NBER WORKING PAPER SERIES

THE INCREDIBLE FLEXIBILITY OF MOMENT MATCHING

Isaiah Andrews
Bas Sanders

Working Paper 35378
http://www.nber.org/papers/w35378

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2026

We thank Pat Kline, Marc Melitz, Mikkel Plagborg-Møller, Dominic Russel and Jesse Shapiro for helpful comments, and Claude and GPT for outstanding research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Incredible Flexibility of Moment Matching
Isaiah Andrews and Bas Sanders
NBER Working Paper No. 35378
June 2026
JEL No. C13, C18, C51

ABSTRACT

We ask how far the choice of which moments to match can push estimates in misspecified structural models. The answer is: very far. Under regularity conditions, an adversarial researcher informed about the data distribution can choose moments that render any parameter value the unique solution to the population moment-matching problem. Moreover, in many cases they can do so with little increase in model-implied standard errors relative to maximum likelihood. We illustrate both results in a menu-cost model.

Isaiah Andrews
Massachusetts Institute of Technology
Department of Economics
and NBER
iandrews@mit.edu

Bas Sanders
Harvard University
bas_sanders@outlook.com

The Incredible Flexibility of Moment Matching

Isaiah Andrews and Bas Sanders*

Abstract

We ask how far the choice of which moments to match can push estimates in misspecified structural models. The answer is: very far. Under regularity conditions, an adversarial researcher informed about the data distribution can choose moments that render any parameter value the unique solution to the population moment-matching problem. Moreover, in many cases they can do so with little increase in model-implied standard errors relative to maximum likelihood. We illustrate both results 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. Many researchers estimate their models by matching a set of hand-selected moments rather than using more traditional statistical approaches like maximum likelihood, and it is common to cite misspecification concern as a reason.[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.

*This version: June 17, 2026. Andrews: MIT Department of Economics and NBER; iandrews@mit.edu. Sanders: SEO Amsterdam Economics; b.sanders@seo.nl. We thank Pat Kline, Marc Melitz, Mikkel Plagborg-Møller, Dominic Russel and Jesse Shapiro for helpful comments, and Claude and GPT for outstanding research assistance.

[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."

At the same time, the flexibility to choose moments introduces additional 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 study how much scope there is for such manipulation by asking what results an adversarial researcher could engineer through their choice of moments. Our main finding is an "anything-goes" result for moment-matching estimation under misspecification: under regularity 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 turns out not to be the case. Specifically, we characterize the choice of moments which minimizes model-implied asymptotic variance subject to delivering a given estimand, and show that in many cases the resulting standard errors are close to those from maximum likelihood, which in turn lower-bound the model-implied standard errors for any asymptotically normal and unbiased estimator. Moreover, we show that matters are even worse if we consider a researcher who may choose their moments after observing the data.[2]

To illustrate our results, following Cocci and Plagborg-Møller (2024) we estimate the model of Alvarez and Lippi (2014), who study menu-cost price setting in multiproduct firms. Fitting the Alvarez and Lippi (2014) model to scanner data used by Cocci and Plagborg-Møller (2024), 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, in many cases with standard errors little larger than those implied by maximum likelihood.

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 additional 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 dan-

[2]Specifically, as we discuss below a researcher who chooses the moments after observing the sample can often engineer any estimate value they wish, along with model-implied standard errors that exactly match maximum likelihood.

gers 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 (2024) 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. Our paper complements this literature by highlighting a different dimension of the moment selection problem: the scope for moment choice to alter the resulting estimands.

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 others 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 rather than any other. 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.

The rest of the paper is organized as follows. Section 2 illustrates the empirical consequences of moment choice in an example building on Alvarez and Lippi (2014) and Cocci and Plagborg-Møller (2024). Section 3 introduces our formal setting and characterizes the class of moments that yield a specific moment-matching estimand. Section 4 finds an efficient moment function in this class. Section 5 discusses routes to mitigate the issues suggested by our theoretical results. Additional details on the empirical example, together with all proofs, are collected in the Online Appendix.

2 Illustration: Moment Choice in a Menu-Cost Model

To illustrate the empirical consequences of moment choice before turning to the theory, we follow Cocci and Plagborg-Møller (2024) and estimate Alvarez and Lippi (2014)'s model of menu-cost price setting in multiproduct firms. To obtain a simplified, two-parameter version of the model in which estimates can easily be plotted, we focus on the cross-sectional distribution of non-zero absolute centered log price changes, $X := |\Delta p - \overline{\Delta p}|$, conditional on $\Delta p \neq 0$.[3] Alvarez and Lippi (2014)'s model implies that the distribution of X depends on the model's structural primitives through two objects: the number of products N (n in the notation of Alvarez and Lippi 2014) and the threshold $\bar{y}$ (for the sum of squared price gaps across products within a firm) beyond which the optimal policy resets prices.[4] We therefore parameterize the model by $\theta = (N, \bar{y})$, and study how the estimated value of θ varies with the choice of moments.

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 (2024). We compute the conservative standard errors of Cocci and Plagborg-Møller (2024) at the resulting estimate, and plot the estimate along with the marginal 95% confidence intervals for the two parameters in black in each of Figures 1-5.

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 1 reports point estimates for θ obtained from all pairs of power moments $\mathrm{E}[X^k]$ with $k \in \{1, \ldots, 8\}$. As the figure highlights, the resulting range of estimates extends well beyond the range of statistical uncertainty captured by the Cocci and Plagborg-Møller (2024) standard errors.[5] This becomes even more apparent as we broaden the range of moments considered. Figure 2 shows results when we also match quantiles of the price change distribution, considering quantiles in the grid $\{0.10, 0.20, \ldots, 0.90\}$. Specifically, we plot results for all power-power pairs, all power-quantile pairs, and all quantile-quantile pairs (restricting to pairs with $|q_1 - q_2| \geq 0.25$ to

[3]Following Cocci and Plagborg-Møller (2024), we use demeaned log changes since Alvarez and Lippi (2014) abstract from inflation. We take absolute values since the absolute log change suffices to compute all of the estimation moments considered in Cocci and Plagborg-Møller (2024).

[4]In the original three-parameter formulation of Alvarez and Lippi (2014), the parameters are 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 $(N, \bar{y})$. Since the model-implied data distribution is well-defined for non-integer N, we follow Cocci and Plagborg-Møller (2024) and allow $N \in \mathbb{R}_+$.

[5]This reflects no flaw in the Cocci and Plagborg-Møller (2024) standard errors, which are designed to upper bound statistical uncertainty given a set of moments, not variability induced by the choice of moments.

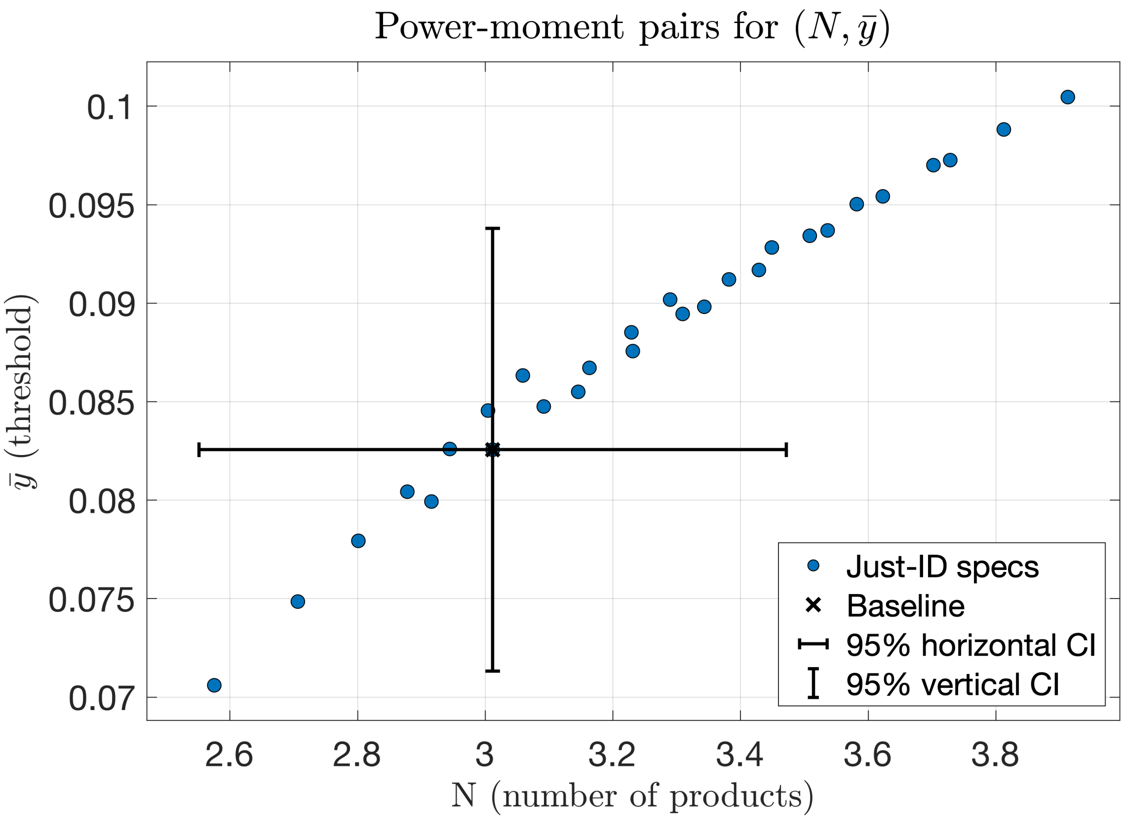

Figure 1: Point estimates for θ, considering powers of the distribution of non-zero absolute centered log price changes. The horizontal and vertical CIs are conservative confidence intervals based on Cocci and Plagborg-Møller (2024).

exclude nearly redundant quantile pairs).

The resulting range of estimates extends far beyond that suggested by the (conservative) Cocci and Plagborg-Møller (2024) confidence intervals. Moreover, different estimates in this set point to substantively different economic interpretations, with the smallest estimate suggesting less than two products per firm, and the largest more than 12. Similarly, the lowest estimate suggests that firms adjust prices once the sum of squared log price deviations exceeds 0.025, while the largest suggests a threshold is ten times larger. The model also suggests a reason for the broadly upward-sloping relationship between the $\bar{y}$ and N estimates - to explain the distribution of observed price changes, it seems plausible the data might "want" either firms with few products, but 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. In fact, however, it shows only our lack of creativity thus far.

We next consider a researcher who, rather than limiting attention to powers and quantiles, exploits the full flexibility of moment matching. For a given target parameter value $\bar{\theta} =$

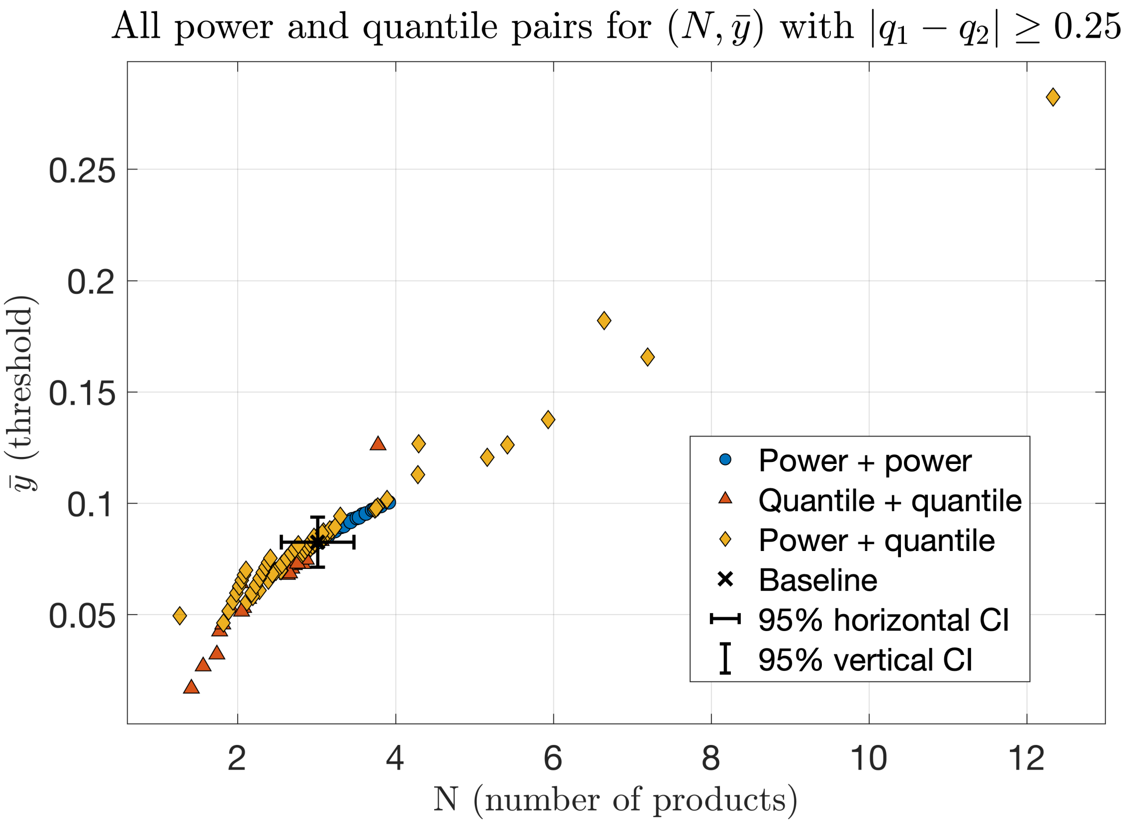

Figure 2: Point estimates for θ, considering powers and quantiles of the distribution of non-zero absolute centered log price changes. The horizontal and vertical CIs are conservative confidence intervals based on Cocci and Plagborg-Møller (2024).

$(N, \bar{y})$, our results in Sections 3 and 4 below 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.[6] Figure 3 applies a version of this construction to a grid of 12 target parameter values $\bar{\theta} = (N, \bar{y})$ with $N \in \{2, 3, 4, 5\}$ and $\bar{y} \in \{0.06, 0.12, 0.18\}$ and a distribution estimate based on the observed data. As expected given the theory, the resulting estimates track their targets up to sampling uncertainty, and show that the upward-sloping relationship in the power and quantile figures was specific to the moments considered there. Moreover, these targeted estimates come with reasonable-looking standard errors: across the 12 sets of moments, the standard errors for N (evaluated at the estimates $\hat{\theta}$) range from 0.2% to 39% larger than the (model-implied) MLE standard error evaluated at the same point, with a median of 9.7%, while the standard errors for $\bar{y}$ range from 3.4% to 214.2%

[6]For this exercise we censor the distribution of observed price changes to ensure that the support of X does not vary with θ, which would introduce even more researcher degrees of freedom. See Online Appendix S1 below for details.

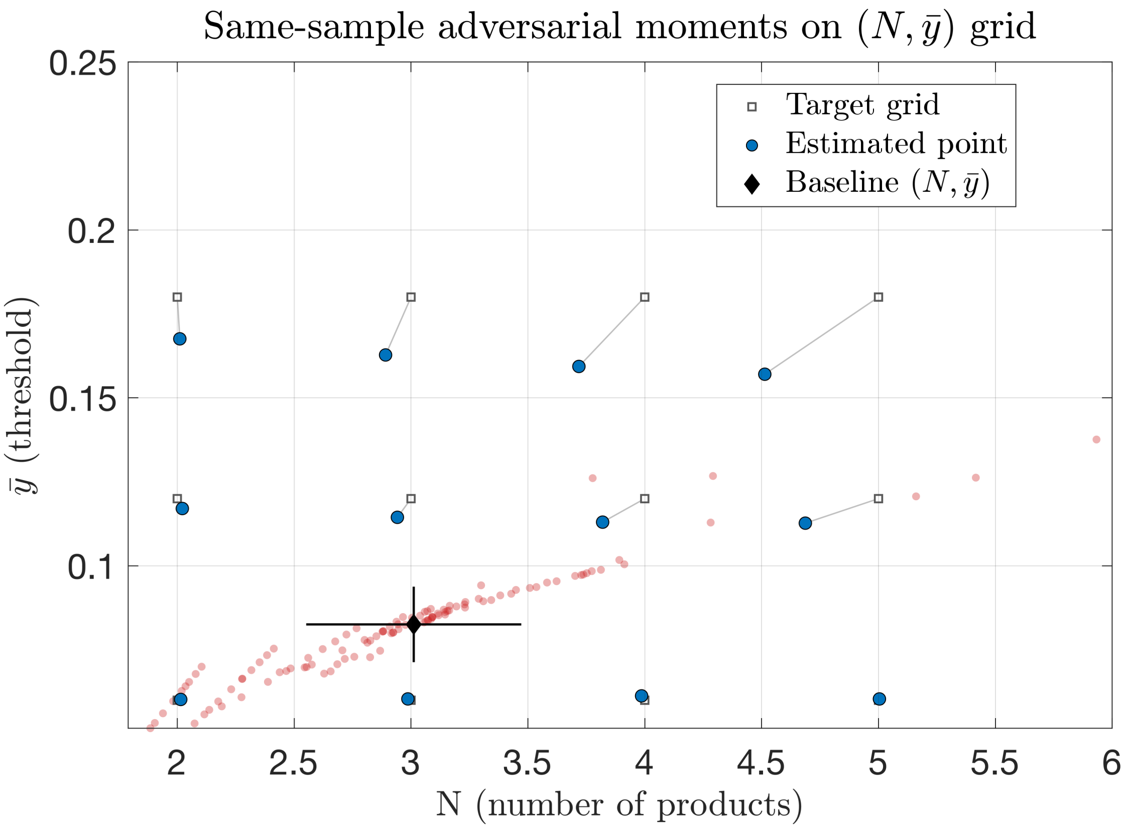

Figure 3: Preview of the adversarial construction in the Alvarez and Lippi (2014) application. Target grid (open squares) and point estimates (blue dots) for $\theta = (N, \bar{y})$ obtained from the target-specific adversarial moments. The gray lines connect each target to its estimate. The red scatter shows, for reference, the varying-moment estimates from Figure 2; the black diamond and bars are the baseline $(N, \bar{y})$ estimate and confidence intervals.

larger, with a median of 17.2%. A full tabulation appears in Table 1 in Online Appendix S1, which also provides additional details on the application and our implementation, along with plots of the adversarial moments.

Overall, these results highlight that even in a simple model, the choice of moments can greatly influence the results obtained without necessarily incurring a large standard error cost. We next show that this is not a peculiarity of this specific example, but is instead a general property of moment matching in misspecified models.

3 The Flexibility of Moment Matching

3.1 Setting

Consider a researcher who observes i.i.d. draws $X_1, ..., X_n \in \mathcal{X}$ from a probability distribution P. The researcher postulates a 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 a common dominating measure (e.g. Lebesgue measure for continuous data, or counting measure for discrete data). 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 $W_n \in \mathbb{R}^{D \times D}$.[7] 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

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

As 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

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

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

[7] Note that g does not depend on θ.

In the special case where the moments g admit "perfect fit", meaning that there exists at least one $\theta \in \Theta$ 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.[8] Maintaining this condition also ensures that the model's over-identifying restrictions hold so long as the researcher limits attention to the moments g.

3.2 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, social conventions on the form of "acceptable" moments, etc.) 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 the target parameter value (i.e. $g \in L^2\left(P_{\bar{\theta}}; \mathbb{R}^D\right)$), define

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

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 next lemma characterizes the function class $\mathcal{C}\left(\bar{\theta}\right)$. It uses the chi-squared divergence, which for distributions Q_1 and Q_2 with Q_1 absolutely continuous with respect to Q_2 (that

[8] 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 estimands for a fixed set of moments.

is, where probability zero events under Q_2 are also probability zero under Q_1), is defined as

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

When Q_1 is not absolutely continuous with respect to Q_2, the χ^2 divergence is infinite.

Lemma 1 (Characterizing $\mathcal{C}\left(\bar{\theta}\right)$)**.** *If* $\chi^2\left(P \| P_{\bar{\theta}}\right) \in (0,\infty)$, *then*

$$\mathcal{C}\left(\bar{\theta}\right) = \left\{ f - \frac{\mathrm{E}_{\bar{\theta}}\left[f\left(X\right)\left(\frac{dP}{dP_{\bar{\theta}}}\left(X\right)-1\right)\right]}{\chi^2\left(P\|P_{\bar{\theta}}\right)}\left(\frac{dP}{dP_{\bar{\theta}}}(X)-1\right) : f \in L^2\left(P_{\bar{\theta}};\mathbb{R}^D\right)\right\}.$$

Note that $\chi^2\left(P\|P_{\bar{\theta}}\right) \in (0,\infty)$ implies (i) $P \neq P_{\bar{\theta}}$, so the model-implied distribution with parameter $\bar{\theta}$ is not the true DGP, and (ii) P is absolutely continuous with respect to $P_{\bar{\theta}}$.

Intuitively, $\mathcal{C}\left(\bar{\theta}\right)$ is constructed by taking all square-integrable functions f and residualizing them against the recentered likelihood-ratio $\frac{dP}{dP_{\bar{\theta}}} - 1$. The recentered likelihood ratio has mean zero under $\bar{\theta}$ by construction, while the residualization corrects for the mean of f under P. If $f\left(x\right) = 1$, the construction yields $g\left(x\right) = 0$, which has no identifying power: every $\theta \in \Theta$ satisfies the resulting population moment equation. If $f\left(x\right) = x$, the resulting moment is

$$g\left(x\right) = x - \frac{\mathrm{E}_P\left[X\right] - \mathrm{E}_{\bar{\theta}}\left[X\right]}{\chi^2\left(P\|P_{\bar{\theta}}\right)}\left(\frac{dP}{dP_{\bar{\theta}}}\left(x\right)-1\right),$$

which corrects for the difference in the mean of X under P and $P_{\bar{\theta}}$.

As the example with $f(x) = 1$ illustrates, a given moment function in $\mathcal{C}\left(\bar{\theta}\right)$ might fail to pin down $\bar{\theta}$ uniquely. We next show that if the model is misspecified then under regularity conditions an adversarial researcher can eliminate such competing solutions by adding finitely many moments in $\mathcal{C}\left(\bar{\theta}\right)$. To state this result, let

$$L := \frac{dP}{dP_{\bar{\theta}}},\ L_\theta := \frac{dP_\theta}{dP_{\bar{\theta}}},$$

denote the likelihood ratio of the true distribution P relative to the target distribution $P_{\bar{\theta}}$, and of a generic model-implied distribution P_θ relative to $P_{\bar{\theta}}$, respectively. Furthermore, let $r := L - 1$ and $r_\theta := L_\theta - 1$ denote the corresponding recentered likelihood-ratios, and define $s_\theta\left(x\right) := \frac{\partial}{\partial\theta}\log p_\theta\left(x\right)$ as the score function (i.e. gradient of the log density) at θ.

Assumption 1 (Uniqueness of estimand)**.** *Suppose* $0 < \mathrm{E}_{P_{\bar{\theta}}}[r(X)^2] < \infty$, *and that* P_θ, $P_{\bar{\theta}}$ *are mutually absolutely continuous for all* $\theta \in \Theta$. *Further assume:*

(i) Θ *is compact.*

(ii) The map $\theta \mapsto r_\theta$ *is continuous from* Θ *into* $L^2(P_{\bar\theta})$.

(iii) The functions $\theta \mapsto \mathrm{E}_{P_\theta}[s_{\bar\theta}(X)]$ *and* $\theta \mapsto \mathrm{E}_{P_\theta}[r(X)]$ *are continuously differentiable in a neighborhood of* $\bar\theta$*, with derivative obtained by differentiating under the integral sign.*

(iv) The score $s_{\bar\theta} \in L^2(P_{\bar\theta}; \mathbb{R}^K)$ *satisfies rank* $\left(\mathrm{E}_{\bar\theta}\left[s_{\bar\theta}(X)s_{\bar\theta}(X)^\top\right]\right) = K$ *and*

$$\mathrm{E}_{P_{\bar\theta}}\left[\left(v^\top s_{\bar\theta}(X) - r(X)\right)^2\right] > 0 \qquad \textit{for every } v \neq 0.$$

(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. Condition (ii) is a continuity condition on the model-implied density, and is sufficient for $\theta' \to \theta$ to imply $\mathrm{E}_{\theta'}[f(X)] \to \mathrm{E}_\theta[f(X)]$ for all $f \in L^2(P_{\bar\theta})$. In a similar spirit, condition (iii) is a differentiability condition which says, for the specific moments $\mathrm{E}_\theta[s_{\bar\theta}] = \int s_{\bar\theta}(x)dP_\theta(x)$ and $\mathrm{E}_\theta[r] = \int r(x)dP_\theta(x)$, we can pass the derivative with respect to θ through the integral. Condition (iv) requires that the recentered likelihood ratio r not lie in the K-dimensional linear subspace spanned by the score, and that the Fisher information has full rank at $\bar\theta$. Loosely speaking, the span requirement means that there is not some direction in which we can move θ, local to $\bar\theta$, so that the change in r_θ is proportional to r. 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 (Uniqueness of estimand)**.** *Under Assumption 1 and* $0 < \chi^2(P \| P_{\bar\theta}) < \infty$*, there exists* $D < \infty$ *and* $g \in L^2(P_{\bar\theta}; \mathbb{R}^D)$ *such that* $\bar\theta$ *is the unique solution to the population moment-matching problem,* $\Theta_g = \{\bar\theta\}$.

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

Remark 1 (Absolute continuity). Our characterization assumes that P is known and absolutely continuous with respect to $P_{\bar\theta}$. If absolute continuity fails, manipulation is even easier. Specifically, if there exists a set A with $P(A) > 0$ and $P_{\bar{\theta}}(A) = 0$, an adversary can add functions such as $c \cdot \mathbb{I}\{X \in A\}$ to the moments, which shift $\mathrm{E}_P[g(X)]$ while leaving both $\mathrm{E}_{\bar{\theta}}[g(X)]$ and $\mathrm{Var}_{\bar{\theta}}(g(X))$ unchanged, and thus create "free" moment shifts which are impossible under absolute continuity.

Remark 2 (Sample-driven moment engineering). Similarly, in finite samples an adversary can often engineer any estimate they wish by choosing a sample-dependent g such that $\mathrm{E}_{\bar{\theta}}[g(X)] = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$. In particular, if the model implies the data are continuous then an adversarial researcher can add indicators for the observed data points to the moments without changing $\mathrm{E}_{\bar{\theta}}[g(X)]$ at all. Our focus on the population mapping $g \mapsto \Theta_g$ thus makes the manipulation problem harder than if we allowed moment choice based on the data.

4 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 moments that minimize asymptotic variance subject to targeting a given pseudo-true value. We find that an adversarial researcher can often engineer not only their desired estimand but also model-implied standard errors close to those of maximum likelihood.

4.1 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

$$\sqrt{n}\left(\hat{\theta}_{g,W} - \theta\right) \overset{P_\theta}{\rightsquigarrow} \mathcal{N}\left(0, \mathrm{AVar}\left(\hat{\theta}_{g,W}\right)\right), \tag{3}$$

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

$$\mathrm{AVar}\left(\hat{\theta}_{g,W}\right) := (G^\top W G)^{-1} G^\top W \Omega W G (G^\top W G)^{-1} \tag{4}$$

for

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

Model-implied standard errors are then based on plug-in estimates for Ω and G. We focus on model-implied standard errors for two reasons: first, it yields a tractable expression that depends only on the chosen moments g and the parametric model. Second, such standard errors 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).

4.2 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 the variance in equation (4) is the same as that obtained from the just-identifying moments $\tilde{g}(X) := (G^\top W G)^{-1} G^\top W g(X)$, 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 $\hat{\theta}_g$.

With this simplification, consider the problem of minimizing the asymptotic variance:

$$g^*_{\bar{\theta}} \in \underset{g \in \mathcal{C}(\bar{\theta})}{\arg\min}\ \mathrm{AVar}\left(\hat{\theta}_g\right).$$

Thanks to our focus on the model-implied variance this problem admits a simple solution.

Proposition 2 (Efficient moment function within $\mathcal{C}\left(\bar{\theta}\right)$)**.** *Suppose that $\chi^2\left(P \| P_{\bar{\theta}}\right) \in (0, \infty)$, and that $\mathrm{E}_{\bar{\theta}}[s_{\bar{\theta},\perp} s_{\bar{\theta},\perp}^\top]$ is positive-definite for*

$$s_{\bar{\theta},\perp} := s_{\bar{\theta}} - \frac{\mathrm{E}_P\left[s_{\bar{\theta}}(X)\right]}{\chi^2\left(P \| P_{\bar{\theta}}\right)} \left(\frac{dP}{dP_{\bar{\theta}}} - 1\right).$$

Then any efficient, just-identifying moment function subject to $g \in \mathcal{C}\left(\bar{\theta}\right)$ takes the form

$$g^*_{\bar{\theta}} := A s_{\bar{\theta},\perp} + c,$$

for some invertible matrix $A \in \mathbb{R}^{K\times K}$ and vector $c \in \mathbb{R}^K$. Moreover, the minimum asymptotic variance is

$$\begin{aligned}\mathrm{AVar}\left(\hat{\theta}_{g^*_{\bar{\theta}}}\right) &= \mathrm{E}_{\bar{\theta}}\left[s_{\bar{\theta},\perp}\left(X\right) s_{\bar{\theta},\perp}\left(X\right)^\top\right]^{-1} \\ &= \left\{\mathrm{E}_{\bar{\theta}}\left[s_{\bar{\theta}}\left(X\right) s_{\bar{\theta}}\left(X\right)^\top\right] - \frac{\mathrm{E}_P\left[s_{\bar{\theta}}\left(X\right)\right]\mathrm{E}_P\left[s_{\bar{\theta}}\left(X\right)\right]^\top}{\chi^2\left(P\|P_{\bar{\theta}}\right)}\right\}^{-1}.\end{aligned}$$

The efficient moments are (up to linear transformation) the score $s_{\bar{\theta}}$ purged of the component aligned with the likelihood-ratio residual r. While these moments need not uniquely pin down $\bar{\theta}$, we next show that provided one uses the efficient weighting matrix, augmenting with additional moments does not increase the asymptotic variance.

Lemma 2 (Augmentation does not affect asymptotic variance)**.** *Fix $\bar{\theta} \in \Theta$ with $\chi^2\left(P\|P_{\bar{\theta}}\right) \in (0,\infty)$ and let $g^*_{\bar{\theta}}$ be an efficient moment function as characterized in Proposition 2. Let $q : \mathcal{X} \to \mathbb{R}^J$ be a finite collection of moments in $\mathcal{C}\left(\bar{\theta}\right)$, and form the augmented moment vector $\tilde{g} := (g^{*\top}_{\bar{\theta}}, q^\top)^\top$. Under efficient weighting $W = \mathrm{Var}_{\bar{\theta}}(\tilde{g})^{-1}$, where we assume for simplicity $\mathrm{Var}_{\bar{\theta}}(\tilde{g})$ has full rank,*

$$\mathrm{AVar}\left(\hat{\theta}_{\tilde{g},W}\right) = \mathrm{AVar}\left(\hat{\theta}_{g^*_{\bar{\theta}}}\right).$$

Lemma 2 shows that Proposition 2's restriction to just-identified moments is purely for analytic convenience: given an efficient set of moments, one can always augment it with moments as in Proposition 1 and weight efficiently to construct an estimator which both delivers the unique estimand $\bar{\theta}$ and minimizes asymptotic variance.

4.3 Variance Cost of Estimand Targeting

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 drive a given conclusion? We close this section by showing that this is, unfortunately, not the case in general.

Under $P_{\bar{\theta}}$, canonical results in statistics tell us that maximum likelihood estimation is asymptotically efficient. Correspondingly, under $P_{\bar{\theta}}$ the score $s_{\bar{\theta}}$ delivers the smallest asymptotic variance among all moments, in the sense that

$$\mathrm{AVar}\left(\hat{\theta}_g\right) \geq \mathrm{AVar}\left(\hat{\theta}_{s_{\bar{\theta}}}\right)$$

for all moments g, where for square matrices A and B we say $A \geq B$ if and only if $A - B$ is positive semidefinite. We can thus use the model-implied variance of the MLE as a benchmark: if the engineered moments in Proposition 2 deliver a variance which is close to the MLE we know that there is little cost of moment engineering, in the sense that a researcher "organically" reaching the same estimate could not have meaningfully smaller model-implied standard errors.

To quantify the efficiency loss relative to the MLE, define the *variance inflation factor* $VIF\left(\bar{\theta}\right) \geq 1$ as the worst-case ratio of scalar asymptotic variances over all one-dimensional linear combinations of parameters,

$$VIF\left(\bar{\theta}\right) := \sup_{\alpha \neq 0} \frac{\alpha^\top \mathrm{AVar}\left(\hat{\theta}_{g^*_{\bar{\theta}}}\right)\alpha}{\alpha^\top \mathrm{AVar}\left(\hat{\theta}_{s_{\bar{\theta}}}\right)\alpha}.$$

The square root of this quantity, $\sqrt{VIF\left(\bar{\theta}\right)}$, 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.

The inflation factor $VIF\left(\bar{\theta}\right)$ has a close connection to measures of statistical distinguishability. To state this connection, we use the variational characterization of the chi-squared divergence, which states that we can also express the root chi-squared divergence $\sqrt{\chi^2\left(P \| P_{\bar{\theta}}\right)}$ as a "maximum population t-statistic" for distinguishing P and $P_{\bar{\theta}}$,

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

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.

This variational representation also suggests a class of generalized chi-squared divergences, where rather than searching over $L^2\left(P_{\bar{\theta}};\mathbb{R}\right)$ we consider more restricted function classes. Formally, for any linear class $\mathcal{H} \subseteq L^2\left(P_{\bar{\theta}}; \mathbb{R}\right)$, define

$$\sqrt{\chi^2_{\mathcal{H}}\left(P \| P_{\bar{\theta}}\right)} := \sup_{h \in \mathcal{H}, \mathrm{Var}_{\bar{\theta}}(h(X))>0} \frac{\left|\mathrm{E}_P\left[h\left(X\right)\right] - \mathrm{E}_{\bar{\theta}}\left[h\left(X\right)\right]\right|}{\sqrt{\mathrm{Var}_{\bar{\theta}}\left(h\left(X\right)\right)}}. \tag{6}$$

Like (5), (6) asks how large a mean shift, measured in standard deviation units, is induced by the change from $P_{\bar{\theta}}$ to P, now taking the upper bound over the restricted class of functions $\mathcal{H}$. Since we take the supremum over a smaller set, $\sqrt{\chi^2_{\mathcal{H}}\left(P \| P_{\bar{\theta}}\right)} \leq \sqrt{\chi^2\left(P \| P_{\bar{\theta}}\right)}$ by construction. The inflation factor $VIF(\bar{\theta})$ turns out to be precisely determined by the ratio of these two divergences with $\mathcal{H}$ the linear span of the score $s_{\bar{\theta}}$.

Proposition 3 (Discrepancy ratio)**.** *Suppose that* $0 < \chi^2(P \| P_{\bar{\theta}}) < \infty$, *and that*

$$I_{\bar{\theta}} := \mathrm{E}_{P_{\bar{\theta}}}[s_{\bar{\theta}}(X) s_{\bar{\theta}}(X)^\top], \quad Q_{\bar{\theta}} := \mathrm{E}_{P_{\bar{\theta}}}[s_{\bar{\theta},\perp}(X) s_{\bar{\theta},\perp}(X)^\top]$$

are 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 $\kappa\left(\bar{\theta}\right) \in [0,1]$ measures the fraction of the total chi-squared separation between P and $P_{\bar{\theta}}$ that is captured by linear combinations of the score. This reflects the ease with which $P_{\bar{\theta}}$ and P can be distinguished using linear combinations of the score, relative to using any square integrable function. Proposition 3 thus implies that the variance penalty is very large (so $\kappa\left(\bar{\theta}\right) \approx 1$ and hence $VIF\left(\bar{\theta}\right) \gg 1$) if and only if tests based on the mean of the score have nearly as much power to distinguish P from $P_{\bar{\theta}}$ as do tests based on any square-integrable functions. Intuitively, however, linear transformations of the score are a (very) small subset of the square-integrable functions, so we will have $\kappa\left(\bar{\theta}\right) \approx 1$ only if the misspecification happens to match what the likelihood "cares about" local to $\bar{\theta}$. Consequently, as suggested by the example in Section 2, VIF need not be large in applications.

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 lead researchers working with the same data and model to reach very different conclusions. In our view, the results above suggest some concrete steps to improve current practice.

Researchers should report (inter alia) estimates which are efficient under their model Estimators efficient under a given model, for instance maximum likelihood, do not in general remain efficient once the model is wrong. Nevertheless, since they are motivated solely by the model such 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 in settings where 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 maximum likelihood estimand, which minimizes the Kullback–Leibler divergence between the data and model-implied distributions (White, 1982), is often hard to interpret when the model is wrong. We agree, and for this reason do not view maximum likelihood 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.

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), all of whom derive optimal procedures under specific forms of misspecification concern. If widely adopted, this would shift the focus from the choice of estimator to the forms of misspecification that researchers wish to guard against.

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, we view this as complementary with the approaches above.

Another 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 (2024): "Standard errors for calibrated parameters," *Review of Economic Studies*, rdae099.

Cooley (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.

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.

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.

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

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, http://dx.doi.org/10.1177/0956797611417632rm 10.1177/0956797611417632.

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.

White, H. (1982): “Maximum likelihood estimation of misspecified models,” *Econometrica: Journal of the econometric society*, 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 provides additional details on the motivating application discussed in Section 2 of the draft, as well as all proofs. Throughout, notation, definitions, and numbered results refer to the main text.

S1 Further Details on Application

This appendix provides further details for the motivating example discussed in Section 2. In particular, we discuss how we apply our adversarial construction to this example and provide further information on the model-implied standard errors.

Since our theoretical results consider an adversary who knows the data distribution, our first step is to construct an estimate for P. This is complicated by the fact that the model-implied distribution of X has parameter-dependent support $[0, \sqrt{\bar{y}}]$. If we estimate P with support extending beyond this range, the chi-squared divergence will be infinite and as discussed in Remark 1, our adversarial construction will mechanically imply very large scope for manipulation.

To avoid this possibility, we 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 then construct an estimate $\hat{P}$ for the unknown law of X_c. Specifically,

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

for the continuous component of $\hat{P}$ on $[0,c]$ we use a logspline density estimator and combine it with the plug-in estimate of the mass at c^+.[2]

We then directly implement our adversarial construction using the estimated distribution $\hat{P}$, constructing the target-specific adversarial moment $g_{\bar{\theta},\hat{P}}$ and computing estimates $\hat{\theta}$ which solve

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

Since we use the estimated distribution $\hat{P}$, the resulting estimates (plotted e.g. in Figures 3 and 4) use the data twice, first to form $\hat{P}$ and then to construct $\hat{\theta}$.

To mitigate re-use of the data, we also compute cross-fit estimates. For this approach, reported in Figure 5, we split the sample into 10 disjoint folds, say $\{\mathcal{I}_1,\ldots,\mathcal{I}_{10}\}$ 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 the two figures).

In both Figures 4 and 5, we include ellipses around each estimate which reflect the model-implied confidence intervals for one-dimensional linear combinations $\alpha^{\top}\theta$, using model-implied standard errors.[3] 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.[4] As these ellipses highlight, the deviation of the engineered estimates

[2]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 knots. For $B(x)'\gamma$ the log-density under coefficients γ, we choose γ to minimize $-\sum_{i:X_i\le c} B(X_i)'\gamma + n_c \log\int_0^c \exp\{B(u)'\gamma\}\,du$ 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^+.

[3]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, our estimates are robust to a grid of 90 starting values extending beyond the range shown in the figure, suggesting that non-uniqueness is not an issue over the range of parameter values we study.

[4]Formally, we plot Wald confidence ellipses constructed from the sandwich variance $\mathrm{AVar}(\hat{\theta}_{g_{\bar{\theta},\hat{P}}})$ 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.

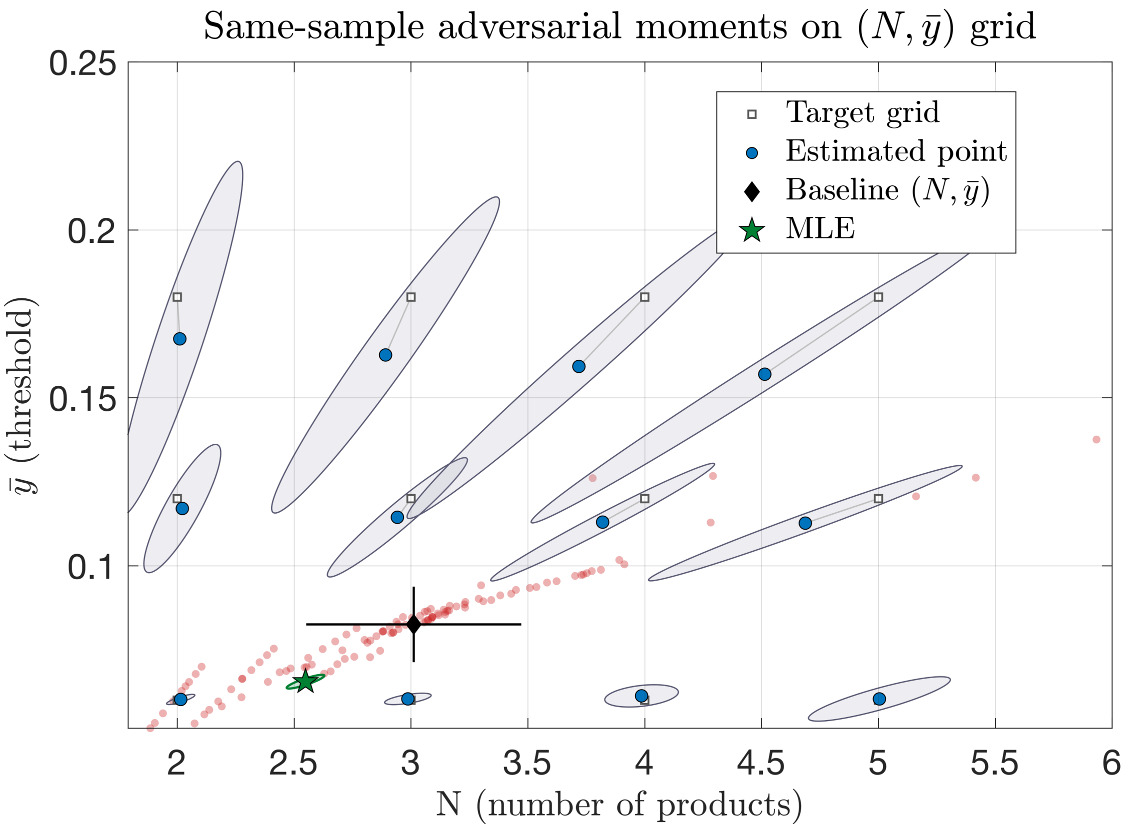

Figure 4: Same-sample adversarial estimates from Proposition 2 with model-implied uncertainty. Shaded regions are model-implied Wald confidence ellipses around each $\hat{\theta}$, using χ^2_1 critical values to ensure correct coverage for one-dimensional projections. The red scatter shows, for reference, 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 misspecification-robust confidence ellipse, again using a χ^2_1 critical value.

from their targets is on the same order as sampling variability. Moreover, for some targets the model-implied standard errors are similar to the misspecification-robust standard errors for the (unconstrained) maximum likelihood estimator, which we plot for comparison.

As further evidence, Table 1 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

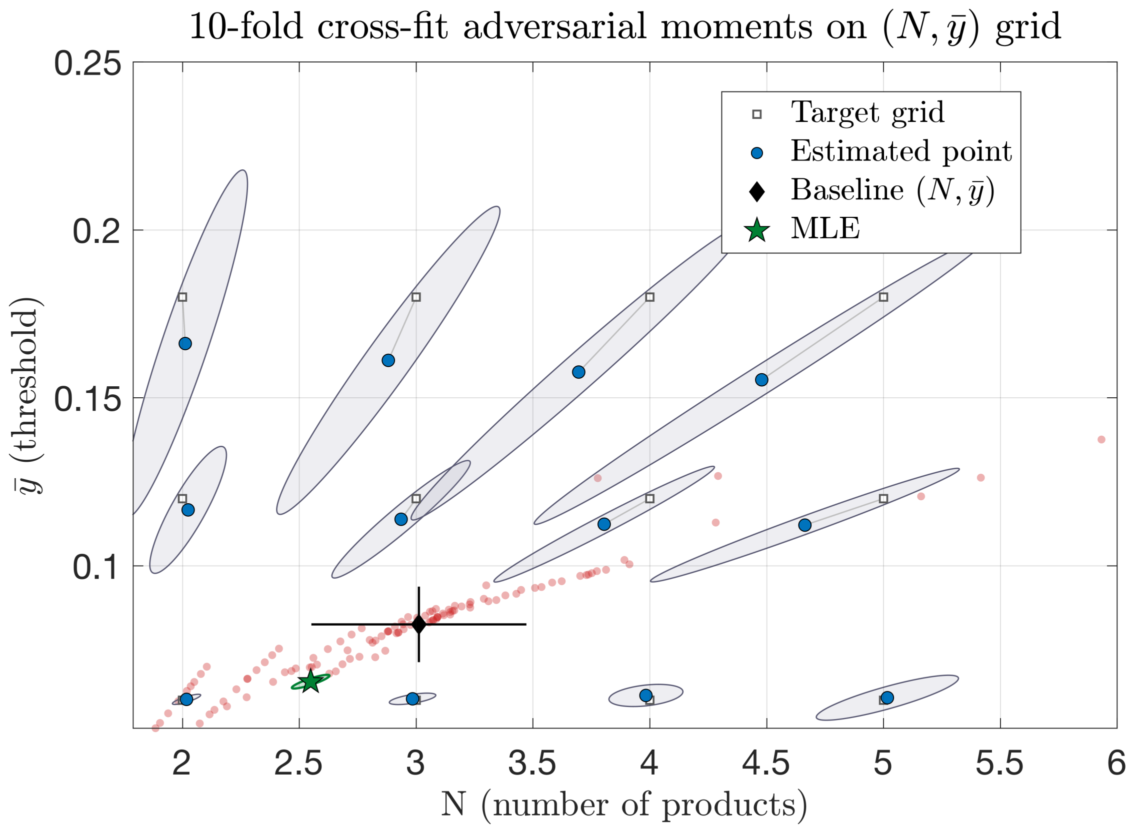

Figure 5: 10-fold cross-fit adversarial estimates from Proposition 2 with model-implied uncertainty. Shaded regions are model-implied Wald confidence ellipses around each $\hat{\theta}$, using χ^2_1 critical values to ensure correct coverage for one-dimensional projections. The red scatter shows, for 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 misspecification-robust confidence ellipse, again using a χ^2_1 critical value.

these quantities at the construction target, $\theta = \bar{\theta}$, and is the direct finite-sample analog of Proposition 3. The right 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 4.

Finally, we plot the moment functions underlying each estimate. For each target $\bar{\theta} = (N, \bar{y})$ on our grid, the adversarial construction of Proposition 2 produces a pair of moment functions, one for each component of θ, evaluated at the same censored-logspline estimate $\hat{P}$. Figure 6 plots these functions. Each facet corresponds to one grid point, laid out as in the parameter grid (rows: $\bar{y}$ decreasing top to bottom; columns: N increasing left to right). The horizontal axis is the censored absolute price change $|\Delta p|$ on $[0, c]$; the filled dot at $|\Delta p| = c$ reports the value each moment takes when $X > c$.

Because the just-identified estimator is invariant to any full-rank linear transformation

		At target $\bar{\theta}$		At estimate $\hat{\theta}_{g_{\bar{\theta},\hat{P}}}$	
N	$\bar{y}$	$\sqrt{VIF_N}-1$	$\sqrt{VIF_{\bar{y}}}-1$	$\sqrt{VIF_N}-1$	$\sqrt{VIF_{\bar{y}}}-1$
2	0.06	24.8%	5.3%	23.1%	4.8%
2	0.12	2.6%	3.3%	2.5%	3.4%
2	0.18	0.1%	5.0%	0.2%	4.9%
3	0.06	4.9%	12.8%	4.8%	10.5%
3	0.12	5.7%	15.4%	4.8%	14.4%
3	0.18	5.7%	15.2%	4.5%	14.4%
4	0.06	0.5%	100.6%	1.9%	75.5%
4	0.12	20.0%	22.8%	18.0%	20.8%
4	0.18	17.0%	24.7%	14.5%	22.9%
5	0.06	41.6%	218.7%	39.0%	214.2%
5	0.12	27.9%	21.8%	26.0%	20.0%
5	0.18	26.8%	31.0%	23.8%	28.4%
Median		11.3%	18.6%	9.7%	17.2%

Table 1: 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 $\theta=\hat{\theta}_{g_{\bar{\theta},\hat{P}}}$ (right panel), with both asymptotic-variance matrices evaluated under P_θ.

of the moment vector (it solves $\frac{1}{n}\sum_i g - \mathrm{E}_\theta[g] = 0$, which is unchanged by $g \mapsto Ag+b$ for invertible A and any b), the pair displayed in each facet is only defined up to such a transformation. We normalize the two displayed functions to be mean-zero, orthogonal, and of unit variance under Lebesgue measure on $[0,c]$, and in particular choose the orthonormal pair closest in mean square to the raw moments (after normalizing those to also be mean zero and unit variance). The overflow point mass enters neither the mean nor the inner product; its value is mapped through the same transformation and shown as the dot at $|\Delta p| = c$.

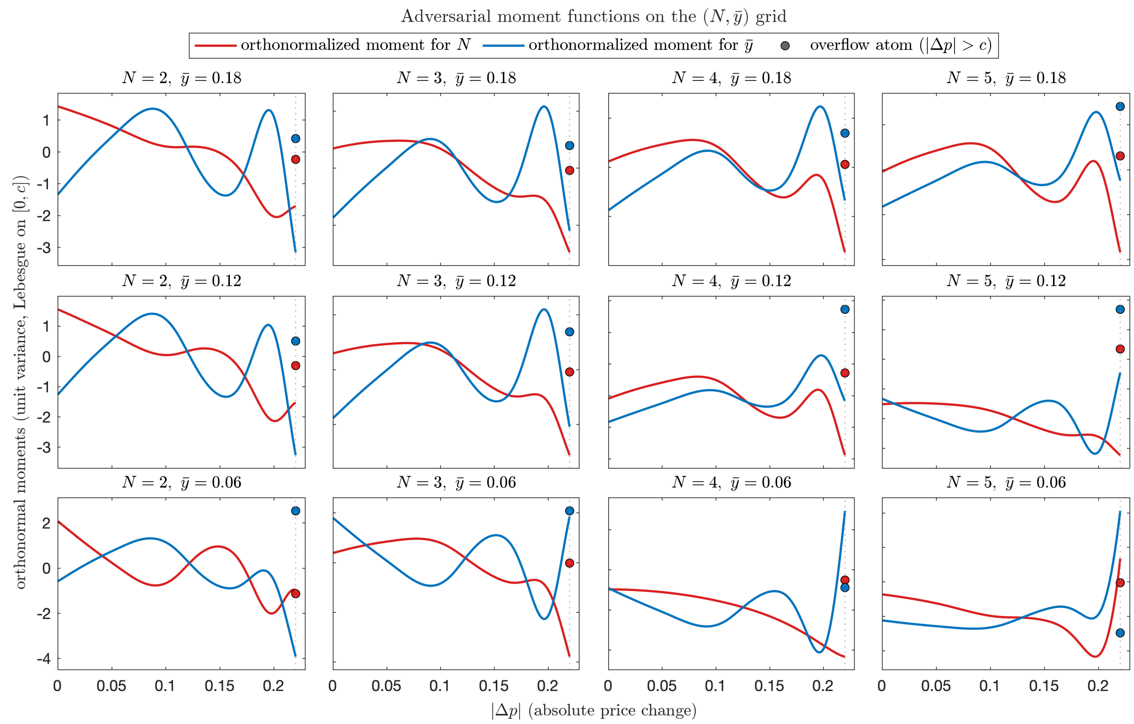

Figure 6: Adversarial moment functions on the $(N, \bar{y})$ grid, evaluated at the same-sample censored-logspline estimate $\hat{P}$ of Section S1. Each facet is one target; within a facet the two moment functions are mean-zero, orthogonal, and of unit variance under Lebesgue measure on $[0, c]$ (symmetric orthogonalization of the raw $(N, \bar{y})$ moment pair). The filled dot at $|\Delta p| = c$ is the moment's value when $X > c$; the overflow point mass does not enter the orthogonalization.

S2 Proofs

S2.1 Proof of Lemma 1

Define the likelihood ratios

$$L := \frac{dP}{dP_{\bar{\theta}}}, \ r := L - 1.$$

Take any $g \in L^2\left(P_{\bar{\theta}}; \mathbb{R}^D\right)$. Since P is absolutely continuous with respect to $P_{\bar{\theta}}$, we have

$$\mathrm{E}_P\left[g\left(X\right)\right] = \mathrm{E}_{\bar{\theta}}\left[g\left(X\right)L\left(X\right)\right].$$

Therefore the equality

$$\mathrm{E}_{\bar{\theta}}\left[g\left(X\right)\right] = \mathrm{E}_P\left[g\left(X\right)\right]$$

holds if and only if

$$\mathrm{E}_{\bar{\theta}}\left[g\left(X\right)r\left(X\right)\right] = 0.$$

So we can rewrite the constraint set as

$$\mathcal{C}\left(\bar{\theta}\right) = \left\{g \in L^2\left(P_{\bar{\theta}}; \mathbb{R}^D\right) : \mathrm{E}_{\bar{\theta}}\left[g\left(X\right)r\left(X\right)\right] = 0\right\}.$$

Now, to show the displayed set is included in $\mathcal{C}\left(\bar{\theta}\right)$, fix any $f \in L^2\left(P_{\bar{\theta}}; \mathbb{R}^D\right)$ and define

$$g := f - \frac{\mathrm{E}_{\bar{\theta}}\left[f\left(X\right)r\left(X\right)\right]}{\mathrm{E}_{\bar{\theta}}\left[r\left(X\right)^2\right]}r.$$

Because $\mathrm{E}_{\bar{\theta}}\left[r\left(X\right)^2\right] = \chi^2\left(P \| P_{\bar{\theta}}\right) \in (0, \infty)$, the denominator is finite and nonzero, so g is well-defined. Moreover, since $f \in L^2\left(P_{\bar{\theta}}; \mathbb{R}^D\right)$ and $r \in L^2\left(P_{\bar{\theta}}; \mathbb{R}\right)$, we have $g \in L^2\left(P_{\bar{\theta}}; \mathbb{R}^D\right)$. Then since $\mathrm{E}_{\bar{\theta}}\left[g\left(X\right)r\left(X\right)\right] = 0$, we have $g \in \mathcal{C}\left(\bar{\theta}\right)$.

To show the reverse inclusion, take any $g \in \mathcal{C}\left(\bar{\theta}\right)$. Then $\mathrm{E}_{\bar{\theta}}\left[g\left(X\right)r\left(X\right)\right] = 0$ by definition of $\mathcal{C}\left(\bar{\theta}\right)$. Set $f = g$. Then

$$g = f - \frac{\mathrm{E}_{\bar{\theta}}\left[f\left(X\right)r\left(X\right)\right]}{\mathrm{E}_{\bar{\theta}}\left[r\left(X\right)^2\right]}r.$$

So g belongs to the displayed set. □

S2.2 Proof of Proposition 1

We make use of the following lemma.

Lemma 3 (Collinearity and mixtures)**.** *Fix $\bar\theta$ and suppose $0 < \chi^2(P\|P_{\bar\theta}) < \infty$. Then*

$$\inf_{a\in\mathbb{R}} \mathrm{E}_{\bar\theta}\left[(r_\theta(X) - ar(X))^2\right] = 0$$

if and only if one of $P_{\bar\theta}$, P_θ, and P is a possibly degenerate mixture of the other two.

Proof. Because $\mathrm{span}\{r\}$ is a finite-dimensional closed subspace of $L^2(P_{\bar\theta})$, the displayed infimum equals zero if and only if there exists $a \in \mathbb{R}$ such that $r_\theta = ar$ $P_{\bar\theta}$-almost surely. Equivalently,

$$\frac{dP_\theta}{dP_{\bar\theta}} = 1 + a\left(\frac{dP}{dP_{\bar\theta}} - 1\right) = (1-a) + a\frac{dP}{dP_{\bar\theta}}.$$

Integrating both sides over an arbitrary measurable set A gives

$$P_\theta(A) = (1-a)P_{\bar\theta}(A) + aP(A),$$

so $P_\theta = (1-a)P_{\bar\theta} + aP$. 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 $P_{\bar\theta} = \frac{1}{1-a}P_\theta + \frac{-a}{1-a}P$. Finally, $a = 0$ gives $P_\theta = P_{\bar\theta}$, and $a = 1$ gives $P_\theta = P$.

Conversely, if one of the three distributions is a possibly degenerate mixture of the other two, elementary rearrangement yields $P_\theta = (1-a)P_{\bar\theta} + aP$ for some $a \in \mathbb{R}$. Taking Radon–Nikodym derivatives with respect to $P_{\bar\theta}$ gives $r_\theta = ar$, and hence the infimum is zero. □

That lemma in hand, note that, for $C_d(\bar\theta)$ the analog of $\mathcal{C}(\bar\theta)$ for $\mathbb{R}^d$-valued functions, $s_{\bar\theta,\perp} = s_{\bar\theta} - \frac{\mathrm{E}_{\bar\theta}[s_{\bar\theta} r]}{\mathrm{E}_{\bar\theta}[r^2]} r \in \mathcal{C}_K(\bar\theta)$. Indeed, $\mathrm{E}_{\bar\theta}[s_{\bar\theta,\perp}(X)] = 0$ because $\mathrm{E}_{\bar\theta}[s_{\bar\theta}(X)] = 0$ and $\mathrm{E}_{\bar\theta}[r(X)] = 0$, while

$$\mathrm{E}_{\bar\theta}[s_{\bar\theta,\perp}(X)r(X)] = 0$$

by construction. Hence $\mathrm{E}_P[s_{\bar\theta,\perp}(X)] = \mathrm{E}_{\bar\theta}[s_{\bar\theta,\perp}(X)L(X)] = \mathrm{E}_{\bar\theta}[s_{\bar\theta,\perp}(X)]$.

Next, define $\Psi(\theta) := \mathrm{E}_{P_\theta}[s_{\bar\theta,\perp}(X)]$, and observe that $\Psi(\bar\theta) = 0$. By differentiating under the integral sign, $\dot\Psi(\bar\theta) = \mathrm{E}_{\bar\theta}\left[s_{\bar\theta,\perp}(X)s_{\bar\theta}(X)^\top\right]$. Since $s_{\bar\theta,\perp}$ is the residual from projecting $s_{\bar\theta}$ on $\mathrm{span}\{r\}$, the last display equals $\mathrm{E}_{\bar\theta}\left[s_{\bar\theta,\perp}(X)s_{\bar\theta,\perp}(X)^\top\right] := Q_{\bar\theta}$. By Assumption 1(iv), $Q_{\bar\theta}$ is nonsingular. The inverse function theorem therefore implies that there exists $\varepsilon > 0$ such that

$$\Psi(\theta) = \Psi(\bar\theta) \quad \text{and} \quad \|\theta - \bar\theta\| < \varepsilon \quad \Longrightarrow \quad \theta = \bar\theta.$$

Moment matching on $s_{\bar\theta,\perp}$ admits no local solutions other than $\bar\theta$.

Now consider the nonlocal solution set based on moments $s_{\bar{\theta},\perp}$

$$S := \left\{\theta \in \Theta : \|\theta - \bar{\theta}\| \geq \varepsilon,\ \mathrm{E}_{\theta}[s_{\bar{\theta},\perp}(X)] = \mathrm{E}_{\bar{\theta}}[s_{\bar{\theta},\perp}(X)]\right\}.$$

By the $L^2(P_{\bar{\theta}})$-continuity of $\theta \mapsto r_\theta$, the moment maps are continuous for every fixed square-integrable moment. Hence S is compact. Fix $\theta \in S$. Define $v_\theta := r_\theta - \frac{\mathrm{E}_{\bar{\theta}}[r_\theta(X)r(X)]}{\mathrm{E}_{\bar{\theta}}[r(X)^2]} r$. Then $v_\theta \in L^2(P_{\bar{\theta}})$ and $\mathrm{E}_{\bar{\theta}}[v_\theta(X)r(X)] = 0$, so $v_\theta \in \mathcal{C}_1(\bar{\theta})$. Moreover,

$$\mathrm{E}_{\theta}[v_\theta(X)] - \mathrm{E}_{\bar{\theta}}[v_\theta(X)] = \mathrm{E}_{\bar{\theta}}[v_\theta(X)r_\theta(X)] = \mathrm{E}_{\bar{\theta}}[v_\theta(X)^2].$$

The last quantity is strictly positive by Assumption 1(v) and Lemma 3, since v_θ is the residual from projecting r_θ on $\operatorname{span}\{r\}$. Therefore v_θ separates this θ from $\bar{\theta}$.

By continuity of the moment map associated with v_θ, this separation persists on an open neighborhood U_θ of θ. The collection $\{U_\theta : \theta \in S\}$ covers the compact set S, so there exist $\theta_1, \ldots, \theta_J \in S$ such that

$$S \subseteq \bigcup_{j=1}^{J} U_{\theta_j}.$$

Let $v_j := v_{\theta_j}$. Each $v_j \in \mathcal{C}_1(\bar{\theta})$, so $\bar{\theta}$ satisfies all augmented moment conditions. Local alternatives are excluded by $s_{\bar{\theta},\perp}$, and nonlocal alternatives are excluded by at least one of $v_1, \ldots, v_J$. Hence for g that collects these moment functions, $\{\theta \in \Theta : \mathrm{E}_\theta[g(X)] = \mathrm{E}_P[g(X)]\} = \{\bar{\theta}\}$, as desired. $\square$

S2.3 Proof of Proposition 2

Again let $C_d(\bar{\theta})$ denote the analog of $\mathcal{C}(\bar{\theta})$ for $\mathbb{R}^d$-valued functions. Adding a constant to g does not affect the asymptotic variance, so we work with centered moments $h(X) := g(X) - \mathrm{E}_{P_{\bar{\theta}}}[g(X)]$. The constraint $g \in \mathcal{C}_K(\bar{\theta})$ is equivalent to

$$\mathrm{E}_{P_{\bar{\theta}}}[h(X)] = 0, \qquad \mathrm{E}_{P_{\bar{\theta}}}[h(X)r(X)] = 0.$$

Let $\mathcal{H}$ denote the closed linear subspace of such centered moments.

For $\mathbb{R}^K$-valued h, define

$$G_h := \mathrm{E}_{P_{\bar{\theta}}}\left[h(X)s_{\bar{\theta}}(X)^\top\right], \qquad \Omega_h := \mathrm{E}_{P_{\bar{\theta}}}\left[h(X)h(X)^\top\right].$$

The asymptotic variance is $G_h^{-1}\Omega_h(G_h^{-1})^\top$. Because this expression is invariant to nonsingular linear transformations of h, and the variance is infinite when G has reduced rank, we may impose the normalization $\mathrm{E}_{P_{\bar\theta}}[h(X)s_{\bar\theta}(X)^\top] = I_K$. Since every $h \in \mathcal{H}$ is orthogonal to r, and $s_{\bar\theta,\perp}$ is the projection of $s_{\bar\theta}$ onto $\mathcal{H}$,

$$\mathrm{E}_{P_{\bar\theta}}\left[h(X)s_{\bar\theta}(X)^\top\right] = \mathrm{E}_{P_{\bar\theta}}\left[h(X)s_{\bar\theta,\perp}(X)^\top\right].$$

Let $Q_{\bar\theta} := \mathrm{E}_{P_{\bar\theta}}[s_{\bar\theta,\perp}(X)s_{\bar\theta,\perp}(X)^\top]$, and consider the normalized candidate $h^*(X) := Q_{\bar\theta}^{-1}s_{\bar\theta,\perp}(X)$. This candidate is feasible and satisfies the normalization because

$$\mathrm{E}_{P_{\bar\theta}}[h^*(X)s_{\bar\theta}(X)^\top] = \mathrm{E}_{P_{\bar\theta}}[h^*(X)s_{\bar\theta,\perp}(X)^\top] = Q_{\bar\theta}^{-1}Q_{\bar\theta} = I_K.$$

Now take any other normalized feasible h. Since $h \in \mathcal{H}$, $\mathrm{E}_{P_{\bar\theta}}[h(X)r(X)] = 0$. Using $s_{\bar\theta,\perp} = s_{\bar\theta} - \frac{\mathrm{E}_{P_{\bar\theta}}[s_{\bar\theta}(X)r(X)]}{\mathrm{E}_{P_{\bar\theta}}[r(X)^2]}r$, we have

$$\mathrm{E}_{P_{\bar\theta}}[h(X)s_{\bar\theta,\perp}(X)^\top] = \mathrm{E}_{P_{\bar\theta}}[h(X)s_{\bar\theta}(X)^\top] - \mathrm{E}_{P_{\bar\theta}}[h(X)r(X)]\frac{\mathrm{E}_{P_{\bar\theta}}[s_{\bar\theta}(X)r(X)]^\top}{\mathrm{E}_{P_{\bar\theta}}[r(X)^2]} = I_K.$$

Therefore, defining $u(X) := h(X) - Q_{\bar\theta}^{-1}s_{\bar\theta,\perp}(X)$, we obtain

$$\begin{aligned}
\mathrm{E}_{P_{\bar\theta}}[u(X)s_{\bar\theta,\perp}(X)^\top] &= \mathrm{E}_{P_{\bar\theta}}[h(X)s_{\bar\theta,\perp}(X)^\top] - \mathrm{E}_{P_{\bar\theta}}[Q_{\bar\theta}^{-1}s_{\bar\theta,\perp}(X)s_{\bar\theta,\perp}(X)^\top] \\
&= I_K - Q_{\bar\theta}^{-1}Q_{\bar\theta} \\
&= 0.
\end{aligned}$$

Taking transposes also gives $\mathrm{E}_{P_{\bar\theta}}[s_{\bar\theta,\perp}(X)u(X)^\top] = 0$.

Hence

$$\begin{aligned}
\mathrm{E}_{P_{\bar\theta}}[h(X)h(X)^\top] &= \mathrm{E}_{P_{\bar\theta}}[\{Q_{\bar\theta}^{-1}s_{\bar\theta,\perp}(X) + u(X)\}\{Q_{\bar\theta}^{-1}s_{\bar\theta,\perp}(X) + u(X)\}^\top] \\
&= Q_{\bar\theta}^{-1} + \mathrm{E}_{P_{\bar\theta}}[u(X)u(X)^\top],
\end{aligned}$$

because the cross terms vanish. Since $\mathrm{E}_{P_{\bar\theta}}[u(X)u(X)^\top] \succeq 0$, we have $\mathrm{E}_{P_{\bar\theta}}[h(X)h(X)^\top] \succeq Q_{\bar\theta}^{-1}$, where equality holds if and only if $u(X) = 0$ $P_{\bar\theta}$-almost surely. Thus $Q_{\bar\theta}^{-1}s_{\bar\theta,\perp}$ is the unique normalized minimizer. $\square$

S2.4 Proof of Lemma 2

Let $\tilde{g} = (g_{\bar{\theta}}^{*\top}, q^\top)^\top$. Consider weight matrices of the form

$$W_\varepsilon = \begin{pmatrix} W_0 & 0 \\ 0 & \varepsilon I_J \end{pmatrix}, \qquad \varepsilon > 0,$$

where W_0 is any positive definite $K \times K$ matrix. For every $\varepsilon > 0$, W_ε is positive definite, and the corresponding asymptotic variance is well-defined. As $\varepsilon \downarrow 0$, the GMM estimator based on $\tilde{g}$ with weight W_ε has asymptotic variance converging to the just-identified asymptotic variance based on $g_{\bar{\theta}}^*$ alone, since the additional moments receive vanishing weight.

Efficient weighting minimizes the asymptotic variance among positive definite weighting matrices. Hence $\mathrm{AVar}\left(\hat{\theta}_{\tilde{g},\mathrm{Var}_{\bar{\theta}}(\tilde{g})^{-1}}\right) \leq \lim_{\varepsilon \downarrow 0} \mathrm{AVar}\left(\hat{\theta}_{\tilde{g},W_\varepsilon}\right) = \mathrm{AVar}\left(\hat{\theta}_{g_{\bar{\theta}}^*}\right)$, where for square matrices A and B, we interpret $A \geq B$ to mean that $A - B$ is positive semidefinite. Conversely, for any positive definite weighting matrix W, the overidentified GMM estimator based on $\tilde{g}$ has the same asymptotic variance as the just-identified estimator based on

$$\left(G^\top W G\right)^{-1} G^\top W \tilde{g}(X),$$

where $G = \partial\, \mathrm{E}_\theta[\tilde{g}(X)]/\partial\theta|_{\theta=\bar{\theta}}$. Proposition 2 therefore implies that no such estimator can have asymptotic variance strictly below $\mathrm{AVar}\left(\hat{\theta}_{g_{\bar{\theta}}^*}\right)$. Combining the two inequalities gives $\mathrm{AVar}\left(\hat{\theta}_{\tilde{g},W}\right) = \mathrm{AVar}\left(\hat{\theta}_{g_{\bar{\theta}}^*}\right)$ for $W = \mathrm{Var}_{\bar{\theta}}(\tilde{g})^{-1}$. □

S2.5 Proof of Proposition 3

Fix $\bar{\theta} \in \Theta$, and denote $I\left(\bar{\theta}\right) := \mathrm{E}_{\bar{\theta}}\left[s_{\bar{\theta}}\left(X\right) s_{\bar{\theta}}\left(X\right)^\top\right]$ and $m\left(\bar{\theta}\right) := \mathrm{E}_P\left[s_{\bar{\theta}}\left(X\right)\right]$. From Proposition 2, we know

$$\mathrm{AVar}\left(\hat{\theta}_{g_{\bar{\theta}}^*}\right) = \left(I\left(\bar{\theta}\right) - \frac{m\left(\bar{\theta}\right) m\left(\bar{\theta}\right)^\top}{\chi^2\left(P \| P_{\bar{\theta}}\right)}\right)^{-1},$$

while the MLE benchmark has $\mathrm{AVar}\left(\hat{\theta}_{s_{\bar{\theta}}}\right) = I\left(\bar{\theta}\right)^{-1}$. For any non-zero $\alpha \in \mathbb{R}^K$, define

$$VIF_\alpha\left(\bar{\theta}\right) := \frac{\alpha^\top \mathrm{AVar}\left(\hat{\theta}_{g_{\bar{\theta}}^*}\right) \alpha}{\alpha^\top \mathrm{AVar}\left(\hat{\theta}_{s_{\bar{\theta}}}\right) \alpha}.$$

Writing $y = I\left(\bar{\theta}\right)^{-1/2}\alpha$ and $w = I\left(\bar{\theta}\right)^{-1/2}\frac{m}{\sqrt{\chi^2(P\|P_{\bar{\theta}})}}$, we have

$$VIF_\alpha\left(\bar{\theta}\right) = \frac{\alpha^\top\left\{I\left(\bar{\theta}\right) - \frac{m(\bar{\theta})m(\bar{\theta})^\top}{\chi^2(P\|P_{\bar{\theta}})}\right\}^{-1}\alpha}{\alpha^\top I\left(\bar{\theta}\right)^{-1}\alpha} = \frac{y^\top\left(I_K - ww^\top\right)^{-1}y}{y^\top y}.$$

Since $I\left(\bar{\theta}\right) - \frac{m(\bar{\theta})m(\bar{\theta})^\top}{\chi^2(P\|P_{\bar{\theta}})}$ is positive definite, $I_K - ww^\top$ is positive definite. So from the Sherman-Morrison formula we know

$$\left(I_K - ww^\top\right)^{-1} = I_K + \frac{ww^\top}{1 - w^\top w},$$

so we obtain

$$VIF_\alpha\left(\bar{\theta}\right) = 1 + \frac{\left(y^\top w\right)^2}{\left(1 - w^\top w\right)y^\top y} = 1 + \frac{\left(\alpha^\top I\left(\bar{\theta}\right)^{-1}m\left(\bar{\theta}\right)\right)^2}{\left(\chi^2\left(P\|P_{\bar{\theta}}\right) - m\left(\bar{\theta}\right)^\top I\left(\bar{\theta}\right)^{-1}m\left(\bar{\theta}\right)\right)\alpha^\top I\left(\bar{\theta}\right)^{-1}\alpha}.$$

To find $VIF\left(\bar{\theta}\right)$, note that

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

But the last expression is the maximum of a Rayleigh quotient and it follows that

$$VIF\left(\bar{\theta}\right) = \lambda_{\max}\left\{\left(I_K - ww^\top\right)^{-1}\right\} = \frac{1}{1 - w^\top w} = \frac{1}{1 - \frac{m(\bar{\theta})^\top I(\bar{\theta})^{-1}m(\bar{\theta})}{\chi^2(P\|P_{\bar{\theta}})}}.$$

Since $\chi^2\left(P\|P_{\bar{\theta}}\right) = \chi^2_{L^2}\left(P\|P_{\bar{\theta}}\right)$, it now remains to show that

$$\chi^2_{\mathcal{S}_{\bar{\theta}}}\left(P\|P_{\bar{\theta}}\right) = m\left(\bar{\theta}\right)^\top I\left(\bar{\theta}\right)^{-1}m\left(\bar{\theta}\right). \tag{S1}$$

Toward this, note that any $h \in \mathcal{S}_{\bar{\theta}}$ will look like $h(X) = \beta^\top s_{\bar{\theta}}(X)$ and that $\mathrm{E}_{\bar{\theta}}\left[s_{\bar{\theta}}(X)\right] = 0$. The left-hand side becomes

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

The maximum is achieved at $\beta \propto I\left(\bar{\theta}\right)^{-1}m\left(\bar{\theta}\right)$ and the conclusion follows. $\square$