The Incredible Flexibility of Moment Matching

Isaiah Andrews and Bas Sanders*

Abstract

We consider parametric models estimated by matching researcher-selected moments of the data to their model-implied analogs. We ask how far the choice of moments to match can push estimates when the model is misspecified. The answer is: very far. Under regularity conditions, an adversarial researcher who knows the data distribution can choose moments making any parameter value the unique moment-matching estimand. Moreover, in some cases they can do so with model-implied standard errors close to those maximum likelihood would report at the same point. We illustrate in a menu-cost model.

Keywords: Moment matching, structural estimation, model misspecification

JEL Codes: C13, C18, C51

1 Introduction

Economists often view structural models as approximations rather than literal descriptions of reality. Motivated in part by this view, researchers sometimes estimate models by matching a set of hand-selected moments rather than using more traditional statistical approaches like maximum likelihood.[1] This practice affords researchers the freedom to base estimation on the predictions of the model that they deem most economically important, and thus seems potentially appealing in settings where researchers think their models may be wrong.

*First version: June 2026. This version: November 2026. Andrews: MIT Department of Economics and NBER; iandrews@mit.edu. Sanders: Coefficient Giving; bas.sanders@coefficientgiving.org. We thank Pat Kline, Marc Melitz, Mikkel Plagborg-Møller, Dominic Russel and Jesse Shapiro for helpful comments. We thank the Kilts Center for Marketing at the University of Chicago Booth School of Business for making the Dominick's data available. We used multiple AI systems in the production of this paper, including for assistance writing and checking the text, proofs, and code.

[1]For instance, Bordalo et al. (2020) write that "We prefer [moment matching] to maximum likelihood for two reasons. First, one advantage of our model is that it is simple and transparent. However, this simplicity comes at the cost of likely misspecification, and it is well known that with misspecification concerns moment estimators are often more reliable."

By moment matching we will mean the practice of estimating the unknown parameters θ in a parametric model by minimizing the (weighted) distance between the sample average of a set of researcher-chosen moments $g(X_i)$ and their model-implied expectation $\mathrm{E}_\theta[g(X)]$, which is often computed by simulation. Parametric models allow this approach to be used for any (suitably integrable) choice of g, but also typically imply the availability of more efficient alternatives such as maximum likelihood. The flexibility to choose moments introduces researcher degrees of freedom (Simmons et al., 2011) which could in principle be used to steer estimates toward a desired conclusion. In this paper, we characterize the scope for such manipulation by asking what results an adversarial researcher could engineer through their choice of moments when the model is misspecified (in the sense that the true data distribution is inconsistent with the model). Our main finding is an "anything-goes" result for moment matching in misspecified models: under some conditions, for any pre-specified parameter value there exists a choice of moments which renders that value the unique solution to the population moment-matching problem. Thus, an adversarial researcher who knows the true data-generating process can construct moments delivering any estimand they wish.

One might wonder if such contrived moment choices will generate large standard errors and thus be statistically unconvincing, but this need not be the case. We characterize the smallest possible model-implied asymptotic variance subject to delivering a given estimand, and compare it to the smallest model-implied variance that any regular estimator may have at that parameter value, which (thanks to the use of model-implied standard errors) corresponds to the variance maximum likelihood would report if it delivered the same estimate. The resulting variance ratio depends on the model and the form of misspecification, but equals one when a particular statistical discrepancy (chi-squared divergence) between the data and the model is infinite at the target parameter value, for instance because the data have heavier tails than the model. We find that it is sometimes close to one in other examples as well. Matters are even worse if the researcher may choose their moments after observing the data.[2]

Our results give an upper bound on what an adversarial researcher could achieve. A researcher facing further constraints, for instance due to limits on what moments may be used without raising suspicons, or a need to pre-commit to a choice of moments before the data-generating process or sample are fully known, would have correspondingly less flexibility. Our main results are also specific to the case of moment matching in a parametric model, where the choice of g affords an adversarial researcher especially high scope for

[2]Specifically, when the model implies continuously distributed data, a researcher who chooses the moments after observing the sample can engineer any estimate value they wish, along with model-implied standard errors that exactly match those maximum likelihood would report at that estimate.

manipulation. This differs from general GMM settings (Hansen, 1982) in which the model implies only a fixed set of unconditional moment equalities, though a researcher may still have the freedom to choose how to weight them.[3] In many prominent applications of GMM, however, the unconditional moments originate from a set of conditional moment restrictions together with researcher-selected instrument functions. We show in the Online Appendix that in misspecified conditional moment equality models, an adversarial researcher may again (under regularity conditions) select any estimand they wish by choosing appropriate instruments.

To illustrate our results, following Cocci and Plagborg-Møller (2025) we estimate the model of Alvarez and Lippi (2014), who study menu-cost price setting in multiproduct firms. Fitting Alvarez and Lippi (2014)'s model to scanner data used by Cocci and Plagborg-Møller (2025), we find that considering alternative, intuitively reasonable moments generates a broad range of point estimates, where the range of estimates is wide both economically and relative to the degree of statistical uncertainty. Going further, we apply our adversarial moment construction in these data, and confirm that an adversarial researcher could steer estimates toward desired values.

We close the paper with a discussion of what steps might be taken to mitigate the concerns we raise. We recommend that researchers report estimators motivated by efficiency considerations under the model, for instance the maximum likelihood estimator when feasible, since these provide a natural point of comparison across papers without the additional variation introduced by moment choice. We think it is natural to complement such model-efficient estimators with both specification tests and alternative estimators motivated by misspecification concerns, but recommend that such concerns be articulated as precisely as possible, and that a formal argument be made for why the particular misspecification concerns in a given application lead to the chosen estimator. A growing literature in econometrics, including Armstrong and Kolesár (2021), Bonhomme and Weidner (2022), Christensen and Connault (2023), Adusumilli (2026), and Andrews et al. (2026), provides formal tools for such analyses, some of which build on the partial identification literature (e.g. Manski, 1989; Molinari, 2020).

Our paper contributes to a literature on best practices for moment matching, a discussion that traces back at least to the calibration literature (Prescott, 1986; Hansen and Heckman, 1996; Cooley, 1997) and that has long recognized the importance of moment choice. For example, Wooldridge (2001) notes: "One problem is that the researcher must choose the ad-

[3] Andrews et al. 2025 characterize the resulting scope for data-driven manipulation.

ditional moment conditions to be added in an ad hoc manner." Prior work in economics, for instance Leamer (1974, 1978) and Lovell (1983), has noted both the prevalence and the dangers of specification search, though without a focus on the moment-based estimation setting we study here. Other work has proposed methods for systematically constructing informative moments. For example, Gallant and Tauchen (1996) propose generating moment conditions from the score of an auxiliary model that approximates the data distribution. More recent work has focused on statistical inference and efficiency in calibration and moment-matching exercises. Cocci and Plagborg-Møller (2025) observe that calibration can be interpreted as minimum distance estimation and derive conservative standard errors when the covariance structure of the empirical moments is unknown. Kaji et al. (2023) introduce adversarial estimation, which estimates structural models by using a discriminator to learn features that best distinguish simulated from observed data. A related literature studies GMM under misspecification, where the estimand depends on the weighting matrix (Hall and Inoue, 2003).

The rest of the paper is organized as follows. Section 2 introduces our formal setting and characterizes the class of moments that yield a specific moment-matching estimand. Section 3 characterizes the attainable precision in this class. Section 4 illustrates the empirical consequences of moment choice in an example building on Alvarez and Lippi (2014) and Cocci and Plagborg-Møller (2025). Section 5 discusses routes to mitigate the issues suggested by our theoretical results. The Online Appendix collects all proofs, as well as supplementary results including those for conditional moment equality models.

2 The Flexibility of Moment Matching

2.1 Setting

Consider a researcher who observes i.i.d. draws $X_1, \ldots, X_n \in \mathcal{X}$ from a probability distribution P. The researcher postulates a dominated parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with parameter space $\Theta \subseteq \mathbb{R}^K$. For each $\theta \in \Theta$, let p_θ denote the density of P_θ with respect to dominating measure μ (which need not dominate P). The model is statistically well-specified if and only if $P \in \mathcal{P}$, or equivalently if there exists $\theta \in \Theta$ such that $P = P_\theta$.

The researcher estimates their model by moment matching. Specifically, they choose target moments $g : \mathcal{X} \to \mathbb{R}^D$ with $D \geq K$ and a (possibly data-dependent) symmetric

positive definite weighting matrix $W_n \in \mathbb{R}^{D\times D}$.[4] Define the *sample moment function* as the difference between the sample average of the target moments and their model-implied mean

$$m_n(g,\theta) := \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathrm{E}_\theta[g(X)],$$

where E_θ denotes the expectation when $X \sim P_\theta$. The researcher estimates θ by minimizing the W_n-weighted distance of the sample moments from zero, yielding the set of minimizers

$$\hat{\Theta}_g(W_n) := \arg\min_{\vartheta\in\Theta} m_n(g,\vartheta)^\top W_n m_n(g,\vartheta). \tag{1}$$

As is conventional, we define population analogs by replacing sample averages with expectations under the true data distribution P. Define the *population moment function* as

$$m(g,\theta) := \mathrm{E}_P[g(X)] - \mathrm{E}_\theta[g(X)].$$

Analogously, provided $W_n \xrightarrow{p} W = W(P)$ as $n\to\infty$ for some symmetric positive definite matrix $W \in \mathbb{R}^{D\times D}$, we define the population *pseudo-true set* as

$$\Theta_g(W) := \arg\min_{\vartheta\in\Theta} m(g,\vartheta)^\top W m(g,\vartheta).$$

Under misspecification, $\Theta_g(W)$ generally depends on both the choice of moments g and the weight matrix W.

In the special case where the moments g admit "perfect fit", meaning that there exists $\theta\in\Theta$ such that $m(g,\theta)=0$, define the set of population solutions to the moments as

$$\Theta_g := \{\vartheta\in\Theta : m(g,\vartheta)=0\}.$$

When the perfect-fit condition holds, $\Theta_g(W)=\Theta_g$ for all positive definite W, so the choice of weights no longer matters for the estimand. Since our interest in this paper is in the choice of moments, rather than the choice of weights, we limit attention to g satisfying the perfect-fit condition. Since this constrains the choice set, relaxing this constraint only increases researcher degrees of freedom.[5] Maintaining this condition also ensures that the model's

[4] Note that g does not depend on θ.

[5] Even if the perfect-fit condition fails, any interior minimizer of the weighted problem will satisfy the perfect-fit condition for new moments formed based on the first-order condition $\frac{\partial m(g,\vartheta)}{\partial\vartheta}^\top W m(g,\vartheta) = 0$. See e.g. Andrews et al. (2025) for a complementary discussion of how weight choices shape estimates and

over-identifying restrictions hold so long as the researcher limits attention to the moments g.

2.2 Moment Engineering

Consider a motivated researcher who would like to report estimates near a particular value $\bar{\theta}$. Since the moment-matching estimand Θ_g depends on the chosen moments g, presumably some choices of moments will deliver estimands closer to the target value than others. Absent other constraints (e.g. from scientific integrity or conventions on the form of "acceptable" moments), how far could such a researcher get?

To answer this question, we consider the extreme case of a fully adversarial researcher who engineers their moments to make $\bar{\theta}$ a population solution of the moment conditions. For the time being we further assume that this researcher knows the true data-generating process P, though we return to this point below. Restricting attention to moments which are square-integrable under $P_{\bar{\theta}}$ and integrable under P, define

$$\mathcal{C}(\bar{\theta}) := \left\{ g \in L^2(P_{\bar{\theta}}; \mathbb{R}^D) \cap L^1(P; \mathbb{R}^D) : \mathrm{E}_P[g(X)] = \mathrm{E}_{\bar{\theta}}[g(X)] \right\}.$$

In words, $\mathcal{C}\left(\bar{\theta}\right)$ collects all $\mathbb{R}^D$-valued moment functions whose population mean under the true distribution P matches their mean under the model-implied distribution $P_{\bar{\theta}}$.

The set $\mathcal{C}(\bar{\theta})$ is nonempty since it always contains constant functions. As constant functions illustrate, however, many moments in $\mathcal{C}(\bar{\theta})$ fail to uniquely pin down $\bar{\theta}$. We next show that if the model is misspecified then under regularity conditions an adversarial researcher can eliminate competing solutions by adding finitely many moments in $\mathcal{C}\left(\bar{\theta}\right)$. To state this result, let $s_\theta = \frac{\partial}{\partial \theta} \log(p_\theta)$ denote the score of the model at θ, and $I_\theta := \mathrm{E}_\theta[s_\theta(X) s_\theta(X)^\top]$ the Fisher information. We impose the following conditions.

Assumption 1 *Suppose that:*

(i) Θ *is compact.*

(ii) P_θ *is continuous in* θ *with respect to total variation.*

(iii) The model is differentiable in quadratic mean (DQM; see e.g. van der Vaart, 1998, Chapter 7) at $\bar{\theta}$*, with score* $s_{\bar{\theta}} \in L^2(P_{\bar{\theta}}; \mathbb{R}^K)$ *such that* $\mathrm{E}_{\bar{\theta}}[s_{\bar{\theta}}(X)] = 0$ *and* $I_{\bar{\theta}}$ *is positive definite.*

estimands for a fixed set of moments.

(iv) There exists no $v \in \mathbb{R}^K$ such that P has density $(1 + v^\top s_{\bar\theta})p_{\bar\theta}$ with respect to μ.

(v) For every $\theta \neq \bar\theta$, none of $P_{\bar\theta}$, P_θ, or P is a (possibly degenerate) mixture of the other two.

Condition (i) requires compactness of the parameter space, which is a common regularity condition for nonlinear estimation results. Conditions (ii) and (iii) are standard continuity, smoothness, and rank conditions on the model and do not restrict P.[6] Condition (iv) requires that P not lie in a K-dimensional set of distributions spanned by the score. Loosely speaking, this means that there is not some direction in which we can move θ, local to $\bar\theta$, so that the change in the model-implied distribution is proportional to $P - P_{\bar\theta}$. Finally, condition (v) is a global condition on the model and implies that (a) the model is misspecified (in the sense that $P \notin \mathcal{P}$) and (b) the model is identified at $\bar\theta$ in the sense that there does not exist $\theta \neq \bar\theta$ with $P_\theta = P_{\bar\theta}$.

Proposition 1 *Under Assumption 1, there exist $D < \infty$ and a bounded $g \in \mathcal{C}(\bar\theta)$ such that $\bar\theta$ is the unique solution to the population moment-matching problem, $\Theta_g = \{\bar\theta\}$. Moreover, it is possible to choose these moments such that $\frac{\partial}{\partial\theta}\mathrm{E}_\theta[g(X)]$ has rank K at $\theta = \bar\theta$.*

Proposition 1 establishes an "anything-goes" feature of moment matching in misspecified parametric models: under regularity conditions, one can construct moments so that any pre-specified parameter value $\bar\theta \in \Theta$ is the unique solution to the population moment conditions, and thus the unique pseudo-true value.

Example: normal location Suppose the model is $P_\theta = N(\theta, 1)$ with Θ a compact interval, while the data follow a standard Laplace distribution, with density $\frac{1}{2}e^{-|x|}$. The model is DQM with score $s_{\bar\theta}(x) = x - \bar\theta$ and $I_{\bar\theta} = 1$, so conditions (i)-(iii) hold, while Online Appendix S3 verifies conditions (iv) and (v). Hence, every $\bar\theta \in \Theta$ is the unique moment-matching estimand for some choice of moments. $\triangle$

Remark 1 (When can the conditions fail?) The restrictions placed on P by parts (iv) and (v) of Assumption 1 hold in many, but not all, settings. For a given target $\bar\theta$ the distributions P for which (iv) or (v) fails lie in a $(K+1)$-parameter family, so when X is continuously distributed the set of possible true distributions is, intuitively speaking, "large" relative to the set of model-implied distributions and there will often be many ways for Assumption 1 to hold. At the other extreme, when X is binary then even for $K = 1$ neither condition (iv)

[6] When $\bar\theta$ lies on the boundary of Θ, limits at $\bar\theta$ are taken along $\theta \in \Theta$.

nor condition (v) can hold. Indeed, in this case every moment in $\mathcal{C}(\bar{\theta})$ is constant so long as $P \neq P_{\bar{\theta}}$. Online Appendix S2 provides further results for contexts with discrete data.

3 Minimizing the Asymptotic Variance

One might hope that the negative result established by Proposition 1 would be more troubling in theory than in practice. In particular, it could be that the moment functions needed to engineer a particular pseudo-true value are either socially unacceptable (e.g. too obviously contrived to be taken seriously), or deliver estimates with such extreme imprecision that they cannot hope to be persuasive. This section addresses the latter possibility by characterizing the smallest possible asymptotic variance subject to targeting a given pseudo-true value.

3.1 Asymptotic Variance for a Fixed Moment Function

Proposition 1 shows that for any $\bar{\theta}$, there exists a finite set of moments that render $\bar{\theta}$ the unique population solution. Indeed, there are many such choices of moments, and different moments in this set imply different standard errors for the resulting estimates.

Under standard regularity conditions as in Newey and McFadden (1994), if we consider estimation based on moments g and weight matrix $W_n \to_p W$, then under the model

$$\sqrt{n}\left(\hat{\theta}_{g,W} - \theta\right) \overset{P_\theta}{\rightsquigarrow} N\left(0, \operatorname{AVar}\left(\hat{\theta}_{g,W}\right)\right), \tag{2}$$

where $\overset{P_\theta}{\rightsquigarrow}$ denotes convergence in distribution under P_θ, and

$$\operatorname{AVar}\left(\hat{\theta}_{g,W}\right) := (G^\top W G)^{-1} G^\top W \Omega W G (G^\top W G)^{-1} \tag{3}$$

for

$$G := \mathrm{E}_\theta\left[g(X) s_\theta(X)^\top\right], \quad \Omega := \operatorname{Var}_\theta\left(g(X)\right),$$

where we take $\operatorname{AVar}(\hat{\theta}_{g,W})$ to be infinite when $G^\top W G$ is singular. Model-implied standard errors are then based on plug-in estimates for Ω and G.[7] We focus on model-implied standard errors for two reasons: first, this yields a tractable expression for the variance that depends only on the chosen moments g and the parametric model. Second, such standard errors

[7]Under DQM at θ, $G = \frac{\partial}{\partial \theta} \mathrm{E}_\theta[g(X)]$ for bounded g or, more generally, when $\|g(X)\|^2$ is uniformly integrable under the model near θ (Lemma S1 in the Online Appendix). For other g this identity requires further regularity conditions.

are often used in practice, and are implicit, for example, in moment-matching papers that quantify uncertainty using the parametric bootstrap (e.g. Su and Judd 2012; Bourreau et al. 2021; Lagakos et al. 2023; Simonovska and Waugh 2025). Online Appendix S3.3 shows, that some of our results in this section also extend to robust standard errors.

3.2 Efficient Moment Function for Specific Pseudo-True Value

We next ask how precise an estimator can be when attention is restricted to moments that target a given pseudo-true value. As a first step, we note that it suffices to consider just-identifying sets of moments, i.e. taking $D = \dim(g) = \dim(\theta) = K$. In particular, observe that when $G^\top WG$ is nonsingular the variance in Equation (3) is the same as that obtained from the just-identifying moments $\tilde{g}(X) := (G^\top WG)^{-1}G^\top Wg(X)$, where if $\bar{\theta}$ solves the moments based on g, it also solves those based on $\tilde{g}$. Thus, to lower-bound the variance it suffices to consider the just-identified case. Since the weight matrix drops out, we write the resulting estimate as $\hat{\theta}_g$.

We thus seek the smallest $\mathrm{AVar}(\hat{\theta}_g)$ over just-identifying $g \in \mathcal{C}(\bar{\theta})$, evaluating the variance at $\theta = \bar{\theta}$. Thanks to our focus on the model-implied variance this problem admits a simple solution, which depends on the chi-squared divergence between P and $P_{\bar{\theta}}$. For distributions Q_1 and Q_2, this divergence is defined as

$$\chi^2(Q_1 \| Q_2) := \mathrm{E}_{Q_2}\left[\left(\frac{dQ_1}{dQ_2}(X) - 1\right)^2\right] \tag{4}$$

when Q_1 is absolutely continuous with respect to Q_2, and is infinite otherwise. When $\chi^2(P\|P_{\bar{\theta}}) < \infty$, let $r := \frac{dP}{dP_{\bar{\theta}}} - 1$ denote the recentered likelihood ratio. Further let $s_{\bar{\theta},\perp}$ denote the elementwise projection of $s_{\bar{\theta}}$ onto the $L^2(P_{\bar{\theta}})$ closure of $\mathcal{C}(\bar{\theta})$, $Q_{\bar{\theta}} := \mathrm{E}_{\bar{\theta}}[s_{\bar{\theta},\perp}(X)s_{\bar{\theta},\perp}(X)^\top]$, and $A \succeq B$ denote that $A - B$ is positive semidefinite.

Proposition 2 *Suppose Assumption 1(iii)-(iv) hold. Then $Q_{\bar{\theta}}$ is positive definite, and:*

(a) $\mathrm{AVar}(\hat{\theta}_{g,W}) \succeq Q_{\bar{\theta}}^{-1}$ *for every $g \in \mathcal{C}(\bar{\theta})$ and positive definite W, and for every $\varepsilon > 0$, there is a bounded, just-identifying $g \in \mathcal{C}(\bar{\theta})$ with* $\mathrm{AVar}(\hat{\theta}_g) \preceq (1+\varepsilon)Q_{\bar{\theta}}^{-1}$.

(b) If $\chi^2(P\|P_{\bar{\theta}}) < \infty$, then

$$s_{\bar{\theta},\perp} = s_{\bar{\theta}} - \frac{\mathrm{E}_P[s_{\bar{\theta}}(X)]}{\chi^2(P\|P_{\bar{\theta}})}r, \qquad Q_{\bar{\theta}} = I_{\bar{\theta}} - \frac{\mathrm{E}_P[s_{\bar{\theta}}(X)]\,\mathrm{E}_P[s_{\bar{\theta}}(X)]^\top}{\chi^2(P\|P_{\bar{\theta}})},$$

and the bound in (a) is attained by just-identifying moments $g^*_{\bar\theta} := As_{\bar\theta,\perp} + c$ *for invertible* $A \in \mathbb{R}^{K\times K}$ *and* $c \in \mathbb{R}^K$*. If* $\chi^2(P\|P_{\bar\theta}) = \infty$*, then* $s_{\bar\theta,\perp} = s_{\bar\theta}$ *and* $Q_{\bar\theta} = I_{\bar\theta}$*, though the bound need not be attained.*

(c) If further Assumption 1(i), (ii), and (v) hold, then for every $\varepsilon > 0$ *there exist a bounded (not necessarily just-identifying)* $g \in \mathcal{C}(\bar\theta)$ *and a positive definite* W *such that* $\Theta_g = \{\bar\theta\}$ *and* $\mathrm{AVar}(\hat\theta_{g,W}) \preceq (1+\varepsilon)Q^{-1}_{\bar\theta}$*.*

Proposition 2(a) shows that $Q^{-1}_{\bar\theta}$ lower-bounds the model-implied variance of estimators targeting $\bar\theta$, and that there exist bounded moments which approach this bound. Part (b) shows that the bound is attained when $\chi^2(P\|P_{\bar\theta}) < \infty$ and characterizes the optimal just-identifying moments, while when the divergence is infinite the optimal model-implied variance agrees with that of the MLE. Finally, part (c) shows that we can also approach the bound using over-identifying moments which pin down $\bar\theta$ as in Proposition 1.

With finite divergence, the efficient moments are (up to affine transformation) equal to $s_{\bar\theta}$ orthogonalized against the recentered likelihood ratio r. By contrast, with infinite divergence efficient moments need not exist, but the proof of Proposition 2 shows that bounded moment sequences which converge to the score $s_{\bar\theta}$ in $L^2(P_{\bar\theta})$ approach the efficiency bound. Infinite divergence can arise for a variety of reasons, but two leading cases are that (i) P is not absolutely continuous with respect to $P_{\bar\theta}$ or (ii) P has heavier tails than $P_{\bar\theta}$. In the former case, moments which are nonzero only on events which have zero probability only under $P_{\bar\theta}$ can shift $\mathrm{E}_P[g(X)]$ arbitrarily without affecting $\mathrm{E}_{\bar\theta}[g(X)]$ or $\mathrm{Var}_{\bar\theta}(g(X))$.[8] In the latter case, a similar effect can be attained using sequences of moments which change behavior only in the tails.

Example: normal location, continued In this example $\chi^2(P\|P_{\bar\theta}) = \infty$ for every $\bar\theta$ due to the heavier tails of the Laplace distribution. Consequently, Proposition 2 implies that for every target $\bar\theta$ there exist moment-matching estimators with model-implied asymptotic variance approaching that of the MLE.[9] To construct one class of near-efficient moments, suppose the adversary is restricted to functions of $X_M := \max\{\min\{X, M\}, -M\}$. The χ^2 divergence for the resulting marginal distributions is finite, so Proposition 2(b) gives efficient moments in this restricted class. Figure 1(a) plots these moments for $\bar\theta = 1$ and

[8]The same logic covers moment selection after observing the data, since we can treat the empirical distribution as a discrete P, so the divergence will be infinite when $P_{\bar\theta}$ is continuous.

[9]Online Appendix S3 shows that the variance bound is attained only at $\bar\theta = 0$, where one can simply use the MLE.

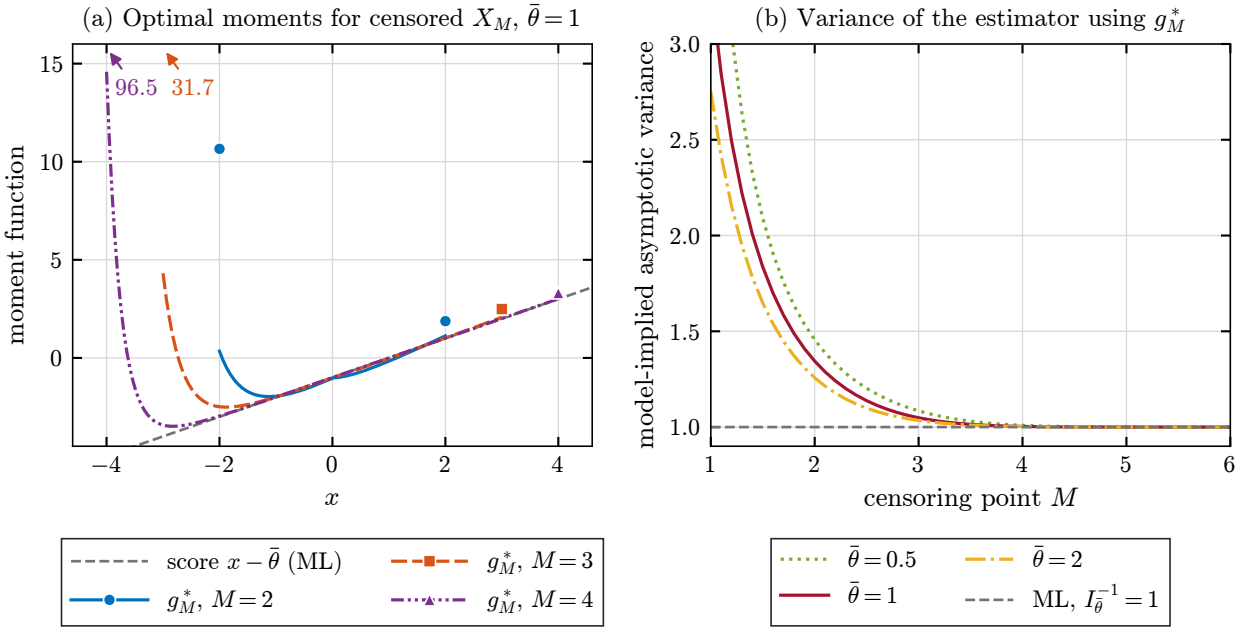


Figure 1: Normal-location example. Panel (a) plots the efficient moment g_M^* for the censored variable X_M at target $\bar{\theta} = 1$, for $M = 2, 3, 4$, normalized to have mean zero and derivative one under the model. Markers show the values at the censoring points $\pm M$, while numbered arrows show censoring points outside y-axis bounds. Panel (b) plots the model-implied asymptotic variance of the moment-matching estimator using g_M^* against M for $\bar{\theta} \in \{0.5, 1, 2\}$. Maximum likelihood has model-implied variance one.

$M \in \{2, 3, 4\}$, and shows that they largely track the score in the middle of the distribution, but behave differently near $-M$. The Laplace distribution assigns much more mass to X_M values at or near $-M$, so this substantially raises $\mathrm{E}_P[g_M^*(X_M)]$ while having little effect on $\mathrm{E}_{\bar{\theta}}[g_M^*(X_M)]$ and $\mathrm{Var}_{\bar{\theta}}(g_M^*(X_M))$. Panel (b) plots the variance of these constrained-efficient moment-matching estimators for different censoring values M and $\bar{\theta} \in \{0.5, 1, 2\}$. We see that the variance cost of targeting declines rapidly as we raise the value of M.[10] $\triangle$

3.3 Variance Cost Under Finite Divergence

We began this section by asking whether statistical considerations could mitigate the impact of moment engineering. In particular, is it the case that the variance inflation due to moment engineering substantially constrains the scope for an adversarial researcher to reach a given conclusion? Proposition 2 shows that when the χ^2 divergence is infinite, there is no variance

[10] If we instead use misspecification-robust standard errors, Online Appendix S3 shows that the variance of the g_M^*-based estimator grows with M. As an alternative, Proposition S3 derives the just-identified moments which minimize the robust asymptotic variance. We find that they yield a smaller robust variance than maximum likelihood for the targets $\bar{\theta}$ considered.

cost to estimand targeting, in the sense that there exist targeted estimators with model-implied variance arbitrarily close to the ML variance at the same point (which is in turn the smallest model-implied variance possible at that point). By contrast, when the divergence is finite there is generally some variance cost.

To quantify the efficiency loss for targeted estimators relative to maximum likelihood, define the *variance inflation factor* $VIF\left(\bar{\theta}\right)$ as

$$VIF\left(\bar{\theta}\right) := \sup_{\alpha \neq 0} \frac{\alpha^\top Q_{\bar{\theta}}^{-1} \alpha}{\alpha^\top I_{\bar{\theta}}^{-1} \alpha},$$

where the numerator corresponds to the minimum asymptotic variance for estimating $\alpha^\top \theta$ over estimators targeting $\bar{\theta}$, while the denominator is the model-implied asymptotic variance of maximum likelihood under $\bar{\theta}$, and we take the maximum over $\alpha \neq 0$. Efficiency of the MLE implies that $VIF\left(\bar{\theta}\right) \geq 1$. The square root $\sqrt{VIF\left(\bar{\theta}\right)}$ is the largest factor by which any one-dimensional confidence interval (i.e. confidence interval for a scalar linear transformation $\alpha^\top \theta$, or more generally for a smooth scalar function $\psi(\theta)$) based on the engineered moments must widen relative to the MLE benchmark.

It is important to be clear about what this measure captures: maximum likelihood delivers the smallest model-implied asymptotic variance at a given parameter value, but for given P the probability limit of maximum likelihood generally differs from that of the engineered estimator targeting $\bar{\theta}$. Consequently, the model-implied variance of the realized MLE could lie above or below that of the engineered estimator.

The inflation factor $VIF\left(\bar{\theta}\right)$ has a close connection to measures of statistical distinguishability. In particular, the root chi-squared divergence is equal to the "maximum population t-statistic" for distinguishing P and $P_{\bar{\theta}}$,

$$\sqrt{\chi^2\left(P \| P_{\bar{\theta}}\right)} = \sup_{h:\ \mathrm{Var}_{\bar{\theta}}(h(X))>0} \frac{\left|\mathrm{E}_P\left[h\left(X\right)\right] - \mathrm{E}_{\bar{\theta}}\left[h\left(X\right)\right]\right|}{\sqrt{\mathrm{Var}_{\bar{\theta}}\left(h\left(X\right)\right)}}, \tag{5}$$

where the supremum is over $h \in L^2(P_{\bar{\theta}}) \cap L^1(P)$ with positive variance under $P_{\bar{\theta}}$. This is a version of a known variational representation of the chi-squared divergence (see e.g. Polyanskiy and Wu 2025, Theorem 7.26 and Equation 7.91), which we prove in the Online Appendix for completeness. Intuitively, this measures the largest shift, in standard deviation units, for the mean of any scalar function of the data when the distribution changes from $P_{\bar{\theta}}$ to P. Analogously, for a linear class of functions $\mathcal{H} \subseteq L^2\left(P_{\bar{\theta}}; \mathbb{R}\right)$, let $\sqrt{\chi^2_{\mathcal{H}}\left(P \| P_{\bar{\theta}}\right)}$ denote the root generalized divergence obtained by taking the supremum in (5) over $h \in \mathcal{H}$ only,

so $\chi^2_{\mathcal{H}}(P\|P_{\bar{\theta}}) \leq \chi^2(P\|P_{\bar{\theta}})$. The inflation factor depends only on the ratio of these two divergences with $\mathcal{H}$ the linear span of the score.

Proposition 3 (Discrepancy ratio) *Suppose Assumption 1(iii) holds and* $0 < \chi^2(P\|P_{\bar{\theta}}) < \infty$, *and* $Q_{\bar{\theta}}$ *is positive definite. If* $\mathcal{S}_{\bar{\theta}} := \left\{v^\top s_{\bar{\theta}} : v \in \mathbb{R}^K\right\}$, *then*

$$VIF\left(\bar{\theta}\right) = \left\{1 - \kappa(\bar{\theta})\right\}^{-1} \text{ for } \kappa(\bar{\theta}) := \frac{\chi^2_{\mathcal{S}_{\bar{\theta}}}(P\|P_{\bar{\theta}})}{\chi^2(P\|P_{\bar{\theta}})}.$$

The quantity $\kappa\left(\bar{\theta}\right)$ measures the fraction of the total chi-squared separation between P and $P_{\bar{\theta}}$ that is captured by linear combinations of the score. Proposition 3 thus implies that the variance penalty is very large ($VIF\left(\bar{\theta}\right) \gg 1$, or equivalently $\kappa\left(\bar{\theta}\right) \approx 1$) if and only if some linear combination of the score separates P and $P_{\bar{\theta}}$ nearly as well as does any square-integrable function.

4 Illustration: Moment Choice in a Menu-Cost Model

To illustrate the empirical consequences of moment choice, we follow Cocci and Plagborg-Møller (2025) and estimate Alvarez and Lippi (2014)'s model of menu-cost price setting in multiproduct firms. Under the model, each firm sells N products whose desired log prices follow Brownian motions. Firms must pay a menu cost to reset prices but, upon paying that cost, may reset prices for all their products, and Alvarez and Lippi (2014) show that firms optimally reset prices once the sum of squared deviations between current and desired prices exceeds a threshold $\bar{y}$. We consider estimation based on the cross-sectional distribution of nonzero absolute centered log price changes, $X := |\Delta p - \overline{\Delta p}|$ conditional on $\Delta p \neq 0$, whose model-implied distribution is parameterized by $(N, \bar{y})$. Since the model-implied distribution is well-defined for non-integer N, we follow Cocci and Plagborg-Møller (2025) and allow any real $N > 1$ - see Online Appendix S4 for further details. We fit the model using data on beer sales at a Chicago-area supermarket (Kilts Center for Marketing, 2018), which we clean and normalize following Plagborg-Møller and Cocci (2024).

As a baseline, we consider matching the second and fourth moments of X, which is a just-identified subset of the moments considered in Cocci and Plagborg-Møller (2025). We compute the conservative standard errors of Cocci and Plagborg-Møller (2025) at the resulting estimate, and plot the estimate along with the marginal 95% confidence intervals for the two parameters in black in Figures 2 and 3. We next explore the scope for obtaining

different estimates of θ by matching alternative moments. Specifically, we consider a range of just-identified specifications, matching different aspects of the distribution of log price changes. Figure 2 reports point estimates for θ obtained from all pairs of power moments $\mathrm{E}[X^k]$ with $k \in \{1, \ldots, 8\}$, all pairs of quantiles in $\{0.10, 0.20, \ldots, 0.90\}$ with $|q_1 - q_2| \geq 0.25$ (to exclude nearly redundant pairs), and all power-quantile pairs. Powers alone move the estimates beyond the confidence intervals of Cocci and Plagborg-Møller (2025), and adding quantiles moves them further still.[11] Moreover, different estimates in this set point to substantively different economic interpretations, with the smallest estimate suggesting fewer than two products per firm, and the largest more than 12. Similarly, the smallest and largest estimates of the adjustment threshold $\bar{y}$ differ by roughly a factor of 17.[12]

Despite the substantial sensitivity to estimation moments found in Figure 2, there is a broadly upward-sloping relationship between the estimates for N and $\bar{y}$. This may be intuitive under the model, since to explain e.g. a distribution consisting mostly of small price changes, one could either use firms with few products and a low reset threshold, or firms with many products and a higher threshold. One might hope that this conclusion, at least, would be robust to the choice of moments. Our theoretical results, however, suggest that this cannot be the case.

To illustrate this point, we next consider a researcher who exploits the full flexibility of moment matching to target a grid of values $\bar{\theta} = (N, \bar{y})$ with $N \in \{2, 3, 4, 5\}$ and $\bar{y} \in \{0.06, 0.12, 0.18\}$. For a given target $\bar{\theta}$, our results in Sections 2 and 3 show how a researcher who knows the distribution of X can choose moments which (a) imply estimand $\bar{\theta}$ and (b) minimize the model-implied variance over moments with that estimand. The original model does not satisfy our Assumption 1(iii), however, since at some parameter values it is not DQM in $\bar{y}$, which determines the support of X. To avoid this issue, as in the normal-location example we consider an adversary restricted to a smaller class of moments, in this case functions of $X_c = \min\{X, c\}$ for $c = 0.22$, which (for a compact parameter space with $N > 1$ and $\bar{y} > c^2$) ensures that Assumption 1(i)-(iii) hold.[13]

Since the efficient moments in Proposition 2(b) depend on the true data distribution, we next construct an estimate for P. We use a logspline density estimator to fit the distribution of X_c for $X_c < c$, and then use the empirical frequency to estimate the probability $X_c =$

[11] This reflects no flaw in the Cocci and Plagborg-Møller (2025) standard errors, which are designed to upper bound statistical uncertainty given a set of moments, not variability induced by the choice of moments.

[12] Nineteen power-quantile pairs have two exact solutions, so for completeness Figure 2 plots both.

[13] Absent this restriction, the adversary could obtain even smaller model-implied variances.

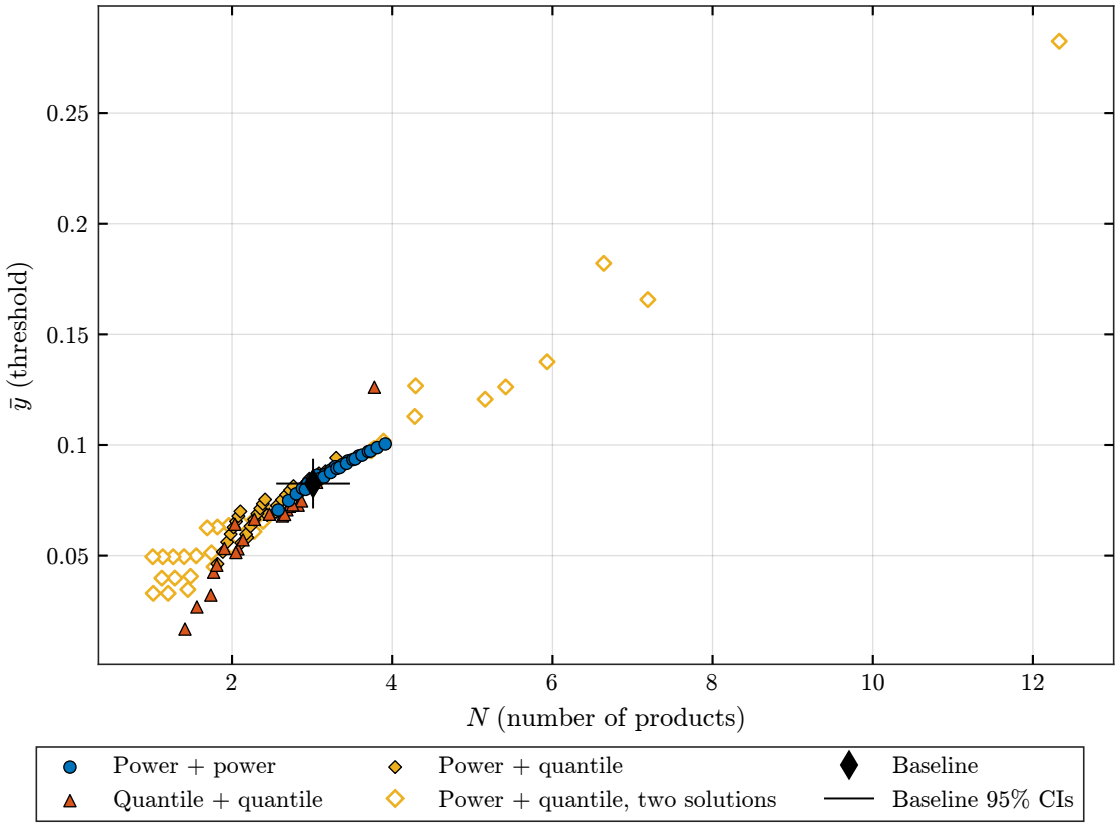


Figure 2: Point estimates for θ, considering powers and quantiles of the distribution of nonzero absolute centered log price changes, with quantile pairs restricted to $|q_1 - q_2| \geq 0.25$. Hollow markers show both solutions for the nineteen pairs with two exact solutions. The horizontal and vertical CIs are conservative confidence intervals based on Cocci and Plagborg-Møller (2025).

c.[14] We then implement our adversarial construction using the estimated distribution $\hat{P}$, constructing the target-specific adversarial moment $g_{\bar{\theta},\hat{P}}$ and computing estimates $\hat{\theta}$ solving

$$m_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_{\bar{\theta},\hat{P}}(X_{c,i}) - \mathrm{E}_\theta[g_{\bar{\theta},\hat{P}}(X_c)] = 0.$$

The resulting estimates are plotted in Figure 3.[15] We include ellipses around each estimate

[14]The logspline estimator models the log conditional density on $[0, c]$ as a cubic B-spline, as in Kooperberg and Stone (1991, 1992), using three evenly spaced interior knots. For $B(x)^\top\gamma$ the log-density under coefficients γ, we choose γ to minimize $-\sum_{i:X_i\leq c} B(X_i)^\top\gamma + n_c \log\int_0^c \exp\{B(u)^\top\gamma\}\,du$ for n_c the number of uncensored observations. The fitted conditional density is multiplied by the empirical mass $\frac{n_c}{n}$ below c, and paired with the empirical mass $\frac{n-n_c}{n}$ at c. Assumption 1(iv)-(v) hold at the estimated distribution for every target (Online Appendix S4).

[15]Since we use the estimated distribution $\hat{P}$, the results in Figure 3 use the data twice, first to construct $\hat{P}$ and then to construct $\hat{\theta}$ and the standard errors. Online Appendix S4 reports similar results based on a

which reflect the model-implied confidence intervals for one-dimensional linear combinations $\alpha^\top\theta$, using model-implied standard errors.[16] To read off the 95% confidence interval for a given parameter, e.g. N, for a given choice of moments, simply take the projection of the corresponding ellipse on the corresponding axis.[17] As these ellipses highlight, the deviation of the engineered estimates from their targets is on the same order as sampling variability. For comparison, we also plot the maximum likelihood estimate along with its model-implied confidence ellipse.[18]

As Figure 3 highlights, we can obtain a wide range of different estimates through the targeted choice of which moments to match. Moreover, in some cases the standard errors of these targeted estimates are comparable to (or even smaller than) the standard errors for maximum likelihood in these data, reflecting that while the MLE model-implied standard errors are the smallest possible at a given parameter value, these standard errors themselves vary with the parameter, so a suitable choice of target can allow moment-matching standard errors smaller than those at the realized MLE. If we instead compare the model-implied standard errors for the engineered estimates $\hat{\theta}_g$ with the MLE standard errors also evaluated at $\hat{\theta}_g$, we find that the engineered standard errors for the number of products N range from 0.2% to 39.1% larger, with a median of 10.6%, while those for $\bar{y}$ range from 3.7% to 215.3% larger, with a median of 18.8%. A full tabulation of both comparisons appears in Table S2 in Online Appendix S4, which also provides additional details on the application, along with plots of the adversarial moments.

Overall, these results highlight that even in a simple model, the choice of moments can move estimates across a wide range with standard errors that, for many targets, are close to those censored-data maximum likelihood would report at the same estimates.

cross-fit procedure designed to mitigate this re-use.

[16]Note that these standard errors take the choice of moments as given and so ignore the randomness induced by estimation of $\hat{P}$. In addition, since we consider a just-identifying set of moments, we are not theoretically enforcing uniqueness as in Proposition 1. Nevertheless, a search over a grid of $(N, \bar{y})$ values with $N \in [1.01, 40]$ and $\bar{y} \in (c^2, 2]$ finds exactly one solution to the moment equations for each target, suggesting that non-uniqueness is not an issue over the range of parameter values we study.

[17]Formally, we plot Wald confidence ellipses constructed from the sandwich variance $\text{AVar}(\hat{\theta}_{g_{\hat{\theta},\hat{P}}})$ and χ^2_1 critical values, where the use of χ^2_1 critical values ensures the projections are non-conservative but does not imply joint coverage of both parameters.

[18]We consider maximum likelihood based on the censored price changes X_c, again to avoid irregularity due to the dependence of the support on $\bar{y}$.

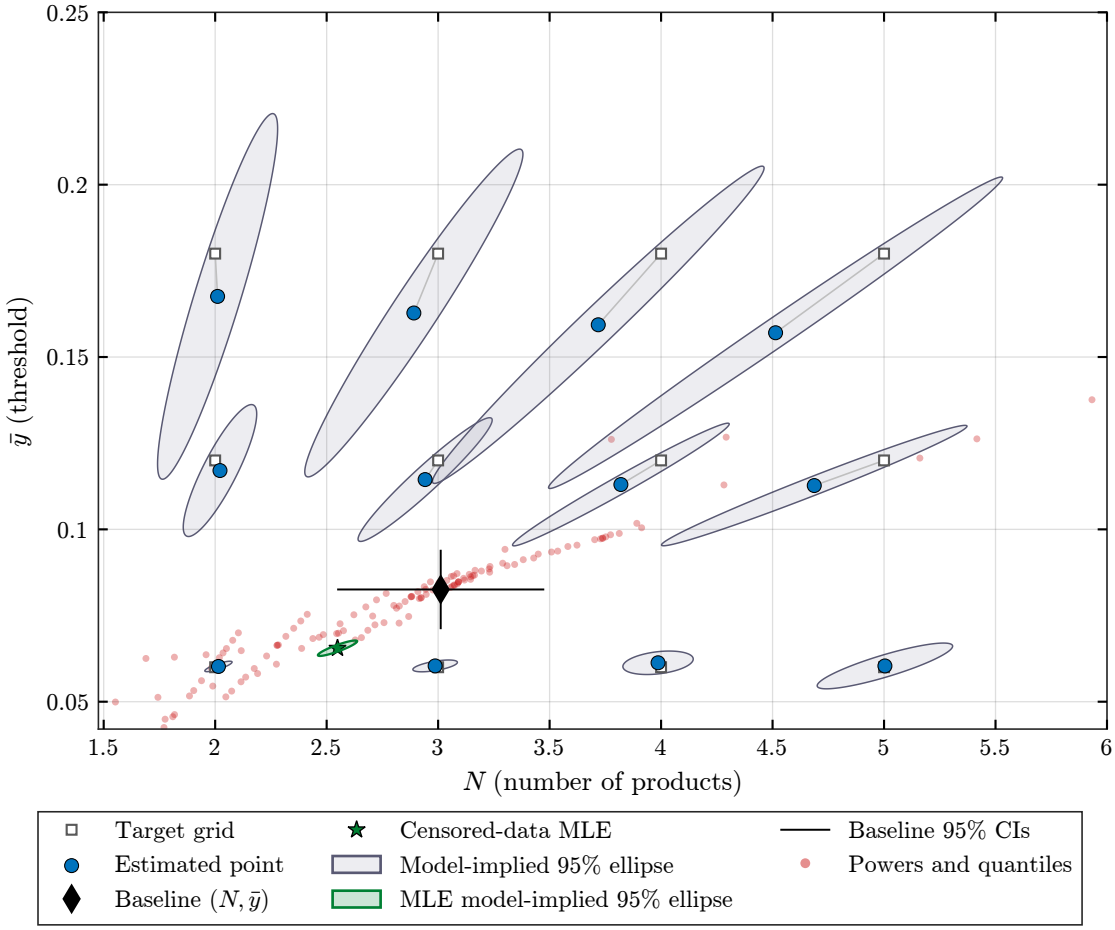


Figure 3: Adversarial estimates from Proposition 2 with model-implied uncertainty. Gray lines connect each target to its estimate. Shaded regions are model-implied Wald confidence ellipses around each $\hat{\theta}$, using χ^2_1 critical values so that their projections onto each axis are nominal 95% model-implied confidence intervals, treating the moments as fixed. The red scatter shows, for reference, the varying-moment estimates from Figure 2, while the black diamond and bars are the baseline $(N, \bar{y})$ estimate and confidence intervals. Finally, the green star and ellipse show the maximum likelihood estimate and model-implied confidence ellipse, again using a χ^2_1 critical value.

5 What to Do with This?

Given concerns about model misspecification, it is natural for researchers to want discretion over which moments to use for estimation. As we highlight above, however, this flexibility also introduces substantial researcher degrees of freedom which could be abused by an adversarial researcher. Even with the best of intentions, these issues may also lead researchers working with the same data and model to reach very different conclusions. Moreover, as we show in Online Appendix S5, these issues are not unique to parametric models estimated using

moment matching, since a similar "anything-goes" result applies to estimands in misspecified conditional moment equality models. In our view, these results suggest some concrete steps to improve current practice.

Researchers should report (inter alia) estimates which are efficient under their model Once the model is wrong, the estimand depends on the estimator, so efficiency comparisons lose their usual meaning. Nevertheless, since they are motivated solely by the model, efficient estimators provide a clear point of comparison. For instance, if two researchers obtain different maximum likelihood estimates for a given model, we know that this must reflect differences in their data or implementation. Even when maximum likelihood is intractable, other estimators motivated by efficiency under the model (e.g. maximum simulated likelihood, or the adversarial estimator of Kaji et al. 2023) seem a good starting point.

One might object that the estimands from model-efficient estimators (e.g. the maximum likelihood estimand, which minimizes the Kullback-Leibler divergence between the data and model-implied distributions, White 1982) are often hard to interpret when the model is wrong. We agree, and for this reason do not view the reporting of model-efficient estimators as a panacea, but only as a useful first step.

Researchers should explicitly motivate their choice of alternative estimator In settings where researchers choose to focus on an estimator which is inefficient under the model while reporting the MLE or another model-efficient estimate in a secondary role, we recommend that they explain what leads them to this choice. If they are motivated by misspecification concerns, they should articulate what these concerns are and why they point to their preferred estimator above all others. This would amount to a formalization of some current practices: researchers often informally discuss why at least some of their choices are well-suited to their substantive objectives, but the motivation is rarely spelled out precisely. Moreover, while formal specification tests are only sporadically reported in some literatures (Andrews et al., 2025), we view such tests as a natural antecedent to many such discussions.

While formally connecting one's misspecification concerns to the choice of estimator may sound daunting, there is a large and growing econometric literature which speaks to this precise issue. Broadly, the papers in this literature start from a precise description of misspecification concerns, and then derive the corresponding "best" estimator, confidence interval, or other procedure. Versions of this approach have been developed for particular misspecification settings by, for example, Armstrong and Kolesár (2021), Bonhomme and Weidner

(2022), Christensen and Connault (2023), Adusumilli (2026), and Andrews et al. (2026). A complementary route is to relax assumptions until one actually believes the model, which may entail partial identification (Manski, 1989; Molinari, 2020). If widely adopted, these approaches would shift the focus from the choice of estimator to the forms of misspecification that researchers wish to guard against, which seems to us a clearer basis for discussion and disagreement.

Together, we think that these steps could help discipline the choice of what to report when one is concerned about misspecification. Another possibility, applicable in settings with a meaningful pre-data stage (e.g. lab or field experiments), could be to pre-commit to the choice of estimator. Since one would presumably still want to ground the choice of estimator either in efficiency considerations or in context-specific misspecification concerns, however, we view this as complementary to the approaches above.

A further possibility, to an extent implicit in current practice, could be to require that researchers use only "reasonable" moments as defined e.g. by social convention. To us, however, the remaining researcher degrees of freedom, together with the ambiguity about what constitutes a "reasonable" moment and why, all seem like limitations of this approach. Overall, we think economics would be well-served by adopting clearer guidelines for what estimators researchers should use when they are worried their model is wrong.

References

Adusumilli, K. (2026): "You've Got to be Efficient: Ambiguity, Misspecification and Variational Preferences," *arXiv preprint arXiv:2604.05327*.

Alvarez, F., and F. Lippi (2014): "Price setting with menu cost for multiproduct firms," *Econometrica*, 82, 89–135.

Andrews, I., J. Chen, and O. Tecchio (2025): "The purpose of an estimator is what it does: Misspecification, estimands, and over-identification," *arXiv preprint arXiv:2508.13076*.

Andrews, I., R. Li, and Y. Shang (2026): "Misspecification-Averse Estimation."

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

Bonhomme, S., and M. Weidner (2022): "Minimizing sensitivity to model misspecification," *Quantitative Economics*, 13, 907–954.

Bordalo, P., N. Gennaioli, Y. Ma, and A. Shleifer (2020): "Overreaction in macroeconomic expectations," *American Economic Review*, 110, 2748–2782.

Bourreau, M., Y. Sun, and F. Verboven (2021): "Market entry, fighting brands, and tacit collusion: Evidence from the French mobile telecommunications market," *American Economic Review*, 111, 3459–3499.

Christensen, T., and B. Connault (2023): "Counterfactual sensitivity and robustness," *Econometrica*, 91, 263–298.

Cocci, M. D., and M. Plagborg-Møller (2025): "Standard Errors for Calibrated Parameters," *Review of Economic Studies*, 92, 2952–2978.

Cooley, T. F. (1997): "Calibrated models," *Oxford Review of Economic Policy*, 13, 55–69.

Gallant, A. R., and G. Tauchen (1996): "Which moments to match?" *Econometric Theory*, 12, 657–681.

Hall, A. R., and A. Inoue (2003): "The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models," *Journal of Econometrics*, 114, 361–394.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P., and J. J. Heckman (1996): "The empirical foundations of calibration," *Journal of Economic Perspectives*, 10, 87–104.

Kaji, T., E. Manresa, and G. Pouliot (2023): "An adversarial approach to structural estimation," *Econometrica*, 91, 2041–2063.

Kilts Center for Marketing (2018): "Dominick's Dataset," University of Chicago Booth School of Business, https://www.chicagobooth.edu/research/kilts/research-data/dominicks.

Kooperberg, C., and C. J. Stone (1991): "A study of logspline density estimation," *Computational Statistics & Data Analysis*, 12, 327–347.

——— (1992): "Logspline density estimation for censored data," *Journal of Computational and Graphical Statistics*, 1, 301–328.

Lagakos, D., A. M. Mobarak, and M. E. Waugh (2023): "The welfare effects of encouraging rural–urban migration," *Econometrica*, 91, 803–837.

Leamer, E. E. (1974): "False Models and Post-Data Model Construction," *Journal of the American Statistical Association*, 69, 122–131.

——— (1978): "Regression Selection Strategies and Revealed Priors," *Journal of the American Statistical Association*, 73, 580–587.

Lovell, M. C. (1983): "Data Mining," *The Review of Economics and Statistics*, 65, 1–12.

Manski, C. F. (1989): "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343–360.

Molinari, F. (2020): "Microeconometrics with Partial Identification," in *Handbook of Econometrics* ed. by Durlauf, S. N., Hansen, L. P., Heckman, J. J., and Matzkin, R. L. Volume 7A: Elsevier, 355–486.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Plagborg-Møller, M., and M. D. Cocci (2024): "Replication package for: Standard Errors for Calibrated Parameters. Version v1," Zenodo, https://doi.org/10.5281/zenodo.12784903.

Polyanskiy, Y., and Y. Wu (2025): *Information Theory: From Coding to Learning*, Cambridge: Cambridge University Press.

Prescott, E. C. (1986): "Theory ahead of business-cycle measurement," in *Carnegie-Rochester conference series on public policy* Volume 25, 11–44, Elsevier.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011): "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366.

Simonovska, I., and M. E. Waugh (2025): "Trade models, trade elasticities, and the gains from trade,"Technical report, National Bureau of Economic Research.

Su, C.-L., and K. L. Judd (2012): "Constrained optimization approaches to estimation of structural models," *Econometrica*, 80, 2213–2230.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge: Cambridge University Press.

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–25.

Wooldridge, J. M. (2001): "Applications of generalized method of moments estimation," *Journal of Economic Perspectives*, 15, 87–100.

Online Appendix to

"The Incredible Flexibility of Moment Matching"

Isaiah Andrews and Bas Sanders

This online appendix contains proofs for results in the main text (Appendix S1), results for data with finite support (Appendix S2), further details on the normal-location example (Appendix S3) and the menu-cost application (Appendix S4), and results for misspecified conditional moment equality models (Appendix S5).

S1 Proofs

Let $\Delta := P - P_{\bar{\theta}}$, and for $u \in L^1(P_{\bar{\theta}})$ let $uP_{\bar{\theta}}$ denote the signed measure which assigns mass $\int_B u \, dP_{\bar{\theta}}$ to sets B. Let

$$\mathcal{B}_\Delta := \left\{ f : \mathcal{X} \to \mathbb{R} \text{ bounded and measurable}, \ \int f \, d\Delta = 0 \right\}$$

denote the bounded scalar matching moments, and $\mathcal{C}_1(\bar{\theta})$ the scalar matching moments.

S1.1 Auxiliary Results

Lemma S1 (DQM implies differentiation of moments) *Under Assumption 1(iii),* $\int |p_{\bar{\theta}+h} - p_{\bar{\theta}} - (h^\top s_{\bar{\theta}}) p_{\bar{\theta}}| \, d\mu = o(\|h\|)$. *Hence for every bounded measurable* $f : \mathcal{X} \to \mathbb{R}^d$,

$$\mathrm{E}_{\bar{\theta}+h}[f(X)] - \mathrm{E}_{\bar{\theta}}[f(X)] = \mathrm{E}_{\bar{\theta}}[f(X) s_{\bar{\theta}}(X)^\top] h + o(\|h\|).$$

The same expansion holds for every measurable f *such that* $\|f(X)\|^2$ *is uniformly integrable under* $\{P_\theta : \theta \in \Theta, \|\theta - \bar{\theta}\| \leq \delta\}$ *for some* $\delta > 0$.

Proof. Let $a_h := \sqrt{p_{\bar{\theta}+h}} - \sqrt{p_{\bar{\theta}}}$ and $e_h := a_h - \frac{1}{2} h^\top s_{\bar{\theta}} \sqrt{p_{\bar{\theta}}}$. DQM gives $\|e_h\|_{L^2(\mu)} = o(\|h\|)$, and since $s_{\bar{\theta}} \in L^2(P_{\bar{\theta}})$, $\|a_h\|_{L^2(\mu)} = O(\|h\|)$. Observe, however, that $p_{\bar{\theta}+h} - p_{\bar{\theta}} - (h^\top s_{\bar{\theta}}) p_{\bar{\theta}} = 2\sqrt{p_{\bar{\theta}}} e_h + a_h^2$ which together with Cauchy-Schwarz gives the first claim, while the second follows from boundedness of f.

For the final claim, it suffices to consider scalar f. Uniform integrability implies $f \in L^2(P_{\bar\theta})$, so Cauchy-Schwarz gives $|\int f\sqrt{p_{\bar\theta}} e_h \, d\mu| \le \|f\|_{L^2(P_{\bar\theta})}\|e_h\|_{L^2(\mu)} = o(\|h\|)$. For $M > 0$ and $\|h\| \le \delta$, since $a_h^2 \le 2p_{\bar\theta+h} + 2p_{\bar\theta}$, Cauchy-Schwarz further gives

$$\left|\int f a_h^2 \, d\mu\right| \le M\|a_h\|_{L^2(\mu)}^2 + 2\|a_h\|_{L^2(\mu)} \sup_{\theta\in\Theta:\|\theta-\bar\theta\|\le\delta} \mathrm{E}_\theta\left[f(X)^2\,\mathbb{I}\{|f(X)| > M\}\right]^{1/2}.$$

Dividing by $\|h\|$ and taking $h \to 0$ and then $M \to \infty$, the right-hand side vanishes by uniform integrability. Hence $\int f\{p_{\bar\theta+h} - p_{\bar\theta} - (h^\top s_{\bar\theta})p_{\bar\theta}\}\,d\mu = o(\|h\|)$, as claimed. □

Lemma S2 (Annihilator of the matching constraint) *Suppose $P \ne P_{\bar\theta}$. A finite signed measure ν satisfies $\int f\,d\nu = 0$ for all $f \in \mathcal{B}_\Delta$ if and only if $\nu \in \mathrm{span}\{\Delta\}$.*

Proof. Since $\Delta \ne 0$, there is a bounded measurable b with $\int b\,d\Delta = 1$. For every bounded measurable f, $f - (\int f\,d\Delta)b \in \mathcal{B}_\Delta$. If ν is orthogonal to $\mathcal{B}_\Delta$ (in the sense that every function in $\mathcal{B}_\Delta$ integrates to zero against ν) then $\int f\,d\nu = (\int b\,d\nu)\int f\,d\Delta$ for every bounded measurable f, which implies that $\nu = (\int b\,d\nu)\Delta$. The converse is immediate. □

Lemma S3 (Closure of the matching moments) *Suppose $P \ne P_{\bar\theta}$, and let $\mathcal{H}$ denote the $L^2(P_{\bar\theta})$ closure of $\mathcal{B}_\Delta$. Then $\mathcal{H}^\perp = \{u \in L^2(P_{\bar\theta}) : uP_{\bar\theta} \in \mathrm{span}\{\Delta\}\}$, $\mathcal{H}$ is the closure of $\mathcal{C}_1(\bar\theta)$, and*

$$\mathcal{H} = \begin{cases} \mathcal{C}_1(\bar\theta) = \{f \in L^2(P_{\bar\theta}) : \mathrm{E}_{\bar\theta}[f(X)r(X)] = 0\} & \text{if } \chi^2(P\|P_{\bar\theta}) < \infty, \\ L^2(P_{\bar\theta}) & \text{if } \chi^2(P\|P_{\bar\theta}) = \infty. \end{cases}$$

Proof. Since $\mathcal{H}$ is the closure of $\mathcal{B}_\Delta$, a function $u \in L^2(P_{\bar\theta})$ belongs to $\mathcal{H}^\perp$ if and only if $\mathrm{E}_{\bar\theta}[uf] = \int f\,d(uP_{\bar\theta}) = 0$ for every $f \in \mathcal{B}_\Delta$, which by Lemma S2 holds if and only if $uP_{\bar\theta} \in \mathrm{span}\{\Delta\}$. If $uP_{\bar\theta} = c\Delta$ with $c \ne 0$, however, then $P = (1+u/c)P_{\bar\theta}$, so $\chi^2(P\|P_{\bar\theta}) = \mathrm{E}_{\bar\theta}[u^2]/c^2 < \infty$. Hence infinite divergence implies $\mathcal{H}^\perp = \{0\}$, while with finite divergence $\Delta = rP_{\bar\theta}$ and $\mathcal{H}^\perp = \mathrm{span}\{r\}$. Taking orthogonal complements gives the displayed characterization, except for the equality with $\mathcal{C}_1(\bar\theta)$. With finite divergence, however, every $f \in L^2(P_{\bar\theta})$ is P-integrable, since $\mathrm{E}_P[|f|] = \mathrm{E}_{\bar\theta}[|f|(1+r)] < \infty$, and $\mathrm{E}_P[f] - \mathrm{E}_{\bar\theta}[f] = \mathrm{E}_{\bar\theta}[fr]$, so $\mathcal{C}_1(\bar\theta) = \{f \in L^2(P_{\bar\theta}) : \mathrm{E}_{\bar\theta}[fr] = 0\}$. In both cases $\mathcal{B}_\Delta \subseteq \mathcal{C}_1(\bar\theta) \subseteq \mathcal{H}$, so $\mathcal{H}$ is the closure of $\mathcal{C}_1(\bar\theta)$. □

Lemma S4 (Variational characterization) *If there exists an event B with $0 < P_{\bar\theta}(B) < 1$, then (5) holds, with the supremum over $h \in L^2(P_{\bar\theta};\mathbb{R}) \cap L^1(P;\mathbb{R})$ with $\mathrm{Var}_{\bar\theta}(h(X)) > 0$.*

Proof. If $\chi^2(P\|P_{\bar\theta}) < \infty$, then for any such h, $\mathrm{E}_P[h] - \mathrm{E}_{\bar\theta}[h] = \mathrm{E}_{\bar\theta}[hr] = \mathrm{Cov}_{\bar\theta}(h, r)$, so the Cauchy-Schwarz inequality gives $|\mathrm{E}_P[h] - \mathrm{E}_{\bar\theta}[h]| \le \sqrt{\mathrm{Var}_{\bar\theta}(h)\chi^2(P\|P_{\bar\theta})}$, with equality for

$h = r$ when $P \neq P_{\bar{\theta}}$. For $\chi^2(P\|P_{\bar{\theta}}) = \infty$, suppose towards contradiction that the supremum equals $c < \infty$. Take any bounded h, and a bounded f with $\mathrm{Var}_{\bar{\theta}}(f) > 0$. Applying the above bound to $h + tf$ and letting $t \to 0$ when $\mathrm{Var}_{\bar{\theta}}(h) = 0$, we obtain $|\int h\, d\Delta| \leq c\|h\|_{L^2(P_{\bar{\theta}})}$. The map $h \mapsto \int h\, d\Delta$ thus extends to a bounded linear functional on $L^2(P_{\bar{\theta}})$, so by the Riesz representation theorem there exists $u \in L^2(P_{\bar{\theta}})$ with $\int h\, d\Delta = \mathrm{E}_{\bar{\theta}}[hu]$ for all bounded h. Considering indicators yields $P = (1+u)P_{\bar{\theta}}$, so $\chi^2(P\|P_{\bar{\theta}}) = \mathrm{E}_{\bar{\theta}}[u^2] < \infty$, a contradiction. □

Lemma S5 (Equivalent forms of Assumption 1(iv)-(v)) *Suppose $P \neq P_{\bar{\theta}}$ and $I_{\bar{\theta}}$ is positive definite. Then:*

(a) Condition (iv) holds if and only if $(v^\top s_{\bar{\theta}})P_{\bar{\theta}} \notin \mathrm{span}\{\Delta\}$ for every nonzero $v \in \mathbb{R}^K$. It holds whenever $\chi^2(P\|P_{\bar{\theta}}) = \infty$, and when $\chi^2(P\|P_{\bar{\theta}}) < \infty$ it holds if and only if r is not a linear combination of the components of $s_{\bar{\theta}}$.

(b) Condition (v) holds if and only if $P_\theta - P_{\bar{\theta}} \notin \mathrm{span}\{\Delta\}$ for every $\theta \neq \bar{\theta}$. If $\chi^2(P\|P_{\bar{\theta}}) = \infty$ and $\chi^2(P_\theta\|P_{\bar{\theta}}) < \infty$ for all θ, it holds if and only if $P_\theta \neq P_{\bar{\theta}}$ for all $\theta \neq \bar{\theta}$.

Proof. (a) If $(v^\top s_{\bar{\theta}})P_{\bar{\theta}} = c\Delta$ for $v \neq 0$ then $c \neq 0$ since $v^\top s_{\bar{\theta}} \neq 0$, so $P = (1 + (v/c)^\top s_{\bar{\theta}})P_{\bar{\theta}}$ and (iv) fails. Conversely if $P = (1 + w^\top s_{\bar{\theta}})P_{\bar{\theta}}$ then $w \neq 0$ since $P \neq P_{\bar{\theta}}$, and $(w^\top s_{\bar{\theta}})P_{\bar{\theta}} = \Delta$. In that case P is absolutely continuous with respect to $P_{\bar{\theta}}$ with $dP/dP_{\bar{\theta}} - 1 = w^\top s_{\bar{\theta}} \in L^2(P_{\bar{\theta}})$, so $\chi^2(P\|P_{\bar{\theta}}) = w^\top I_{\bar{\theta}} w < \infty$ and $r = w^\top s_{\bar{\theta}}$, while with finite divergence $r = w^\top s_{\bar{\theta}}$ implies $P = (1 + w^\top s_{\bar{\theta}})P_{\bar{\theta}}$. Hence (iv) holds whenever the divergence is infinite, and when it is finite (iv) fails if and only if r is a linear combination of the components of $s_{\bar{\theta}}$. (b) $P_\theta - P_{\bar{\theta}} = a\Delta$ is equivalent to $P_\theta = (1-a)P_{\bar{\theta}} + aP$. If $0 < a < 1$, then P_θ is a nontrivial mixture of $P_{\bar{\theta}}$ and P; if $a > 1$, then $P = \frac{1}{a}P_\theta + \frac{a-1}{a}P_{\bar{\theta}}$; if $a < 0$, then $P_{\bar{\theta}} = \frac{1}{1-a}P_\theta + \frac{-a}{1-a}P$; while $a = 0$ and $a = 1$ give $P_\theta = P_{\bar{\theta}}$ and $P_\theta = P$. Conversely, each mixture relation can be rearranged as $P_\theta = (1-a)P_{\bar{\theta}} + aP$ for some a. For the last claim, if $P_\theta - P_{\bar{\theta}} = a\Delta$ with $a \neq 0$, then $dP/dP_{\bar{\theta}} - 1 = a^{-1}(dP_\theta/dP_{\bar{\theta}} - 1)$ would be square-integrable, contradicting infinite divergence. □

S1.2 Proof of Proposition 1

By Lemma S5, Assumption 1 conditions (iv) and (v) are equivalent to $(v^\top s_{\bar{\theta}})P_{\bar{\theta}} \notin \mathrm{span}\{\Delta\}$ for $v \neq 0$ and $P_\theta - P_{\bar{\theta}} \notin \mathrm{span}\{\Delta\}$ for $\theta \neq \bar{\theta}$. Consider the linear map $T : \mathcal{B}_\Delta \to \mathbb{R}^K$ defined by $Tf = \mathrm{E}_{\bar{\theta}}[s_{\bar{\theta}} f]$. This map is onto, since otherwise there exists $v \neq 0$ such that $\mathrm{E}_{\bar{\theta}}[(v^\top s_{\bar{\theta}})f] = 0$ for all $f \in \mathcal{B}_\Delta$, in which case Lemma S2 gives $(v^\top s_{\bar{\theta}})P_{\bar{\theta}} \in \mathrm{span}\{\Delta\}$, contradicting (iv). Choose $f_j \in \mathcal{B}_\Delta$ with $Tf_j = e_j$ for $j = 1, \ldots, K$, and let g_0 collect

$f_j - \mathrm{E}_{\bar\theta}[f_j]$. Since constants belong to $\mathcal{B}_\Delta$ and $\mathrm{E}_{\bar\theta}[s_{\bar\theta}] = 0$, g_0 is bounded, matching, and satisfies $\mathrm{E}_{\bar\theta}[g_0 s_{\bar\theta}^\top] = I_K$. Lemma S1 then gives $\mathrm{E}_{\bar\theta+h}[g_0] - \mathrm{E}_{\bar\theta}[g_0] = h + o(\|h\|)$, so there exists $\varepsilon > 0$ such that no $\theta \in \Theta$ with $0 < \|\theta - \bar\theta\| < \varepsilon$ solves the moment conditions based on g_0.

$F := \{\theta \in \Theta : \|\theta - \bar\theta\| \geq \varepsilon\}$ is compact. For each $\eta \in F$, condition (v) and Lemma S2 provide $v_\eta \in \mathcal{B}_\Delta$ with $\mathrm{E}_\eta[v_\eta] - \mathrm{E}_{\bar\theta}[v_\eta] \neq 0$. Since $\theta \mapsto P_\theta$ is continuous in total variation and v_η is bounded, $\theta \mapsto \mathrm{E}_\theta[v_\eta]$ is continuous, so this separation persists on an open neighborhood of η. A finite subcover of F yields $v_{\eta_1}, \ldots, v_{\eta_J}$, and appending these to g_0 gives a bounded $g \in \mathcal{C}(\bar\theta)$ with $\Theta_g = \{\bar\theta\}$, whose Jacobian at $\bar\theta$ has rank K since its first K rows form I_K. □

S1.3 Proof of Proposition 2

As in Lemma S3, let $\mathcal{H}$ denote the $L^2(P_{\bar\theta})$ closure of $\mathcal{C}_1(\bar\theta)$. The characterization in Lemma S3 implies that $s_{\bar\theta,\perp}$ is the (component-wise) projection of $s_{\bar\theta}$ onto $\mathcal{H}$. Consequently $s_{\bar\theta} - s_{\bar\theta,\perp}$ is orthogonal to $\mathcal{H}$, so $\mathrm{E}_{\bar\theta}\left[s_{\bar\theta,\perp} s_{\bar\theta}^\top\right] = \mathrm{E}_{\bar\theta}\left[s_{\bar\theta,\perp} s_{\bar\theta,\perp}^\top\right] = Q_{\bar\theta}$, while since $\mathrm{E}_{\bar\theta}[s_{\bar\theta}] = 0$ and $\mathcal{H}$ contains constants, $\mathrm{E}_{\bar\theta}[s_{\bar\theta,\perp}] = 0$. If $v^\top Q_{\bar\theta} v = 0$ then $v^\top s_{\bar\theta,\perp} = 0$, and thus $v^\top s_{\bar\theta} \in \mathcal{H}^\perp$. In this case Lemma S3 implies that $(v^\top s_{\bar\theta}) P_{\bar\theta} \in \mathrm{span}\{\Delta\}$, which by Lemma S5 and condition (iv) implies $v = 0$. Hence $Q_{\bar\theta}$ is positive definite.

Lemma S6 *Suppose Assumption 1(iii)-(iv) hold. For just-identifying $g \in \mathcal{C}(\bar\theta)$ with G full-rank,* $\mathrm{AVar}(\hat\theta_g) = Q_{\bar\theta}^{-1} + \mathrm{E}_{\bar\theta}[u(X)u(X)^\top]$ *for* $u := G^{-1}(g - \mathrm{E}_{\bar\theta}[g(X)]) - Q_{\bar\theta}^{-1} s_{\bar\theta,\perp}$. *In particular,* $\mathrm{AVar}(\hat\theta_g) = Q_{\bar\theta}^{-1}$ *if and only if $g = A s_{\bar\theta,\perp} + c$ $P_{\bar\theta}$-almost surely for some invertible A and vector c.*

Proof. Let $h := G^{-1}(g - \mathrm{E}_{\bar\theta}[g])$. Each component of h belongs to $\mathcal{C}_1(\bar\theta) \subseteq \mathcal{H}$, while $\mathrm{E}_{\bar\theta}[h s_{\bar\theta}^\top] = I_K$ since $G = \mathrm{E}_{\bar\theta}[g s_{\bar\theta}^\top]$ as in (3). Since $s_{\bar\theta} - s_{\bar\theta,\perp}$ is orthogonal to $\mathcal{H}$, it follows that $\mathrm{E}_{\bar\theta}[h s_{\bar\theta,\perp}^\top] = I_K$ and thus that $\mathrm{E}_{\bar\theta}[u s_{\bar\theta,\perp}^\top] = I_K - Q_{\bar\theta}^{-1} Q_{\bar\theta} = 0$. Consequently, $\mathrm{AVar}(\hat\theta_g) = \mathrm{E}_{\bar\theta}[hh^\top] = Q_{\bar\theta}^{-1} + \mathrm{E}_{\bar\theta}[uu^\top]$, while the last claim follows from the fact that $u = 0$ if and only if $g = G Q_{\bar\theta}^{-1} s_{\bar\theta,\perp} + \mathrm{E}_{\bar\theta}[g]$. □

For part (a), consider $g \in \mathcal{C}(\bar\theta)$ and positive definite W. If $G^\top W G$ is singular the lower bound holds by convention, so suppose $G^\top W G$ is nonsingular. We then have that $\tilde g := (G^\top W G)^{-1} G^\top W (g - \mathrm{E}_{\bar\theta}[g])$ is a just-identifying element of $\mathcal{C}(\bar\theta)$ with $\mathrm{E}_{\bar\theta}[\tilde g s_{\bar\theta}^\top] = I_K$ and $\mathrm{AVar}(\hat\theta_{\tilde g}) = \mathrm{AVar}(\hat\theta_{g,W})$, so the lower bound follows from Lemma S6. For the upper approximation, by Lemma S3 there exists a sequence $f_l \in \mathcal{B}_\Delta^K$ with $f_l \to s_{\bar\theta,\perp}$ in $L^2(P_{\bar\theta})$. The moments $g_l := f_l - \mathrm{E}_{\bar\theta}[f_l]$ are bounded members of $\mathcal{C}(\bar\theta)$, and since $\mathrm{E}_{\bar\theta}[s_{\bar\theta,\perp}] = 0$, $g_l \to s_{\bar\theta,\perp}$ in $L^2(P_{\bar\theta})$. Cauchy-Schwarz then implies that $G_l = \mathrm{E}_{\bar\theta}[g_l s_{\bar\theta}^\top] \to Q_{\bar\theta}$ and $\mathrm{Var}_{\bar\theta}(g_l) \to Q_{\bar\theta}$,

so $\mathrm{AVar}(\hat{\theta}_{g_l}) \to Q_{\bar{\theta}}^{-1}$. Since $Q_{\bar{\theta}}^{-1}$ is positive definite, however, it follows that $\mathrm{AVar}(\hat{\theta}_{g_l}) \preceq (1+\varepsilon)Q_{\bar{\theta}}^{-1}$ for l sufficiently large.

For part (b), if $\chi^2(P\|P_{\bar{\theta}}) < \infty$ then Lemma S3 shows that $\mathcal{H} = \{f : \mathrm{E}_{\bar{\theta}}[fr] = 0\}$, so $s_{\bar{\theta},\perp} = s_{\bar{\theta}} - \mathrm{E}_{\bar{\theta}}[s_{\bar{\theta}}r]r/\mathrm{E}_{\bar{\theta}}[r^2]$, where $\mathrm{E}_{\bar{\theta}}[s_{\bar{\theta}}r] = \mathrm{E}_P[s_{\bar{\theta}}] - \mathrm{E}_{\bar{\theta}}[s_{\bar{\theta}}] = \mathrm{E}_P[s_{\bar{\theta}}]$. The expression for $Q_{\bar{\theta}}$ follows by direct calculation, while since $\mathcal{H} = \mathcal{C}_1(\bar{\theta})$ by Lemma S3, $s_{\bar{\theta},\perp} \in \mathcal{C}(\bar{\theta})$, while the rest of the result is immediate from Lemma S6. If instead $\chi^2(P\|P_{\bar{\theta}}) = \infty$ then $\mathcal{H} = L^2(P_{\bar{\theta}})$ so $s_{\bar{\theta},\perp} = s_{\bar{\theta}}$ and $Q_{\bar{\theta}} = I_{\bar{\theta}}$, though the bound need not be attained.

For part (c), fix $\varepsilon > 0$ and take a bounded, just-identifying g_l from the sequence constructed in the proof of part (a), selecting l sufficiently large that $\mathrm{AVar}(\hat{\theta}_{g_l}) \preceq (1+\varepsilon/2)Q_{\bar{\theta}}^{-1}$ and G_l is nonsingular. Let g° denote the bounded moments from Proposition 1, and $\check{g} := (g_l^\top, g^{\circ\top})^\top$, where $\Theta_{\check{g}} = \{\bar{\theta}\}$ since $\Theta_{g^\circ} = \{\bar{\theta}\}$ and $g_l \in \mathcal{C}(\bar{\theta})$. For W_0 positive definite and $\eta > 0$ define $W_\eta := \mathrm{diag}(W_0, \eta I)$. As $\eta \downarrow 0$, the asymptotic variance (3) based on $(\check{g}, W_\eta)$ converges to $G_l^{-1}\mathrm{Var}_{\bar{\theta}}(g_l)G_l^{-\top} = \mathrm{AVar}(\hat{\theta}_{g_l})$. Choosing η sufficiently small proves the result. □

S1.4 Proof of Proposition 3

Condition (iv) holds, since otherwise, by Lemma S5(a), $r = \lambda^\top s_{\bar{\theta}}$ for some λ, with $\lambda \neq 0$ since $\chi^2(P\|P_{\bar{\theta}}) > 0$. Since $rP_{\bar{\theta}} = \Delta$, Lemma S3 would then give $\lambda^\top s_{\bar{\theta}} \in \mathcal{H}^\perp$, so $\lambda^\top s_{\bar{\theta},\perp} = 0$ and $\lambda^\top Q_{\bar{\theta}}\lambda = 0$, contradicting positive definiteness of $Q_{\bar{\theta}}$. From Proposition 2(b), we know

$$Q_{\bar{\theta}}^{-1} = \left(I_{\bar{\theta}} - \frac{\mathrm{E}_P[s_{\bar{\theta}}]\,\mathrm{E}_P[s_{\bar{\theta}}]^\top}{\chi^2\left(P\|P_{\bar{\theta}}\right)}\right)^{-1}.$$

Writing $y = I_{\bar{\theta}}^{-1/2}\alpha$ and $w = I_{\bar{\theta}}^{-1/2}\,\mathrm{E}_P[s_{\bar{\theta}}]/\sqrt{\chi^2(P\|P_{\bar{\theta}})}$, we have $Q_{\bar{\theta}} = I_{\bar{\theta}}^{1/2}(I_K - ww^\top)I_{\bar{\theta}}^{1/2}$, so

$$VIF\left(\bar{\theta}\right) = \sup_{\alpha\neq 0}\frac{\alpha^\top Q_{\bar{\theta}}^{-1}\alpha}{\alpha^\top I_{\bar{\theta}}^{-1}\alpha} = \max_{y\neq 0}\frac{y^\top\left(I_K - ww^\top\right)^{-1}y}{y^\top y}$$

$$= \frac{1}{1-w^\top w} = \frac{1}{1-\mathrm{E}_P[s_{\bar{\theta}}]^\top I_{\bar{\theta}}^{-1}\,\mathrm{E}_P[s_{\bar{\theta}}]/\chi^2(P\|P_{\bar{\theta}})},$$

since $I_K - ww^\top$ has eigenvalues 1 and $1 - w^\top w$, and $1 - w^\top w > 0$ because $Q_{\bar{\theta}}$ is positive definite.

Since $\chi^2\left(P\|P_{\bar{\theta}}\right) = \chi^2_{L^2}\left(P\|P_{\bar{\theta}}\right)$ by Lemma S4 (whose hypothesis holds since $I_{\bar{\theta}}$ is positive

definite), it remains only to show that

$$\chi^2_{\mathcal{S}_{\bar\theta}}(P\|P_{\bar\theta}) = \mathrm{E}_P[s_{\bar\theta}]^\top I_{\bar\theta}^{-1}\,\mathrm{E}_P[s_{\bar\theta}]. \tag{S1}$$

Towards this end, note that any $h \in \mathcal{S}_{\bar\theta}$ can be written as $h(X) = \beta^\top s_{\bar\theta}(X)$ and that $\mathrm{E}_{\bar\theta}[s_{\bar\theta}] = 0$. The left-hand side becomes

$$\sup_{\beta\neq 0}\frac{\left(\mathrm{E}_P\left[\beta^\top s_{\bar\theta}\right]\right)^2}{\mathrm{Var}_{\bar\theta}\left(\beta^\top s_{\bar\theta}\right)} = \sup_{\beta\neq 0}\frac{\left(\beta^\top \mathrm{E}_P[s_{\bar\theta}]\right)^2}{\beta^\top I_{\bar\theta}\beta}.$$

If $\mathrm{E}_P[s_{\bar\theta}] = 0$, both sides of (S1) are zero. Otherwise, by the Cauchy-Schwarz inequality the maximum is achieved at $\beta \propto I_{\bar\theta}^{-1}\,\mathrm{E}_P[s_{\bar\theta}]$ and the conclusion follows. □

S2 Data with Finite Support

This appendix studies when Assumption 1 can, and cannot, hold in settings with discrete data.

Suppose that $P \neq P_{\bar\theta}$, that $I_{\bar\theta}$ is positive definite, and that the model is identified at $\bar\theta$, in the sense that $P_\theta \neq P_{\bar\theta}$ for all other parameter values $\theta \in \Theta$. By Lemma S5 in Appendix S1, Assumption 1(iv) or (v) fails only if P belongs to

$$\mathcal{E}(\bar\theta) := \left\{(1 + v^\top s_{\bar\theta})P_{\bar\theta} : v \in \mathbb{R}^K\right\} \cup \bigcup_{\theta\in\Theta\setminus\{\bar\theta\}} \left\{P_{\bar\theta} + c(P_\theta - P_{\bar\theta}) : c \in \mathbb{R}\right\},$$

the union of the tangent plane of the model at $P_{\bar\theta}$ with all lines through $P_{\bar\theta}$ and another model point. Provided the model is sufficiently smooth in θ (as we assume below), this set is of dimension at most $K + 1$.

Now suppose that X takes values in $\{x_1, \ldots, x_J\}$, $J \geq 2$. We can identify distributions with the probability simplex $S := \{q \in [0,1]^J : \mathbf{1}^\top q = 1\}$ and signed measures with zero total mass with the set $H_0 := \{d \in \mathbb{R}^J : \mathbf{1}^\top d = 0\}$. Write the model as $\theta \mapsto p(\theta) \in S$ and the true distribution as $q \in S$, and let $\dot p(\theta) := \partial p(\theta)/\partial\theta^\top \in \mathbb{R}^{J\times K}$. We assume that Θ is compact, $\bar\theta$ lies in the interior of Θ, $p(\theta)$ is continuously differentiable on an open set containing Θ, $p(\bar\theta) > 0$ has full support, $\dot p(\bar\theta)$ has full rank (which requires that $J \geq K + 1$), and $p(\bar\theta) \neq p(\theta)$ for all $\theta \neq \bar\theta$. These conditions directly impose Assumption 1(i)-(iii), while the next result gives necessary and sufficient conditions for parts (iv) and (v) to hold as well.

Lemma S7 *Under the above conditions and $q \neq p(\bar{\theta})$, condition (iv) holds if and only if $q - p(\bar{\theta})$ does not lie in the column space of $\dot{p}(\bar{\theta})$, and condition (v) holds if and only if $q \notin \{p(\bar{\theta}) + c(p(\theta) - p(\bar{\theta})) : c \in \mathbb{R}\}$ for every $\theta \in \Theta \setminus \{\bar{\theta}\}$.*

Proof. The distribution $(1 + v^\top s_{\bar{\theta}})P_{\bar{\theta}}$ puts mass $p_j(\bar{\theta}) + (\dot{p}(\bar{\theta})v)_j$ on x_j, which gives the first claim. The second follows from Lemma S5(b), since $p(\theta) \neq p(\bar{\theta})$ for $\theta \neq \bar{\theta}$. □

Proposition S1 *Under the above conditions:*

(a) If $J = K+1$, Assumption 1(iv) fails for every $q \neq p(\bar{\theta})$, and every vector g of moments with $\mathrm{E}_q[g(X)] = \mathrm{E}_{\bar{\theta}}[g(X)]$ has a rank-deficient Jacobian $\mathrm{E}_{\bar{\theta}}[g(X)s_{\bar{\theta}}(X)^\top]$. When $K = 1$ and $J = 2$, every such moment is constant, so $\Theta_g = \Theta$.

(b) If $J \geq K+3$, the set of q for which Assumption 1(iv) or (v) fails has Lebesgue measure zero, so Assumption 1 holds at $\bar{\theta}$ for almost every q.

(c) If $J = K + 2$, the set of q for which Assumption 1 fails can have positive Lebesgue measure.

Proof. (a) Since $\mathbf{1}^\top p(\theta) = 1$ for all θ, the column space of $\dot{p}(\bar{\theta})$ lies in H_0. Both have dimension $K = J - 1$, so they are equal and $q - p(\bar{\theta}) = \dot{p}(\bar{\theta})v^*$ for some $v^* \neq 0$. Assumption 1(iv) thus fails by Lemma S7. For any g with $\mathrm{E}_q[g] = \mathrm{E}_{\bar{\theta}}[g]$, however, it follows that

$$0 = \mathrm{E}_q[g] - \mathrm{E}_{\bar{\theta}}[g] = \sum_j g(x_j)(\dot{p}(\bar{\theta})v^*)_j = \mathrm{E}_{\bar{\theta}}[g s_{\bar{\theta}}^\top]v^*,$$

so the Jacobian annihilates v^*. When $K = 1$ and $J = 2$, however, the Jacobian is a nonzero multiple of $g(x_2) - g(x_1)$, which must therefore vanish.

(b) By Lemma S7, Assumption 1(iv) or (v) fails only for q in $A \cup B$, where

$$A := \left\{p(\bar{\theta}) + \dot{p}(\bar{\theta})v : v \in \mathbb{R}^K\right\}, \qquad B := \left\{p(\bar{\theta}) + c\left(p(\theta) - p(\bar{\theta})\right) : \theta \in \Theta,\ c \in \mathbb{R}\right\}.$$

A is an affine set of dimension $K < J - 1$, and so has $(J-1)$-dimensional Lebesgue measure zero. Consequently, its intersection with the simplex S also has zero Lebesgue measure. The set B, the union of the lines through $p(\bar{\theta})$ and the other model points, is the image of $\Theta \times \mathbb{R}$ under the map $\varphi(\theta, c) := p(\bar{\theta}) + c(p(\theta) - p(\bar{\theta}))$. Since p is continuously differentiable on an open set containing Θ, φ is locally Lipschitz on an open subset of $\mathbb{R}^{K+1}$ containing $\Theta \times \mathbb{R}$, so B has Hausdorff dimension at most $K + 1 < J - 1$, and so also has $(J-1)$-dimensional Lebesgue measure zero.

(c) See Example S1 below. □

Example S1 (Three support points) Suppose $X \in \{0,1,2\}$, and that P_θ corresponds to the Binomial$(2,\theta)$ model $p(\theta) = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)$, and $\Theta = [a,b] \subset (0,1)$. Direct calculation yields

$$p(\theta) - p(\bar{\theta}) = (\theta - \bar{\theta})\, d(\theta), \quad d(\theta) = \big(\theta + \bar{\theta} - 2,\ 2(1-\theta-\bar{\theta}),\ \theta + \bar{\theta}\big)\,,$$

from which it is immediate that $d(\bar{\theta}) = \dot{p}(\bar{\theta})$. Every $d \in H_0$ is determined by $u := d_3 - d_1$ and $w := d_2$, and $d(\theta)$ has $(u,w) = (2, 2(1-\theta-\bar{\theta}))$. Let $u_q := (q_3 - q_1) - (p_3(\bar{\theta}) - p_1(\bar{\theta}))$ and $w_q := q_2 - p_2(\bar{\theta})$ denote the coordinates of $q - p(\bar{\theta})$. By Lemma S7, for $q \neq p(\bar{\theta})$ Assumption 1 fails at $\bar{\theta}$ if and only if $u_q \neq 0$ and $\frac{w_q}{u_q} \in \big[1-b-\bar{\theta},\ 1-a-\bar{\theta}\big]$. Both the failure set and its complement contain nonempty open subsets of the simplex near $p(\bar{\theta})$.

S3 Normal-Location Example and Robust Variance

This section provides details on the normal-location example from the main text. Write ϕ and Φ for the standard normal density and distribution function, and $p(x) = \frac{1}{2}e^{-|x|}$ for the density of P.

S3.1 Verifying Assumption 1

Since $p(x)^2/\phi(x-\bar{\theta}) = \frac{\sqrt{2\pi}}{4}\exp\{(x-\bar{\theta})^2/2 - 2|x|\}$ is not integrable, $\chi^2(P\|P_{\bar{\theta}}) = \infty$ for every $\bar{\theta}$. The normal location model is continuous in total variation and DQM, with score $s_{\bar{\theta}}(x) = x - \bar{\theta}$ and $I_{\bar{\theta}} = 1$, so Assumption 1(i)-(iii) hold. Assumption 1(iv) holds because the divergence is infinite (Lemma S5(a) in Appendix S1). Since also $\chi^2(P_\theta\|P_{\bar{\theta}}) = \exp\{(\theta-\bar{\theta})^2\} - 1 < \infty$ and $P_\theta \neq P_{\bar{\theta}}$ for $\theta \neq \bar{\theta}$, (v) holds by Lemma S5(b).

S3.2 Near-Efficient Moments

By Proposition 2, the model-implied variance cannot be smaller than $I_{\bar{\theta}}^{-1} = 1$. Considering just-identifying moments as in the proof of Proposition 2(a), if a just-identifying $g \in \mathcal{C}(\bar{\theta})$ attained the bound, then by Lemma S6 we know $g = As_{\bar{\theta}} + c$ $P_{\bar{\theta}}$-almost surely for some c and A full-rank. Since P is absolutely continuous with respect to $P_{\bar{\theta}}$, the same holds P-almost surely, and matching then requires $\mathrm{E}_P[s_{\bar{\theta}}(X)] = -\bar{\theta} = 0$. Hence the bound is attained only at $\bar{\theta} = 0$, the probability limit of maximum likelihood in this example.

To address cases where the bound cannot be attained exactly, for $M > 0$ let $X_M := \max\{\min\{X, M\}, -M\}$. Under P_θ, X_M has density $\phi(x-\theta)$ on $(-M, M)$, and atoms $\Phi(-M-\theta)$ at $-M$ and $1-\Phi(M-\theta)$ at M. Under P it instead has density $\frac{1}{2}e^{-|x|}$ on $(-M, M)$ and atoms $\frac{1}{2}e^{-M}$ at $\pm M$. The censored model is DQM, with score

$$s_M(x) = x - \bar{\theta} \text{ for } |x| < M, \qquad s_M(-M) = -\frac{\phi(M+\bar{\theta})}{\Phi(-M-\bar{\theta})}, \qquad s_M(M) = \frac{\phi(M-\bar{\theta})}{1-\Phi(M-\bar{\theta})}.$$

Let r_M denote the recentered density ratio for the censored data (with respect to Lebesgue measure on $(-M, M)$ plus counting measure on $\{-M, M\}$). Since r_M is bounded, $\chi^2_M := \mathrm{E}_{\bar{\theta}}[r_M(X_M)^2] < \infty$, and Proposition 2(b), applied to the censored data, implies that the efficient moment among functions of X_M is $g^*_M = s_M - \frac{\mathrm{E}_P[s_M(X_M)]}{\chi^2_M} r_M$, with model-implied variance $\{I_M - \mathrm{E}_P[s_M(X_M)]^2/\chi^2_M\}^{-1}$ for I_M the censored-data Fisher information. Figure 1(a) plots examples of the functions g^*_M, while Figure 1(b) plots examples of the resulting (model-implied) estimator variance.

S3.3 Misspecification-Robust Variance

For $g \in L^2(P; \mathbb{R}^D) \cap L^1(P_{\bar{\theta}}; \mathbb{R}^D)$ with $\mathrm{E}_P[g(X)] = \mathrm{E}_{\bar{\theta}}[g(X)]$, finite G, and weight matrix W, the misspecification-robust asymptotic variance replaces Ω in (3) with its analog under P,

$$\mathrm{AVar}_R(\hat{\theta}_{g,W}) := (G^\top W G)^{-1} G^\top W \,\mathrm{Var}_P(g(X)) W G (G^\top W G)^{-1},$$

where we again take $\mathrm{AVar}_R(\hat{\theta}_{g,W})$ to be infinite when $G^\top W G$ is singular. The following result shows that, in the special case of a scalar parameter and just-identifying moments, moments with model-implied variance approaching $I^{-1}_{\bar{\theta}}$ necessarily have diverging robust variance.

Proposition S2 *Suppose Assumption 1(iii) holds, $K = 1$, P is absolutely continuous with respect to $P_{\bar{\theta}}$, $\mathrm{E}_P[s_{\bar{\theta}}(X)^2] < \infty$, and $\mathrm{E}_P[s_{\bar{\theta}}(X)] \neq 0$. If $g_l \in \mathcal{C}(\bar{\theta})$ are just-identifying moments with $G_l = \mathrm{E}_{\bar{\theta}}[g_l(X) s_{\bar{\theta}}(X)] \neq 0$ and $\mathrm{AVar}(\hat{\theta}_{g_l}) \to I^{-1}_{\bar{\theta}}$, then $\mathrm{Var}_P(g_l(X))/G_l^2 \to \infty$.*

Proof. Let $h_l := (g_l - \mathrm{E}_{\bar{\theta}}[g_l])/G_l$, so that $\mathrm{E}_{\bar{\theta}}[h_l] = \mathrm{E}_P[h_l] = 0$, $\mathrm{E}_{\bar{\theta}}[h_l s_{\bar{\theta}}] = 1$, and $\mathrm{Var}_P(g_l)/G_l^2 = \mathrm{E}_P[h_l^2]$. Write $h_l = I^{-1}_{\bar{\theta}} s_{\bar{\theta}} + u_l$, where $\mathrm{E}_{\bar{\theta}}[u_l s_{\bar{\theta}}] = 0$. Then $\mathrm{AVar}(\hat{\theta}_{g_l}) = \mathrm{E}_{\bar{\theta}}[h_l^2] = I^{-1}_{\bar{\theta}} + \mathrm{E}_{\bar{\theta}}[u_l^2]$, so $u_l \to 0$ in $L^2(P_{\bar{\theta}})$, and hence in $P_{\bar{\theta}}$-probability. By absolute continuity, $u_l \to 0$ in P-probability as well. Matching implies $\mathrm{E}_P[u_l] = -I^{-1}_{\bar{\theta}} \mathrm{E}_P[s_{\bar{\theta}}] \neq 0$ for all l. Suppose that $\mathrm{E}_P[h_l^2]$ were bounded along some subsequence. Since $\mathrm{E}_P[u_l^2] \leq 2\,\mathrm{E}_P[h_l^2] + 2 I^{-2}_{\bar{\theta}} \mathrm{E}_P[s^2_{\bar{\theta}}]$, the u_l would then be uniformly integrable under P along that subsequence, so $\mathrm{E}_P[u_l] \to 0$, a

contradiction. Hence $\mathrm{E}_P[h_l^2] \to \infty$. □

In the normal example $\mathrm{E}_P[s_{\bar\theta}(X)] = -\bar\theta$, while $\mathrm{E}_P[s_{\bar\theta}(X)^2] = 2 + \bar\theta^2$, so Proposition S2 applies for every $\bar\theta \neq 0$. Table S1 illustrates this divergence for the estimates based on moments g_M^*. At $\bar\theta = 0$, by contrast, estimates based on g_M^* approach the MLE (which has asymptotic variance 2 in this example) as $M \to \infty$.

	g_M^*: robust			g_R
$\bar\theta$	$M = 2$	$M = 3$	$M = 4$	V_R
0	1.4	1.7	1.9	1.01
0.25	1.6	2.1	4.7	1.40
0.5	3.0	6.3	19.8	1.97
1	9.2	26.8	87.7	1.47
1.5	21.5	66.3	208.2	0.77
2	51.4	132.3	384.6	0.42

Table S1: Normal-location example with Laplace data. Misspecification-robust asymptotic variance of estimates based on the moments g_M^* and g_R. Maximum likelihood has a robust variance of 2 and a pseudo-true value of zero.

The explosive variance found in Table S1 reflects the disagreement between the model-implied and robust asymptotic variance. If we instead aim to minimize the robust variance, we can obtain a result similar to Proposition 2. For simplicity, here we state the result for the special case of a finite (reverse) χ^2 divergence.

Proposition S3 *Suppose Assumption 1(iii) holds, $P_{\bar\theta}$ is absolutely continuous with respect to P with $\ell := dP_{\bar\theta}/dP$, $0 < \chi^2(P_{\bar\theta}\|P) < \infty$, and $J_R := \mathrm{E}_{\bar\theta}[s_{\bar\theta}(X)s_{\bar\theta}(X)^\top \ell(X)]$ is finite. Let*

$$t := s_{\bar\theta}\ell - \frac{\mathrm{E}_{\bar\theta}[s_{\bar\theta}(X)\ell(X)]}{\chi^2(P_{\bar\theta}\|P)}(\ell - 1), \qquad Q_R := J_R - \frac{\mathrm{E}_{\bar\theta}[s_{\bar\theta}(X)\ell(X)]\ \mathrm{E}_{\bar\theta}[s_{\bar\theta}(X)\ell(X)]^\top}{\chi^2(P_{\bar\theta}\|P)},$$

and suppose Q_R is positive definite. Then for every $g \in L^2(P;\mathbb{R}^D)$ with $\mathrm{E}_P[g(X)] = \mathrm{E}_{\bar\theta}[g(X)]$ and positive definite W, $\mathrm{AVar}_R(\hat\theta_{g,W}) \succeq V_R := Q_R^{-1}$, with equality for $g_R := t$.

Proof. Since $P_{\bar\theta}$ is absolutely continuous with respect to P, $\mathrm{E}_{\bar\theta}[f] = \mathrm{E}_P[f\ell]$ for $P_{\bar\theta}$-integrable f. Let $\tilde s := s_{\bar\theta}\ell$, so that $\mathrm{E}_P[\tilde s] = \mathrm{E}_{\bar\theta}[s_{\bar\theta}] = 0$, $\mathrm{E}_P[\tilde s\tilde s^\top] = J_R$, and $G = \mathrm{E}_P[g\tilde s^\top]$, which is finite by the Cauchy-Schwarz inequality, as is $\mathrm{E}_{\bar\theta}[g] = \mathrm{E}_P[g\ell]$ since $\mathrm{E}_P[\ell^2] = 1 + \chi^2(P_{\bar\theta}\|P)$. Minimizing $\mathrm{AVar}_R(\hat\theta_{g,W})$ subject to $\mathrm{E}_P[g] = \mathrm{E}_{\bar\theta}[g]$ is thus the same as the finite-divergence problem in Proposition 2, but with the roles of $P_{\bar\theta}$ and P interchanged and $s_{\bar\theta}$ replaced by $\tilde s$, so that r becomes $\ell - 1$, $\mathrm{E}_P[s_{\bar\theta}]$ becomes $\mathrm{E}_{\bar\theta}[s_{\bar\theta}\ell]$, $I_{\bar\theta}$ becomes J_R, and $Q_{\bar\theta}$ becomes Q_R.

However, the relevant portions in the proof of Proposition 2 are not specific to the score, using only that it is square-integrable with mean zero under the base measure, and that G is its covariance with g, together with positive definiteness of the analog of $Q_{\bar{\theta}}$, which we assume directly. These arguments thus continue to apply, from which the result is immediate. □

One may similarly generalize the infinite-divergence case and part (c) of Proposition 2. Note, however, that since we now consider the robust variance the ($\bar{\theta}$-specific) asymptotic variance of maximum likelihood is no longer a lower bound. To illustrate Proposition S3, the final column of Table S1 reports the minimized robust variance, which for the targets $\bar{\theta}$ considered we find to be smaller than the variance of maximum likelihood in this example (which is equal to two).

S4 Further Details on Application

This appendix provides further details for the application discussed in Section 4. In particular, we discuss how we apply our adversarial construction to this example and provide further information on the model-implied standard errors. Following the replication code and data of Cocci and Plagborg-Møller (2025), we use weekly UPC-level beer prices for Dominick's store 122, keep prices between \$0.20 and \$25, treat price changes smaller than one cent as zero, and drop observations above the 99th percentile of absolute log price changes, leaving 11,111 nonzero price changes. Again following Cocci and Plagborg-Møller (2025), we also demean price changes since Alvarez and Lippi (2014) abstract from inflation.

Alvarez and Lippi (2014) consider a model parameterized by N, the volatility of desired prices σ, and a scaled menu-cost parameter s. The threshold in the optimal policy is then $\bar{y} \approx \sigma s\sqrt{2(N+2)}$. Proposition 6 in Alvarez and Lippi (2014) implies that the density of X depends on (N, σ, s) only through $(N, \bar{y})$. Once we condition on nonzero price-changes, the density of X is given by

$$p_\theta(x) = \frac{2}{\sqrt{\bar{y}}\,\mathrm{B}\left(\frac{1}{2}, \frac{N-1}{2}\right)} \left(1 - \frac{x^2}{\bar{y}}\right)^{(N-3)/2} \quad \text{on } [0, \sqrt{\bar{y}}],$$

where B denotes the beta function.

One can show that due to parameter-dependence of the support, the model is not DQM in $\bar{y}$ when $N \leq 5$. We thus consider an adversary who knows only the distribution of the

censored absolute centered log price change

$$X_c = \begin{cases} X, & X \leq c, \\ c, & X > c, \end{cases}$$

where we use the censoring value $c = 0.22$, which is somewhat below the smallest upper bound on the support, $\sqrt{0.06} \approx 0.245$, obtained over our grid of target parameters.[1] Under P_θ, the continuous part on $[0, c]$ is the original model density $p_\theta(x)$, while the point c has mass $P_\theta(X > c)$.

We next verify Assumption 1. Let Θ be a compact set containing our targets such that $N > 1$ and $\bar{y} > c^2$, so that $c < \sqrt{\bar{y}}$ for all $\theta \in \Theta$. On the set of θ values with $N > 1$ and $\bar{y} > c^2$, the distribution of X_c is continuous on $[0, c)$, and has a point mass at c. Moreover, the density is continuously differentiable in θ, where the score is uniformly bounded local to each θ. Consequently, the censored model is continuous in total variation by Scheffé's lemma, its Fisher information is continuous by dominated convergence, and it is DQM by van der Vaart (1998, Lemma 7.6). Moreover, since the scores with respect to $\bar{y}$ and N are linearly independent for all $N > 1$, the Fisher information has full rank. Thus, the censored model satisfies Assumption 1(i)-(iii) at every $\bar{\theta} \in \Theta$.

Conditions (iv) and (v) of Assumption 1 restrict the true distribution of X_c, which we do not know, but they hold for $\hat{P}$ at every $\bar{\theta} \in \Theta$. In particular, since $c < \sqrt{\bar{y}}$ on Θ, the model densities and scores are analytic in x on $(0, c)$. By contrast, the density of $\hat{P}$ on $(0, c)$ is e^S for a cubic spline S whose third derivative jumps at each interior knot. Since S, S', and S'' are continuous, the third derivative of e^S jumps by e^S times the jump in S''', so e^S is not almost everywhere equal to any analytic function on $(0, c)$. Hence $\hat{P}$ does not have density $(1 + v^\top s_{\bar{\theta}})p_{\bar{\theta}}$, so Assumption 1(iv) holds. Likewise, if any of $\hat{P}$, $P_{\bar{\theta}}$, and P_θ can be written as a non-trivial mixture of the other two, we can solve for $\hat{P}$ in terms of the other two, which would mean that $\hat{P}$ has an analytic density on $(0, c)$. Assumption 1(v) can thus fail only if point identification fails for θ. However, the derivative of the log density on $(0, c)$ is given by $-(N-3)x/(\bar{y} - x^2)$ which determines N and, for $N \neq 3$, $\bar{y}$. If instead $N = 3$ the density is $1/\sqrt{\bar{y}}$, which again identifies $\bar{y}$.

To mitigate the data re-use noted in Section 4, we also compute cross-fit estimates. For this approach, reported in Figure S1, we split the sample into 10 disjoint folds, say

[1] Censoring in this way preserves more information than truncating to price changes below the threshold, since it also preserves the tail probability $P_\theta(X > c)$.

$\{\mathcal{I}_1, \ldots, \mathcal{I}_{10}\}$ with $\cup_j \mathcal{I}_j = \{1, \ldots, n\}$, where for each fold we estimate $\hat{P}_{-j}$ on the data excluding fold j, evaluate $g_{\bar{\theta},\hat{P}_{-j}}$ on fold j, and solve the average moment equation

$$\frac{1}{10}\sum_{j=1}^{10}\left(\frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} g_{\bar{\theta},\hat{P}_{-j}}(X_{c,i}) - \mathrm{E}_\theta[g_{\bar{\theta},\hat{P}_{-j}}(X_c)]\right) = 0.$$

Cross-fitting reduces the dependence between the estimated data distribution and the sample moments, but comes at the cost of higher variability in the estimated data distribution due to the smaller sample size (though with 10-fold cross-fitting this cost is limited, as reflected by the similarity of Figures 3 and S1).

Table S2 reports the variance inflation factors for the specific parameters N and $\bar{y}$. In particular, for an evaluation point θ, define

$$\sqrt{VIF_N(\theta;\bar{\theta})} - 1 \;=\; \sqrt{\frac{\mathrm{AVar}_{g_{\bar{\theta},\hat{P}}}(\theta)_{N,N}}{\mathrm{AVar}_{\mathrm{MLE}}(\theta)_{N,N}}} - 1, \qquad \sqrt{VIF_{\bar{y}}(\theta;\bar{\theta})} - 1 \;=\; \sqrt{\frac{\mathrm{AVar}_{g_{\bar{\theta},\hat{P}}}(\theta)_{\bar{y},\bar{y}}}{\mathrm{AVar}_{\mathrm{MLE}}(\theta)_{\bar{y},\bar{y}}}} - 1,$$

where both asymptotic-variance matrices are evaluated under P_θ, while the adversarial moment function is held fixed at the one constructed for target $\bar{\theta}$. The left panel evaluates these quantities at the construction target, $\theta = \bar{\theta}$, and is the direct finite-sample analog of Proposition 3. The middle panel evaluates the same fixed moment function at the same-sample adversarial estimate, $\theta = \hat{\theta}_{g_{\bar{\theta},\hat{P}}}$, and so parallels the construction in Figure 3. The right panel reports $\sqrt{\mathrm{AVar}_N/\mathrm{AVar}_{\mathrm{MLE},N}} - 1$ and $\sqrt{\mathrm{AVar}_{\bar{y}}/\mathrm{AVar}_{\mathrm{MLE},\bar{y}}} - 1$, where $\mathrm{AVar}_N := \mathrm{AVar}_{g_{\bar{\theta},\hat{P}}}(\hat{\theta}_{g_{\bar{\theta},\hat{P}}})_{N,N}$ is the model-implied asymptotic variance of the adversarial estimator at its own estimate and $\mathrm{AVar}_{\mathrm{MLE},N} := \mathrm{AVar}_{\mathrm{MLE}}(\hat{\theta}_{ML})_{N,N}$ that of the MLE at its own estimate, and analogously for $\bar{y}$. Since the Fisher information varies considerably over the parameter space, these columns are more dispersed than the others.

Finally, we plot the moment functions underlying each estimate. For each target $\bar{\theta} = (N, \bar{y})$ on our grid, Proposition 2(b) gives the adversarial moment function

$$g_{\bar{\theta},\hat{P}} = s_{\bar{\theta}} - \frac{\mathrm{E}_{\hat{P}}\left[s_{\bar{\theta}}(X_c)\right]}{\chi^2(\hat{P}\|P_{\bar{\theta}})} r.$$

We thus obtain pair of moment functions for each target value. The estimate is unchanged if we replace g by $Ag + b$ for invertible A and any b. Any two functions with the same span (up to a constant shift) thus yield the same estimate. To choose a pair to plot, we first restrict attention to pairs which are mean-zero, orthogonal, and of unit variance under

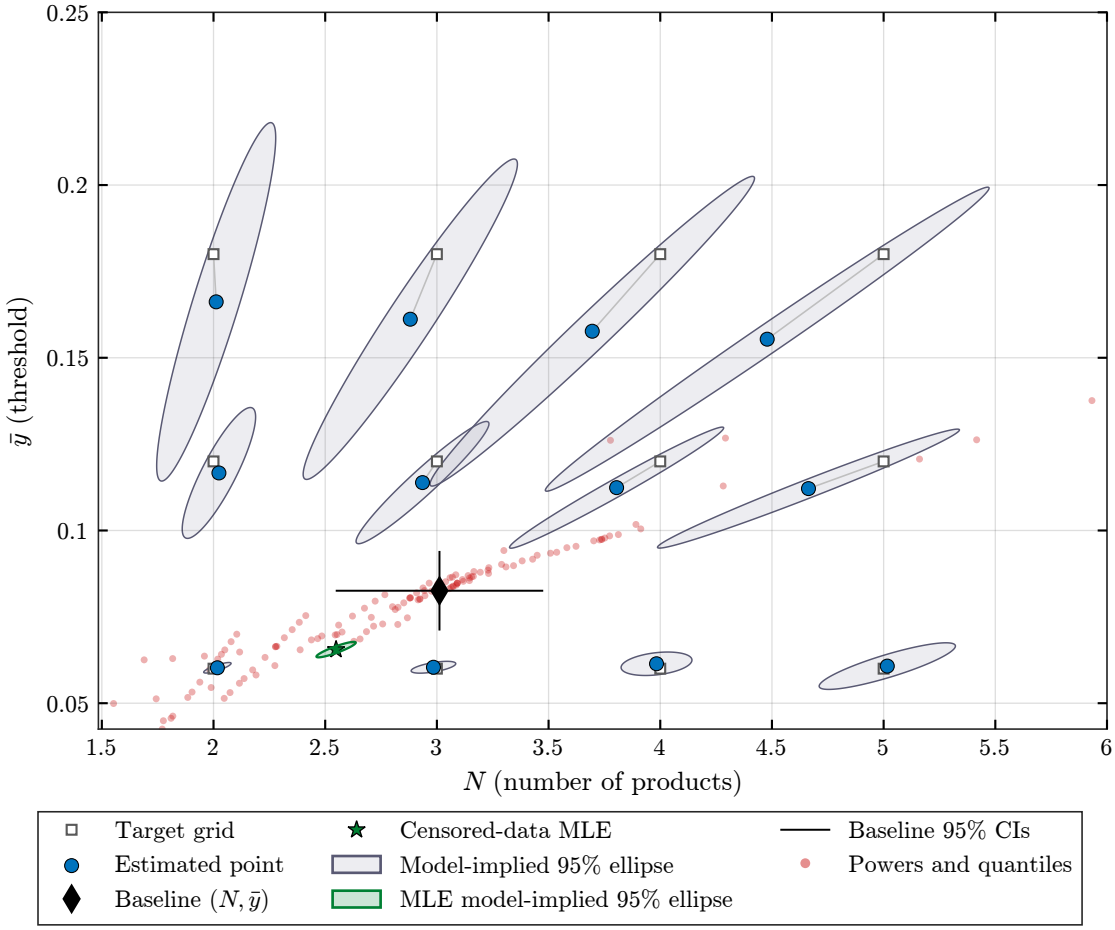


Figure S1: 10-fold cross-fit adversarial estimates from Proposition 2 with model-implied uncertainty. Gray lines connect each target to its estimate. Shaded regions are model-implied Wald confidence ellipses around each $\hat{\theta}$, using χ^2_1 critical values so that their projections onto each axis are nominal 95% model-implied confidence intervals, treating the moments as fixed. The red scatter shows, for reference, the varying-moment estimates from Figure 2, while the black diamond and bars are the baseline $(N, \bar{y})$ estimate and confidence intervals. Finally, the green star and ellipse show the maximum likelihood estimate and model-implied confidence ellipse, again using a χ^2_1 critical value.

Lebesgue measure on $[0, c]$. Among these, we plot the pair which minimizes the sum of the total variations of the two functions on $[0, c]$, so the plotted functions oscillate as little as possible. The point mass at $X_c = c$ plays no role in the normalization, but we apply the same affine transformation to the value of $g_{\bar{\theta},\hat{P}}$ at $X_c = c$ and plot the result as a dot. Figure S2 plots the normalized moments, with a separate facet for each target.

		At target $\bar{\theta}$		At estimate $\hat{\theta}_{g_{\bar{\theta},\hat{P}}}$		Relative to MLE at $\hat{\theta}_{ML}$	
N	$\bar{y}$	$\sqrt{VIF_N}-1$	$\sqrt{VIF_{\bar{y}}}-1$	$\sqrt{VIF_N}-1$	$\sqrt{VIF_{\bar{y}}}-1$	$\sqrt{\frac{\text{AVar}_N}{\text{AVar}_{\text{MLE},N}}}-1$	$\sqrt{\frac{\text{AVar}_{\bar{y}}}{\text{AVar}_{\text{MLE},\bar{y}}}}-1$
2	0.06	24.8%	5.3%	24.5%	5.2%	−30.5%	−32.2%
2	0.12	2.6%	3.3%	2.7%	3.7%	87.5%	754.0%
2	0.18	0.1%	5.0%	0.2%	5.2%	206.4%	2266.5%
3	0.06	4.9%	12.8%	5.1%	11.4%	13.6%	−23.6%
3	0.12	5.7%	15.4%	5.2%	15.6%	242.3%	700.6%
3	0.18	5.7%	15.2%	4.9%	15.6%	456.9%	2021.8%
4	0.06	0.5%	100.6%	2.0%	80.0%	79.5%	51.8%
4	0.12	20.0%	22.8%	19.7%	22.8%	453.1%	694.0%
4	0.18	17.0%	24.7%	15.9%	24.9%	746.6%	1953.6%
5	0.06	41.6%	218.7%	39.1%	215.3%	245.7%	196.2%
5	0.12	27.9%	21.8%	28.7%	22.1%	679.9%	679.4%
5	0.18	26.8%	31.0%	26.1%	31.1%	1059.0%	1916.1%
Median		11.3%	18.6%	10.6%	18.8%	244.0%	697.3%

Table S2: Left and middle panels: proportional increase $\sqrt{VIF_N}-1$ and $\sqrt{VIF_{\bar{y}}}-1$ in the model-implied standard error of the adversarial estimator relative to the MLE, evaluated at the target $\theta=\bar{\theta}$ (left panel) or at the same-sample adversarial estimate $\theta=\hat{\theta}_{g_{\bar{\theta},\hat{P}}}$ (middle panel). Right panel: proportional difference $\sqrt{\text{AVar}_N/\text{AVar}_{\text{MLE},N}}-1$ and $\sqrt{\text{AVar}_{\bar{y}}/\text{AVar}_{\text{MLE},\bar{y}}}-1$ between the root model-implied asymptotic variance of the adversarial estimator at $\hat{\theta}_{g_{\bar{\theta},\hat{P}}}$ and that of the censored-data MLE at its own estimate.

S5 Misspecified Conditional Moment Equality Models

This appendix develops analogs of Propositions 1 and 2 for misspecified conditional moment equality models, where the researcher now selects what functions to use as instruments, rather than what moments to match. We show that if the conditional moment equality is misspecified then, under regularity conditions, any target can again be made the unique population solution. We then characterize the variance cost of targeting relative to the usual semiparametric efficiency formula, evaluated as if the conditional moments held at the target.

S5.1 Setting and Uniqueness of Estimand

Suppose the researcher postulates the conditional moment equalities

$$\mathrm{E}_P[\rho(X,\theta)\mid Z]=0,\qquad \rho(X,\theta)\in\mathbb{R}^d,\quad \theta\in\Theta\subseteq\mathbb{R}^K, \tag{S2}$$

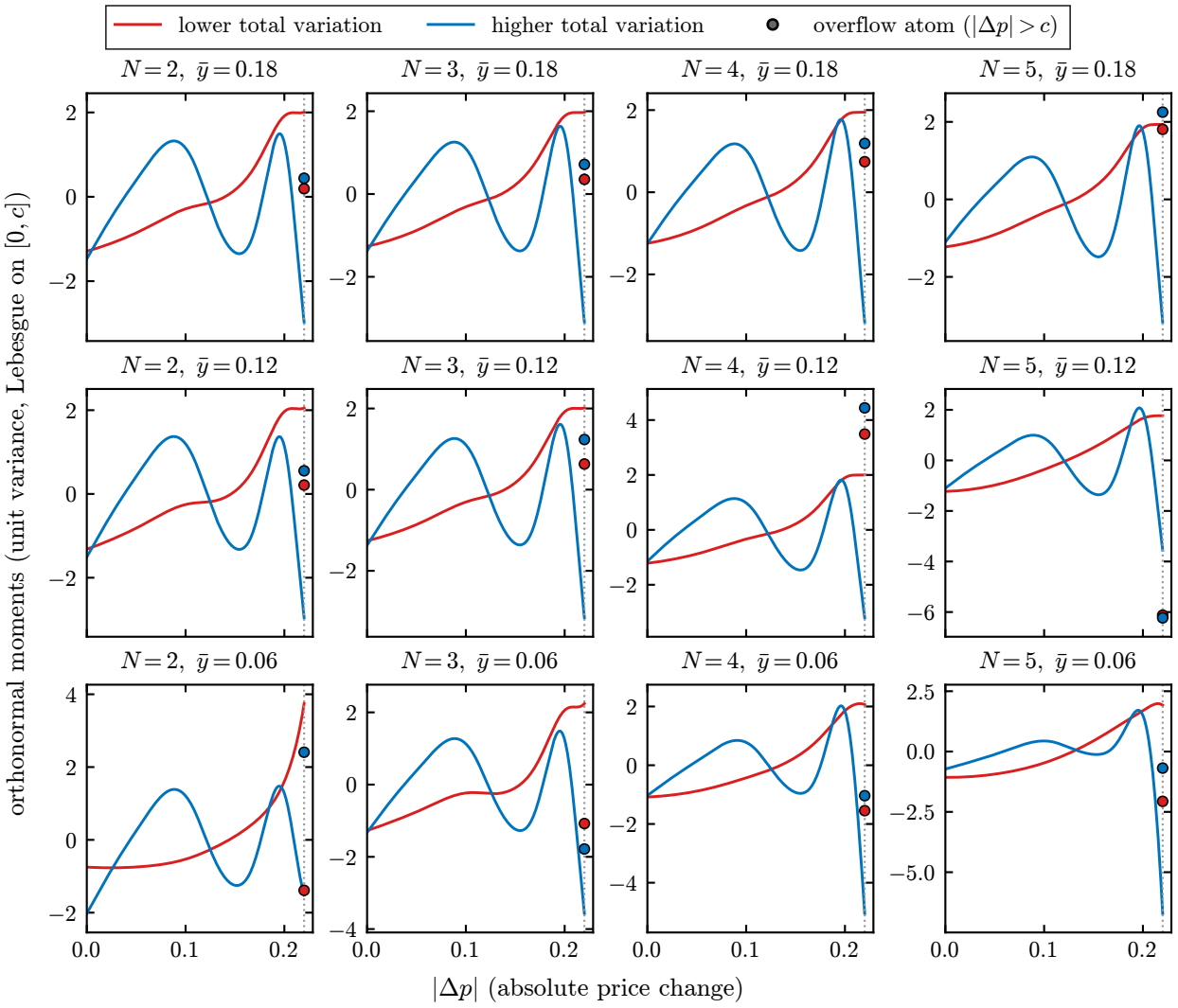


Figure S2: Adversarial moment functions on the $(N, \bar{y})$ grid, evaluated at the same-sample censored-logspline estimate $\hat{P}$ of Section 4. Each facet corresponds to one target $\bar{\theta}$, and applies the linear normalization described in the text. The dots at $|\Delta p| = c$ show the values for $X > c$.

where Z is a component of X. Let P_Z denote the distribution of Z under P, $\mathcal{Z}$ its support, and $m_\theta(z) := \mathrm{E}_P[\rho(X, \theta) \mid Z = z]$, where we assume that $\rho(X, \theta)$ is integrable under P for all $\theta \in \Theta$. We consider the misspecified case where there is no $\theta \in \Theta$ with $m_\theta = 0$ P_Z-almost surely. Fix a target $\bar{\theta}$, and note that $m_{\bar{\theta}}$ is nonzero as an element of $L^1(P_Z; \mathbb{R}^d)$.

The researcher chooses a finite matrix of instruments $A : \mathcal{Z} \to \mathbb{R}^{D \times d}$ (which may depend on P and $\bar{\theta}$) and uses estimators which minimize the (weighted) deviation of the sample moments $\frac{1}{n}\sum_{i=1}^n A(Z_i)\rho(X_i, \theta)$ from zero. By the law of iterated expectations, $\mathrm{E}_P[A(Z)\rho(X, \theta)] = \mathrm{E}_{P_Z}[A(Z)m_\theta(Z)]$ whenever $A(Z)\rho(X, \theta)$ is integrable under P (for in-

stance when A is bounded), so a target $\bar{\theta}$ solves the population moments exactly if and only if

$$\mathrm{E}_{P_Z}[A(Z)m_{\bar{\theta}}(Z)] = 0. \tag{S3}$$

This condition requires that $m_{\bar{\theta}}(Z)$ be orthogonal to the rows of $A(Z)$ under P_Z. We call the rows of such A *targeting instruments* for $\bar{\theta}$. Let $\mathcal{B}$ denote the set of bounded measurable instruments $a : \mathcal{Z} \to \mathbb{R}^d$, and let

$$\mathcal{T}(\bar{\theta}) := \left\{a \in \mathcal{B} : \ \mathrm{E}_{P_Z}[a(Z)^\top m_{\bar{\theta}}(Z)] = 0\right\}$$

denote the set of bounded targeting instruments. Parallel to Proposition 1, we first ask when there exists a finite-dimensional A (which is to say, a finite collection of targeting instruments) yielding a unique solution $\bar{\theta}$.

Assumption S1 *Assume:*

(i) Θ *is compact.*

(ii) The map $\theta \mapsto m_\theta$ *is continuous from* Θ *into* $L^1(P_Z; \mathbb{R}^d)$.

(iii) This map is differentiable at $\bar{\theta}$ *in* $L^1(P_Z; \mathbb{R}^d)$*, with derivative* $\Gamma_{\bar{\theta}} \in L^1(P_Z; \mathbb{R}^{d\times K})$*:*

$$\mathrm{E}_{P_Z}\left[\|m_{\bar{\theta}+h}(Z) - m_{\bar{\theta}}(Z) - \Gamma_{\bar{\theta}}(Z)h\|\right] = o(\|h\|).$$

(iv) There exists $A : \mathcal{Z} \to \mathbb{R}^{K\times d}$ *with rows in* $\mathcal{T}(\bar{\theta})$ *such that* $\mathrm{E}_{P_Z}[A(Z)\Gamma_{\bar{\theta}}(Z)]$ *is nonsingular.*

(v) For every $\theta \neq \bar{\theta}$*, there exists* $a \in \mathcal{T}(\bar{\theta})$ *with* $\mathrm{E}_{P_Z}[a(Z)^\top m_\theta(Z)] \neq 0$.

Conditions (i)-(iii) are regularity conditions paralleling those in Assumption 1. By condition (iii), the Jacobian of the population moments $\mathrm{E}_{P_Z}[A(Z)m_\theta(Z)]$ at $\bar{\theta}$ is $\mathrm{E}_{P_Z}[A(Z)\Gamma_{\bar{\theta}}(Z)]$ for bounded A, so condition (iv) requires that some bounded targeting instruments imply a nonsingular Jacobian for the resulting population moment conditions, which is in turn necessary for $\sqrt{n}$-consistency of the corresponding GMM estimator. Thus the existence of such moments is necessary for the existence of "nice" moments in the usual sense. Finally, condition (v) requires that for any $\theta \neq \bar{\theta}$ there exist $a \in \mathcal{T}(\bar{\theta})$ such that the resulting moment is nonzero at θ. Absent this condition, it is mechanically impossible to isolate $\bar{\theta}$ since any moments which are solved there are also solved at θ. The next lemma shows that Assumption S1(iv) and (v) are equivalent to requiring that m_θ and $\Gamma_{\bar{\theta}}v$ not be multiples of $m_{\bar{\theta}}$.

Lemma S8 *For $f \in L^1(P_Z; \mathbb{R}^d)$, there exists $a \in \mathcal{T}(\bar{\theta})$ with $\mathrm{E}_{P_Z}[a(Z)^\top f(Z)] \neq 0$ if and only if f is not a scalar multiple of $m_{\bar{\theta}}$. Consequently, condition (v) of Assumption S1 holds if and only if m_θ is not a scalar multiple of $m_{\bar{\theta}}$ for any $\theta \neq \bar{\theta}$, and under condition (iii), condition (iv) holds if and only if $\Gamma_{\bar{\theta}} v$ is not a scalar multiple of $m_{\bar{\theta}}$ for any nonzero $v \in \mathbb{R}^K$.*

Proof. Towards the first claim, if $f = cm_{\bar{\theta}}$ for $c \in \mathbb{R}$, then

$$\mathrm{E}_{P_Z}[a(Z)^\top f(Z)] = c\,\mathrm{E}_{P_Z}[a(Z)^\top m_{\bar{\theta}}(Z)] = 0$$

for every $a \in \mathcal{T}(\bar{\theta})$. Since $m_{\bar{\theta}} \neq 0$, there exists a bounded measurable b_0 with $\mathrm{E}_{P_Z}[b_0(Z)^\top m_{\bar{\theta}}(Z)] = 1$. Conversely, suppose that $\mathrm{E}_{P_Z}[a(Z)^\top f(Z)] = 0$ for every $a \in \mathcal{T}(\bar{\theta})$, and let $c := \mathrm{E}_{P_Z}[b_0(Z)^\top f(Z)]$. For any bounded measurable a, the instrument $\tilde{a} := a - \mathrm{E}_{P_Z}[a(Z)^\top m_{\bar{\theta}}(Z)]\, b_0$ is bounded, and since $\mathrm{E}_{P_Z}[b_0(Z)^\top m_{\bar{\theta}}(Z)] = 1$ it satisfies $\mathrm{E}_{P_Z}[\tilde{a}(Z)^\top m_{\bar{\theta}}(Z)] = 0$, so $\tilde{a} \in \mathcal{T}(\bar{\theta})$. Hence

$$\begin{aligned} 0 = \mathrm{E}_{P_Z}[\tilde{a}(Z)^\top f(Z)] &= \mathrm{E}_{P_Z}[a(Z)^\top f(Z)] - c\;\mathrm{E}_{P_Z}[a(Z)^\top m_{\bar{\theta}}(Z)] \\ &= \mathrm{E}_{P_Z}\left[a(Z)^\top \{f(Z) - cm_{\bar{\theta}}(Z)\}\right], \end{aligned}$$

which since this should hold for all bounded a implies that $f = cm_{\bar{\theta}}$ P_Z-almost surely. This proves the first claim, and the claim for condition (v) follows by taking $f = m_\theta$. For condition (iv), if $\Gamma_{\bar{\theta}} v = cm_{\bar{\theta}}$ for some nonzero v, then $\mathrm{E}_{P_Z}[A(Z)\Gamma_{\bar{\theta}}(Z)]v = c\,\mathrm{E}_{P_Z}[A(Z)m_{\bar{\theta}}(Z)] = 0$ for every A with rows in $\mathcal{T}(\bar{\theta})$, so condition (iv) fails. Conversely, if for all nonzero v, $\Gamma_{\bar{\theta}} v$ is not a scalar multiple of $m_{\bar{\theta}}$, then consider the linear map $T : \mathcal{T}(\bar{\theta}) \to \mathbb{R}^K$ given by $Ta := \mathrm{E}_{P_Z}[\Gamma_{\bar{\theta}}(Z)^\top a(Z)]$. If T were not onto, some nonzero v would be orthogonal to its range, so that $\mathrm{E}_{P_Z}[a(Z)^\top \Gamma_{\bar{\theta}}(Z)v] = 0$ for every $a \in \mathcal{T}(\bar{\theta})$, and the first claim would imply that $\Gamma_{\bar{\theta}} v$ is a scalar multiple of $m_{\bar{\theta}}$. Hence there exist $a_1, \ldots, a_K \in \mathcal{T}(\bar{\theta})$ with $Ta_k = e_k$, and the matrix A with rows $a_k^\top$ satisfies $\mathrm{E}_{P_Z}[A(Z)\Gamma_{\bar{\theta}}(Z)] = I_K$. □

Proposition S4 *Under Assumption S1, there exist $D < \infty$ and a matrix $A : \mathcal{Z} \to \mathbb{R}^{D \times d}$ of bounded instruments such that $\{\theta \in \Theta : \mathrm{E}_P[A(Z)\rho(X, \theta)] = 0\} = \{\bar{\theta}\}$. The Jacobian of the resulting moments at $\bar{\theta}$ has rank K.*

Proof. By Assumption S1(iv), there exists A with rows in $\mathcal{T}(\bar{\theta})$ and $\mathrm{E}_{P_Z}[A(Z)\Gamma_{\bar{\theta}}(Z)]$ nonsingular. Let $A_0 := \mathrm{E}_{P_Z}[A(Z)\Gamma_{\bar{\theta}}(Z)]^{-1} A$, and observe that $\mathrm{E}_{P_Z}[A_0(Z)m_{\bar{\theta}}(Z)] = 0$ and $\mathrm{E}_{P_Z}[A_0(Z)\Gamma_{\bar{\theta}}(Z)] = I_K$. Since A_0 is bounded, Assumption S1(iii) implies that there exists $\varepsilon > 0$ such that $\bar{\theta}$ is the only solution with $\|\theta - \bar{\theta}\| < \varepsilon$. The set $F := \{\theta \in \Theta :$

$\|\theta - \bar{\theta}\| \geq \varepsilon\}$ is compact, while for each $\eta \in F$, Assumption S1(v) yields $a_\eta \in \mathcal{T}(\bar{\theta})$ with $\mathrm{E}_{P_Z}[a_\eta(Z)^\top m_\eta(Z)] \neq 0$. Since a_η is bounded, Assumption S1(ii) implies that this inequality persists on an open neighborhood of η. Extracting a finite subcover of F and appending the associated instruments to A_0 yields bounded instruments for which $\bar{\theta}$ is the unique solution, and whose Jacobian at $\bar{\theta}$ has rank K since its first K rows are I_K. □

Proposition S4 is the conditional moment equality analog of Proposition 1, and shows that when the conditional moment equalities are misspecified, there exists a choice of instruments which makes any given target $\bar{\theta}$ the unique solution to the moment conditions.

S5.2 Efficient Moments for a Given Target

We next consider a conditional moment equality analog of Proposition 2, examining the minimum asymptotic variance attainable while targeting a given estimand. Since, unlike in the setting in the main text, the model does not supply a full distribution for the data, we compute asymptotic variances under P. Let

$$S_{\bar{\theta}}(z) := \mathrm{E}_P[\rho(X, \bar{\theta})\rho(X, \bar{\theta})^\top \mid Z = z]$$

denote the conditional second moment of the residual at $\bar{\theta}$, which equals its conditional variance when the conditional moment equalities hold at $\bar{\theta}$. For a choice of $A : \mathcal{Z} \to \mathbb{R}^{D \times d}$ satisfying (S3) and a positive-definite weighting matrix $W \in \mathbb{R}^{D \times D}$, define

$$G_{\bar{\theta}}(A) := \mathrm{E}_{P_Z}[A(Z)\Gamma_{\bar{\theta}}(Z)], \qquad \Omega_{\bar{\theta}}(A) := \mathrm{E}_{P_Z}[A(Z)S_{\bar{\theta}}(Z)A(Z)^\top],$$
$$V_{\bar{\theta}}(A, W) := (G^\top W G)^{-1} G^\top W \Omega W G (G^\top W G)^{-1} \quad \text{for} \quad G = G_{\bar{\theta}}(A),\ \Omega = \Omega_{\bar{\theta}}(A),$$

when $G^\top W G$ is nonsingular, and take $V_{\bar{\theta}}(A, W)$ to be infinite otherwise. Since $A(Z)\rho(X, \bar{\theta})$ has mean zero under P, $\Omega_{\bar{\theta}}(A)$ is their variance, and under standard regularity conditions $V_{\bar{\theta}}(A, W)$ is the asymptotic variance under P of the estimator based on instruments A and weighting matrix W. As a benchmark, let

$$\mathcal{I}_{\bar{\theta}} := \mathrm{E}_{P_Z}[\Gamma_{\bar{\theta}}(Z)^\top S_{\bar{\theta}}(Z)^{-1}\Gamma_{\bar{\theta}}(Z)].$$

When the conditional moment equalities hold at $\bar{\theta}$, $\mathcal{I}_{\bar{\theta}}^{-1}$ is the semiparametric efficiency bound and is attained by the optimal instruments $\Gamma_{\bar{\theta}}(Z)^\top S_{\bar{\theta}}(Z)^{-1}$. Under misspecification, by contrast, $\mathcal{I}_{\bar{\theta}}^{-1}$ will play a role similar to that played by the model-implied variance of max-

imum likelihood in Section 3, and will lower-bound the attainable variance over estimators targeting $\bar{\theta}$.

Assumption S2 *Suppose that* $\mathrm{E}_{P_Z}[\operatorname{tr} S_{\bar{\theta}}(Z)] < \infty$, $S_{\bar{\theta}}(z)$ *is positive definite for* P_Z*-almost every* z*, and* $\mathcal{I}_{\bar{\theta}}$ *is finite.*

To characterize the optimal instruments, we work in the Hilbert space $\mathcal{H}_{\bar{\theta}}$ consisting of instruments $a : \mathcal{Z} \to \mathbb{R}^d$ such that $\mathrm{E}_{P_Z}[a(Z)^\top S_{\bar{\theta}}(Z)a(Z)] < \infty$, with inner product $\langle a, c\rangle_{\bar{\theta}} := \mathrm{E}_{P_Z}[a(Z)^\top S_{\bar{\theta}}(Z)c(Z)]$. Let $\mathcal{H}^0_{\bar{\theta}} := \{a \in \mathcal{H}_{\bar{\theta}} : \mathrm{E}_{P_Z}[a(Z)^\top m_{\bar{\theta}}(Z)] = 0\}$ denote the subset of $\mathcal{H}_{\bar{\theta}}$ that target $\bar{\theta}$. The following lemma establishes the properties of these spaces that we use to characterize the variance bound.

Lemma S9 *Under Assumption S1(iii) and Assumption S2,*

$$0 < \mathrm{E}_{P_Z}[m_{\bar{\theta}}(Z)^\top S_{\bar{\theta}}(Z)^{-1} m_{\bar{\theta}}(Z)] \leq 1,$$

$\mathcal{H}^0_{\bar{\theta}}$ *is a closed subspace of* $\mathcal{H}_{\bar{\theta}}$*, and* $\mathcal{T}(\bar{\theta})$ *is dense in* $\mathcal{H}^0_{\bar{\theta}}$*.*

Proof. By Jensen's inequality, $S_{\bar{\theta}}(z) \succeq m_{\bar{\theta}}(z) m_{\bar{\theta}}(z)^\top$, so $m_{\bar{\theta}}(z)^\top S_{\bar{\theta}}(z)^{-1} m_{\bar{\theta}}(z) \leq 1$, while the expectation of this quantity is strictly positive since $m_{\bar{\theta}} \neq 0$ and $S_{\bar{\theta}}^{-1}$ is positive definite almost surely. For $a \in \mathcal{H}_{\bar{\theta}}$, Cauchy-Schwarz thus gives $|\mathrm{E}_{P_Z}[a(Z)^\top m_{\bar{\theta}}(Z)]| \leq \|a\|_{\bar{\theta}}$, so the $a \mapsto \mathrm{E}_{P_Z}[a(Z)^\top m_{\bar{\theta}}(Z)]$ functional is continuous and $\mathcal{H}^0_{\bar{\theta}}$ is closed. Since $\mathrm{E}_{P_Z}[\operatorname{tr} S_{\bar{\theta}}(Z)] < \infty$, bounded instruments belong to $\mathcal{H}_{\bar{\theta}}$ and are dense in it: in particular, the truncations $a_l = a \min\{1, l/\|a\|\}$ satisfy $\|a_l - a\|^2_{\bar{\theta}} \leq \mathrm{E}_{P_Z}[a(Z)^\top S_{\bar{\theta}}(Z) a(Z)\, \mathbb{I}\{\|a(Z)\| > l\}] \to 0$. For $a \in \mathcal{H}^0_{\bar{\theta}}$ and such a_l, let b_0 be bounded with $\mathrm{E}_{P_Z}[b_0(Z)^\top m_{\bar{\theta}}(Z)] = 1$, as in the proof of Lemma S8. Then $\tilde{a}_l := a_l - \mathrm{E}_{P_Z}[a_l(Z)^\top m_{\bar{\theta}}(Z)] b_0 \in \mathcal{T}(\bar{\theta})$, and $\mathrm{E}_{P_Z}[a_l(Z)^\top m_{\bar{\theta}}(Z)] \to 0$ by continuity, so $\tilde{a}_l \to a$ in $\mathcal{H}_{\bar{\theta}}$. □

As the proof shows, the functional $a \mapsto \mathrm{E}_{P_Z}[a(Z)^\top m_{\bar{\theta}}(Z)]$ is continuous in the norm $\|\cdot\|_{\bar{\theta}}$. By contrast, in the parametric case the functional $h \mapsto \mathrm{E}_P[h(X)] - \mathrm{E}_{\bar{\theta}}[h(X)]$ is continuous in the model-implied standard deviation only when $\chi^2(P\|P_{\bar{\theta}}) < \infty$ (Lemma S4). Consequently, there is no analog here of the infinite-divergence case, in which discontinuity of the matching functional makes targeting costless.

The usual optimal instruments $\Gamma_{\bar{\theta}}(Z)^\top S_{\bar{\theta}}(Z)^{-1}$ satisfy (S3) if and only if

$$\mathrm{E}_{P_Z}[\Gamma_{\bar{\theta}}(Z)^\top S_{\bar{\theta}}(Z)^{-1} m_{\bar{\theta}}(Z)] = 0,$$

that is, if and only if the columns of $\Gamma_{\bar\theta}$ are orthogonal to $m_{\bar\theta}$ once we weight by $S_{\bar\theta}^{-1}$. To obtain targeting instruments more generally, we project $m_{\bar\theta}$ out of $\Gamma_{\bar\theta}$. Let

$$\Gamma_{\bar\theta}^{\perp}(z) := \Gamma_{\bar\theta}(z) - m_{\bar\theta}(z)\beta_{\bar\theta}^{\top} \quad \text{for} \quad \beta_{\bar\theta} := \frac{\mathrm{E}_{P_Z}[\Gamma_{\bar\theta}(Z)^{\top} S_{\bar\theta}(Z)^{-1} m_{\bar\theta}(Z)]}{\mathrm{E}_{P_Z}[m_{\bar\theta}(Z)^{\top} S_{\bar\theta}(Z)^{-1} m_{\bar\theta}(Z)]}$$

denote the residual from the population weighted least squares projection of the columns of $\Gamma_{\bar\theta}$ on $m_{\bar\theta}$, with weight $S_{\bar\theta}^{-1}$, so that $\mathrm{E}_{P_Z}[\Gamma_{\bar\theta}^{\perp}(Z)^{\top} S_{\bar\theta}(Z)^{-1} m_{\bar\theta}(Z)] = 0$ by construction. Let

$$\mathcal{J}_{\bar\theta} := \mathrm{E}_{P_Z}[\Gamma_{\bar\theta}^{\perp}(Z)^{\top} S_{\bar\theta}(Z)^{-1} \Gamma_{\bar\theta}^{\perp}(Z)]$$

denote the analog of $\mathcal{I}_{\bar\theta}$ which replaces $\Gamma_{\bar\theta}$ by $\Gamma_{\bar\theta}^{\perp}$. The next result shows that $\mathcal{J}_{\bar\theta}^{-1}$ is the smallest variance attainable when targeting $\bar\theta$, and that this bound can be approached.

Proposition S5 *Under Assumption S1(iii)-(iv) and Assumption S2, $\mathcal{J}_{\bar\theta}$ is positive definite, and:*

(a) $V_{\bar\theta}(A, W) \succeq \mathcal{J}_{\bar\theta}^{-1}$ for every A that satisfies (S3) *and has finite $\Omega_{\bar\theta}(A)$, and every positive definite W. Moreover, for every $\varepsilon > 0$ there exist bounded, just-identifying instruments A satisfying* (S3) *with $V_{\bar\theta}(A) \preceq (1+\varepsilon)\mathcal{J}_{\bar\theta}^{-1}$.*

(b) The bound in (a) is attained by the just-identifying instruments $A_{\bar\theta}^{*}(z) := \Gamma_{\bar\theta}^{\perp}(z)^{\top} S_{\bar\theta}(z)^{-1}$.

(c) If further Assumption S1(i), (ii), and (v) hold, then for every $\varepsilon > 0$ there exist bounded (not necessarily just-identifying) instruments A satisfying (S3) *and a positive definite W such that $\{\theta \in \Theta : \mathrm{E}_P[A(Z)\rho(X,\theta)] = 0\} = \{\bar\theta\}$ and $V_{\bar\theta}(A, W) \preceq (1+\varepsilon)\mathcal{J}_{\bar\theta}^{-1}$.*

Proof. For $a \in \mathcal{H}_{\bar\theta}$ and $v \in \mathbb{R}^K$, Cauchy-Schwarz implies that $|\mathrm{E}_{P_Z}[a(Z)^{\top}\Gamma_{\bar\theta}(Z)v]| \leq \|a\|_{\bar\theta}(v^{\top}\mathcal{I}_{\bar\theta}v)^{1/2}$, while $|\mathrm{E}_{P_Z}[a(Z)^{\top} m_{\bar\theta}(Z)]| \leq \|a\|_{\bar\theta}$ by the proof of Lemma S9. Hence, $G_{\bar\theta}(A)$ and $\mathrm{E}_{P_Z}[A(Z)m_{\bar\theta}(Z)]$ are well-defined for all A with finite $\Omega_{\bar\theta}(A)$. Since $\mathcal{I}_{\bar\theta}$ is finite and $\mathrm{E}_{P_Z}[m_{\bar\theta}(Z)^{\top} S_{\bar\theta}(Z)^{-1} m_{\bar\theta}(Z)] \in (0, 1]$ by Lemma S9, the same argument implies that $\beta_{\bar\theta}$ and $\mathcal{J}_{\bar\theta}$ are well-defined and finite. For nonzero $v \in \mathbb{R}^K$, Assumption S1(iv) and Lemma S8 imply that $\Gamma_{\bar\theta}v$ is not a scalar multiple of $m_{\bar\theta}$, so $\Gamma_{\bar\theta}^{\perp}v = \Gamma_{\bar\theta}v - (\beta_{\bar\theta}^{\top}v)m_{\bar\theta}$ is nonzero with positive probability under P_Z. Since $S_{\bar\theta}(z)$ is positive definite for P_Z-almost every z, it follows that $v^{\top}\mathcal{J}_{\bar\theta}v = \mathrm{E}_{P_Z}[(\Gamma_{\bar\theta}^{\perp}(Z)v)^{\top} S_{\bar\theta}(Z)^{-1}\Gamma_{\bar\theta}^{\perp}(Z)v] > 0$, so $\mathcal{J}_{\bar\theta}$ is positive definite.

By the definition of $\beta_{\bar\theta}$, $\mathrm{E}_{P_Z}[A_{\bar\theta}^{*}(Z)m_{\bar\theta}(Z)] = \mathrm{E}_{P_Z}[\Gamma_{\bar\theta}^{\perp}(Z)^{\top} S_{\bar\theta}(Z)^{-1} m_{\bar\theta}(Z)] = 0$, so $A_{\bar\theta}^{*}$ satisfies (S3). Since $\Gamma_{\bar\theta} = \Gamma_{\bar\theta}^{\perp} + m_{\bar\theta}\beta_{\bar\theta}^{\top}$, this further implies that

$$\mathrm{E}_{P_Z}[A_{\bar\theta}^{*}(Z)\Gamma_{\bar\theta}(Z)] = \mathrm{E}_{P_Z}[A_{\bar\theta}^{*}(Z)\Gamma_{\bar\theta}^{\perp}(Z)] = \mathcal{J}_{\bar\theta} = \mathrm{E}_{P_Z}[A_{\bar\theta}^{*}(Z)S_{\bar\theta}(Z)A_{\bar\theta}^{*}(Z)^{\top}],$$

which gives $V_{\bar\theta}(A^*_{\bar\theta}) = \mathcal{J}_{\bar\theta}^{-1}$, and so establishes part (b) given the lower bound in part (a).

For part (a), write $G = G_{\bar\theta}(A)$. If $G^\top W G$ is singular the bound holds by convention. Otherwise, let $\tilde A := (G^\top W G)^{-1} G^\top W A$. $\tilde A$ is just-identifying with rows in $\mathcal{H}_{\bar\theta}$, $\mathrm{E}_{P_Z}[\tilde A(Z)\Gamma_{\bar\theta}(Z)] = I_K$, $\mathrm{E}_{P_Z}[\tilde A(Z) m_{\bar\theta}(Z)] = 0$, and $V_{\bar\theta}(A, W) = \mathrm{E}_{P_Z}[\tilde A(Z) S_{\bar\theta}(Z) \tilde A(Z)^\top]$. For $U := \tilde A - \mathcal{J}_{\bar\theta}^{-1} A^*_{\bar\theta}$, we have $\mathrm{E}_{P_Z}[U(Z)\Gamma_{\bar\theta}(Z)] = 0$ and $\mathrm{E}_{P_Z}[U(Z) m_{\bar\theta}(Z)] = 0$, and since $S_{\bar\theta} A^{*\top}_{\bar\theta} = \Gamma^{\perp}_{\bar\theta} = \Gamma_{\bar\theta} - m_{\bar\theta}\beta_{\bar\theta}^\top$,

$$\mathrm{E}_{P_Z}[U(Z) S_{\bar\theta}(Z) A^*_{\bar\theta}(Z)^\top] = \mathrm{E}_{P_Z}[U(Z)\Gamma_{\bar\theta}(Z)] - \mathrm{E}_{P_Z}[U(Z) m_{\bar\theta}(Z)]\beta_{\bar\theta}^\top = 0.$$

Hence $V_{\bar\theta}(A, W) = \mathcal{J}_{\bar\theta}^{-1} + \mathrm{E}_{P_Z}[U(Z) S_{\bar\theta}(Z) U(Z)^\top] \succeq \mathcal{J}_{\bar\theta}^{-1}$. For the bounded approximation, Lemma S9 applied row by row gives bounded A_l with rows in $\mathcal{T}(\bar\theta)$ and $A_l \to A^*_{\bar\theta}$ in $\mathcal{H}_{\bar\theta}$, and the Cauchy-Schwarz inequality together with $\mathrm{E}_{P_Z}[\Gamma_{\bar\theta}(Z)^\top S_{\bar\theta}(Z)^{-1}\Gamma_{\bar\theta}(Z)] < \infty$ implies $\mathrm{E}_{P_Z}[A_l(Z)\Gamma_{\bar\theta}(Z)] \to \mathcal{J}_{\bar\theta}$ and $\mathrm{E}_{P_Z}[A_l(Z) S_{\bar\theta}(Z) A_l(Z)^\top] \to \mathcal{J}_{\bar\theta}$. Hence $V_{\bar\theta}(A_l) \to \mathcal{J}_{\bar\theta}^{-1}$, so since $\mathcal{J}_{\bar\theta}^{-1}$ is positive definite, $V_{\bar\theta}(A_l) \preceq (1+\varepsilon)\mathcal{J}_{\bar\theta}^{-1}$ for l sufficiently large.

For part (c), fix $\varepsilon > 0$ and take A_l from the sequence constructed in the proof of part (a), selecting l sufficiently large that $G_{\bar\theta}(A_l)$ is nonsingular and $V_{\bar\theta}(A_l) \preceq (1 + \varepsilon/2)\mathcal{J}_{\bar\theta}^{-1}$. Let A° denote the bounded instruments from Proposition S4, and $\check A := (A_l^\top, A^{\circ\top})^\top$. Since $\bar\theta$ is the unique solution to $\mathrm{E}_P[A^\circ(Z)\rho(X,\theta)] = 0$ and A_l satisfies (S3), $\bar\theta$ is also the unique solution to $\mathrm{E}_P[\check A(Z)\rho(X,\theta)] = 0$, while $\Omega_{\bar\theta}(\check A)$ is finite since $\check A$ is bounded and $\mathrm{E}_{P_Z}[\operatorname{tr} S_{\bar\theta}(Z)] < \infty$. For $\eta > 0$ define $W_\eta := \operatorname{diag}(I_K, \eta I)$. As $\eta \downarrow 0$, $G_{\bar\theta}(\check A)^\top W_\eta G_{\bar\theta}(\check A) \to G_{\bar\theta}(A_l)^\top G_{\bar\theta}(A_l)$, which is nonsingular, and $W_\eta \Omega_{\bar\theta}(\check A) W_\eta \to \operatorname{diag}(\Omega_{\bar\theta}(A_l), 0)$, so $V_{\bar\theta}(\check A, W_\eta) \to G_{\bar\theta}(A_l)^{-1}\Omega_{\bar\theta}(A_l) G_{\bar\theta}(A_l)^{-\top} = V_{\bar\theta}(A_l)$. Since $\mathcal{J}_{\bar\theta}^{-1}$ is positive definite, $V_{\bar\theta}(\check A, W_\eta) \preceq (1+\varepsilon)\mathcal{J}_{\bar\theta}^{-1}$ for η sufficiently small. □