OPTIMAL INFERENCE IN A CLASS OF NONPARAMETRIC MODELS

Timothy Armstrong (Yale University)

Michal Kolesár (Princeton University)

September 2015

- Interested in inference on linear functional Lf in regression model

$$y_i = f(x_i) + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2(x_i)).$$

  x_i is fixed, $\sigma^2(x_i)$ is known.

- Important special cases:
  1. Inference at a point: $Lf = f(0)$
  2. Regression discontinuity: $Lf = f(0_+) - f(0_-)$
  3. ATE under unconfoundedness: $x_i = \{w_i, d_i\}$,
     $Lf = \frac{1}{n} \sum_i (f(w_i, 1) - f(w_i, 0))$
  4. Partially linear model

Convexity Assumption

$f \in \mathcal{F}$, a known convex set

Rules out e.g. sparsity, but not usual shape/smoothness restrictions:

Monotonicity $\mathcal{F} = \{f : f \text{ non-increasing}\}$

Lipschitz class $\mathcal{F}_{\text{Lip}}(C) = \{f : |f(x_1) - f(x_2)| \leq C|x_1 - x_2|\}$ (or Hölder class generalizations).

Taylor class $\mathcal{F}_{\text{T},2}(C) = \left\{f : |f(x) - f(0) - f'(x)x| \leq Cx^2\right\}$ (useful for RD / inference at point)

Sign restrictions in linear regression $\left\{f(x) = x'\beta : \beta_j \geq 0, j \in \mathcal{J}\right\}$

- Will take C as known if necessary, and ask later if this can be relaxed.

- Normality $\implies$ can derive *finite-sample* procedures that minimize the worst case loss over $\mathcal{G} \subseteq \mathcal{F}$
- without Normality, procedures will be valid and optimal asymptotically under regularity conditions, uniformly over $\mathcal{F}$

1. Setting $\mathcal{G} = \mathcal{F}$ yields minimax procedures.
    - Problem well-studied if loss is MSE, general solution in Donoho (1994), used to derive optimal kernels and rates of convergence (Stone, 1980; Fan, 1993; Cheng, Fan, and Marron, 1997)
    - Donoho (1994) derives fixed-length confidence intervals (CI) that are almost optimal
2. $\mathcal{G} \subset \mathcal{F}$ "smoother" functions: adaptive inference ("directing power")
    - For two-sided CIs, Cai and Low (2004) give bounds

Derive one-sided CIs, $[\hat{c}, \infty)$, that minimize maximum quantiles of excess length over $\mathcal{G}$, with $\hat{c} = \hat{L} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}) - z_{1-\alpha}\,\text{sd}(\hat{L})$, for optimal estimator $\hat{L}$

- For case $\mathcal{F} = \mathcal{G}$ (minimax CIs), $\hat{L}$ has same form as minimax MSE estimators / fixed-length CIs of Donoho (1994)
- We show that if $\mathcal{F}$ is symmetric, adaptation severely limited.
  - Adaptation requires non-convexity or shape restrictions: otherwise, cannot do better at smaller C while maintaining coverage for larger C
  - Conversely *any* inference method that claims to do better than minimax CIs when f is smooth *must be size distorted for some* $f \in \mathcal{F}(C)$
  - Related to Low (1997), who shows adapting to derivative smoothness classes limited for two-sided (random-length) CIs.

We derive two-sided CIs that minimize expected length over $\mathcal{G} = \{g\}$, solving the problem of "adaptation to a function" posed in Cai, Low, and Xia (2013)

- Can be used to bound scope for adaptivity

Asymptotically, optimal procedures often correspond to kernel estimators with fixed (optimal) kernel, and bw that depends on optimality criterion. We find that for RD and inference at a point:

- Optimal 95% fixed-length CIs use *larger bandwidth* than minimax MSE estimators.
  - Undersmoothing cannot be optimal
  - Recentering CIs by estimating bias cannot be optimal—it's essentially equivalent to using higher-order kernel and undersmoothing (Calonico, Cattaneo, and Titiunik, 2014).
- Difference is small: CI around minimax MSE estimator only 1% longer
  - In practice, can keep the same bandwidth as for estimation, and construct CI around it using worst-case bias correction

We apply the general results to:

1. RD with $\mathcal{F} = \{f_+ - f_- : f_\pm \in \mathcal{F}_{\mathrm{T},2}(C)\}$ as in Cheng, Fan, and Marron (1997)
    - Optimal bandwidths balance number of "effective observations" on each side of cutoff
    - Illustrate with empirical application from Lee (2008)
2. Linear regression with β possibly constrained (sign restrictions, sparsity, elliptical constraints)
3. Sample average treatment effect under unconfoundedness under Hölder class (separate paper)

- Stats literature on minimax estimation/inference/rates of convergence/adaptivity: Ibragimov and Khas'minskii (1985), Donoho and Liu (1991), Donoho and Low (1992), Donoho (1994), Low (1995), Low (1997), Cai and Low (2004), Cai, Low, and Xia (2013), Cheng, Fan, and Marron (1997), Fan (1993), Fan, Gasser, Gijbels, Brockmann, and Engel (1997), Lepski and Tsybakov (2000)
- "non-standard" CIs: Imbens and Manski (2004), Müller and Norets (2012), Calonico, Cattaneo, and Titiunik (2014), Calonico, Cattaneo, and Farrell (2015), Rothe (2015)
- Adaptive estimation/inference in econometrics: Sun (2005), Armstrong (2015), Chernozhukov, Chetverikov, and Kato (2014)

Finite-Sample results

Asymptotic results

Applications

Conclusion

- Consider the problem of inference on $f(0)$ when f is restricted to be in Lipschitz class $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C) = \{f : |f(x_1) - f(x_2)| \leq C|x_1 - x_2|\}$.
- Assume $\sigma(x) = \sigma$, known

- To measure performance of $(1-\alpha)\%$ one-sided CIs $[\hat{c}, \infty)$, we use maximum quantiles of excess length

$$\mathrm{EL}_\beta(\hat{c}, \mathcal{G}) = \sup_{g \in \mathcal{G}} q_{g,\beta}(Lg - \hat{c}),$$

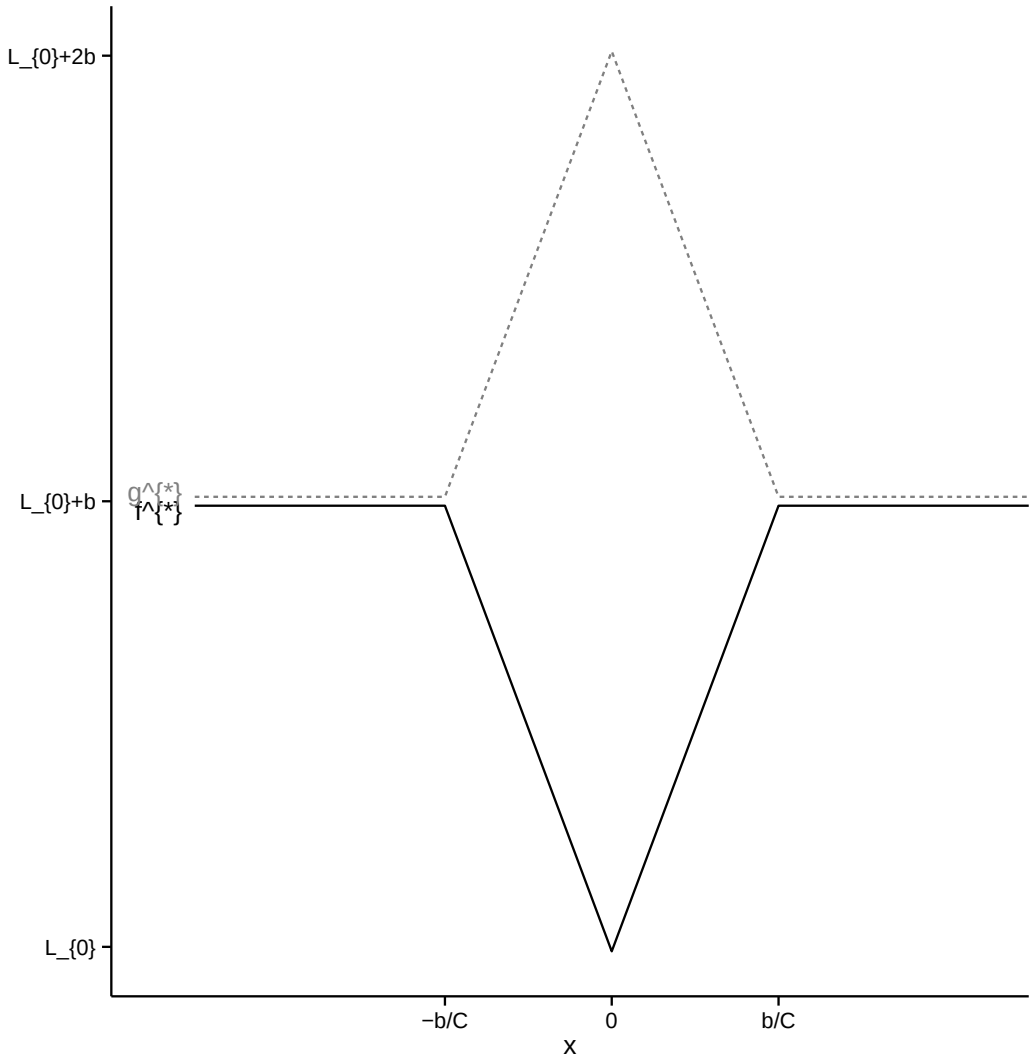
  where $q_{g,\beta}$ is βth quantile under g.
- For two-sided CIs, we focus on fixed-length CIs $\hat{L} \pm \chi$, where $\hat{L}$ is estimator, and χ is chosen to satisfy coverage:

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

- For estimation, we use maximum MSE, $R_{\mathrm{MSE}}(\hat{L}) = \sup_{f \in \mathcal{F}} E_f(\hat{L} - Lf)^2$

- In running example, $Lf = f(0)$, $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$, consider minimax test of $H_0 : Lf \leq L_0$ against $H_1 : Lf \geq L_0 + 2b$
  - Inverting minimax tests yields CI that minimizes $\text{EL}_\beta(\hat{c}, \mathcal{F})$, where β is minimax power of the test.
- First need to find least favorable null and alternative. Problem equivalent to $Y \sim \mathcal{N}(\mu, \sigma^2 I)$, $\mu = (f(x_1), \ldots, f(x_n)) \in M$ convex
- Both $M_0 = M \cap \{f : Lf \leq L_0\}$ and $M_1 = M \cap \{g : Lg \geq L_0 + 2b\}$ are convex—least favorable functions minimize distance between them (Ingster and Suslina, 2003)

$$(g^*, f^*) = \operatorname*{argmin}_{g \in M_1, f \in M_0} \sum_{i=1}^{n} (g(x_i) - f(x_i))^2.$$



$$g^*(x) =$$

$$L_0 + b + (b - C|x|)_+$$

$$f^*(x) =$$

$$L_0 + b - (b - C|x|)_+$$

- $g^*(x) = L_0 + b + (b - C|x|)_+$, $f^*(x) = L_0 + b - (b - C|x|)_+$
- Minimax test then given by LR test of $\mu_0 = (f^*(x_1), \ldots, f^*(x_n))$ against $\mu_1 = (g^*(x_1), \ldots, g^*(x_n))$: reject for large values of $Y'(\mu_1 - \mu_0)$
- Test can be written as rejecting whenever

$$\tilde{L}(h) - L_0 - b\left(1 - \frac{\sum_{i=1}^n k_T(x_i/h)^2}{\sum_{i=1}^n k_T(x_i/h)}\right) \geq \frac{(\sum_{i=1}^n k_T(x_i/h)^2)^{1/2}}{\sum_{i=1}^n k_T(x_i/h)} \sigma z_{1-\alpha}.$$

  where $k_T(u) = (1 - |u|)_+$, $h = b/C$ and

$$\tilde{L}(h) = \frac{\sum_{i=1}^n (g^*(x_i) - f^*(x_i)) Y_i}{\sum_{i=1}^n (g^*(x_i) - f^*(x_i))} = \frac{\sum_{i=1}^n k_T(x_i/h) Y_i}{\sum_{i=1}^n k_T(x_i/h)}$$

- Key feature: non-random bias correction based on worst-case bias, doesn't disappear asymptotically

- In general, we observe $Y = Kf + \sigma\epsilon$, ϵ is standard Normal and K linear operator, with $\langle Kg, Kf \rangle = \sum_i (Kg)(x_i)(Kf)(x_i)$,
- Heteroscedasticity handled by setting
  $Kf = (f(x_1)/\sigma(x_1), \ldots, f(x_n)/\sigma(x_n))$, $Y = (Y_1/\sigma(x_1), \ldots, Y_n/\sigma(x_n))$.
- Define modulus of continuity (Donoho and Liu, 1991):

$$\omega(\delta; \mathcal{F}) = \sup \{ L(g - f) : \|K(g - f)\| \leq \delta, \;\; g, f \in \mathcal{F} \}$$

  Denote solutions by g^*_δ, f^*_δ, and let $f^*_{M,\delta} = (g^*_\delta + f^*_\delta)/2$
- Problem of finding LF functions equivalent to finding $\omega^{-1}(\cdot; \mathcal{F})$, so for running example, $g^* = g^*_{\omega^{-1}(2b)}$, $f^* = f^*_{\omega^{-1}(2b)}$

Define

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

These estimators minimize maximum bias given variance bound (and vice versa) (Low, 1995). Their maximum and minimum bias over $\mathcal{F}$ satisfies

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

In running example: $\tilde{L}(h) = \hat{L}_{\omega^{-1}(2hC),\mathcal{F}_{\text{Lip}}(C)}$

When $\mathcal{F}$ has additional structure, $\hat{L}_\delta$ simplifies:

- If $\mathcal{F}$ is *translation invariant* (for some $\iota \in \mathcal{F}$ with $L\iota = 1$, $c\iota \in \mathcal{F}$ for all $c \in \mathbb{R}$), then $\delta/\omega'(\delta;\mathcal{F}) = \left\langle K(g_\delta^* - f_\delta^*), \iota \right\rangle$, and estimator has Nadaraya-Watson form:

$$\hat{L}_{\delta,\mathcal{F}} = Lf_{M,\delta}^* + \frac{\langle K(g_\delta^* - f_\delta^*), Y - Kf_{M,\delta}^* \rangle}{\langle K(g_\delta^* - f_\delta^*), K\iota \rangle}.$$

- If $\mathcal{F}$ is *centrosymmetric* ($f \in \mathcal{F} \implies -f \in \mathcal{F}$), then $f_\delta^* = -g_\delta^*$, and

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

Theorem 1 (One-sided minimax CI)

Let

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

Then $[\hat{c}_{\alpha,\delta,\mathcal{F}}, \infty)$ is a $1-\alpha$ CI for Lf, with coverage minimized at f^*_δ. For $\beta = \Phi(\delta/\sigma - z_{1-\alpha})$, it minimizes $\text{EL}_\beta(\hat{c}, \mathcal{F})$ among all one sided $1-\alpha$ CIs. All quantiles of excess length are maximized at g^*_δ. The minimax excess length at quantile β is $\text{EL}_\beta(\hat{c}_{\alpha,\delta,\mathcal{F}}; \mathcal{F}) = \omega(\delta;\mathcal{F})$.

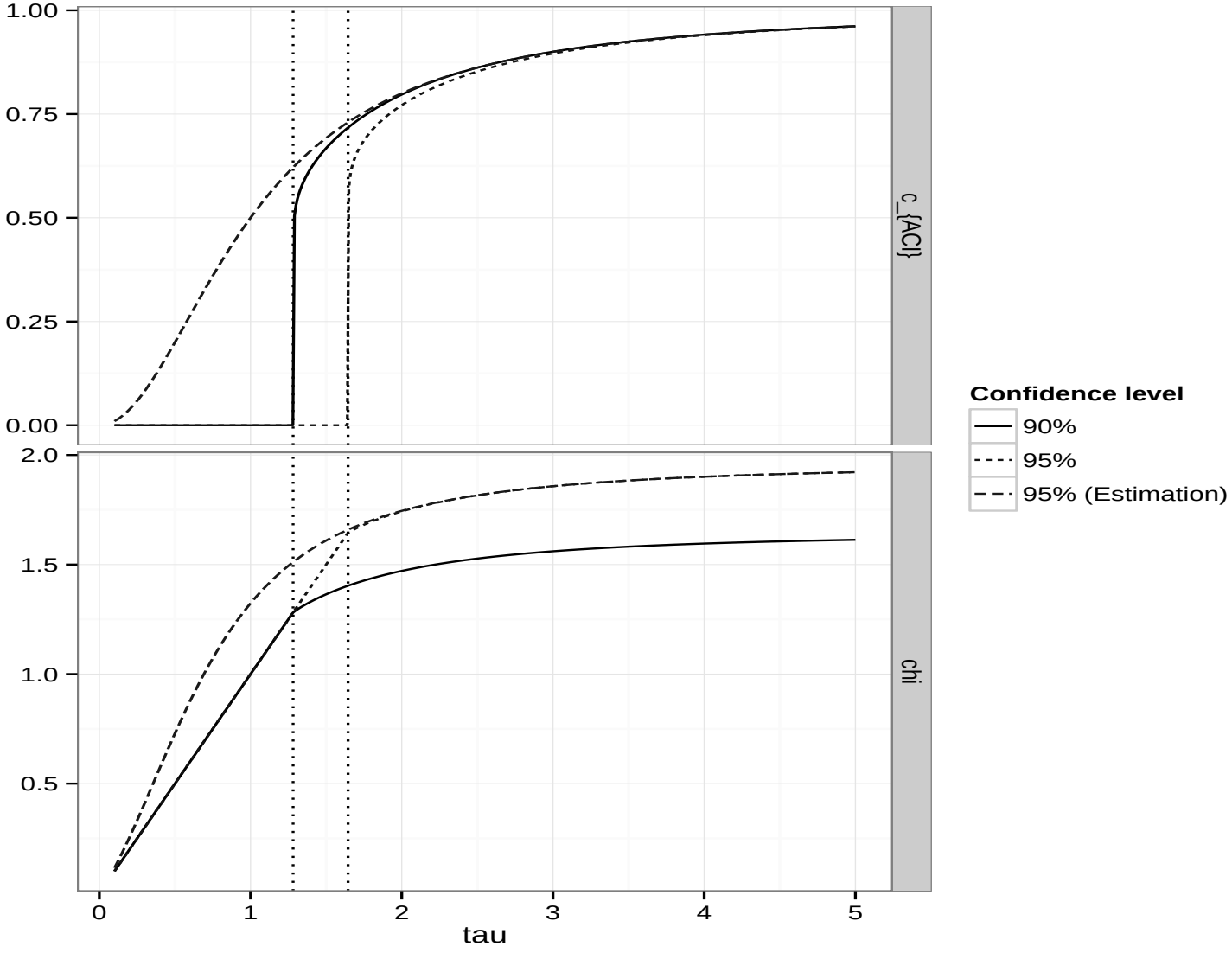
- β is minimax power of underlying tests (under translation invariance)
- Bias-correction based on worst-case bias under $\mathcal{F}$, non-random
- In running example, using bw h minimizes β quantile of excess length at $\beta = \Phi\left(\frac{\omega^{-1}(2hC)}{\sigma} - z_{1-\alpha}\right)$

- For estimation and two-sided CI, exact optimality results hard
- Donoho (1994) shows that procedures based on $\hat{L}_{\delta,\mathcal{F}}$ minimax optimal if we restrict attention to affine estimators
- Results use the fact that problem is just as hard if we know that f is in one-dimensional subfamily $\left\{\lambda f_\delta^* + (1-\lambda) g_\delta^* : 0 \leq \lambda \leq 1\right\}$

  To state these results, consider $Z \sim \mathcal{N}(\theta, 1)$, $\theta \in [-\tau, \tau]$

- Minimax linear estimator is $c_\rho(\tau)Z$, $c_\rho(\tau) = \tau^2/(1+\tau^2)$ with minimax risk $\rho(\tau) = \tau^2/(1+\tau^2)$
- Shortest fixed-length CI is $c_\chi(\tau)Z \pm \chi_\alpha(c_\chi(\tau)Z)$, solution characterized in Drees (1999), similar in spirit to Imbens and Manski (2004)

OPTIMAL SHRINKAGE IN BOUNDED NORMAL MEANS



Theorem (Donoho (1994))

minimax MSE affine estimator is $\hat{L}_{\delta,\mathcal{F}}$ where δ solves

$$\max_{\delta>0} \frac{\omega(\delta;\mathcal{F})}{\delta} \sqrt{\rho\left(\frac{\delta}{2\sigma}\right)}\sigma.$$

and the optimal δ satisfies $c_\rho(\delta/(2\sigma)) = \delta\omega'(\delta;\mathcal{F})/\omega(\delta;\mathcal{F})$.

The shortest fixed-length affine CI is $\hat{L}_{\delta,\mathcal{F}} \pm \frac{\omega(\delta;\mathcal{F})}{\delta}\chi_\alpha\left(\frac{\delta}{2\sigma}\right)\sigma$ where δ solves

$$\max_{\delta>0} \frac{\omega(\delta;\mathcal{F})}{\delta}\chi_\alpha\left(\frac{\delta}{2\sigma}\right)\sigma.$$

and the optimal δ satisfies $c_\chi(\delta/(2\sigma)) = \delta\omega'(\delta;\mathcal{F})/\omega(\delta;\mathcal{F})$.

- For example, to find minimax MSE optimal bandwidth in running example, solve

$$\frac{\delta^2}{4\sigma^2+\delta^2} = c_\rho(\delta/(2\sigma)) = \frac{\delta\omega'(\delta;\mathcal{F})}{\omega(\delta;\mathcal{F})} = \frac{\delta^2}{2\omega(\delta;\mathcal{F})\sum_i g^*_\delta(x_i)}$$

  which yields

$$\sigma^2 = C^2h^2\left(\sum_i k_T(x_i/h) - \sum_i k_T(x_i/h)^2\right).$$

  Asymptotically

$$h_{opt,MSE} = \left(\frac{3\sigma^2}{C^2 n f_X(0)}\right)^{1/3} + o_p(1)$$

- Can also use these results to derive optimal rates of convergence (eg Fan (1993); Cheng, Fan, and Marron (1997))—$n^{-1/3}$ here

- onesided CIs focus on good performance under least favorable $f \in \mathcal{F}$, which may be too pessimistic
- Alternative: optimize excess length over smaller class $\mathcal{G}$ of smoother functions

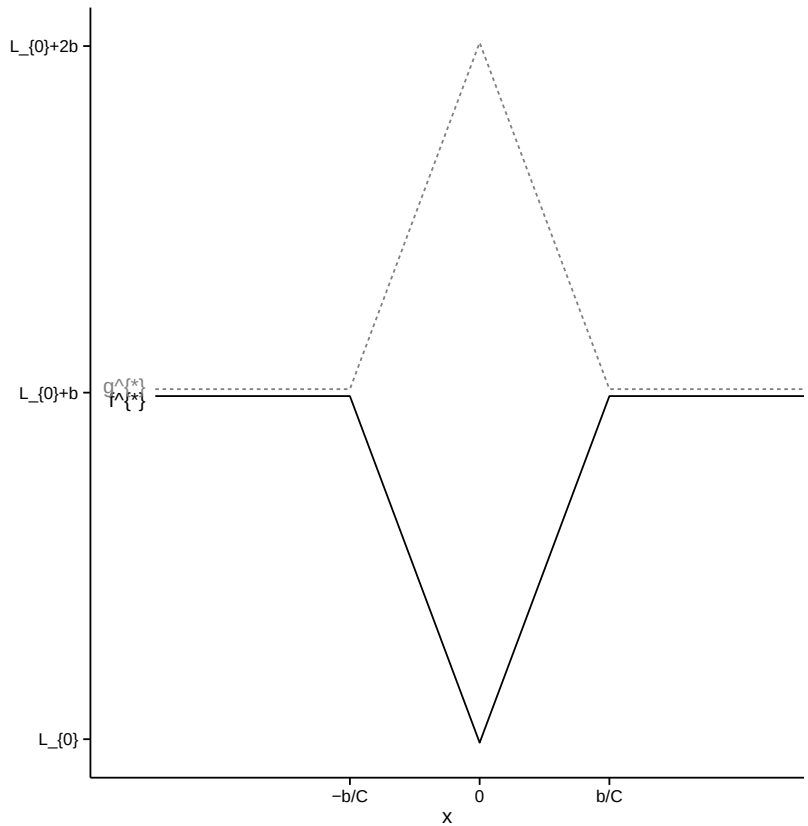
$$\inf_{\hat{c}} \sup_{g \in \mathcal{G}} q_{q,\beta}(Lg - \hat{c}),$$

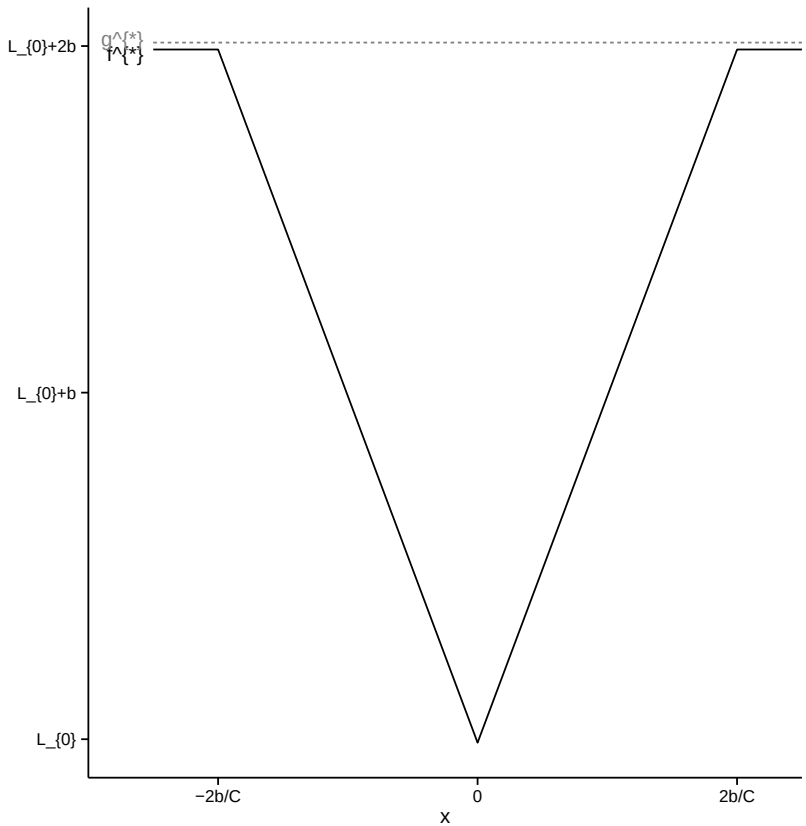
  among $\hat{c}$ that satisfy $\sup_{f \in \mathcal{F}} P(Lf \geq \hat{c}) \geq 1 - \alpha$.
- Amounts to "directing power" at smooth alternatives, while maintaining size over all of $\mathcal{F}$

- Associated testing problem in running example: $H_0 : Lf \leq L_0$ against $H_1 : \{Lf \geq L_0 + 2b\} \cap \{f \in \mathcal{G}\}$
- Inverting these minimax tests will yield CI that minimizes β quantile of excess length over $\mathcal{G}$, where β is minimax power of the test.
- As long as $\mathcal{G}$ is convex, this is still equivalent to testing convex null against convex alternative $\implies$ LF functions minimize distance between sets:

$$(f^*, g^*) = \underset{f \in \mathcal{F}, g \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} (g(x_i) - f(x_i))^2, \quad Lg \geq L_0 + 2b,\ Lf \leq L_0$$

- To make this concrete, consider $\mathcal{G} = \{g(x) : g(x) = c, c \in \mathbb{R}\}$ (i.e. $g(x) = c\iota$), and suppose $Lf \geq L_0 + b$ under alternative
- Solution: $f^* = L_0 + b - (b - X|x|)_+$ (as before), $g^*(x) = L_0 + b$





- But $g^* - f^*$ same as before, so estimator as before

$$\tilde{L}(h) = \frac{\sum_{i=1}^n (g^*(x_i) - f^*(x_i))Y_i}{\sum_{i=1}^n (g^*(x_i) - f^*(x_i))} = \frac{\sum_{i=1}^n k_T(x_i/h)Y_i}{\sum_{i=1}^n k_T(x_i/h)}$$

- Worst case-bias under the null and variance same as before $\implies$ Same CI as before

Summary

One sided CI that minimizes maximum excess length over $\mathcal{F}$ for $\beta = \Phi(\delta/\sigma - z_{1-\alpha})$ subject to $1-\alpha$ coverage also minimizes $\mathrm{EL}_\beta(\hat{c}; \mathrm{span}(\iota))$ for $\beta = \Phi(\delta/(2\sigma) - z_{1-\alpha})$

- Define order modulus of continuity Cai and Low (2004):

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

  so that $\omega(\delta; \mathcal{F}) = \omega(\delta; \mathcal{F}, \mathcal{F})$, and define

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

  so that $L_{\delta,\mathcal{F},\mathcal{F}} = L_{\delta,\mathcal{F}}$

- bias formulas generalize:

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

- In running example, $\tilde{L}(h) = \hat{L}_{\omega^{-1}(hC;\mathcal{F},\mathcal{G}),\mathcal{F},\mathcal{G}}$

Theorem 2 (One-sided adaptive CIs)

Let $\mathcal{F}$ and $\mathcal{G} \subseteq \mathcal{F}$ be convex, and suppose that f_δ^* and g_δ^* achieve the ordered modulus at δ. 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 $\text{EL}_\beta(\hat{c},\mathcal{G})$ among all one-sided $1-\alpha$ CIs, where Φ denotes the standard normal cdf. 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 $\text{EL}_\beta(\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}},\mathcal{G}) = \omega(\delta;\mathcal{F},\mathcal{G})$.

- Suppose $\mathcal{F}$ is centrosymmetric and

$$f^*_{\delta,\mathcal{F},\mathcal{G}} - g^*_{\delta,\mathcal{F},\mathcal{G}} \in \mathcal{F}\,. \tag{1}$$

  Holds for $\mathcal{G}$ "smooth enough", e.g. $\mathcal{G} = \operatorname{span}(\iota)$ under translation invariance as in running example

- Then 0 and $f^*_{\delta,\mathcal{F},\mathcal{G}} - g^*_{\delta,\mathcal{F},\mathcal{G}}$ also solve the modulus, and since $\omega(\delta;\mathcal{F}) = \sup\{-2Lf : \|Kf\| \le \delta/2, f \in \mathcal{F}\}$ under centrosymmetry,

$$\omega(\delta;\mathcal{F},\mathcal{G}) = \omega(\delta;\mathcal{F},\{0\}) = \sup_{f\in\mathcal{F}}\{-Lf : \|Kf\| \le \delta\} = \frac{1}{2}\omega(2\delta;\mathcal{F}),$$

- Implies $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}} = \hat{c}_{\alpha,\delta,\mathcal{F},\{0\}} = \hat{c}_{\alpha,2\delta,\mathcal{F}}$.

Theorem 3 (Non-adaptivity of one-sided CIs under centrosymmetry)

Let $\mathcal{F}$ be centrosymmetric. Then the one-sided CI that is minimax for the βth quantile also optimizes $\mathrm{EL}_{\tilde{\beta}}(\hat{c}; \mathcal{G})$ for any $\mathcal{G}$ such that the solution to the ordered modulus problem exists and satisfies (1), where

$$\tilde{\beta} = \Phi((z_\beta - z_{1-\alpha})/2).$$

In particular, the minimax CI optimizes $\mathrm{EL}_{\tilde{\beta}}(\hat{c}; \{0\})$.

- CI that is minimax for median excess length among 95% CIs also optimizes $\Phi(-1.645/2) \approx 0.205$ quantile under the zero function.

- CI $[\hat{c}_{\alpha,\sigma(z_\beta+z_{1-\alpha}),\mathcal{F}})$ that is minimax for βth quantile of excess length is unbiased at 0, and satisfies

$$q_{0,\beta}(L0 - \hat{c}_{\alpha,\sigma(z_\beta+z_{1-\alpha})}) = \frac{1}{2}(\omega'(\delta;\mathcal{F})\delta + \omega(\delta;\mathcal{F})).$$

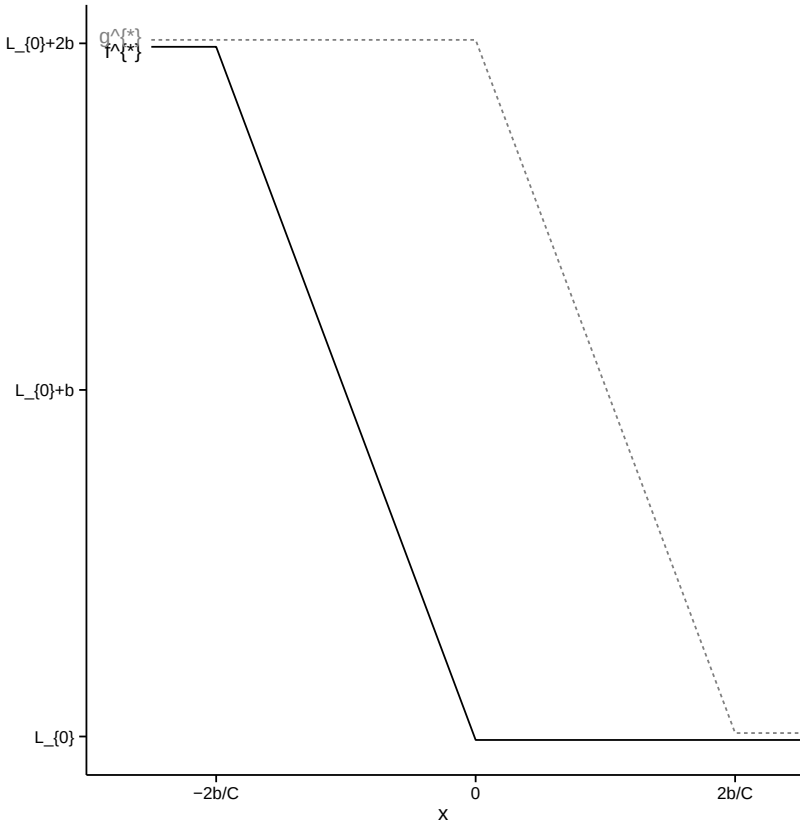
  Hence,

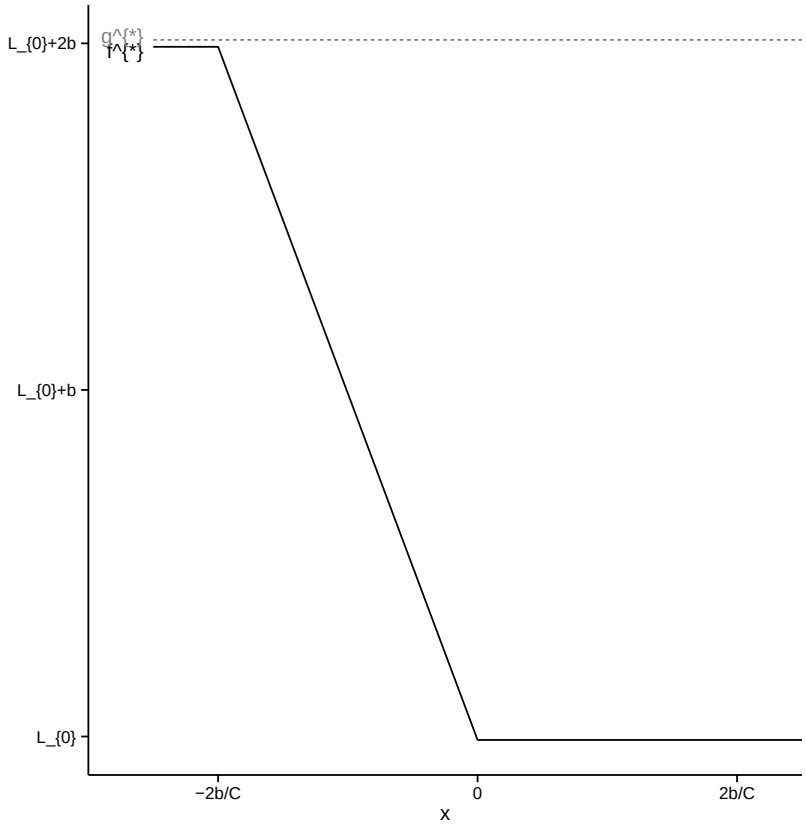
$$\frac{\omega(\delta;\mathcal{F},\mathcal{G})}{q_{0,\beta}(L0 - \hat{c}_{\alpha,\sigma(z_\beta+z_{1-\alpha})})} = \frac{\omega(\delta;\mathcal{F},\mathcal{G})}{\frac{1}{2}(\omega'(\delta)\delta + \omega(\delta))} = \frac{\omega(2\delta)}{\omega'(\delta)\delta + \omega(\delta)}.$$

- Typically, $\omega(\delta;\mathcal{F}) = A\delta^r(1 + o(1))$ as $n \to \infty$ for some where r is the optimal rate of convergence of the MSE. Then for $1/2 \leq r \leq 1$, minimax CI has asymptotic efficiency at least 94.3% when indeed $f = 0$.
- Adapting to $\mathcal{G}$ that includes 0 at least as hard as adapting to zero

- Need shape restriction or non-convexity for adaptation
- Similar to impossibility results in Low (1997) and Cai and Low (2004) for two-sided CIs, and in contrast to positive results for MSE
  - Minimax rate of shrinkage describes the actual rate for all functions in the class
  - Possible to construct estimators that do better when f is smoother, but impossible to tell how well you did
- For valid inference in cases where $\mathcal{F}$ is convex and centrosymmetric, one has to think hard about appropriate C
  - Not possible to try to estimate it from the data and to better than if we assume worst possible case

- Suppose, in running example, that we know f is non-increasing
- Least favorable functions without and with directing power:





- Without directing power, optimal estimator again given by triangular kernel, but now includes bias correction (to ensure $\overline{\text{bias}} = -\underline{\text{bias}}$)

$$\tilde{L}(h) = \frac{\sum_i k_i Y_i/\sigma_i^2}{\sum_i k_i/\sigma_i^2} + b\frac{\sum_i \text{sign}(x_i) k_i(1-k_i)/\sigma_i^2}{\sum_i k_i/\sigma_i^2},$$

  where $k_i = k_\beta(x_i/h)$, and optimal bw bigger than without monotonicity. About 20% reduction in quantiles of excess length
- With directing power, optimal estimator averages *all positive observations*, and averages negative observations using triangular kernel. *Excess length shrinks at parametric rate.*
- When Lipschitz assumption dropped and only monotonicity maintained, optimal estimator averages all positive observations, and excess length *still* shrinks at *parametric rate*

- Fixed-length confidence intervals cannot be adaptive
- Cai and Low (2004) construct random-length confidence intervals that are within a constant factor of lower bound on expected length
- Cai, Low, and Xia (2013) construct random-length confidence intervals under shape constraints that have near minimum expected length for each individual function (again within constant)

- Natural best-case scenario for two-sided CIs: optimize expected length at a single function $\mathcal{G} = \{g\}$
- By Pratt (1961), inverting UMP tests against $\mathcal{G}$ achieve exactly this
- Again amounts to testing convex null against convex alternative, LF function under null solves

$$\overline{f}^*_\theta = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2, \quad Lf \leq \theta$$

Theorem 4 (Adaptation to a function)

CI with minimum expected measure $E_g \lambda(C)$ st $1 - \alpha$ coverage on $\mathcal{F}$ inverts family of tests ϕ_θ, where ϕ_θ rejects for large values of $\langle K(g - \overline{f}^*_\theta), Y \rangle$ with critical value given by $1 - \alpha$ quantile under $\overline{f}^*_\theta$.

- What is efficiency loss of CIs around suboptimal affine estimators?
- Affine estimators are Normal, with variance that doesn't depend on f, and bias that does
- For each performance criterion, only worst-case bias and variance matter: if we can calculate them, then can also calculate maximum MSE, and form of one- and two-sided CIs
- Let $\tilde{\chi}_\alpha(B)$ solve $P(|Z+B| \leq \chi) = \Phi(\chi - B) - \Phi(\chi + B) = 1 - \alpha$. Then for estimator $\hat{L}$ with variance V and maximum bias B, is the shortest CI is

$$\hat{L} \pm V^{1/2}\tilde{\chi}_\alpha(B/V^{1/2})$$

Theorem 5 (Suboptimal estimators)

Let $\hat{L} = a + \langle w, Y \rangle$ be an affine estimator. Then
$[\hat{L} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}) - \|w\| z_{1-\alpha}\sigma, \infty)$ is valid CI and
$\hat{L} \pm \sigma\|w\| \tilde{\chi}_{\alpha}(\overline{\text{bias}}_{\mathcal{F}}(\hat{L})/\sigma\|w\|)$ is the shortest fixed-length $1-\alpha$ CI centered at $\hat{L}$.

- Not deep result, but very useful: allows to compute exact efficiency loss from using suboptimal estimators, or size-distortion of CIs with (pointwise) asymptotic justification
- Asymptotic version of this theorem can be used to calculate asymptotic efficiency loss from using suboptimal kernel, and/or suboptimal bandwidth

- Consider some other kernel k in running example, $\hat{L} = \frac{\sum_i k(x_i/h) Y_i}{\sum_i k(x_i/h)}$
- Variance: $\frac{\sigma^2 \sum_i k(x_i/h)^2}{(\sum_i k(x_i/h))^2}$
- Maximum bias, since $f \in \mathcal{F}_{\text{Lip}}(C)$.

$$\left| \frac{\sum_i k(x_i/h)(f(x_i) - f(0))}{\sum_i k(x_i/h)} \right| \leq C \frac{\sum_i |k(x_i/h)||x_i|}{\sum_i k(x_i/h)}.$$

Bound attained at $f(x) = C|x_i|$ if $k \geq 0$, otherwise gives an upper bound.

Finite-Sample results

Asymptotic results

Applications

Conclusion

- In many cases (depending on L and smoothness of $\mathcal{F}$, but including inference at a point and RD), nonparametric regression problem equivalent to White noise model $Y(dt) = f(t) + \sigma\epsilon(t)$
  - See Brown and Low (1996) and Donoho and Low (1992)
  - In running example, this holds with $\sigma^2 = \sigma(0)^2/nf_X(0)$
- Suppose $\mathcal{F} = \{f : J(f) \leq C\}$ for some J (as in running example), and that for the white noise model, following functionals are *homogeneous*

$$J(af(\cdot/h)) = ah^{-s_J} J(f)$$

$$\langle Ka_1 f(\cdot/h), Ka_2 g(\cdot/h)\rangle = a_1 a_2 h^{-2s_K} \langle Kf, Kg\rangle$$

$$L(af(\cdot/h)) = ah^{-s_L} Lf$$

- In running example, we have $s_L = 0$, $s_J = 2$, $s_K - 1/2$

- (single-class) modulus problem then renormalizes: if $g^*_{C,\delta}, f^*_{C,\delta}$ minimize $\min|L(f_1 - f_0)|$ st $\|K(f_1 - f_0)\| \leq \delta$, $J(f_1) \leq C$, $J(f_0) \leq C$, then

$$g^*_{C,\delta} = a g^*_{1,1}(\cdot/h) \qquad f^*_{C,\delta} = a f^*_{1,1}(\cdot/h)$$
$$\omega_C(\delta) = C^{1-r}\delta^r \omega_1(1)$$

  where $a = \delta^{-s_J/(s_K - s_J)} C^{s_K/(s_K - s_J)}$, $h = (C/\delta)^{1/(s_K - s_J)}$ and

$$r = \frac{s_L - s_J}{s_K - s_J}.$$

- root of minimax MSE, and (excess) length of CIs will shrink at rate $n^{-r/2}$

- Class of optimal estimators can be written as

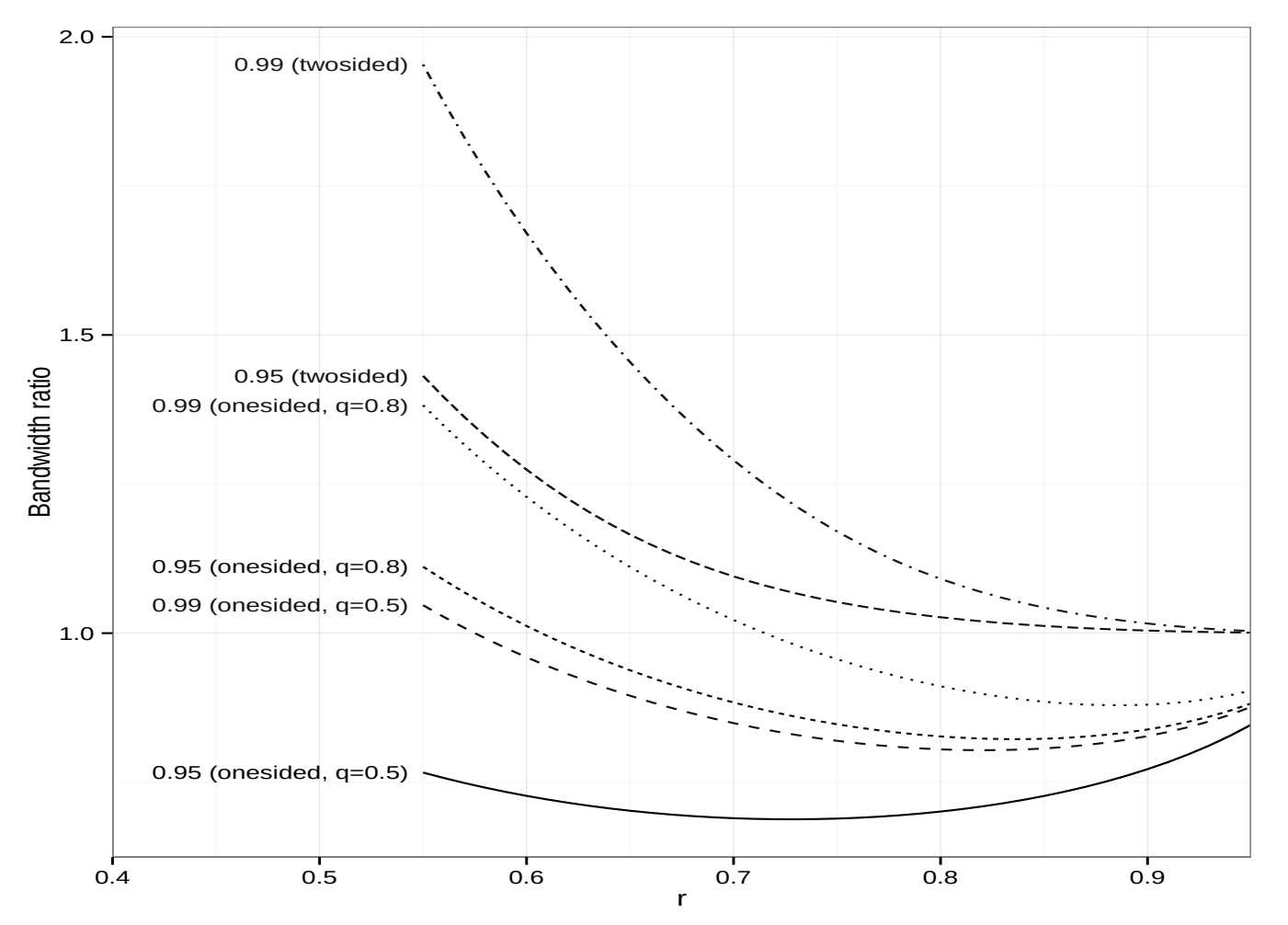
$$\hat{L}_\delta = \tilde{L}(h) = h^{2s_K - s_L} \langle Kk(\cdot/h), Y \rangle + Ch^{s_J - s_L} \left(Lf_{M,1,1} - \langle Kk, Kf_{M,1,1} \rangle \right),$$

  with $h = (C/\delta)^{1/(s_K - s_J)}$ and kernel $k(u) = r\omega_1(1)(g^*_{1,1} - f^*_{1,1})$.
- Recall that optimal δ given by $c_\ell(\delta/(2\sigma)) = \delta\omega'(\delta)/\omega(\delta)$. Plugging in definition of h yields optimal bandwidth

$$h = (2\sigma c_\ell^{-1}(r)/C)^{\frac{1}{s_J - s_K}},$$

  where, for one-sided CIs, $c_\beta^{-1}(r) = (z_\beta - z_{1-\alpha})/2$

RATIOS OF OPTIMAL BANDWIDTHS, $s_k = -1/2$, $s_l = 0$



Ratios of of optimal bandwidths for CIs to optimal MSE bandwidths

- Optimal bandwidth ratios depend only on dilation exponents s_L, s_K and s_J:

$$\frac{h_\ell}{h_{\ell'}} = \left(\frac{c_\ell^{-1}(r)}{c_{\ell'}^{-1}(r)} \right)^{\frac{1}{s_J - s_K}}$$

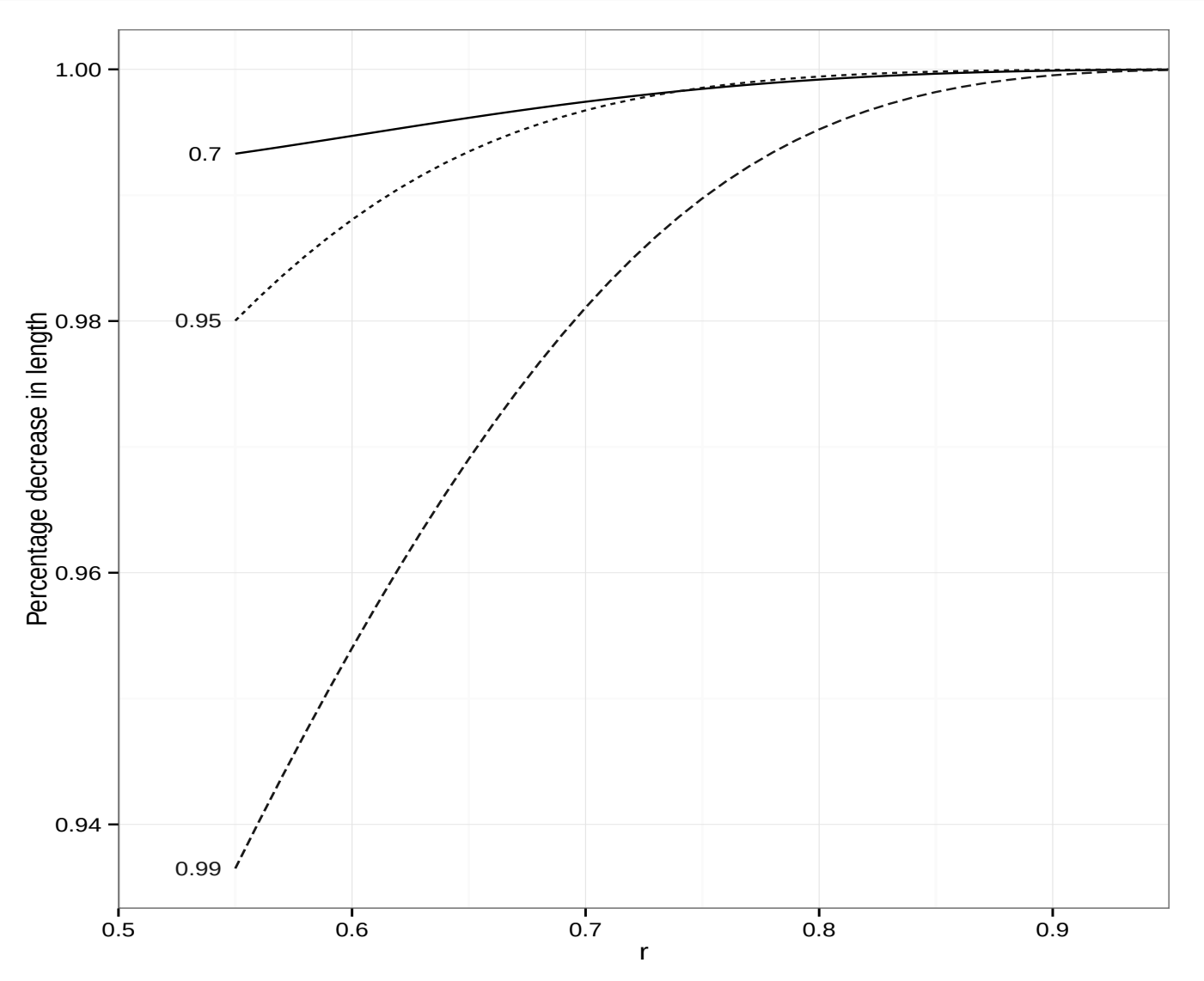
- Bandwidths of same order in all cases: *no undersmoothing*
- For one-sided CIs, bandwidth gets larger with quantile that we are minimizing
- For 95+% two-sided CIs, if $s_L = 0$ and $s_K = -1/2$, optimal fixed-length CI uses a *larger* bandwidth than optimal MSE bandwidth

- For any bandwidth h, worst-case bias is $\frac{C}{2}\frac{1-r}{r}h^{s_J-s_L}(\int k^2)^{1/2}$
- Can use this worst case bias to construct CIs around $\tilde{L}(h)$
- How much bigger are two-sided CIs around minimax MSE bandwidth? Ratio of CI lengths given by

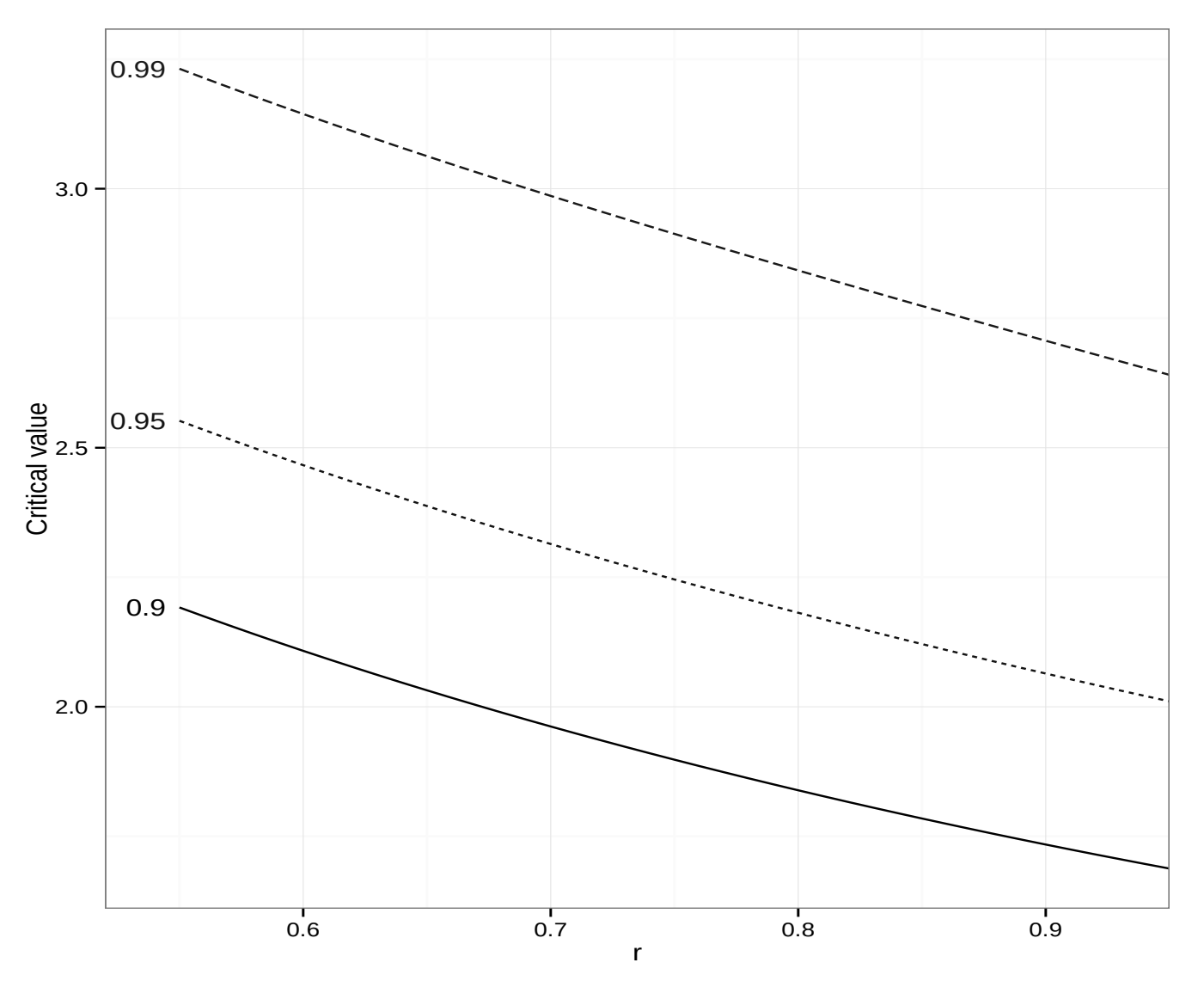
$$\left(\frac{c^{-1}_{\chi,\alpha}(r)}{c^{-1}_{\rho}(r)}\right)^{r-1} \cdot \frac{\tilde{\chi}_\alpha(c^{-1}_{\chi,\alpha}(r)(1-1/r))}{\tilde{\chi}_\alpha(c^{-1}_{\rho}(r)(1-1/r))},$$

  where $\tilde{\chi}_\alpha(B)$ solves $P(|\mathcal{N}(0,1)+B| \leq \chi) = \Phi(\chi - B) - \Phi(\chi + B) = 1-\alpha$
- Need to use $\tilde{\chi}_\alpha\left(\sqrt{\frac{1-r}{r}}\right)$ instead of $|z_{\alpha/2}|$ as a critical value to ensure coverage for CI around minimax MSE bandwidth

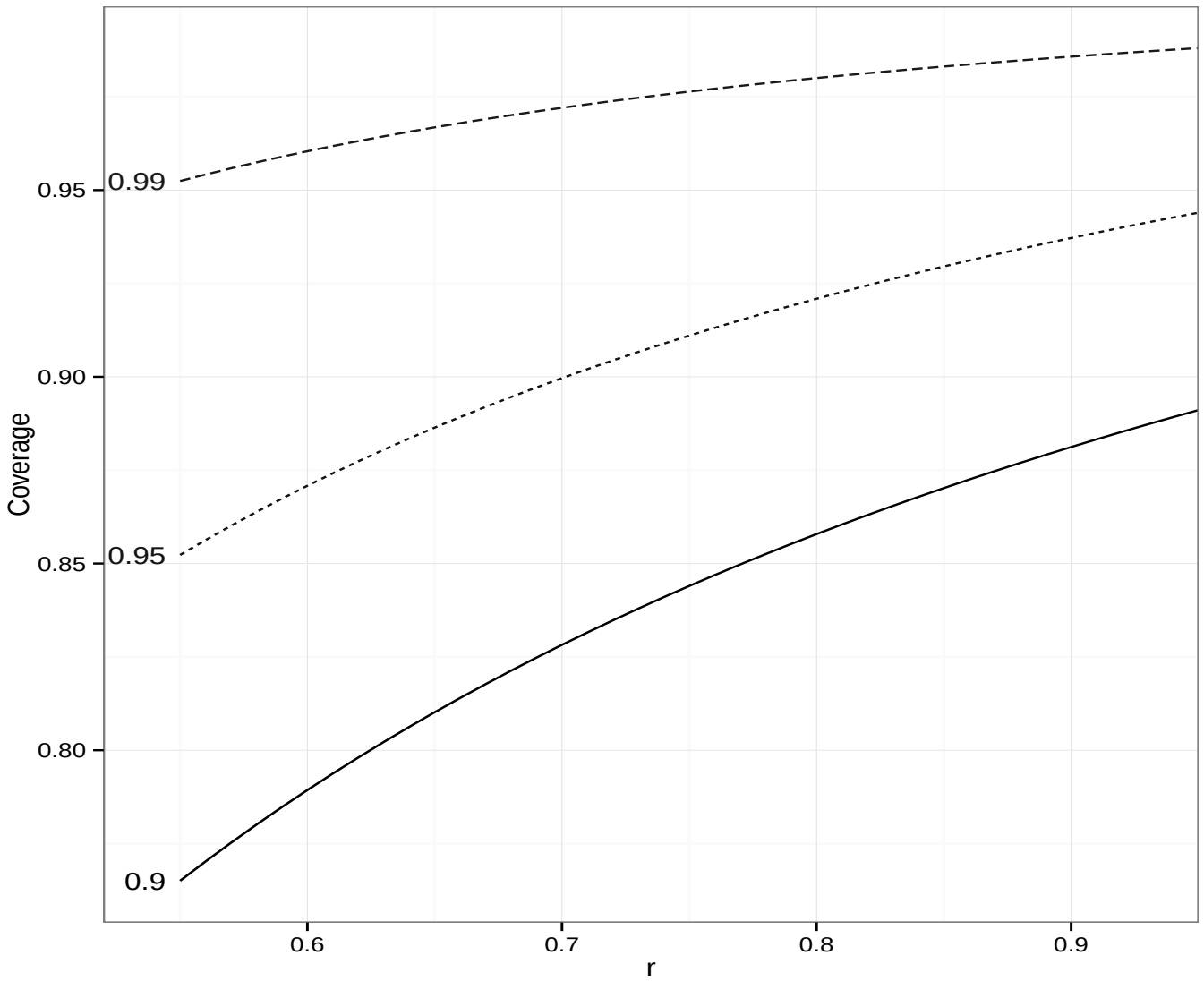
LENGTH OF OPTIMAL CIS RELATIVE TO CIS AROUND MSE BW



"CRITICAL VALUES" FOR CI AROUND MSE BANDWIDTH



UNDERCOVERAGE WITH USUAL CRITICAL VALUES



- To construct two-sided CIs, can keep the same bandwidth as for estimation, price is $< 2\%$ for 95% CIs
- Need to use a slightly higher critical value to ensure proper coverage

- Results so far assumed using optimal kernel
- Under renormalization, maximum bias and variance renormalize in similar way for suboptimal kernels
- For any kernel k, let $\tilde{h}_k$ be bandwidth that equates the maximum bias and root variance, and let $w(k) = se(\tilde{L}_k(\tilde{h}_k)) = \sup_f \text{bias}_f(\tilde{L}_k(\tilde{h}_k))$
- Suppose criterion scales linearly with maximum bias and root variance

Theorem 6 (Efficiency loss of suboptimal kernels)

1. Relative efficiency of k and $\tilde{k}$ (where the optimal bandwidth is used in both cases) does not depend on the performance criterion, and is given by $w(k)/w(\tilde{k})$
2. Results for ratios of optimal bandwidths remain unchanged for suboptimal kernels
3. Efficiency loss from using bandwidth optimal for a different criterion rather than bandwidth optimal for criterion of interest remains unchanged for suboptimal kernels

- bounds for minimax MSE efficiency of different kernels of Cheng, Fan, and Marron (1997) 1. are tight; and 2. hold for other efficiency criteria
- Using minimax MSE bandwidth for two-sided CIs a good idea no matter what kernel one uses

Finite-Sample results

Asymptotic results

Applications

Conclusion

- Interested in $Lf = \lim_{x \downarrow 0} f(x) - \lim_{x \uparrow 0} f(x)$.
- Let $f_+(x) = f(x)I(x > 0)$ and $f_-(x) = -f(x)I(x < 0)$ so that $f = f_+ - f_-$.
- We consider class

$$\mathcal{F}_{RDT,2}(C) = \{f_+ - f_- : f_+ \in \mathcal{F}_{\mathrm{T},2}(C; \mathbb{R}_+), f_- \in \mathcal{F}_{\mathrm{T},2}(C; \mathbb{R}_-),\}$$

  where $\mathcal{F}_2(C; \mathcal{X})$, is the class from Sacks and Ylvisaker (1978),

$$\mathcal{F}_{\mathrm{T},2}(C; \mathcal{X}) = \left\{ f : |f(x) - f(0) - f'(0)x| \leq Cx^2 \text{ all } x \in \mathcal{X} \right\}.$$

- $\mathcal{F}_{\mathrm{T},2}$ also used in Cheng, Fan, and Marron (1997) for estimation at a point that justifies much of empirical RD practice

Least favorable functions are symmetric, $g^*_\delta(x) = -f^*_\delta(x)$ and have form

$$g^*_\delta(x) = [(b - b_- + d_+x - Cx^2)_+ - (b - b_- + d_+x + Cx^2)_-]1(x > 0)$$
$$[(b_- + d_-x - Cx^2)_+ - (b_- + d_-x + Cx^2)_-]1(x < 0)$$

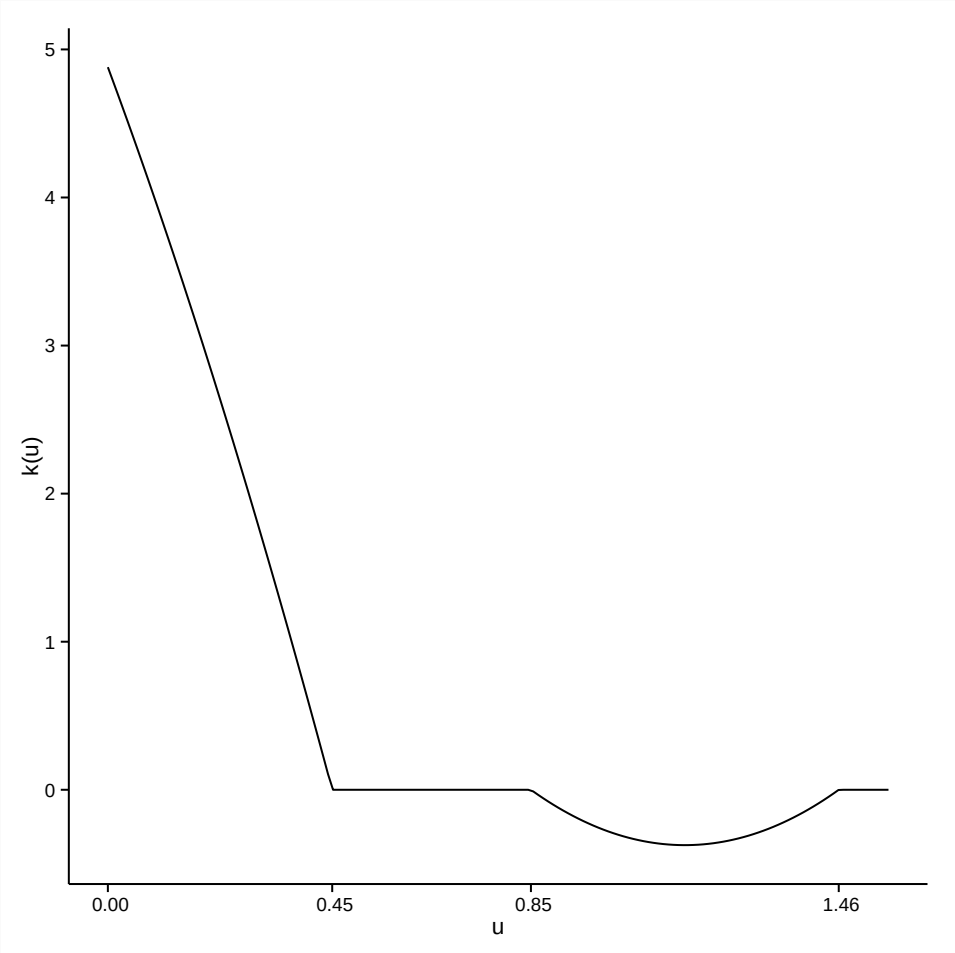
with b_-, d_+, d_- chosen to solve

$$0 = \sum_{i=1}^n \frac{g^n_{-,b,C}(x_i)x_i}{\sigma^2(x_i)}, \qquad 0 = \sum_{i=1}^n \frac{g^n_{+,b,C}(x_i)x_i}{\sigma^2(x_i)},$$

and

$$\sum_{i=1}^n \frac{g_{+,b,C}(x_i)}{\sigma^2(x_i)} = \sum_{i=1}^n \frac{g_{-,b,C}(x_i)}{\sigma^2(x_i)}$$

OPTIMAL KERNEL

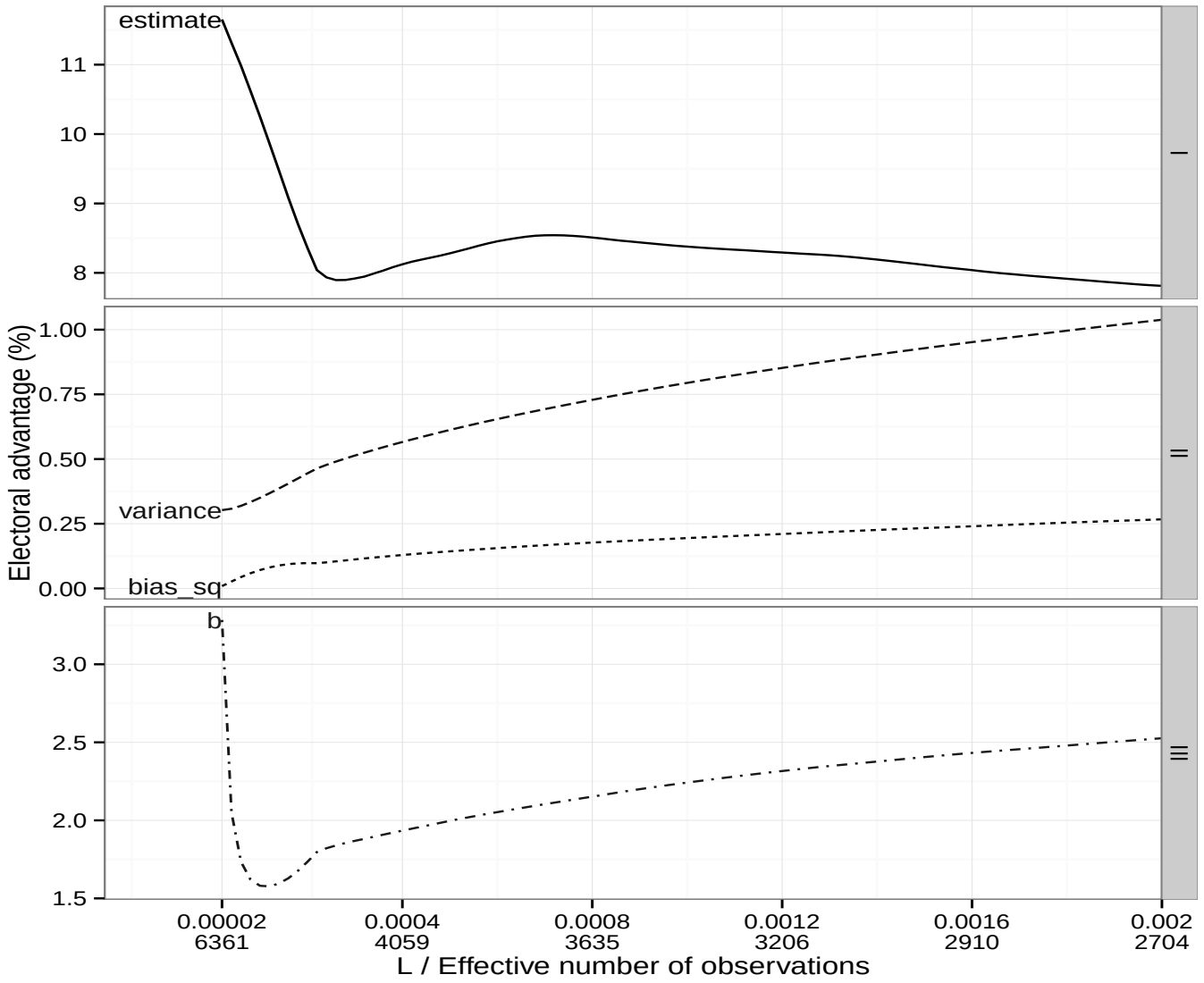


- Asymptotically, g^*_δ corresponds to difference between two kernel estimators, with bandwidths chosen to equate number of effective observations
- Optimal kernel same as for inference at a point, derived in Cheng, Fan, and Marron (1997) using upper bound on minimax MSE

application to Lee (2008)

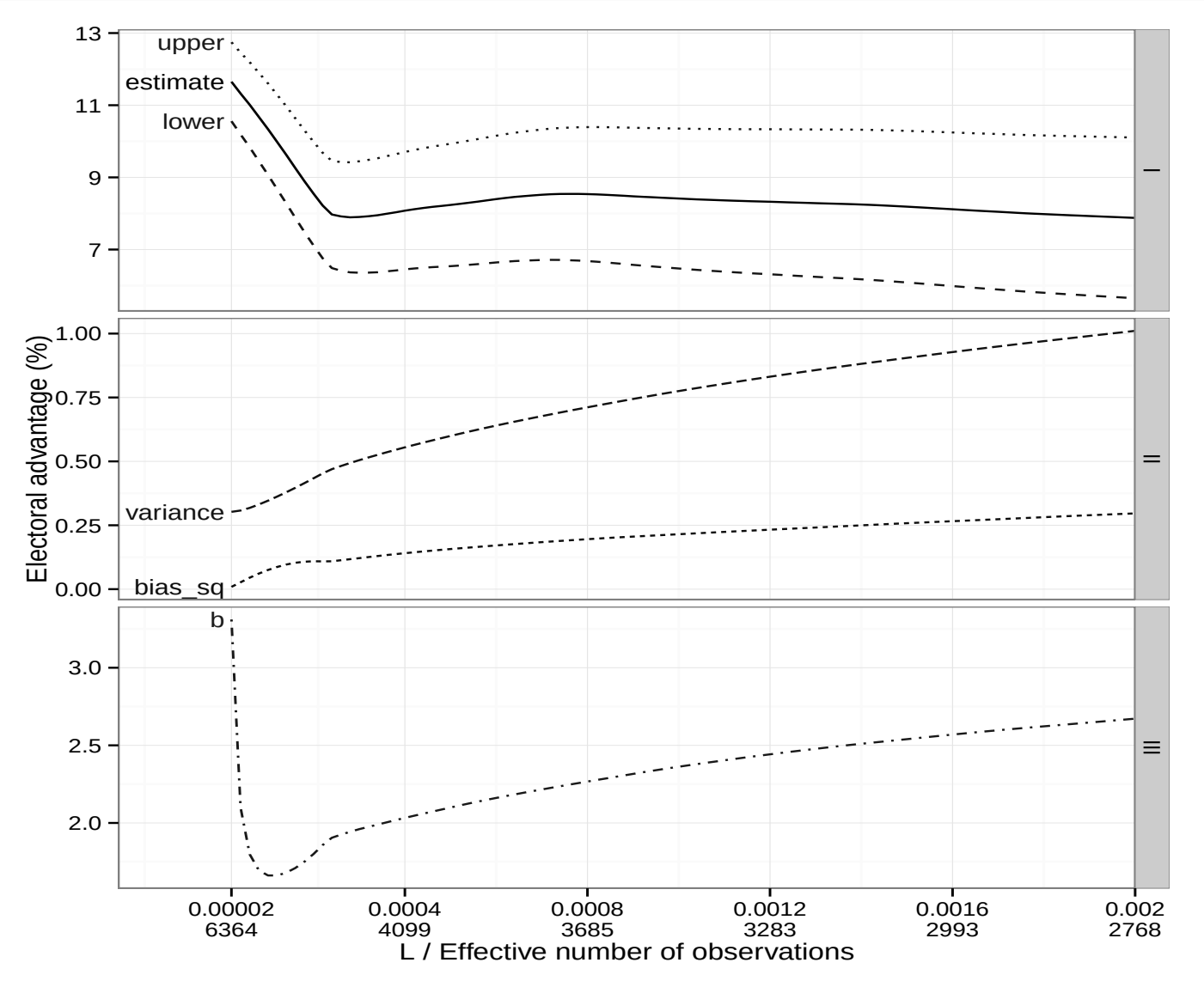
- RD design:
  - X_i = margin of victory in previous election for Democratic party (negative for Republican victory)
  - Y_i = Democratic vote share in given election
  - $D_i = I(X_i \geq 0)$ = indicator for Democratic incumbency
  - $n = 6558$ observations of elections between 1946 and 1998
- For simplicity, assume homoscedastic errors, use estimates $\hat{\sigma}^2_-(0) = 155.3$ and $\hat{\sigma}_+(0) = 210.3$ derived using Imbens and Kalyanaraman (2012) bandwidth
- LF functions very close to scaled versions of optimal bandwidth
- Unless C very small, results in line with Lee (2008) and Imbens and Kalyanaraman (2012)

MINIMAX MSE ESTIMATOR AS FUNCTION OF c



Note $L = C$

OPTIMAL FIXED-LENGTH CIS



Finite-Sample results

Asymptotic results

Applications

Conclusion

1. give exact results for 1. minimax optimal and 2. adaptive one-sided CIs.
    - CIs use non-random bias correction based on worst-case bias
    - Adaptivity without shape restrictions severely limited, like in two-sided case.
    - Impossible to avoid thinking hard about appropriate C
2. give exact solution to problem of "adaptation to a function"
3. use these finite-sample results to characterize optimal tuning parameters for different performance criteria
    - building CIs around minimax MSE bandwidth nearly optimal
    - undersmoothing cannot be optimal