FINITE-SAMPLE OPTIMAL ESTIMATION AND INFERENCE ON AVERAGE TREATMENT EFFECTS UNDER UNCONFOUNDEDNESS

Timothy Armstrong (Yale University)

Michal Kolesár (Princeton University)

September 2017

- Treatment effects often estimated under "unconfoundedness" or "selection-on-observables" assumption

  $\Longrightarrow$ Systematic differences in outcomes between treated and control units with same values of covariates x_i due only to treatment

- When x_i continuously distributed, estimation requires *regularization*:
  - matching with imperfect matches
  - kernel/series/sieve estimation of conditional means

  Regularization causes finite-sample bias

- Conventional approach to choosing estimators and constructing confidence intervals (CIs) based on first-order asymptotics
  - Ignores bias; many estimators asymptotically efficient
- This paper: take finite-sample approach to address these problems

Conventional approach

1. Pick estimator $\hat{\beta}$ that achieves semiparametric efficiency bound
2. Construct std error $\widehat{\text{se}}(\hat{\beta})$ by estimating efficient asymptotic variance
3. Construct nominal 95% CI as $\hat{\beta} \pm 1.96\widehat{\text{se}}(\hat{\beta})$

Problems:

- Many estimators achieve semiparametric efficiency bound
  - Estimators that do not achieve the bound (e.g. nearest neighbor matching) often more intuitively appealing than estimators that do (e.g. matching with 10th order kernel)
- Ignores bias due to smoothing and lack of perfect overlap $\implies$ potential undercoverage of CIs
- Theory requires a lot of smoothness (e.g. bound on 10th derivative)

Our approach

- ~~Theory requires a lot of smoothness (e.g. bound on 10th derivative)~~
- Start out with smoothness assumption on conditional mean of outcome given covariates x_i and treatment d_i: okay to use bound on lower order derivative
- Consider performance of estimators and CIs conditional on (x_i, d_i)

1. Pick estimator $\hat{\beta}$ that ~~achieves semiparametric efficiency bound~~ solves finite-sample worst-case bias-variance tradeoff
    - ~~Many estimators achieve semiparametric efficiency bound~~
    - Unique optimal estimator, linear in outcomes
    - Under Lipschitz (first-derivative) constraint on conditional mean, with large enough bound, matching estimator with single match optimal

2. Construct std error $\widehat{\text{se}}(\hat{\beta})$ by estimating ~~efficient asymptotic variance~~ its conditional variance $\sum_i k_i^2 \operatorname{var}(y_i \mid x_i, x_i)$, where $\hat{\beta} = \sum_i k_i y_i$ is optimal estimator
3. Construct nominal 95% CI ~~as $\hat{\beta} \pm 1.96\widehat{\text{se}}(\hat{\beta})$~~ by using larger critical value that explicitly takes into account possible bias (substantial in empirical application)
    - ~~Ignores bias due to smoothing and lack of perfect overlap $\implies$ potential undercoverage of CIs~~
    - Finite-sample coverage under normal errors and known variance
    - Feasible CIs with estimated variance valid and asymptotically efficient when error distribution unknown

- Derive minimal smoothness conditions for achieving semiparametric efficiency bound when conditioning on realized treatments and covariates
  - Need a bound on derivative of order $\dim(X_i)/2$
  - Matches unconditional bound of Robins et al. (2009) when no smoothness is imposed on propensity score
- Our CIs asymptotically valid and efficient with unknown error distribution even when semiparametric efficiency bound cannot be achieved
  - In this case, critical value > 1.96 even asymptotically

Related literature

- Apply framework of Armstrong and Kolesár (2017, hereafter AK17) and Donoho (1994)
    - Similar motivation to RD application in AK17 and Kolesár and Rothe (2017): finite sample approach means we don't have to worry about discrete vs continuous regressors, etc.
- Semiparametrics with alternative asymptotics (papers by Cattaneo, Farrell, Jansson, Newey; Abadie and Imbens)
- Low regularity semiparametrics (Robins et al., 2009; Khan and Tamer, 2010)
- Classical selection on observables literature (Rosenbaum and Rubin, 1983; Hahn, 1998; Dehejia and Wahba, 1999; Heckman et al., 1997, 1998a,b; Hirano et al., 2003, ...)

Setup

Optimal estimators/CIs

Numerical illustration

Application

Conclusion

Setup

- For $i = 1, \ldots, n$, observe covariates x_i, treatment indicator d_i, and outcome $y_i = (1 - d_i)y_i(0) + d_i y_i(1)$, where $y_i(0), y_i(1)$ are potential outcomes
- Condition on $\{x_i, d_i\}_{i=1}^n$ throughout: all expectations and probability statements are conditional [↗ Discussion of conditioning on (d_i, x_i)]
- Leads to fixed design regression model

$$y_i = f(x_i, d_i) + u_i, \quad E(u_i) = 0,$$

with u_i independent over i

- To obtain finite sample results, further assume $u_i \sim N(0, \sigma^2(x_i, d_i))$ with $\sigma^2(x_i, d_i)$ known

- Under unconfoundedness (selection on observables assumption), conditional average treatment effect (CATE) given by

$$\text{CATE}(f) = \frac{1}{n}\sum_{i=1}^{n} E[(Y_i(1) - Y_i(0)) \mid x_i] = \frac{1}{n}\sum_{i=1}^{n} [f(x_i, 1) - f(x_i, 0)]$$

[↗ Details of CATE definition]

Key Assumption

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

- For many results, focus on Lipschitz class, for some norm $\|\cdot\|$,

$$\mathcal{F} = \mathcal{F}_{\text{Lip}}(C) = \{f : |f(x, d) - f(\tilde{x}, d)| \leq C\|x - \tilde{x}\|,\ d \in \{0, 1\}\}$$

Criteria for relative efficiency

- Interval C is $100 \cdot (1 - \alpha)\%$ CI for CATE if

$$\inf_{f \in \mathcal{F}} P_f(\text{CATE}(f) \in C) \geq 1 - \alpha, \tag{1}$$

  where P_f denotes probability under f
- Among CIs that satisfy (1), minimize worst-case (expectation or quantile of) length over $\mathcal{F}$ (minimax)
  - By results in Armstrong and Kolesár (2016), not much to be gained by making directing power at smoother functions [↗ Details]
- For estimation, focus on maximum (worst-case over $\mathcal{F}$) mean squared error (MSE)

Relation to asymptotic efficiency bounds

- We use same notion of efficiency as asymptotic results in the literature
- For example, saying

  > *matching estimators achieve semiparametric efficiency bound when regression function has at least one derivative and there is only one continuous covariate*

  can be formalized as stating

  > *ratio of the minimax MSE of a matching estimator to the optimal minimax MSE converges to one when* $\mathcal{F} = \mathcal{F}_{Lip}(C)$ *and there is only one continuous covariate*

- Our approach is "uniform in the underlying distribution" version of nonparametric smoothness

Setup

Optimal estimators/CIs

Numerical illustration

Application

Conclusion

- For notational ease, let $Lf = \text{CATE}(f)$
- Optimal CIs are based on *linear estimators* (AK17) $\hat{L}_k = \sum_{i=1}^{n} k(x_i, d_i) y_i$
- By linearity, $\hat{L}_k$ is normal with variance

$$\text{sd}(\hat{L}_k)^2 = \sum_{i=1}^{n} k(x_i, d_i)^2 \sigma^2(x_i, d_i)$$

and bias bounded in absolute value by

$$\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} k(x_i, d_i) f(x_i, d_i) - Lf \right|$$

- One-sided CI based on $\hat{L}_k$ given by $[\hat{c}, \infty)$, where

$$\hat{c} = \hat{L}_k - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) - \text{sd}(\hat{L}_k) z_{1-\alpha}$$

- For two-sided CI, could add and subtract $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) + \text{sd}(\hat{L}_k) z_{1-\alpha/2}$, but this is conservative. Instead, note that

$$\frac{\hat{L}_k - Lf}{\text{sd}(\hat{L}_k)} \sim N(t, 1) \text{ where } |t| \leq \frac{\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)}{\text{sd}(\hat{L}_k)}.$$

  Letting $\text{cv}_\alpha(t)$ denote $1 - \alpha$ quantile of $|N(t, 1)|$, can therefore use *fixed length CI* (FLCI)

$$\hat{L}_k \pm \text{cv}_\alpha(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)/\text{sd}(\hat{L}_k)) \, \text{sd}(\hat{L}_k)$$

- For two-sided FLCI, choose k to minimize length

$$2\text{cv}_\alpha(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)/\text{sd}(\hat{L}_k))\text{sd}(\hat{L}_k)$$

- When using $\hat{L}_k$ as a point estimate, one can minimax the MSE:

$$\sup_{f\in\mathcal{F}} E_f\left[(\hat{L}_k - Lf)^2\right] = \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)^2 + \text{sd}(\hat{L}_k)^2$$

- For one-sided CI, we use minimax quantiles of excess length (see paper)

- All of these criteria increasing in $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$ and $\text{sd}(\hat{L}_k)$, so it suffices to
  1. minimize variance subject to bound on bias
  2. vary this bound on bias to find optimal bias-variance tradeoff for given criterion (FLCI, MSE, etc.)
- Same idea as usual nonparametric bias-variance tradeoff, but in finite samples.
  - Unlike, say, estimating conditional mean at a point or RD parameter (Armstrong and Kolesár, 2017), no closed form for this problem in general, even asymptotically.
- Using results from Donoho (1994), AK17, we can trace out this bias-variance frontier using convex optimization.
- We give results for $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$. See paper for general convex $\mathcal{F}$.

- Bias-variance frontier can be traced out by solving

$$\max_{f \in \mathcal{F}_{Lip}(C)} 2\frac{1}{n}\sum_{i=1}^{n}[f(x_i,1) - f(x_i,0)] \text{ s.t. } \sqrt{\sum_{i=1}^{n}\frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)}} \leq \frac{\delta}{2}$$

  Let f^*_δ denote solution
- Constraint that $f \in \mathcal{F}_{Lip}(C)$ is equivalent to requiring that for $d \in \{0,1\}$ and $i,j \in \{1,\ldots,n\}$ (Beliakov, 2006)

$$|f(x_i,d) - f(x_j,d)| \leq C\|x_i - x_j\|$$

- Developed algorithm similar to LARS/LASSO algorithm that traces out solution path as function of δ: piecewise linear solution

- Let

$$\hat{L}_\delta = \hat{L}_{k^*_\delta(\cdot)} = \sum_{i=1}^n k^*_\delta(x_i, d_i) y_i, \qquad k^*_\delta(x_i, d_i) = \frac{\frac{f^*_\delta(x_i, d_i)}{\sigma^2(x_i, d_i)}}{\sum_{j=1}^n \frac{d_j f^*_\delta(x_j, d_j)}{\sigma^2(x_j, d_j)}}.$$

  $\{\hat{L}_\delta\}_{\delta > 0}$ traces out the optimal bias-variance frontier.

- Bias of $\hat{L}_\delta$ maximized at $-f^*_\delta$ and minimized at f^*_δ:

$$\overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(C)}(\hat{L}_\delta) = \frac{1}{n} \sum_{i=1}^n \left[f^*_\delta(x_i, 1) - f^*_\delta(x_i, 0) \right] - \sum_{i=1}^n k^*_\delta(x_i, d_i) f^*_\delta(x_i, d_i)$$

- Variance is

$$\text{sd}(\hat{L}_\delta)^2 = \sum_{i=1}^n k^*_\delta(x_i, d_i)^2 \sigma^2(x_i, d_i).$$

Theorem 1

Let $\hat{L}_\delta$ and $\overline{\text{bias}}_\delta = \overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}}(\hat{L}_\delta)$ be given above and let

$$\hat{c}_{\alpha,\delta} = \hat{L}_\delta - \overline{\text{bias}}_\delta - \text{sd}(\hat{L}_\delta) z_{1-\alpha}.$$

Then

1. $[\hat{c}_{\alpha,\delta}, \infty)$ is a $1-\alpha$ CI over $\mathcal{F}_{\text{Lip}}(C)$, and it has optimal minimax βth quantile excess length over all $1-\alpha$ CIs where $\beta = \Phi(\delta - z_{1-\alpha})$ and Φ denotes the standard normal cdf.
2. $\left\{\hat{L}_\delta \pm \text{cv}_\alpha(\overline{\text{bias}}_\delta / \text{sd}(\hat{L}_\delta))\text{sd}(\hat{L}_\delta)\right\}$ is a $1-\alpha$ CI over $\mathcal{F}_{\text{Lip}}(C)$ for any δ. Optimal FLCI centered at an affine estimator takes this form with $\delta = \delta_\chi$ where δ_χ minimizes $\text{cv}_\alpha(\overline{\text{bias}}_\delta / \text{sd}(\hat{L}_\delta))\text{sd}(\hat{L}_\delta)$ over δ.

- For one-sided CIs, optimality follows by application of results in AK17, who use results on minimax testing of convex null against convex alternative
- For two-sided CI, the theorem above only gives optimality among affine FLCIs. However, the optimal affine FLCI is close to optimal among all CIs, under both minimax criterion and at smooth functions (see AK17).
    - In our application, we calculate that our CI is at least 94% efficient

Optimal estimators

- Estimation under MSE criterion: minimax affine estimator is $\hat{L}_{\delta_\rho}$ where δ_ρ minimizes $\overline{\text{bias}}_\delta^2 + \text{sd}(\hat{L}_\delta)^2$. It is near-minimax among all estimators (follows from Donoho, 1994)
- If we want CI centered at optimal point estimate, we can form a FLCI based on MSE optimal estimate: $\left\{\hat{L}_{\delta_\rho} \pm \text{cv}_\alpha(\overline{\text{bias}}_{\delta_\rho}/\text{sd}(\hat{L}_{\delta_\rho}))\text{sd}(\hat{L}_{\delta_\rho})\right\}$.
- In our application, we δ_ρ and δ_χ are close to each other, so not much efficiency is lost by doing this.
- Intuition: better to widen CI to take into account bias than to undersmooth.

- Since $\sigma^2(x_i, d_i)$ is unknown, we recommend replacing it with estimate or guess $\tilde{\sigma}^2(x_i, d_i)$ to compute the optimal weights and variance of estimator
  - We assume homoscedasticity for computing optimal weights, but drop this assumption for estimating variance
  - Analogous to OLS with heteroscedasticity robust standard errors

- Nearest neighbor matching estimator with M matches is linear estimator with weights

$$k_{\text{match},M}(x_i, d_i) = \frac{1}{n}(2d_i - 1)\left(1 + \frac{K_M(i)}{M}\right).$$

  where $K_M(i)$ is number of times observation i is matched [↗ Details]

- To compute CIs, evaluate minimax MSE, etc. for matching estimators, we just need to compute $\overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(C)}$ for these weights. This is another convex programming problem.
- We show that, when C is large enough, optimal estimator $\hat{L}_\delta$ is the matching estimator with $M = 1$.

Optimality of matching with low smoothness

Theorem 2

Suppose $\sigma^2(x_i, d_i) = \sigma$ is constant and that distances $\|x_i - x_j\|_{\mathcal{X}}$ take on unique values as i and j vary. Then there exists a constant K depending on σ and $\{x_i, d_i\}_{i=1}^n$ such that, if $C/\delta > K$, optimal estimator $\hat{L}_\delta$ is given by the matching estimator with $M = 1$.

- Intuition: Matching estimators with $M = 1$ minimize bias
- For lower values of C, optimal estimator takes form of matching estimator with variable number of matches (follows from our algorithm)
  - Observations matched more times considered "further away"
  - NN estimate weights observations by their distance

Conventional approach uses asymptotics based on the assumption that semiparametric efficiency bound is achieved.

- Suppose x_i and d_i drawn from a distribution where $P(d_i = 1|x_i = x) = p(x)$, and we have enough smoothness (i.e. $\mathcal{F}$ bounds high enough order derivative).
- Then optimal rate of convergence is $\sqrt{n}$, bias asymptotically negligible, and optimal weights asymptotically proportional to efficient influence function k_{seb} (Hahn, 1998)
- We show that optimal rate is slower than $\sqrt{n}$ for Lipschitz when $\dim X_i > 2$, in general need derivative of order $\dim(X_i)/2$

Setup

Optimal estimators/CIs

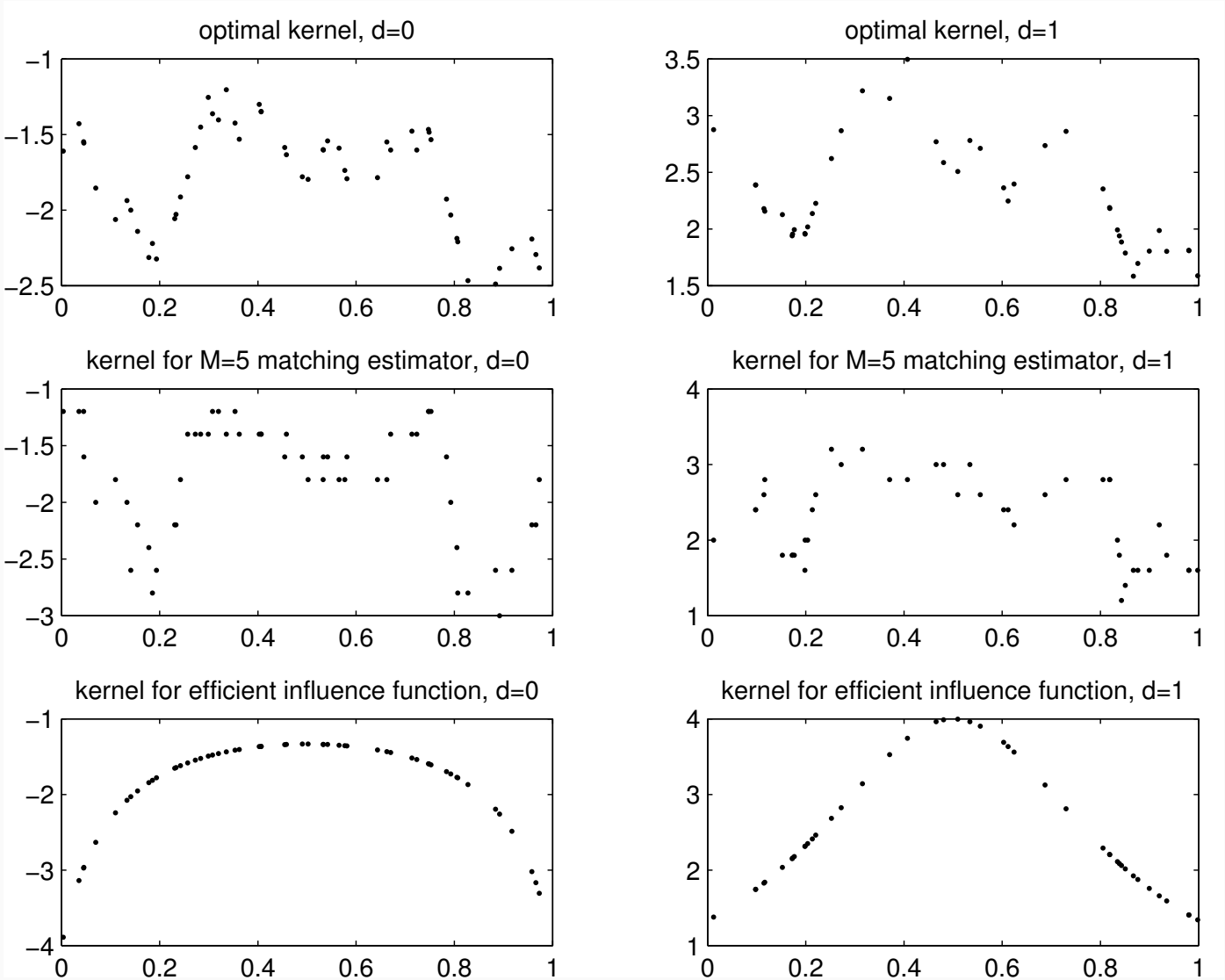
Numerical illustration

Application

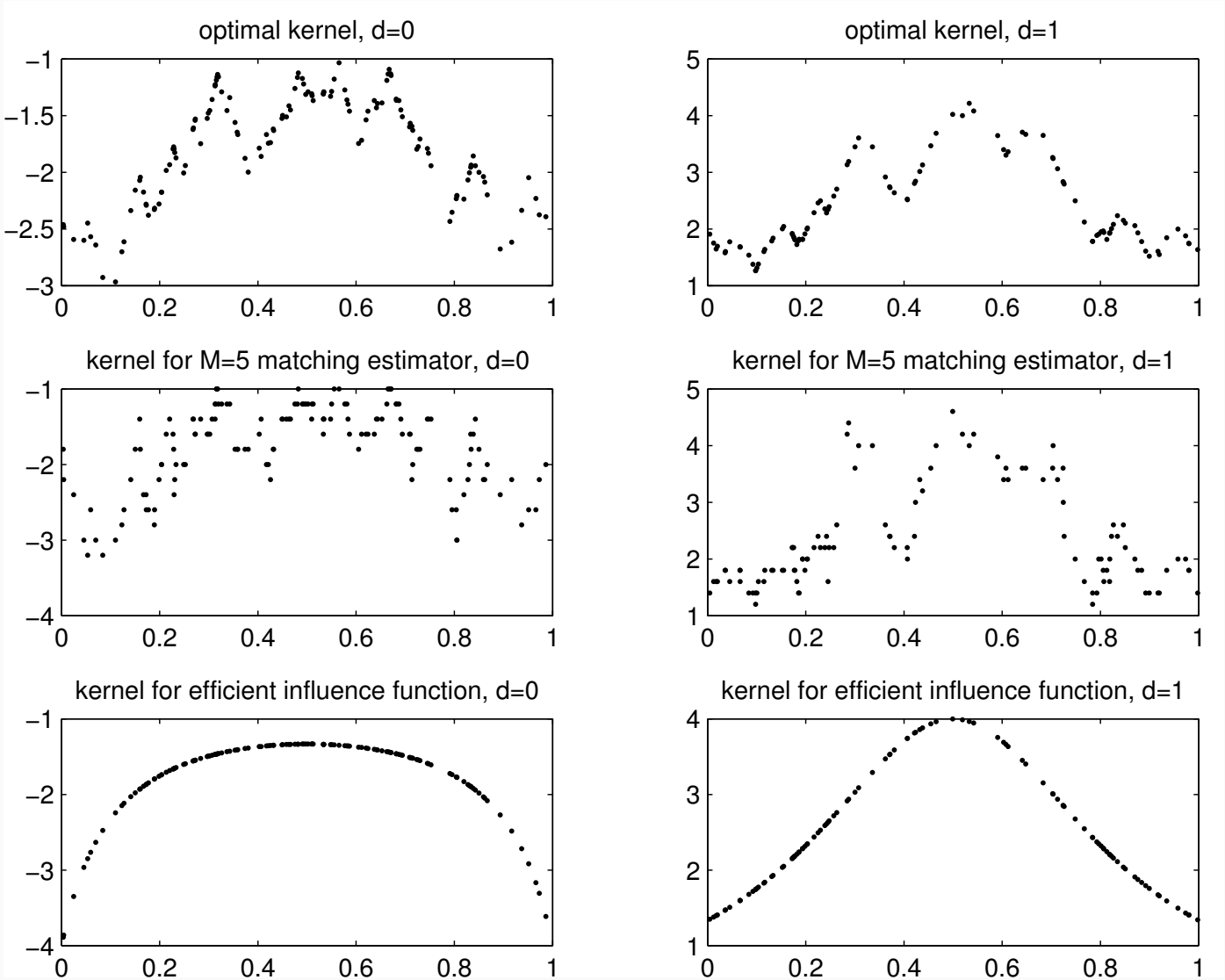
Conclusion

- To see what the optimal weights look like, we take single draw of $\{x_i, d_i\}_{i=1}^n$ from a known data generating process with a single covariate. Then plot the optimal weights k_δ^*
  - We take $\delta = 2z_{.95}$ (optimizes .95 quantile of excess length for 95% CI)
- Since $p(x) = P(d_i = 1|x_i = x)$ is known, we can also plot the efficient influence function k_{seb}.
- Also plot the matching weights $k^*_{\text{match},M}$ with $M = 5$.
- Multiply all weights by n for comparison across sample sizes

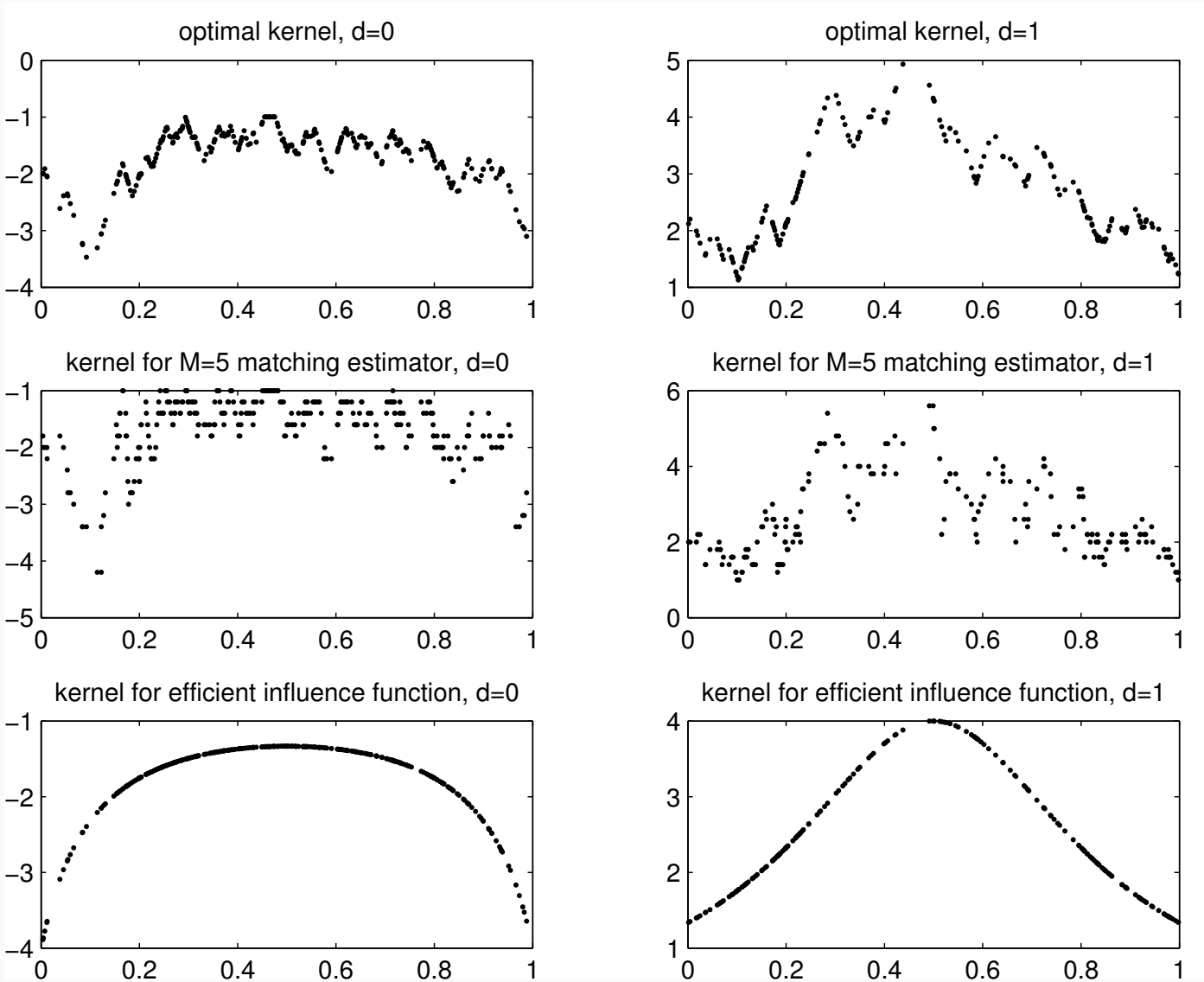
Estimator weights for $n = 100$



Estimator weights for $n = 250$



Estimator weights for $n = 500$



Numerical Illustration

	Optimal estimator		Matching estimator	
n	bias	sd	bias	sd
100	0.0201	0.2053	0.0202	0.2081
250	0.0087	0.1331	0.0079	0.1353
500	0.0057	0.0963	0.0048	0.0983

Notes on numerical illustration

- Dimension of x_i is one, so we achieve semiparametric efficiency bound with optimal estimator.
- Matching achieves the bound if $M \to \infty$ at the appropriate rate.
- Numerical results seem to reflect this:
  - Weights are not too far off from efficient influence function for $n = 500$.
  - Worst-case bias is small relative to standard error (about 5–10% of standard error)
- On the other hand, will see that bias is substantial relative to standard error in NSW data even though n is much larger—reflects greater dimension of x_i

Setup

Optimal estimators/CIs

Numerical illustration

Application

Conclusion

National Supported Work Demonstration (NSW)

- y_i: earnings in 1978 (after training program) in \$1,000s
- d_i: indicator for program participation
- x_i: age, education, indicators for Black and Hispanic, indicator for marriage, earnings in 1974, earnings in 1975, and employment indicators for 1974 and 1975

Sample of treated ($d_i = 1$) individuals from NSW, sample of untreated ($d_i = 0$) individuals from PSID (data from Dehejia and Wahba, 1999)

- NSW also included randomized controls, which we don't use.
- Can non-experimental sample replicate experimental results? (Lalonde (1985), Dehejia and Wahba (1999), Smith and Todd (2001, 2005), Abadie and Imbens (2011))

  Following this literature, focus on average treatment effect on the treated (CATT):

$$Lf = \frac{\sum_{i=1}^{n} [f(x_i, 1) - f(x_i, 0)] d_i}{\sum_{i=1}^{n} d_i}$$

- Results for ATE go through with obvious adjustments (see paper)

- Definition of $\mathcal{F}_{\text{Lip}}(C)$ depends on norm that defines distance on x
- Definition of matching estimators also depends on this
- Let A be positive definite symmetric matrix and define norm

$$\|x\|_{A,p} = \left(\sum_{i=1}^{n} |(A^{1/2}x)_i|^p \right)^{1/p}$$

- Focus on $p = 1$ and A diagonal: Lipschitz condition easier to interpret
- With $C = 1$ and A diagonal, j, jth element gives a priori bound on derivative of regression function with respect to jth element of x

Ideally, choose A based on a priori restrictions one is willing to impose

- Here, focus on $A^{1/2} = A^{1/2}_{\text{main}} = \text{diag}(0, 1, 20, 20, 0, 1, 0, 0, 0)$

  $\Longrightarrow$ suffices to control for education, previous year's earnings and Black/Hispanic indicators when making the selection-on-observables assumption

- With $C = 1$
  - No more than one-to-one effect of previous year's earnings on this year's earnings
  - Bound of \$1,000 increase in wages per year of education
  - Bound of \$20,000 for wage gap between Black or Hispanic and others

Earnings are less than \$5,000 for most of treated sample and less than \$20,000 for most of sample overall, so these bounds are large in percentage terms

Issues with this approach to choosing norm

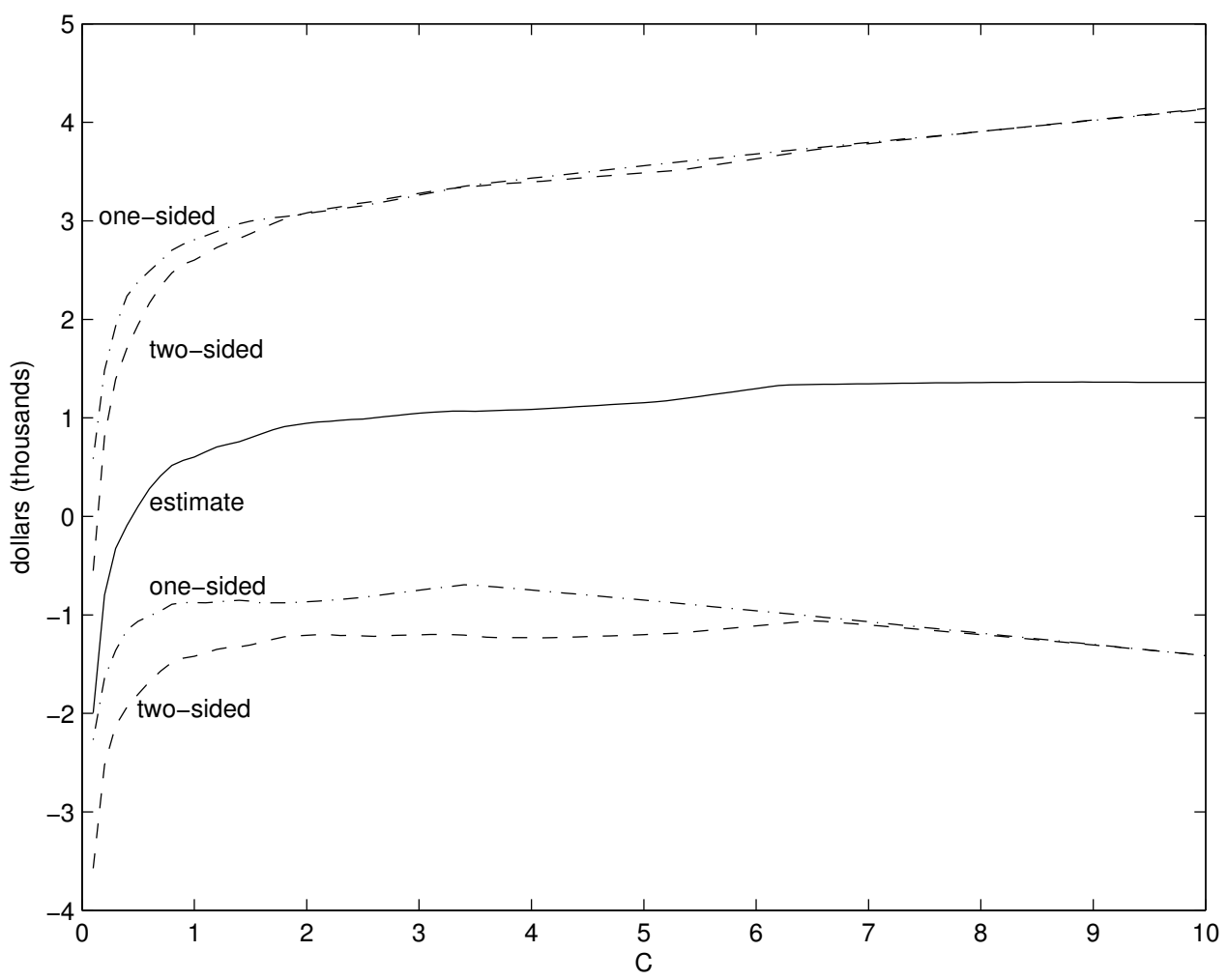
- Asking a lot of researcher
- Concern one might try many different choices and justify one ex post

  Alternative: use invariance considerations to pick a norm

- For example, let $A^{1/2}$ be diagonal matrix with j, jth given by $1/\text{sd}(x_{ij})$ (invariant to choice of units for x) [↗ More]

1. Choice of C
    - By results in AK17, one must choose C a priori: one cannot start with a conservative choice and "let the data tell us" that C is, in fact, smaller than we thought [↗ Details]
    - We recommend plotting the CIs and estimates as a function of C
    - We have argued that $C = 1$ is plausible, so we take this as a benchmark and include it in our range of choices
2. Choice of quantile of excess length for one-sided CI
    - We take 0.8 since it corresponds to a benchmark in statistical power analysis (Cohen, 1988)

Optimal estimator and CIs for CATT in NSW data

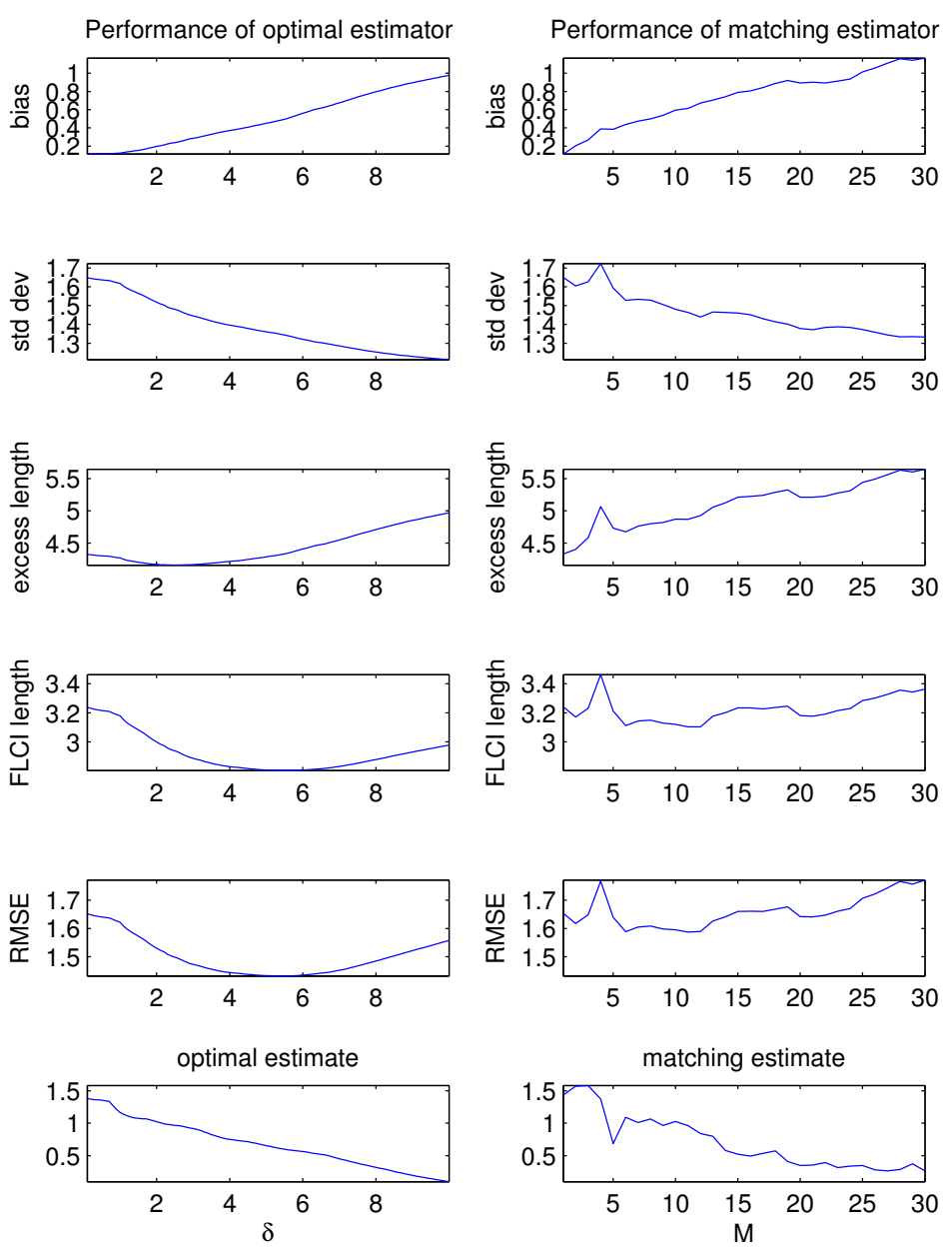


Comments on results

- For C small, Lipschitz assumption implies that selection on pretreatment variables does not lead to substantial bias.
- Estimator incorporates this by tending towards raw difference-in-means estimate (which is negative for this data set).
- For larger C, estimator gets larger. This suggests that the estimator is accounting for the possibility of selection bias by controlling for observables.
- As required by our results, the estimator is identical to matching with $M = 1$ for large enough C.

Closer look at results for $C = 1$

- Optimal estimator is $\hat{L}_\delta$, where δ is chosen to optimize bias-variance tradeoff for the given criterion (one-sided CI/FLCI/MSE).
- To illustrate this, we plot bias, standard deviation, one-sided excess length, FLCI length and MSE against δ.
- For comparison M plays same role as δ for matching estimator (trades off bias and variance). We plot these quantities against M for the matching estimator.


Performance of optimal estimator
Performance of matching estimator
bias
std dev
excess length
FLCI length
RMSE
optimal estimate
matching estimate
δ
M

Optimal estimator

criterion	δ	$\hat{L}$	$\overline{\text{bias}}$	se	robust se
one-sided CI	2.4865	0.9652	0.2376	1.4806	0.9752
FLCI	5.5988	0.5906	0.5060	1.3399	0.9009
MSE	5.4672	0.6019	0.4920	1.3449	0.9033

Matching estimator

criterion	M	$\hat{L}$	$\overline{\text{bias}}$	se	robust se
one-sided CI	1	1.4353	0.1137	1.6496	0.9950
FLCI	12	0.8409	0.6739	1.4393	0.8540
MSE	11	0.9609	0.6132	1.4641	0.8541

Comments

- Bias nonnegligible proportion of CI length (about half of standard deviation)
- FLCI and RMSE make similar choices for bias-variance tradeoff: one can check that FLCI centered at the RMSE optimal estimator is near optimal for FLCI in this data
    - This is a numerical result for this data set and application rather than a general result. However, we have found similar results in other settings.
    - Don't have to worry about CI and estimate giving radically different results
- On the other hand, the one-sided CI chooses a smaller bias/standard-deviation ratio

Comparison to experimental estimates

Difference-in-means estimate of ATT using experimental controls (not used here) is 1.794. Our estimates can differ due to:

1. Difference between CATT and ATT
2. Failure of selection-on-observables assumption (so that $\frac{\sum_{i=1}^{n}[f(x_i,1)-f(x_i,0)]d_i}{\sum_{i=1}^{n} d_i}$ does not give the actual CATT)
3. Bias and variance in estimating $\frac{\sum_{i=1}^{n}[f(x_i,1)-f(x_i,0)]d_i}{\sum_{i=1}^{n} d_i}$

Bias in 3. may play a substantial role in explaining difference:

- For example, with $C = 1$, optimal MSE estimate is 0.60, with worst-case bias 0.49: bias accounts for almost half of difference

Setup

Optimal estimators/CIs

Numerical illustration

Application

Conclusion

Conclusion

- Derived optimal estimators/CIs for average treatement effects under unconfoundedness.
- Estimators are linear in outcomes and trade off bias and variance optimally.
- With conservative Lipschitz (first-derivative) constraint, matching with a single match is optimal.
- In general, estimators can be computed using convex optimization.
- CIs take the form of estimator plus-or-minus critical value, but wider than conventional CIs since they take into account worst-case bias explicitly.

- Our general results apply to any convex function class:
  - Partly linear model
  - Higher order smoothness
  - Smoothness assumptions on residuals of best linear predictor (natural for regression adjusted matching)
  - Smoothness assumptions on differenced conditional means (natural for difference-in-difference matching)
- Other applications

Pros and Cons of conditioning

(+) Takes into account realized covariate imbalance between treated and control groups.

- Consider RCT with d_i randomly assigned, but most men end up in treatment group: conditioning takes into account resulting bias concerns

(-) Cannot use smoothness of propensity score $p(x) = P(d_i = 1 \mid x_i = x)$.

- Our view: worth working out sharp efficiency bounds in both cases and comparing them.
- If one prefers not to condition, our CIs still valid. Also rate optimal if no smoothness on $p(x)$ imposed (conjecture: they are close to optimal in finite samples)

[↗ Back]

Formal definition of CATE

- Assume: $\{(x_i, d_i, y_i(0), y_i(1))\}_{i=1}^n$ are i.i.d. and $\{y_i(1), y_i(0)\} \perp\!\!\!\perp d_i \mid x_i$
- Then

$$\frac{1}{n}\sum_{i=1}^{n} E[y_i(1) - y_i(0) \mid d_i, \ldots, d_n, x_1, \ldots, x_n] = \frac{1}{n}\sum_{i=1}^{n}(f(X_i, 1) - f(X_i, 0)),$$

  where $f(x, d) = E(y_i(d) \mid x_i = x) = E(y_i | d_i = d, x_i = x)$, and
  $y_i = f(x_i, d_i) + u_i$
- Assumption that u_i is (conditionally) normal follows from assumption that $y_i(0)$ and $y_i(1)$ are normal (but not necessarily joint normal) conditional on $\{(x_i, d_i)\}_{i=1}^n$

[↗ Back]

Adaptation bounds

- Minimax may be "too pessimistic" and requires knowing $\mathcal{F}$ (e.g., need to know C to use $\mathcal{F}_{\text{Lip}}(C)$)
- Potential solution: adaptive inference
  - Require coverage over $\mathcal{F}$, but optimize length simultaneously over $\mathcal{F}$ and one or more smaller sets $\mathcal{G}_j \subsetneq \mathcal{F}$ (e.g. $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C_0)$ and $\mathcal{G}_j = \mathcal{F}_{\text{Lip}}(C_j)$, $C_J < \cdots < C_1 < C_0$
- When $\mathcal{F}$ centrosymmetric ($f \in \mathcal{F}$ implies $-f \in \mathcal{F}$), results in AK17 show that very little can be gained by adaptive inference.

  $\Longrightarrow$ CIs must depend explicitly on C: one cannot be conservative and then "let the data show" that C is in fact smaller.
- If $\mathcal{F}$ is asymmetric (e.g. monotonicity restrictions) or nonconvex (e.g. Giné-Nickl style restrictions), adaptive CIs may be possible.

[↗ Back to Efficiency Criteria] [↗ Back to Application]

- Matching estimators take the form

$$\frac{1}{n}\sum_i \left[d_i(y_i - \hat{f}(x_i, 0)) + (1 - d_i)(\hat{f}(x_i, 1) - y_i)\right],$$

  where $\hat{f}(x, d)$ is estimate of $f(x, d)$.
- We consider nearest neighbor matching: $\hat{f}(x_i, 0)$ is M-nearest neighbor estimate of $f(x_i, 0)$ among observations with $d_j = 0$, similarly for $f(x_i, 1)$.
- If $d_i = 1$, $d_j = 0$ and y_j is used in forming $\hat{f}(x_i, 0)$, we say that j is used as a match (and similarly for $d_i = 0$, $d_j = 1$). $K_M(j)$ denotes number of times j is used as a match.

[↗ Back]

Alternative choice of norm

- Diagonal elements of $A_{ne}^{1/2} = \text{diag}(1/\text{std}(x_1), \ldots, 1/\text{std}(x_{d_x}))$:

Age	Educ.	Black	Hispanic	Married
0.0952	0.3275	2.1998	5.4864	2.5993

1974 earnings	197 earnings	1974 emp.	1975 emp.
0.0729	0.0721	2.9793	2.9297

- Each entry gives bound on derivative of conditional mean wrt x_j (when $C = 1$).
  - Probably doesn't reflect relative magnitude of a priori bounds for most researchers: e.g., wage gap is much larger for Hispanics than Blacks.
- To allow for a one-to-one effect of last year's earnings (as with the main specification with $C = 1$), need C above 10, which leads to very wide CIs (for $C = 10$, optimal FLCI is 1.7176 ± 7.6797). [↗ Back]