Online Appendix to "Adapting to Misspecification"

Timothy B. Armstrong, Patrick Kline and Liyang Sun

August 2024

Appendix C Additional details

C.1 Constrained adaptation

The constrained adaptive estimator solves the problem

$$A^*(\mathcal{B}; \overline{R}) = \inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \hat{\theta})}{R^*(B)} \quad \text{s.t.} \quad \sup_{B \in \mathcal{B}} R_{\max}(B, \hat{\theta}) \leq \overline{R}. \tag{19}$$

We can rewrite this formulation as a weighted minimax problem similar to the one in Section 4.1 by setting $t = \overline{R}/A^*(\mathcal{B}; \overline{R})$ and considering the problem

$$\inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} \max \left\{ \frac{R_{\max}(B, \hat{\theta})}{R^*(B)}, \frac{R_{\max}(B, \hat{\theta})}{t} \right\} = \inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \hat{\theta})}{\min\{R^*(B), t\}}. \tag{20}$$

Indeed, any solution to (19) must also be a solution to (20) with $t = \overline{R}/A^*(\mathcal{B}; \overline{R})$, since any decision function achieving a strictly better value of (20) would satisfy the constraint in (19) and achieve a strictly better value of the objective in (19). Conversely, letting $\tilde{A}^*(t)$ be the value of (20), any solution to (20) will achieve the same value of the objective (19) and will satisfy the constraint for $\bar{R} = t \cdot \tilde{A}^*(t)$. In fact, this solution to (20) will also solve (19) for $\bar{R} = t \cdot \tilde{A}^*(t)$ so long as this value of $\bar{R}$ is large enough to allow some scope for adaptation.

Arguing as in Section 4.1, we can write the optimization problem (20) as

$$\inf_{\hat{\theta}} \sup_{(\theta,b) \in \cup_{B' \in \mathcal{B}} \mathcal{C}_{B'}} \tilde{\omega}(\theta, b, t) R(\theta, b, \hat{\theta}), \tag{21}$$

$$\text{where } \tilde{\omega}(\theta, b, t) = \left(\inf_{B \in \mathcal{B} \text{ s.t. } (\theta,b) \in \mathcal{C}_B} \min\{R_{\max}(B), t\} \right)^{-1} = \max\{\omega(\theta, b), 1/t\}$$

and $\omega(\theta, b)$ is given in Lemma 4.1 in Section 4.1. Thus, we can solve (20) by solving for the minimax estimator under the loss function $(\theta, b, d) \mapsto \tilde{\omega}(\theta, b, t)L(\theta, b, d)$. Letting $A^*(t)$ be the optimized objective function, we can then solve (19) by finding a t such that $\bar{R} = t \cdot A^*(t)$.

We summarize these results in the following lemma, which is proved in Section C.1.1 of the appendix.

Lemma C.1. *Any solution to (19) is also a solution to (21) with $t = \overline{R}/A^*(\mathcal{B}; \overline{R})$. Conversely, let $\tilde{A}^*(t)$ denote the value of (21) and let $\tilde{R}(t) = \tilde{A}^*(t) \cdot t$. If $\tilde{R}(t) > \inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} R_{\max}(B, \hat{\theta})$ and $\inf_{B \in \mathcal{B}} R^*(B) > 0$, then $A^*(\mathcal{B}; \tilde{R}(t)) = \tilde{A}^*(t)$ and any solution to (21) is also a solution to (19) with $\bar{R} = \tilde{R}(t)$.*

C.1.1 Details for constrained adaptation

We provide proof for Lemma C.1, which shows the constrained adaption problem is equivalent to the weighted minimax problem with a particular set of weights. The first statement is immediate from the arguments proceeding the statement of the lemma in Section 4.4. For the second statement, let $\bar{\delta}$ be a decision rule with $\sup_{B \in \mathcal{B}} R_{\max}(B, \bar{\delta}) < \tilde{R}(t)$. Such a decision rule exists and satisfies $\sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \bar{\delta})}{R^*(B)} < \infty$ by the assumptions of the lemma. Let δ_t^* be a solution to (20).

Suppose, to get a contradiction, that a decision δ' satisfies the constraint in (19) with $\bar{R} = \tilde{R}(t)$ and achieves a strictly better value of the objective than $\tilde{A}^*(t)$. For $\lambda \in (0, 1)$, let δ'_λ be the randomized decision rule that places probability λ on $\bar{\delta}$ and probability $1 - \lambda$ on δ', independently of the data Y. Note that $R_{\max}(B, \delta'_\lambda) = \sup_{(\theta,b) \in \mathcal{C}_B} R(\theta, b, \delta'_\lambda) = \sup_{(\theta,b) \in \mathcal{C}_B} \left[\lambda R(\theta, b, \bar{\delta}) + (1 - \lambda)R(\theta, b, \delta')\right] \leq \sup_{(\theta,b) \in \mathcal{C}_B} \lambda R(\theta, b, \bar{\delta}) + \sup_{(\theta,b) \in \mathcal{C}_B} (1-\lambda)R(\theta, b, \delta') = \lambda R_{\max}(B, \bar{\delta}) + (1 - \lambda)R_{\max}(B, \delta')$ so that, for $\lambda \in (0, 1)$,

$$\sup_{B \in \mathcal{B}} R_{\max}(B, \delta_\lambda) \leq \lambda \sup_{B \in \mathcal{B}} R_{\max}(B, \bar{\delta}) + (1 - \lambda) \sup_{B \in \mathcal{B}} R_{\max}(B, \delta') < \tilde{R}(t) = \tilde{A}^*(t) \cdot t$$

and

$$\sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta_\lambda)}{R^*(B)} \leq \lambda \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \bar{\delta})}{R^*(B)} + (1 - \lambda) \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta')}{R^*(B)}.$$

Since $\sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \bar{\delta})}{R^*(B)}$ is finite and $\frac{\sup_{B \in \mathcal{B}} R_{\max}(B, \delta')}{R^*(B)} < \tilde{A}^*(t)$, the above display is strictly less than $\tilde{A}^*(t)$ for small enough λ. Thus, for small enough λ, the objective function in (21)

evaluated at the decision function δ_λ evaluates to

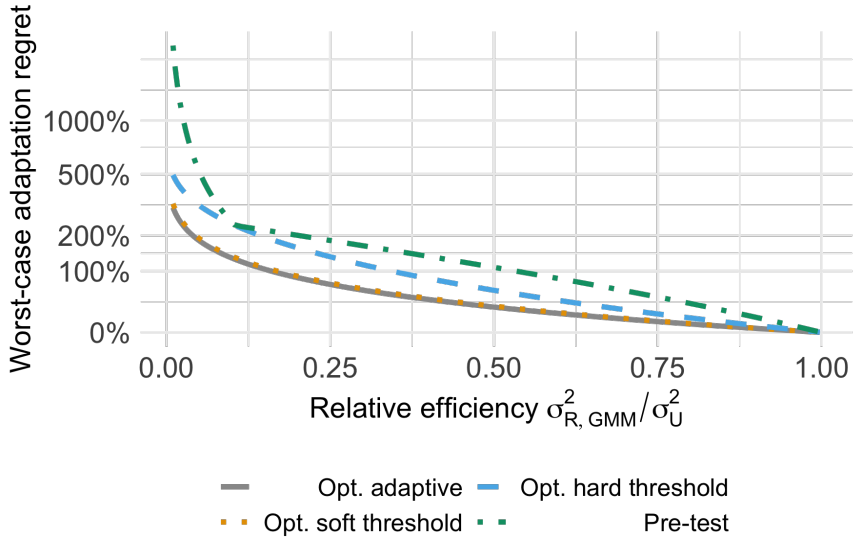
$$\max\left\{\sup_{B\in\mathcal{B}}\frac{R_{\max}(B,\delta_\lambda)}{R^*(B)},\sup_{B\in\mathcal{B}}\frac{R_{\max}(B,\delta_\lambda)}{t}\right\}<\max\left\{\tilde{A}^*(t),\tilde{R}(t)/t\right\}=\tilde{A}^*(t),$$

a contradiction.

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

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 ρ^2 is close to one as shown in Figure A1. To minimize the worst-case adaptation regret, the adaptive hard-threshold estimator needs to use a threshold that would increase to infinity as ρ^2 gets closer to one.

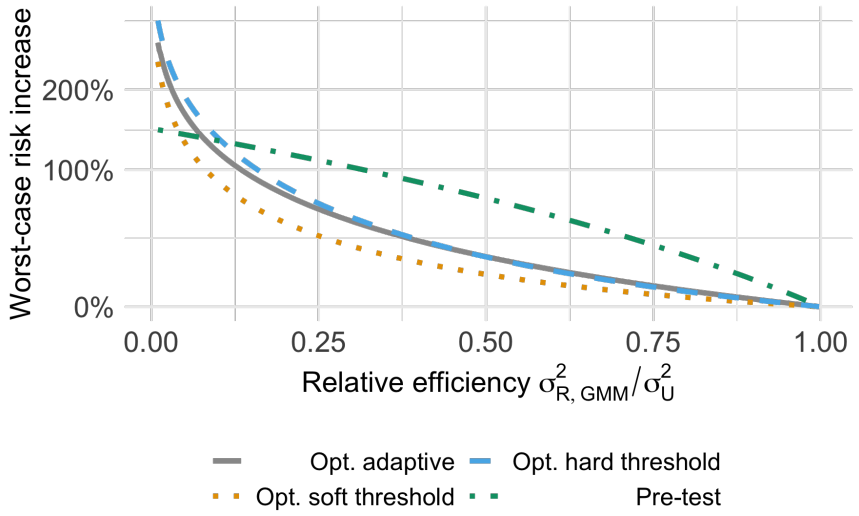
Figure A1: Worst case adaptation regret as function of relative efficiency



Notes: Vertical axis plots $(A_{\max}(\mathcal{B},\hat{\theta})-1)\times 100$ on $\log_{10}$ scale.

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 A2, the pre-test estimator tends to exhibit substantially greater worst-case risk than the class of adaptive estimators for most values of ρ^2. As discussed in Section C.3 below, adaptive estimators have large worst-case risk when ρ^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.

Figure A2: Worst case risk increase relative to Y_U



Notes: Vertical axis plots $(R_{\max}(\infty, \hat{\theta}) - \sigma_U)/\sigma_U \times 100$ on $\log_{10}$ scale.

C.3 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. Recall that $1 - \rho^2$ is equal to $\sigma^2_{R,GMM}/\sigma^2_U$, so that $|\rho| \to 1$ corresponds to the case where $\sigma^2_{R,GMM}/\sigma^2_U \to 0$. It will be convenient to phrase our results in terms of $\rho^{-2} - 1 = (1 - \rho^2)/\rho^2 = (1 + o(1)) \cdot \sigma^2_{R,GMM}/\sigma^2_U$ as $|\rho| \to 1$.

Let $A(\delta, \rho)$ denote the worst-case adaptation regret of the estimator given by (4) under the given value of ρ, 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 ρ. 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 λ, where $\delta_{S,\lambda}$ are $\delta_{H,\lambda}$ are defined in Section 4.3. 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.

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.3.1) and lower bounds (Section C.3.2). The lower bounds in Section C.3.2 are essentially immediate from results in Bickel (1983) for adapting to $B \in \mathcal{B} = \{0, \infty\}$, whereas the upper bounds in Section C.3.1 involve new arguments to deal with intermediate values of B.

C.3.1 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.2. *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 λ 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.2, $A_S^*(\rho)$ and $A_H^*(\rho)$ are, for $(\rho^{-2} - 1)^{-1}$ large enough, bounded from above by the supremum over μ of

$$\frac{\bar{r}(\tilde{\lambda}_\rho, \mu) + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\mu|) + \rho^{-2} - 1} \tag{22}$$

Let $c(\rho)$ be such that $c(\rho)/\tilde{\lambda}_\rho \to 0$ and $c(\rho) \to \infty$ as $|\rho| \uparrow 1$. We bound (22) separately for $|\mu| \le c(\rho)$ and for $|\mu| \ge c(\rho)$. For $|\mu| \le c(\rho)$, we use the bound $r^{\mathrm{BNM}}(|\mu|) \ge .8 \cdot \mu^2/(\mu^2+1)$ (Donoho, 1994), which gives an upper bound for (22) of

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

As $|\rho| \uparrow 1$, this increases more slowly than $\log(\rho^{-2}-1)^{-1}$. For $|\mu| \ge c(\rho)$, we use the bound $r^{\mathrm{BNM}}(|\mu|) \ge r^{\mathrm{BNM}}(c(\rho))$ which gives an upper bound for (22) of

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

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.3.2 Lower bounds

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

Given an estimator $\delta(Y)$ of μ 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.3 (Bickel 1983, Theorem 4.1)**.** *For $t \in (0,1]$, let δ_t be an estimator that satisfies $m(\delta_t) \le 1-t$. Then, as $t \uparrow 1$, $M(\delta_t) \ge (1+o(1)) \cdot 2\log(1-t)$.*

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.3.1. The required lower bound $A^*(\rho) \ge (1+o(1))2\log(\rho^{-2}-1)^{-1}$ follows from this result.

Lemma C.4. *For $\rho \in (-1,1)$, let $\delta_\rho : \mathbb{R} \to \mathbb{R}$ be an estimator of μ in the normal means problem $Y \sim N(\mu,1)$. Suppose that the worst-case adaptation regret $A(\delta_\rho, \rho)$ of the corre-*

sponding 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 B.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.3 with $1 - t = (\rho^{-2} - 1) \cdot (2 + \varepsilon)\log(\rho^{-2} - 1)^{-1}$, it follows that

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

as required.

□

Appendix D Computational details

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

D.1 Computing minimax estimators

As shown in Sections 4.1 and 4.2, one can compute adaptive estimators by solving a weighted minimax problem which, in our setting, can be further simplified using invariance. To solve

these problems, we use the insight that the minimax estimator can be characterized as a Bayes estimator for a *least favorable prior*. We first give a brief review of this approach before going into details for our setting.

Consider the generic problem of computing a minimax decision over the parameter space $\mathcal{C}$ for a parameter ϑ under loss $\bar{L}(\vartheta, \delta)$. We use E_ϑ and P_ϑ to denote expectation under ϑ and the probability distribution of the data Y under ϑ. Letting π denote a *prior* distribution on $\mathcal{C}$, the *Bayes risk* of δ is given by

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

The *Bayes decision*, which we will denote $\delta_\pi^{\text{Bayes}}$, optimizes $R_{\text{Bayes}}(\pi, \delta)$ over δ. It can be computed by optimizing expected loss under the posterior distribution for ϑ taking π as the prior. Under squared error loss, the Bayes decision is the posterior mean.

$R_{\text{Bayes}}(\pi, \delta)$ gives a lower bound for the worst-case risk of δ under $\mathcal{C}$ and $R_{\text{Bayes}}(\pi, \delta_\pi^{\text{Bayes}})$ gives a lower bound for the minimax risk. Under certain conditions, a *minimax theorem* applies, which tells us that this lower bound is in fact sharp. In this case, letting Γ denote the set of priors π supported on $\mathcal{C}$, the minimax risk over $\mathcal{C}$ is given by

$$\min_\delta \max_{\pi \in \Gamma} R_{\text{Bayes}}(\pi, \delta) = \max_{\pi \in \Gamma} \min_\delta R_{\text{Bayes}}(\pi, \delta) = \max_{\pi \in \Gamma} R_{\text{Bayes}}(\pi, \delta_\pi^{\text{Bayes}}).$$

The distribution π that solves this maximization problem is called the *least favorable prior*. When the minimax theorem applies, the Bayes decision for this prior is the minimax decision over $\mathcal{C}$.

The expression $R_{\text{Bayes}}(\pi, \delta_\pi^{\text{Bayes}})$ is convex as a function of π if the set of possible decision functions is sufficiently unrestricted and the set Γ is convex. While one may need to allow randomized decisions in general, the estimation problems we consider will be such that the Bayes decision is nonrandomized. Thus, we can use convex optimization software to compute the least favorable prior and minimax estimator so long as we have a way of approximating π with a finite dimensional object that retains the convex structure of the problem.

In our setting, we use invariance arguments to obtain the objective function (6), which is a minimax problem over the unknown parameter $\tilde{b} = b/\sigma_O$ (the noncentrality parameter of the overidentification statistic T_O). We solve (6), as well as the bounded normal mean problem used to obtain the scaling in (6), by solving for a least favorable prior over $\tilde{b}$ using

a finite dimensional approximation $\pi(\tilde{b}_1), \ldots, \pi(\tilde{b}_J)$ to the prior over a grid of J values of $\tilde{b}$. The least favorable prior for (θ, b) is then given by a flat (improper) prior for θ along with the corresponding prior for $\tilde{b} = b/\sigma_O$, with the flat prior for θ following from invariance. We now discuss the details of this approximation.

D.2 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

$$E_{T\sim N(b,1)} \left(\sum_{k=1}^{K} \psi_k 1\{T = t_k\} - b \right)^2. \tag{23}$$

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

$$\mu_{kb} = \Phi\left((t_k + t_{k+1})/2 - b\right) - \Phi\left((t_k + t_{k-1})/2 - b\right), \tag{24}$$

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 (23) is therefore

$$\sum_{k=1}^{K} \psi_k^2 \mu_{kb} - 2b \sum_{k=1}^{K} \psi_k \mu_{kb} + b^2. \tag{25}$$

D.3 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 as described in (5) for each $B \in \mathcal{B}$.

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

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

whose solution is the minimax estimator $\delta^{\text{BNM}}\left(T;|\tilde{b}|\right)$. In particular, for each $|\tilde{b}| = B/\sigma_O \in \{0.1, 0.2, \ldots, 9\}$ 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 π 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 (25), 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

$$\max_{\pi\in\Delta^J} \min_{\{\psi_k\}_{k=1}^K} \sum_{\ell=1}^{J} \pi_\ell \left(\sum_{k=1}^{K} \psi_k^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^{K} \psi_k \mu_{kb_\ell} + b_\ell^2 \right). \tag{26}$$

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

Since the objective is concave in π (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.4 Computing the optimally adaptive estimator for a given ρ^2

As explained in the main text, the adaptive problem only depends on Σ through the correlation coefficient ρ^2. For a given value of ρ^2, we use convex programming methods to solve for the function $\delta^*(t;\rho)$ based on the steps described below.

1. Approximate the prior π with the finite dimensional vector $\pi \in \Delta^J$, where the parameter space for b/σ_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 (25), 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

$$\max_{\pi\in\Delta^J} \min_{\{\psi_k\}_{k=1}^K} \sum_{\ell=1}^J \pi_\ell \omega_\ell \left(\sum_{k=1}^K \psi_k^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^K \psi_k \mu_{kb_\ell} + b_\ell^2 \right) + \rho^{-2} - 1 \tag{27}$$

   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 problem is then

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

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

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

This algorithm is a finite dimensional approximation to the optimization problem in Theorem 4.1(iii). While Theorem 4.1(iii) does not formally show the existence of a solution to this infinite dimensional problem, we find that the algorithm reliably converges to a global maximum, and that the least favorable prior stabilizes as the number of gridpoints

and range of the grid increase. Based on this numerical finding, we conjecture that the minimax problem in Theorem 4.1(iii) admits a least favorable prior, and that this solution can be approximated arbitrarily well using the our grid approach.

D.5 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 ρ^2 approaches one, the solution becomes sensitive to small changes in ρ. The uneven spacing of the ρ grid allows for more accurate interpolation based on the simple pre-tabulated lookup table that we describe next.

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

D.6 Computing the analytic adaptive estimators

To find the analytic adaptive estimators in the class of ERM estimators, soft thresholding estimators and hard thresholding estimators, it suffices to solve the two dimensional minimax problem in threshold λ and scaled bias level $\tilde{b}$. We provide details for the claim in the main text that this two dimensional minimax problem can be easily solved even though the minimax theorem does not apply to these restricted classes of estimators. To simplify the computation of the analytic adaptive estimator in practice, we pre-calculate the adaptive thresholds λ 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, for example, soft-thresholding estimator $\delta_{S,\lambda}(T_O; \rho)$ for a given application, we conduct a spline interpolation across ρ^2 values to tailor the threshold to the exact parameter values desired. For example, we obtain $\delta_{S,\lambda}(T_O; -0.524)$ firstly based on spline interpolation at $\rho^2 = (-0.524)^2$ to obtain the threshold λ, and then with the observed test statistic T_O.

The derivation for soft and hard thresholding 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

$$\int_a^b x^2\phi(x)dx = \Phi(b) - \Phi(a) - (b\phi(b) - a\phi(a)). \tag{28}$$

D.6.1 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

$$\begin{aligned}
& E_{T_O\sim N(\tilde{b},1))}\left(\delta_{S,\lambda}(T_O) - \tilde{b}\right)^2 \\
= \; & E_{T_O\sim N(\tilde{b},1)}\left(\mathbf{1}\{T_O > \lambda\}\left(T_O - \lambda - \tilde{b}\right) + \mathbf{1}\{T_O < -\lambda\}\left(T_O + \lambda - \tilde{b}\right) - \mathbf{1}\{-\lambda < T_O < \lambda\}\tilde{b}\right)^2 \\
= \; & \tilde{b}^2\left(\Phi\left(\lambda - \tilde{b}\right) - \Phi\left(-\lambda - \tilde{b}\right)\right) + \int_{\lambda-\tilde{b}}^{\infty}(x-\lambda)^2\phi(x)dx + \int_{-\infty}^{-\lambda-\tilde{b}}(x+\lambda)^2\phi(x)dx
\end{aligned} \tag{29}$$

The integrals in (29) 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 (28).

The analytic adaptive objective function

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

can now be easily solved by Matlab's `fminimax` function when the risk function is evaluated based on the simplified expression derived above, and the parameter space for $\tilde{b}$ is approximated by an equally spaced grid values spanning $[-9, 9]$ with a step size of 0.025.

D.6.2 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 (28)

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

D.6.3 Adaptive ERM

For the adaptive ERM estimator $\delta_{ERM,\lambda}(T_O) = \frac{T_O^2}{T_O^2+\lambda} \cdot T_O$, we evaluate the risk function based on 10^5 simulations draws from $T_O \sim N(\tilde{b}, 1)$ and similarly optimize λ for the analytic adaptive objective function.

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 ob-

servations. 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.

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

Table A1: Estimates of the impact of NSW job training on earnings.

	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

Notes: 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 $(A_{\max}(\mathcal{B}, \hat{\theta}) - 1) \times 100$.

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 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

$$R^*(\mathcal{C}_{(B_1,B_2)}) = \begin{cases} \Sigma_U & \text{for } (B_1, B_2) = (\infty, \infty) \\ \Sigma_U - \Sigma_{UO,2}\Sigma_{O,2}^{-1}\Sigma_{UO,2} & \text{for } (B_1, B_2) = (\infty, 0) \\ \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO} & \text{for } (B_1, B_2) = (0, 0) \end{cases} \tag{30}$$

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

$$\tilde{R}(\tilde{\mathcal{S}}(b)) = \begin{cases} \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO} & \text{if } b_1 = b_2 = 0 \\ \Sigma_U - \Sigma_{UO,2}\Sigma_{O,2}^{-1}\Sigma_{UO,2} & \text{if } b_1 \neq 0, b_2 = 0 \\ \Sigma_U & \text{if } b_1 \neq 0, b_2 \neq 0 \end{cases} \tag{31}$$

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

F.1 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 ∞.

Table A2: Pairwise adaptive estimates

	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	∞	∞	1.071	1.078	1.541
Max Regret	24%	∞	∞	7.1%	7.8%	54%
Max Regret (rel. to multivariate)	26%	∞	∞	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	∞	∞	1.05	1.055	2.051
Max Regret	16.4%	∞	∞	5%	5.5%	105%
Max Regret (rel. to multivariate)	26%	∞	∞	13.6%	14.2%	105%
Threshold					0.62	1.96

Notes: 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 $(A_{\max}(\mathcal{B}, \hat{\theta}) - 1) \times 100$. "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.

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.

Table A3: Adapting pairwise with GMM composite

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

Notes: 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 $Y_{\text{comp}} = \arg\min_\theta (Y_R - \theta)' \Sigma_R^{-1} (Y_R - \theta)$. GMM combines Y_{comp} and Y_U optimally under the assumption that Y_{comp} is unbiased. If applicable, the adaptive thresholds are reported. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \hat{\theta}) - 1) \times 100$. "Max Regret (rel. to multivariate)" refers to the worst case adaptation regret relative to the multivariate oracle in (30). The correlation coefficient between Y_U and $Y_{\text{comp}} - Y_U$ is -0.45.

F.2 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_{comp}. We then adapt between Y_U and Y_{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 set $\mathcal{B}$ of bounds B on the bias consists of the two elements 0 and ∞.

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 (30). 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 $\mathcal{B}$.

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.