Adapting to Misspecification

Tim Armstrong[1] Patrick Kline[2] Liyang Sun[3]

[1]USC; [2]UC Berkeley and NBER; [3]CEMFI

August 2023

Motivation: robustness-efficiency tradeoff

- When studying a scalar estimand θ, empirical researchers commonly report "robustness exercises" comprised of
  - an asymptotically unbiased estimate Y_U
  - a restricted estimate Y_R with asymptotic bias b, but asymptotically more efficient if no bias
- This paper: a novel way to combine Y_U and Y_R into an optimally *adaptive* estimate

Local misspecification framework

- More generally, the bias can arise from using (potentially misspecified) overidentifying restrictions
- Let $Y_O = Y_R - Y_U$ be an estimate of the bias b
- Asymptotic approximation:

$$\begin{pmatrix} Y_U \\ Y_O \end{pmatrix} \sim_a N\left(\begin{pmatrix} \theta \\ b \end{pmatrix}, \Sigma\right), \quad \Sigma = \begin{pmatrix} \Sigma_U & \rho\sqrt{\Sigma_U}\sqrt{\Sigma_O} \\ \rho\sqrt{\Sigma_U}\sqrt{\Sigma_O} & \Sigma_O \end{pmatrix}$$

- Most of our examples are linear so the asymptotic approximation is also valid globally
- Common to report $T_O = Y_O/\Sigma_O^{1/2}$ as an *over-identification* test

This paper: *Adapting* to misspecification (1/2)

Overview of logic:

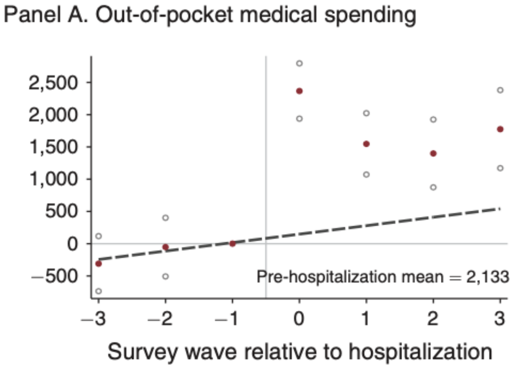
i) If b were known, efficient to use GMM imposing that

$$\mathbb{E}[Y_R - b] = \mathbb{E}[Y_U] = \theta$$

ii) If only know $|b| \leq B$, natural to seek an estimator that minimizes the *worst-case risk* over the parameter space $|b| \leq B$

We call such an estimator the "B-minimax" estimator and denote $R^*(B)$ to be its *worst-case risk*

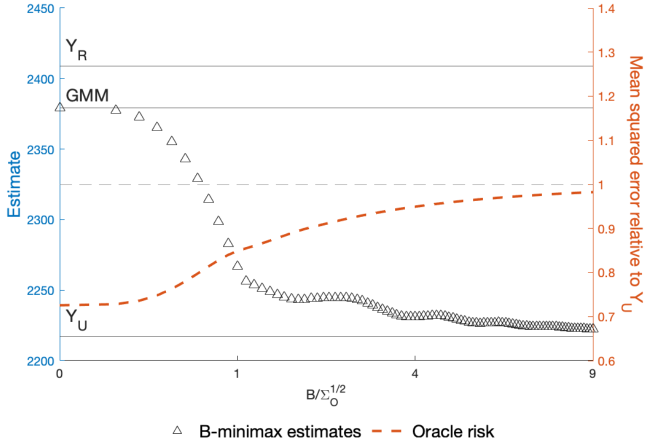
Dobkin, Finkelstein, Kluender and Notowidigdo (2018)



- Estimates the effects of an unexpected hospitalization on medical spending
- The researchers report Y_U, assuming a linear pre-trend
- Imposing no pre-trend, can also report a more precise Y_R

Illustration: *B-minimax* estimates

- To assess the sensitivity of Y_R to bias, we can also report different *B-minimax* estimates varying the bound B on the amount of bias



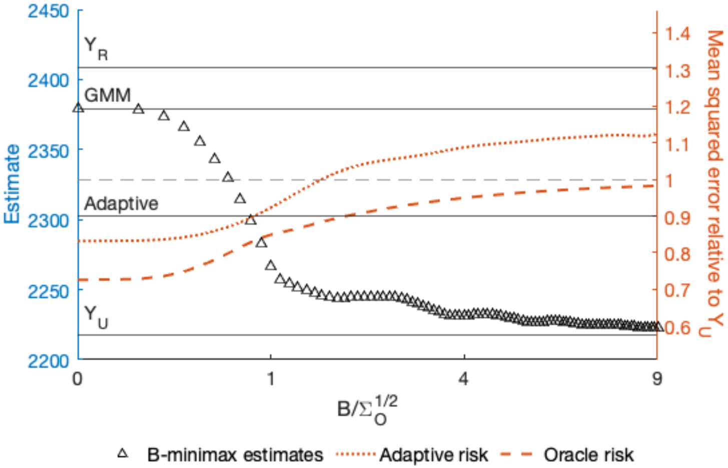
This paper: *Adapting* to misspecification (2/2)

Overview of logic:

iii) Propose an optimally *adaptive* estimator without reference to the upper bound B

- *Adaptation*: how well can a single estimator mimic the oracle that knows B ?
- When B is unknown, will always regret not being able to implement relevant B-minimax decision
- The *optimally adaptive estimator* tries to get as close as possible to the oracle risk function over $B \in \mathcal{B}$

Illustration: *Adaptive* estimate



Related literature: adaptive estimation

- Common to define a procedure to be "adaptive" over a set of parameter spaces if it is simultaneously near-minimax for all of these parameter spaces.
  - Our definition (best constant multiplicative factor) corresponds to Tsybakov (1998)
- Typically used for *asymptotic* results (rates, sometimes constants) in *nonparametric/high-dimensional* problems: adapting to smoothness/sparsity using cross-validation, etc.
- In contrast to this literature, we focus on *exact* results in *low dimensional* setting
- Exception: Bickel (1984) considers adapting between $B = 0$ and $B = \infty$
- Robustness-efficiency tradeoffs: Hodges and Lehmann (1952); Bickel (1983, 1984)

Related literature

- Specification testing: Hausman (1978); Breusch and Pagan (1980); Sargan (1988); Guggenberger (2010)
- Model averaging: Akaike (1973); Mallows (1973); Schwarz (1978); Claeskens and Hjort (2003); Hansen (2007); Hansen and Racine (2012); Cheng, Liao and Shi (2019); Fessler and Kasy (2019); de Chaisemartin and D'Haultfœuille (2020)
- Adaptive *confidence interval*: impossible to tighten minimax CI and maintain coverage for all b (Low, 1997; Armstrong and Kolesar, 2018)
  - Worst-case asym. variance of our adaptive estimator is bounded by Σ_U
  - Adaptive estimate is always included in the robust CI
- Computation: Chamberlain (2000); Elliott, Müller and Watson (2015); Müller and Wang (2019); Kline and Walters (2021)

Outline

General setup

- Researcher observes data or initial estimates $Y = (Y_U, Y_O)$
- Distribution of Y depends on unknown parameters (θ, b)
- Use $P_{\theta,b}$ and $E_{\theta,b}$ to denote probability and expectation under (θ, b)
- *Decision rule* $\delta()$ maps Y to an action a

General setup

- Researcher observes data or initial estimates $Y = (Y_U, Y_O)$
- Distribution of Y depends on unknown parameters (θ, b)
- Use $P_{\theta,b}$ and $E_{\theta,b}$ to denote probability and expectation under (θ, b)
- *Decision rule* $\delta()$ maps Y to an action a
- Loss function $L(\theta, b, a)$ measures disutility of action a when the parameter is (θ, b)
- Risk:

$$R(\theta, b, \delta) = E_{\theta,b} L(\theta, b, \delta(Y)) = \int L(\theta, b, \delta(y))\, dP_{\theta,b}(y)$$

- E.g. estimation with squared error loss $L(\theta, b, a) = (a - \theta)^2$, implying risk is given by mean squared error (MSE)

Constrained minimax estimation

- The researcher entertains multiple parameter spaces:

$$\mathcal{C}_B = \{(\theta, b) : |b| \leq B\} = \mathbb{R} \times [-B, B]$$

- The parameter spaces are indexed by $B \in \mathcal{B}$ for $\mathcal{B} = [0, \infty]$

Constrained minimax estimation

- The researcher entertains multiple parameter spaces:

$$\mathcal{C}_B = \{(\theta, b) : |b| \leq B\} = \mathbb{R} \times [-B, B]$$

- The parameter spaces are indexed by $B \in \mathcal{B}$ for $\mathcal{B} = [0, \infty]$
- Given a parameter space $\mathcal{C}_B$, the *worst-case risk* is

$$R_{\max}(B, \delta) = \sup_{(\theta, b) \in \mathcal{C}_B} R(\theta, b, \delta)$$

- The minimax estimator under $\mathcal{C}_B$ ("B-minimax estimator" for short) optimizes $R_{\max}(B, \delta)$, thereby obtaining the *minimax risk* for $\mathcal{C}_B$:

$$R^*(B) = \inf_{\delta} R_{\max}(B, \delta) = \inf_{\delta} \sup_{(\theta, b) \in \mathcal{C}_B} R(\theta, b, \delta)$$

- Sensitivity analysis: compute "B-minimax estimator" for a range of $B \in \mathcal{B}$
  - Under MSE, the "B-minimax estimator" is GMM when $B = 0$ and Y_U when $B = \infty$

Outline

Outline

What is the cost of committing to a single decision δ ?

- We define the *adaptation regret* under $\mathcal{C}_B$ of δ to be :

$$A(B,\delta) = \frac{R_{\max}(B,\delta)}{R^*(B)}$$

that is, the proportional increase in worst-case risk relative to the *minimax risk* (oracle)

What is the cost of committing to a single decision δ ?

- We define the *adaptation regret* under $\mathcal{C}_B$ of δ to be :

$$A(B, \delta) = \frac{R_{\max}(B, \delta)}{R^*(B)}$$

that is, the proportional increase in worst-case risk relative to the *minimax risk* (oracle)

- The *optimally adaptive estimator* δ^{adapt} minimizes over $B \in \mathcal{B}$ the *worst-case adaptation regret* :

$$A^*(\mathcal{B}) = \inf_{\delta} \sup_{B \in \mathcal{B}} A(B, \delta) = \inf_{\delta} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta)}{R^*(B)}$$

Following Tsybakov (1998), we refer to $A^*(\mathcal{B})$ as the *loss of efficiency under adaptation*

Adaptation as weighted minimax

Optimally adaptive estimator solves

$$\inf_{\delta} \sup_{B \in \mathcal{B}} A(B, \delta) = \inf_{\delta} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta)}{R^*(B)} \qquad \text{(def. of adaptation regret)}$$

$$= \inf_{\delta} \sup_{B \in \mathcal{B}} \sup_{(\theta, b) \in \mathcal{C}_B} \frac{R(\theta, b, \delta)}{R^*(B)} \qquad \text{(def. of worst-case risk)}$$

$$= \inf_{\delta} \sup_{(\theta, b) \in \cup_{B' \in \mathcal{B}'} \mathcal{C}_{B'}} \omega(\theta, b) R(\theta, b, \delta) \qquad \text{(reordering of sup)}$$

- We derive the general formula for the weights in the paper
- For nested sets $\mathcal{C}_B = \mathbb{R} \times [-B, B]$ and $R^*(B)$ is increasing in B
- The weight is $\omega(\theta, b) = R^*(|b|)^{-1}$

Solving minimax problems

- Consider the general problem of computing a minimax decision over a parameter space $\mathcal{C}$ for a parameter ϑ under loss $\bar{L}(\vartheta, a)$
- Let $R(\vartheta, \delta) = E_\vartheta \bar{L}(\vartheta, \delta)$ denote risk
- Letting π denote a *prior* distribution on $\mathcal{C}$, the *Bayes risk* of δ is given by

$$R_{\text{Bayes}}(\pi, \delta) = \int R(\vartheta, \delta)\, d\pi(\vartheta) = \int E_\vartheta \bar{L}(\vartheta, \delta(Y))\, d\pi(\vartheta)$$

Solving minimax problems

- Let Γ denote the set of priors π supported on $\mathcal{C}$.
- Under regularity conditions, a *minimax theorem* says that

$$\min_{\delta} \max_{\vartheta \in \mathcal{C}} R(\vartheta, \delta) = \min_{\delta} \max_{\pi \in \Gamma} R(\vartheta, \delta) = \max_{\pi \in \Gamma} \min_{\delta} R(\vartheta, \delta)$$

  The maximizing π is called *least favorable prior*
- The inner minimization is solved by the *Bayes decision* $\delta_{\pi}^{\text{Bayes}}$, e.g. for estimation under squared error loss, it's the posterior mean
- The outer maximization is concave in π.

We solve for least favorable π by using a grid approximation and convex optimization, following, e.g., Chamberlain (2000)

B-minimax estimator

Further simplification under squared error loss: $L(\theta, b, a) = (\theta - a)^2$.

Theorem 1.

In the local misspecification setting, the B-minimax estimator is given by

$$Y_U - \rho\sqrt{\Sigma_U} T_O + \rho\sqrt{\Sigma_U}\delta^{BNM}(T_O; B/\sqrt{\Sigma_O})$$

and the minimax risk is

$$R^*(B) = \rho^2 \Sigma_U r^{BNM}\left(\frac{B}{\sqrt{\Sigma_O}}\right) + \Sigma_U - \rho^2\Sigma_U$$

where $\delta^{BNM}(y;\tau)$ denotes the minimax estimator and $r^{BNM}(\tau)$ denotes minimax risk in the bounded normal mean problem, in which we observe $Y \sim N(\vartheta, 1)$ and impose $|\vartheta| \leq \tau$

details

Optimally adaptive estimator

Theorem 2.

In the local misspecification setting, the optimally adaptive estimator is given by

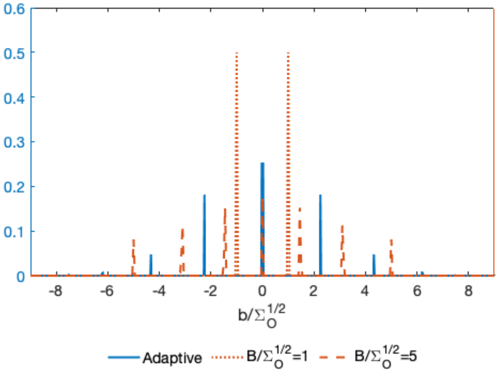
$$Y_U - \rho\sqrt{\Sigma_U}T_O + \rho\sqrt{\Sigma_U}\tilde{\delta}^{adapt}(T_O;\rho)$$

where $\tilde{\delta}^{adapt}(T_O;\rho)$ *solves*

$$\inf_{\delta}\sup_{\tilde{b}\in\mathbb{R}} \frac{E_{T\sim N(\tilde{b},1)}(\tilde{\delta}(T)-\tilde{b})^2+\rho^{-2}-1}{r^{BNM}(|\tilde{b}|)+\rho^{-2}-1}.$$

The loss of efficiency under adaptation is given by the value of the above display.

Illustration: Adaptive prior over b



Note: Adaptive prior works especially well when $b \approx 0$ and limits risk when $|b|$ is large

Outline

1. Setup
2. Main results
   - The adaptation problem
   - Behavior of the adaptive estimator as a function of ρ^2
   - Behavior of the adaptive estimator when $\rho^2 \to 1$
3. Empirical exercise
4. Conclusion

Weighted average interpretation

Adaptive estimator takes the form:

$$\underbrace{Y_U - \rho\sqrt{\Sigma_U} T_O}_{\text{optimal GMM using } b=0} \quad - \quad \underbrace{(-\rho)\sqrt{\Sigma_U}\tilde{\delta}^{\text{adapt}}(T_O;\rho)}_{\text{shrinkage estimate of bias}}$$

- Bias estimator $\tilde{\delta}^{adapt}(\cdot;\rho)$ yields non-linear shrinkage
- Equivalently: a weighted average of Y_U and GMM, with convex weighting function $w(T_O) = \tilde{\delta}^{adapt}(T_O;\rho)/T_O$
- Tuning free shrinkage with $n < 3$!

Weighted average interpretation

Adaptive estimator takes the form:

$$\underbrace{Y_U - \rho\sqrt{\Sigma_U}T_O}_{\text{optimal GMM using } b=0} - \underbrace{(-\rho)\sqrt{\Sigma_U}\tilde{\delta}^{\text{adapt}}(T_O;\rho)}_{\text{shrinkage estimate of bias}}$$

- Bias estimator $\tilde{\delta}^{adapt}(\cdot;\rho)$ yields non-linear shrinkage
- Equivalently: a weighted average of Y_U and GMM, with convex weighting function $w(T_O) = \tilde{\delta}^{adapt}(T_O;\rho)/T_O$
- Tuning free shrinkage with $n < 3$!
- Shape depends only on correlation ρ between Y_U and Y_O
- Compute via convex programming and provide a simple "lookup table" taking as inputs (Y_U, Y_R, Σ)

Simple "nearly-adaptive" estimator

$$\underbrace{Y_U - \rho\sqrt{\Sigma_U}T_O}_{\text{optimal GMM using } b=0} \quad - \quad \underbrace{(-\rho)\sqrt{\Sigma_U}\tilde{\delta}^{\text{adapt}}(T_O;\rho)}_{\text{shrinkage estimate of bias}}$$

- Simple "nearly-adaptive" estimator: replace $\tilde{\delta}^{\text{adapt}}(T_O;\rho)$ with *soft thresholding* estimator:

$$\delta_{S,\lambda}(T_O) = \max\{|T_O| - \lambda, 0\}\,\text{sgn}(T_O)$$

where the threshold λ is optimized over the same criterion, and therefore depends on ρ

Simple "nearly-adaptive" estimator

$$\underbrace{Y_U - \rho\sqrt{\Sigma_U}T_O}_{\text{optimal GMM using } b=0} \quad - \quad \underbrace{(-\rho)\sqrt{\Sigma_U}\tilde{\delta}^{\text{adapt}}(T_O;\rho)}_{\text{shrinkage estimate of bias}}$$

- Simple "nearly-adaptive" estimator: replace $\tilde{\delta}^{\text{adapt}}(T_O;\rho)$ with *soft thresholding* estimator:

$$\delta_{S,\lambda}(T_O) = \max\{|T_O| - \lambda, 0\}\,\text{sgn}(T_O)$$

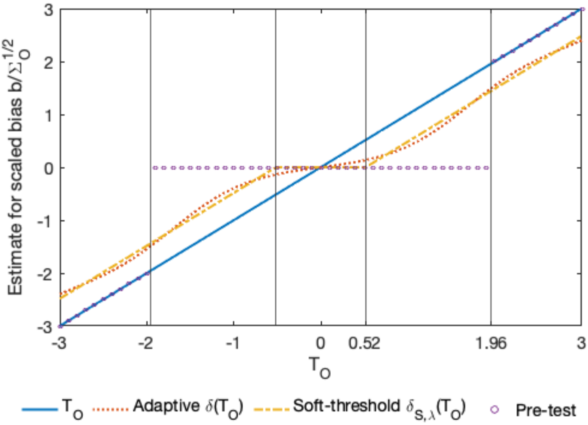
  where the threshold λ is optimized over the same criterion, and therefore depends on ρ

- Naive *pre-test* estimator corresponds to *hard thresholding*

$$\delta_{H,\lambda}(T_O) = T_O \cdot I(|T_O| > \lambda)$$

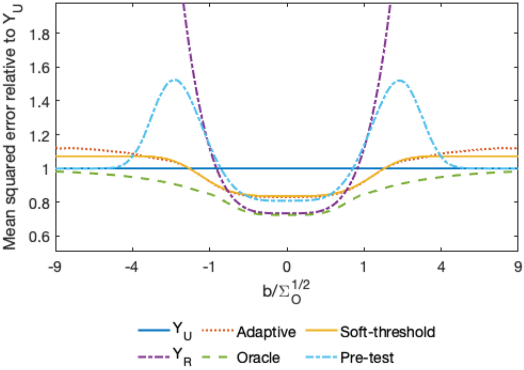
  with $\lambda = 1.96$

Illustration: Three estimators of bias



Note: Estimators of scaled bias for Dobkin et al. example ($\rho = -0.524$)

Illustration: Risk functions



Note: MSE as a function of $b/\sqrt{\Sigma_O}$ for Dobkin et al. example ($\rho = -0.524$)

Intuition for optimally adaptive estimator

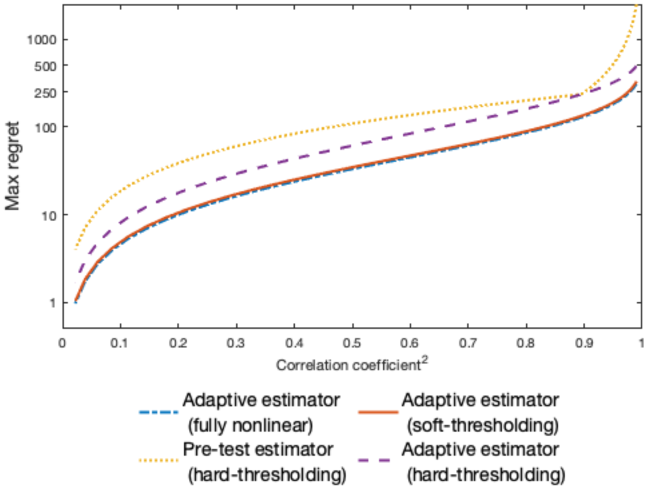
- Hard thresholding (pre-test) seems like a reasonable candidate for an adaptive estimator:
  - $|b|$ small $\Longrightarrow$ pre-test fails to reject $\Longrightarrow$ use Y_R (or optimal GMM)
  - $|b|$ large $\Longrightarrow$ pre-test rejects $\Longrightarrow$ use Y_U
- Actually, it's pretty terrible but we can improve it by
  - i) smoothing the indicator function (soft thresholding instead of hard thresholding)
  - ii) optimizing λ for adaptation (rather than just using 1.96)
- Optimally adaptive estimator improves further, but it basically looks like soft thresholding, which is already close to optimally adaptive
- This matches findings of Bickel (1984) for the version of our problem where $\mathcal{B} = \{0, \infty\}$

Outline

Behavior of the adaptive estimator when $\rho^2 \to 1$

- For a fixed ρ^2, the adaptive estimator δ^{adapt} achieves *worst-case risk* as close to the Oracle as possible
- Intuition for why the *worst-case adaptation regret* increases in ρ^2:
  - Suppose Y_R is efficient when $b = 0$, then we have $\rho^2 = \frac{\Sigma_U - \Sigma_R}{\Sigma_U}$
  - As Y_R gets more precise (and ρ^2 gets larger), adaptation places more weight on Y_R
- We show that the *asymptotic worst-case adaptation regret* and the *worst-case risk* grow at the rate of (log of) $\frac{\Sigma_U - \Sigma_R}{\Sigma_R} = \frac{\rho^2}{1-\rho^2}$
- We also show that the adaptive soft-thresholding estimator attains the same rate

Soft-thresholding tracks nonlinear adaptive estimator well



Extension: constrained adaptation

- We can bound the increase in minimax risk when the worst-case adaptation regret is large

$$\inf_{\delta} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta)}{R^*(B)} \quad \text{and} \quad \sup_{B \in \mathcal{B}} R_{\max}(B, \delta) \leq \overline{R}$$

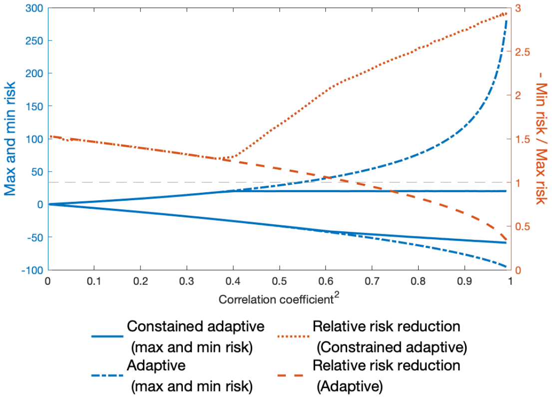
- Operationally, this constrained adaptation is equivalent to solving

$$\inf_{\delta} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta)}{\min\{R^*(B), t\}}$$

where $t = \overline{R}/A^*(\mathcal{B})$

- A variant of the original adaptation problem by putting more weights on adaptation regret at B where $R^*(B)$ is large

Still achieves relative risk reduction



Note: Always more to gain than to lose with constrained adaptation

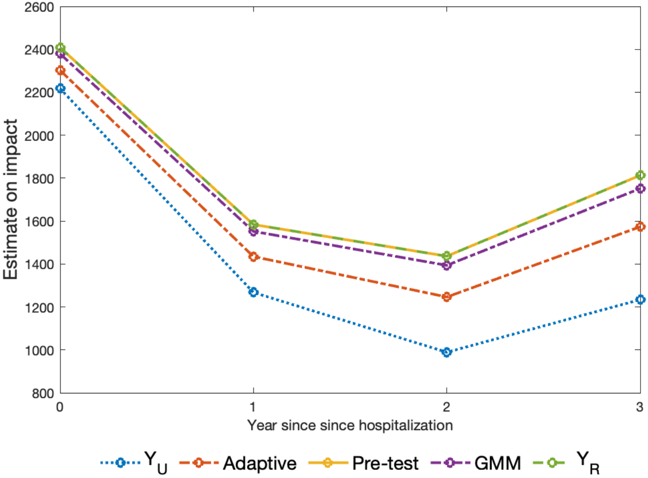
Outline

Outline

Dobkin et al (2018) original estimates

- Parameter of interest θ are impact of hospitalization on out-of-pocket (OOP) spending for the non-elderly insured (ages 50 to 59) in the Health and Retirement Studies (HRS) in the US
- We take Y_U to be the “parametric event study” estimates, where the authors control for a linear pretrend.
- We take Y_R to be estimates that impose a zero pretrend.
- The bias b is the amount of a linear pretrend. Without a pretrend, the bias is zero.

Dobkin et al (2018) adaptive estimates

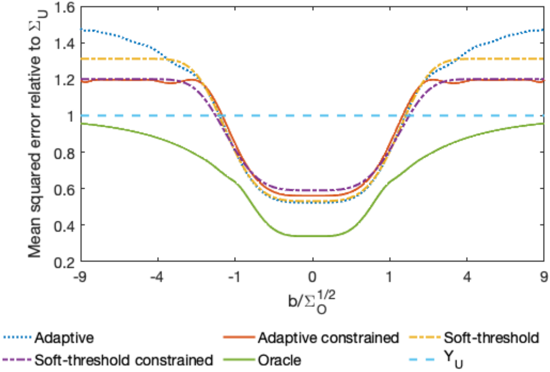


Dobkin et al (2018) first year

	Y_U	Y_R	Y_O	GMM	Adaptive	Soft-threshold	Pre-test
Estimate	2,217	2,409	192	2,379	2,302	2,287	2,409
Std Error	(257)	(221)	(160)	(219)			
Max Regret	38%	∞		∞	15%	15%	68%
Threshold						0.52	1.96

Dobkin et al (2018) risk profiles

The correlation coefficient ranges from -0.524 in the early years to -0.813 in the later years.



Risk functions for $\rho = -0.813$

Outline

Conclusion

- Applied researchers often estimate a sequence of nested specifications (robustness checks)
- We generalize the classical robustness-efficiency tradeoffs to a continuum of models, indexed by different degrees of misspecification
- Adaptive estimator uses a specification test to refine — rather than choose — estimate of a parameter by minimizing the worst case "adaptation regret"
- Pre-tabulated solutions $\rightarrow$ researcher only needs to report correlation coefficient ρ with specification test. MATLAB / R code at: `https://github.com/lsun20/MissAdapt`
- Thank you !

Using invariance

- We focus on an invariant loss function $L(\theta, b, a) = (a - \theta)^2$
- Applying invariance arguments (Lehmann and Casella, 1998, pp. 159-161), it follows that the solutions to the minimax problems take the form

$$Y_U + \bar{\delta}(T_O)$$

- Risk of this estimator doesn't depend on θ, so we can search for least favorable prior over b only.
- Least favorable prior for (θ, b) combines this with flat (improper) prior for θ.

return

Outline

- Adapting to heterogeneous effects
- Adapting to non-experimental controls

Negative weights in TWFE specifications

- Recent literature emphasizes that TWFE estimators can identify non-convex weighted averages of treatment effects $\rightarrow$ potential for biases large enough to flip sign.
- Gentzkow, Shapiro, and Sinkinson (2011) study effect of newspapers on voter turnout by estimating TWFE model via OLS.
- de Chaisemartin and D'Haultfœuille (2020) estimate that 46% of the weights underlying their TWFE specification are negative.
  - We take the GSS TWFE specification as Y_R.
  - They propose a convex weighted alternative that identifies a form of ATT. We take their estimator as Y_U.

Gentzkow, Shapiro, and Sinkinson (2011)

- Y_U exhibits large max regret bc std error $\sim$ 50% above GMM.
- Pre-test chooses non-convex Y_R but also has large regret.
- Adaptive approach puts roughly 60% of weight on Y_U.

	Y_U	Y_R	Y_O	GMM	Adaptive	Soft-threshold	Pre-test
Estimate	0.0043	0.0026	-0.0017	0.0024	0.0036	0.0036	0.0026
Std Error	(0.0014)	(0.0009)	(0.001)	(0.0009)			
Max Regret	145%	∞		∞	44%	46%	118%
Threshold						0.64	1.96

Outline

- Adapting to heterogeneous effects
- Adapting to non-experimental controls

Adapting to non-experimental controls

- LaLonde (1991): compare experimental and quasi-experimental estimates of effects of training
  - Conclusion: estimates highly sensitive to choice of specification
  - Heckman and Hotz (1989): pre-tests would have guarded against bias.
  - But how much bias was there?
- Adapt over finite set of bounds $\mathcal{B} = \{(0,0),(\infty,0),(\infty,\infty)\}$ (assumes Y_{R2} less biased than Y_{R1})
  - Y_U – experimental contrast
  - Y_{R1} – regression adjusted contrast with non-experimental control ("CPS-1")
  - Y_{R2} – regression adjusted contrast with pscore screened non-experimental control (Angrist and Pischke, 2007)

LaLonde (1991) (as in Angrist and Pischke, 2007)

- Substantial gains to combining all 3 estimates via GMM (GMM_3) but J-test rejects at 5% level.
- J-test fails to reject that Y_U and Y_{R2} have same probability limit.
- Adaptive estimate close to GMM_2. Near oracle performance.

	Y_U	Y_{R1}	Y_{R2}	GMM_2	GMM_3	Adaptive	Pre-test
Estimate	1794	794	1362	1629	1210	1597	1629
Std error	(668)	(618)	(741)	(619)	(595)		
Max Regret	26%	∞	∞	∞	∞	7.77%	47.5%
Risk rel. to Y_U							
when $b_1 = 0$ and $b_2 = 0$	1	0.853	1.23	0.858	0.793	0.855	0.80
when $b_1 \neq 0$ and $b_2 = 0$	1	∞	1.23	0.858	∞	0.925	0.993
when $b_1 \neq 0$ and $b_2 \neq 0$	1	∞	∞	∞	∞	1.077	1.475