Adapting to Misspecification

Timothy Armstrong, Patrick Kline, and Liyang Sun*

February 2023

Abstract

Empirical research typically involves an efficiency-robustness 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 if the assumptions are violated, or they can relax some of these assumptions to motivate a more variable unrestricted estimator that is asymptotically unbiased. 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 is not known, or when that bound may not be sharp, we propose shrinkage estimators that are adaptive: they minimize the percentage increase in worst case risk relative to an oracle that knows the magnitude of the restricted estimator's bias. We show how to compute the adaptive estimator by solving for a least favorable prior in a weighted convex minimax problem. A simple lookup table is provided for computing the adaptive estimates from the restricted and unrestricted estimates, their standard errors, and their correlation. We revisit several influential empirical papers and study how estimates of economic parameters change when adapting to misspecification.

*We thank Isaiah Andrews, Manuel Arellano, Stéphane Bonhomme, Bryan Graham, 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 audiences at Aarhus Workshop Econometrics I, CEMFI Fall Econometrics Conference, and ASSA 2023. Liyang Sun gratefully acknowledges support from the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010.

1 Introduction

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful. – George Box (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 George Box 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. Supposing this noise is normally distributed, however, the researcher is unable to infer a bound on the bias's magnitude.

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 *minimax* under this constraint: it is the function of the restricted and unrestricted estimators that minimizes worst case risk subject to the bound B. We use the shorthand "B-minimax" to refer to this estimator. It has a particularly simple structure that can be shown to correspond to a Bayes estimator under a discrete least favorable prior on the bias. 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 efficiently weighted GMM. 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 maximal risk an estimation procedure yields over the oracle's B-minimax procedure. Intuitively, adaptation regret captures the regret an ambiguity averse researcher feels over having exposed themselves to an unnecessarily high level of maximal risk. 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 procedure achieves worst case risk near that of the oracle regardless of the true bias magnitude and can be thought of as a conventional minimax 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 GMM estimate plus a "shrinkage" estimate of the scaled bias. As with the B-minimax estimator, the shrinkage estimate can be viewed as a Bayes estimate of bias involving 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. An appealing feature of the prior is that it depends only on the correlation between the restricted and unrestricted estimators. Enumerating these solutions 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 lookup table, it is not analytic. Building on the intuition in Efron and Morris (1972), 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. 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://lsun20.github.io/MissAdapt.

To illustrate the advantages of adapting to, rather than testing for, misspecification, we revisit three 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 final example, from Angrist and Krueger (1991), considers whether to instrument for years of schooling when estimating the returns to education.

Related literature Our analysis builds on seminal 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, 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 (see Johnstone (2015), Tsybakov (2009) and references therein). 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, our framework applies to other definitions of this problem, and we consider in detail an extension that places a bound on worst-case risk under the unconstrained parameter space.

While this high dimensional literature has focused on asymptotic rates and constants (with the papers by Hodges, Lehmann and Bickel cited above standing out as an important exception), we focus on the 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, including Chamberlain (2000), Elliott et al. (2015), Müller and Wang (2019) and Kline and Walters (2019), among others.

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 not needed in linear settings, which include many of our applications. In particular, the proposed adaptive procedures give global risk guarantees in these settings. 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; Hjort and Claeskens, 2003). In contrast to recent Empirical Bayesian proposals engineered for forecasting problems (Hansen, 2007; Hansen and Racine, 2012) our analysis considers a scalar estimand, which renders Stein style shrinkage arguments inapplicable.

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 of our 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 we note that extensions to vector-valued b are possible as well.

It will be convenient to work with $Y_O = Y_R - Y_U$, which gives an estimate of the bias 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.

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:

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

Here, 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 we can do using only the constraint $(\theta, b) \in \mathcal{C}_B$. To achieve this benchmark, the researcher must specify an appropriate parameter space $\mathcal{C}_B$ in order to calculate the B-minimax estimator. In our main example, the parameter spaces are indexed by an a priori bound on the bias magnitude $|b|$ of the constrained estimator Y_R.

How much do we have to 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 who knows B and is able to compute the B-minimax estimator, this estimator 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 $\mathcal{C}_B$. The minimum possible worst-case adaptation regret 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) we refer to $A^*(\mathcal{B})$ as the *loss of efficiency under adaptation.* We refer to an estimator that achieves this bound as *optimally adaptive*, and we use the notation δ^{adapt} for this estimator. To measure the efficiency of an ad hoc estimator δ relative to the optimally adaptive estimator, we can compute the quantity

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

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 $\mathcal{B} = [0, \infty]$.

Remark 2.1. 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 $B \in \mathcal{B}$, without knowledge of B, and the loss of efficiency under adaptation gives the reciprocol of this best possible simultaneous relative efficiency.

Remark 2.2. The general question of adaptation considered in the literature can be phrased in our setting as obtaining a single estimator δ that is "nearly B-minimax" for all $B \in \mathcal{B}$. We obtain the specific setup above by defining "near" to mean "up to the smallest uniform multiplicative factor." This gives a simple and intuitive setting in which statements about adaptation correspond directly to relative efficiency statements, as described in Remark 2.1.

While we consider this a useful baseline case, the approach in this paper extends to other ways of setting up the problem such as constraining increase in the worst case risk of the estimator to be small relative to the unbiased estimator.

3 Computing adaptive estimators

To compute the optimally adaptive estimator, we must solve (1). As we now show, this can be phrased as a minimax problem with a scaled loss function, thereby allowing us to leverage results from the literature on computation of minimax estimators.

3.1 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 B,θ,b such that $B\in\mathcal{B}$ and $(\theta,b)\in\cup_{B'\in\mathcal{B}}\mathcal{C}_{B'}$. 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 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 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, the optimal adaptive estimator can be computed via the following algorithm, once we have a general way of computing minimax estimators.

Algorithm 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.

3.2 Computing minimax estimators

Algorithm 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*. This can be phrased as a convex optimization problem over distributions on (θ, b), which can be solved 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 (2019) for recent applications in econometrics). In this section, we briefly summarize some of the relevant ideas as they apply to our setting, leaving details for the appendix. While we treat the general problem in this section, invariance can also be used to further simplify the problem. Indeed, in our main example, we use the fact that minimax and adaptive estimators are invariant to certain transformations to reduce the problem to finding a least favorable prior over b, with a flat "prior" on θ.

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 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 (this may require allowing randomized decisions in general, but the estimation problems we consider will be such that the Bayes decision is nonrandomized), and the set Γ is convex. 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 $(\pi(\vartheta_1), \dots, \pi(\vartheta_J))$ for a grid of J values of ϑ, following Chamberlain (2000).

3.3 Computation 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 3.2. Details and formal statements are given in Appendix A. These results apply to general loss functions for estimation of the form $L(\theta, b, \delta) = \ell(\theta - \delta)$, but we focus in the main text on the case of squared error loss $L(\theta, b, \delta) = (\theta - \delta)^2$.

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

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

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

$$\rho\sqrt{\Sigma_U}\delta\left(T_O\right) + Y_U - \rho\sqrt{\Sigma_U}T_O \tag{3}$$

with $\delta\left(T_O\right)$ given by $\delta^{\text{BNM}}\left(T_O; \frac{B}{\sqrt{\Sigma_O}}\right)$ where $\delta^{\text{BNM}}(y; \tau)$ denotes the minimax estimator in the bounded normal mean problem, in which we observe $Y \sim N(\vartheta, 1)$ and impose the parameter space $\mathcal{C} = [-\tau, \tau]$ on ϑ. Furthermore, the B-minimax risk is given by

$$R^*(B) = \rho^2\Sigma_U r^{\text{BNM}}\left(\frac{B}{\sqrt{\Sigma_O}}\right) + \Sigma_U - \rho^2\Sigma_U \tag{4}$$

where $r^{\text{BNM}}(\tau)$ denotes minimax risk in the bounded normal mean problem given above.

We compute $r^{\mathrm{BNM}}(\tau)$ by computing a least favorable prior on a grid approximating $[-\tau,\tau]$, following the methods described in Section 3.2 above. The bounded normal means problem has been considered in several papers and applications to other minimax problems; see Lehmann and Casella (1998, Section 9.7(i), p. 425).

The scaling function (2) can now be written $\omega(\theta,b) = R^*(|b|)$, where R^* for our problem is given in (4). 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 (3), but with δ given by the estimator $\tilde{\delta}^{\mathrm{adapt}}(t;\rho)$, which minimizes

$$\max_{\tilde{b}\in\mathbb{R}} \frac{E_{T\sim N(\tilde{b},1)}(\tilde{\delta}(T)-\tilde{b})^2+\rho^{-2}-1}{r^{\mathrm{BNM}}(|\tilde{b}|)+\rho^{-2}-1} \tag{5}$$

The loss of efficiency under adaptation $A^*([0,\infty])$ is then given by the minimized value of (5). 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.

The main computational step of computing $\tilde{\delta}^{\mathrm{adapt}}(t;\rho)$ can be performed once for each value of the scalar parameter ρ and tabulated. Thus, in our main application, the optimally adaptive estimator is easily computed using a lookup table.

To get some intuition for the form of the B-minimax estimator and the optimally adaptive estimator, note that $Y_U - \frac{\rho\sqrt{\Sigma_U}}{\sqrt{\Sigma_O}}Y_O$ is the optimal GMM estimator of θ under the restriction $b=0$. In particular, if $\rho\sqrt{\Sigma_O}\sqrt{\Sigma_U} = -\Sigma_O$, then optimal GMM is simply the restricted estimator Y_R, which is efficient in this case. If $b \neq 0$, then this estimator will have bias $\frac{\rho\sqrt{\Sigma_U}}{\sqrt{\Sigma_O}}b$. The estimator in (3) adds the estimate $\rho\sqrt{\Sigma_U}\delta\left(\frac{Y_O}{\sqrt{\Sigma_O}}\right)$ of this bias term $\frac{\rho\sqrt{\Sigma_U}}{\sqrt{\Sigma_O}}b$. In particular, $\delta(Y_O/\sqrt{\Sigma_O})$ is an estimate of $b/\sqrt{\Sigma_O}$. 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 $\delta(\cdot)$ to be an estimator that attempts to adapt to different values of B in this constraint.

3.4 Simple "nearly adaptive" estimators

While the optimally adaptive estimator is easy to compute using convex programming and even easier 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 (3) 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

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

which leads to the estimator

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

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 (5) 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 discussed further in Section 4, we find that soft thresholding yields nearly optimal performance for the adaptation problem relative to the optimally adaptive estimators. In contrast, hard thresholding performs much worse. This mirrors 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 ∞.

4 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 specification. In practice, it is often difficult to commit to a restricted estimator when a specification test clearly rejects, as referees and others may demand that the rejected model be discarded. Likewise, commitment to an unrestricted specification can be complicated by the difficulty of publishing imprecise results.

While pre-analysis plans can be used to deter the tendency to hide results, the editorial process may force a researcher to emphasize one specification over another. The adaptive approach counters this tendency by combining information across reported specifications. Because the only inputs required to compute the adaptive estimator are the restricted and unrestricted estimators themselves along with their 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.

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

We begin with 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 OLS estimates of μ_ℓ including the trend as Y_U and the estimates omitting the trend as Y_R. 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							Soft-	Pre-
hospital		Y_U	Y_R	Y_O	GMM	Adaptive	threshold	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. "Yrs since hospital" refers to years since hospitalization. "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 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.

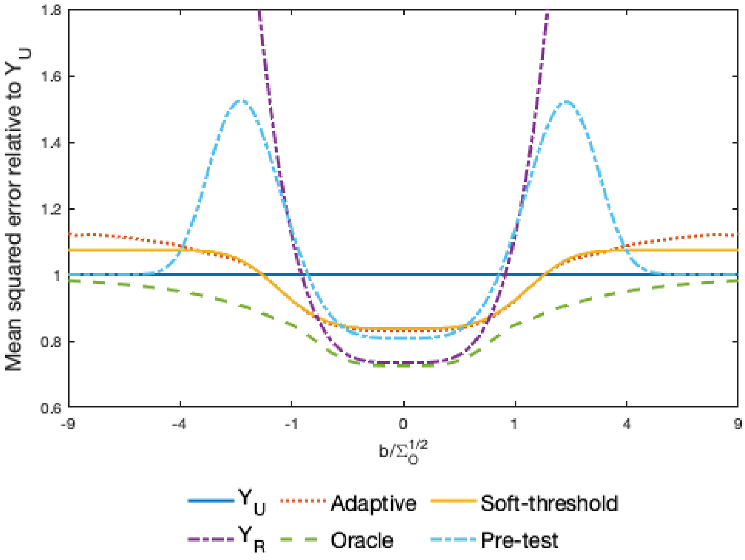

Figure 1: Risk functions for μ_0 ($\rho = -0.524$)

To assess the tradeoffs involved in adapting to misspecification, Figure 1 depicts the risk functions of the various estimation approaches listed in the first row of Table 1. Here, the correlation coefficient ρ between Y_R and Y_U 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 bias is moderate 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 adaptation penalty. 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 that of the pre-test estimator.

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,

		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 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.

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 the following Figure 2 depicting the risk functions of the respective 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. It seems likely most researchers would view this tradeoff favorably, leading us to recommend constrained adaptation as a default option. 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.

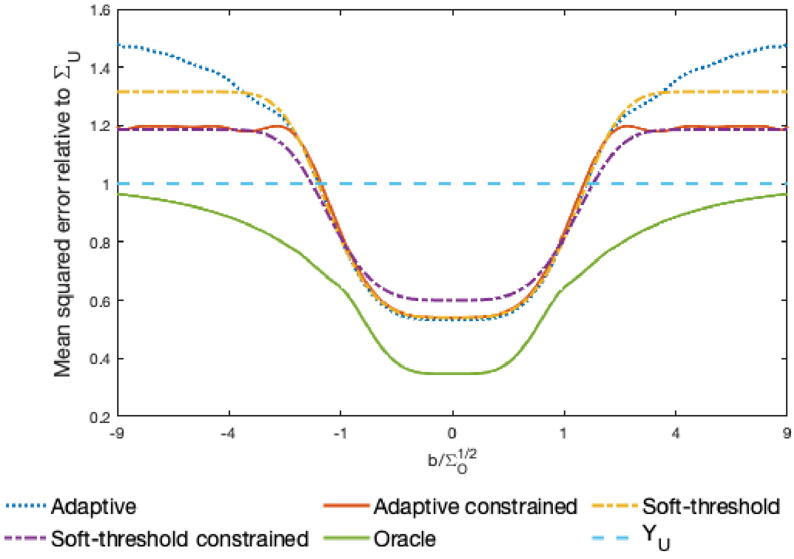

Figure 2: Risk functions for μ_3 ($\rho = -0.813$)

4.2 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 markdowns. Y_U is taken as the average markdown implied by optimally weighted GMM estimation using a set of demand-side instruments, while Y_R is the GMM estimate adding to the demand side instruments a set of supply-side instruments.

	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)	(3.17)			
Max Regret	96%	∞		32%	34%	107%
Threshold					0.59	1.96

Table 3: Adaptive estimates for the average markup (in percent). "Max Regret" refers to *worst-case* adaptation regret in percentage terms $(A_{\max}(\delta) - 1) \times 100$. The correlation coefficient is $\rho = -0.7$.

Table 3 lists estimates under different estimation approaches. While adding the supply side instruments reduces the standard errors by nearly 30%, the difference Y_O between the restricted and unrestricted estimates is large and statistically significant, with $T_O \approx 6$. 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.

4.3 Adapting to endogeneity (Angrist and Krueger, 1991)

In a landmark study, Angrist and Krueger (1991) estimated 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 4 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.

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 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).

	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 4: Returns to schooling. Standard errors in parentheses. "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 $\rho = -0.9998$.

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 estimator wish to avoid the regret associated with missing out on the enormous efficiency gains if OLS 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, this is a setting where it is wise to take to other considerations into account. Committing to Y_R exposes the researcher to potentially unlimited risk. The adaptive and soft-thresholding estimators avoid commitment but still expose the researcher to an approximately five fold maximal risk increase. As shown in Table 5, if we instead follow our rule of thumb of limiting ourselves to a 20% increase in maximal risk, we find that 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, reflect the potential efficiency costs of weighting Y_U so heavily. These efficiency concerns are outweighed in this case by the potential for extremely large biases.

	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 5: Adaptive estimates of returns to schooling. "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 coefficient is $\rho = -0.9998$.

5 Conclusion

Empirical research inevitably involves 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 adaptive estimates requires only point estimates, standard errors, and the correlation between estimators, objects that are easily produced by standard statistical packages.

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.

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. (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.

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.

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

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.

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.

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. (2015). *Gaussian estimation: Sequence and wavelet models.* Online manuscript available at http://statweb.stanford.edu/˜imj/.

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

Kline, P. and C. Walters (2019, July). Audits as Evidence: Experiments, Ensembles, and Enforcement. *arXiv:1907.06622 [econ, stat]*. arXiv: 1907.06622.

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

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

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

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

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.

A Details for main example

We provide details and formal results for the results in Section 3.3 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 3.3, using a weighted loss function and Lemma 1 to obtain the optimally adaptive estimator.

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

$$Y = \begin{pmatrix} \underset{1\times 1}{Y_U} \\ \underset{p\times 1}{Y_O} \end{pmatrix} \sim N\left(\begin{pmatrix} \underset{1\times 1}{\theta} \\ \underset{p\times 1}{b} \end{pmatrix}, \Sigma \right), \quad \Sigma = \begin{pmatrix} \underset{1\times 1}{\Sigma_U} & \underset{1\times p}{\Sigma_{UO}} \\ \underset{p\times 1}{\Sigma_{UO}'} & \underset{p\times p}{\Sigma_O} \end{pmatrix}. \tag{6}$$

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 1. We assume that the loss function satisfies the invariance condition

$$L(\theta+t,b,d+t)=L(\theta,b,d) \quad \text{all } t\in\mathbb{R}. \tag{7}$$

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

$$\inf_{\delta}\sup_{\theta\in\mathbb{R},b\in\mathcal{C}} R(\theta,b,\delta). \tag{8}$$

Theorem 1. *Suppose that the loss function $L(\theta,b,d)$ is convex in d and that (7) holds. Then the minimax risk (8) is given by*

$$\begin{aligned}&\inf_{\bar\delta}\sup_{b\in\mathcal{C}} E_{0,b}[\tilde L(b,\bar\delta(Y_O)-\Sigma_{UO}\Sigma_O^{-1}b)] \qquad (9)\\ &=\sup_{\pi \textit{ supported on } \mathcal{C}}\ \inf_{\bar\delta}\int E_{0,b}[\tilde L(b,\bar\delta(Y_O)-\Sigma_{UO}\Sigma_O^{-1}b)]\,d\pi(b)\end{aligned}$$

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 (8) 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 (9).

Proof. The minimax problem (8) 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 transormation 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

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

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 (9), 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 (8). 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 (9).

It remains to show that the infimum in the first line of (9) is achieved, and that the equality claimed in (9) holds. The equality in (9) 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 (9) 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 specialize this result to derive the formula for the minimax estimator and the optimally adaptive estimator under squared error loss in Section 3.3. 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 (8) 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, (9) 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 $|t| \leq B/\sqrt{\Sigma_O}$. 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

$$\frac{\Sigma_{UO}^2}{\Sigma_O} r^{\text{BNM}}(B/\sqrt{\Sigma_O}) + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O} \tag{10}$$

and the B-minimax estimator as

$$\frac{\Sigma_{UO}}{\sqrt{\Sigma_O}}\delta^{\mathrm{BNM}}(Y_O/\sqrt{\Sigma_O}; B/\sqrt{\Sigma_O}) + Y_U - \frac{\Sigma_{UO}}{\Sigma_O}Y_O. \tag{11}$$

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 (3) and (4).

We summarize these results in a the following corollary.

Corollary 1. *In the case where the dimension p of Y_O is 1, $L(\theta, b, t) = (\theta - t)^2$ and $\mathcal{C} = [-B, B]$, the minimax problem (8) has a solution δ given by (11). The minimax risk is given by (10).*

To find the optimally adaptive estimator and loss of efficiency under adaptation in our main example, we apply Lemma 1 with $\omega(\theta, b) = R^*(|b|)^{-1}$, with $R^*(B)$ given by (10). This leads to the minimax problem (8) with $\mathcal{C} = \mathbb{R}$ and $L(\theta, b, d) = R^*(|b|)^{-1}(\theta - d)^2$. The function $\tilde{L}$ in Theorem 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 (9) as

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

This is minimized by $\bar{\delta}$ satisfying $\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}}\bar{\delta}(Y_O) = \tilde{\delta}^{\mathrm{adapt}}(Y_O/\sqrt{\Sigma}; \rho)$ where $\tilde{\delta}^{\mathrm{adapt}}(T; \rho)$ is a solution to

$$\inf_{\tilde{\delta}} \sup_{\tilde{b}\in\mathbb{R}} \frac{E_{T\sim N(\tilde{b},1)}\left[\left(\tilde{\delta}(T) - \tilde{b}\right)^2\right] + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\tilde{b}|) + \rho^{-2} - 1}. \tag{12}$$

By Theorem 1, the optimally adaptive estimator is given by

$$\frac{\Sigma_{UO}}{\sqrt{\Sigma_O}}\tilde{\delta}^{\mathrm{adapt}}(Y_O/\sqrt{\Sigma}; \rho) + Y_U - \frac{\Sigma_{UO}}{\Sigma_O}Y_O = \rho\sqrt{\Sigma_U}\tilde{\delta}^{\mathrm{adapt}}(T_O; \rho) + Y_U - \rho\sqrt{\Sigma_U}T_O. \tag{13}$$

We summarize these results in the following corollary.

Corollary 2. *For adaptation over the parameter spaces $\mathcal{C}_B = \mathbb{R} \times [-B, B]$ in the main example, the loss of efficiency under adaptation is given by the value of (12), and an optimally adaptive estimator is given by the formula in (13).*