Adapting to Misspecification*

Timothy B. Armstrong†, Patrick Kline‡, and Liyang Sun§

May 2023

Abstract

Empirical research typically involves a robustness-efficiency tradeoff. A researcher seeking to estimate a scalar parameter can invoke strong assumptions to motivate a restricted estimator that is precise but may be heavily biased, or they can relax some of these assumptions to motivate a more robust, but variable, unrestricted estimator. When a bound on the bias of the restricted estimator is available, it is optimal to shrink the unrestricted estimator towards the restricted estimator. For settings where a bound on the bias of the restricted estimator is unknown, we propose adaptive shrinkage estimators that minimize the percentage increase in worst case risk relative to an oracle that knows the bound. We show that adaptive estimators solve a weighted convex minimax problem and provide lookup tables facilitating their rapid computation. Revisiting five empirical studies where questions of model specification arise, we examine the advantages of adapting to—rather than testing for—misspecification.

Keywords: Adaptive estimation, Minimax procedures, Specification testing, Shrinkage, Robustness.

JEL classification codes: C13, C18.

*We thank Isaiah Andrews, Manuel Arellano, Dmitry Arkhangelsky, Stéphane Bonhomme, Bryan Graham, Lihua Lei, and Aleksey Tetenov for helpful discussions on this project. Replication code is available at `https://github.com/lsun20/MissAdapt`. The paper also benefited from the comments of conference and seminar audiences at the Aarhus Workshop Econometrics I, the CEMFI Fall Econometrics Conference, the University of Mannheim, Michigan State University, the University of Amsterdam, the University of Chicago, UC Berkeley, and ASSA 2023. Timothy Armstrong gratefully acknowledges support from National Science Foundation Grant SES-2049765. Liyang Sun gratefully acknowledges support from the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010, and Ayudas Juan de la Cierva Formación.

†USC. Email: timothy.armstrong@usc.edu

‡UC Berkeley and NBER. Email: pkline@berkeley.edu

§CEMFI. Email: lsun20@cemfi.es

1 Introduction

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful. – Box and Draper (1987)

Empirical research is typically characterized by a robustness-efficiency tradeoff. The researcher can either invoke strong assumptions to motivate an estimator that is precise, but sensitive to violations of model assumptions, or they can employ a less precise estimator that is robust to these violations. Familiar examples include the choice of whether to add a set of controls to a regression, whether to exploit over-identifying restrictions in estimation, and whether to allow for endogeneity or measurement error in an explanatory variable.

As the quote from Box and Draper illustrates, decisions of this nature are often approached with a degree of pragmatism: imposing a false restriction may be worthwhile if doing so yields improvements in precision that are not outweighed by corresponding increases in bias. While precision is readily assessed with asymptotic standard errors, the measurement of bias is less standardized. A popular informal approach is to conduct a series of "robustness exercises," whereby estimates from models that add or subtract assumptions from some baseline are reported and examined for differences. While robustness exercises of this nature can be informative, they can also be perplexing. How should the results of this exercise be used to refine the baseline estimate of the parameter of interest?

The traditional answer offered in econometrics textbooks and graduate courses is to use a specification test to select a model. Specification tests offer a form of asymptotic insurance against bias: as the degree of misspecification grows large relative to the noise in the data, the test rejects with near certainty. Yet when biases are modest, as one might expect of models that serve as useful approximations to the world, the price of this insurance in terms of increased variance can be exceedingly high.

In this paper we explore an alternative to specification testing: *adapting* to misspecification. Rather than selecting estimates from a single model, the adaptive approach combines estimates from multiple models in order to optimize a robustness-efficiency tradeoff. The robustness notion considered is the procedure's worst case risk. In the canonical case of squared error loss, the risk of relying on a potentially misspecified estimator is the sum of its variance and the square of its (unknown) bias. Contrasting a credible *unrestricted* estimator with a potentially misspecified *restricted* estimator provides a noisy estimate of the restricted estimator's bias.

At first blush, it would appear difficult to trade off a combination procedure's robustness against its variance when the bias of one of its inputs is potentially infinite. Consider, however, an oracle who knows a bound B on the magnitude of the restricted estimator's bias. Such an oracle, if sufficiently ambiguity averse, will seek an estimator that is *min-*

imax under this constraint: i.e., a function of the restricted and unrestricted estimators that minimizes worst case risk subject to the bound B. Such B*-minimax* estimators have a particularly simple structure, corresponding to a Bayes estimator utilizing a discrete least favorable prior on the restricted estimator's bias and an independent flat prior on the parameter of interest. When $B = 0$, the oracle knows that the unrestricted and restricted estimators are unbiased for the same parameter; consequently, the 0-minimax estimator amounts to the efficiently weighted Generalized Method of Moments (GMM) estimator. By contrast, when $B = \infty$, the oracle knows the restricted estimator is hopelessly biased; hence, the ∞-minimax estimator corresponds to the unrestricted estimator. For intermediate values of B, the B-minimax estimator involves a type of shrinkage of the bias estimate towards zero that is used to adjust the GMM estimator for expected biases.

Now consider a researcher who does not know a bound on the bias. To quantify the disadvantage this researcher faces relative to the oracle, we introduce the notion of *adaptation regret*, which gives the percentage increase in worst case risk an estimation procedure yields over the oracle's B-minimax procedure. Because adaptation regret depends on the true bias magnitude, it is unknown at the time of estimation. However, it is typically possible to deduce the maximal (i.e., the "worst case") adaptation regret of a procedure across all possible bias magnitudes ex-ante. Importantly, the worst case adaptation regret of a procedure can often be bounded even when the bias cannot.

Our proposal for optimizing the robustness-efficiency tradeoff is to employ an *adaptive* estimator that minimizes the worst case adaptation regret. The adaptive estimator achieves worst case risk near that of the oracle regardless of the true bias magnitude. We show that the adaptive estimator can equivalently be written as a conventional minimax estimation procedure featuring a scaled notion of risk. The adaptive estimator blends the insurance properties of specification tests with the potential for efficiency gains when the restriction being considered is approximately satisfied. Like a pre-test estimator, the risk of the adaptive estimator remains bounded as the bias grows large. When biases are modest, however, the risk of the adaptive estimator is correspondingly modest. And when biases are negligible, the adaptive estimator performs nearly as well as could be achieved if prior knowledge of the bias had been available.

We show that the adaptive estimator takes a simple functional form, amounting to a weighted average of the GMM estimator and the unrestricted estimator. The combination weights depend on a shrinkage estimate of the restricted estimator's bias. As with the B-minimax estimator, the shrinkage estimate can be viewed as a Bayes estimate of bias under a discrete least favorable prior. In contrast with the B-minimax case, however, this prior requires no input from the researcher and is robust in the sense that the risk of the procedure remains bounded as the bias grows. Another appealing feature of the prior is that it depends only on the correlation between the restricted and unrestricted estimators. Enumerating these priors over a grid of correlation coefficients, we provide a lookup table that facilitates near instantaneous computation of the adaptive combination procedure.

Though the adaptive estimator is conceptually simple and easy to compute using our automated lookup table, it is not analytic. Building on insights from Efron and Morris (1972) and Bickel (1984), we explore the potential of a soft-thresholding estimator to approximate the adaptive estimator's behavior. Interestingly, we find that optimizing the soft threshold to mimic the oracle yields worst-case regret comparable to the fully adaptive estimator, while typically delivering lower worst case risk. We also devise constrained versions of both the adaptive estimator and its soft-thresholding approximation that limit the increase in maximal risk to a pre-specified level, an extension that turns out to be important in cases where the restricted estimator is orders of magnitude more precise than the unrestricted estimator. MATLAB and R code implementing the adaptive estimator, its soft-thresholding approximation, and their risk limited variants is provided online at `https://github.com/lsun20/MissAdapt`. We also provide routines for computing B-minimax estimates, which may be useful in settings where prior information about the magnitude of biases is available.

To illustrate the advantages of adapting to—rather than testing for—misspecification, we revisit five empirical examples where questions of model specification arise. The first example, drawn from Dobkin et al. (2018), considers whether to control for a linear trend in an event study analysis. A second example from Berry et al. (1995) considers whether to exploit potentially invalid supply side instruments in demand estimation. A third example drawn from Gentzkow et al. (2011) compares a two-way fixed effects estimator that exhibits negative weights in many periods to a more variable convex weighted estimator proposed by de Chaisemartin and D'Haultfœuille (2020b). A fourth example revisits LaLonde (1986)'s seminal evaluation of the National Supported Work demonstration, pooling models utilizing experimental and non-experimental controls to obtain improved estimates of treatment effects. A final example, drawn from Angrist and Krueger (1991), considers whether to instrument for years of schooling when estimating the returns to education.

In all of the above examples, adapting between models is found to yield substantially lower worst case risk and worst case adaptation regret than selecting a single model via pre-testing. The automatic procedures developed in this paper therefore provide an attractive alternative to using specification tests to summarize robustness exercises, particularly given that pre-tests have long been criticized for also leading to selective reporting of results (Leamer, 1978; Miguel, 2021). While researchers planning prospectively (e.g., in a pre-analysis plan) to entertain multiple specifications may wish to commit ex-ante to reporting adaptive summaries of the specifications considered, consumers of statistical research can also easily compute adaptive estimates from reported point estimates, standard errors, and the correlation between estimators. We find in the majority of our examples that the restricted estimators considered are nearly efficient, suggesting that accurate adaptive estimates can often be recovered from published tables ex-post even when correlations between estimators are not reported and replication data are unavailable.

Related literature. Our analysis builds on early contributions by Hodges and Lehmann (1952) and Bickel (1983, 1984) who consider families of robustness-efficiency tradeoffs defined over pairs of nested models. The main application to misspecified models generalizes this work by considering a continuum of models, indexed by different degrees of misspecification. Our general framework also allows for other sets of parameter spaces indexed by a regularity parameter, although computational constraints limit us to low dimensional applications in practice.

We follow a large statistics literature on the problem of adaptation, defined as the search for an estimator that does "nearly as well" as an oracle with additional knowledge of the problem at hand. Adaptation has been of particular interest in the nonparametric and high dimensional statistics literature (e.g., Tsybakov, 2009; Johnstone, 2019), in which adaptive estimators mimic oracles that use knowledge of the true smoothness or sparsity structure of a regression function to pick the correct bandwidth or regressors. We focus on the case where "nearly as well as an oracle" is defined formally as "up to the smallest constant multiplicative factor," which follows the definition used in Tsybakov (1998) and leads to simple risk guarantees and statements about relative efficiency. However, we also consider in detail an important departure from this definition that further restricts worst-case risk under the unconstrained parameter space.

While the high dimensional statistics literature has mostly focused on asymptotic rates and constants, we focus on exact computation of quantities of interest in low dimensional settings. In particular, we apply methods for numerical computation of optimal procedures using least favorable priors similar to those used in the recent econometrics literature (e.g., Chamberlain, 2000; Elliott et al., 2015; Müller and Wang, 2019; Kline and Walters, 2021).

To model bias, we work within a local asymptotic misspecification framework of the sort popularized recently by Andrews et al. (2017). We note, however, that this local approximation is unnecessary in linear settings of the sort that characterize many of the applications we consider. In particular, the proposed adaptive procedures give global risk guarantees for linear estimation problems. Armstrong and Kolesár (2021) study optimal inference in such settings under a known constraint on the bias of a potentially misspecified moment condition.

A large literature considers Bayesian and empirical Bayesian schemes for either model selection or model averaging (Akaike, 1973; Mallows, 1973; Schwarz, 1978; Leamer, 1978; Hjort and Claeskens, 2003). The proposed adaptive estimator can be viewed as a Bayes estimator that utilizes a "robust" prior guaranteeing bounded influence of specification biases on risk. In contrast to recent empirical Bayesian proposals engineered for forecasting problems (e.g., Hansen, 2007; Hansen and Racine, 2012) our analysis considers a scalar estimand, which renders Stein style shrinkage arguments inapplicable.

de Chaisemartin and D'Haultfœuille (2020a) apply an empirical MSE minimization approach in a setting like ours with a scalar parameter and misspecification; they show that the maximum decrease in MSE of this approach over the unrestricted estimator is greater than the maximum increase in MSE over the unrestricted estimator. We demonstrate numerically that the risk-limited variants of our adaptive estimators also satisfy this property.

It is natural to wonder if adaptive estimators can be used to construct adaptive confidence intervals (CIs) that exhibit nearly the same length as CIs based on efficient GMM when $B = 0$, while still maintaining coverage when B is large. Unfortunately, work dating back to Low (1997) establishes that this goal cannot be achieved; see Armstrong and Kolesár (2018) for impossibility results applicable to our main examples. Hence, while it is possible to construct an estimator that closely mimics an oracle, it is not possible to construct an analogous CI that adapts to biases while maintaining uniform size control. Replacing size control with other criteria amenable to adaptation is an interesting topic that we leave for future research.

Plan for paper. The rest of the paper is organized as follows. Section 2 introduces the main concepts and notation used in this paper. Section 3 illustrates the ideas through an empirical example. Section 4 presents our main results, including methods for computing adaptive estimators. Section 5 presents empirical examples. Section 6 concludes.

2 Preliminaries

Consider a researcher who observes data or initial estimate Y taking values in a set $\mathcal{Y}$, following a distribution $P_{\theta,b}$ that depends on unknown parameters (θ, b). We use $E_{\theta,b}$ to denote expectation under the distribution $P_{\theta,b}$. While we develop many results in a general setting, our main interest is in possibly misspecified models in a normal or asymptotically normal setting.

Main example. The random variable $Y = (Y_U, Y_R)$ consists of an "unrestricted" estimator Y_U of a scalar parameter $\theta \in \mathbb{R}$ and a "restricted" estimator Y_R that is predicated upon additional model assumptions. The additional restrictions required to motivate the restricted estimator make it less robust but potentially more efficient. To capture this tradeoff, we assume that Y_U is asymptotically unbiased for θ, while Y_R may exhibit a bias of b stemming from violation of the additional restrictions. We focus on the case where Y_R is a single scalar-valued estimate, but extensions to vector-valued b are possible as well.

It will often be convenient to work with the quantity $Y_O = Y_R - Y_U$, which gives an estimate of the bias in Y_R that can be used in a test of overidentifying restrictions. We work with the large sample approximation

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

The variance matrix Σ is treated as known, which arises as a local approximation to misspecification. In practice, the asymptotic variance will typically be measured via a consistent ("misspecification robust") variance estimate. In the special case where Y_R is fully efficient the restriction $\rho\sqrt{\Sigma_U}\sqrt{\Sigma_O} = -\Sigma_O$ ensues because the unrestricted estimator equals the restricted estimator plus uncorrelated noise. As famously noted by Hausman (1978), one can compute Σ_O in this case simply by subtracting the squared standard error of the restricted estimator from that of the unrestricted estimator.

Commonly encountered examples of restricted versus unrestricted specifications include (respectively) "short" versus "long" regressions containing nested sets of covariates, estimators imposing linearity/additive separability versus "saturated" specifications, and estimators motivated by exogeneity/ignorability assumptions versus those motivated by models accommodating endogeneity.

Other settings. While our main example considers a local misspecification setting with a single restricted estimator, the proposed approach applies more generally to other adaptation problems involving an unknown regularity parameter. Section 5.4 considers an application with two restricted estimators, while Appendix B.1 considers a general setting with multiple restricted estimates.

2.1 Decision rules, loss and risk

A decision rule $\delta : \mathcal{Y} \to \mathcal{A}$ maps the data Y to an action $a \in \mathcal{A}$. The loss of taking action a under parameters (θ, b) is given by the function $L(\theta, b, a)$. While it is possible to analyze many types of loss functions in our framework, we will focus on the familiar case of estimation of a scalar parameter θ with squared error loss: $\theta \in \mathbb{R}$, $\mathcal{A} = \mathbb{R}$ and the loss function is $L(\theta, b, \hat{\theta}) = (\hat{\theta} - \theta)^2$.

The risk of a decision rule is given by the function

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

A decision δ is *minimax* over the set $\mathcal{C}$ for the parameter (θ, b) if it minimizes the maximum risk over $(\theta, b) \in \mathcal{C}$. We are interested in a setting where the researcher entertains multiple parameter spaces $\mathcal{C}_B$, indexed by $B \in \mathcal{B}$, which may restrict the parameters (θ, b) in different ways. The maximum risk over the set $\mathcal{C}_B$ is

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

A decision δ is *minimax* over $\mathcal{C}_B$ if it minimizes $R(B, \delta)$. The *minimax risk* for the parameter space $\mathcal{C}_B$ is the risk of this decision:

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

We use the term B-*minimax* as shorthand for "minimax over $\mathcal{C}_B$" and B-minimax risk for "minimax risk for the parameter space $\mathcal{C}_B$." At times, we will use "minimax" or "B-minimax" for "maximum risk of δ over $(\theta, b) \in \mathcal{C}_B$" even when δ is not actually the minimax decision.

Main example (continued). In our main example, we define $\mathcal{C}_B$ to place a bound B on the magnitude of the bias of the restricted estimator:

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

We consider the sets $\mathcal{C}_B$ for $B \in [0, \infty]$. Thus, $B = \infty$ corresponds to the unrestricted parameter space, while $B = 0$ corresponds to the restricted parameter space. It follows from the theory of minimax estimation in linear models that the ∞-minimax estimator (the B-minimax estimator when $B = \infty$) is Y_U, while the 0-minimax estimator (the B-minimax estimator when B=0) is $Y_U - (\rho\sqrt{\Sigma_U}/\sqrt{\Sigma_O})Y_O$. Inspection of this formula reveals that the 0-minimax estimator is the efficient GMM estimator exploiting the restriction $b = 0$. In the special case where the restricted estimator is fully efficient, the 0-minimax estimator is additionally equal to the restricted estimator $Y_R = Y_U + Y_O$.

2.2 Adaptation

The B-minimax risk gives a benchmark for how well one can do using only the constraint $(\theta, b) \in \mathcal{C}_B$. To calculate the B-minimax estimator achieving this benchmark, the researcher must specify an appropriate parameter space $\mathcal{C}_B$. In our main example, the parameter spaces are indexed by an a priori bound on the magnitude $|b|$ of the constrained estimator's bias.

How much must one give up in order to avoid specifying B? Consider an estimator δ formed without reference to a particular parameter space $\mathcal{C}_B$. Relative to an oracle that knows B and is able to compute the B-minimax estimator, δ yields a proportional increase in worst-case risk over $\mathcal{C}_B$ given by

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

We refer to $A(B, \delta)$ as the *adaptation regret* of the estimator δ under the set $\mathcal{C}_B$. This regret may be as large as $A_{\max}(\mathcal{B}, \delta) = \sup_{B \in \mathcal{B}} A(B, \delta)$, a quantity we term the *worst case adaptation regret.* The lowest possible value $A_{\max}(\mathcal{B}, \delta)$ can take is given by

$$A^*(\mathcal{B}) = \inf_{\delta} \sup_{B \in \mathcal{B}} A(B, \delta) = \inf_{\delta} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta)}{R^*(B)}. \tag{1}$$

Following Tsybakov (1998) $A^*(\mathcal{B})$ gives the *loss of efficiency under adaptation.* An estimator δ is *optimally adaptive* if $A_{\max}(\mathcal{B}, \delta) = A^*(\mathcal{B})$. We use the notation δ^{adapt} to denote such an estimator. To measure the efficiency of an ad hoc estimator δ relative to the optimally adaptive estimator, one can compute

$$\frac{A^*(\mathcal{B})}{A_{\max}(\mathcal{B}, \delta)} = \frac{\inf_{\delta} A_{\max}(\mathcal{B}, \delta)}{A_{\max}(\mathcal{B}, \delta)}.$$

We refer to this quantity as the *adaptive efficiency* of the estimator δ.

Main example (continued). In our main example, $\mathcal{C}_B = \mathbb{R} \times [-B, B]$, and we seek estimators that perform well even in the worst case when $B = \infty$. Thus, we take the set of values of B under consideration to be $\mathcal{B} = [0, \infty]$.

Granular $\mathcal{B}$. Bickel (1984) considered adapting over the finite set $\mathcal{B}^{gran} = \{0, \infty\}$. Naturally, it is easier to adapt to the elements of $\mathcal{B}^{gran}$ than to the infinite set $\mathcal{B} = [0, \infty]$. Consequently, $A^*(\mathcal{B}^{gran}) \leq A^*(\mathcal{B})$. However, consideration of $\mathcal{B}^{gran}$ may leave efficiency gains on the table for $0 < b < \infty$ because $R^*(b) \leq R^*(\infty)$.

Note that $A(B, \delta)^{-1} = R^*(B)/R_{\max}(B, \delta)$ gives the *relative efficiency* of the estimator δ under the minimax criterion for parameter space $\mathcal{C}_B$, according to the usual definition. Thus, the optimally adaptive estimator obtains the best possible relative efficiency that can be obtained simultaneously for all $B \in \mathcal{B}$. The loss of efficiency under adaptation gives the reciprocal of this best possible simultaneous relative efficiency. Bickel (1982) studied an asymptotic regime where $A(B, \delta^{adapt})$ tended to one, implying no asymptotic loss of efficiency under adaptation.

2.3 Discussion

Fundamentally, an optimally adaptive estimator is one that is "nearly B-minimax" for all $B \in \mathcal{B}$, a notion that accords closely with the usual definitions in the literature (e.g., Tsybakov, 1998, 2009; Johnstone, 2019). The definition in (1) operationalizes "near" as "up to the smallest uniform multiplicative factor," which provides an intuitive link between statements about adaptation and relative efficiency. However, the approach developed in this paper is easily extended to other definitions of near, such as the smallest absolute distance from the relevant B-minimax risk. In Section 4.5 we also consider an extension that places a bound on worst-case risk relative to the unbiased estimator.

Adaptive estimators, like their minimax antecedents, provide convenient alternatives to Bayesian estimation that avoid the requirement to fully specify a prior. It is well known that minimax strategies can be justified on decision theoretic grounds by various axiomatizations of ambiguity aversion (Gilboa and Schmeidler, 1989; Schmeidler, 1989). Adaptation regret can be thought of as capturing the regret an ambiguity averse researcher feels over having exposed themselves to an unnecessarily high level of worst case risk, regardless of what losses were actually realized.

A different sort of justification for minimax decisions—attributable to Savage (1954)—involves the potential of such decisions to foster consensus in settings where priors differ among members of a group. In Appendix A we develop a stylized extension of Savage (1954)'s argument that illustrates the ability of adaptive decisions to foster consensus among "committees" characterized by different sets of beliefs. Taking the committees to represent different camps of researchers, the model suggests adaptive estimation can help to forge consensus between researchers with varying beliefs about the suitability of different econometric models. In accord with the notion that the desirability of an optimally adaptive decision derives from its resemblance to the relevant B-minimax decision, the model suggests the prospects for achieving consensus decrease with the loss of efficiency under adaptation $A^*(\mathcal{B})$.

3 An Illustration

To build some intuition for B-minimax and optimally adaptive estimators, we consider an example drawn from Dobkin et al. (2018) concerning whether to detrend a quasi-experimental estimator of treatment effects. In this case Y_R corresponds to a two-way fixed effects estimator of the effect of unexpected hospitalization on medical spending, while Y_U corresponds to a linearly detrended estimate of the same quantity. We return to this example in Section 5 where further details on the econometric specification under consideration are provided.

The B-minimax and optimally adaptive estimators are depicted in Figure 1. Both estimators have been computed numerically assuming squared error loss, implying risk is given by mean squared error (MSE). The first y-axis reports point estimates of θ, which is measured in dollars. Realized values of Y_R, Y_U, the efficient GMM estimator, and the optimally adaptive estimator are depicted by horizontal lines. Realized values of the B-minimax estimators are plotted as triangles.

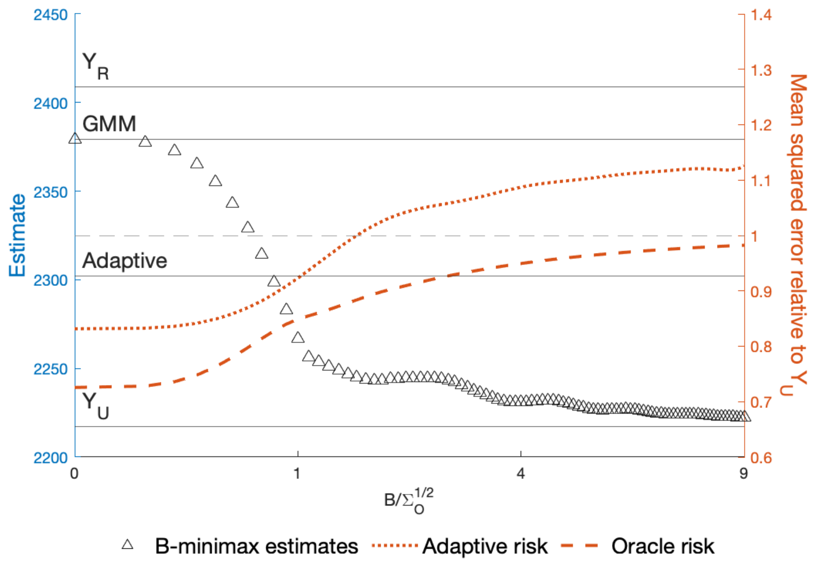

Figure 1: B-minimax and adaptive estimators

In this example Y_R is not fully efficient, leading the GMM estimator to place positive weight on Y_U. When $B = 0$, the B-minimax estimator coincides with efficient GMM. As B grows, the B-minimax estimator scales nonlinearly towards Y_U, reflecting the tradeoff between robustness and efficiency. The adaptive estimator lies roughly halfway between the efficient GMM estimate and the realized value of Y_U, coming very close ex-post to the B-minimax estimate that arises when $B = \Sigma_O^{1/2}$.

The second y-axis of Figure 1 measures worst case MSE scaled in terms of Σ_U (i.e., in terms of the risk of Y_U). The dashed line gives the worst case risk of an oracle that knows the bound B and computes the B-minimax estimator. When $B = 0$ the B-minimax oracle achieves a sizable 27% worst case MSE reduction relative to Y_U. As B grows large, the minimax risk of the B-minimax oracle converges with that of Y_U. Hence, by exploiting prior knowledge of the bound B, the oracle can obtain an estimator with risk weakly lower than Y_U.

The adaptive estimator tries to limit worst case risk without prior knowledge of B. The worst case risk of the optimally adaptive estimator is given by the dotted line, which follows a profile mimicking that of the B-minimax oracles. The price of not knowing the bound B in advance is that the worst case risk of the adaptive estimator lies everywhere above that of the corresponding oracle's risk. Fortunately, the worst case risk of δ^{adapt} remains bounded as B approaches infinity. In fact, the adaptation regret $A(B, \delta^{adapt})$ is nearly constant in the oracle bound B. Consequently, the adaptation regret associated with not having used Y_U when $B/\Sigma_O^{1/2} = 9$ roughly equals the adaptation regret associated with not having used GMM when $B = 0$. Moreover, the reduction in risk relative to Y_U when $B = 0$ exceeds the increase in worst-case risk relative to Y_U when $B/\Sigma_O^{1/2} = 9$, a property emphasized by de Chaisemartin and D'Haultfœuille (2020a).

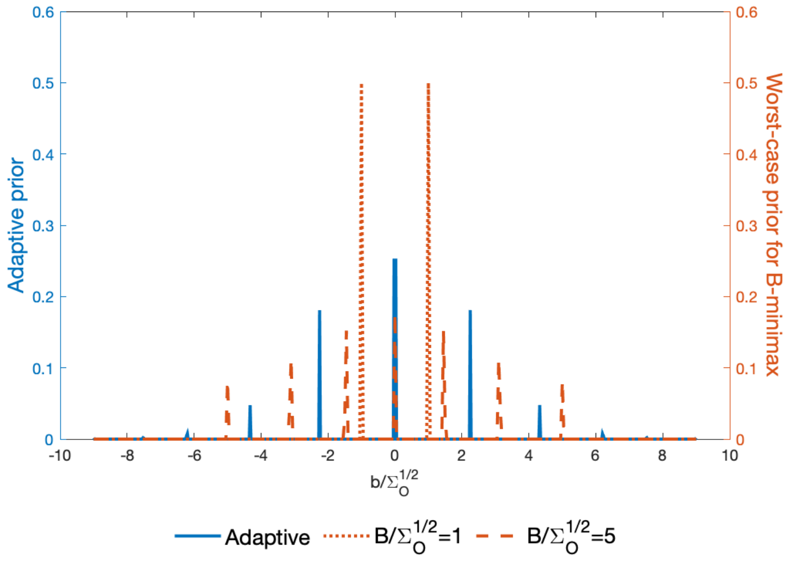

Figure 2: Least favorable priors when $\rho = -0.524$

As we show in the next section, both the adaptive estimator and its B-minimax antecedents can be thought of as Bayes estimators motivated by particular least favorable priors. Figure 2 depicts the least favorable priors utilized by the B-minimax estimator for two values of B along with the least favorable prior of the adaptive estimator. These distributions depend on the data only through the estimated value of ρ, which takes the value -0.524 in this example. All three priors on $b/\Sigma_O^{1/2}$ are discrete, symmetric about zero, and decreasing in $|b|$. Hence, all three estimators will tend to be more efficient than Y_U when the true bias magnitude $|b|$ is small. The adaptive prior has the important advantage over B-minimax priors of not requiring specification of the bound B. A second advantage of the adaptive prior is that it is *robust*: the risk of δ^{adapt} remains bounded as $|b|$ grows large. In contrast, the risk of a B-minimax estimator grows rapidly and without limit once $|b|$ exceeds the posited bound B.

4 Main results

Computing the optimally adaptive estimator requires solving (1). As we now show, this task amounts to solving a minimax problem with a scaled loss function, thereby allowing us to leverage results from the literature on computation of minimax estimators.

4.1 Adaptation as minimax with scaled loss

Plugging in the definition of $R_{\max}(B, \delta)$, the criterion that the optimally adaptive estimator δ^{adapt} minimizes can be written

$$\sup_{B\in\mathcal{B}} \frac{R_{\max}(B,\delta)}{R^*(B)} = \sup_{B\in\mathcal{B}} \sup_{(\theta,b)\in\mathcal{C}_B} \frac{R(\theta,b,\delta)}{R^*(B)} = \sup_{(\theta,b)\in\cup_{B'\in\mathcal{B}}\mathcal{C}_{B'}} \ \sup_{B\in\mathcal{B} \text{ s.t. } (\theta,b)\in\mathcal{C}_B} \frac{R(\theta,b,\delta)}{R^*(B)}$$

where the last equality follows by noting that the double supremum on either side of this equality is over the same set of values of (B, θ, b). Letting

$$\omega(\theta,b) = \left(\inf_{B\in\mathcal{B} \text{ s.t. } (\theta,b)\in\mathcal{C}_B} R^*(B)\right)^{-1}, \tag{2}$$

we obtain the following lemma.

Lemma 4.1. *The loss of efficiency under adaptation (1) is given by*

$$A^*(\mathcal{B}) = \inf_{\delta} \sup_{(\theta,b)\in\cup_{B'\in\mathcal{B}}\mathcal{C}_{B'}} \omega(\theta,b)R(\theta,b,\delta)$$

and a decision δ^{adapt} that achieves this infimum (if it exists) is optimally adaptive.

Lemma 4.1 shows that finding an optimally adaptive decision can be written as a minimax problem with a weighted version of the original loss function. In particular, δ is found to minimize the maximum (over θ, b) of the objective $\omega(\theta,b)R(\theta,b,\delta) = E_{\theta,b}\omega(\theta,b)L(\theta,b,\delta(Y))$. Hence, the optimal adaptive estimator corresponds to a minimax estimator under the loss function $\omega(\theta,b)L(\theta,b,\delta(Y))$. Of course, $\omega(\theta,b)$ must be computed, but this also amounts to computing a family of minimax problems.

Main example (continued). In our main example, the sets $\mathcal{C}_B = \mathbb{R} \times [-B, B]$ are nested so that $R^*(B)$ is increasing in B and $\omega(\theta,b) = R^*(|b|)^{-1}$.

To summarize, provided that we have a general method for constructing minimax estimators, the optimally adaptive estimator can be computed via the following algorithm.

Algorithm 4.1 (General computation of optimally adaptive estimator)**.**

Input Set of parameter spaces $\mathcal{C}_B$, loss function, (Y, Σ) as described in Section 2, along with a generic method for computing minimax estimators

Output Optimally adaptive estimator δ^{adapt} and loss of efficiency under adaptation $A^*(\mathcal{B})$

1. Compute the minimax risk $R^*(B)$ for each $B \in \mathcal{B}$ and use this to form the weight $\omega(\theta, b)$ as in (2).

2. Form the loss function $(\theta, b, a) \mapsto \omega(\theta, b)L(\theta, b, a)$. Compute the optimally adaptive estimator δ^{adapt} as the minimax estimator under the parameter space $\cup_{B\in\mathcal{B}}\mathcal{C}_B$, and compute the loss of efficiency under adaptation $A^*(\mathcal{B})$ as the corresponding minimax risk.

4.2 Computing minimax estimators

Algorithm 4.1 allows us to compute adaptive estimators once we have a generic method for solving minimax estimation problems. A typical approach to this problem is to use the insight that the minimax estimator can often be characterized as a Bayes estimator for a *least favorable prior*. Such estimators can be formulated as solving a convex optimization problem over distributions on (θ, b) that can be evaluated numerically using discretization or other approximation techniques so long as the dimension of (θ, b) is sufficiently low (see Chamberlain (2000), Elliott et al. (2015), Müller and Wang (2019) and Kline and Walters (2021) for recent applications in econometrics).

We now summarize the relevant ideas as they apply to our general setup. In the next subsection, we use the fact that in our main example the minimax and adaptive estimators are invariant to certain transformations to reduce the problem to finding a least favorable prior over b, with a flat (improper) prior on θ. Details on the choices made to evaluate the estimators numerically are provided in Appendix C.

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. To implement Algorithm 4.1, $\mathcal{C}_B$ plays the role of $\mathcal{C}$ and $L(\theta, b, \delta)$ plays the role of $\bar{L}(\vartheta, \delta)$ for a B on a grid approximating $\mathcal{B}$. We then solve this problem with $\cup_{B\in\mathcal{B}}\mathcal{C}_B$ playing the role of $\mathcal{C}$ and $\omega(\theta, b)L(\theta, b, \delta)$ playing the role of $\bar{L}(\vartheta, \delta)$.

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 applications, we approximate π with the finite dimensional vector $(\pi(\vartheta_1), \ldots, \pi(\vartheta_J))$ for a grid of J values of ϑ, following Chamberlain (2000).

4.3 Adaptive estimation in main example

In our main example, we use invariance to further simplify the problem before applying the methods for computing minimax estimators in Section 4.2. We focus in the main text on the case of squared error loss $L(\theta, b, \delta) = (\theta - \delta)^2$. Appendix B.1 provides proofs of the results in this section and includes general loss functions for estimation of the form $L(\theta, b, \delta) = \ell(\theta - \delta)$.

It will be useful to transform the data to Y_U, T_O where $T_O = Y_O/\sqrt{\Sigma_O}$ is the t-statistic for a specification test of the null that $b = 0$. We observe

$$\begin{pmatrix} Y_U \\ T_O \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ b/\sqrt{\Sigma_O} \end{pmatrix}, \begin{pmatrix} \Sigma_U & \rho\sqrt{\Sigma_U} \\ \rho\sqrt{\Sigma_U} & 1 \end{pmatrix} \right). \tag{3}$$

where Σ_U, Σ_O and $\rho = \text{corr}(Y_U, T_O) = \text{corr}(Y_U, Y_O)$ are treated as known. This representation is equivalent to our original setting, as Σ_O is known and can be used to transform T_O to Y_O.

Applying invariance arguments and the Hunt-Stein theorem, it follows that the B-minimax estimator $\delta_B^*(Y_U, T_O)$ takes the form

$$\rho\sqrt{\Sigma_U}\delta\left(T_O\right) + Y_U - \rho\sqrt{\Sigma_U}T_O. \tag{4}$$

To build some intuition for this expression, note that $Y_U - \rho\sqrt{\Sigma_U}T_O$ is the optimal GMM estimator of θ under the restriction $b = 0$. When $\rho\sqrt{\Sigma_O}\sqrt{\Sigma_U} = -\Sigma_O$, optimal GMM reduces to the restricted estimator Y_R, which is efficient in this case. If $b \neq 0$, then GMM will exhibit a bias of $-\frac{\rho\sqrt{\Sigma_U}}{\sqrt{\Sigma_O}}b$. The estimator in (4) subtracts from the GMM estimate a corresponding estimate $-\rho\sqrt{\Sigma_U}\delta\left(\frac{Y_O}{\sqrt{\Sigma_O}}\right)$ of this bias term.

The $\delta\left(T_O\right)$ employed by the B-minimax estimator can be shown to evaluate to the *bounded normal mean* estimator $\delta^{\text{BNM}}\left(T_O; \frac{B}{\sqrt{\Sigma_O}}\right)$, where $\delta^{\text{BNM}}(y; \tau)$ denotes the minimax estimator of $\vartheta \in \mathcal{C} = [-\tau, \tau]$ when $Y \sim N(\vartheta, 1)$. The bounded normal mean problem has been studied extensively (see, e.g., Lehmann and Casella, 1998, Section 9.7(i), p. 425) and we detail its computation in Appendix C.2. The corresponding B-minimax risk is

$$R^*(B) = \rho^2 \Sigma_U r^{\text{BNM}} \left(\frac{B}{\sqrt{\Sigma_O}} \right) + \Sigma_U - \rho^2 \Sigma_U, \tag{5}$$

where $r^{\text{BNM}}(\tau)$ denotes minimax risk in the bounded normal mean problem. This expression was used to construct the oracle risk curve displayed in Figure 1. We evaluate $r^{\text{BNM}}(\tau)$ numerically by computing a least favorable prior on a grid approximating $[-\tau, \tau]$, following the methods described in Section 4.2 above.

The scaling function (2) can now be written $\omega(\theta, b) = R^*(|b|)$, where R^* for our problem is given in (5). To compute the optimally adaptive estimator for squared error loss, it therefore suffices to compute the minimax estimator for θ under the scaled loss function $R^*(|b|)^{-1}(\theta - \delta)^2$. Invariance arguments can again be applied to show that the optimally adaptive estimator takes the same form as in (4), but with δ given by the estimator $\tilde{\delta}^{\text{adapt}}(t; \rho)$, which minimizes

$$\sup_{\tilde{b} \in \mathbb{R}} \frac{E_{T \sim N(\tilde{b},1)}(\tilde{\delta}(T) - \tilde{b})^2 + \rho^{-2} - 1}{r^{\text{BNM}}(|\tilde{b}|) + \rho^{-2} - 1} \tag{6}$$

and we detail its computation in Appendix C.3. The loss of efficiency under adaptation $A^*([0, \infty])$ is then given by the minimized value of (6). Computation is performed by searching for a least favorable prior over $\tilde{b}$ on a grid approximation of $[-K, K]$ for a large value K. The least favorable prior for $\tilde{b}$ corresponds to a prior on $b/\sqrt{\Sigma_O}$, and the invariance arguments for θ lead to a flat (improper) prior for θ.

We summarize these results in the following theorem, which is proved in Appendix B.1.

Theorem 4.1. *Consider our main example, given by the model in (3) with parameter spaces $\mathcal{C}_B = \mathbb{R} \times [-B, B]$ for $B \in \mathcal{B} = [0, \infty]$ and squared error loss $L(\theta, b, d) = (d - \theta)^2$. The following results hold:*

(i) The B-minimax estimator takes the form in (4) with $\delta(\cdot)$ given by $\delta^{\text{BNM}}\left(\cdot; \frac{B}{\sqrt{\Sigma_O}}\right)$ and the minimax risk $R^*(B)$ is given by (5).

(ii) An optimally adaptive estimator is given by (4) with $\delta(\cdot)$ given by a function $\tilde{\delta}^{\text{adapt}}(t; \rho)$ that minimizes (6).

(iii) The loss of efficiency under adaptation is

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

where the supremum is over all probability distributions π on $\mathbb{R}$.

4.3.1 Weighted average interpretation

One can write the estimator in (4) as a weighted average:

$$w(T_O) \cdot Y_U + (1 - w(T_O)) \cdot \underbrace{(Y_U - \rho\sqrt{\Sigma_U} \cdot T_O)}_{\text{Optimal GMM}}, \tag{7}$$

where $w(T_O) = \delta(T_O)/T_O$ is a data-dependent weight. The B-minimax estimator takes $\delta(\cdot)$ to be a minimax estimator that uses the constraint $|b| \leq B$ with known B, whereas the optimally adaptive estimator takes as $\delta(\cdot)$ an estimator engineered to adapt to different values of B in this constraint. As detailed in Appendix C.4, we find numerically that the adaptive estimator "shrinks" T_O towards zero, leading the weight $\delta(T_O)/T_O$ to fall between zero and one for all values of ρ.

The data dependent nature of the weight $w(T_O)$ is clearly crucial for the robustness properties of the optimally adaptive estimator. As T_O grows large, less weight is placed on the optimal GMM estimator and more weight is placed on the unrestricted estimator Y_U. If one were to commit ex-ante to a fixed (i.e., non-stochastic) weight on Y_U, the worst-case risk of the procedure would become unbounded as the optimal GMM estimator can exhibit arbitrarily large bias. Consequently, worst case adaptation regret would also become unbounded.

4.3.2 Impossibility of consistently estimating the asymptotic distribution

Recall that (3) provides the asymptotic distribution of (Y_U, T_O) under local misspecification. In this asymptotic regime, b gives the limit of the bias of the restricted estimator divided by $\sqrt{n}$ and cannot be consistently estimated. In contrast, consistent estimates for ρ and Σ_U are available via the usual asymptotic variance formulas used in overidentification tests for GMM.

To obtain the sampling distribution of the optimally adaptive estimator, one can plug the distribution of (Y_U, T_O) stipulated in (3) into expression (7). Unfortunately, this distribution cannot be consistently estimated, as it depends on the local asymptotic bias b. For instance, the asymptotic variance of the optimally adaptive estimator δ^{adapt} takes the form $\rho^2\Sigma_U v(b/\sqrt{\Sigma_O}) + \Sigma_U - \rho^2\Sigma_U$, where $v(\tilde{b}) = \text{var}_{T_O \sim N(\tilde{b},1)}(\tilde{\delta}^{\text{adapt}}(T_O;\rho))$ denotes the variance of $\tilde{\delta}^{\text{adapt}}(T_O;\rho))$ when $T_O \sim N(\tilde{b}, 1)$. Because $\tilde{\delta}^{\text{adapt}}(T_O;\rho)$ is a nonlinear function of T_O, this variance formula is a nonconstant function of b. Since b cannot be consistently estimated, it is not possible to consistently estimate the asymptotic variance of δ^{adapt}. See Leeb and Pötscher (2005) for a discussion of these issues in the context of pre-test estimators. Related arguments (Low, 1997; Armstrong and Kolesár, 2018) establish the impossibility of constructing adaptive CIs.

When b is given, one can construct consistent estimates of the sampling distribution of the adaptive estimator, which is useful for assessing its theoretical risk properties. In particular, the mean squared error of the estimator (4) is given by

$$\rho^2\Sigma_U r(b/\sqrt{\Sigma_U}) + \Sigma_U - \rho^2\Sigma_U \quad \text{where} \quad r(\tilde{b}) = E_{T\sim N(\tilde{b},1)}(\delta(T) - \tilde{b})^2.$$

In our applications, we report these asymptotic risk functions by plotting them as a function of b.

4.3.3 Lookup table

To ease computation of the optimally adaptive estimator, we solved for the function $\tilde{\delta}^{\text{adapt}}(t;\rho)$ numerically at a grid of values of the scalar parameter ρ using convex programming methods, the details of which are provided in Appendix C.4. Tabulating these solutions yields a simple lookup table that allows rapid retrieval of the empirically relevant function. Computation of the final estimator is extremely fast, taking only milliseconds to implement.

4.4 Simple "nearly adaptive" estimators

While the optimally adaptive estimator is straightforward to compute via convex programming and is trivial to implement once the solution is tabulated, it lacks a simple closed form. To reduce the opacity of the procedure, one can replace the term $\delta(T_O)$ in (4) with an analytic approximation.

A natural choice of approximations for $\delta(T_O)$ is the class of *soft-thresholding* estimators, which are indexed by a threshold $\lambda \geq 0$ and given by

$$\delta_{S,\lambda}(T) = \max\{|T| - \lambda, 0\}\,\text{sgn}(T) = \begin{cases} T - \lambda & \text{if } T > \lambda \\ T + \lambda & \text{if } T < -\lambda \\ 0 & \text{if } |T| \leq \lambda, \end{cases}$$

which leads to the estimator

$$\rho\sqrt{\Sigma_U}\delta_{S,\lambda}(T_O) + Y_U - \rho\sqrt{\Sigma_U}T_O = \begin{cases} Y_U - \rho\sqrt{\Sigma_U}\lambda & \text{if } T_O > \lambda \\ Y_U + \rho\sqrt{\Sigma_U}\lambda & \text{if } T_O < -\lambda \\ Y_U - \rho\sqrt{\Sigma_U}T_O & \text{if } |T_O| \leq \lambda. \end{cases}$$

We also consider the class of *hard-thresholding* estimators, which are given by

$$\delta_{H,\lambda}(T) = T \cdot I(|t| \geq \lambda) = \begin{cases} T & \text{if } |T| > \lambda \\ 0 & \text{if } |T| \leq \lambda, \end{cases}$$

which leads to the estimator

$$\rho\sqrt{\Sigma_U}\delta_{H,\lambda}(T_O) + Y_U - \rho\sqrt{\Sigma_U}T_O = \begin{cases} Y_U & \text{if } |T_O| > \lambda \\ Y_U - \rho\sqrt{\Sigma_U}T_O & \text{if } |T_O| \leq \lambda. \end{cases}$$

Note that hard-thresholding leads to a simple pre-test rule: use the unrestricted estimator if $|T_O| > \lambda$ (i.e. if we reject the null that $b = 0$ using critical value λ) and otherwise use the GMM estimator that is efficient under the restriction $b = 0$. The soft-thresholding estimator uses a similar idea, but avoids the discontinuity at $T_O = \lambda$.

To compute the hard and soft-thresholding estimators that are optimally adaptive in these classes of estimators, we minimize (6) numerically over λ. The minimax theorem does not apply to these restricted classes of estimators. Fortunately, however, the resulting two dimensional minimax problem in λ and $\tilde{b}$ is easily solved in practice as explained in Appendix C.5. The optimized value of (6) then gives the worst-case adaptation regret of the optimally adaptive soft or hard-thresholding estimator.

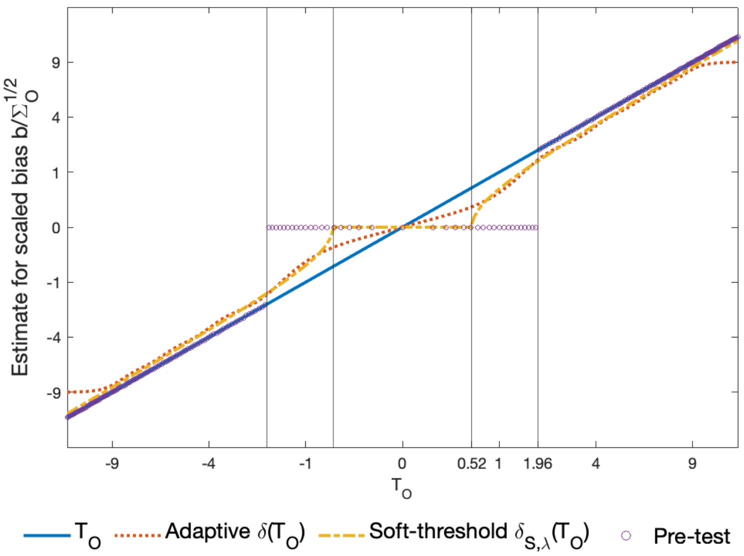

Figure 3: Estimators of scaled bias when $\rho = -0.524$

Figure 3 plots the optimally adaptive and soft-thresholding estimators of the scaled bias as functions of T_O. These functions depend on the data only through the estimated value of ρ, which takes the value -0.524 here, as in the two-way fixed effects example introduced in Section 3. The optimal soft-threshold λ yielding the lowest worst cast adaptation regret in this example is 0.52. Both the adaptive and soft-thresholding estimators continously shrink small values of T_O towards zero. However, the soft-thresholding estimator sets all values of $|T_O|$ less than 0.52 to zero, while the optimally adaptive estimator avoids flat regions. In contrast to the continuous nature of these two adaptive estimators, a conventional pre-test using $\lambda = 1.96$ exhibits large discontinuities at the hard threshold.

Like the optimally adaptive estimator δ^{adapt}, the worst-case adaptation regret of the optimally adaptive soft and hard-thresholding estimators depends only on ρ. We report comparisons between these estimators in our empirical applications in Section 5 and provide a more detailed analysis in Appendix B.3. As discussed in Appendix B.3, soft-thresholding yields nearly optimal performance for the adaptation problem relative to δ^{adapt} in a wide range of settings. In contrast, hard-thresholding typically exhibits both substantially elevated worst case adaptation regret and worst case risk driven by the possibility that the scaled bias has magnitude near λ. In Appendix B.4 we consider the behavior of these adaptive estimators as $|\rho| \to 1$ and show that the worst-case adaptation regret of δ^{adapt}, as well as the optimally adaptive soft and hard-thresholding estimators, increases at a logarithmic rate.

These conclusions mirror the findings of Bickel (1984) for the case where the set $\mathcal{B}$ of bounds B on the bias consists of the two elements 0 and ∞. When $|\rho|$ is close to 1, using the constraint $b = 0$ leads to a very large efficiency gain relative to the unconstrained estimator. As $|\rho| \to 1$, it become increasingly difficult to achieve this large efficiency gain when b is small while retaining robustness to large values of b. This dilemma leads to increasing loss of efficiency under adaptation for $|\rho|$ near 1. In particular, the optimally adaptive estimator exhibits increasing worst-case risk relative to Y_U as $|\rho| \to 1$ (see Lemma B.3). In such settings, it may be desirable to resolve this tradeoff in different ways, a topic we turn to in the next section.

4.5 Constrained adaptation

If the loss of efficiency under adaptation $A^*(\mathcal{B})$ is large, then we face a nontrivial decision about which parameter space $\mathcal{C}_B$ to use. One way of resolving this tradeoff is to impose an upper bound on the increase in maximum risk over the union of parameter spaces $\cup_{B\in\mathcal{B}}\mathcal{C}_B$. This leads to the problem

$$A^*(\mathcal{B}; \overline{R}) = \inf_{\delta} \sup_{B\in\mathcal{B}} \frac{R_{\max}(B,\delta)}{R^*(B)} \quad \text{s.t.} \quad \sup_{B\in\mathcal{B}} R_{\max}(B,\delta) \le \overline{R}, \tag{8}$$

where $\overline{R}$ is a constraint on the maximum risk over the union of the parameter spaces $\cup_{B\in\mathcal{B}}\mathcal{C}_B$. We can relate this to 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

$$\inf_{\delta} \sup_{B \in \mathcal{B}} \max \left\{ \frac{R_{\max}(B, \delta)}{R^*(B)}, \frac{R_{\max}(B, \delta)}{t} \right\} = \inf_{\delta} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta)}{\min\{R^*(B), t\}}. \tag{9}$$

Indeed, any solution to (8) must also be a solution to (9) with $t = \overline{R}/A^*(\mathcal{B}; \overline{R})$, since any decision function achieving a strictly better value of (9) would satisfy the constraint in (8) and achieve a strictly better value of the objective in (8). Conversely, letting $\tilde{A}^*(t)$ be the value of (9), any solution to (9) will achieve the same value of the objective (8) and will satisfy the constraint for $\bar{R} = t \cdot \tilde{A}^*(t)$. In fact, this solution to (9) will also solve (8) 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 (see Lemma 4.2 below).

Arguing as in Section 4.1, we can write the optimization problem (9) as

$$\inf_{\delta} \sup_{(\theta, b) \in \cup_{B' \in \mathcal{B}} \mathcal{C}_{B'}} \tilde{\omega}(\theta, b, t) R(\theta, b, \delta), \tag{10}$$

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

and $\omega(\theta, b)$ is given in (2) in Section 4.1. Thus, we can solve (9) 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 (8) by finding t such that $\bar{R} = t \cdot A^*(t)$.

We summarize these results in the following lemma, which is proved in Section B.2 of the appendix.

Lemma 4.2. *Any solution to (8) is also a solution to (10) with* $t = \overline{R}/A^*(\mathcal{B}; \overline{R})$. *Conversely, let* $\tilde{A}^*(t)$ *denote the value of (10) and let* $\tilde{R}(t) = \tilde{A}^*(t) \cdot t$. *If* $\tilde{R}(t) > \inf_\delta \sup_{B \in \mathcal{B}} R_{\max}(B, \delta)$ *and* $\inf_{B \in \mathcal{B}} R^*(B) > 0$, *then* $A^*(\mathcal{B}; \tilde{R}(t)) = \tilde{A}^*(t)$ *and any solution to (10) is also a solution to (8) with* $\bar{R} = \tilde{R}(t)$.

How should the bound $\overline{R}$ on worst-case risk be chosen? This choice depends on how one trades off efficiency when b is small against robustness when b is large. As noted by Bickel (1984) in his analysis of the granular case where $\mathcal{B} = \{0, \infty\}$, it is often possible to greatly improve the risk at $b = 0$ relative to the unbiased estimator Y_U in exchange for modest increases in risk in the worst case. Similarly, we find that moderate choices of $\overline{R}$ equal to 20% or 50% above the risk of Y_U yield large efficiency improvements in our applications when b is small.

One way of measuring these tradeoffs, suggested by de Chaisemartin and D'Haultfœuille (2020a), is to look for an estimator where the best-case decrease in risk relative to Y_U is greater than the worst-case increase in risk over Y_U. We show numerically in Appendix B.3 that this property holds for the constrained soft-thresholding version of our estimator so long as $\overline{R}$ is less than 70% above the risk of Y_U, and that it holds even for unconstrained soft-thresholding ($\overline{R} = \infty$) when ρ^2 is less than 0.86. The optimally adaptive estimator exhibits similar properties: depictions of its performance as a function of ρ^2—both when unconstrained and when $\overline{R}$ is set at 120% of the risk of Y_U—are provided in Appendix Figure A5.

Our approach can also be generalized to explore other ways of trading off risk across different values of b or different parameter spaces $\mathcal{C}_B$. The constrained adaptation problem (9) can be interpreted as an adaptation problem that places weights on the parameter spaces $\mathcal{C}_B$ under consideration by rewriting it as

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

where $g(B) = \max \{1, R^*(B)/t\}$ is a weight on the parameter space $\mathcal{C}_B$. One can use different weight functions $g(B)$ to trade off risk in different ways.

5 Examples

We now consider a series of examples where questions of specification arise and examine how adapting to misspecification compares to pre-testing and other strategies such as committing ex-ante to either the unrestricted or restricted estimator. Because the only inputs required to compute the adaptive estimator are the restricted and unrestricted point estimates along with their estimated covariance matrix, the burden on researchers of reporting adaptive estimates is very low. In the examples below, we draw on published tables of point estimates and standard errors whenever possible, in most cases using the replication data only to derive estimates of the covariance between the estimators. In the majority of these examples, we find that the restricted estimator is nearly efficient, implying the relevant covariances could have been inferred from published standard errors.

5.1 Adapting to a pre-trend (Dobkin et al., 2018)

We begin by returning to an example from Dobkin et al. (2018) who study the effects of unexpected hospitalization on out of pocket (OOP) spending. They consider a panel specification of the form

$$OOP_{it} = \gamma_t + X_{it}'\alpha + \sum_{\ell=0}^{3} \mu_\ell D_{it}^\ell + \varepsilon_{it},$$

where OOP_{it} is the OOP spending of individual i in calendar year t, $D_{it}^\ell = 1\{t - e_i = r\}$ is an event time indicator, e_i is the date of hospitalization, X_{it} is a vector of time varying covariates, and the $\{\mu_\ell\}_{\ell=0}^3$ are meant to capture the causal effect of hospitalization on OOP spending at various horizons, with $\ell = 0$ giving the contemporaneous impact.

Concerned that the parallel trends assumption required of their event study design might be violated, the authors add a linear trend $t - e_i$ to X_{it} in their baseline specification but also report results dropping the trend.

Table 1 shows the results of this robustness exercise at each horizon $\ell \in \{0, 1, 2, 3\}$, where we have denoted the ordinary least squares (OLS) estimates of μ_ℓ including the trend as Y_U and the estimates omitting the trend as Y_R. These point estimates exactly replicate the numbers underlying Panel A of Dobkin et al. (2018)'s Figure 1. The restricted estimates of μ_0 exhibit standard errors about 25% lower than the corresponding unrestricted estimates, with larger precision gains present at longer horizons. The GMM estimator that imposes $b = 0$ tracks Y_R closely and yields trivial improvements in precision, suggesting the restricted estimator is fully efficient. Consequently, the variability of the difference Y_O between the restricted and unrestricted estimators can be closely approximated by the difference between the squared standard error of Y_U and that of Y_R. At each horizon, we find a standardized difference T_O between the estimators of approximately 1.2.

Yrs since hospital		Y_U	Y_R	Y_O	GMM	Adaptive	Soft-threshold	Pre-test
0	Estimate	2,217	2,409	192	2,379	2,302	2,287	2,409
	Std Error	(257)	(221)	(160)	(219)			
	Max Regret	38%	∞		∞	15%	15%	68%
	Threshold						0.52	1.96
1	Estimate	1,268	1,584	316	1,552	1,435	1,408	1,584
	Std Error	(337)	(241)	(263)	(239)			
	Max Regret	98%	∞		∞	33%	34%	124%
	Threshold						0.59	1.96
2	Estimate	989	1,436	447	1,394	1,246	1,210	1,436
	Std Error	(430)	(270)	(373)	(267)			
	Max Regret	159%	∞		∞	47%	49%	161%
	Threshold						0.66	1.96
3	Estimate	1,234	1,813	579	1,752	1,574	1,530	1,813
	Std Error	(530)	(313)	(482)	(309)			
	Max Regret	195%	∞		∞	54%	57%	180%
	Threshold						0.69	1.96

Table 1: Impact of unexpected hospitalization on out of pocket (OOP) expenditures of the non-elderly insured (ages 50 to 59) from Dobkin et al. (2018). Standard errors in parentheses clustered by individual as in original study. "Yrs since hospital" refers to years since hospitalization. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$. The correlation coefficients between Y_U and Y_O by years since hospitalization are -0.524, -0.703, -0.784 and -0.813 respectively.

Since the difference Y_O between the restricted and unrestricted estimators is not statistically differentiable from zero at conventional levels of significance, the pre-test estimator simply discards the noisy estimates that include a trend and selects the restricted model. However, Y_O offers a fairly noisy assessment of the restricted estimator's bias. While zero bias can't be rejected at the 5% level in the year after hospitalization, neither can a bias equal to 50% of the restricted estimate.

The adaptive estimator balances these considerations regarding robustness and precision, generating an estimate roughly halfway between Y_R and Y_U. The worst case adaptation regret of the adaptive estimator rises from only 15% for the contemporaneous impact to 54% three years after hospitalization. The large value of $A^*(\mathcal{B})$ found at $\ell = 3$ is attributable to the elevated precision gains associated with Y_R at that horizon: in exchange for bounded risk, we miss out on the potentially very large risk reductions if $b = 0$. By contrast, the low adaptation regret provided at horizon $\ell = 0$ reflects the milder precision gains offered by Y_R when considering contemporaneous impacts. In effect, the near oracle performance found at this horizon reflects that the efficiency cost of robustness is low here.

The soft-thresholding estimator arrives at an estimate very similar to the adaptive estimator. By construction, the adaptive estimator exhibits lower worst case adaptation regret than the soft-thresholding estimator. Standard errors are not reported for the soft-thresholding, adaptive, or pre-test estimators because the variability of these procedures depends on the unknown bias level b.

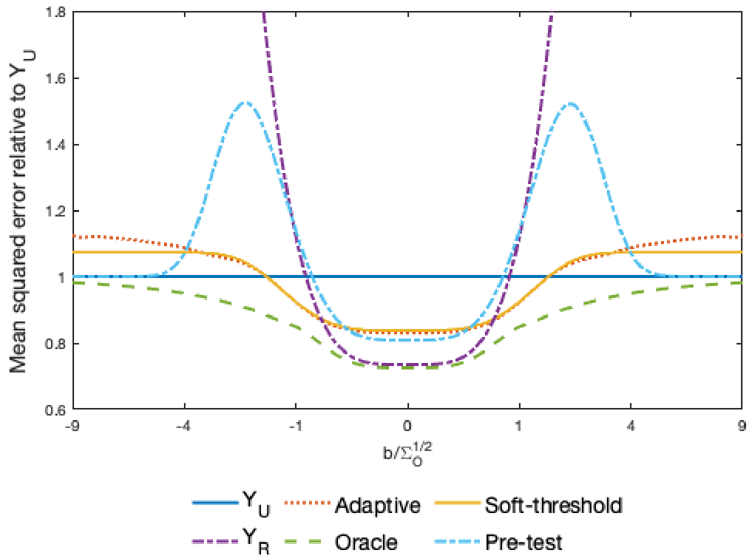

Figure 4: Risk functions for μ_0 ($\rho = -0.524$)

To assess the tradeoffs involved in adapting to misspecification, Figure 4 depicts the risk functions of the various estimation approaches listed in the first row of Table 1. Here, the correlation coefficient ρ between Y_U and Y_O equals -0.524: the value we estimated for the contemporaneous impact μ_0. As a normalization, the risk of the unrestricted estimator has been set to 1. The restricted estimator exhibits low risk when the bias is small but very high risk when the bias is large. Pre-testing yields good performance when the bias is either very large or very small. When the scaled bias is near the threshold value of 1.96 the pre-test estimator's risk becomes very large, as the results of the initial test become highly variable.

The line labeled "oracle" plots the B-minimax risk for $B = |b|$. The oracle's prior knowledge of the bias magnitude yields uniformly lower risk than any other estimator. The adaptive estimator mirrors the oracle, with nearly constant worst case adaptation regret. When the bias in the restricted estimator is small, the adaptive estimator yields large risk reductions relative to Y_U. When the bias is large, the adaptive estimator's risk remains bounded at a level substantially below the worst case risk experienced by the pre-test estimator.

		Unconstrained		Constrained $\bar{R}/\Sigma_U \leq 1.2$	
Years since hosp.		Adaptive	Soft-threshold	Adaptive	Soft-threshold
0	Estimates	2,302	2,287	2,302	2,287
	Max Regret	15%	15%	15%	15%
	Max Risk	13%	7%	13%	7%
	Threshold		0.52		0.52
1	Estimates	1,435	1,408	1,429	1,408
	Max Regret	33%	34%	41%	34%
	Max Risk	28%	17%	19%	17%
	Threshold		0.59		0.59
2	Estimates	1,246	1,210	1,248	1,176
	Max Regret	47%	49%	54%	60%
	Max Risk	41%	26%	19%	19%
	Threshold		0.66		0.56
3	Estimates	1,574	1,530	1,569	1,463
	Max Regret	54%	57%	60%	77%
	Max Risk	48%	31%	19%	19%
	Threshold		0.69		0.53

Table 2: Impact of unexpected hospitalization on out of pocket (OOP) expenditures of the non-elderly insured (ages 50 to 59) from Dobkin et al. (2018). "Yrs since hospital" refers to years since hospitalization. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$. "Max risk" refers to the worst case risk increase relative to Y_U in percentage terms $(R_{\max}(\delta) - \Sigma_U) \times 100$. The correlation coefficients between Y_U and Y_O by years since hospitalization are -0.524, -0.703, -0.784 and -0.813 respectively.

Table 2 shows the results from constrained adaptation limiting the worst case risk to no more than 20% above the risk of Y_U. This constraint results in relatively minor adjustments to the point estimates of both the adaptive and soft-thresholding estimators, even at horizon $\ell = 3$ in which unconstrained adaptation yields a 31-48% increase in worst case risk over Y_U. Of course, larger adjustments would have occurred if more extreme values of T_O had been realized. Ex-ante, constraining the adaptive estimator cuts its worst case risk by more than half while yielding only a modest increase of 6 percentage points in its worst case adaptation regret. The tradeoff between worst case risk and adaptation regret is somewhat less favorable for the soft-thresholding estimator: reducing its worst case risk by roughly a third raises its worst case adaptation regret by a third.

These worst case risk / adaptation regret tradeoffs are illustrated in Figure 5, which depicts the risk functions of the estimators at horizon $\ell = 3$. Remarkably, the risk constrained adaptive estimator exhibits substantially lower risk than the unconstrained adaptive and soft-thresholding estimators at most bias levels, while exhibiting only slightly elevated risk when the bias is small. We expect most researchers would view this tradeoff favorably. Constraining the soft-thresholding estimator yields similar risk reductions when the bias is large but generates more substantial risk increases when the bias magnitude is negligible. Overall, however, the constrained soft-thresholding estimator provides a reasonably close approximation to the constrained adaptive estimator.

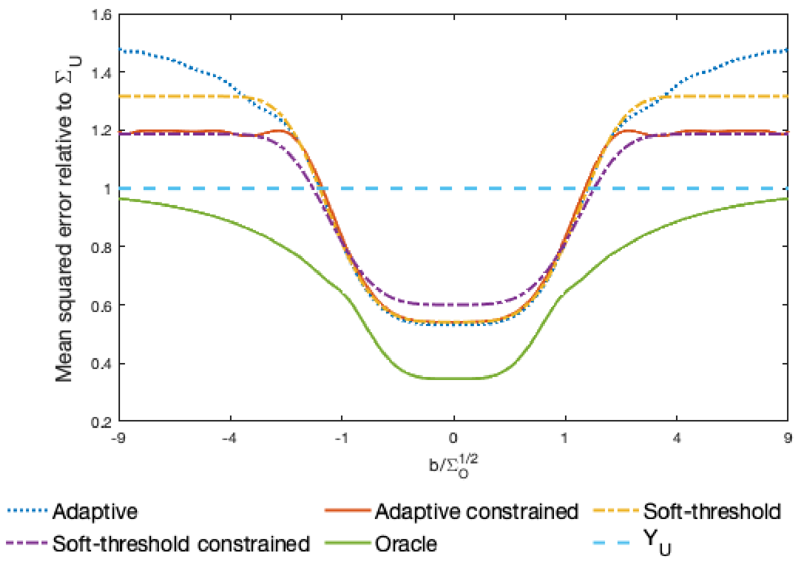

Figure 5: Risk functions for μ_3 ($\rho = -0.813$)

5.2 Adapting to an invalid instrument (Berry et al., 1995)

Our second example comes from Berry et al. (1995)'s seminal study of the equilibrium determination of automobile prices. As in Andrews et al. (2017) and Armstrong and Kolesár (2021), we focus on their analysis of average price-cost markups. Y_U is taken as the average markup implied by optimally weighted GMM estimation using a set of 8 demand-side instruments described in Andrews et al. (2017). We take as Y_R the GMM estimator that adds to the demand side instruments a set of 12 additional supply-side instruments. Following Armstrong and Kolesár (2021), we compute the GMM estimates in a single step using a weighting matrix allowing for unrestricted misspecification ($B = \infty$).

	Y_U	Y_R	Y_O	Adaptive	Soft-threshold	Pre-test
Estimate	52.95	33.53	-19.42	49.44	51.89	52.95
Std Error	(2.54)	(1.81)	(1.78)			
Max Regret	96%	∞		32%	34%	107%
Threshold					0.59	1.96

Table 3: Adaptive estimates for the average markup (in percent). Point estimates and standard errors calculated using misspecification robust weighting matrix as in Armstrong and Kolesár (2021). "Max Regret" refers to worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B},\delta)-1)\times 100$. The correlation coefficient between Y_U and Y_O is $\rho=-0.7$.

Table 3 lists estimates under different estimation approaches. The realizations of Y_R and Y_U correspond, respectively, to the estimates labeled "all excluded supply" and "none" in Figure 1 of Armstrong and Kolesár (2021). Because both Y_U and Y_R are computed using an efficient weighting matrix, the variance of their difference Y_O is given by the difference in their squared standard errors. While relying on demand side instruments alone implies automobile prices average 53% above marginal cost, adding supply side instruments yields much lower markups, with prices approximately 34% above marginal cost on average. Adding the supply side instruments not only decreases the average markup estimate but also reduces the standard error by nearly 30%. However, the difference Y_O between the restricted and unrestricted estimates is large and statistically significant, with $T_O\approx -11$.

Detecting what appears to be severe misspecification, the adaptive estimator shrinks strongly towards Y_U, as does the soft-thresholding estimator. The chosen soft-threshold is very low, indicating a relatively high level of robustness to bias: only scaled bias estimates smaller than 0.59 in magnitude are zeroed out. Consequently, even realizations of T_O near 3 would have yielded soft-thresholding point estimates close to Y_U in this setting. Evidently, entertaining instruments that turn out to be heavily biased yields little adaptation regret in this scenario, as both the soft-thresholding and optimally adaptive estimators are highly robust. Had the realized value of Y_O been small, these estimators would have placed significant weight on Y_R, potentially yielding substantial efficiency gains relative to relying on Y_U alone.

5.3 Adapting to heterogeneous effects (Gentzkow et al., 2011)

An influential recent literature emphasizes the potential for two-way fixed effects estimators to identify non-convex weighted averages of heterogeneous treatment effects (de Chaisemartin and D'Haultfœuille, 2020b; Sun and Abraham, 2021; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021). Convexity of the weights defining a causal estimand θ is generally agreed to be an important desideratum, guaranteeing that when treatment effects are of uniform sign, θ will also possess that sign. Hence, an estimator exhibiting asymptotically convex weights limits the scope of potential biases when treatment effects are all of the same sign. However, when treatment effect heterogeneity is mild, an estimator exhibiting asymptotic weights of mixed sign may yield negligible asymptotic bias and substantially lower asymptotic variance than a convex weighted alternative. Consequently, researchers choosing between standard two-way fixed effects estimators and recently proposed convex weighted estimators often face a non-trivial robustness-efficiency tradeoff.

An illustration of this tradeoff comes from Gentzkow et al. (2011) who study the effect of newspapers on voter turnout in US presidential elections between 1868 and 1928. They consider the following linear model relating the first-difference of the turnout rate to the first difference of the number of newspapers available in different counties:

$$\Delta y_{ct} = \beta \Delta n_{ct} + \Delta \gamma_{st} + \delta \Delta x_{ct} + \lambda \Delta z_{ct} + \Delta \varepsilon_{ct},$$

where Δ is the first difference operator, γ_{st} is a state-year effect, x_{ct} is a vector of observable county characteristics, and z_{ct} denotes newspaper profitability. The parameter β is meant to capture a causal effect of newspapers on voter turnout. In what follows, we take the OLS estimator of β as Y_R.

de Chaisemartin and D'Haultfœuille (2020b) establish that Y_R yields a weighted average of average causal effects across different time periods and different counties, estimating that 46% of the relevant weights are negative. To guard against the potential biases stemming from reliance on negative weights, they propose a convex weighted estimator of average treatment effects featuring weights that are treatment shares. We take this convex weighted estimator as Y_U, implying our estimand of interest θ is average treatment on the treated.

Table 4 reports the realizations of (Y_U, Y_R) and their standard errors, which exactly replicate those given in Table 3 of de Chaisemartin and D'Haultfœuille (2020b). Once again the estimated variance of Y_O is closely approximated by the difference in squared standard errors between Y_U and Y_R, suggesting Y_R is nearly efficient. Hence, the downstream GMM, adaptive, and soft-thresholding estimators could have been computed using only the published point estimates and standard errors.

Though the realized value of Y_U is nearly twice as large as that of Y_R, the two estimators are not statistically distinguishable from one another at the 5% level. Hence, a conventional pre-test suggests ignoring the perils of negative weights and confining attention to Y_R on account of its substantially increased precision. Like Y_R, GMM exhibits a standard error roughly 35% below that of Y_U. Consequently, relying solely on the convex-weighted but highly inefficient estimator Y_U exposes the researcher to a large worst-case adaptation regret of 145%.

In contrast to the pre-test, both the optimally adaptive estimator and its soft-thresholding

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

Table 4: Estimates of the effect of one additional newspaper on turnout. Bootstrap standard errors in parentheses computed using the same 100 bootstrap samples utilized by de Chaisemartin and D'Haultfœuille (2020b). "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$. The correlation coefficient between Y_U and Y_O is -0.77.

approximation place substantial weight $w(T_O)$ on the convex estimator, yielding estimates roughly 60% of the way towards Y_U from GMM. This phenomenon owes to the fact that with $T_O = -1.7$ both estimators detect the presence of a non-trivial amount of bias in Y_R. We can easily compute the soft-thresholding bias estimate from the figures reported in the table as $(-1.7 + .64) \times 0.001 \approx -.001$, suggesting that Y_R exhibits a bias of nearly 40%. Balancing this bias against the estimator's increased precision leads the soft-thresholding estimator to essentially split the difference between the convex and non-convex weighted estimators, which yields a near optimal worst case adaptation regret of 46%.

5.4 Adapting to non-experimental 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. Table 5 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%	∞	∞	∞	∞	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

Table 5: 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 D for further details.

From Table 5, 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 update in our best estimate of the effect of NSW training away from the experimental benchmark. In Appendix Table A1 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 Appendix Table A2, 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.

5.5 Adapting to endogeneity (Angrist and Krueger, 1991)

Our final example comes from Angrist and Krueger (1991)'s classic analysis of the returns to schooling using quarter of birth as an instrument for schooling attainment. Documenting that individuals born in the first quarter of the year acquire fewer years of schooling than those born later in the year, they demonstrate that the earnings of those born in the first quarter of the year also earn less than those born later in the year.

Table 6 replicates exactly the estimates reported in Angrist and Krueger (1991, Panel B, Table III) for men born 1930-39. Y_U gives the Wald-IV estimate of the returns to schooling using an indicator for being born in the first quarter of the year as an instrument for years of schooling completed, while Y_R gives the corresponding OLS estimate. Neither estimator controls for additional covariates. The first stage relationship between quarter of birth and years of schooling exhibits a z-score of 8.24, suggesting an asymptotic normal approximation to Y_U is likely to be highly accurate. As in our previous examples, the variance of the difference between Y_U and Y_R is very closely approximated by the difference in their squared standard errors, indicating this exercise could have been computed using only the information reported in the original published tables.

While the IV estimator accounts for endogeneity, it is highly imprecise, with a standard error two orders of magnitude greater than OLS. Consequently, the maximal regret associated with using IV instead of OLS is extremely large, as the variability of Y_U is more than 5,000 times that of Y_R. IV and OLS cannot be statistically distinguished

	Y_U	Y_R	Y_O	Adaptive	Soft-threshold	Pre-test
Estimate	0.102	0.0709	-0.0311	0.071	0.071	0.071
Std Error	(0.0239)	(0.0003)	(0.0239)			
Max Regret	500145%	∞		493%	537%	17882%
Thresholds					2.07	1.96

Table 6: Returns to schooling. Standard errors in parentheses computed under homoscedasticity as in original study. "Max regret" refers to the worst case adaptation regret in percentage terms $(A^*(\mathcal{B}) - 1) \times 100$. The correlation coefficients between Y_U and Y_O is $\rho = -0.9998$.

at conventional significance levels, with $T_O \approx 1.3$. The inability to distinguish IV from OLS estimates of the returns to schooling is characteristic not only of the specifications reported in Angrist and Krueger (1991) but of the broader quasi-experimental literature spawned by their landmark study (Card, 1999).

The confluence of extremely large maximal regret for Y_U with a statistically insignificant difference Y_O, leads the adaptive estimator, the soft-thresholding estimator and the pre-test estimator to all coincide with Y_R. The motives for this coincidence are of course quite different. The adaptive and soft-thresholding estimators seek to avoid the regret associated with missing out on the enormous efficiency gains of OLS if it is essentially unconfounded. By contrast, the pre-test estimator simply fails to reject the null hypothesis that years of schooling is exogenous at the proper significance level.

Despite the agreement of the three approaches, the extremely large adaptation regret exhibited by the optimally adaptive estimator suggests it is unlikely to garner consensus in this setting. Committing to Y_R exposes the researcher to potentially unlimited risk. The adaptive and soft-thresholding estimators avoid committing to either Y_U or Y_R before observing the data but still expose the researcher to an approximately five fold maximal risk increase relative to Y_U. A skeptic concerned with the potential biases in OLS is therefore unlikely to be willing to rely on such an estimator.

	Unconstrained		Constrained $\bar{R}/\Sigma_U \leq 1.2$	
	Adaptive	Soft-threshold	Adaptive	Soft-threshold
Estimate (fully nonlinear)	0.071	0.071	0.087	0.091
Max Regret	493%	537%	30089%	34086%
Max Risk	455%	427%	20%	20%
Threshold		2.07		0.45

Table 7: Adaptive estimates of returns to schooling. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$. "Max risk" refers to the worst case risk increase relative to Y_U in percentage terms $(R_{\max}(\delta) - \Sigma_U)/\Sigma_U \times 100$. The correlation coefficient is $\rho = -0.9998$.

As shown in Table 7, if we instead follow the rule of thumb of limiting ourselves to a 20% increase in maximal risk, both the adaptive and soft-threshold estimators yield returns to schooling estimates of roughly 9%, approximately halfway between OLS and IV. The maximal regret of these estimates is extremely high, reflecting the potential efficiency costs of weighting Y_U so heavily. These efficiency concerns are likely outweighed in this case by the potential for extremely large biases. Though these estimates are unlikely to garner consensus across camps of researchers with widely different beliefs, the risk-limited adaptive estimator should yield wider consensus than proposals to discard Y_R and rely on Y_U alone.

6 Conclusion

Empiricists routinely encounter robustness-efficiency tradeoffs. The reporting of estimates from different models has emerged as a best practice at leading journals. The methods introduced here provide a scientific means of summarizing what has been learned from such exercises and arriving at a preferred estimate that trades off considerations of bias against variance.

Computing the adaptive estimators proposed in this paper requires only point estimates, standard errors, and the covariance between estimators, objects that are easily produced by standard statistical packages. As our examples revealed, in many cases the restricted estimator is nearly efficient, implying the relevant covariance can be deduced from the standard errors of the restricted and unrestricted estimators.

In line with earlier results from Bickel (1984), we found that soft-thresholding estimators closely approximate the optimally adaptive estimator in the scalar case, while requiring less effort to compute. An interesting topic for future research is whether similar approximations can be developed for higher dimensional settings where the curse of dimensionality renders direct computation of optimally adaptive estimators infeasible.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory, 1973*, pp. 267–281. Akademiai Kiado.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017). Measuring the Sensitivity of Parameter Estimates to Estimation Moments. *The Quarterly Journal of Economics 132*(4), 1553–1592.

Angrist, J. D. and A. B. Krueger (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

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

Armstrong, T. B. and M. Kolesár (2018). Optimal Inference in a Class of Regression Models. *Econometrica 86*(2), 655–683.

Armstrong, T. B. and M. Kolesár (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics 12*(1), 77–108.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile Prices in Market Equilibrium. *Econometrica 63*(4), 841–890.

Bickel, P. J. (1982, September). On Adaptive Estimation. *The Annals of Statistics 10*(3), 647–671.

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.

Bickel, P. J. (1984, September). Parametric Robustness: Small Biases can be Worthwhile. *The Annals of Statistics 12*(3), 864–879. Publisher: Institute of Mathematical Statistics.

Box, G. E. and N. R. Draper (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

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

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics 3*, 1801–1863.

Chamberlain, G. (2000, November). Econometric applications of maxmin expected utility. *Journal of Applied Econometrics 15*(6), 625–644.

de Chaisemartin, C. and X. D'Haultfœuille (2020a, June). Empirical MSE Minimization to Estimate a Scalar Parameter.

de Chaisemartin, C. and X. D'Haultfœuille (2020b, September). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–2996.

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.

Dobkin, C., A. Finkelstein, R. Kluender, and M. J. Notowidigdo (2018). The Economic Consequences of Hospital Admissions. *American Economic Review 108*(2), 308–52.

Donoho, D. L. (1994, March). Statistical Estimation and Optimal Recovery. *The Annals of Statistics 22*(1), 238–270.

Efron, B. and C. Morris (1972). Empirical Bayes on Vector Observations: An Extension of Stein's Method. *Biometrika 59*(2), 335–347.

Elliott, G., U. K. Müller, and M. W. Watson (2015, March). Nearly Optimal Tests When a Nuisance Parameter Is Present Under the Null Hypothesis. *Econometrica 83*(2), 771–811.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2011, December). The Effect of Newspaper Entry and Exit on Electoral Politics. *American Economic Review 101*(7), 2980–3018.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics 18*(2), 141–153.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica 46*(6), 1251–1271.

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.

Hjort, N. L. and G. Claeskens (2003). Frequentist Model Average Estimators. *Journal of the American Statistical Association 98*(464), 879–899.

Hodges, J. L. and E. L. Lehmann (1952). The use of Previous Experience in Reaching Statistical Decisions. *The Annals of Mathematical Statistics 23*(3), 396–407.

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

Kline, P. and C. Walters (2021). Reasonable doubt: Experimental detection of job-level employment discrimination. *Econometrica 89*(2), 765–792.

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

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*, Volume 53. John Wiley & Sons Incorporated.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(01), 21–59.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd edition ed.). New York: Springer.

Low, M. G. (1997, December). On nonparametric confidence intervals. *The Annals of Statistics 25*(6), 2547–2554.

Mallows, C. L. (1973). Some Comments on CP. *Technometrics 15*(4), 661–675.

Miguel, E. (2021). Evidence on research transparency in economics. *Journal of Economic Perspectives 35*(3), 193–214.

Müller, U. K. and Y. Wang (2019, March). Nearly weighted risk minimal unbiased estimation. *Journal of Econometrics 209*(1), 18–34.

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

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica: Journal of the Econometric Society*, 571–587.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Sun, L. and S. Abraham (2021). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics 225*(2), 175–199.

Tsybakov, A. B. (1998, December). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *The Annals of Statistics 26*(6), 2420–2469.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* New York: Springer.

Appendix A 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.

A.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 δ 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 δ allowing payment of the smallest G that ensures consensus.

A committee member who is certain of the parameters (θ, 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.

A.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 δ. 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 δ.

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 δ, 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 δ, 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 δ 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 δ that minimizes $\sup_{B\in\mathcal{B}} \ln\left(R_{max}(B,\delta)/R^*(B)\right) = \sup_{B\in\mathcal{B}} \ln A(B,\delta)$. The CoC will therefore propose the optimally adaptive decision δ^{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.

A.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 B Details and proofs for Section 4

B.1 Details for main example

We provide details and formal results for the results in Section 4.3 giving B-minimax and optimally adaptive estimators in our main example. We first provide a general theorem characterizing minimax estimators in a setting that includes our main example. We then specialize this result to derive the the formula for the B-minimax estimator and optimally adaptive estimator for our main example given in Section 4.3, using a weighted loss function and Lemma 4.1 to obtain the optimally adaptive estimator. This proves Theorem 4.1.

We consider a slightly more general setting with p misspecified estimates, leading to a $p \times 1$ vector Y_O:

$$Y = \begin{pmatrix} \underset{1\times 1}{Y_U} \\ \underset{p\times 1}{Y_O} \end{pmatrix} \sim N\left(\begin{pmatrix} \underset{1\times 1}{\theta} \\ \underset{p\times 1}{b} \end{pmatrix}, \Sigma \right), \quad \Sigma = \begin{pmatrix} \underset{1\times 1}{\Sigma_U} & \underset{1\times p}{\Sigma_{UO}} \\ \underset{p\times 1}{\Sigma'_{UO}} & \underset{p\times p}{\Sigma_O} \end{pmatrix}. \tag{11}$$

In our main example, $p = 1$ and $\rho = \Sigma_{UO}/\sqrt{\Sigma_U \Sigma_O}$. We are interested in the minimax risk of an estimator $\delta : \mathbb{R}^{p+1} \to \mathbb{R}$ under the loss function $L(\theta, b, d)$, which may incorporate a scaling to turn the minimax problem into a problem of finding an optimally adaptive estimator, following Lemma 4.1. We assume that the loss function satisfies the invariance condition

$$L(\theta + t, b, d + t) = L(\theta, b, d) \quad \text{all } t \in \mathbb{R}. \tag{12}$$

We consider minimax estimation over a parameter space $\mathbb{R} \times \mathcal{C}$:

$$\inf_{\delta} \sup_{\theta \in \mathbb{R}, b \in \mathcal{C}} R(\theta, b, \delta). \tag{13}$$

Theorem B.1. *Suppose that the loss function $L(\theta, b, d)$ is convex in d and that (12) holds. Then the minimax risk (13) is given by*

$$\begin{aligned} &\inf_{\bar{\delta}} \sup_{b \in \mathcal{C}} E_{0,b}[\tilde{L}(b, \bar{\delta}(Y_O) - \Sigma_{UO}\Sigma_O^{-1} b)] \\ &= \sup_{\pi \text{ supported on } \mathcal{C}} \inf_{\bar{\delta}} \int E_{0,b}[\tilde{L}(b, \bar{\delta}(Y_O) - \Sigma_{UO}\Sigma_O^{-1} b)] \, d\pi(b) \end{aligned} \tag{14}$$

where $\tilde{L}(b, t) = EL(0, b, t + V)$ with $V \sim N(0, \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma'_{UO})$. Furthermore, the minimax problem (13) has at least one solution, and any solution δ^* takes the form

$$\delta^*(Y_U, Y_O) = Y_U - \Sigma_{UO}\Sigma_O^{-1} Y_O + \bar{\delta}^*(Y_O)$$

where $\bar{\delta}^*$ achieves the infimum in (14).

Proof. The minimax problem (13) is invariant (in the sense of pp. 159-161 of Lehmann and Casella (1998)) to the transformations $(\theta, b) \mapsto (\theta + t, b)$ and the associated transformation of the data $(Y_U, Y_O) \mapsto (Y_U + t, Y_O)$, where t varies over $\mathbb{R}$. Equivariant estimators for this group of transformations are those that satisfy $\delta(y_U + t, y_O) = \delta(y_U, y_O) + t$, which is equivalent to imposing that the estimator takes the form $\delta(y_U, y_O) = \delta(0, y_O) + y_U$. The risk of such an estimator does not depend on θ and is given by

$$R(\theta, b, \delta) = R(0, b, \delta) = E_{0,b}\left[L(0, b, \delta(0, Y_O) + Y_U)\right].$$

Using the decomposition $Y_U - \theta = \Sigma_{UO}\Sigma^{-1}(Y_O - b) + V$ where $V \sim N(0, \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma'_{UO})$

is independent of Y_O, the above display is equal to

$$E_{0,b}\left[L(0,b,\delta(0,Y_O)+\Sigma_{UO}\Sigma_O^{-1}(Y_O-b)+V)\right]=E_{0,b}\tilde{L}(b,\delta(0,Y_O)+\Sigma_{UO}\Sigma_O^{-1}(Y_O-b)).$$

Letting $\bar{\delta}(Y_O)=\delta(0,Y_O)+\Sigma_{UO}\Sigma_O^{-1}Y_O$, the above display is equal to $E_{0,b}[\tilde{L}(b,\bar{\delta}(Y_O)-\Sigma_{UO}\Sigma_O^{-1}b)]$. Thus, if an estimator $\bar{\delta}^*$ achieves the infimum in (14), the corresponding estimator $\delta(Y_U,Y_O)=\delta(0,Y_O)+Y_U=\bar{\delta}^*(Y_O)-\Sigma_{UO}\Sigma_O^{-1}Y_O+Y_U$ will be minimax among equivariant estimators for (13). It will then follow from the Hunt-Stein Theorem (Lehmann and Casella, 1998, Theorem 9.2) that this minimax equivariant estimator is minimax among all estimators, that any other minimax estimator takes this form and that the minimax risk is given by the first line of (14).

It remains to show that the infimum in the first line of (14) is achieved, and that the equality claimed in (14) holds. The equality in (14) follows from the minimax theorem, as stated in Theorem A.5 in Johnstone (2019) (note that $d\mapsto\tilde{L}(b,d-\Sigma_{UO}\Sigma_O^{-1}b)$ is convex since it is an integral of the convex functions $d\mapsto L(0,b,d-\Sigma_{UO}\Sigma_O^{-1}b+v)$ over the index v). The existence of an estimator $\bar{\delta}^*$ that achieves the infimum in the first line of (14) follows by noting that the set of decision rules (allowing for randomized decision rules) is compact in the topology defined on p. 405 of Johnstone (2019), and the risk $E_{0,b}[\tilde{L}(b,\bar{\delta}(Y_O)-\Sigma_{UO}\Sigma_O^{-1}b)]$ is continuous in $\bar{\delta}$ under this topology. As noted immediately after Theorem A.1 in Johnstone (2019), this implies that $\bar{\delta}\mapsto\sup_b E_{0,b}[\tilde{L}(b,\bar{\delta}(Y_O)-\Sigma_{UO}\Sigma_O^{-1}b)]$ is a lower semicontinuous function on the compact set of possibly randomized decision rules under this topology, which means that there exists a decision rule that achieves the minimum. From this possibly randomized decision rule, we can construct a nonrandomized decision rule that achieves the minimum by constructing a nonrandomized decision rule with uniformly smaller risk by averaging, following Johnstone (2019, p. 404). □

We now prove Theorem 4.1 by specializing this result. The notation is the same as in the main text, with ρ in the main text given by $\Sigma_{UO}/\sqrt{\Sigma_U\Sigma_O}$.

First, we derive the minimax estimator and minimax risk in (13) when $L(\theta,b,d)=(\theta-d)^2$ and $\mathcal{C}=[-B,B]$. We have $\tilde{L}(b,t)=E(t+V)^2=t^2+\Sigma_U-\Sigma_{UO}^2/\Sigma_O$. Thus, (14) becomes

$$\begin{aligned}
&\inf_{\bar{\delta}}\sup_{b\in[-B,B]}E_{0,b}\left[\left(\bar{\delta}(Y_O)-\frac{\Sigma_{UO}}{\Sigma_O}b\right)^2\right]+\Sigma_U-\frac{\Sigma_{UO}^2}{\Sigma_O}\\
&=\inf_{\bar{\delta}}\sup_{b\in[-B,B]}\frac{\Sigma_{UO}^2}{\Sigma_O}E_{0,b}\left[\left(\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}}\bar{\delta}(Y_O)-\frac{b}{\sqrt{\Sigma_O}}\right)^2\right]+\Sigma_U-\frac{\Sigma_{UO}^2}{\Sigma_O}.
\end{aligned}$$

This is equivalent to observing $T_O=Y_O/\sqrt{\Sigma_O}\sim N(t,1)$ and finding the minimax estimator of t under the constraint $|t|\le B/\sqrt{\Sigma_O}$. Letting $\delta^{\text{BNM}}(T_O;B/\sqrt{\Sigma_O})$ denote the solution to this minimax problem and letting $r^{\text{BNM}}(B/\sqrt{\Sigma_O})$ denote the value of this minimax problem, the optimal $\bar{\delta}$ in the above display satisfies $\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}}\bar{\delta}(Y_O) = \delta^{\text{BNM}}(Y_O/\sqrt{\Sigma_O}; B/\sqrt{\Sigma_O})$, which gives the value of the above display as

$$\frac{\Sigma_{UO}^2}{\Sigma_O} r^{\text{BNM}}(B/\sqrt{\Sigma_O}) + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O} \tag{15}$$

and the B-minimax estimator as

$$\frac{\Sigma_{UO}}{\sqrt{\Sigma_O}} \delta^{\text{BNM}}(Y_O/\sqrt{\Sigma_O}; B/\sqrt{\Sigma_O}) + Y_U - \frac{\Sigma_{UO}}{\Sigma_O} Y_O. \tag{16}$$

Substituting $T_O = Y_O/\sqrt{\Sigma_O}$ and the notation $\rho = \Sigma_{UO}/\sqrt{\Sigma_U \Sigma_O}$ used in the main text gives (4) and (5). This proves part (i) of Theorem 4.1.

To find the optimally adaptive estimator and loss of efficiency under adaptation in our main example, we apply Lemma 4.1 with $\omega(\theta, b) = R^*(|b|)^{-1}$, with $R^*(B)$ given by (15). This leads to the minimax problem (13) with $\mathcal{C} = \mathbb{R}$ and $L(\theta, b, d) = R^*(|b|)^{-1}(\theta - d)^2$. The function $\tilde{L}$ in Theorem B.1 is then given by $\tilde{L}(b, t) = ER^*(|b|)^{-1}(t + V)^2 = R^*(|b|)^{-1}(t^2 + \Sigma_U - \Sigma_{UO}^2/\Sigma_O)$, which gives (14) as

$$\inf_{\bar{\delta}} \sup_{b \in \mathbb{R}} \frac{E_{0,b}\left[\left(\bar{\delta}(Y_O) - \frac{\Sigma_{UO}}{\Sigma_O} b\right)^2\right] + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O}}{\frac{\Sigma_{UO}^2}{\Sigma_O} r^{\text{BNM}}(|b|/\sqrt{\Sigma_O}) + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O}} = \inf_{\bar{\delta}} \sup_{b \in \mathbb{R}} \frac{E_{0,b}\left[\left(\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}}\bar{\delta}(Y_O) - \frac{b}{\sqrt{\Sigma_O}}\right)^2\right] + \rho^{-2} - 1}{r^{\text{BNM}}(|b|/\sqrt{\Sigma_O}) + \rho^{-2} - 1}.$$

This proves part (iii) of Theorem 4.1. The above display is minimized by $\bar{\delta}$ satisfying $\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}}\bar{\delta}(Y_O) = \tilde{\delta}^{\text{adapt}}(Y_O/\sqrt{\Sigma_O}; \rho)$ where $\tilde{\delta}^{\text{adapt}}(T; \rho)$ minimizes (6) in the main text. By Theorem B.1, the optimally adaptive estimator is given by

$$\frac{\Sigma_{UO}}{\sqrt{\Sigma_O}} \tilde{\delta}^{\text{adapt}}(Y_O/\sqrt{\Sigma}; \rho) + Y_U - \frac{\Sigma_{UO}}{\Sigma_O} Y_O = \rho\sqrt{\Sigma_U}\tilde{\delta}^{\text{adapt}}(T_O; \rho) + Y_U - \rho\sqrt{\Sigma_U} T_O. \tag{17}$$

This proves the part (ii) of Theorem 4.1.

B.2 Details for constrained adaptation

We provide proof for Lemma 4.2, 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 4.5. 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 $\tilde{\delta}_t^*$ be a solution to (9).

Suppose, to get a contradiction, that a decision δ' satisfies the constraint in (8) 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 (10) 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.

B.3 Numerical results on estimators as a function of ρ^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 ρ^2. As ρ^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 ρ^2 increase, which narrows the range of true bias for which the adaptive estimator beats Y_U in terms of risk.

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 A2. 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.

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 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 4.4, adaptive estimators have large worst-case risk when ρ^2 is close to one. The pre-test estimator has lower worst-case

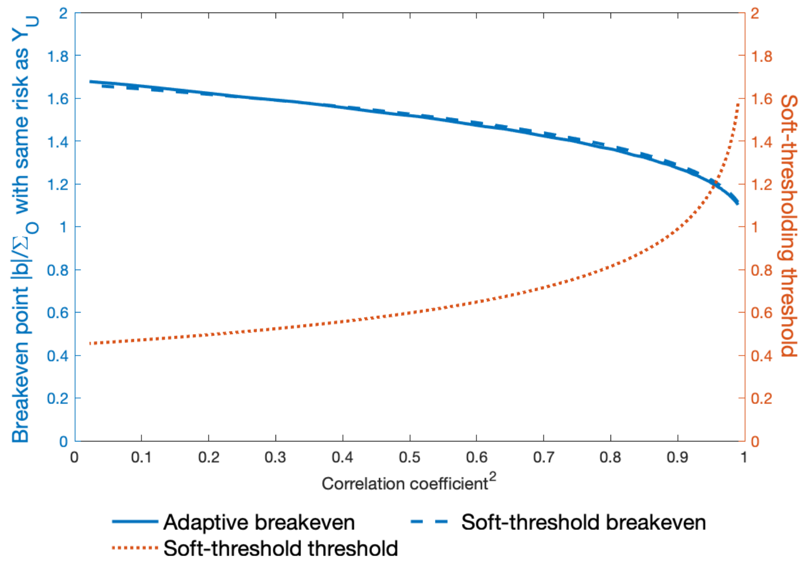

Figure A1: Threshold for adaptive soft-thresholding estimator

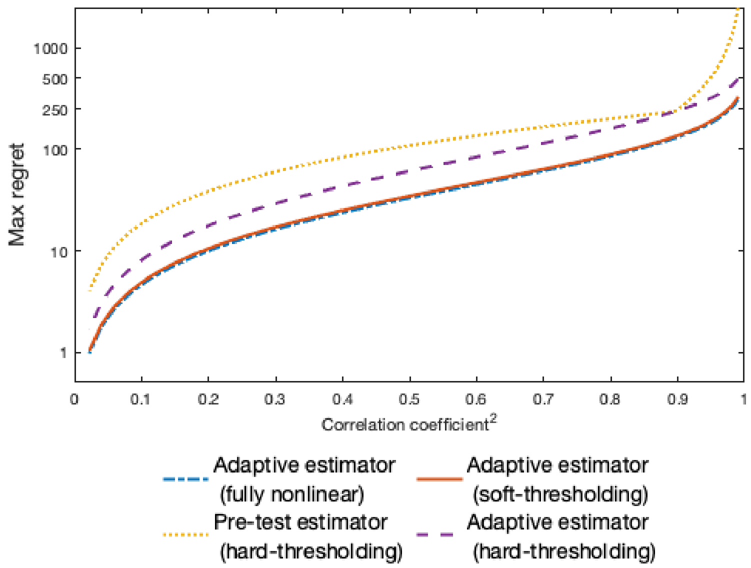

Figure A2: "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \delta) - 1) \times 100$.

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

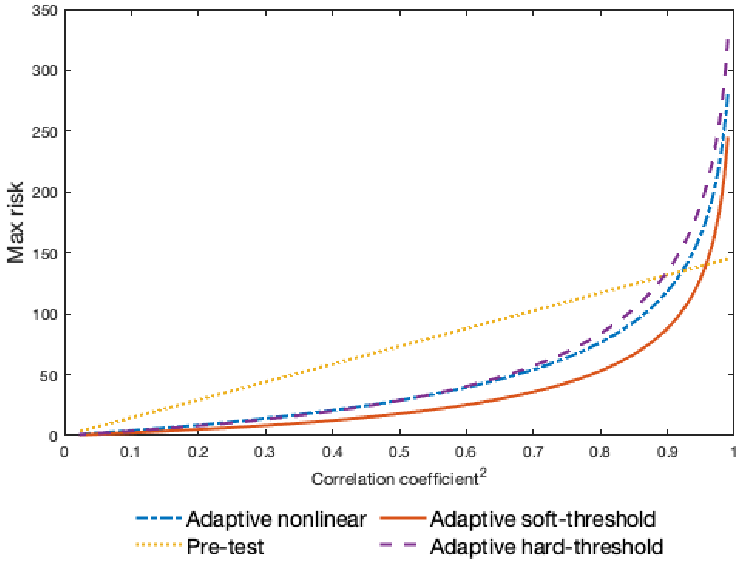

Figure A3: "Max risk" refers to the worst case risk increase relative to Y_U in percentage terms $(R_{\max}(\delta) - \Sigma_U)/\Sigma_U \times 100$.

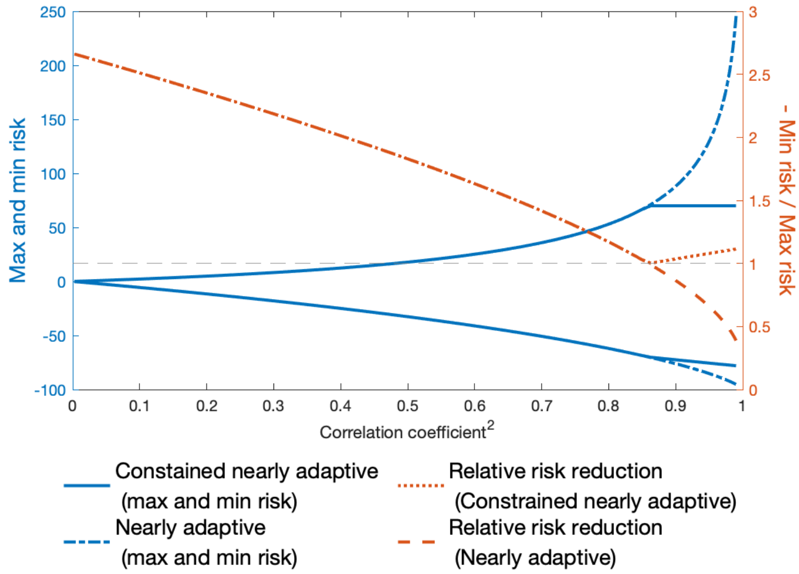

Figure A4: "Max risk" refers to the worst case risk increase relative to Y_U in percentage terms $(R_{\max}(\infty, \delta) - \Sigma_U)/\Sigma_U \times 100$. "Min risk" refers to the best case risk decrease relative to Y_U in percentage terms $(\min_b R(\theta, b, \delta) - \Sigma_U)/\Sigma_U \times 100$. 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.

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.

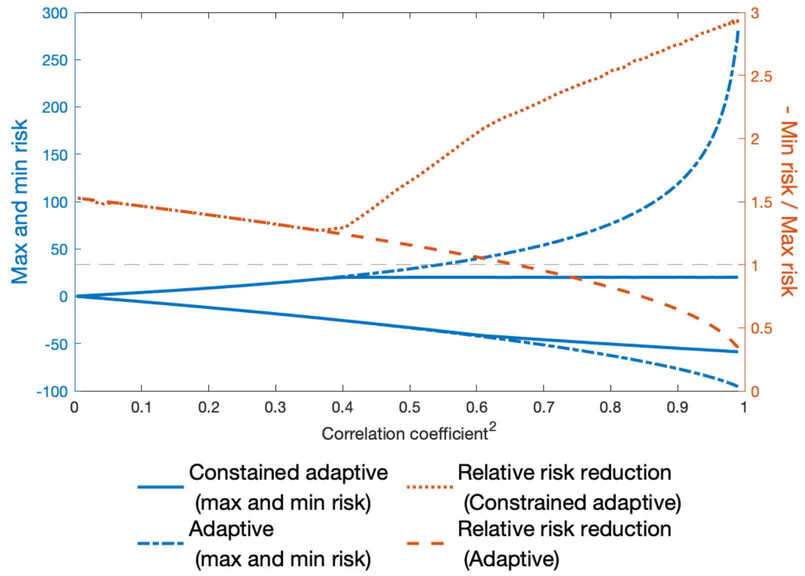

Figure A5: "Max risk" refers to the worst case risk increase relative to Y_U in percentage terms $(R_{\max}(\infty, \delta) - \Sigma_U)/\Sigma_U \times 100$. "Min risk" refers to the best case risk decrease relative to Y_U in percentage terms $(\min_b R(\theta, b, \delta) - \Sigma_U)/\Sigma_U \times 100$. 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.

B.4 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 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 4.4. 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 ρ close to 1, consider the Hausman setting where Y_R is efficient under the restriction $b = 0$. In this case, we have $\operatorname{var}(Y_R) = \operatorname{cov}(Y_R, Y_U)$, $\operatorname{cov}(Y_O, Y_U) = \operatorname{cov}(Y_R - Y_U, Y_U) = \operatorname{var}(Y_R) - \operatorname{var}(Y_U)$ and $\operatorname{var}(Y_O) = \operatorname{var}(Y_R) + \operatorname{var}(Y_U) - 2\operatorname{cov}(Y_R, Y_U) = \operatorname{var}(Y_U) - \operatorname{var}(Y_R)$. It follows that

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

and

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

Therefore, $|\rho| \to 1$ corresponds to the case where $\operatorname{var}(Y_R)/\operatorname{var}(Y_U) \to 0$. Furthermore, $\rho^{-2} - 1 = \frac{\operatorname{var}(Y_R)}{\operatorname{var}(Y_U)}(1 + o(1))$ 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 B.2. *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 B.2. We split the proof into upper bounds (Section B.4.1) and lower bounds (Section B.4.2). The lower bounds in Section B.4.2 are essentially immediate from results in Bickel (1983) for adapting to $B \in \mathcal{B} = \{0, \infty\}$, whereas the upper bounds in Section B.4.1 involve new arguments to deal with intermediate values of B.

B.4.1 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 B.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 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). $\square$

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

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

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 $|\mu| \le c(\rho)$ and for $|\mu| \ge c(\rho)$. For $|\mu| \le c(\rho)$, we use the bound $r^{\text{BNM}}(|\mu|) \ge .8 \cdot \mu^2/(\mu^2+1)$ (Donoho, 1994), which gives an upper bound for (18) 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 $r^{\text{BNM}}(|\mu|) \ge r^{\text{BNM}}(c(\rho))$ which gives an upper bound for (18) of

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

As $|\rho| \uparrow 1$, $c(\rho) \to \infty$ and $r^{\text{BNM}}(|c(\rho)|) \to 1$, 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.

B.4.2 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 B.2 (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 B.4.1. The required lower bound $A^*(\rho) \ge (1+o(1))2\log(\rho^{-2}-1)^{-1}$ follows from this result.

Lemma B.3. *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 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 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 B.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 C Computational details

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

C.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

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

Define $\pi_{kb}=\Pr_{T\sim N(b,1)}\left(T=t_k\right)$ 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

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

where we define $t_0=-\infty$ and $t_{K+1}=\infty$, which ensures that $\sum_{k=1}^{K}\pi_{kb}=1$. The discretized approximation to the risk function (19) is therefore

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

C.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 π with the finite dimensional vector $\mu\in\Delta^J$, 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

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

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

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

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

Since the objective is concave in μ (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.

C.3 Computing the optimally adaptive estimator for a given ρ^2

As explained in the main text, the adaptive problem in the main example 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 $\tilde{\delta}^{\text{adapt}}(t;\rho)$ based on the steps described below, which is a specialized version of the second step of Algorithm 4.1.

1. Approximate the prior π with the finite dimensional vector $\mu\in\Delta^J$, 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

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

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

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

problem is then

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

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

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

C.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 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 $\tilde{\delta}^{\text{adapt}}(T_O;\rho)$ 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 $\tilde{\delta}\left(T_O;-0.524\right)$ based on spline interpolation at $\rho^2=(-0.524)^2$ together with the observed test statistic T_O based on the 2D grid of ρ^2 and t values.

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

C.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 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 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

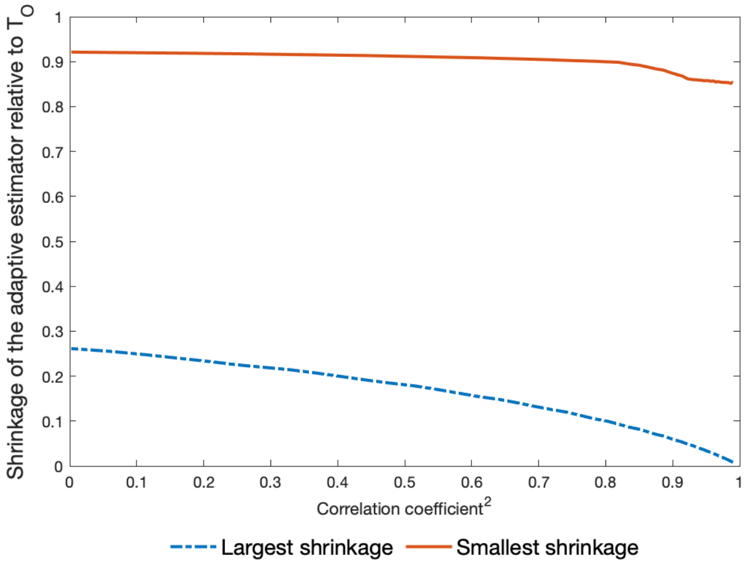

Figure A6: Shrinkage pattern for the adaptive estimator

distribution. Then for any $a < b$, we have

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

C.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 ρ^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 λ, and then with the observed test statistic T_O.

C.5.2 Hard thresholding

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

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

where the last term greatly simplifies due to Equation (24).

Appendix D Details of (LaLonde, 1986) example

In Section 5.4, 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.

D.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.

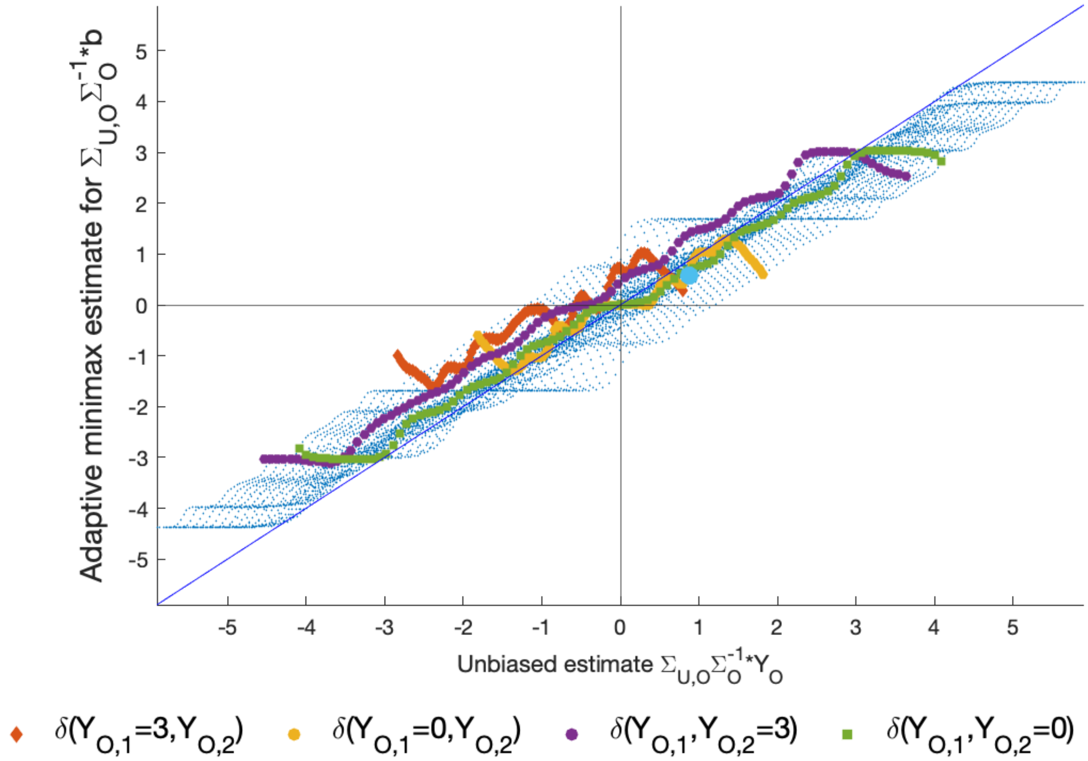

Figure A7: The adaptive minimax estimator of bias are illustrated by blue dots in the background, plotted against the their unbiased counterparts. The highlights are the estimates holding Y_{O1} and Y_{O2} constant respectively. In particular, the big blue dot highlights the adaptive estimate for the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

D.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 ∞.

Table A1 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.

	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

Table A1: 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.

	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

Table A2: 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.

D.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_{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 A2 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 $\mathcal{B}$.