# Online Appendix to "Adapting to Misspecification"

Timothy B. Armstrong, Patrick Kline and Liyang Sun

June 27, 2023

## Appendix B Group decision making interpretation

This appendix develops a simple model of group decision making inspired by Savage (1954)'s arguments regarding the ability of minimax decisions to foster consensus among individuals with heterogeneous beliefs. Extending these arguments, we illustrate how adaptive decisions can serve to foster consensus across groups of individuals with different sets of beliefs.

### B.1 Consensus in a single committee

Suppose there is a committee comprised of members with heterogeneous beliefs that include all priors supported on the set $\mathcal{C}_B$. The committee chair, who we will call the *$B$-chair*, offers a take it or leave it proposal that her committee follow a decision rule $\delta$ in exchange for the provision of a public good providing payoff $G$ to each member of the committee. This public good might consist of a persuasive speech, a reduction in committee work, or an offer to end the meeting early.

If the committee agrees to the proposal, the $B$-chair earns a payoff $K - C(G)$, where $K$ is the value of consensus and $C(\cdot)$ is an increasing cost function. If some member of the committee does not agree to the proposal, the chair and all committee members receive payoff zero. The $B$-chair therefore seeks a rule $\delta$ allowing payment of the smallest $G$ that ensures consensus.

A committee member who is certain of the parameters $(\theta, b)$ will accept the chair's offer if and only if $R(\theta, b, \delta) \leq G$. However, the committee member with the most pessimistic beliefs regarding these parameters will require a public goods provision level of at least $R_{\max}(B, \delta)$ to agree to the offer. To achieve consensus at minimal cost, the $B$-chair can propose the $B$-minimax decision, which requires public goods provision level $R^*(B)$ to achieve consensus.

The $B$-chair will be willing to provide this level of public goods if and only if $K \geq C(R^*(B))$, in which case consensus ensues. If this condition does not hold, the chair deems the $B$-minimax decision too costly to implement and consensus is not achieved. Hence, when no individual holds beliefs that are too extreme, the minimax decision fosters consensus.

## B.2 Consensus among committees

Now suppose there is a collection $\mathcal{B}$ of committees that is led by a *chair of chairs* (CoC) who would like for the $B$-chairs to agree on a common decision making rule $\delta$. Suppose also that $K > \sup_{B \in \mathcal{B}} C(R^*(B))$, so that each $B$-chair would privately prefer to implement the $B$-minimax decision. The CoC has a fixed budget that can be used to persuade the chairs to instead coordinate on a common rule $\delta$.

By the arguments above, each $B$-chair must pay a cost $C(R_{max}(B, \delta))$ to secure consensus regarding the CoC's proposed plan $\delta$, leaving her with payoff $K - C(R_{max}(B, \delta))$. However, each chair can also defy the CoC and propose the $B$-minimax decision to her committee, yielding payoff $K - C(R^*(B))$. Hence, to compel a $B$-chair to propose a decision $\delta$, the CoC must offer a transfer of at least $\Delta_B = C(R_{max}(B, \delta)) - C(R^*(B))$. To economize on transfer costs, the CoC searches for a $\delta$ that minimizes the maximal required payment $\sup_{B \in \mathcal{B}} \Delta_B$ across all committees.

Different functional forms for the cost function $C$ yield different notions of adaptation. To motivate the formulation in (1), we assume $C(G) = \ln G$, which suggests chairs produce the public good according to an increasing returns to scale technology that is exponential in effort costs. With this choice of $C(\cdot)$, the CoC's problem is to find a $\delta$ that minimizes $\sup_{B \in \mathcal{B}} \ln(R_{max}(B, \delta)/R^*(B)) = \sup_{B \in \mathcal{B}} \ln A(B, \delta)$. The CoC will therefore propose the optimally adaptive decision $\delta^{\text{adapt}}$, which yields $\sup_{B \in \mathcal{B}} \Delta_B = \ln A^*(\mathcal{B})$. When $A^*(\mathcal{B})$ is too large, the CoC balks at the cost and consensus fails.

## B.3 Discussion

Taking the committees to represent different camps of researchers, our stylized model suggests adaptive estimation can help to forge consensus between researchers with varying beliefs about the suitability of different econometric models. The prospects for achieving consensus are governed by the loss of efficiency under adaptation. When $A^*(\mathcal{B})$ is small, consensus

is likely, as the adaptive decision will yield maximal risk similar to each camp's perceived $B$-minimax risk. When $A^*(\mathcal{B})$ is large, however, consensus is unlikely to emerge, as the optimally adaptive estimator will be perceived as excessively risky by camps with extreme beliefs.

# Appendix C Additional details

## C.1 Numerical results on estimators as a function of $\rho^2$

Section 4.4 introduces the class of soft thresholding estimators and hard thresholding estimators. In Figure A1, we plot the solution to the nearly adaptive objective function for soft-thresholding, which corresponds to a threshold that increases with $\rho^2$. As $\rho^2$ increases, to minimize the worst-case adaptation regret, more weight needs to be placed on the optimal GMM estimator, which explains the increase in the adaptive threshold. Correspondingly, the adaptive estimator incurs more bias as $\rho^2$ increase, which narrows the range of true bias for which the adaptive estimator beats $Y_U$ in terms of risk.

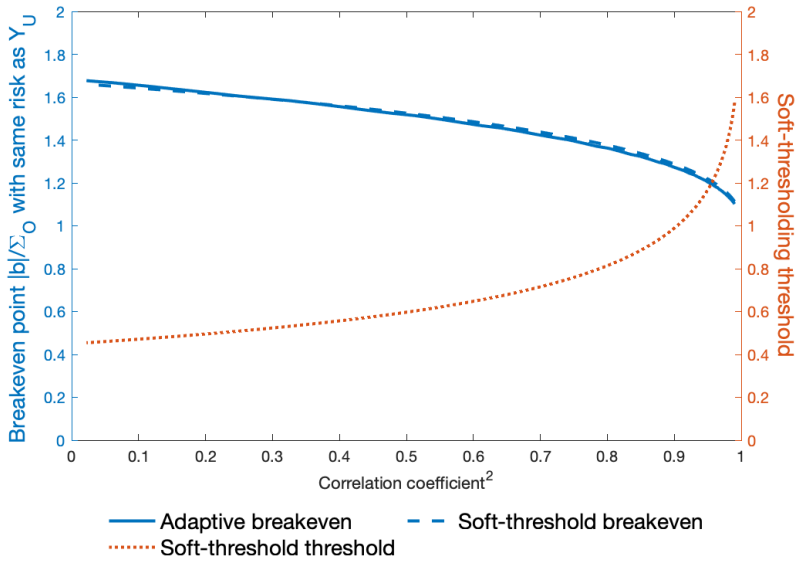


Figure A1: Threshold for adaptive soft-thresholding estimator

In practice, it is common to use a fixed threshold of 1.96, which corresponds to a pre-test rule that switches between the unrestricted estimator and the GMM estimator based on the result of the specification test. Doing so leads to high level of worst-case adaptation

regret especially when $\rho^2$ is close to one as shown in Figure A2. To minimize the worst-case adaptation regret, the adaptive hard-threshold estimator needs to use a threshold that would increase to infinity as $\rho^2$ gets closer to one.

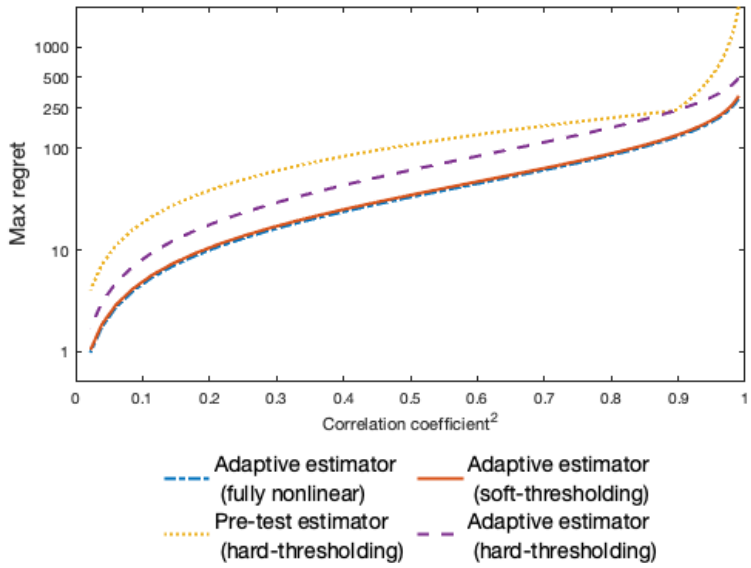


Figure A2: "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$.

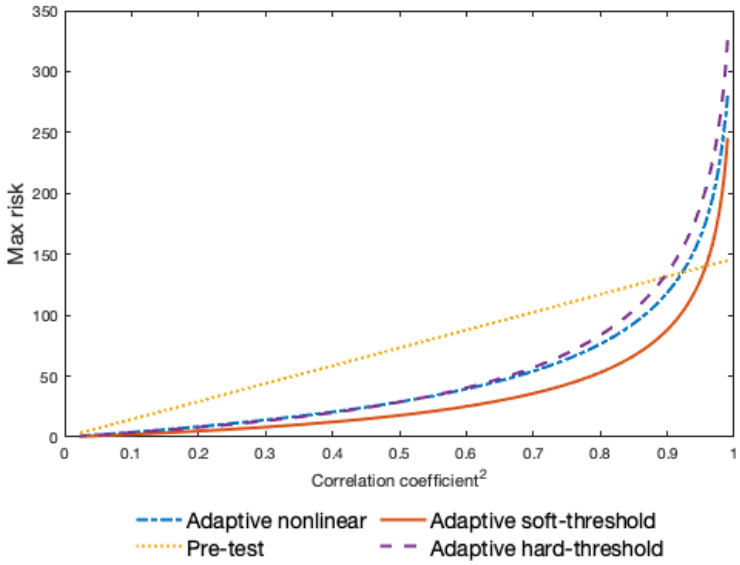


Figure A3: "Max risk" refers to the worst case risk increase relative to $Y_U$ in percentage terms $(R_{\max}(\delta) - \Sigma_U)/\Sigma_U \times 100$.

A pre-test estimator utilizing a fixed threshold at 1.96 realizes its worst-case risk when the scaled bias $\tilde{b}$ is itself near the 1.96 threshold. As shown in Figure A3, the pre-test

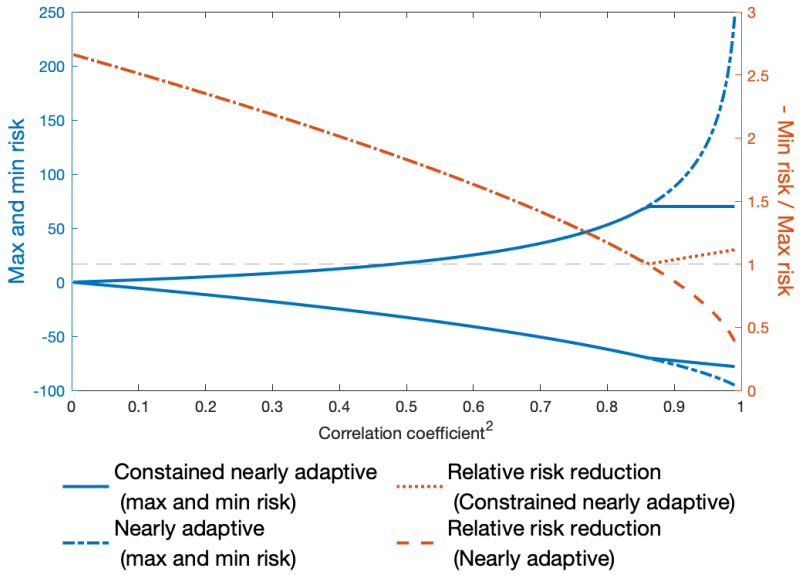


Figure A4: "Max risk" refers to the worst case risk increase relative to $Y_U$ in percentage terms $(R_{\max}(\infty, \delta) - \Sigma_U)/\Sigma_U \times 100$. "Min risk" refers to the best case risk decrease relative to $Y_U$ in percentage terms $(\min_b R(\theta, b, \delta) - \Sigma_U)/\Sigma_U \times 100$. The calculations are based on the soft thresholding nearly adaptive estimator. The constrained variant bounds the worst-case risk to be less than 70% above the risk of $Y_U$.

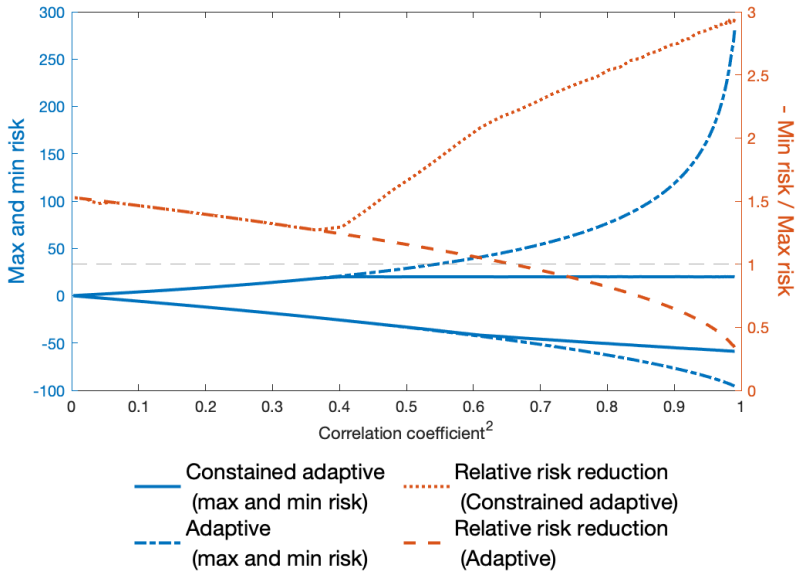


Figure A5: "Max risk" refers to the worst case risk increase relative to $Y_U$ in percentage terms $(R_{\max}(\infty, \delta) - \Sigma_U)/\Sigma_U \times 100$. "Min risk" refers to the best case risk decrease relative to $Y_U$ in percentage terms $(\min_b R(\theta, b, \delta) - \Sigma_U)/\Sigma_U \times 100$. The calculations are based on the optimally adaptive estimator. The constrained variant bounds the worst-case risk to be less than 20% above the risk of $Y_U$.

estimator tends to exhibit substantially greater worst-case risk than the class of adaptive estimators for most values of $\rho^2$. As discussed in Section 4.4, adaptive estimators have large worst-case risk when $\rho^2$ is close to one. The pre-test estimator has lower worst-case risk in these cases, due to the fixed threshold at 1.96. However, one can achieve the same worst-case risk while achieving a much lower worst-case adaptation regret by constraining the worst-case risk directly as in Section 4.5. For example, Figure A4 shows that for the constrained soft-thresholding version of the adaptive estimator, even as we constrain the worst-case risk to be less than 70% above the risk of $Y_U$, the best-case decrease in risk relative to $Y_U$ is still greater than the worst-case increase in risk over $Y_U$. Figure A5 shows that this property holds for the unconstrained optimally adaptive estimator so long as $\rho^2 \leq 0.65$ and also when the optimally adaptive estimator is constrained to exhibit risk no greater than 120% of the risk of $Y_U$.

## C.2 Asymptotics as $|\rho| \to 1$

This section considers the behavior of the worst-case adaptation regret as $|\rho| \to 1$ for the optimally adaptive estimator as well as for the hard and soft-thresholding estimators. Let $A(\delta, \rho)$ denote the worst-case adaptation regret of the estimator given by (4) under the given value of $\rho$, so that $A(\delta, \rho)$ returns the value of (6) with $\tilde{\delta} = \delta$. We use $A^*(\rho) = \inf_\delta A(\delta, \rho)$ (where the infimum is over all estimators) to denote the loss of efficiency under adaptation for the given value of $\rho$. Likewise, we denote by $A_S(\lambda, \rho) = A(\delta_{S,\lambda}, \rho)$ and $A_H(\lambda, \rho) = A(\delta_{H,\lambda}, \rho)$ the worst-case adaptation regret for soft and hard-thresholding respectively with threshold $\lambda$, where $\delta_{S,\lambda}$ are $\delta_{H,\lambda}$ are defined in Section 4.4. Finally, we use $A_S^*(\rho) = \inf_\lambda A_S(\lambda, \rho)$ and $A_H^*(\rho) = \inf_\lambda A_H(\lambda, \rho)$ to denote the minimum worst-case adaptation regret for soft and hard-thresholding respectively.

To get some intuition for the interpretation of $\rho$ close to 1, consider the Hausman setting where $Y_R$ is efficient under the restriction $b = 0$. In this case, we have $\text{var}(Y_R) = \text{cov}(Y_R, Y_U)$, $\text{cov}(Y_O, Y_U) = \text{cov}(Y_R - Y_U, Y_U) = \text{var}(Y_R) - \text{var}(Y_U)$ and $\text{var}(Y_O) = \text{var}(Y_R) + \text{var}(Y_U) - 2\,\text{cov}(Y_R, Y_U) = \text{var}(Y_U) - \text{var}(Y_R)$. It follows that

$$\rho^2 = \frac{\text{cov}(Y_O, Y_U)^2}{\text{var}(Y_U)\,\text{var}(Y_O)} = \frac{\text{var}(Y_U) - \text{var}(Y_R)}{\text{var}(Y_U)}$$

and

$$\rho^{-2} - 1 = \frac{\mathrm{var}(Y_U)}{\mathrm{var}(Y_U) - \mathrm{var}(Y_R)} - 1 = \frac{\mathrm{var}(Y_R)}{\mathrm{var}(Y_U) - \mathrm{var}(Y_R)} = \frac{\mathrm{var}(Y_R)/\mathrm{var}(Y_U)}{1 - \mathrm{var}(Y_R)/\mathrm{var}(Y_U)}.$$

Therefore, $|\rho| \to 1$ corresponds to the case where $\mathrm{var}(Y_R)/\mathrm{var}(Y_U) \to 0$. Furthermore, $\rho^{-2} - 1 = \frac{\mathrm{var}(Y_R)}{\mathrm{var}(Y_U)}(1 + o(1))$ as $|\rho| \to 1$, revealing that this quantity captures the relative efficiency of the restricted estimator under proper specification.

The following theorem characterizes the behavior of $A^*(\rho)$, $A^*_S(\rho)$ and $A^*_H(\rho)$ as $|\rho| \to 1$.

**Theorem C.1.** *We have*

$$\lim_{|\rho|\uparrow 1} \frac{A^*(\rho)}{2\log(\rho^{-2} - 1)^{-1}} = \lim_{|\rho|\uparrow 1} \frac{A^*_S(\rho)}{2\log(\rho^{-2} - 1)^{-1}} = \lim_{|\rho|\uparrow 1} \frac{A^*_H(\rho)}{2\log(\rho^{-2} - 1)^{-1}} = 1.$$

In the remainder of this section, we prove Theorem C.1. We split the proof into upper bounds (Section C.2.1) and lower bounds (Section C.2.2). The lower bounds in Section C.2.2 are essentially immediate from results in Bickel (1983) for adapting to $B \in \mathcal{B} = \{0, \infty\}$, whereas the upper bounds in Section C.2.1 involve new arguments to deal with intermediate values of $B$.

### C.2.1 Upper bounds

In this section, we show that $A^*_S(\rho) \leq (1 + o(1))2\log(\rho^{-2} - 1)^{-1}$ and $A^*_H(\rho) \leq (1 + o(1))2\log(\rho^{-2} - 1)^{-1}$. Since $A^*(\rho)$ is bounded from above by both $A^*_S(\rho)$ and $A^*_H(\rho)$, this also implies $A^*(\rho) \leq (1 + o(1))2\log(\rho^{-2} - 1)^{-1}$.

Let $r_S(\lambda, t) = E_{T\sim N(\mu,1)}(\delta_{S,\lambda}(T) - \mu)^2$ and $r_S(\lambda, t) = E_{T\sim N(\mu,1)}(\delta_{H,\lambda}(T) - \mu)^2$ denote the risk of soft and hard-thresholding. Then

$$A_S(\lambda, \rho) = \sup_{\mu\in\mathbb{R}} \frac{r_S(\lambda, \mu) + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\mu|) + \rho^{-2} - 1}$$

and similarly for $A_H(\lambda, \rho)$. We use the following upper bound for $r_H(\lambda, \mu)$ and $r_S(\lambda, \mu)$, which follows immediately from results given in Johnstone (2019).

**Lemma C.1.** *There exists a constant $C$ such that, for $\lambda > C$, both $r_S(\lambda, \mu)$ and $r_H(\lambda, \mu)$*

*are bounded from above by* $\bar{r}(\lambda, \mu)$ *where*

$$\bar{r}(\lambda, \mu) = \begin{cases} \min\left\{\lambda \exp\left(-\lambda^2/2\right) + 1.2\mu^2, 1 + \mu^2\right\} & |\mu| \leq \lambda \\ 1 + \lambda^2 & |\mu| > \lambda. \end{cases}$$

*Proof.* The bound for $r_H(\lambda, \mu)$ follows from Lemma 8.5 in Johnstone (2019) along with the bound $r_H(\lambda, 0) \leq \frac{2+\varepsilon}{\sqrt{2\pi}} \lambda \exp\left(-\lambda^2/2\right)$ which holds for any $\varepsilon > 0$ for $\lambda$ large enough by (8.15) in Johnstone (2019). The bound for $r_L(\lambda, \mu)$ follows from Lemma 8.3 and (8.7) in Johnstone (2019). □

Let $\tilde{\lambda}_\rho = \sqrt{2\log(\rho^{-2} - 1)^{-1}}$. By Lemma C.1, $A_S^*(\rho)$ and $A_H^*(\rho)$ are, for $(\rho^{-2} - 1)^{-1}$ large enough, bounded from above by the supremum over $\mu$ of

$$\frac{\bar{r}(\tilde{\lambda}_\rho, \mu) + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\mu|) + \rho^{-2} - 1} \tag{18}$$

Let $c(\rho)$ be such that $c(\rho)/\tilde{\lambda}_\rho \to 0$ and $c(\rho) \to \infty$ as $|\rho| \uparrow 1$. We bound (18) separately for $|\mu| \leq c(\rho)$ and for $|\mu| \geq c(\rho)$. For $|\mu| \leq c(\rho)$, we use the bound $r^{\mathrm{BNM}}(|\mu|) \geq .8 \cdot \mu^2/(\mu^2 + 1)$ (Donoho, 1994), which gives an upper bound for (18) of

$$\begin{aligned} &\frac{\bar{r}(\tilde{\lambda}_\rho, \mu) + \rho^{-2} - 1}{.8 \cdot \mu^2/(\mu^2 + 1) + \rho^{-2} - 1} \leq \frac{\sqrt{2\log(\rho^{-2} - 1)^{-1}} \cdot (\rho^{-2} - 1) + 1.2\mu^2 + \rho^{-2} - 1}{.8 \cdot \mu^2/(\mu^2 + 1) + \rho^{-2} - 1} \\ &\leq \sqrt{2\log(\rho^{-2} - 1)^{-1}} + (1.2/.8) \cdot (\mu^2 + 1) + 1 \leq \sqrt{2\log(\rho^{-2} - 1)^{-1}} + (1.2/.8) \cdot (c(\rho)^2 + 1) + 1. \end{aligned}$$

As $|\rho| \uparrow 1$, this increases more slowly than $\log(\rho^{-2} - 1)^{-1}$. For $|\mu| \geq c(\rho)$, we use the bound $r^{\mathrm{BNM}}(|\mu|) \geq r^{\mathrm{BNM}}(c(\rho))$ which gives an upper bound for (18) of

$$\frac{\bar{r}(\tilde{\lambda}_\rho, \mu) + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|c(\rho)|) + \rho^{-2} - 1} \leq \frac{\bar{r}(\tilde{\lambda}_\rho, \mu)}{r^{\mathrm{BNM}}(|c(\rho)|)} + 1 \leq \frac{1 + \tilde{\lambda}_\rho^2}{r^{\mathrm{BNM}}(|c(\rho)|)} + 1.$$

As $|\rho| \uparrow 1$, $c(\rho) \to \infty$ and $r^{\mathrm{BNM}}(|c(\rho)|) \to 1$, so that the above display is equal to a $1 + o(1)$ term times $\tilde{\lambda}_\rho^2 = 2\log(\rho^{-2} - 1)^{-1}$ as required.

### C.2.2 Lower bounds

In this section, we show that $A^*(\rho) \geq (1+o(1))2\log(\rho^{-2}-1)^{-1}$. Since $A_S^*(\rho)$ and $A_H^*(\rho)$ are bounded from below by $A^*(\rho)$ , this also implies $A_S^*(\rho) \geq (1+o(1))2\log(\rho^{-2}-1)^{-1}$ and $A_H^*(\rho) \geq (1+o(1))2\log(\rho^{-2}-1)^{-1}$.

Given an estimator $\delta(Y)$ of $\mu$ in the normal means problem $Y \sim N(\mu,1)$, let $m(\delta) = E_{T\sim N(0,1)}\delta(Y)^2$ denote the risk at $\mu=0$ and let $M(\delta) = \sup_{\mu\in\mathbb{R}} E_{T\sim N(\mu,1)}(\delta(Y)-\mu)^2$ denote worst-case risk. The following lemma is immediate from Bickel (1983, Theorem 4.1).

**Lemma C.2** (Bickel 1983, Theorem 4.1)**.** *For $t \in (0,1]$, let $\delta_t$ be an estimator that satisfies $m(\delta_t) \leq 1-t$. Then, as $t \uparrow 1$, $M(\delta_t) \geq (1+o(1))\cdot 2\log(1-t)$.*

Using this result, we prove the following lemma, which gives a lower bound for the worst-case adaptation regret and the worst-case risk of any estimator achieving the upper bound in Section C.2.1. The required lower bound $A^*(\rho) \geq (1+o(1))2\log(\rho^{-2}-1)^{-1}$ follows from this result.

**Lemma C.3.** *For $\rho \in (-1,1)$, let $\delta_\rho : \mathbb{R} \to \mathbb{R}$ be an estimator of $\mu$ in the normal means problem $Y \sim N(\mu,1)$. Suppose that the worst-case adaptation regret $A(\delta_\rho,\rho)$ of the corresponding estimator (4) satisfies $A(\delta_\rho,\rho) \leq (1+o(1))2\log(\rho^{-2}-1)^{-1}$ as $|\rho| \to 1$. Then the following results hold as $|\rho| \to 1$.*

*i.) The worst-case risk of the corresponding estimator (4) is bounded from below by a $1+o(1)$ term times $2\Sigma_U \log(\rho^{-2}-1)^{-1}$*

*ii.) $A(\delta_\rho,\rho) \geq (1+o(1))\cdot 2\log(\rho^{-2}-1)^{-1}$.*

*Proof.* By the arguments Section A.1, the worst-case risk of the estimator (4) with $\delta = \delta_\rho$ is given by $\Sigma_U \cdot \left[\rho^2 \sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T)-\mu)^2 + 1 - \rho^2\right]$. As $|\rho| \uparrow 1$, this is bounded from below by a $1+o(1)$ term times $\Sigma_U \sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T)-\mu)^2$. Similarly, $A(\delta_\rho,\rho)$ is bounded from below by a $1+o(1)$ term times $\sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T)-\mu)^2$ as $|\rho| \uparrow 1$. Thus, it suffices to show that $\sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T)-\mu)^2 \geq (1+o(1))\cdot 2\log(\rho^{-2}-1)^{-1}$.

To show this, note that it follows from plugging in $\tilde{b}=0$ to the objective in (6) that, for any $\varepsilon > 0$, we have, for $|\rho|$ close enough to 1,

$$\frac{E_{T\sim N(0,1)}\delta_\rho(T)^2}{\rho^{-2}-1} \leq A(\delta_\rho,\rho) \leq (2+\varepsilon)\log(\rho^{-2}-1)^{-1}.$$

Applying Lemma C.2 with $1-t=(\rho^{-2}-1)\cdot(2+\varepsilon)\log(\rho^{-2}-1)^{-1}$, it follows that

$$\sup_{\mu} E_{T\sim N(\mu,1)}(\delta_\rho(T)-\mu)^2 \geq (1+o(1))\cdot 2\log\left[(\rho^{-2}-1)\cdot(2+\varepsilon)\log(\rho^{-2}-1)^{-1}\right]$$
$$=(1+o(1))\cdot\left[2\log(\rho^{-2}-1)+\log(2+\varepsilon)+\log\log(\rho^{-2}-1)^{-1}\right]=(1+o(1))\cdot 2\log(\rho^{-2}-1)$$

as required.

□

# Appendix D Computational details

In this section, we provide additional details on our computation of the adaptive estimator.

## D.1 Discrete approximation to estimators and risk function

Operationally, discretizing the support of the random variable $T\in\mathcal{T}$ into $K$ points, finding an estimator $\delta(T)$ is equivalent to finding a "policy" function $\delta(t):\mathcal{T}\to\mathbb{R}$:

$$\delta(t)=\sum_{k=1}^{K}\psi_k 1\{t=t_k\}.$$

Hence, we can rewrite the risk of estimator $\delta(T)$ when $T\sim N(b,1)$ as

$$E_{T\sim N(b,1)}\left(\sum_{k=1}^{K}\psi_k 1\{T=t_k\}-b\right)^2. \tag{19}$$

Define $\mu_{kb}=\Pr_{T\sim N(b,1)}(T=t_k)$ as the probability of falling into the $k$'th grid point given bias $b$, which can be evaluated analytically via the following discrete approximation to the normal distribution

$$\mu_{kb}=\Phi\left((t_k+t_{k+1})/2-b\right)-\Phi\left((t_k+t_{k-1})/2-b\right), \tag{20}$$

where we define $t_0=-\infty$ and $t_{K+1}=\infty$, which ensures that $\sum_{k=1}^{K}\mu_{kb}=1$. The discretized

approximation to the risk function (19) is therefore

$$\sum_{k=1}^{K} \psi_k^2 \mu_{kb} - 2b \sum_{k=1}^{K} \psi_k \mu_{kb} + b^2. \tag{21}$$

## D.2 Computing minimax risk in the bounded normal mean problem

We now provide details on how to compute the minimax risk $r^{\text{BNM}}(|\tilde{b}|)$ in the bounded normal mean problem, which allows us to easily compute the $B$-minimax risk for the main example as described in 5 for each $B \in \mathcal{B}$. This subsection is a specialized version of the first step of Algorithm 4.1.

By definition, the minimax risk $r^{\text{BNM}}(|\tilde{b}|)$ is the minimized value of the following minimax problem

$$\min_{\delta} \max_{b \in [-|\tilde{b}|, |\tilde{b}|]} E_{T \sim N(b,1)} (\delta(Y) - b)^2$$

whose solution is the minimax estimator $\delta^{\text{BNM}}\left(T; |\tilde{b}|\right)$. In particular, for each $|\tilde{b}| = B/\sqrt{\Sigma_O} \in \{0.1, 0.2, \ldots, 9\}$ we calculate the minimax risk $r^{\text{BNM}}(|\tilde{b}|)$ following the steps below. To compute the minimax risk function $r^{\text{BNM}}(|\tilde{b}|)$ for values of $|\tilde{b}|$ that are not included in the fine grid, we rely on spline interpolation.

1. Approximate the prior $\pi$ with the finite dimensional vector $\pi \in \Delta^J$, where the parameter space $[-|\tilde{b}|, |\tilde{b}|]$ is approximated by an equally spaced grid of $b$ values spanning $[-|\tilde{b}|, |\tilde{b}|]$ with a step size of 0.05, totaling to $J$ grid values. Approximate the conditional risk function as in (21), where the support for $T \sim N(b, 1)$ is approximated by an equally spaced grid of $t$ values spanning $[-|\tilde{b}| - 3, |\tilde{b}| + 3]$ with a step size of 0.1, totaling to $K$ grid values. The minimax problem becomes

$$\max_{\pi \in \Delta^J} \min_{\{\psi_k\}_{k=1}^K} \sum_{\ell=1}^{J} \pi_\ell \left( \sum_{k=1}^{K} \psi_k^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^{K} \psi_k \mu_{kb_\ell} + b_\ell^2 \right). \tag{22}$$

2. The solution to the inner optimization yields the posterior mean $\psi_k^*(\pi) = \frac{\sum_{\ell=1}^{J} \pi_\ell \mu_{kb_\ell} b_\ell}{\sum_{\ell=1}^{J} \pi_\ell \mu_{kb_\ell}}$.

The outer problem is then

$$\max_{\pi \in \Delta^J} \sum_{\ell=1}^{J} \pi_\ell \left( \sum_{k=1}^{K} \left(\psi_k^*(\pi)\right)^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^{K} \psi_k^*(\pi)\, \mu_{kb_\ell} + b_\ell^2 \right).$$

3. Solve the outer problem for the least favorable prior $\pi^*$ based on sequential quadratic programming via MATLAB's fmincon routine. The minimax estimator $\delta^{\mathrm{BNM}}\left(T; |\tilde{b}|\right)$ is therefore $\sum_{k=1}^{K} \psi_k^*(\pi^*)\, 1\{t = t_k\}$ and the minimax risk $r^{\mathrm{BNM}}(|\tilde{b}|)$ is the minimized value.

Since the objective is concave in $\pi$ (it is the pointwise infimum over a set of linear functions; see Boyd and Vandenberghe, 2004, p. 81), we can check that the algorithm has found a global maximum by checking for a local maximum.

## D.3 Computing the optimally adaptive estimator for a given $\rho^2$

As explained in the main text, the adaptive problem in the main example only depends on $\Sigma$ through the correlation coefficient $\rho^2$. For a given value of $\rho^2$, we use convex programming methods to solve for the function $\tilde{\delta}^{\mathrm{adapt}}(t; \rho)$ based on the steps described below, which is a specialized version of the second step of Algorithm 4.1.

1. Approximate the prior $\pi$ with the finite dimensional vector $\pi \in \Delta^J$, where the parameter space for $b/\sqrt{\Sigma_O}$ is approximated by an equally spaced grid of $\tilde{b}$ values spanning $[-9, 9]$ with a step size of 0.025, totaling to $J$ grid values. Approximate the conditional risk function as in (21), where the support for $T \sim N(\tilde{b}, 1)$ is approximated by an equally spaced grid of $t$ values spanning $[-12, 12]$ with a step size of 0.05, totaling to $K$ grid values. The adaptation problem (6) becomes

$$\max_{\pi \in \Delta^J} \min_{\{\psi_k\}_{k=1}^K} \sum_{\ell=1}^{J} \pi_\ell \omega_\ell \left( \sum_{k=1}^{K} \psi_k^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^{K} \psi_k \mu_{kb_\ell} + b_\ell^2 \right) + \rho^{-2} - 1 \tag{23}$$

   where $\omega_\ell = \left( r^{\mathrm{BNM}}(|\tilde{b}_\ell|) + \rho^{-2} - 1 \right)^{-1}$ using output from the previous subsection.

2. The solution to the inner optimization yields $\psi_k^*(\pi) = \frac{\sum_{\ell=1}^{J} \pi_\ell \mu_{kb_\ell} \omega_\ell b_\ell}{\sum_{\ell=1}^{J} \pi_\ell \mu_{kb_\ell} \omega_\ell}$. The outer prob-

lem is then

$$\max_{\pi\in\Delta^J}\sum_{\ell=1}^{J}\pi_\ell\omega_\ell\left(\sum_{k=1}^{K}\left(\psi_k^*\left(\pi\right)\right)^2\mu_{kb_\ell}-2b_\ell\sum_{k=1}^{K}\psi_k^*\left(\pi\right)\mu_{kb_\ell}+b_\ell^2\right)+\rho^{-2}-1.$$

3. Solve the outer problem for the least favorable (adaptive) prior $\pi^*$ based on sequential quadratic programming via Matlab's fmincon routine. The adaptive estimator $\tilde{\delta}^{\text{adapt}}(t;\rho)$ is therefore $\sum_{k=1}^{K}\psi_k^*\left(\pi^*\right)1\left\{t=t_k\right\}$. The loss of efficiency under adaptation is the minimized value.

As with the bounded normal mean problem, the objective is concave in $\pi$, so we can check that the algorithm has found a global maximum by checking for a local maximum.

## D.4 Computing the optimally adaptive estimator based on the lookup table

To simplify the computation of the optimally adaptive estimator, we pre-calculate the adaptive estimates over an unequally spaced grid $\tanh([0, 0.05, 0.10, \ldots, 3])$ of correlation coefficients using the algorithm described above. As $\rho^2$ approaches one, the solution becomes sensitive to small changes in $\rho$. The uneven spacing of the $\rho$ grid allows for more accurate interpolation based on the simple pre-tabulated lookup table that we describe next.

To rapidly obtain a final estimator $\tilde{\delta}^{\text{adapt}}(T_O;\rho)$ for a given application, we conduct 2D interpolation across $\rho^2$ and $t$ values to tailor the adaptive estimates to the exact parameter values desired. For example, we obtain $\tilde{\delta}\left(T_O;-0.524\right)$ based on spline interpolation at $\rho^2=(-0.524)^2$ together with the observed test statistic $T_O$ based on the 2D grid of $\rho^2$ and $t$ values.

Figure A6 plots the maximum and minimum values of $\delta(T_O)/T_O$ against $\rho^2$. For all enumerated values of $\rho^2$, the adaptive estimator "shrinks" $T_O$ towards zero.

## D.5 Computing the nearly adaptive estimators

To find the nearly adaptive estimators in the class of soft thresholding estimators and hard thresholding estimators, it suffices to solve the two dimensional minimax problem in threshold $\lambda$ and scaled bias level $\tilde{b}$. We provide details for the claim in the main text that this two

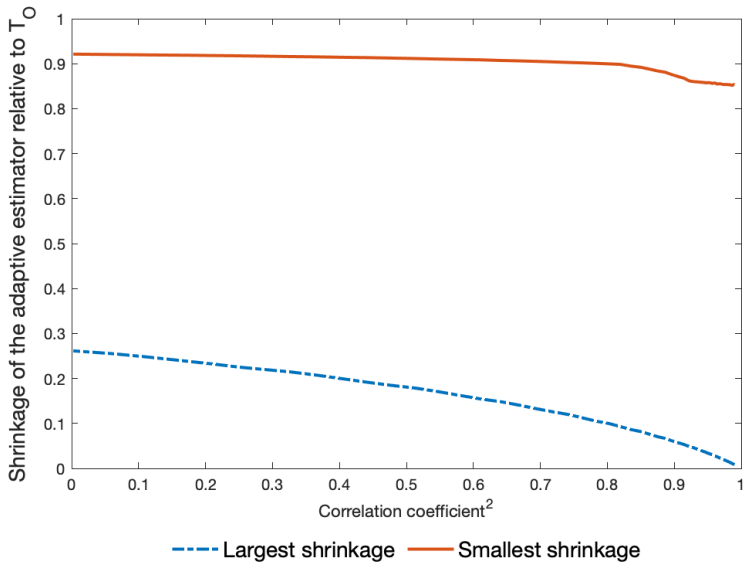


Figure A6: Shrinkage pattern for the adaptive estimator

dimensional minimax problem can be easily solved in practice even though the minimax theorem does not apply to these restricted classes of estimators. The derivation is largely based on the following equality using moments of a truncated standard normal $X_i \mid a < X_i < b$. Let $\phi(x)$ and $\Phi(x)$ denote the pdf and cdf of a standard normal distribution. Then for any $a < b$, we have

$$\int_a^b x^2 \phi(x) dx = \Phi(b) - \Phi(a) - (b\phi(b) - a\phi(a)). \tag{24}$$

### D.5.1 Soft thresholding

Rewrite the soft thresholding estimator as $\delta_{S,\lambda}(T_O) = \mathbf{1}\{T_O > \lambda\}(T_O - \lambda) + \mathbf{1}\{T_O < -\lambda\}(T_O + \lambda)$ and its risk function can be expressed as

$$\begin{aligned} & E_{T_O \sim N(\tilde{b},1))} \left(\delta_{S,\lambda}(T_O) - \tilde{b}\right)^2 \\ = \; & E_{T_O \sim N(\tilde{b},1)} \left(\mathbf{1}\{T_O > \lambda\}\left(T_O - \lambda - \tilde{b}\right) + \mathbf{1}\{T_O < -\lambda\}\left(T_O + \lambda - \tilde{b}\right) - \mathbf{1}\{-\lambda < T_O < \lambda\}\tilde{b}\right)^2 \\ = \; & \tilde{b}^2 \left(\Phi\left(\lambda - \tilde{b}\right) - \Phi\left(-\lambda - \tilde{b}\right)\right) + \int_{\lambda - \tilde{b}}^{\infty} (x - \lambda)^2 \phi(x) dx + \int_{-\infty}^{-\lambda - \tilde{b}} (x + \lambda)^2 \phi(x) dx \end{aligned} \tag{25}$$

The integrals in (25) simplify to

$$
\begin{aligned}
&\int_{\lambda-\tilde{b}}^{\infty}(x-\lambda)^2\,\phi(x)dx+\int_{-\infty}^{-\lambda-\tilde{b}}(x+\lambda)^2\,\phi(x)dx\\
=&\int_{\lambda-\tilde{b}}^{\infty}x^2\phi(x)dx+\int_{-\infty}^{-\lambda-\tilde{b}}x^2\phi(x)dx\\
&-2\lambda\left(\int_{\lambda-\tilde{b}}^{\infty}x\phi(x)dx-\int_{-\infty}^{-\lambda-\tilde{b}}x\phi(x)dx\right)\\
&+\lambda^2\left(1-\Phi\left(\lambda-\tilde{b}\right)+\Phi\left(-\lambda-\tilde{b}\right)\right)\\
=&1-\Phi\left(\lambda-\tilde{b}\right)+\Phi\left(-\lambda-\tilde{b}\right)+\left((\lambda-\tilde{b})\phi(\lambda-\tilde{b})-(-\lambda-\tilde{b})\phi(-\lambda-\tilde{b})\right)\\
&-2\lambda\left(\phi(\lambda-\tilde{b})+\phi(-\lambda-\tilde{b})\right)+\lambda^2\left(1-\Phi\left(\lambda-\tilde{b}\right)+\Phi\left(-\lambda-\tilde{b}\right)\right)
\end{aligned}
$$

where we use the fact that $\int_{\lambda-\tilde{b}}^{\infty}x^2\phi(x)dx+\int_{-\infty}^{-\lambda-\tilde{b}}x^2\phi(x)dx=\int_{-\infty}^{\infty}x^2\phi(x)dx-\int_{-\lambda-\tilde{b}}^{\lambda-\tilde{b}}x^2\phi(x)dx$ and Equation (24).

The nearly adaptive objective function

$$
\min_{\lambda}\max_{\tilde{b}}\frac{E_{T_O\sim N(\tilde{b},1))}\left(\delta_{S,\lambda}\left(T_O\right)-\tilde{b}\right)^2+\rho^{-2}-1}{r^{\mathrm{BNM}}(|\tilde{b}|)+\rho^{-2}-1},
$$

can now be easily solved by Matlab's fminimax function when the risk function is evaluated based on the simplified expression derived above.

To simplify the computation of the nearly adaptive estimator, we pre-calculate the adaptive thresholds over an unequally spaced grid $\tanh([0,0.05,0.10,\ldots,3])$ of correlation coefficients as explained above. To rapidly obtain a final estimator $\delta_{S,\lambda}\left(T_O;\rho\right)$ for a given application, we conduct a spline interpolation across $\rho^2$ values to tailor the threshold to the exact parameter values desired. For example, we obtain $\delta_{S,\lambda}\left(T_O;-0.524\right)$ firstly based on spline interpolation at $\rho^2=(-0.524)^2$ to obtain the threshold $\lambda$, and then with the observed test statistic $T_O$.

### D.5.2 Hard thresholding

Similarly rewrite hard thresholding as $\delta_{H,\lambda}(T_O) = (1 - \mathbf{1}\{-\lambda < T_O < \lambda\}) T_O$ and its risk function can be simplified due to Equation (24)

$$\begin{aligned} & E_{T_O \sim N(\tilde{b},1))} \left( \delta_{H,\lambda}(T_O) - \tilde{b} \right)^2 \\ = \; & E_{T_O \sim N(\tilde{b},1)} \left( (1 - \mathbf{1}\{-\lambda < T_O < \lambda\}) \left( T_O - \tilde{b} \right) - \mathbf{1}\{-\lambda < T_O < \lambda\} \tilde{b} \right)^2 \\ = \; & \tilde{b}^2 \left( \Phi\left( \lambda - \tilde{b} \right) - \Phi\left( -\lambda - \tilde{b} \right) \right) + \int_{-\infty}^{\infty} x^2 \phi(x) dx - \int_{-\lambda - \tilde{b}}^{\lambda - \tilde{b}} x^2 \phi(x) dx. \end{aligned}$$

# Appendix E Pooling controls (LaLonde, 1986)

LaLonde (1986) contrasted experimental estimates of the causal effects of job training derived from the National Supported Work (NSW) demonstration with econometric estimates derived from observational controls, concluding that the latter were highly sensitive to modeling choices. Subsequent work by Heckman and Hotz (1989) argued that proper use of specification tests would have guarded against large biases in LaLonde (1986)'s setting. An important limitation of the NSW experiment, however, is that its small sample size inhibits a precise assessment of the magnitude of selection bias associated with any given non-experimental estimator. In what follows, we explore the prospects of improving experimental estimates of the NSW's impact on earnings by utilizing additional non-experimental control groups and adapting to the biases their inclusion engenders.

We consider three analysis samples differentiated by the origin of the untreated ("control") observations. All three samples include the experimental NSW treatment group observations. In the first sample the untreated observations are given by the experimental NSW controls. In a second sample the controls come from LaLonde (1986)'s observational "CPS-1" sample, as reconstructed by Dehejia and Wahba (1999). In the third sample, the controls are a propensity score screened subsample of CPS-1. To estimate treatment effects in the samples with observational controls, we follow Angrist and Pischke (2009) in fitting linear models for 1978 earnings to a treatment dummy, 1974 and 1975 earnings, a quadratic in age, years of schooling, a dummy for no degree, a race and ethnicity dummies, and a dummy for marriage status. The propensity score is generated by fitting a probit model of treatment status on the same covariates and dropping observations with predicted treatment

probabilities outside of the interval $[0.1, 0.9]$.

Let $Y_U$ be the mean treatment / control contrast in the experimental NSW sample. We denote by $Y_{R1}$ the estimated coefficient on the treatment dummy in the linear model described above when the controls are drawn from the CPS-1 sample. Finally, $Y_{R2}$ gives the corresponding estimate obtained from the linear model when the controls come from the propensity score screened CPS-1 sample. We follow the applied literature in assuming trimming does not meaningfully change the estimand, a perspective that can be formalized by viewing the trimmed estimator as one realization of a sequence of estimators with trimming shares that decrease rapidly with the sample size (Huber et al., 2013).

Table A1 reports point estimates from all three estimation approaches along with standard errors derived from the pairs bootstrap. The realizations of $(Y_{R1}, Y_{R2})$ exactly reproduce those found in the last row of Table 3.3.3 of Angrist and Pischke (2009) but the reported standard errors are somewhat larger due to our use of the bootstrap, which accounts both for heteroscedasticity and uncertainty in the propensity score screening procedure. The realization of $Y_U$ matches the point estimate reported in the first row of Angrist and Pischke (2009)'s Table 3.3.3 but again exhibits a modestly larger standard error reflecting heteroscedasticity with respect to treatment status.

| | $Y_U$ | $Y_{R1}$ | $Y_{R2}$ | $GMM_2$ | $GMM_3$ | Adaptive | Pre-test |
|---|---|---|---|---|---|---|---|
| Estimate | 1794 | 794 | 1362 | 1629 | 1210 | 1597 | 1629 |
| Std error | (668) | (618) | (741) | (619) | (595) | | |
| Max Regret | 26% | $\infty$ | $\infty$ | $\infty$ | $\infty$ | 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 | $\infty$ | 1.23 | 0.858 | $\infty$ | 0.925 | 0.993 |
| when $b_1 \neq 0$ and $b_2 \neq 0$ | 1 | $\infty$ | $\infty$ | $\infty$ | $\infty$ | 1.077 | 1.475 |

Table A1: Estimates of the impact of NSW job training on earnings. Bootstrap standard errors in parentheses computed using 1,000 bootstrap samples. The $GMM_2$ estimate imposes $b_2 = 0$ only while the $GMM_3$ estimate imposes $b_1 = 0$ and $b_2 = 0$. A $J$-test of the null $b_1 = b_2 = 0$ motivating $GMM_3$ yields a p-value at 0.04. A corresponding test of the null $b_2 = 0$ motivating $GMM_2$ yields a p-value of 0.51. "Risk rel. to $Y_U$" gives worst case risk scaled by the risk (i.e. variance) of $Y_U$. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$.

While the experimental mean contrast $(Y_U)$ of \$1,794 is statistically distinguishable from zero at the 5% level, considerable uncertainty remains about the magnitude of the average treatment effect of the NSW program on earnings. The propensity trimmed CPS-1 estimate

lies closer to the experimental estimate than does the estimate from the untrimmed CPS-1 sample. However, the untrimmed estimate has a much smaller standard error than its trimmed analogue. Though the two restricted estimators are both derived from the CPS-1 sample, our bootstrap estimate of the correlation between them is only 0.75, revealing that each measure contains substantial independent information.

Combining the three estimators together via GMM, a procedure we denote $GMM_3$, yields roughly an 11% reduction in standard errors relative to relying on $Y_U$ alone. However, the $J$-test associated with the $GMM_3$ procedure rejects the null hypothesis that the three estimators share the same probability limit at the 5% level ($p = 0.04$). Combining only $Y_U$ and $Y_{R2}$ by GMM, a procedure we denote $GMM_2$, yields a standard error 7% below that of $Y_U$ alone. The $J$-test associated with $GMM_2$ fails to reject the restriction that $Y_U$ and $Y_{R2}$ share a common probability limit ($p = 0.51$). Hence, sequential pre-testing selects $GMM_2$.

Letting $b_1 \equiv \mathbb{E}[Y_{R1} - \theta]$ and $b_2 \equiv \mathbb{E}[Y_{R2} - \theta]$ our pre-tests reject the null that $b_1 = b_2 = 0$ and fail to reject that $b_2 = 0$. However, it seems plausible that both restricted estimators suffer from some degree of bias. The adaptive estimator seeks to determine the magnitude of those biases and make the best possible use of the observational estimates. In adapting to misspecification, we operate under the assumption that $|b_1| \geq |b_2|$, which is in keeping with the common motivation of propensity score trimming as a tool for bias reduction (e.g., Angrist and Pischke, 2009, Section 3.3.3). Denoting the bounds on $(|b_1|, |b_2|)$ by $(B_1, B_2)$, we adapt over the finite collection of bounds $\mathcal{B} = \{(0, 0), (\infty, 0), (\infty, \infty)\}$, the granular nature of which dramatically reduces the computational complexity of finding the optimally adaptive estimator. Note that the scenario $(B_1, B_2) = (0, \infty)$ has been ruled out by assumption, reflecting the belief that propensity score trimming reduces bias. See Appendix F for further details.

From Table A1, the multivariate adaptive estimator yields an estimated training effect of \$1,597: roughly two thirds of the way towards $Y_U$ from the efficient $GMM_3$ estimate. Hence, the observational evidence, while potentially quite biased, leads to a non-trivial (11%) adjustment of our best estimate of the effect of NSW training away from the experimental benchmark. In Table A2 we show that pairwise adaptation using only $Y_U$ and $Y_{R1}$ or only $Y_U$ and $Y_{R2}$ yields estimates much closer to $Y_U$. A kindred approach, which avoids completely discarding the information in either restricted estimator, is to combine $Y_{R1}$ and $Y_{R2}$ together via optimally weighted GMM and then adapt between $Y_U$ and the composite GMM estimate.

As shown in Table A3, this two step approach yields an estimate of \$1,624, extremely close to the multivariate adaptive estimate of \$1,597, but comes with substantially elevated worst case adaptation regret relative to a multivariate oracle who knows which pair of bounds in $\mathcal{B}$ prevails.

While the multivariate adaptive estimate of \$1,597 turns out to be very close to the pre-test estimate of \$1,629, the adaptive estimator's worst case adaptation regret of 7.7% is substantially lower than that of the pre-test estimator, which exhibits a maximal regret of 47.5%. The adaptive estimator achieves this advantage by equalizing the maximal adaptation regret across the three bias scenarios $\{(b_1 = 0, b_2 = 0), (b_1 \neq 0, b_2 = 0), (b_1 \neq 0, b_2 \neq 0)\}$ allowed by our specification of $\mathcal{B}$. When both restricted estimators are unbiased, the adaptive estimator yields a 14.5% reduction in worst case risk relative to $Y_U$. However, an oracle that knows both restricted estimators are unbiased would choose to employ $GMM_3$, implying maximal adaptation regret of $0.855/0.793 \approx 1.077$. When $Y_{R1}$ is biased, but $Y_{R2}$ is not, the adaptive estimator yields a 7.5% reduction in worst case risk. An oracle that knows only $Y_{R1}$ is biased will rely on $GMM_2$, which yields worst case scaled risk of 0.858; hence, the worst case adaptation regret of not having employed $GMM_2$ in this scenario is $0.925/0.858 \approx 1.077$. Finally, when both restricted estimators are biased, the adaptive estimator can exhibit up to a 7.7% increase in risk relative to $Y_U$.

The near oracle performance of the optimally adaptive estimator in this setting suggests it should prove attractive to researchers with a wide range of priors regarding the degree of selection bias present in the CPS-1 samples. Both the skeptic that believes the restricted estimators may be immensely biased and the optimist who believes the restricted estimators are exactly unbiased should face at most a 7.7% increase in maximal risk from using the adaptive estimator. In contrast, an optimist could very well object to a proposal to rely on $Y_U$ alone, as doing so would raise risk by 26% over employing $GMM_3$.

# Appendix F Details of bivariate adaptation

In Section Appendix E, we report the results of adapting simultaneously to the bias in two restricted estimators when the bias spaces take a nested structure. Denoting the bounds on $(|b_1|, |b_2|)$ of the two restricted estimators by $(B_1, B_2)$, we adapt over the finite collection of bounds $\mathcal{B} = \{(0, 0), (\infty, 0), (\infty, \infty)\}$. Note that the scenario $(B_1, B_2) = (0, \infty)$ has been

ruled out by assumption, reflecting the belief that propensity score trimming reduces bias. The minimax risk over each bias space $\mathcal{C}_{(B_1,B_2)}$ is therefore

$$R^*(\mathcal{C}_{(B_1,B_2)}) = \begin{cases} \Sigma_U & \text{for } (B_1,B_2) = (\infty,\infty) \\ \Sigma_U - \Sigma_{UO,2}\Sigma_{O,2}^{-1}\Sigma_{UO,2} & \text{for } (B_1,B_2) = (\infty,0) \\ \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO} & \text{for } (B_1,B_2) = (0,0) \end{cases} \tag{26}$$

Then $\delta(Y_O)$ is the solution to the following problem

$$\inf_{\delta} \max_{(B_1,B_2)\in\mathcal{B}} \frac{\max_{b\in\mathcal{C}_{(B_1,B_2)}} E_{Y_O\sim N(b,\Sigma_O)}(\delta(Y_O) - \Sigma_{UO}\Sigma_O^{-1}b)^2 + \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO}}{R^*(\mathcal{C}_{(B_1,B_2)})}$$

Since the three spaces are nested, we can rewrite the adaptation problem as

$$\inf_{\delta} \sup_{b\in\mathbb{R}\times\mathbb{R}} \frac{E_{Y_O\sim N(b,\Sigma_O)}(\delta(Y_O) - \Sigma_{UO}\Sigma_O^{-1}b)^2 + \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO}}{\tilde{R}(\tilde{\mathcal{S}}(b))}$$

where the scaling is

$$\tilde{R}(\tilde{\mathcal{S}}(b)) = \begin{cases} \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO} & \text{if } b_1 = b_2 = 0 \\ \Sigma_U - \Sigma_{UO,2}\Sigma_{O,2}^{-1}\Sigma_{UO,2} & \text{if } b_1 \neq 0, b_2 = 0 \\ \Sigma_U & \text{if } b_1 \neq 0, b_2 \neq 0 \end{cases} \tag{27}$$

Given the high dimensionality of the adaptation problem, we use CVX instead of Matlab's fmincon to solve the scaled minimax problem.

## F.1 Shrinkage pattern

To illustrate the shrinkage properties of the multivariate adaptive estimator, Figure A7 plots the adaptive minimax estimator of bias against its unbiased counterpart $\Sigma_{U,O}\Sigma_O^{-1}Y_O$. The figure reveals a complex shrinkage pattern reflecting the asymmetric nature of $\mathcal{C}_B$. When $Y_{O1} = Y_{R1} - Y_U$ is small, $Y_{O2} = Y_{R2} - Y_U$ is shrunk aggressively towards zero. However when $Y_{O2}$ is small, $Y_{O1}$ is shrunk less aggressively towards zero. When both $Y_{O1}$ and $Y_{O2}$ are large, the biases exhibit little shrinkage.

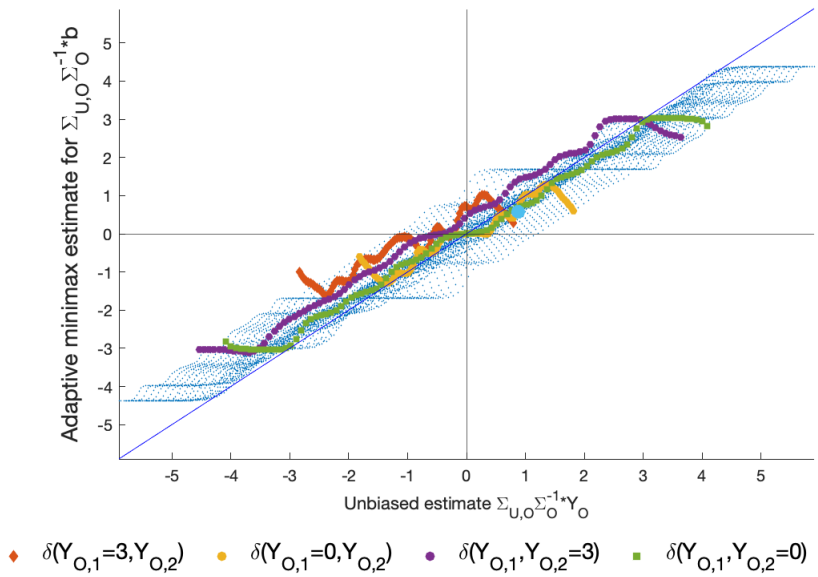


Figure A7: The adaptive minimax estimator of bias are illustrated by blue dots in the background, plotted against the their unbiased counterparts. The highlights are the estimates holding $Y_{O1}$ and $Y_{O2}$ constant respectively. In particular, the big blue dot highlights the adaptive estimate for the LaLonde example, which involves shrinkage.

## F.2 Pairwise adaptation

For comparison with the trivariate adaptation estimates reported in the text, we also consider pairwise adaptation using only $Y_U$ and $Y_{R1}$ or only $Y_U$ and $Y_{R2}$, keeping the bias spaces as before. Specifically to adapt using only $Y_U$ and $Y_{Rj}$, we consider an oracle where the set $\mathcal{B}$ of bounds $B$ on the bias consists of the two elements 0 and $\infty$.

Table A2 shows that pairwise adaptation produces estimates much closer to $Y_U$ than the multivariate adaptive estimate. While pairwise adaptive estimates both incur smaller adaptation regret, the efficiency gain when the model is correct is smaller than with the multivariate adaptive estimate.

## F.3 Bivariate adaptation with GMM composite

For another comparison with the trivariate adaptation estimates reported in the text, we also consider combining $Y_{R1}$ and $Y_{R2}$ first via optimally weighted GMM, which is a composite of the two $Y_{\text{comp}}$. We then adapt between $Y_U$ and $Y_{\text{comp}}$. The bias space is now also a composite of the two-dimensional bias space $\mathcal{C}_{(B_1,B_2)}$, and we consider an oracle where the

| | $Y_U$ | $Y_R$ | GMM | Adaptive | Soft-threshold | Pre-test |
|---|---|---|---|---|---|---|
| CPS-1 untrimmed | 1794 | 794 | 1123 | 1659 | 1608 | 1794 |
| Std error | (668) | (617) | (600) | | | |
| Rel. risk when $b = 0$ | 1 | 0.85 | 0.81 | 0.863 | 0.869 | 0.894 |
| Rel. risk when $b \neq 0$ | 1 | $\infty$ | $\infty$ | 1.071 | 1.078 | 1.541 |
| Max Regret | 24% | $\infty$ | $\infty$ | 7.1% | 7.8% | 54% |
| Max Regret (rel. to multivariate) | 26% | $\infty$ | $\infty$ | 24.8% | 25.6% | 79.5% |
| Threshold | | | | | 0.63 | 1.96 |
| CPS-1 trimmed | 1794 | 1362 | 1629 | 1657 | 1638 | 1362 |
| Std error | (668) | (741) | (619) | | | |
| Rel. risk when $b = 0$ | 1 | 1.23 | 0.86 | 0.9 | 0.91 | 1.166 |
| Rel. risk when $b \neq 0$ | 1 | $\infty$ | $\infty$ | 1.05 | 1.055 | 2.051 |
| Max Regret | 16.4% | $\infty$ | $\infty$ | 5% | 5.5% | 105% |
| Max Regret (rel. to multivariate) | 26% | $\infty$ | $\infty$ | 13.6% | 14.2% | 105% |
| Threshold | | | | | 0.62 | 1.96 |

Table A2: Estimates of the impact of NSW job training on earnings. Bootstrap standard errors in parentheses computed using 1,000 bootstrap samples. In the top panel $Y_R$ corresponds to estimates using the untrimmed CPS-1 as controls, which are referred to as $Y_{R1}$ in the main text. In the bottom panel, $Y_R$ corresponds to estimates derived from the propensity score trimmed CPS-1 sample, which are referred to as $Y_{R2}$ in the main text. Adaptive estimates adapt pairwise between $Y_U$ and $Y_R$ within panel. If applicable, the adaptive thresholds are reported. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$. "Max Regret (rel. to multivariate)" refers to the worst case adaptation regret in terms of the multivariate oracle. "Rel. risk" gives worst case risk scaled by the risk (i.e. variance) of $Y_U$. The correlation between $Y_U$ and $Y_{Rj} - Y_U$ is -0.44 in the top panel and -0.38 in the bottom panel.

set $\mathcal{B}$ of bounds $B$ on the bias consists of the two elements 0 and $\infty$.

Table A3 shows that composite adaptation produces estimates very similar to the multivariate adaptive estimate. The adaptation regret relative to an oracle who knows a bound on the bias of composite is also small. However, for a fair comparison with multivariate adaptation, one should compare its efficiency loss relative to the multivariate oracle with minimax risk specified in (26). This notion of worst case regret is substantially higher at 25% because bivariate adaptation against the GMM composite cannot leverage the nested structure of the multivariate parameter space $\mathcal{B}$.

| | $Y_U$ | $Y_{\text{comp}}$ | GMM | Adaptive | Soft-threshold | Pre-test |
|---|---|---|---|---|---|---|
| Estimate | 1794 | 882 | 1173 | 1624 | 1601 | 1794 |
| Std error | (668) | (612) | (595) | | | |
| Max Regret | 26% | $\infty$ | $\infty$ | 8% | 8.3% | 56% |
| Max Regret (rel. to multivariate) | 26% | $\infty$ | $\infty$ | 25.4% | 26.3% | 81.5% |
| Threshold | | | $\infty$ | | 0.64 | 1.96 |

Table A3: Adaptive estimates for the impact of job training, adapting to $B_{\text{comp}} \in \{0, \infty\}$, which is the bound on the bias of the composite estimator $Y_{\text{comp}} = \arg\min_\theta (Y_R - \theta)' \Sigma_R (Y_R - \theta)$. If applicable, the adaptive thresholds are reported. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$. "Max Regret (rel. to multivariate)" refers to the worst case adaptation regret relative to the multivariate oracle in (26). The correlation coefficient between $Y_U$ and $Y_{\text{comp}} - Y_U$ is -0.45.

# References for Online Appendix


Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bickel, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In M. H. Rizvi, J. S. Rustagi, and D. Siegmund (Eds.), *Recent Advances in Statistics*, pp. 511–528. Academic Press.

Boyd, S. P. and L. Vandenberghe (2004, March). *Convex Optimization*. Cambridge University Press.

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association 94*(448), 1053–1062.

Heckman, J. J. and V. J. Hotz (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American statistical Association 84*(408), 862–874.

Huber, M., M. Lechner, and C. Wunsch (2013). The performance of estimators based on the propensity score. *Journal of Econometrics 175*(1), 1–21.

Johnstone, I. M. (2019). *Gaussian estimation: Sequence and wavelet models*. Online manuscript available at https://imjohnstone.su.domains/.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons.