Adapting to Misspecification

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

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

May 2023, Michigan Labor Celebration

Robustness-efficiency tradeoff

- Empiricists go to great lengths to obtain precise and *credible* estimates.
- Conventional to report standard errors to provide assessment of variability.
- Proliferation of "robustness" checks to assess possible biases.
- What to take away from such exercises?

Robustness-efficiency tradeoff

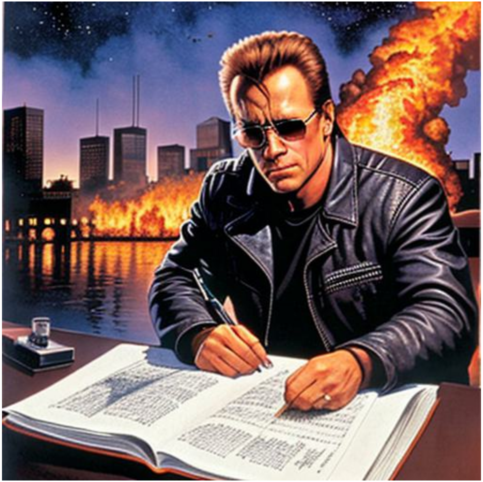
- Empiricists go to great lengths to obtain precise and *credible* estimates.
- Conventional to report standard errors to provide assessment of variability.
- Proliferation of "robustness" checks to assess possible biases.
- What to take away from such exercises?

TABLE 4—ROBUSTNESS TO ALTERNATIVE SPECIFICATIONS AND SAMPLE RESTRICTIONS FOR THE NON-ELDERLY INSURED (*Ages 50 to 59*) IN HRS

Specification	[Baseline] (1)	Individual FEs (2)	Balanced panel (3)	Wave FEs only (4)	Additional demographic controls (cubic in age; dummies for gender, race, and education) (5)	No restriction for pre-period obsevation (6)	Poisson (7)
Panel A. Out-of-pocket medical spending							
12-month effect	3,275	3,461	2,362	3,286	3,244	3,486	1.00
	(373)	(409)	(663)	(349)	(373)	(356)	(0.130)
	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Average annual effect over 36 months	1,429	1,531	1,426	1,395	1,389	1,363	0.47
	(202)	(228)	(485)	(191)	(203)	(209)	(0.083)
	[<0.001]	[<0.001]	[0.0033]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Pre-hospitalization mean	2,133	2,133	1,967	2,133	2,133	2,170	2,133

Source: Dobkin et al (2018, AER)

A minimax approach to interpreting robustness exercises



The Terminator scrutinizes the statistical tables in an issue of the Quarterly Journal of Economics while Cambridge, Mass burns in the background.

Local misspecification framework

- Consider two estimates of a scalar target parameter θ
  - an asymptotically unbiased estimate Y_U
  - a restricted estimate Y_R with asymptotic bias b, but lower variance
- Example: long vs short regression
- Let $Y_O = Y_R - Y_U$ be an estimate of the bias b
- Asymptotic approximation:

$$\begin{pmatrix} Y_U \\ Y_O \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ b \end{pmatrix}, \Sigma \right), \quad \Sigma = \begin{pmatrix} \Sigma_U & \Sigma_{UO} \\ \Sigma_{UO} & \Sigma_O \end{pmatrix}$$

- Common to report $T_O = Y_O / \Sigma_O^{1/2}$ as an *over-identification* test

Adapting to misspecification

Today: Combine Y_U and Y_R into a single optimal estimate

Overview of logic:

- If b were known, efficient to use GMM imposing that

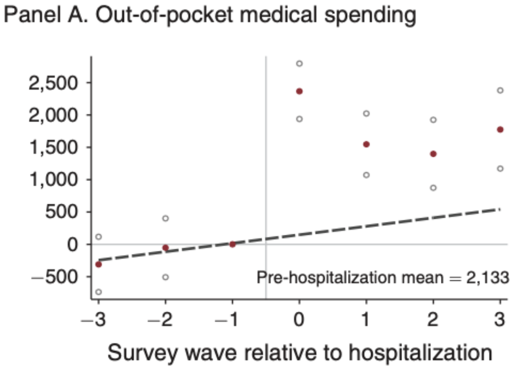
$$\mathbb{E}[Y_R - b] = \mathbb{E}[Y_U] = \theta$$

- If only know $|b| \leq B$, minimax estimation attractive
- Propose *adaptive* estimator for setting where B is unknown
  - "Shrink" Y_O to estimate b and adjust GMM accordingly
  - Achieves maximal risk near the minimax level *uniformly* in B

Related literature

- Specification testing: Hausman (1978); Breusch and Pagan (1980); Sargan (1988); Guggenberger (2010)
- Model averaging: Akaike (1973), Mallows (1973), Schwarz (1978), Leamer (1978), Claeskens and Hjort (2003), Hansen (2007), Hansen and Racine (2012), de Chaisemartin and D'Haultfœuille (2022)
- Robustness-efficiency tradeoffs: Hodges and Lehmann (1952), Bickel (1983, 1984)
- Adaptive estimation: Bickel (1982), Tsybakov (1998)
  - 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.
  - Armstrong and Kolesar (2018): Impossible to tighten minimax CI and maintain coverage for all b
- Computation: Chamberlain (2000); Elliott, Müller and Watson (2015); Müller and Wang (2019); Kline and Walters (2021)

Dobkin, Finkelstein, Kluender and Notowidigdo (2018)



- θ is effect of unexpected hospitalization on medical spending
- The researchers report Y_U, allowing a linear pre-trend
- Omitting trend yields a more precise (but less credible) Y_R

A minimax approach

If we know $|b| \leq B$ then reasonable to compute *B-minimax* estimator δ_B^* that minimizes worst case risk (Wald, 1950; Savage, 1954)

$$R_{\text{max}}(B, \delta) = \sup_{(\theta, b) \in \mathbb{R} \times [-B, B]} R(\theta, b, \delta)$$

where $R(\theta, b, \delta)$ gives MSE of an estimator δ.

A minimax approach

If we know $|b| \leq B$ then reasonable to compute *B-minimax* estimator δ_B^* that minimizes worst case risk (Wald, 1950; Savage, 1954)

$$R_{\text{max}}(B, \delta) = \sup_{(\theta, b) \in \mathbb{R} \times [-B, B]} R(\theta, b, \delta)$$

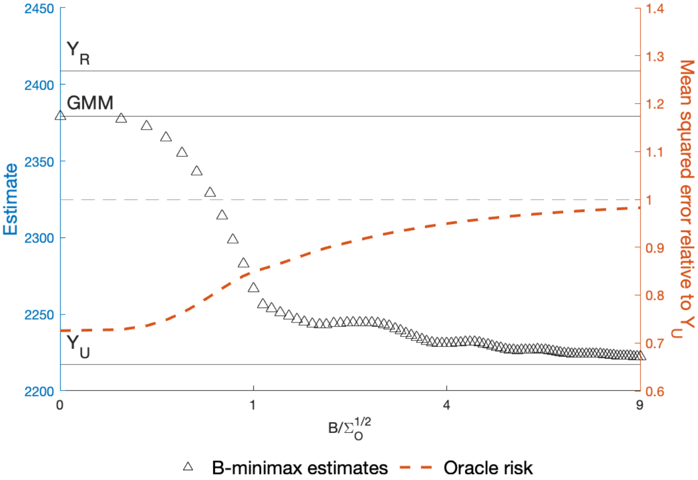
where $R(\theta, b, \delta)$ gives MSE of an estimator δ.

Choice of B trades off robustness against efficiency

- $\delta_0^* = Y_U - \Sigma_{UO}/\Sigma_O \cdot Y_O$ (GMM / "least robust")
- $\delta_\infty^* = Y_U$ ("most robust")

Sensitivity analysis: compute δ_B^* and $R_{\text{max}}(B, \delta_B^*)$ for a range of $B \in \mathcal{B}$

B-minimax estimates



Note: Oracle risk is $R_{\max}(B, \delta_B^*) \leq \Sigma_U$

Which B to choose?

An Oracle that knows a (true) bound B faces maximal risk

$$R^*(B) = R_{\max}(B, \delta_B^*)$$

Define the *adaptation regret* of any estimator δ as the proportional increase in worst-case risk over the Oracle

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

Which B to choose?

An Oracle that knows a (true) bound B faces maximal risk

$$R^*(B) = R_{\max}(B, \delta_B^*)$$

Define the *adaptation regret* of any estimator δ as the proportional increase in worst-case risk over the Oracle

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

Key idea: mimic the Oracle by minimizing *worst case* adaptation regret

$$A_{\max}(\mathcal{B}, \delta) = \sup_{B \in \mathcal{B}} A(B, \delta)$$

Resulting *adaptive* estimator gets as close as possible to the Oracle simultaneously for all $B \in \mathcal{B} = [0, \infty]$. i.e., it is uniformly *near*-minimax.

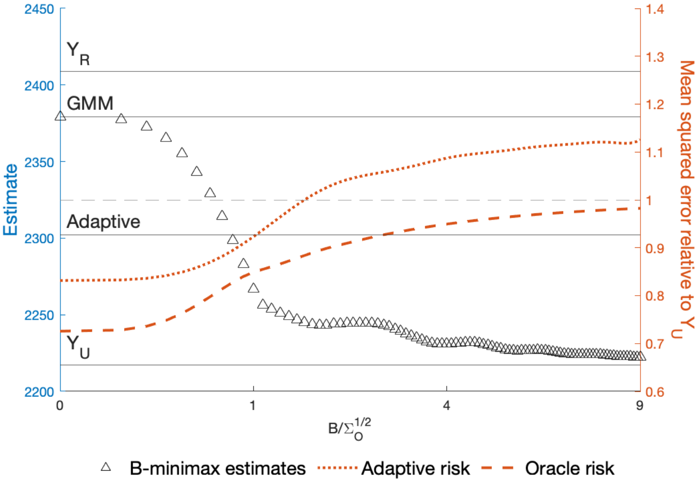
A scaled risk interpretation

For $\mathcal{B} = [0, \infty]$, the worst-case adaptation regret is equivalent to the worst-case *scaled* risk with scaling $R^*(|b|)$

$$A_{\max}(\mathcal{B}, \delta) = \sup_{B \in \mathcal{B}} \frac{\sup_{|b| \le B} R(\theta, b, \delta)}{R^*(B)} = \sup_{b \in \mathbb{R}} \frac{R(\theta, b, \delta)}{R^*(|b|)}$$

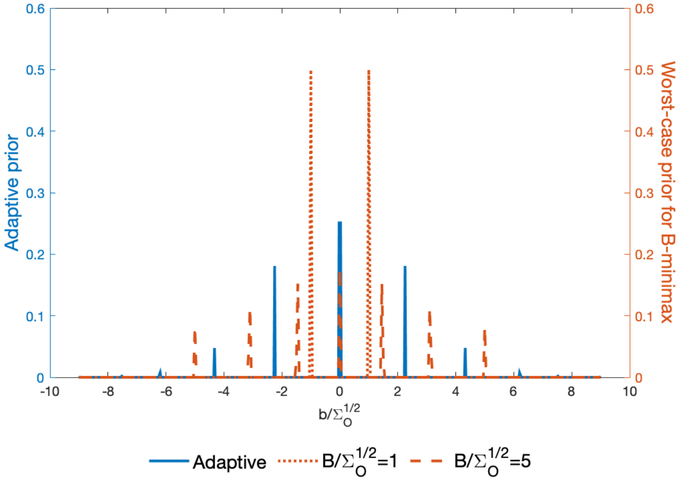
- Problem has been reduced to minimax on new objective.
- For $0 < \varepsilon \le \rho^2 \le 1 - \varepsilon < 1$ the minimax theorem still applies.
- Discretize $\mathbb{R}$ and solve for π using a convex optimization routine.
- Solution will exhibit constant adaptation regret at all points of support of least favorable prior.

Adaptive estimate



Note: worst case risk of adaptive estimator is bounded!

Least favorable priors over b



Adaptive prior works especially well when $b \approx 0$ and requires no tuning

Overview of results

Adaptive estimator takes the form:

$$\underbrace{\frac{\Sigma_{UO}}{\sqrt{\Sigma_O}}\delta(T_O)}_{\text{Bias estimator}} + \underbrace{Y_U - \Sigma_{UO}/\Sigma_O \cdot Y_O}_{\text{GMM estimator of } \theta}$$

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

Dobkin et al (2018) original estimates

- Omitting pre-trend lowers std errs by $14 - 30\%$
- Can't reject absence of trend ($T_O \approx 1.2$)

Yrs since hosp.	Y_U	Y_R	Y_O	ρ
0	2,217	2,409	192	-0.524
	(257)	(221)	(160)	
1	1,268	1,584	316	-0.703
	(337)	(241)	(263)	
2	989	1,436	447	-0.784
	(430)	(270)	(373)	
3	1,234	1,813	579	-0.813
	(530)	(313)	(482)	

Table: Impact of hospitalization on out of pocket (OOP) expenditures for the non-elderly insured (ages 50 to 59) in the HRS. Standard errors in parentheses. "Yrs since hosp." refers to years since hospitalization.

Dobkin et al (2018) adaptive estimates

Adaptive estimate roughly half way between trend and no trend models.

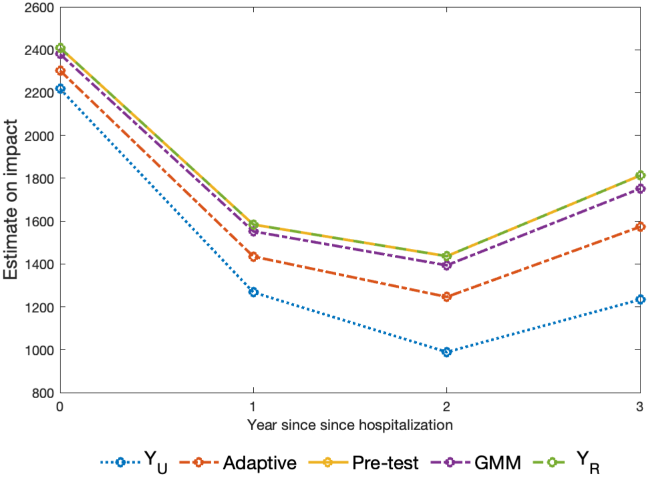


Figure: Estimates of the impact of hospitalization on OOP spending

Dobkin et al (2018) risk profiles

Adaptive yields (much) lower worst case risk than pre-test of $|T_O| > 1.96$

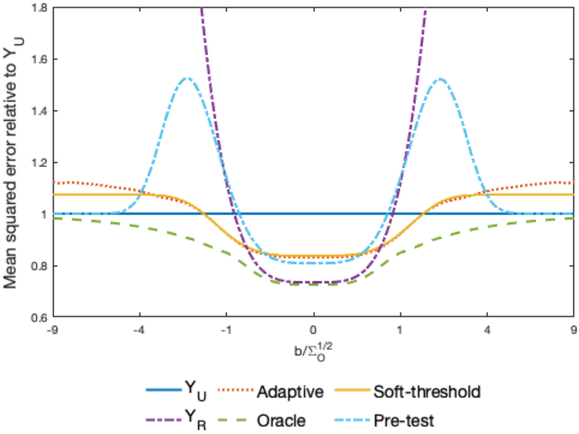


Figure: Risk functions for θ_0 ($\rho = -0.524$)

BLP estimates (as in Andrews, Gentzkow, Shapiro, 2017)

- Parameter of interest θ is the average price markup.
- Y_R is GMM estimate using demand and supply side instruments.
- Y_U is GMM using only the demand side instruments.
- If the demand side instruments are valid, the bias b is zero.

Y_U	Y_R	Y_O = Difference	ρ
52.95	33.53	-19.42	-0.7
(2.54)	(1.81)	(1.78)	

Table: Average markup (in percent)

Note: Y_R has much lower std errs but $T_O \approx -11$!

BLP adaptive estimates

- Huge T_O leads both adaptive and soft-threshold estimators to place nearly all weight on Y_U.
- Soft-threshold ($\lambda = .59$) is much lower than 1.96 used by pre-test but regret is also much lower.

	Y_U	Y_R	Adaptive	Soft-threshold	Pre-test
Estimate	52.95	33.53	49.44	51.89	52.95
Max Regret	96%	∞	32%	34%	107%
Threshold				0.59	1.96

Table: Adaptive estimates for the average markup (in percent). "Max Regret" refers to *worst-case* adaptation regret $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$.

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'Haultfoeuille (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

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?
- Today: estimate bias to refine effects of training
  - 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.
- Adapt over finite set of bounds $\mathcal{B} = \{(0,0),(\infty,0),(\infty,\infty)\}$ (assumes Y_{R2} less biased than Y_{R1})
- 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

Angrist and Krueger (1991) estimates

- Suppose parameter of interest θ is return to schooling (presumed constant)
- Take Y_U to be the Wald-IV estimate, and Y_R to be the OLS estimate
- When schooling is exogenous, the bias is zero.

Wald Y_U	OLS Y_R	Y_O = Difference	ρ
0.102	0.071	-0.0311	-0.9998
(0.0239)	(0.0003)	(0.0239)	

Table: Returns to schooling

Note: Y_R is orders of magnitude more precise than $Y_U \rightarrow$ huge adaptation regret.

Angrist and Krueger (1991): limiting max risk

- Unconstrained adaptive estimator puts nearly all weight on OLS
- Huge (5x) increase in max risk over IV
- When limited to $\sim$ 20% increase in max risk, shrinks half way to IV.

	Unconstrained	$\bar{R}/\Sigma_U \leq 1.5$	$\bar{R}/\Sigma_U \leq 1.2$
Estimate (fully nonlinear)	0.071	0.0794	0.0855
Maximum risk	5.55	1.51	1.23
Estimate (soft-threshold)	0.071	0.0836	0.0893
Threshold	2.07	0.7686	0.5283
Maximum risk	5.27	1.59	1.28

Table: Adaptive estimates of returns to schooling with bounds on minimax risk. Maximum risk is reported relative to Σ_U.

Angrist and Krueger (1991) risk profile

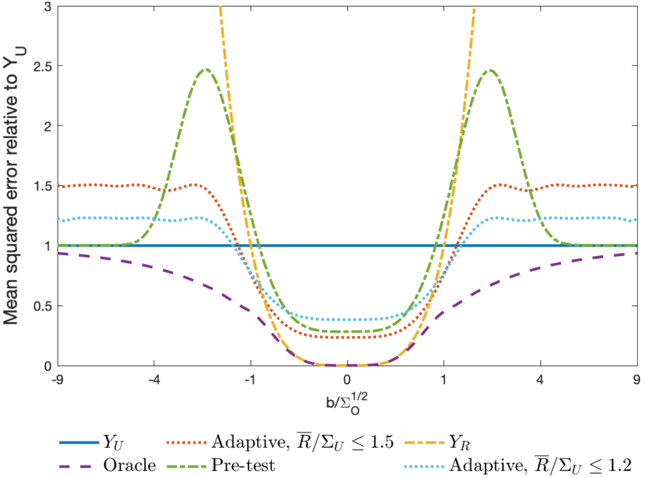


Figure: Risk profiles ($\rho = -0.9998$)

Conclusion

- Economists love models. But models are never quite right.
- Consequently, we are asked to report specification tests.
- Adaptive estimator uses a specification test to refine 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`
- Ongoing work: adaptive binary decisions.

B-minimax estimator

Claim.

The *B-minimax* estimator δ_B^* takes the form:

$$\frac{\Sigma_{UO}}{\sqrt{\Sigma_O}} \underbrace{\delta^{\text{BNM}}(T_O)}_{\text{Scaled bias estimator}} + \underbrace{Y_U - \Sigma_{UO}/\Sigma_O \cdot Y_O}_{\text{GMM estimator of } \theta},$$

where $\delta^{\text{BNM}}(T_O)$ solves

$$\inf_{\delta} \sup_{|\tilde{b}| \le B/\sqrt{\Sigma_O}} E_{T_O \sim N(\tilde{b},1)} \left[\left(\delta(T_O) - \tilde{b} \right)^2 \right].$$

In other words, $\delta^{\text{BNM}}(T_O)$ is (MSE) minimax for estimating $\tilde{b} = b/\sqrt{\Sigma_O}$ in the parameter space $|\tilde{b}| \le B/\sqrt{\Sigma_O}$. Proof

Computation

Compute δ^{BNM} by solving for least favorable prior ala Chamberlain (2000)

- The *Bayes risk* of a decision $\delta()$ under prior π on b is

$$R_{\text{Bayes}}(\pi, \delta) = \int R(b, \delta)\, d\pi(b) = \int \int L(b, \delta(y))\, dP_b(y)\, d\pi(b).$$

- Let Γ be the set of priors supported on a set $[-B, B]$. By the minimax theorem, we have

$$\min_{\delta} \max_{b \in [-B,B]} R(b, \delta) = \min_{\delta} \max_{\pi \in \Gamma} R_{\text{Bayes}}(\pi, \delta) = \max_{\pi \in \Gamma} \min_{\delta} R_{\text{Bayes}}(\pi, \delta).$$

- The inner minimization is solved by the *Bayes decision* δ_π. Under squared loss, δ_π given by posterior mean.
- Outer problem solved by discretizing prior and convex optimizer.

Two extensions

1. Adaptive estimator is complex. A simpler soft-thresholding estimator

$$\delta(T_O) = \mathbf{1}\{T_O > \lambda\}(T_O - \lambda) + \mathbf{1}\{T_O < -\lambda\}(T_O + \lambda)$$

achieves comparable risk performance when λ is chosen to minimize the worst-case adaptation regret.

Two extensions

1. Adaptive estimator is complex. A simpler soft-thresholding estimator

$$\delta\left(T_O\right) = \mathbf{1}\left\{T_O > \lambda\right\}\left(T_O - \lambda\right) + \mathbf{1}\left\{T_O < -\lambda\right\}\left(T_O + \lambda\right)$$

achieves comparable risk performance when λ is chosen to minimize the worst-case adaptation regret.

2. As Σ_R gets smaller, worst-case adaptation regret grows. Possible to bound the increase in minimax risk by solving the constrained problem

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

Alternative (not for today) minimize *additive* notion of worst-case adaptation regret: $\sup_{B \in \mathcal{B}} R_{\max}(B, \delta) - R^*(B)$

Three estimators of bias

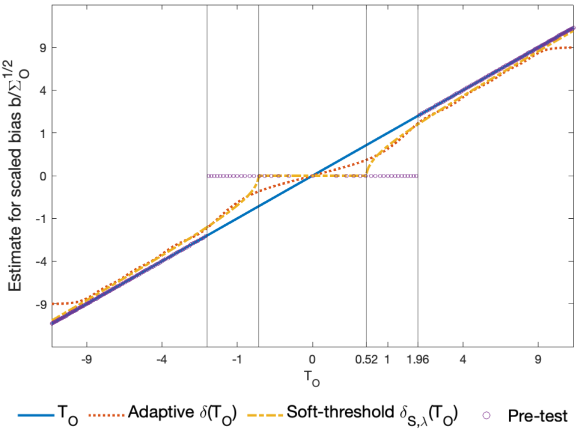


Figure: Estimators of scaled bias when $\rho = -0.524$

Proof sketches: using invariance

- We consider squared loss function $L(\theta, d) = (\theta - d)^2$
- Estimation for θ is (location) invariant because
  - $L(\theta + t, d + t) = L(\theta, d)$
  - for $(\theta, b) \mapsto (\theta + t, b)$, the same transformation on the data $(Y_U, Y_O) \mapsto (Y_U + t, Y_O)$ leads to the same transformation of the distribution $P_{\theta,b}$
- Hunt-Stein Theorem implies we can search for minimax rules among equivariant estimators, which in our setting takes the form $\delta(Y_U, Y_O) = \tilde{\delta}(Y_O) + Y_U$
- Note that we can orthogonalize

$$Y_U - \theta = \frac{\Sigma_{UO}}{\Sigma_O}(Y_O - b) + V$$

where $V \sim N\left(0, \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O}\right)$ is independent of Y_O return

Proof sketches: further simplification

- Thus,

$$
\begin{aligned}
&E_{\theta,b}\left[L\left(\theta, b, \tilde{\delta}(Y_O)+Y_U\right)\right] \\
&= E_{\theta,b}\left[L\left(0, b, \tilde{\delta}(Y_O)+\frac{\Sigma_{UO}}{\Sigma_O}(Y_O-b)+V\right)\right]
\end{aligned}
$$

- The risk function does not depend on θ anymore, so we can evaluate the minimax problem as

$$
\inf_{\tilde{\delta}} \sup_{|b|\le B} R(0,b,\tilde{\delta}) = \inf_{\tilde{\delta}} \sup_{|b|\le B} E_{0,b}L(0,\tilde{\delta}(Y_O)+\frac{\Sigma_{UO}}{\Sigma_O}(Y_O-b)+V)
$$

- Let $\tilde{L}(b,d) = EL(0,d+V)$ and $\bar{\delta}(y_O) = \tilde{\delta}(y_O)+\frac{\Sigma_{UO}}{\Sigma_O}y_O$, we can write

$$
E_{0,b}\left[\tilde{L}\left(b, \bar{\delta}(Y_O)-\frac{\Sigma_{UO}}{\Sigma_O}b\right)\right]
$$

- The modified loss function $\tilde{L}(b,d) = d^2+\Sigma_U-\frac{\Sigma_{UO}^2}{\Sigma_O}$ is a shifted squared error loss

Proof sketches: modified minimax problem

- It follows that the estimator that solves the original minimax problem

$$\inf_{\delta} \sup_{|b| \leq B} E_{\theta,b} L(\theta, \delta(Y_U, Y_O))$$

is given by

$$\tilde{\delta}^*(Y_O) + Y_U = \bar{\delta}^*(Y_O) + Y_U - \frac{\Sigma_{UO}}{\Sigma_O} Y_O$$

where $\bar{\delta}^*(Y_O)$ solves

$$\inf_{\delta} \sup_{|b| \leq B} E_{0,b} \left[\tilde{L} \left(b, \bar{\delta}(Y_O) - \frac{\Sigma_{UO}}{\Sigma_O} b \right) \right].$$

Proof sketches: reparameterization

- It follows that the estimator that solves the original minimax problem is given by

$$\bar{\delta}^*(Y_O) + Y_U - \frac{\Sigma_{UO}}{\Sigma_O} Y_O$$

- Applying a reparameterization $\bar{\delta}(Y_O) = \frac{\Sigma_{UO}}{\sqrt{\Sigma_O}} \bar{\delta}(Y_O/\sqrt{\Sigma_O}) = \frac{\Sigma_{UO}}{\sqrt{\Sigma_O}} \bar{\delta}(T_O)$.
- $\bar{\delta}^*(T_O)$ solves

$$\inf_{\delta} \sup_{|b| \leq B} E_{0,b} \frac{\Sigma_{UO}^2}{\Sigma_O} E_{0,b} \left(\delta(T_O) - \frac{b}{\sqrt{\Sigma_O}} \right)^2 + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O}$$

$$\inf_{\delta} \sup_{|\tilde{b}| \leq B/\sqrt{\Sigma_O}} \frac{\Sigma_{UO}^2}{\Sigma_O} \left(E_{T_O \sim N(\tilde{b},1)} \left(\delta(T_O) - \tilde{b} \right)^2 + 1/\rho^2 - 1 \right)$$

- In other words, $\bar{\delta}^*(T_O)$ is minimax for estimating $\tilde{b}$, the mean of a normal r.v. with s.d. 1, in the parameter space $[-B/\sqrt{\Sigma_O}, B/\sqrt{\Sigma_O}]$