$, and similarly for the negative observations, so that n_e measures how much the variance shrinks. The results in Armstrong and Kolesár (2016) imply that in large samples, $n_e = O(C^{-2/5})$ for any performance criterion (so that doubling C reduces the effective sample size by about 25%), which predicts n_e in the Lee data almost exactly. The x-axis in Figure 4 reports n_e for the minimax MSE estimator.

The range of minimax MSE estimates varies between 5.8% and 7.3% for $C \in [0.005, 0.1]$, which is close to the original Lee estimate of 7.7% that was based on a global fourth degree polynomial. Interestingly, the lower and upper limits $\hat{c}_u$ and $\hat{c}_\ell$ of the one-sided CIs $[\hat{c}_\ell, \infty)$ and $(-\infty, \hat{c}_u]$ are not always within the corresponding limits for the two-sided CIs. The reason for this is that for any given C, the optimal δ is lower for one-sided CIs than for

two-sided fixed-length CIs—it equals 2.49 for one-sided CIs independently of the value of C, but for two-sided CIs, it varies between 4.1 and 11.7, depending on the exact value of C. Consequently, the effective number of observations for one-sided CIs is between 3% and 22% lower than for fixed-length CIs. Thus, when the point estimate decreases with the amount of smoothing as is the case for low values of C, then one-sided CIs are effectively centered around a lower estimate, which explains why at first the one-sided CI limits are both below the two-sided limits. This reverses once the point estimate starts increasing with the amount of smoothing.

On the other hand, the effective number of observations for the minimax MSE estimator is very close to that for fixed-length CIs throughout the entire range of Cs, never differing by more than 3%. This matches the asymptotic predictions in Armstrong and Kolesár (2016).

5.4 Confidence intervals based on suboptimal estimators

The minimax optimal procedures in Section 5.3 require that δ be chosen optimal for each performance criterion. In practice, a researcher may have multiple criteria in mind for a single estimate (e.g. one may want to report an estimator with good MSE, while also reporting a CI centered at this estimator). How much worse is the performance of confidence intervals when δ is not optimally chosen? Such confidence intervals can be constructed using Theorem 3.4. Figure 5 gives the resulting confidence intervals for the Lee data, with δ chosen so that the $\hat{L}_\delta$ is the minimax MSE estimator. In contrast with Figure 4, the limits of the one-sided CIs are now contained within the two-sided CIs, as they are both based on the same estimator, although they are less than $(z_{1-\alpha/2} - z_{1-\alpha})\,\mathrm{sd}(\hat{L}_\delta)$ apart as would be the case if $\hat{L}_\delta$ were unbiased.

The half-length of the two-sided fixed-length CI is at least 99.92% efficient relative to choosing δ optimally for fixed-length confidence intervals over the range of Cs reported in the graph. Similarly, the maximum excess length at the 0.8 quantile of the one-sided CIs is at least 97.3% efficient relative to minimax optimal CI. These results are in line with the asymptotic efficiency of confidence intervals based on the minimax MSE estimator that we compute in Armstrong and Kolesár (2016), which imply that the asymptotic efficiency of two-sided fixed-length CIs is 99.9%, and it is 98.0% for one-sided CIs.

Another natural question is: how much worse do CIs based on a different class of estimators perform? Cheng, Fan, and Marron (1997) show that local polynomial estimators achieve high asymptotic efficiency for the minimax MSE criterion $R(\hat{L})$. Consequently, these estimators have been recommended as an attractive choice in practice (see, e.g. Imbens and

Lemieux, 2008), and they have been very popular in recent applied work. Below, we use the results in Section 3 to derive relative efficiency of these estimators in the finite-sample normal model for the confidence interval criteria introduced in that section.

Consider a linear estimator

$$\hat{L}^l_{w_+,w_-} = \frac{\sum_{i=1}^n w_+(x_i)y_i}{\sum_{i=1}^n w_+(x_i)} - \frac{\sum_{i=1}^n w_-(x_i)y_i}{\sum_{i=1}^n w_-(x_i)},$$

where the weights w_+ and w_- satisfy $w_+(-x) = w_-(x) = 0$ for $x > 0$, and $\sum_i w_+(x_i)x_i = \sum_i w_-(x_i)x_i = 0$, so that the estimator is unbiased for piecewise linear functions. This covers, in particular, local polynomial estimators of at least linear order. For instance, local linear estimators with kernel k and bandwidths h_+ and h_-, use the weights

$$w_+(x) = k_+(x/h_+)\sum_{i=1}^n k_+(x_i/h_+)(x^2 - x\cdot x_i), \qquad k_+(u) = k(u)1(u>0),$$

and similarly for w_-.

The maximum bias of $\hat{L}^l_{w_+,w_-}$ is attained at $g^*_{w_+,w_-}(x) = \text{sign}(w_+(x))Cx^2 1(x > 0) - \text{sign}(w_-(x))Cx^2 1(x < 0)$. This follows since any $f \in \mathcal{F}_{RDT,2}(C)$ can be written as $(a_1 + a_2 x + r_+(x))1(x > 0) + (a_3 + a_4 x + r_-(x))1(x < 0)$, for some r_+, r_- such that $|r_\pm(x)| \le Cx^2$, so that the bias of the estimator under f can be upper-bounded by the bias at $g^*_{w_+,w_-}$. The minimum bias attains at $-g^*_{w_+,w_-}$. Hence,

$$\begin{aligned}\overline{\text{bias}}_{\mathcal{F}_{RDT,2}(C)}(\hat{L}^l_{w_+,w_-}) &= -\underline{\text{bias}}_{\mathcal{F}_{RDT,2}(C)}(\hat{L}^l_{w_+,w_-}) \\ &= C\frac{\sum_i |w_+(x_i)|x_i^2}{\sum_i w_+(x_i)} + C\frac{\sum_i |w_-(x_i)|x_i^2}{\sum_i w_-(x_i)}. \end{aligned} \tag{28}$$

The variance of the estimator doesn't depend on f and it is given by

$$\text{var}(\hat{L}^l_{w_+,w_-}) = \frac{\sum_i w_+(x_i^+)^2\sigma^2(x_i)}{(\sum_i w_+(x_i))^2} + \frac{\sum_i w_-(x_i)^2\sigma^2(x_i)}{(\sum_i w_-(x_i))^2}.$$

Therefore, given the weights w_+ and w_-, we can again use Theorem 3.4 to construct one and two-sided CIs around $\hat{L}^l_{w_+,w_-}$.

For local linear estimators, optimal bandwidths h_+ and h_- can be computed by minimizing the maximum excess length (for one-sided CIs) and half-length (for fixed-length CIs) over the bandwidths. We compute these in the Lee application using the triangular kernel. The resulting CIs are very close to the optimal CIs in Figure 4 (see Figure S1 in Supple-

mental Appendix H). Comparing half-length and excess length of the CIs based on local linear estimates to the optimal CIs over the range of C reported in the graph, we find that the two-sided CIs are at least 96.9% efficient, and one-sided CIs (based on optimizing the 0.8 quantile of excess length) are at least 96.9% efficient. This is very close to the asymptotic efficiency result in Armstrong and Kolesár (2016) that the local linear estimator with a triangular kernel is 97.2% efficient, independently of the performance criterion.

5.5 Asymptotic validity

We now give a theorem showing asymptotic validity of the CIs constructed in this section under an unknown error distribution. We consider uniform validity over regression functions in $\mathcal{F}$ and error distributions in a sequence of sets $\mathcal{Q}_n$, and we index probability statements with $f \in \mathcal{F}$ and $Q \in \mathcal{Q}_n$. We make the following assumptions on the x_is and the class of error distributions $\mathcal{Q}_n$.

Assumption 5.1. *For some $p_{X,+}(0) > 0$ and $p_{X,-}(0) > 0$, the sequence $\{x_i\}_{i=1}^n$ satisfies $\frac{1}{nh_n}\sum_{i=1}^n m(x_i/h_n)I(x_i > 0) \to p_{X,+}(0)\int_0^\infty m(u)\,du$ and $\frac{1}{nh_n}\sum_{i=1}^n m(x_i/h_n)I(x_i < 0) \to p_{X,-}(0)\int_{-\infty}^0 m(u)\,du$ for any bounded function m with bounded support and any h_n with $0 < \liminf_n h_n n^{1/(2p+1)} \leq \limsup_n h_n n^{1/(2p+1)} < \infty$.*

Assumption 5.2. *For some $\sigma(x)$ with $\lim_{x\downarrow 0}\sigma(x) = \sigma_+(0) > 0$ and $\lim_{x\uparrow 0}\sigma(x) = \sigma_-(0) > 0$,*

(i) the u_is are independent under any $Q \in \mathcal{Q}_n$ with $E_Q u_i = 0$, $var_Q(u_i) = \sigma^2(x_i)$

(ii) for some $\eta > 0$, $E_Q|u_i|^{2+\eta}$ is bounded uniformly over n and $Q \in \mathcal{Q}_n$.

While the variance function $\sigma^2(x)$ is unknown, the definition of $\mathcal{Q}_n$ is such that the variance function is the same for all $Q \in \mathcal{Q}_n$. This is done for simplicity. One could consider uniformity over classes $\mathcal{Q}_n$ that place only smoothness conditions on $\sigma^2(x)$ at the cost of introducing additional notation and making the optimality statements more cumbersome.

The estimators and CIs in this section are plug-in versions of procedures in Section 3, where an estimate $\hat{\sigma}(x)$ is used in place of the unknown true variance function. We make the following assumption on this estimate. As discussed above, this assumption holds for the variance estimate used here, as well as allowing for other consistent variance estimates.

Assumption 5.3. *The estimate $\hat{\sigma}(x)$ is given by $\hat{\sigma}(x) = \hat{\sigma}_+(0)I(x > 0) + \hat{\sigma}_-(0)I(x < 0)$ where $\hat{\sigma}_+(0)$ and $\hat{\sigma}_-(0)$ are consistent for $\sigma_+(0)$ and $\sigma_-(0)$ uniformly over $f \in \mathcal{F}$ and $Q \in \mathcal{Q}_n$.*

For asymptotic coverage, we consider uniformity over both $\mathcal{F}$ and $\mathcal{Q}_n$. Thus, a confidence set $\mathcal{C}$ is said to have asymptotic coverage at least $1-\alpha$ if

$$\liminf_{n\to\infty} \inf_{f\in\mathcal{F},Q\in\mathcal{Q}_n} P_{f,Q}\left(Lf \in \mathcal{C}\right) \geq 1-\alpha.$$

Theorem 5.1. *Under Assumptions 5.1, 5.2 and 5.3, the confidence intervals given in Sections 5.3 and 5.4 based on $\hat{L}_\delta$ have asymptotic coverage at least $1-\alpha$. The confidence intervals given in Section 5.4 based on local polynomial estimators have asymptotic coverage at least $1-\alpha$ so long as the kernel is bounded and uniformly continuous with bounded support and the bandwidths h_+ and h_- satisfy $h_+ n^{1/(2p+1)} \to h_{+,\infty}$ and $h_- n^{1/(2p+1)} \to h_{-,\infty}$ for some $h_{+,\infty} > 0$ and $h_{-,\infty} > 0$.*

Let $\hat{\chi}$ denote the half-length of the optimal fixed-length CI based on $\hat{\sigma}(x)$. For χ_∞ given in Supplemental Appendix G, the scaled half-length $n^{p/(2p+1)}\hat{\chi}$ converges in probability to χ_∞ uniformly over $\mathcal{F}$ and $\mathcal{Q}_n$. If, in addition, each $\mathcal{Q}_n$ contains a distribution where the u_is are normal, then for any sequence of confidence sets $\mathcal{C}$ with asymptotic coverage at least $1-\alpha$, we have the following bound on the asymptotic efficiency improvement at any $f \in \mathcal{F}_{RDT}(0)$:

$$\liminf_{n\to\infty} \sup_{Q\in\mathcal{Q}_n} \frac{n^{p/(2p+1)} E_{f,Q}\lambda(\mathcal{C})}{\chi_\infty} \geq \frac{(1-\alpha)E[(z_{1-\alpha}-Z)^{2p/(2p+1)} \mid Z \leq z_{1-\alpha}]}{(c_\chi^{-1}(2p/(2p+1)))^{2p/(2p+1)-1}\chi_{A,\alpha}(c_\chi^{-1}(2p/(2p+1)))}$$

where $Z \sim N(0,1)$.

Letting $\hat{c}_{\alpha,\delta}$ denote the lower endpoint of the one-sided CI corresponding to $\hat{L}_\delta$, the CI $[\hat{c}_{\alpha,\delta}, \infty)$ has asymptotic coverage at least $1-\alpha$. If δ is chosen to minimax the β quantile excess length, (i.e. $\delta = z_\beta + z_{1-\alpha}$), then, if each Q_n contains a distribution where the u_is are normal, any other one-sided CI $[\hat{c}, \infty)$ with asymptotic coverage at least $1-\alpha$ must satisfy the efficiency bound

$$\liminf_{n\to\infty} \frac{\sup_{f\in\mathcal{F},Q\in\mathcal{Q}_n} q_{f,Q,\beta}\left(Lf - \hat{c}\right)}{\sup_{f\in\mathcal{F},Q\in\mathcal{Q}_n} q_{f,Q,\beta}\left(Lf - \hat{c}_{\alpha,\delta}\right)} \geq 1.$$

In addition, we have the following bound on the asymptotic efficiency improvement at any $f \in \mathcal{F}_{RDT}(0)$:

$$\liminf_{n\to\infty} \frac{\sup_{Q\in\mathcal{Q}_n} q_{f,Q,\beta}\left(Lf - \hat{c}\right)}{\sup_{Q\in\mathcal{Q}_n} q_{f,Q,\beta}\left(Lf - \hat{c}_{\alpha,\delta}\right)} \geq \frac{2^{2p/(2p+1)}}{1+2p/(2p+1)}.$$

The proof of Theorem 5.1 is given in Supplemental Appendix G. Theorem 5.1 gives

asymptotic validity of the plug-in optimal procedures in this section, and shows that they are efficient when the class of possible distributions $\mathcal{Q}_n$ contains a normal law. The latter assumption is standard in the literature on efficiency bounds in nonparametric models (see, e.g., Fan, 1993, pp. 205-206), and we leave the question of relaxing it for future research. The asymptotic efficiency bounds correspond to those in Section 3 under (20) with $r = 2p/(2p+1)$.

5.6 Comparison with other methods

A naïve, but popular approach to inference in RD is to form a nominal $100\cdot(1-\alpha)\%$ CI around a local polynomial estimator by adding and subtracting the $1-\alpha/2$ quantile of the $N(0,1)$ distribution times the standard error, thereby ignoring bias. Typically, local linear estimators are used, and the justification is based on the accuracy of a linear approximation, often with a citation to Cheng, Fan, and Marron (1997) or other papers that consider minimax MSE in the class $\mathcal{F}_{T,2}(C;\mathbb{R}_+)$ for estimation of $f(0)$. Thus, it is natural to consider the parameter space $\mathcal{F}_{RDT,2}(C)$ and to ask: "what is the largest value of C for which this CI has good coverage?" Since this method ignores bias, there will always be some undercoverage, so we formalize this by finding the largest value of C such that a nominal 95% CI has true coverage 90%. This calculation is easily done using the results in Section 3.5: the naïve approach uses the critical value $z_{1-.05/2} = \text{cv}_{.05}(0)$ to construct a nominal 95% CI, while a valid 90% CI uses $\text{cv}_{.1}(\overline{\text{bias}}_{\mathcal{F}_{RDT,2}(C)}(\hat{L})/\text{se}(\hat{L}))$ (where $\hat{L}$ denotes the estimator and $\text{se}(\hat{L})$ denotes its standard error), so we equate these two critical values and solve for C.

The resulting value of C for which undercoverage is controlled will depend on the bandwidth. If a sequence of bandwidths h_n is chosen so that $h_n n^{1/5} \to 0$ (the researcher makes an "asymptotic promise" to undersmooth), this will lead to a sequence C_n that increases with the sample size. Alternatively, if one chooses a sequence where $h_n n^{1/5}$ converges to a constant (e.g. the researcher forms a CI around the estimate that is MSE optimal for a fixed value of C), C_n will converge to a constant as well. If the bandwidth choice is data dependent, the estimator is non-linear and computing the value of C analytically is more complicated. We consider this case in a Monte Carlo analysis in Appendix A. To provide a simple numerical comparison to commonly used procedures, we consider the (data dependent) Imbens and Kalyanaraman (2012) bandwidth $\hat{h}_{IK}$, but treat it as if it were fixed a priori. We consider CIs based on the local linear estimator with the triangular kernel and this bandwidth, as well as particular strategies for "undersmoothing" relative to this bandwidth.

For the Lee application, the IK bandwidth selector leads to $\hat{h}_{IK} = 29.4$. The naïve two-sided CI based on this bandwidth is given by 7.99 ± 1.97. Treating the bandwidth as

nonrandom, it achieves coverage of at least 90% over $\mathcal{F}_{RDT,2}(C)$ as long as $C \leq C_{\text{naïve}} = 0.0022$. This is a rather low value, implying that even when $x = 20\%$, the prediction error based on a linear Taylor approximation to f differs at most by 0.9% from the true conditional expectation.

To deal with the coverage problem of the naïve CI (or, equivalently, to relax the high level of smoothness it requires), a popular approach is to undersmooth. As discussed above, this leads to a sequence $C_n \to \infty$ under which size distortion is controlled by a given amount, and our methods can be used to compute this sequence. Another popular approach is to subtract an estimate of the bias. In an important paper, Calonico, Cattaneo, and Titiunik (2014) link these two approaches in the context of RD. They derive a novel standard error formula that accounts for the additional variability introduced by the estimated bias, and show that if the pilot bandwidth and the kernel used by the bias estimator equal those used by the local linear estimator of Lf, their procedure amounts to running a quadratic instead of a linear local regression, and then using the usual CI. In the Lee application, this method delivers the CI 6.68 ± 2.91, increasing the half-length substantially relative to the naïve CI. This increase is due to the fact that a local quadratic estimator uses a much smaller effective sample size than a local linear estimator at the same bandwidth resulting in 330 and 718 effective observations, respectively. The maximum smoothness parameter under which these CIs have coverage at least 90% is given by $C_{CCT} = 0.0027$, which is larger than $C_{\text{naïve}}$. By way of comparison, the optimal 95% fixed-length CIs at C_{CCT} leads to a much narrower CI given by 7.59 ± 2.36.

While the CCT CI maintains good coverage for a larger smoothness constant than the naïve CI ($C_{CCT} > C_{\text{naïve}}$), both constants are rather small (equivalently, coverage is bad for moderate values of C). This is an artifact of the large realized value of $\hat{h}_{IK}$: the CCT CI essentially "undersmooths" relative to a given bandwidth by making the bias-standard deviation ratio smaller. Since $\hat{h}_{IK}$ is large to begin with, the amount of undersmoothing is not enough to make the procedure robust to moderate values of C. In fact, the IK bandwidth is generally quite sensitive to tuning parameter choices: we show in a Monte Carlo study in Appendix A that the CCT implementation of the IK bandwidth yields smaller bandwidths and achieves good coverage over a much larger set of functions, at the cost of larger length. In finite samples, the tuning parameters drive the maximum bias of the estimator, and hence its coverage properties, even though under standard pointwise asymptotics, the tuning parameters shouldn't affect coverage.

In contrast, if one performs the CCT procedure starting from a minimax MSE optimal

bandwidth based on a known smoothness constant C, the asymptotic coverage will be quite good (above 94%), although the CCT CI ends up being about 30% longer than the optimal CI (see Armstrong and Kolesár, 2016). Thus, while using a data driven bandwidth selector such as IK for inference can lead to severe undercoverage for smoothness classes used in RD (even if one undersmooths or bias-corrects as in CCT), procedures such as CCT can have good coverage if based on an appropriate bandwidth choice that is fixed ex ante.

6 Conclusion

This paper considered the problem of constructing one- and two-sided confidences intervals for a linear functional of an unknown regression function f in a broad class of regression models under the assumption that f lies in a convex function class $\mathcal{F}$. We showed that, when $\mathcal{F}$ is centrosymmetric, one-sided CIs that minimax a given quantile of excess length that we derive here are also highly efficient at smooth functions relative to CIs that optimize excess length at these smooth functions. Likewise, the fixed-length two-sided CIs of Donoho (1994) are shown to be highly efficient relative to confidence sets that optimize expected length at smooth functions. Both types of CIs are simple to construct. They require an explicit choice of the function class $\mathcal{F}$, which sometimes involves placing an explicit bound the smoothness of f. The above efficiency results imply, however, that specifying this bound can only be avoided at the expense of sacrificing coverage.

Appendix A Monte Carlo evidence

Corollaries 3.2 and 3.3 imply that confidence intervals based on data-driven bandwidths must either undercover or else cannot be shorter than fixed-length confidence intervals that assume worst-case smoothness. In this appendix, we illustrate this implication with a Monte Carlo study in the context of inference in a sharp regression discontinuity design.

As in Section 5, the data are generated from the nonparametric regression model $y_i = f(x_i) + u_i$, and the parameter of interest is the jump in the regression function at zero, $Lf = \lim_{x\downarrow 0} f(x) - \lim_{x\uparrow 0} f(x)$. To help separate the difficulty in constructing CIs for Lf due to unknown smoothness of f from that due to irregular design points or heteroscedasticity, for all designs below, the distribution of x_i is uniform on $[-1, 1]$, and u_i is independent of x_i, distributed $\mathcal{N}(0, \sigma^2)$. The sample size is $n = 500$ in each case.

For σ^2, we consider two values, $\sigma^2 = 0.1295$, and $\sigma^2 = 4 \times 0.1295 = 0.518$. We consider conditional mean functions f that lie in the smoothness class

$$\mathcal{F}_{RDH,2}(C) = \{f_+ - f_- \colon f_+ \in \mathcal{F}_{H,2}(C; \mathbb{R}_+),\ f_- \in \mathcal{F}_{H,2}(C; \mathbb{R}_-)\},$$

where $\mathcal{F}_{H,p}(C; \mathcal{X})$ is the second-order Hölder class, the closure of twice-differentiable functions with second derivative bounded by $2C$, uniformly over $\mathcal{X}$:

$$\mathcal{F}_{H,p}(C; \mathcal{X}) = \{f \colon |f'(x_1) - f'(x_2)| \leq 2C|x_1 - x_2| \text{ all } x_1, x_2 \in \mathcal{X}\}.$$

Unlike the class $\mathcal{F}_{RDT,2}(C)$ considered in Section 5, the class $\mathcal{F}_{RDH,2}(C)$ also imposes smoothness away from the cutoff, so that $\mathcal{F}_{RDH,2}(C) \subseteq \mathcal{F}_{RDT,2}(C)$. Imposing smoothness away from the cutoff is natural in many empirical applications. We consider $C = 1$ and $C = 3$, and for each C, we consider 4 different shapes for f. In each case, f is odd, $f_+ = -f_-$. In Designs 1 through 3, f_+ is given by a quadratic spline with two knots, at b_1 and b_2,

$$f_+(x) = 1(x > 0) \cdot C \left(x^2 - 2(x - b_1)_+^2 + 2(x - b_2)_+^2\right).$$

In Design 1 the knots are given by $(b_1, b_2) = (0.45, 0.75)$, in Design 2 by $(0.25, 0.65)$, and in Design 3 by $(0.4, 0.9)$. The function $f_+(x)$ is plotted in Figure 6 for $C = 1$. For $C = 3$, the function f is identical up to scale. It is clear from the figure that although locally to the cutoff, the functions are identical, they differ away from the cutoff (for $|x| \geq 0.25$), which, as we demonstrate below, affects the performance of data-driven methods. Finally, in Design 4, we consider $f(x) = 0$ to allow us to compare the performance of CIs when f is as smooth

as possible.

We consider four methods for constructing CIs based on data-driven bandwidths, and two fixed-length CIs. All CIs are based on local polynomial regressions with a triangular kernel. The variance estimators used to construct the CIs are heteroscedasticity-robust and based on the nearest neighbor method proposed by Abadie and Imbens (2006) in a different context. This method was also studied in Calonico, Cattaneo, and Titiunik (2014) in an RD setting. The results based on Eicker-Huber-White variance estimators are very similar and not reported here.

The first two methods correspond to naïve CIs based on local linear regression described in Section 5.6. The first CI uses Imbens and Kalyanaraman (2012, IK) bandwidth selector $\hat{h}_{IK}$, and the second CI uses a bandwidth selector proposed in Calonico, Cattaneo, and Titiunik (2014, CCT), $\hat{h}_{CCT}$. The third CI uses the robust bias correction (RBC) studied in CCT, with both the pilot and the main bandwidth given by $\hat{h}_{IK}$ (the main estimate is based on local linear regression, and the bias correction is based on local quadratic regression), so that the bandwidth ratio is given by $\rho = 1$. The fourth CI is also based on RBC, but with the main and pilot bandwidth potentially different and given by the Calonico, Cattaneo, and Titiunik (2014) bandwidth selectors. Finally, we consider two fixed-length CIs with uniform coverage under the class $\mathcal{F}_{RDH,2}(C)$, with $C = 1, 3$, and bandwidth chosen to minimize their half-length. Their construction is similar to the CIs considered in Section 5.4, except they use the fact that under $\mathcal{F}_{RDH,2}(C)$, the maximum bias for local linear estimators based on a fixed bandwidth is attained at $g^*(x) = Cx^2 1(x > 0) - Cx^2 1(x < 0)$ (see Armstrong and Kolesár, 2016, for derivation).

The results are reported in Tables 2 for $C = 1$ and 3 for $C = 3$. One can see from the tables that CIs based on $\hat{h}_{IK}$ may undercover severely even at the higher level of smoothness, $C = 1$. In particular, the coverage of naïve CIs based on $\hat{h}_{IK}$ is as low as 10.1% for 95% nominal CIs in Design 1, and the coverage of RBC CIs is as low as 64.4%, again in Design 1. The undercoverage is even more severe when $C = 3$.

In contrast, CIs based on the CCT bandwidth selector perform much better in terms of coverage under $C = 1$, with coverage over 90% for all designs. These CIs only start undercovering once $C = 3$, with 80.7% coverage in Design 3 for naïve CIs, and 86.2% coverage for RBC CIs. The cost for the good coverage properties, as can be seen from the tables, is that the CIs are longer, sometimes much longer than optimal fixed-length CIs. As discussed in Section 5.6, the dramatically different coverage properties of the CIs based on the IK and CCT bandwidths illustrates the point that the coverage of CIs based on data-

driven bandwidths is governed by the tuning parameters used in defining the bandwidth selector.

Appendix B Proofs for main results

This section contains proofs of the results in Section 3. Section B.1 contains auxiliary lemmas used in the proofs. The proofs of the results in Section 3 are given in the remainder of the section. Section B.2 contains the proof of Theorem 3.1. Section B.3 contains the proof of Theorem 3.3. The corollaries to these theorems follow immediately from the theorems and arguments in the main text, and their proofs are omitted from this section. Theorem 3.4 is immediate from Lemma B.1 below (which, as discussed below, reduces to Lemma 4 in Donoho (1994) in this case, since $\mathcal{G} = \mathcal{F}$).

Before proceeding, we recall that $\omega'(\delta; \mathcal{F}, \mathcal{G})$ was defined in Section 3 to be an arbitrary element of the superdifferential. Here, we introduce notation to denote this set. The superdifferential is defined as

$$\partial\omega(\delta; \mathcal{F}, \mathcal{G}) = \{d\text{: for all } \eta > 0, \omega(\eta; \mathcal{F}, \mathcal{G}) \leq \omega(\delta; \mathcal{F}, \mathcal{G}) + d(\eta - \delta)\}\,.$$

It is nonempty since $\omega(\cdot; \mathcal{F}, \mathcal{G})$ is concave (if f_δ^*, g_δ^* attain the modulus at δ and similarly for $\tilde{\delta}$, then, for $\lambda \in [0,1]$, $f_\lambda = \lambda f_\delta^* + (1-\lambda) f_{\tilde{\delta}}^*$ and $g_\lambda = \lambda g_\delta^* + (1-\lambda) g_{\tilde{\delta}}^*$ satisfy $\|K(g_\lambda - f_\lambda)\| \leq \lambda\delta + (1-\lambda)\tilde{\delta}$ so that $\omega(\lambda\delta + (1-\lambda)\tilde{\delta}) \geq Lg_\lambda - Lf_\lambda = \lambda\omega(\delta) + (1-\lambda)\omega(\tilde{\delta})$).

B.1 Auxiliary Lemmas

The following lemma extends Lemma 4 in Donoho (1994) to the two class modulus (see also Theorem 2 in Cai and Low, 2004b, for a similar result in the Gaussian white noise model). The proof is essentially the same as for the single class case.

Lemma B.1. *Let f^* and g^* solve the optimization problem for $\omega(\delta_0; \mathcal{F}, \mathcal{G})$ with $\|K(f^* - g^*)\| = \delta_0$, and let $d \in \partial\omega(\delta_0; \mathcal{F}, \mathcal{G})$. Then, for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,*

$$Lg - Lg^* \leq d\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|} \text{ and } Lf - Lf^* \geq d\frac{\langle K(g^* - f^*), K(f - f^*)\rangle}{\|K(g^* - f^*)\|}. \quad (29)$$

In particular, the test statistic $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ achieves maximum bias over $\mathcal{F}$ at f^* and minimum bias over $\mathcal{G}$ at g^*.

Proof. In this proof, we use $\omega(\delta)$ to denote the ordered modulus $\omega(\delta; \mathcal{F}, \mathcal{G})$. Suppose that the first inequality does not hold for some g. Then, for some $\varepsilon > 0$,

$$Lg - Lg^* > (d + \varepsilon) \frac{\langle K(g^* - f^*), K(g - g^*) \rangle}{\|K(g^* - f^*)\|}. \tag{30}$$

Let $g_\lambda = (1 - \lambda)g^* + \lambda g$. Since $g_\lambda - g^* = \lambda(g - g^*)$, multiplying by λ gives

$$Lg_\lambda - Lg^* > \lambda(d + \varepsilon) \frac{\langle K(g^* - f^*), K(g - g^*) \rangle}{\|K(g^* - f^*)\|}.$$

The left hand side is equal to $Lg_\lambda - Lf^* - L(g^* - f^*) = Lg_\lambda - Lf^* - \omega(\delta_0)$. Since $g_\lambda \in \mathcal{G}$ by convexity, $Lg_\lambda - Lf^* \leq \omega(\|K(g_\lambda - f^*)\|)$. Note that

$$\frac{d}{d\lambda} \|K(g_\lambda - f^*)\| \Big|_{\lambda=0} = \frac{1}{2} \frac{\frac{d}{d\lambda} \|K(g_\lambda - f^*)\|^2 \Big|_{\lambda=0}}{\|K(g^* - f^*)\|} = \frac{\langle K(g^* - f^*), K(g - g^*) \rangle}{\|K(g^* - f^*)\|} \tag{31}$$

so that $\|K(g_\lambda - f^*)\| = \delta_0 + \lambda \frac{\langle K(g^* - f^*), K(g - g^*) \rangle}{\|K(g^* - f^*)\|} + o(\lambda)$. Putting this all together, we have

$$\omega \left(\delta_0 + \lambda \frac{\langle K(g^* - f^*), K(g - g^*) \rangle}{\|K(g^* - f^*)\|} + o(\lambda) \right) > \omega(\delta_0) + \lambda(d + \varepsilon) \frac{\langle K(g^* - f^*), K(g - g^*) \rangle}{\|K(g^* - f^*)\|},$$

which is a contradiction unless $\langle K(g^* - f^*), K(g - g^*) \rangle = 0$.

If $\langle K(g^* - f^*), K(g - g^*) \rangle = 0$, then (30) gives $Lg - Lg^* > 0$, which implies

$$\omega(\|K(g_\lambda - f^*)\|) \geq Lg_\lambda - Lf^* = \lambda c + \omega(\delta_0)$$

where $c = Lg - Lg^* > 0$. But in this case (31) implies $\|K(g_\lambda - f^*)\| = \delta_0 + o(\lambda)$, again giving a contradiction. This proves the first inequality, and a symmetric argument applies to the inequality involving $Lf - Lf^*$, thereby giving the first result.

Now consider the test statistic $\hat{L}_{\delta, \mathcal{F}, \mathcal{G}}$. Under $g \in \mathcal{G}$, the bias of this statistic is equal to a constant that does not depend on g plus

$$d \frac{\langle K(g^* - f^*), K(g - g^*) \rangle}{\|K(g^* - f^*)\|} - (Lg - Lg^*).$$

It follows from (29) that this is minimized over $g \in \mathcal{G}$ by taking $g = g^*$. Similarly, the maximum bias over $\mathcal{F}$ is taken at f^*. □

The next lemma is a result from the literature on nonparametric testing. It is used in the proof of Theorem 3.3.

Lemma B.2. *Let $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$ be convex sets, and suppose that f^* and g^* minimize $\|K(f-g)\|$ over $f \in \tilde{\mathcal{F}}$ and $g \in \tilde{\mathcal{G}}$. Then, for any level α, the minimax test of $H_0 : \tilde{\mathcal{F}}$ vs $H_1 : \tilde{\mathcal{G}}$ is given by the Neyman-Pearson test of f^* vs g^*. It rejects when $\langle K(f^* - g^*), Y\rangle$ is greater than its $1-\alpha$ quantile under f^*. The minimum power of this test over $\tilde{\mathcal{G}}$ is taken at g^*.*

Proof. The result is immediate from results stated in Section 2.4.3 in Ingster and Suslina (2003), since the sets $\{Kf \colon f \in \tilde{\mathcal{F}}\}$ and $\{Kg \colon g \in \tilde{\mathcal{G}}\}$ are convex. □

The following lemma derives the form of the derivative of ω under translation invariance, and is used in deriving the form of $\hat{L}_\delta$ given in the main text.

Lemma B.3. *Let f^* and g^* solve the modulus problem with $\delta_0 = \|K(g^* - f^*)\| > 0$, and suppose that $f^* + c\iota \in \mathcal{F}$ for all c in a neighborhood of zero, where $L\iota = 1$. Then $\partial\omega(\delta_0; \mathcal{F}, \mathcal{G}) = \left\{ \frac{\delta_0}{\langle K(g^*-f^*), K\iota\rangle} \right\}$.*

Proof. Let $d \in \partial\omega(\delta_0; \mathcal{F}, \mathcal{G})$ and let $f_c = f^* - c\iota$. Let η be small enough so that $f_c \in \mathcal{F}$ for $|c| \leq \eta$. Then, for $|c| \leq \eta$,

$$L(g^* - f^*) + d\left[\|K(g^* - f_c)\| - \delta_0\right] \geq \omega(\|K(g^* - f_c)\|; \mathcal{F}, \mathcal{G}) \geq L(g^* - f_c) = L(g^* - f^*) + c$$

where the first inequality follows from the definition of the superdifferential and the second inequality follows from the definition of the modulus. Since the left hand side of the above display is greater than or equal to the right hand side of the display for c in a neighborhood of zero, and the two sides are equal at $c = 0$, the derivatives of both sides with respect to c must be equal. The result follows since

$$\frac{d}{dc}\|K(g^* - f_c)\|\Big|_{c=0} = \frac{\frac{d}{dc}\|K(g^* - f_c)\|^2\Big|_{c=0}}{2\delta_0} = \frac{\langle K(g^* - f^*), K\iota\rangle}{\delta_0}.$$

□

B.2 Proof of Theorem 3.1

For ease of notation in this proof, let $f^* = f^*_\delta$ and $g^* = g^*_\delta$ denote the functions that solve the modulus problem with $\|K(f^* - g^*)\| = \delta$, and let $d = \omega'(\delta; \mathcal{F}, \mathcal{G}) \in \partial\omega(\delta; \mathcal{F}, \mathcal{G})$ so that

$$\hat{c}_\alpha = \hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}} = Lf^* + d\frac{\langle K(g^* - f^*), KY\rangle}{\|K(g^* - f^*)\|} - d\frac{\langle K(g^* - f^*), Kf^*\rangle}{\|K(g^* - f^*)\|} - z_{1-\alpha}\sigma d.$$

Note that $\hat{c}_\alpha = \hat{L}_{\delta,\mathcal{F},\mathcal{G}} + a$ for a chosen so that the $1-\alpha$ quantile of $\hat{c}_\alpha - Lf^*$ under f^* is zero. Thus, it follows from Lemma B.1 that $[\hat{c}_\alpha, \infty)$ is a valid $1-\alpha$ CI for Lf over $\mathcal{F}$, and that all quantiles of excess coverage $Lg - \hat{c}_\alpha$ are maximized over $\mathcal{G}$ at g^*. In particular, $q_\beta(\hat{c}_\alpha; \mathcal{G}) = q_{g^*,\beta}(Lg^* - \hat{c}_\alpha)$. To calculate this, note that, under g^*, $Lg^* - \hat{c}_\alpha$ is normal with variance $d^2\sigma^2$ and mean

$$Lg^* - Lf^* - d\frac{\langle K(g^* - f^*), K(g^* - f^*)\rangle}{\|K(g^* - f^*)\|} + z_{1-\alpha}\sigma d = \omega(\delta; \mathcal{F}, \mathcal{G}) + d(z_{1-\alpha}\sigma - \delta).$$

The probability that this normal variable is less than or equal to $\omega(\delta; \mathcal{F}, \mathcal{G})$ is given by the probability that a normal variable with mean $d(z_{1-\alpha}\sigma - \delta)$ and variance $d^2\sigma^2$ is less than or equal to zero, which is $\Phi(\delta/\sigma - z_{1-\alpha}) = \beta$. Thus $q_\beta(\hat{c}_\alpha; \mathcal{G}) = \omega(\delta; \mathcal{F}, \mathcal{G})$ as claimed.

It remains to show that no other $1-\alpha$ CI can strictly improve on this. Suppose that some other $1-\alpha$ CI $[\tilde{c}, \infty)$ obtained a strictly shorter βth quantile of excess length for all $g \in \mathcal{G}$. Applying this with $g = g^*$, we would have, for some $\eta > 0$,

$$P_{g^*}(Lg^* - \tilde{c} \leq \omega(\delta; \mathcal{F}, \mathcal{G}) - \eta) \geq \beta.$$

Let $\tilde{f}$ be given by a convex combination between g^* and f^* such that $Lg^* - L\tilde{f} = \omega(\delta; \mathcal{F}; \mathcal{G}) - \eta/2$. Then the above display gives

$$P_{g^*}(\tilde{c} > L\tilde{f}) \geq P_{g^*}(\tilde{c} \geq L\tilde{f} + \eta/2) = P_{g^*}(Lg^* - \tilde{c} \leq Lg^* - L\tilde{f} - \eta/2) \geq \beta.$$

But this would imply that the test that rejects when $\tilde{c} > L\tilde{f}$ is level α for $H_0 : \tilde{f}$ and has power β at g^*. This can be seen to be impossible by calculating the power of the Neyman-Pearson test of $\tilde{f}$ vs g^*, since β is the power of the Neyman-Pearson test of f^* vs g^*, and $\tilde{f}$ is a strict convex combination of these functions.

B.3 Proof of Theorem 3.3

Following Pratt (1961), note that, for any confidence set $\mathcal{C}$ for $\vartheta = Lf$, we have

$$E_g\lambda(\mathcal{C}) = E_g \int (1 - \phi_{\mathcal{C}}(\vartheta))\, d\vartheta = \int E_g(1 - \phi_{\mathcal{C}}(\vartheta))\, d\vartheta$$

by Fubini's theorem, where $\phi_{\mathcal{C}}(\vartheta) = I(\vartheta \notin \mathcal{C})$. Thus, the CI that minimizes this inverts the family of most powerful tests of $H_0 \colon Lf = \vartheta, f \in \mathcal{F}$ against $H_1 \colon f = g$. By Lemma B.2 since the sets $\{f \colon Lf = \vartheta, f \in \mathcal{F}\}$ and $\{g\}$ are convex, the least favorable function f_ϑ minimize $\|K(g - f)\|$ subject to $Lf = \vartheta$, which gives the first part of the theorem.

To derive the expression for expected length, note that if $Lg \leq \vartheta$, then the minimization problem is equivalent to solving the inverse ordered modulus problem $\omega^{-1}(\vartheta - Lg; \{g\}, \mathcal{F})$, and if $Lg \geq \vartheta$, it is equivalent to solving $\omega^{-1}(Lg - \vartheta; \mathcal{F}, \{g\})$. This follows because if the ordered modulus $\omega(\delta; \mathcal{F}, \{g\})$ attained at some f_δ^* and g, then the inequality $\|K(f - g)\| \leq \delta$ must be binding: otherwise a convex combination of $\tilde{f}$ and f_δ^*, where $\tilde{f}$ is such that $L(g - f_\delta^*) < L(g - \tilde{f})$ would achieve a strictly larger value, and similarly for $\omega(\delta; \{g\}, \mathcal{F})$. Such $\tilde{f}$ always exists since by the assumption that f_ϑ exists for all ϑ. Consequently, it also follows that that the modulus and inverse modulus are strictly increasing.

Next, it follows from the proof of Theorem 3.1 that the power of the test ϕ_ϑ at g is given by $\Phi(\delta_\vartheta/\sigma - z_{1-\alpha})$. Therefore,

$$\begin{aligned} E_g[\lambda(\mathcal{C}_g(Y))] &= \int \Phi\left(z_{1-\alpha} - \frac{\delta_\vartheta}{\sigma}\right) \mathrm{d}\vartheta \\ &= \iint 1\left(\delta_\vartheta \leq \sigma(z_{1-\alpha} - z)\right) \mathrm{d}\vartheta\, \mathrm{d}\Phi(z), \end{aligned}$$

where the second line swaps the order of integration. Splitting the inner integral, using fact that $\delta_\vartheta = \omega^{-1}(Lg - \vartheta; \mathcal{F}, \{g\})$ for $\vartheta \leq Lg$ and $\delta_\vartheta = \omega^{-1}(\vartheta - Lg; \{g\}, \mathcal{F})$ for $\vartheta \geq Lg$, and taking a modulus on both sides of the inequality of the integrand then yields

$$\begin{aligned} E_g[\lambda(\mathcal{C}_g(Y))] &= \iint_{\vartheta \leq Lg} 1\left(Lg - \vartheta \leq \omega\left(\sigma(z_{1-\alpha} - z); \mathcal{F}, \{g\}\right)\right) 1(z \leq z_{1-\alpha})\, \mathrm{d}\vartheta\, \mathrm{d}\Phi(z) \\ &\quad + \iint_{\vartheta > Lg} 1\left(\vartheta - Lg \leq \omega\left(\sigma(z_{1-\alpha} - z); \{g\}, \mathcal{F}\right)\right) 1(z \leq z_{1-\alpha})\, \mathrm{d}\vartheta\, \mathrm{d}\Phi(z) \\ &= (1 - \alpha) E\left[\left(\omega(\sigma(z_{1-\alpha} - Z); \mathcal{F}, \{g\}) + \omega(\sigma(z_{1-\alpha} - Z); \{g\}, \mathcal{F})\right) \mid Z \leq z_{1-\alpha}\right], \end{aligned}$$

where Z is standard normal, which yields the result.

Appendix C Fixed-length CIs in the bounded normal mean model

Let $Z \sim N(\mu, 1)$, with $\mu \in [-\tau, \tau]$. Consider first the problem of finding an affine estimator $\hat{\mu} = b + cY$ that maximizes the coverage of the confidence intervals $\hat{\mu} \pm \chi$, with χ given. This problem is equivalent to minimizing the maximum risk under the 0–1 loss $\ell_{0\text{–}1,\chi}(\hat{\mu}, \mu) = 1(|\hat{\mu} - \mu| \geq \chi))$. By symmetry, it is clear that $b = 0$ is optimal. The coverage of a linear estimator cZ if $c \neq 0$ is

$$1 - \sup_{|\mu| \leq \tau} E_\mu[\ell_{0\text{–}1,\chi}(cZ, \mu)] = \Phi\left(\frac{\tau + \chi}{c} - \tau\right) - \Phi\left(\frac{\tau - \chi}{c} - \tau\right). \tag{32}$$

Drees (1999) shows that the affine estimator that is minimax for the 0–1 loss is given by

$$c_{0\text{–}1,\chi}(\tau) = \begin{cases} \left(1/2 + \sqrt{\frac{1}{4} + \frac{1}{2\tau\chi}\log\left(\frac{\tau+\chi}{\tau-\chi}\right)}\right)^{-1} & \text{if } \chi < \tau, \\ 0 & \text{otherwise.} \end{cases}$$

so that the coverage is 1 if $\chi \geq \tau$ and it is given by (32) otherwise. Hence, the shortest fixed-length $1 - \alpha$ affine confidence interval is given by $c_\chi(\tau)Z \pm \chi_{A,\alpha}(\tau)$, where $c_\chi(\tau) = c_{0\text{–}1,\chi_{A,\alpha}(\tau)}(\tau)$, and $\chi_{A,\alpha}(\tau)$ solves $E[\ell_{0\text{–}1,\chi}(c_{0\text{–}1,\chi}(\tau))] = \alpha$ if $\tau \geq \Phi^{-1}(1-\alpha)$, and $\chi_{A,\alpha}(\tau) = \tau$ otherwise.

References

Abadie, A., and G. W. Imbens (2006): "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74(1), 235–267.

Andrews, D. W. K. (2001): "Testing When a Parameter is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69(3), 683–734.

Andrews, D. W. K., and P. Guggenberger (2009a): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77(3), 721–762.

——— (2009b): "Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities," *Econometric Theory*, 25(03), 669–709.

Armstrong, T. B. (2015): "Adaptive testing on a regression function at a point," *The Annals of Statistics*, 43(5), 2086–2101.

Armstrong, T. B., and M. Kolesár (2016): "Simple and honest confidence intervals in nonparametric regression," Unpublished manuscript.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81(2), 608–650.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90(430), 443 – 450.

Brown, L. D., and M. G. Low (1996): "Asymptotic equivalence of nonparametric regression and white noise," *Annals of Statistics*, 24(6), 2384–2398.

Brown, L. D., M. G. Low, and L. H. Zhao (1997): "Superefficiency in nonparametric function estimation," *The Annals of Statistics*, 25(6), 2607–2625.

Cai, T. T., and Z. Guo (2015): "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," Discussion paper, arXiv: 1506.05539.

Cai, T. T., and M. G. Low (2004a): "An adaptation theory for nonparametric confidence intervals," *Annals of Statistics*, 32(5), 1805–1840.

——— (2004b): "Minimax estimation of linear functionals over nonconvex parameter spaces," *Annals of Statistics*, 32(2), 552–576.

Cai, T. T., M. G. Low, and Z. Ma (2014): "Adaptive Confidence Bands for Nonparametric Regression Functions," *Journal of the American Statistical Association*, 109(507), 1054–1070.

Cai, T. T., M. G. Low, and Y. Xia (2013): "Adaptive confidence intervals for regression functions under shape constraints," *The Annals of Statistics*, 41(2), 722–750.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2015): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," working paper.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): "On automatic boundary corrections," *The Annals of Statistics*, 25(4), 1691–1708.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Anti-concentration and honest, adaptive confidence bands," *The Annals of Statistics*, 42(5), 1787–1818.

Cohen, J. (1988): *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Donoho, D. L. (1994): "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 22(1), 238–270.

Donoho, D. L., and R. C. Liu (1991): "Geometrizing Rates of Convergence, III," *The Annals of Statistics*, 19(2), 668–701.

Donoho, D. L., and M. G. Low (1992): "Renormalization Exponents and Optimal Pointwise Rates of Convergence," *The Annals of Statistics*, 20(2), 944–970.

Drees, H. (1999): "On fixed-length confidence intervals for a bounded normal mean," *Statistics & Probability Letters*, 44(4), 399–404.

Fan, J. (1993): "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21(1), 196–216.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.

Gao, W. (2016): "Minimax linear estimation at a boundary point," Discussion paper, Unpublished manuscript, Yale University.

Giné, E., and R. Nickl (2010): "Confidence bands in density estimation," *The Annals of Statistics*, 38(2), 1122–1170.

Hall, P., and J. Horowitz (2013): "A simple bootstrap method for constructing nonparametric confidence bands for functions," *The Annals of Statistics*, 41(4), 1892–1921.

Ibragimov, I. A., and R. Z. Khas'minskii (1985): "On Nonparametric Estimation of the Value of a Linear Functional in Gaussian White Noise," *Theory of Probability & Its Applications*, 29(1), 18–32.

Ichimura, H., and P. E. Todd (2007): "Implementing nonparametric and semiparametric estimators," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6B of *Handbook of Econometrics*, pp. 5369–5468. Elsevier.

Imbens, G. W., and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of Economic Studies*, 79(3), 933–959.

Imbens, G. W., and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142(2), 615–635.

Ingster, Y. I., and I. A. Suslina (2003): *Nonparametric goodness-of-fit testing under Gaussian models*. Springer.

Klemelä, J. (2014): *Multivariate Nonparametric Regression and Visualization: With R and Applications to Finance*. Wiley.

Lee, D. S. (2008): "Randomized experiments from non-random selection in U.S. House elections," *Journal of Econometrics*, 142(2), 675–697.

Lee, D. S., and D. Card (2008): "Regression discontinuity inference with specification error," *Journal of Econometrics*, 142(2), 655–674.

Leeb, H., and B. M. Pötscher (2005): "Model selection and inference: Facts and fiction," *Econometric Theory*, 21(1), 21–59.

Legostaeva, I. L., and A. N. Shiryaev (1971): "Minimax weights in a trend detection problem of a random process," *Theory of Probability & Its Applications*, 16(2), 344–349.

Lehmann, E. L., and J. P. Romano (2005): *Testing statistical hypotheses*. Springer, third edn.

Li, K.-C. (1982): "Minimaxity of the Method of Regularization of Stochastic Processes," *The Annals of Statistics*, 10(3), 937–942.

Low, M. G. (1997): "On nonparametric confidence intervals," *The Annals of Statistics*, 25(6), 2547–2554.

McCloskey, A. (2012): "Bonferroni-Based Size-Correction for Nonstandard Testing Problems," Unpublished Manuscript, Brown University.

Nussbaum, M. (1996): "Asymptotic equivalence of density estimation and Gaussian white noise," *The Annals of Statistics*, 24(6), 2399–2430.

Pratt, J. W. (1961): "Length of confidence intervals," *Journal of the American Statistical Association*, 56(295), 549–567.

Robins, J., and A. van der Vaart (2006): "Adaptive nonparametric confidence sets," *The Annals of Statistics*, 34(1), 229–253.

Sacks, J., and D. Ylvisaker (1978): "Linear Estimation for Approximately Linear Models," *The Annals of Statistics*, 6(5), 1122–1137.

Schennach, S. M. (2015): "A bias bound approach to nonparametric inference," Discussion paper, Cemmap working paper CWP71/15.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

Stone, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8(6), 1348–1360.

Sun, Y. (2005): "Adaptive Estimation of the Regression Discontinuity Model," working paper.

Tsybakov, A. B. (2009): *Introduction to nonparametric estimation*. Springer.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014): "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, 42(3), 1166–1202.

Zhang, C.-H., and S. S. Zhang (2014): "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 217–242.

	α		
b	0.01	0.05	0.1
0.0	2.576	1.960	1.645
0.1	2.589	1.970	1.653
0.2	2.626	1.999	1.677
0.3	2.683	2.045	1.717
0.4	2.757	2.107	1.772
0.5	2.842	2.181	1.839
0.6	2.934	2.265	1.916
0.7	3.030	2.356	2.001
0.8	3.128	2.450	2.093
0.9	3.227	2.548	2.187
1.0	3.327	2.646	2.284
1.5	3.826	3.145	2.782
2.0	4.326	3.645	3.282

Table 1: Critical values $\mathrm{cv}_\alpha(b)$ for selected confidence levels and values of maximum absolute bias b. For $b \geq 2$, $\mathrm{cv}_\alpha(b) \approx b + z_{1-\alpha}$ up to 3 decimal places for these values of α.

	$\sigma^2 = 0.1295$			$\sigma^2 = 4 \cdot 0.1295$		
CI method	Cov. (%)	Bias	RL	Cov. (%)	Bias	RL
Design 1, $(b_1, b_2) = (0.45, 0.75)$						
Naïve, $\hat{h}_{IK}$	10.1	-0.098	0.54	81.7	-0.099	0.72
RBC, $\hat{h}_{IK}, \rho = 1$	64.4	-0.049	0.80	93.9	-0.050	1.06
Naïve, $\hat{h}_{CCT}$	91.2	-0.010	1.01	92.7	-0.010	1.26
RBC, $\hat{h}_{CCT}$	93.7	0.003	1.18	93.6	0.007	1.49
FLCI, $C = 1$	94.6	-0.023	1	94.8	-0.069	1
FLCI, $C = 3$	96.6	-0.009	1.25	96.4	-0.028	1.25
Design 2, $(b_1, b_2) = (0.4, 0.9)$						
Naïve, $\hat{h}_{IK}$	54.2	-0.063	0.68	89.6	-0.085	0.77
RBC, $\hat{h}_{IK}, \rho = 1$	94.8	-0.006	1.00	95.9	-0.043	1.13
Naïve, $\hat{h}_{CCT}$	91.4	-0.009	1.02	92.7	-0.009	1.26
RBC, $\hat{h}_{CCT}$	93.6	0.003	1.19	93.6	0.007	1.49
FLCI, $C = 1$	94.6	-0.023	1	95	-0.065	1
FLCI, $C = 3$	96.6	-0.009	1.25	96.4	-0.028	1.25
Design 3, $(b_1, b_2) = (0.25, 0.65)$						
Naïve, $\hat{h}_{IK}$	87.8	-0.030	0.74	91.4	-0.009	0.76
RBC, $\hat{h}_{IK}, \rho = 1$	94.8	-0.014	1.09	95.0	-0.044	1.12
Naïve, $\hat{h}_{CCT}$	90.9	-0.014	0.97	92.8	-0.013	1.25
RBC, $\hat{h}_{CCT}$	92.2	-0.009	1.14	93.5	-0.007	1.48
FLCI, $C = 1$	94.8	-0.022	1	96.5	-0.028	1
FLCI, $C = 3$	96.6	-0.009	1.25	96.5	-0.025	1.25
Design 4, $f(x) = 0$						
Naïve, $\hat{h}_{IK}$	93.2	0.000	0.54	93.2	-0.001	0.72
RBC, $\hat{h}_{IK}, \rho = 1$	95.2	0.000	0.80	95.2	0.001	1.06
Naïve, $\hat{h}_{CCT}$	93.1	0.001	0.94	93.1	0.003	1.25
RBC, $\hat{h}_{CCT}$	93.5	0.001	1.12	93.5	0.004	1.48
FLCI, $C = 1$	96.8	0.001	1	96.9	0.000	1
FLCI, $C = 3$	96.8	0.001	1.25	96.7	0.002	1.25

Table 2: Monte Carlo simulation, $C = 1$. Coverage ("Cov") and relative length relative to optimal fixed-length confidence interval for $\mathcal{F}_{RDH,2}(1)$ ("RL"). "Bias" refers to bias of estimator around which CI is centered. 11,000 simulation draws.

	$\sigma^2 = 0.1295$			$\sigma^2 = 4 \cdot 0.1295$		
CI method	Cov. (%)	Bias	RL	Cov. (%)	Bias	RL
Design 1, $(b_1, b_2) = (0.45, 0.75)$						
Naïve, $\hat{h}_{IK}$	0.1	-0.292	0.45	22.4	-0.296	0.58
RBC, $\hat{h}_{IK}, \rho = 1$	27.1	-0.127	0.65	77.8	-0.149	0.85
Naïve, $\hat{h}_{CCT}$	89.3	-0.019	0.94	91.6	-0.031	1.05
RBC, $\hat{h}_{CCT}$	93.7	0.004	1.06	93.7	0.012	1.22
FLCI, $C = 1$	71.5	-0.071	0.80	73.5	-0.208	0.80
FLCI, $C = 3$	94.3	-0.029	1	94.6	-0.088	1
Design 2, $(b_1, b_2) = (0.4, 0.9)$						
Naïve, $\hat{h}_{IK}$	60.0	-0.071	0.71	71.4	-0.193	0.72
RBC, $\hat{h}_{IK}, \rho = 1$	93.5	0.000	1.04	95.1	-0.020	1.06
Naïve, $\hat{h}_{CCT}$	89.7	-0.018	0.95	91.7	-0.029	1.05
RBC, $\hat{h}_{CCT}$	93.6	0.004	1.09	93.6	0.012	1.24
FLCI, $C = 1$	71.5	-0.071	0.80	76.4	-0.195	0.80
FLCI, $C = 3$	94.3	-0.029	1	94.6	-0.088	1
Design 3, $(b_1, b_2) = (0.25, 0.65)$						
Naïve, $\hat{h}_{IK}$	79.9	-0.052	0.76	89.2	-0.085	0.73
RBC, $\hat{h}_{IK}, \rho = 1$	93.3	0.001	1.13	94.6	-0.072	1.07
Naïve, $\hat{h}_{CCT}$	80.7	-0.032	0.87	91.8	-0.042	1.01
RBC, $\hat{h}_{CCT}$	86.2	-0.017	1.00	92.7	-0.027	1.20
FLCI, $C = 1$	74.0	-0.068	0.8	93.8	-0.083	0.80
FLCI, $C = 3$	94.3	-0.029	1	95.1	-0.078	1
Design 5, $f(x) = 0$						
Naïve, $\hat{h}_{IK}$	93.2	0.000	0.43	93.2	-0.001	0.57
RBC, $\hat{h}_{IK}, \rho = 1$	95.2	0.000	0.64	95.2	0.001	0.85
Naïve, $\hat{h}_{CCT}$	93.1	0.001	0.75	93.1	0.003	1.00
RBC, $\hat{h}_{CCT}$	93.5	0.001	0.89	93.5	0.004	1.18
FLCI, $C = 1$	96.8	0.001	0.80	96.9	0.000	0.80
FLCI, $C = 3$	96.8	0.001	1	96.7	0.002	1

Table 3: Monte Carlo simulation, $C = 3$. Coverage ("Cov") and relative length relative to optimal fixed-length confidence interval for $\mathcal{F}_{RDH,2}(1)$ ("RL"). "Bias" refers to bias of estimator around which CI is centered. 11,000 simulation draws.

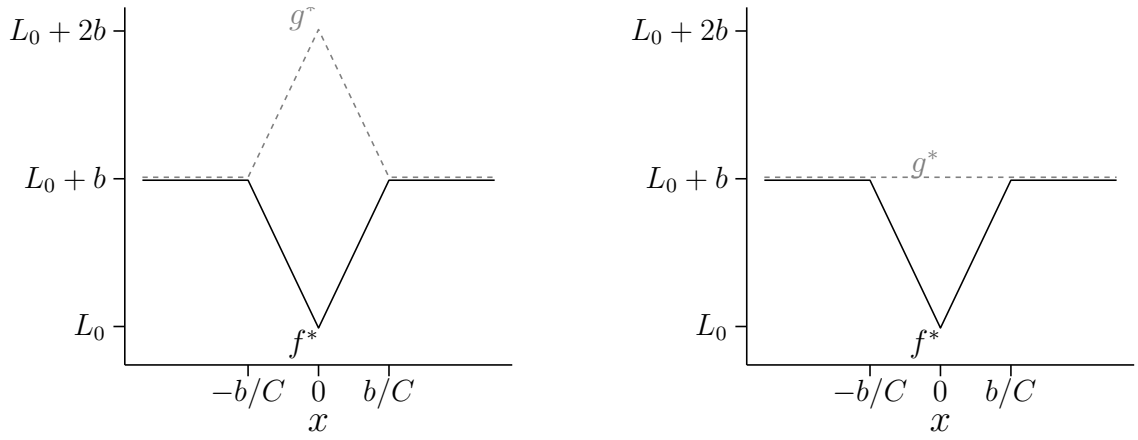


Figure 1: Least favorable null and alternative functions in the simple example for the minimax test (left), and test that directs power at constant alternatives (right).

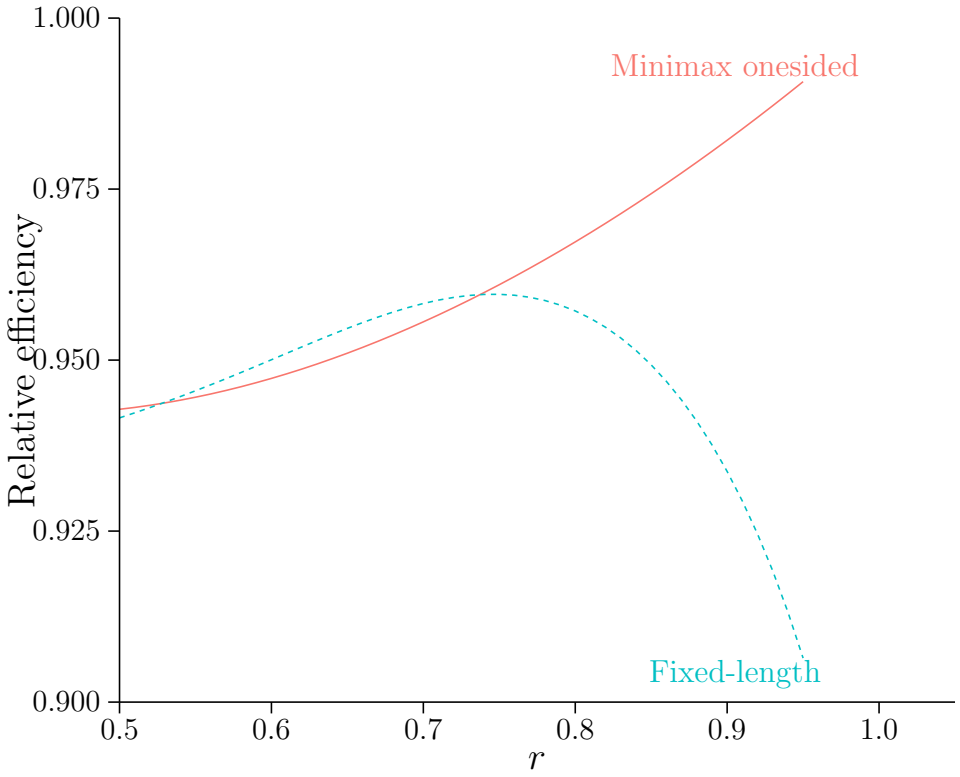


Figure 2: Asymptotic efficiency bounds for one-sided and fixed-length confidence intervals as function of the optimal rate of convergence r under centrosymmetry. Minimax one-sided refers to ratio of β-quantile of excess length of CIs that direct power at smooth functions relative to minimax one-sided CIs given in (19). Shortest fixed-length refers the ratio of expected length of CIs that direct power at a given smooth function relative to shortest fixed-length affine CIs given in (22).

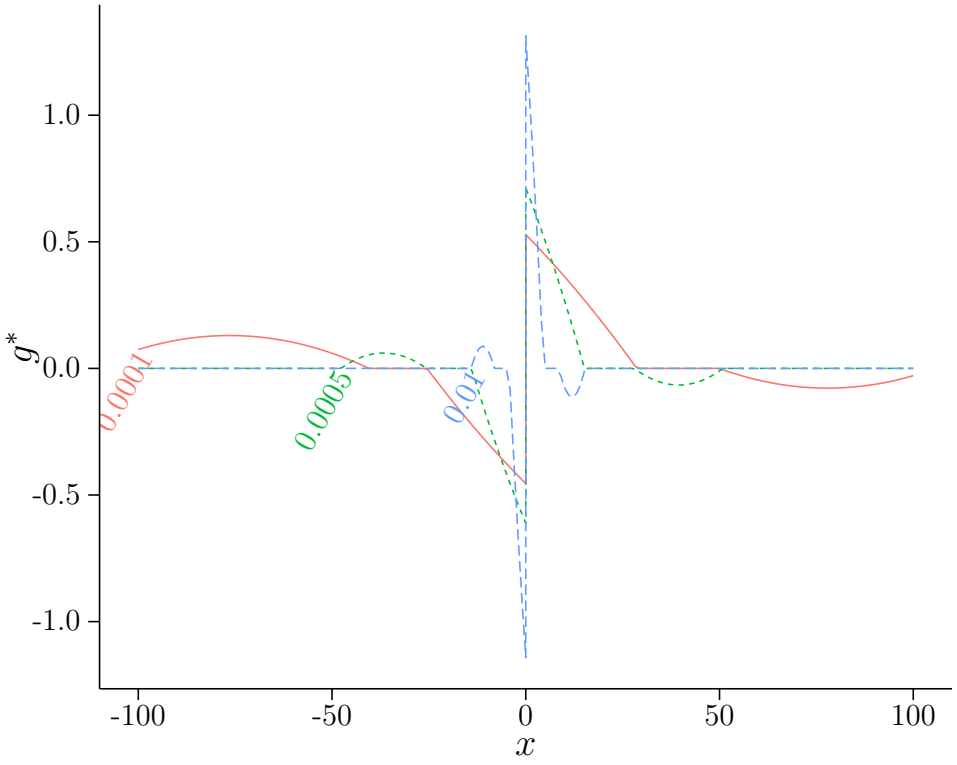


Figure 3: Least favorable function $g^*_{\delta,C}$ that solves the modulus problem for in the class $\mathcal{F}_{RDT,2}(C)$ in the Lee (2008) RD example for different values of the smoothness parameter C. $\delta = 2.49$, which is optimal for constructing minimax one-sided 95% CIs at 0.8 quantile.

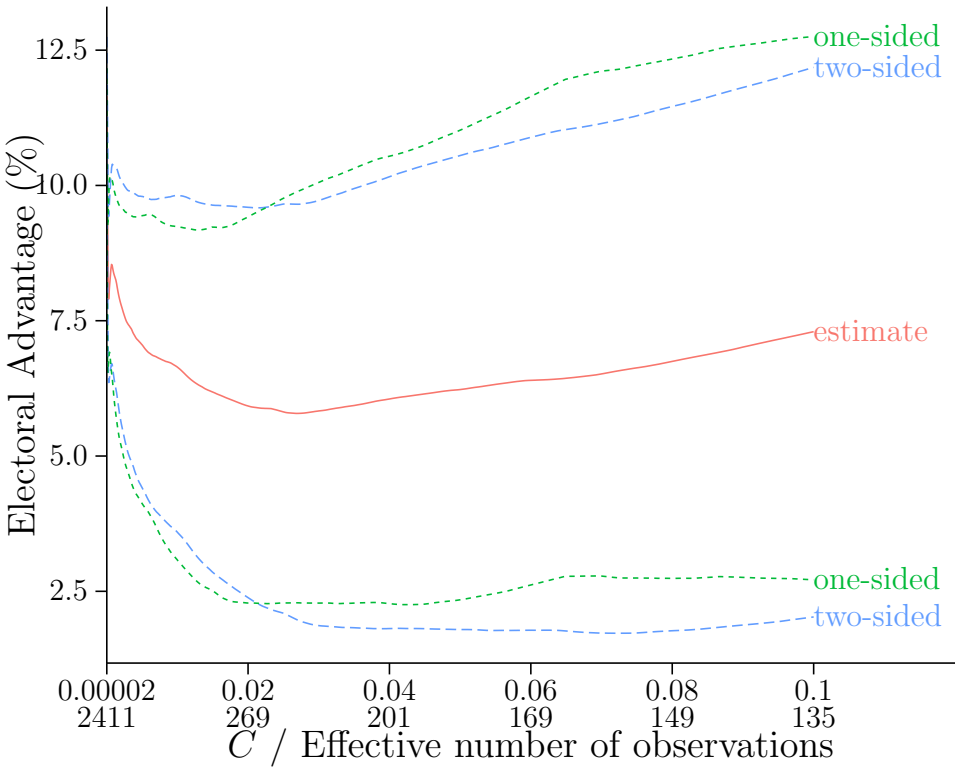


Figure 4: Lee (2008) RD example: minimax MSE estimator (estimator), lower and upper limits of minimax one-sided confidence intervals for 0.8 quantile (one-sided), and fixed-length CIs (two-sided) as function of smoothness C. Effective number of observations corresponds to n_e for the minimax MSE estimator as defined in Equation (27) in the text.

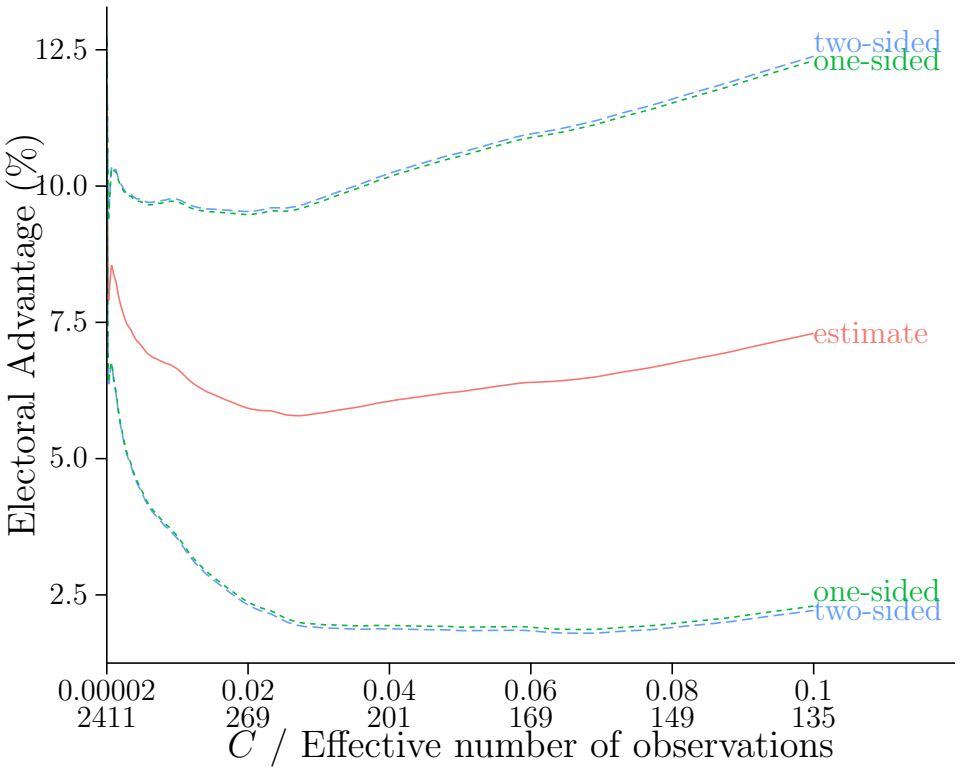


Figure 5: Lee (2008) RD example: minimax MSE estimator (estimator) with two-sided CI (two-sided) as well as lower and upper limits of one-sided CIs around it as function of smoothness C. Effective number of observations corresponds to n_e for the minimax MSE estimator as defined in Equation (27) in the text.

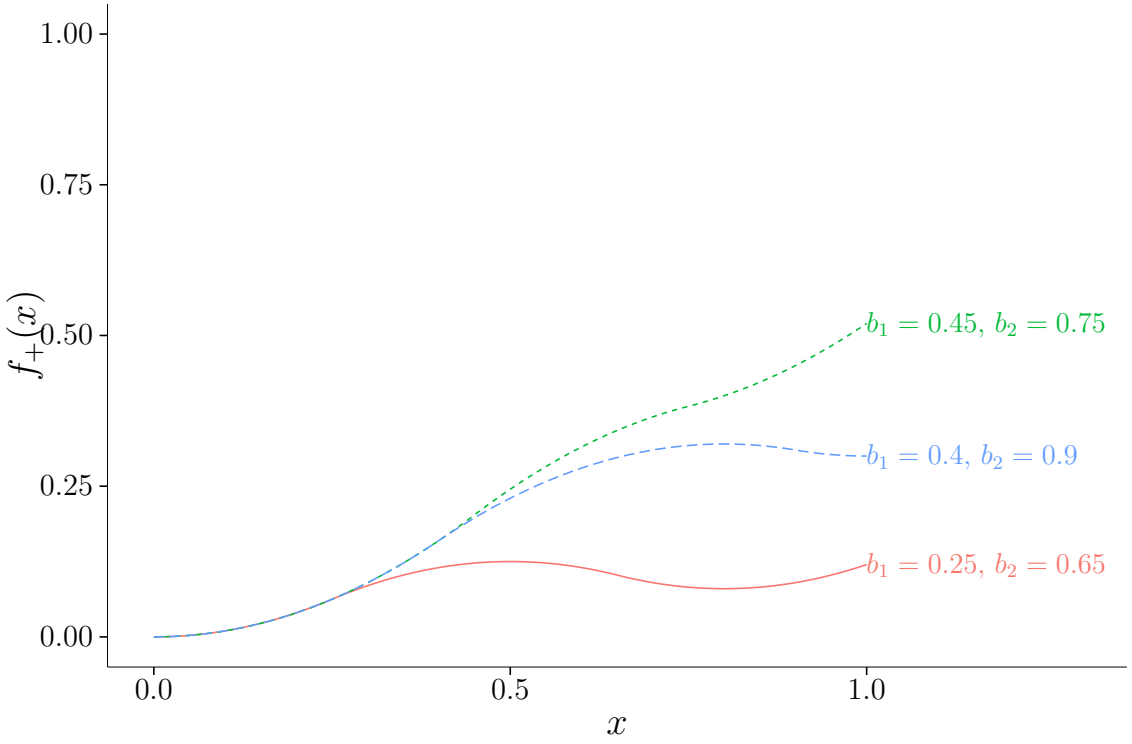


Figure 6: Regression function for Monte Carlo simulation, Designs 1–3, and $C = 1$. Knots $b_1 = 0.45, b_2 = 0.75$ correspond to Design 1, $b_1 = 0.4, b_2 = 0.9$ to Design 2, and $b_1 = 0.25, b_2 = 0.65$ to Design 3.