cemmap

centre for microdata methods and practice

Inference for parameters identified by conditional moment restrictions using a generalized Bierens maximum statistic

Xiaohong Chen
Sokbae Lee
Myung Hwan Seo
Myunghyun Song

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP26/24

Inference for parameters identified by conditional moment restrictions using a generalized Bierens maximum statistic*

Xiaohong Chen† Sokbae Lee‡ Myung Hwan Seo§
Myunghyun Song¶

Abstract

Many economic panel and dynamic models, such as rational behavior and Euler equations, imply that the parameters of interest are identified by conditional moment restrictions. We introduce a novel inference method without any prior information about which conditioning instruments are weak or irrelevant. Building on Bierens (1990), we propose penalized maximum statistics and combine bootstrap inference with model selection. Our

*This is a revised version of the ArXiv:2008.11140 paper entitled "Powerful Inference", which has been online since 25 August 2020. We are grateful to the editor, four anonymous referees, and Jesse Shapiro for helpful comments. This work was supported in part by the Cowles Foundation, the European Research Council (ERC-2014-CoG- 646917-ROMIA) and by the UK Economic and Social Research Council (ESRC) through research grant (ES/P008909/1) to the CeMMAP. Part of this research was carried out when Seo was visiting Cowles Foundation in 2018/2019 and was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A01033487) and the Seoul National University Research Grant in 2023.

†Yale University and Cowles Foundation for Research in Economics; xiaohong.chen@yale.edu

‡Columbia University and Centre for Microdata Methods and Practice; sl3841@columbia.edu

§Author Correspondance; Seoul National University and Institute of Economic Research; myunghseo@snu.ac.kr; +82 2 880 4108

¶Columbia University ; ms6347@columbia.edu

method optimizes asymptotic power by solving a data-dependent max-min problem for tuning parameter selection. Extensive Monte Carlo experiments, based on an empirical example, demonstrate the extent to which our inference procedure is superior to those available in the literature. [C12, C36].

Keywords: conditional moment restrictions, conditional instruments, hypothesis testing, penalization, multiplier bootstrap, max-min.

1 Introduction

Conditional moment restrictions of the following forms are ubiquitous in economics:

$$\mathbb{E}\left[g\left(X_i,\theta\right)|W_i\right]=0 \quad a.s. \quad \text{if and only if } \theta=\theta_0, \tag{1}$$

where $g(x,\theta)$ is a real-valued known function with a finite-dimensional parameter θ, and $W_i \in \mathbb{R}^p$ is a p-dimensional vector of conditioning instruments so that the parameter of interest θ_0 is strongly identified by (1). There is now a mature literature on optimal estimation and inference for θ_0 in (1) when the conditioning instruments W_i are all relevant and of low-dimension. Most existing approaches are firstly to estimate θ_0 efficiently and secondly to develop suitable test statistics based on the estimators (see, e.g., Chamberlain, 1987; Donald, Imbens, and Newey, 2003; Ai and Chen, 2003; Domínguez and Lobato, 2004; Kitamura, Tripathi, and Ahn, 2004, among many others). However, when the dimension p of conditioning instruments W_i is *boundedly large*[1] and when some of W_i are weak or irrelevant, semiparametrically efficient estimators for θ_0, though theoretically optimal according to asymptotic theory, could perform poorly in small samples. In this paper, we

[1]We will use expression "boundedly large" throughout the paper to emphasize that we focus on the setting where the dimension (p) of the conditioning instruments is relatively large but does not grow with sample size (n). We are grateful to the editor for suggesting this expression.

take a different path and conduct inference for θ_0 in (1) directly by skipping the first step efficient estimation of θ_0.

To construct a more informative confidence set for θ_0, we propose an ℓ_1-penalized Bierens (1990)'s maximum statistic for hypothesis testing on θ_0 satisfying model (1). Our penalized inference method is shown to be asymptotically valid when the null hypothesis is true and can be calibrated to optimize the asymptotic power against a set of $n^{-1/2}$-local alternatives of interest. The penalization tuning parameter is selected by solving a data-dependent max-min problem. Specifically, we elaborate on the choice of the penalty parameter λ under a limit of experiments, formally define optimal λ, and establish consistency of our proposed calibration method.

Our new test can be viewed as a generalization of the original Bierens' max test with $\lambda = 0$ to allow for $\lambda \geq 0$. Most of the existing theoretical literature has simply ignored the importance of the choice of Γ in maximizing finite sample power of Bierens' max test, although Bierens and Wang (2012) discuss it for the integrated conditional moment test. Intuitively, a larger Γ will not reduce the finite-sample power of the test but will increase computational costs. Conversely, a smaller Γ may not incur high computational costs but could have an adverse effect on the finite-sample power. Our test is to let Γ be a large region so that it is not a binding constraint and then to choose a scalar $\lambda \geq 0$ in a data-driven way to maximize power. To get the same empirical power, it is much easier to optimize over a scalar λ than over a binding set $\Gamma \subset \mathbb{R}^p$ when p is boundedly large. That is, our penalized test statistic is easier to compute than the one without penalization.[2] The com-

[2]In convex optimization, a penalized approach is equivalent to a constrained one: that is, penalized optimization with parameter λ corresponds to constrained optimization with a different tuning parameter (see, e.g., Boyd and Vandenberghe, 2004). A smaller λ expands the feasible parameter space in the constrained problem, increasing computational time. For non-convex problems like ours, the equivalence is not guaranteed, but we speculate a similar phenomenon because the ℓ_1 penalty term is convex and a larger λ leads the overall objective function to behave more akin to a convex function. To confirm, we

putational gains by penalization are practically important since the p-value is constructed by a multiplier bootstrap procedure.

We demonstrate the usefulness of our method by applying it to a couple of empirical examples. First, as our main example, we revisit Yogo (2004) and find that an uninformative confidence interval (resulting from unconditional moment restrictions) for the elasticity of intertemporal substitution, based on annual US series ($n \approx 100$ and $p = 4$), can turn into an informative one. We provide further supporting evidence via extensive Monte Carlo experiments that mimic Yogo (2004). Second, as our supplementary example, in the online appendix, we revisit the test of Benítez-Silva, Buchinsky, Chan, Cheidvasser, and Rust (2004) and show that our method yields a rejection of the null hypothesis of rational unbiased reporting of ability status at the conventional level even with a small sample size $n \approx 350$ and relatively large $p = 21$. Interestingly, Benítez-Silva, Buchinsky, Chan, Cheidvasser, and Rust (2004) already implemented their version of Bierens' test and failed to reject the null. This is also consistent with an anonymous referee's point that the original Bierens' tests are not recommended for applications with small sample sizes. Both empirical examples suggest that there is substantive evidence for the efficacy of our proposed method.

Closely related literature The original Bierens' statistics (Bierens, 1982, 1990) are designed for consistent specification tests of a parametric null regression model against nonparametric regression alternatives. See, e.g., de Jong (1996); Andrews (1997); Bierens and Ploberger (1997); Stinchcombe and White (1998); Chen and Fan (1999); Fan and Li (2000); Lavergne and Patilea (2008) for earlier various extensions of Bierens' consistent tests of parametric, semiparametric null against nonparametric alternatives in regression settings. Horowitz (2006) is perhaps the first to extend Bierens (1982)'s integrated conditional moment (ICM) statistic to test a null of a parametric IV regression $E[Y - f(X, \theta_0)|W] = 0$

analyzed computation times and present the results in Table 2.

(with $X \neq W$) against a nonparametric IV regression alternative. Our motivation is different as we are concerned about testing the null hypothesis regarding θ_0 in (1), assuming that the underlying IV model (1) is correctly specified while some of the conditioning instruments W_i could be irrelevant.

Recently, Antoine and Lavergne (2023) leveraged Bierens (1982)'s ICM statistic to develop an inference procedure for a finite-dimensional parameter of interest within a linear IV regression framework. The first arXiv version of our work is contemporaneous with theirs and shares the same objective of obtaining more informative confidence sets for θ_0 using conditional instruments, accommodating weak IV scenarios. Importantly, our tests are based on Bierens (1990)'s maximum statistic, whereas the tests of Antoine and Lavergne (2023) are built on Bierens (1982)'s ICM statistic. Thus, our tests are complements rather than serving as substitutes, proposing distinct statistics with differing power properties. However, our tests demonstrate superior size and power properties compared to those of Antoine and Lavergne (2023), particularly as the dimension of irrelevant instruments increases.[3]

The remainder of the paper is organized as follows. In Section 2, we define the test statistic and describe how to obtain bootstrap p-values. Section 3 establishes bootstrap validity. In Section 4, we derive consistency and local power and propose how to calibrate the penalization parameter to optimize the power of the test. In Section 5, we summarize our proposed inference procedure and provide pseudo-code. In Section 6, we extend our method to inference for $\theta_{1,0}$, for which we use plug-in estimation of $\theta_{2,0}$, where $\theta_0 = (\theta_{1,0}, \theta_{2,0})$. In Sections 7 and 8, respectively, we present our main empirical example and Monte Carlo experiments based on Yogo (2004). Section 9 gives concluding remarks. All the proofs are in Section A. Online Appendix B gives our supplementary empirical example using data from Benítez-Silva, Buchinsky, Chan, Cheidvasser, and Rust (2004).

[3]Detailed results of Monte Carlo experiments can be found in an earlier version of this paper, available at `https://arxiv.org/pdf/2008.11140v4`.

2 Test Statistic

In this section, we introduce our test statistic and describe how to carry out bootstrap inference.

Before we present our test statistic, we first assume the following conditions.

Assumption 1. (i) *Θ is a compact convex nonempty subset of $\mathbb{R}^d$.*

(ii) *The time series $\{X_i, W_i\}$ is strictly stationary and ergodic, where W_i is adapted to the natural filtration $\mathcal{F}_{i-1}$ up to time $i-1$, and $\{U_i := g(X_i, \theta_0)\}$ is a martingale difference sequence (mds). Furthermore, $\mathbb{E}[|g(X_i, \theta)|^c]$ is bounded on Θ for some $c > \max\{2, d\}$.*

(iii) *$\mathbb{E}\left[g\left(X_i, \theta\right)|W_i\right] = 0$ a.s. if and only if $\theta = \theta_0$, where $\theta_0 \in \Theta$.*

(iv) *The function $\mathbb{E}\left[|g\left(X_i, \theta\right)|^2\right]$ is bounded and bounded away from zero on Θ.*

(v) *W_i is a bounded random vector in $\mathbb{R}^p$.*

The boundedness assumption on W_i is without loss of generality since we can take a one-to-one transformation to ensure that each component of W_i is bounded (for instance, $x \mapsto \tan^{-1}(x)$ componentwise, as used in Bierens (1990)). Condition (ii) is standard and ensures the weak convergence of the stochastic processes $\sqrt{n}M_n(\gamma)$ and $s_n^2(\gamma)$, which will be introduced later in this section.

Let Γ denote a compact nonempty subset in $\mathbb{R}^p$. Define

$$M\left(\theta, \gamma\right) := \mathbb{E}\left[g\left(X_i, \theta\right)\exp(W_i'\gamma)\right]. \tag{2}$$

Bierens (1990) established the following result.

Lemma 1 (Bierens (1990))**.** *Let Assumption 1 hold. Then, $M\left(\theta, \gamma\right) = 0$ if $\theta = \theta_0$ and $M\left(\theta, \gamma\right) \neq 0$ for almost every $\gamma \in \Gamma$ if $\theta \neq \theta_0$.*

To minimize notational complexity, we often abbreviate $M(\gamma) := M(\theta_0, \gamma)$ as for $U_i := g(X_i, \theta_0)$ in condition (ii) throughout this paper. In order to test a hypothesis

$$H_0 : \theta_0 = \bar{\theta}$$

against its negation $H_1 : \theta_0 \neq \bar{\theta}$, we construct a test statistic as follows. First, define

$$\begin{aligned} M_n(\theta, \gamma) &:= \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta) \exp(W_i' \gamma), \\ s_n^2(\theta, \gamma) &:= \frac{1}{n} \sum_{i=1}^{n} \left[g(X_i, \theta) \exp(W_i' \gamma) \right]^2, \\ Q_n(\theta, \gamma) &:= \sqrt{n} \frac{|M_n(\theta, \gamma)|}{s_n(\theta, \gamma)}, \end{aligned} \tag{3}$$

where we let $Q_n = 0$ if $s_n = 0$ for a given θ and γ. Note that $g(X_i, \theta_0) \exp(W_i' \gamma)$ is a centered random variable and the criterion Q_n is motivated by the t-statistic and it is asymptotically folded standard Normal at $\theta = \theta_0$ for each γ. Define the test statistic for the null hypothesis as

$$T_n(\bar{\theta}, \lambda) := \sup_{\gamma \in \Gamma} \left[Q_n(\bar{\theta}, \gamma) - \lambda \|\gamma\|_1 \right], \tag{4}$$

where $\|a\|_1$ is the ℓ_1 norm of a vector a and $\lambda \geq 0$ is the penalization parameter.[4] We regard $T_n(\bar{\theta}, \lambda)$ as a stochastic process indexed by $\lambda \in \Lambda$, where Λ is a compact nonempty subset in $\mathbb{R}_+ := \{\lambda \in \mathbb{R} | \lambda \geq 0\}$.

We note that our test can be viewed as a generalization of the original Bierens' max test with $\lambda = 0$ to allow for $\lambda \geq 0$. Under our Assumption 1(iv), one can let $\Gamma = \prod_{j=1}^{p} [-a_j, a_j]$ for any $a_j > 0, \ j = 1, ..., p$. The motivation of our test is to let Γ be a large region so that Γ is not a binding constraint and then to choose a scalar $\lambda \geq 0$ in a data-driven way to maximize power. It is much easier to optimize over a scalar λ then over a set of tuning

[4] Instead of a weighted version of the ℓ_1-norm, we focus on the simple case for brevity.

parameters $a_j > 0,\ j = 1, ..., p$ when p is large.

We end this section by commenting that our test statistic $T_n(\bar{\theta}, \lambda)$ is substantially different from the LASSO's criterion. First of all, the term $M_n(\bar{\theta}, \gamma)$ is not the least squares objective function; second, γ is different from regression coefficients. Furthermore, if we mimic LASSO more closely, the test statistic would be

$$T_{n,\text{alt}}(\bar{\theta}, \lambda) := \sup_{\gamma \in \Gamma} \Big\{ n[M_n(\bar{\theta}, \gamma)]^2 - \lambda \|\gamma\|_1 \Big\}. \tag{5}$$

We have opted to consider $T_n(\bar{\theta}, \lambda)$ in the paper because it is properly studentized and comparable in scale to the ℓ_1-penalty term, unlike $T_{n,\text{alt}}(\bar{\theta}, \lambda)$ in (5). In addition, it would be easier to specify Γ with $T_n(\bar{\theta}, \lambda)$ because the magnitude of $M_n(\bar{\theta}, \gamma)$ tends to get larger as the scale of γ increases but $Q_n(\bar{\theta}, \gamma)$ may not.

Finally, we conclude this section by commenting that our choice of the ℓ_1-penalty is on an ad hoc basis. As is well known for LASSO, the ℓ_1-penalty is more likely to exclude irrelevant instruments than the ℓ_2-penalty and shrinks the unpenalized optimizer by the same amount while the ℓ_2-penalty shrinks it proportionally. Also, the penalty term affects the test statistic differently under the null and the alternative hypotheses. Ideally, the best penalty should reduce the value of the test statistic more under the null than under the alternatives to maximize the power, while maintaining some computational gains. This issue of choosing an optimal penalty term is a very challenging topic, which deserves further independent research.

2.1 Bootstrap Critical Values

We consider the multiplier bootstrap to carry out inference. Define

$$\begin{aligned} M_{n,*}(\theta,\gamma) &:= \frac{1}{n}\sum_{i=1}^{n} \eta_i^* g(X_i,\theta)\exp(W_i'\gamma), \\ s_{n,*}^2(\theta,\gamma) &:= \frac{1}{n}\sum_{i=1}^{n} \left[\eta_i^* g(X_i,\theta)\exp(W_i'\gamma)\right]^2, \\ Q_{n,*}(\theta,\gamma) &:= \sqrt{n}\frac{|M_{n,*}(\theta,\gamma)|}{s_{n,*}(\theta,\gamma)}, \\ T_{n,*}(\theta,\lambda) &:= \sup_{\gamma\in\Gamma}\left[Q_{n,*}(\theta,\gamma)-\lambda\|\gamma\|_1\right], \end{aligned} \tag{6}$$

where η_i^* is drawn from $N(0,1)$ and independent from data $\{(X_i,W_i) : i=1,\ldots,n\}$.[5] For each bootstrap replication r, let

$$T_{n,*}^{(r)}(\theta,\lambda) := \sup_{\gamma\in\Gamma}\left[Q_{n,*}^{(r)}(\theta,\gamma)-\lambda\|\gamma\|_1\right]. \tag{7}$$

For each λ, the bootstrap p-value is defined as

$$p_*(\bar\theta,\lambda) := \frac{1}{R}\sum_{r=1}^{R} 1\{T_{n,*}^{(r)}(\bar\theta,\lambda) > T_n(\bar\theta,\lambda)\}$$

for a large R. We reject the null hypothesis at the α level if and only if $p_*(\bar\theta,\lambda)<\alpha$. Then, a bootstrap confidence interval for θ_0 can be constructed by inverting a pointwise test of $H_0 : \theta_0 = \bar\theta$.

3 Bootstrap Validity

We now introduce some additional notation. Let $K(\theta,\gamma_1,\gamma_2) := \mathbb{E}\left[g(X_i,\theta)^2\exp\left(W_i'\left(\gamma_1+\gamma_2\right)\right)\right]$ and $s^2(\theta,\gamma) := \mathbb{E}\left[g(X_i,\theta)^2\exp\left(2W_i'\gamma\right)\right]$. As before, we suppress the dependence on θ when

[5] Instead of using $s_{n,*}^2(\theta,\gamma)$, we may employ $s_n^2(\theta,\gamma)$ to define $Q_{n,*}(\theta,\gamma)$. The resulting first-order asymptotic theory would be equivalent.

they are evaluated at θ_0. The boundedness of W_i and Γ and the moment condition for U_i together imply that

$$\sup_{(\gamma_1,\gamma_2)\in\Gamma^2} K(\gamma_1,\gamma_2) < \infty \text{ and } \inf_{\gamma\in\Gamma} s^2(\gamma) > 0.$$

Let $\{\mathcal{M}(\theta,\gamma) : \gamma \in \Gamma\}$ be a centered Gaussian process with the covariance kernel $\mathbb{E}[\mathcal{M}(\theta,\gamma_1)\mathcal{M}(\theta,\gamma_2)] = K(\theta,\gamma_1,\gamma_2)$. Also, let $\Rightarrow$ denote the weak convergence in the space of uniformly bounded functions on the parameter space that is endowed with the uniform metric. Additionally, we define $L_\infty(A)$ as the space of all bounded functions defined on $A = \Gamma$ or $A = \Lambda$.

We first establish the weak convergence of $T_n(\lambda)$.

Theorem 1. *Let Assumptions 1 hold. Then,*

$$\sqrt{n}M_n(\gamma) \Rightarrow \mathcal{M}(\gamma) \quad \textit{in } L_\infty(\Gamma), \tag{8}$$

$$s_n(\gamma) \xrightarrow{p} s(\gamma) \quad \textit{uniformly in } \Gamma. \tag{9}$$

Furthermore,

$$T_n(\lambda) \Rightarrow T(\lambda) := \sup_{\gamma\in\Gamma}\left[\frac{|\mathcal{M}(\gamma)|}{s(\gamma)} - \lambda\|\gamma\|_1\right] \quad \textit{in } L_\infty(\Lambda).$$

We now show that the bootstrap analog $T_{n,*}(\lambda)$ of $T_n(\lambda)$ converges weakly to the same limit. The definition of the conditional weak convergence, $\Rightarrow^*$ in P, and conditional convergence in probability, $\xrightarrow{p*}$ in P, employed in the following theorem is referred to e.g. Section 2.9 in van der Vaart and Wellner (1996).

Theorem 2. *Let Assumption 1 hold. Then, for any fixed* $\bar{\theta}$*,*

$$\sqrt{n}M_{n,*}(\bar{\theta},\gamma) \Rightarrow^* \mathcal{M}\left(\bar{\theta},\gamma\right) \quad \textit{in } L_\infty(\Gamma) \quad \textit{in } P, \tag{10}$$

$$s_{n,*}(\bar{\theta},\gamma) \xrightarrow{p*} s\left(\bar{\theta},\gamma\right) \quad \textit{uniformly in } \Gamma \quad \textit{in } P. \tag{11}$$

Furthermore,

$$T_{n,*}(\bar{\theta},\lambda) \Rightarrow^* T(\bar{\theta},\lambda) := \sup_{\gamma\in\Gamma}\left[\frac{\left|\mathcal{M}\left(\bar{\theta},\gamma\right)\right|}{s\left(\bar{\theta},\gamma\right)} - \lambda\|\gamma\|_1\right] \quad \textit{in } L_\infty(\Lambda) \quad \textit{in } P.$$

Recall that we abbreviate $M_n(\gamma) := M_n(\theta_0,\gamma)$, $s_n^2(\gamma) := s_n^2(\theta_0,\gamma)$ and $T_n(\lambda) := T_n(\theta_0,\lambda)$. While Theorem 1 holds at $\theta = \theta_0$, Theorem 2 holds for any $\bar{\theta}$. Therefore, Theorems 1 and 2 imply that the bootstrap critical values are valid for any drifting sequence of $\lambda_n \in \Lambda$ under some smoothness condition on the limiting distribution and thus for the calibration of Λ described in Section 4.3. See the proof of Theorem 4 for more details.

4 Consistency, Local Power and Calibration of λ

4.1 Consistency

Suppose that the null hypothesis is set as $H_0 : \theta_0 = \bar{\theta}$ for some $\bar{\theta} \neq \theta_0$. Then, being explicit about the null, we write

$$M_n(\bar{\theta},\gamma) = \frac{1}{n}\sum_{i=1}^{n} g\left(X_i,\bar{\theta}\right)\exp(W_i'\gamma) \xrightarrow{p} \mathbb{E}\left[g\left(X_i,\bar{\theta}\right)\exp(W_i'\gamma)\right],$$
$$s_n(\bar{\theta},\gamma) \xrightarrow{p} \sqrt{\mathbb{E}\left[g^2\left(X_i,\bar{\theta}\right)\exp(2W_i'\gamma)\right]}.$$

Therefore, for any $\lambda \in \Lambda$,

$$n^{-1/2}T_n(\bar{\theta},\lambda) \xrightarrow{p} \sup_{\gamma\in\Gamma}\frac{\left|\mathbb{E}\left[g\left(X_i,\bar{\theta}\right)\exp(W_i'\gamma)\right]\right|}{\sqrt{\mathbb{E}\left[g^2\left(X_i,\bar{\theta}\right)\exp(2W_i'\gamma)\right]}}.$$

On the other hand, the bootstrap statistic is always $O_p(1)$, yielding the consistency of the test based on $T_n(\bar{\theta},\lambda)$, as in the following theorem.

Theorem 3. *Let Assumptions 1 hold. Then, for $\bar{\theta} \neq \theta_0$, $T_n(\bar{\theta},\lambda) \xrightarrow{p} +\infty$ for any $\lambda \in \Lambda$.*

4.2 Local Power

Consider a sequence of local hypotheses of the following form: for some nonzero constant vector B,

$$\theta_n := \theta_0 + B\,n^{-1/2},$$

which leads to the following leading term after linearization:

$$U_{i,n} := U_i + G\left(X_i, \theta_0\right) B\,n^{-1/2}, \tag{12}$$

where $G\left(X_i, \theta\right) := \partial g\left(X_i, \theta\right)/\partial\theta'$, assuming the continuous differentiability of $g\left(\cdot\right)$ at θ_0. Unless $g\left(X_i, \theta\right)$ is linear in θ, the term $G\left(X_i, \theta_0\right)$ depends on θ_0. However, under the null hypothesis, $G\left(X_i, \theta_0\right)$ is completely specified. For B, we may set it as, for example, a vector of ones times a constant. The form of (12) will be intimately related to our proposal regarding how to calibrate the penalization parameter λ.

As before, write $G_i := G\left(X_i, \theta_0\right)$. Under (12), we have

$$\begin{aligned}\sqrt{n}M_n(\theta_n, \gamma) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \exp(W_i'\gamma) + \frac{1}{n}\sum_{i=1}^{n}\exp(W_i'\gamma)G_i B + o_p\left(1\right) \\ &\Rightarrow \mathcal{M}\left(\gamma\right) + \mathbb{E}\left[\exp(W_i'\gamma)G_i B\right] \text{ in } L_\infty(\Gamma).\end{aligned}$$

Then, we can establish that

$$T_n(\theta_n, \lambda) \Rightarrow \sup_{\gamma\in\Gamma}\left[\left|\frac{\mathcal{M}\left(\gamma\right) + \mathbb{E}\left[\exp(W_i'\gamma)G_i B\right]}{s\left(\gamma\right)}\right| - \lambda\left\|\gamma\right\|_1\right] \text{ in } L_\infty(\Lambda), \tag{13}$$

using arguments identical to those to prove Theorem 1.

Define the noncentrality term

$$\kappa\left(\gamma, B\right) := \frac{\mathbb{E}\left[\exp(W_i'\gamma)G_i B\right]}{s(\gamma)} = \frac{\mathbb{E}\left[\exp(W_i'\gamma)G_i B\right]}{\sqrt{\mathbb{E}[U_i^2\exp(2W_i'\gamma)]}}.$$

Define $\mathcal{Q}(\gamma) := \frac{\mathcal{M}(\gamma)}{s(\gamma)}$ and $T(\lambda, B) := \sup_{\gamma \in \Gamma} |\mathcal{Q}(\gamma) + \kappa(\gamma, B)| - \lambda \|\gamma\|_1$. For the test to have a nontrivial power, we need that $\kappa(\gamma^*(B, \lambda), B) \neq 0$ with a positive probability, where $\gamma^*(B, \lambda)$ denotes a (random) maximizer of the stochastic process in (13). Since the penalty affects $\gamma^*(B, \lambda)$ in different ways under the null of $B = 0$ and alternatives of $B \neq 0$, its implication on power of the test is not straightforward to analyze. The subsequent subsection proposes a method to select λ in a more systematic way to increase power.

We discuss some sufficient conditions, under which the presence of penalty increases the power of the test in the setting of the preceding limit experiment. Heuristically, if the unpenalized criterion is maximized at γ near zero, then the maximum would be less affected by the introduction of the ℓ_1 penalty on γ. And a suitable noncentrality function κ, like e.g. a concave function with a unique maximum at zero, can force the maximizing γ closer to zero. That is, the power gain is obtained through the penalization if the limit experiment under the null is maximized at bigger γ than under the alternative. The following lemma is a more formal treatment of the heuristic.

Lemma 2. *Suppose that there exist a unique $\tilde{\gamma}(b)$ maximizing $|\mathcal{Q}(\gamma) + \kappa(\gamma, b)|$ for $b = 0$ or B. If*

$$\|\tilde{\gamma}(B)\|_1 < \|\tilde{\gamma}(0)\|_1 \quad a.s. \tag{14}$$

then, we can find a more powerful test with a strictly positive λ than the test with $\lambda = 0$ in the limit experiment with $B \neq 0$.

The assumption that $\mathcal{Q}(\gamma) + \kappa(\gamma, b)$ is uniquely maximized a.s. on a compact space is standard due to Kim and Pollard (1990, Lemma 2.6), requiring only that the increments of the Gaussian process exhibit non-negligible variances.

An example that may meet the condition (14) is the case where the noncentrality term $\kappa(\gamma, B)$ under the alternative hypothesis induces a sparse solution. This happens when the set of instrumental variables W contains redundant elements. It is similar to the well-known fact that the presence of an irrelevant variable in the linear regression results in loss of power in the tests based on the OLS estimates. Specifically, suppose for simplicity that $W_i = (Z_i, F_i)$ and F_i is a pure noise that is independent of everything else. Then, the noncentrality term can be rewritten as

$$\kappa(\gamma, B) = \frac{\mathbb{E}\left[\exp(Z_i'\gamma_1)G_iB\right]}{\sqrt{\mathbb{E}[U_i^2\exp(2Z_i'\gamma_1)]}}\frac{\mathbb{E}\left[\exp(F_i'\gamma_2)\right]}{\sqrt{\mathbb{E}[\exp(2F_i'\gamma_2)]}} =: \kappa_1(\gamma_1, B)\,\kappa_2(\gamma_2).$$

Then, Jensen's inequality yields that $\kappa_2(\gamma_2) \le 1$ and the equality holds if and only if $\gamma_2 = 0$. If the dimension of F_i is relatively large compared to that of Z_i and the magnitude of $\kappa_2(\gamma_2)$ dominates that of $\kappa_1(\gamma_1, B)$ and $\mathcal{M}(\gamma)$ then $\tilde{\gamma}(B)$ would be closer to zero than $\tilde{\gamma}(0)$, which is a maximizer of a centered Gaussian process of equal marginal variance.

Since analytical derivation is involved in general cases, we provide visual representation of the preceding discussion via some Monte Carlo simulation as below.

Specifically, we generate a random sample with $n = 1000$ from a simple linear regression model with normal random variables. Specifically, we draw independent variables uniformly distributed on $[-1, 1]$, $\varepsilon_{ji},\ j = 1, 4$ and independent standard normal variables $\varepsilon_{ji},\ j = 2, 3$ and generate $X_i = \varepsilon_{1i} + \varepsilon_{2i}/2$, $Y_i = X_i\theta_n + (\varepsilon_{2i} + \varepsilon_{3i})/2$ with $\theta_n = \theta_0 + B\,n^{-1/2}$ with some nonzero constant B, and $Z_i = (W_i, F_i)$, where $W_i = \varepsilon_{1i} - \varepsilon_{4i}$, while F_i is a $(p-1)$-dimensional independent vector uniformly distributed on $[-1, 1]$, that is independent of all the others. Figure 1 plots the "theoretical" power functions of our proposed test, where the power curves are obtained via Monte Carlo simulations with $\Gamma = [-5, 5]^p$. There is only one endogenous regressor here and the three lines in the figure represent the power curves as a function of the penalty level λ for three different values of the dimension p of Z_i, whose first element is strongly correlated to U_i while the others are irrelevant. Thus, $p - 1$ represents the number of irrelevant instruments. The power decreases as the number of irrelevant variables increases when the penalty $\lambda = 0$. This is analogous to the textbook treatment of hypothesis testing with the linear regression model. Next, the power increases gradually up to a certain point as the penalty grows for each p, in line with the preceding discussion. In addition, the power gain from the penalization, that is, the difference between the maximum power and the power at $\lambda = 0$, is bigger for larger p.

4.3 Calibration of λ

The penalty function works differently on how it shrinks the maximizer $\tilde{\gamma}$ under the alternatives. Ideally, it should induce sparse solutions that force zeros for the coefficients of the irrelevant conditioning variable to maximize the power of the test.

Although it is demanding to characterize the optimal choice of λ analytically, we can elaborate on the choice of the penalty parameter λ under the limit of experiments

$$\frac{\mathcal{M}(\gamma) + \mathbb{E}\left[\exp(W_i'\gamma)G_iB\right]}{s(\gamma)},$$

for which we parametrize the size of the deviation by B. Then, our test becomes

$$\mathcal{T}(\lambda, B, \alpha) := 1\left\{\sup_{\gamma\in\Gamma}\left|\frac{\mathcal{M}(\gamma) + \mathbb{E}\left[\exp(W_i'\gamma)G_iB\right]}{s(\gamma)}\right| - \lambda\left\|\gamma\right\|_1 > c_\alpha(\lambda)\right\} \tag{15}$$

for a critical value $c_\alpha(\lambda)$, which is the $(1-\alpha)$ quantile of $\sup_{\gamma\in\Gamma}\left[\frac{|\mathcal{M}(\gamma)|}{s(\gamma)} - \lambda\|\gamma\|_1\right]$.

Let $\mathcal{R}(\lambda, B, \alpha) := \mathbb{E}\left[\mathcal{T}(\lambda, B, \alpha)\right]$ denote the power function of the test under the limit experiment for given λ, B and α, where $0 < \alpha < 1$ is a prespecified level of the test. We propose to select λ by solving the max-min problem:

$$\sup_{\lambda\in\Lambda}\inf_{B\in\mathcal{B}}\mathcal{R}(\lambda, B, \alpha), \tag{16}$$

where Λ is a set of possible values of λ and $\mathcal{B}$ is a set of possible values of B. In some applications, where B is one-dimensional and $|\mathcal{M}(\gamma) + \mathbb{E}\left[\exp(W_i'\gamma)G_iB\right]|$ is stochastically monotone in $|B|$, the inner minimization over $B \in \mathcal{B}$ is simple and easy to characterize. For Λ, we can take a discrete set of possible values of λ, including 0, if suitable. The idea behind (16) is as follows. For each candidate λ, the size of the test is constrained properly because $\mathcal{R}(\lambda, 0, \alpha) \leq \alpha$. Then we look at the least-favorable local power among possible values of B and choose λ that maximizes the least-favorable local power.

To operationalize our proposal, we again rely on a multiplier bootstrap. Define

$$\begin{aligned} M_{n,*,B}(\gamma) &:= \frac{1}{n}\sum_{i=1}^{n}\left(\eta_i^* U_i + \frac{B}{\sqrt{n}} G_i\right)\exp(W_i'\gamma), \\ s^2_{n,*,B}(\gamma) &:= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\eta_i^* U_i + \frac{B}{\sqrt{n}} G_i\right)\exp(W_i'\gamma)\right]^2, \\ Q_{n,*,B}(\gamma) &:= \sqrt{n}\frac{|M_{n,*,B}(\gamma)|}{s_{n,*,B}(\gamma)}, \end{aligned} \tag{17}$$

where η_i^* is drawn from $N(0,1)$ and independent from data $\{(X_i, W_i) : i = 1, \ldots, n\}$. The quantities above are just shifted versions of (6).[6] For each bootstrap replication r, let

$$T^{(r)}_{n,*,B}(\lambda) := \sup_{\gamma\in\Gamma}\left[Q^{(r)}_{n,*,B}(\gamma) - \lambda\|\gamma\|_1\right]. \tag{18}$$

Then the critical value $c_\alpha(\lambda)$ is approximated by $c^*_\alpha(\lambda)$, the $(1-\alpha)$-quantile of $T_{n,*,0}(\lambda) \equiv T_{n,*}(\lambda)$. Once $c^*_\alpha(\lambda)$ is obtained, $\mathcal{R}(\lambda, B, \alpha)$ is approximated by

$$\frac{1}{R}\sum_{r=1}^{R} 1\left\{T^{(r)}_{n,*,B}(\lambda) > c^*_\alpha(\lambda)\right\}. \tag{19}$$

Similarly, let $\widehat{\lambda}$ denote a maximizer of $\min_{B\in\mathcal{B}}\mathcal{R}_n(\lambda, B, \alpha)$ over Λ, where

$$\mathcal{R}_n(\lambda, B, \alpha) := \Pr{}^*\{T_{n,*,B}(\lambda) > c^*_\alpha(\lambda)\},$$

and $\Pr^*$ denotes the conditional probability of the bootstrap sample given the sample. Recall $T(\lambda) = \sup_{\gamma\in\Gamma}\left[\frac{|\mathcal{M}(\gamma)|}{s(\gamma)} - \lambda\|\gamma\|_1\right]$. Let F_λ and F^*_λ denote the distribution function of $T(\lambda)$ and that of $T_{n,*}(\lambda)$ conditional on the sample $\mathcal{X}_n$, respectively. We make the following

[6]Notice that we use a shifted version of $s^2_{n,*,B}(\gamma)$ instead of $s^2_{n,*,0}(\gamma)$ when we define (17). This is because we would like to mimic more closely the finite-sample distribution of the test statistic under the alternative.

regularity condition on F_λ. Let $c_\alpha(\lambda)$ denote the $(1-\alpha)$-quantile of $T(\lambda)$ and $A_\alpha = \{(x,\lambda) : |c_\alpha(\lambda) - x| \le c, \lambda \in \Lambda\}$ for some positive constant c.

Assumption 2. *The partial derivative $\partial F_\lambda(x)/\partial x$ is positive and continuous on A_α.*

When $\lambda = 0$, $T(\lambda)$ is the maximum of a centered Gaussian process. In view of the well-known anti-concentration property of the maximum of a centered Gaussian process (see, e.g., Chernozhukov, Chetverikov, and Kato, 2014), it is not restrictive to assume the presence of a density (with respect to Lebesgue measure) of T at $\lambda = 0$. Furthermore, the extension to noncentral Gaussian process is given by Chernozhukov, Chetverikov, and Kato (2016). Thus, we may assume $T(\lambda)$ has a bounded density for all $\lambda \in \Lambda$. Thus, its Lipschitz property in λ implies that its distribution function indexed by λ is also Lipschitz in λ because

$$
\begin{aligned}
\Pr\{T(\lambda_2) \le x\} &= \Pr\{T(\lambda_1) \le x + T(\lambda_1) - T(\lambda_2)\} \\
&\le \Pr\{T(\lambda_1) \le x + c|\lambda_1 - \lambda_2|\} \le \Pr\{T(\lambda_1) \le x\} + C|\lambda_1 - \lambda_2|,
\end{aligned}
$$

where constants c and C depend only on the Lipschitz constant and the densities, respectively. Then, the continuity assumption in Assumption 2 is reasonable.

To sum up, we formally define optimal λ assuming $\mathcal{R}(\lambda, B, \alpha)$ is continuous on $\Lambda \times \mathcal{B}$, which is possibly set-valued, in the following way.

Definition 1. Let $\Lambda_0 \subset \Lambda$ denote a set of the global solution in (16) so that

$$
\min_{B\in\mathcal{B}} \mathcal{R}(\lambda_0, B, \alpha) \ge \min_{B\in\mathcal{B}} \mathcal{R}(\lambda, B, \alpha) \quad \text{for any } \lambda \in \Lambda \text{ and } \lambda_0 \in \Lambda_0,
$$

where the inequality is strict for $\lambda \notin \Lambda_0$.

The optimal set Λ_0 depends on the set Λ of possible values of λ, the set $\mathcal{B}$ of possible values of B in (12), and the level α of the test.

The following theorem shows that the bootstrap critical values $c^*_\alpha(\lambda)$ are uniformly consistent for $c_\alpha(\lambda)$. Then, it establishes consistency of our proposed calibration method in the sense that $d(\widehat{\lambda}, \Lambda_0) \xrightarrow{p} 0$, where $d(x, X) := \inf\{|x - y| : y \in X\}$.

Theorem 4. *Let Assumptions 1 and 2 hold. Then, $c^*_\alpha(\lambda) \xrightarrow{p} c_\alpha(\lambda)$ uniformly in Λ. Furthermore, $d(\widehat{\lambda}, \Lambda_0) \xrightarrow{p} 0$ and $\Pr\{T_n(\widehat{\lambda}) \geq c^*_\alpha(\widehat{\lambda})\} \to \alpha$.*

5 Implementation

In this section, we summarize our inference procedure and provide pseudo-code in Algorithm 1, which describes how to conduct the pointwise test of $H_0 : \theta = \bar{\theta}$.

In addition to the usual input such as the confidence level α and the number of bootstrap replications R, we need to specify the search space $\Gamma \subset \mathbb{R}^p$, the grid for penalty levels $\Lambda \subset \mathbb{R}_+$, and the set of local alternatives $\mathcal{B} \subset \mathbb{R}^d$. In our numerical work, we choose $\Gamma = [-a, a]^p$ with some constant a (e.g., $a = 1, 5$). For Λ, we recommend excluding $\lambda = 0$ if p is somewhat large (e.g., $p > 5, 10$). Regarding $\mathcal{B}$, it is necessary to know the structure of the inference problem in hand. In our Monte Carlo experiments as well as empirical applications, $d = 1$ and we need to choose $\mathcal{B}$ as a subset of $\mathbb{R}$. We provide details in Section 7.

We now make several remarks on computation of $T_n(\bar{\theta}, \lambda)$, $T^{(r)}_{n,*}(\bar{\theta}, \lambda)$, and $T^{(r)}_{n,*,B}(\lambda)$ in Algorithm 1. For the first empirical application to Yogo (2004) data set in Section 7, we use both the grid search (GS) and the `particleswarm` particle swarm optimization (PSO) solver available in `Matlab`.[7] We find our tests computed using GS and PSO on the set $\Gamma = [-a, a]^4$ perform similarly. For the second empirical application in Section B, we only use the `particleswarm` for optimization over $\Gamma = [-a, a]^{21}$. PSO is a stochastic population-based optimization method proposed by Kennedy and Eberhart (1995). It conducts gradient-free global searches and has been successfully used in economics

[7]Specifically, the `particleswarm` solver is included in the global optimization toolbox software of `Matlab`.

Algorithm 1: Inference for $H_0 : \theta_0 = \bar{\theta}$ versus $H_1 : \theta_0 \neq \bar{\theta}$ under the conditional moment model $\mathbb{E}[g(X_i, \theta_0)|W_i] = 0$

Input: the data set $\{(X_i, W_i) : i = 1, \ldots, n\}$, the hypothesized parameter $\bar{\theta}$, the search space $\Gamma \subset \mathbb{R}^p$, the grid for penalty levels $\Lambda \subset \mathbb{R}_+$.

for $k = 1, 2, \ldots, p$ **do**
  Calculate $\bar{W}_k = \frac{1}{n}\sum_{i=1}^n W_{ki}$ and $s_n(W_k) = \frac{1}{n-1}\sum_{i=1}^n (W_{ki} - \bar{W}_k)^2$.
  Transform $W_{ki} \leftarrow \tan^{-1}\big((W_{ki} - \bar{W}_k)/s_n(W_k)\big)$.
end
Declare variables $U_i \leftarrow g(X_i, \bar{\theta})$.
Declare a function $\gamma \mapsto Q_n(\bar{\theta}, \gamma)$ according to (3).
for *each* λ *in* Λ **do**
  Calculate $T_n(\bar{\theta}, \lambda) = \sup_{\gamma \in \Gamma}\big[Q_n(\bar{\theta}, \gamma) - \lambda\|\gamma\|_1\big]$.
end

Output: the set of test statistics $\{T_n(\bar{\theta}, \lambda) : \lambda \in \Lambda\}$.

Input for bootstrap test: the number of bootstrap replications R, the set of local alternatives $\mathcal{B} \subset \mathbb{R}^d$, the significance level α.

for $r = 1, 2, \ldots, R$ **do**
  Generate $\{\eta_i^* : i = 1, \ldots, n\}$ from i.i.d. $N(0, 1)$ independently of the data.
  for *each* λ *in* Λ **do**
    Declare a function $\gamma \mapsto Q_{n,*}(\bar{\theta}, \gamma)$ according to (6).
    Calculate $T_{n,*}^{(r)}(\bar{\theta}, \lambda) = \sup_{\gamma \in \Gamma}\big[Q_{n,*}(\bar{\theta}, \gamma) - \lambda\|\gamma\|_1\big]$.
    for *each* B *in* $\mathcal{B}$ **do**
      Declare a function $\gamma \mapsto Q_{n,*,B}(\gamma)$ according to (17).
      Calculate $T_{n,*,B}^{(r)}(\lambda) = \sup_{\gamma \in \Gamma}[Q_{n,*,B}(\gamma) - \lambda\|\gamma\|_1]$.
    end
  end
end
for *each* λ *in* Λ **do**
  Declare $cv_\alpha^*(\lambda) \leftarrow (1-\alpha)$-quantile of $\{T_{n,*}^{(r)}(\bar{\theta}, \lambda) : r = 1, \ldots, R\}$.
  for *each* B *in* $\mathcal{B}$ **do**
    Compute the simulated local power as
    $\mathcal{R}_n(\lambda, B, \alpha) = \frac{1}{R}\sum_{r=1}^R 1\{T_{n,*,B}^{(r)}(\lambda) > cv_\alpha^*(\lambda)\}$.
  end
end
Select the optimal penalty level as $\hat{\lambda}(\bar{\theta}) \in \arg\max_{\lambda \in \Lambda} \min_{B \in \mathcal{B}} \mathcal{R}_n(\lambda, B, \alpha)$.
Compute the p-value as $p_*(\bar{\theta}, \hat{\lambda}(\bar{\theta})) = \frac{1}{R}\sum_{r=1}^R 1\{T_{n,*}^{(r)}(\bar{\theta}, \hat{\lambda}(\bar{\theta})) > T_n(\bar{\theta}, \hat{\lambda}(\bar{\theta}))\}$.

Output: Reject H_0 at the significance level α iff $p_*(\bar{\theta}, \hat{\lambda}(\bar{\theta})) < \alpha$.

(for example, see Qu and Tkachenko, 2016).[8] Hence, PSO can be viewed as a more refined approach to global optimization than simple grid search. For the Monte Carlo experiments that mimic Yogo (2004) data set in Section 8, we only apply GS on the set $\Gamma = [-a, a]^p,\ p = 4, 6$. This is because PSO is based on a heuristic procedure and requires careful monitoring to check whether it produces reasonable solutions. Furthermore, it is easy to vectorize using GS but harder using the `particleswarm` solver. In short, it was too costly to monitor the `particleswarm` solver in the Monte Carlo experiments; however, it was possible with empirical applications because we did not have to regenerate data.

We end this section by recalling that the confidence interval for θ_0 can be constructed by inverting a pointwise test of $H_0 : \theta_0 = \bar{\theta}$. For this purpose, one could generate R collections of $\{\eta_i^* : i = 1, \ldots, n\}$ and use the same collections across different values of $\bar{\theta}$ to reduce the random noise in bootstrap inference.

6 Inference with Pre-Estimated Parameters

Partition $\theta = (\theta_1', \theta_2')'$ and $\theta_0 = \left(\theta_{1,0}', \theta_{2,0}'\right)'$. We now consider inference for $\theta_{1,0}$. We assume that for each θ_1, there exists a preliminary estimator $\widehat{\theta}_2(\theta_1) = \psi_n\left(\{X_i, W_i\}_{i=1}^n\right)$ of $\theta_2(\theta_1)$, so that $\theta_{2,0} = \theta_2(\theta_{1,0})$. For example, suppose that $g\left(X_i, \theta\right)$ can be written as $g\left(X_i, \theta\right) = g_1\left(X_i, \theta_1\right) - \theta_2$. Then, $\theta_{2,0} = \mathbb{E}\left[g_1\left(X_i, \theta_{1,0}\right)\right]$, thereby yielding the following estimator of θ_2 given θ_1:

$$\widehat{\theta}_2(\theta_1) = n^{-1} \sum_{i=1}^{n} g_1\left(X_i, \theta_1\right). \tag{20}$$

In what follows, we assume standard regularity conditions on $\widehat{\theta}_2(\theta_1)$. Let Θ_1 denote the

[8] It is possible to adopt the two-step approach used in Qu and Tkachenko (2016). That is, we start with the PSO solver, followed by multiple local searches. Further, the genetic algorithm (GA) can be used in the first step instead of PSO and both GA and PSO methods can be compared to check whether a global solution is obtained. We do not pursue these refinements to save the computational times of bootstrap inference.

parameter space for θ_1, which is a compact set with a non-empty interior. Let $\|a\|$ denote the Euclidean norm of a vector a.

Assumption 3. *Suppose that there exists a $\sqrt{n}$-consistent estimator $\widehat{\theta}_2(\theta_1)$ of $\theta_2(\theta_1)$ that has the following representation: uniformly in $\theta_1 \in \Theta_1$, which is compact and has a non-empty interior,*

$$\sqrt{n}\left(\widehat{\theta}_2(\theta_1) - \theta_2(\theta_1)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_{ni}(\theta_1) + o_p(1),$$

where $\{\psi_{ni}(\theta_1) = (\zeta_{ni}(\theta_1), U_i)\}$ is a strictly stationary ergodic mds array. Also, let $V_{\psi(\theta_1)} := \lim_{n\to\infty} \mathbb{E}[\psi_{ni}(\theta_1)\psi_{ni}(\theta_1)']$ be positive definite. Furthermore, assume that there exists $G(x,\theta)$ such that $\mathbb{E}\|G(X_i,\theta_0)\| < \infty$, $\theta \mapsto G(X_i,\theta)$ is continuous at θ_0 almost surely, and

$$g(X_i,\theta) - g(X_i,\theta_0) - G(X_i,\theta_0)'(\theta-\theta_0) = o_p(\|\theta-\theta_0\|), \tag{21}$$

and that $\theta_1 \mapsto \theta_2(\theta_1)$ is continuously differentiable.

Define $\widehat{U}_i(\theta_1) := g[X_i, \{\theta_1, \widehat{\theta}_2(\theta_1)\}]$, $\widehat{U}_i := \widehat{U}_i(\theta_{1,0})$. We introduce the following demeaned statistics[9]

$$\begin{aligned}
\widehat{M}_n(\theta_1,\gamma) &:= \frac{1}{n}\sum_{i=1}^{n} \widehat{U}_i(\theta_1)\left(\exp(W_i'\gamma) - \frac{1}{n}\sum_{j=1}^{n}\exp(W_j'\gamma)\right),\\
\widehat{s}_n^2(\theta_1,\gamma) &:= \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{U}_i(\theta_1)\left(\exp(W_i'\gamma) - \frac{1}{n}\sum_{j=1}^{n}\exp(W_j'\gamma)\right)\right]^2,\\
\widehat{Q}_n(\theta_1,\gamma) &:= \sqrt{n}\frac{|\widehat{M}_n(\theta_1,\gamma)|}{\widehat{s}_n(\theta_1,\gamma)},
\end{aligned}$$

and the test statistic

$$\widehat{T}_n(\theta_1,\lambda) := \sup_{\gamma\in\Gamma}\left[\widehat{Q}_n(\theta_1,\gamma) - \lambda\|\gamma\|_1\right]. \tag{22}$$

[9]The test statistic $\widehat{Q}_n(\theta_1,\gamma)$ is effectively based on an orthogonalized residual after estimating the unconditional mean.

Partition $G(X_i, \theta) = \left[G_1(X_i, \theta)', G_2(X_i, \theta)'\right]'$, corresponding to the partial derivatives with respect to θ_1 and θ_2. Suppressing the dependence on θ_1 in the notation when $\widehat{T}_n$ or ζ_{ni} is evaluated at $\theta_{1,0}$, we obtain the following result.

Theorem 5. *Let Assumptions 1 and 3 hold. Then,*

$$\widehat{T}_n(\lambda) \Rightarrow \sup_{\gamma \in \Gamma} \left[\frac{\left| \bar{\mathcal{M}}(\gamma) + Z' \mathrm{cov}\left[G_2(X_i, \theta_0), \exp(W_i'\gamma)\right] \right|}{s(\gamma)} - \lambda \|\gamma\|_1 \right] \quad \text{in } L_\infty(\Lambda),$$

where $(Z, \bar{\mathcal{M}}(\gamma))$ is a centered Gaussian random vector with $\mathbb{E}[ZZ'] = \lim_{n\to\infty} \mathbb{E}[\zeta_{ni}\zeta_{ni}']$ and $\mathbb{E}\left[Z\bar{\mathcal{M}}(\gamma)\right] = \lim_{n\to\infty} \mathbb{E}\left[U_i\zeta_{ni}(\exp(W_i'\gamma) - \mathbb{E}\exp(W_i'\gamma))\right]$ for each γ and $\bar{\mathcal{M}}(\gamma)$ is a centered Gaussian process such that $\mathbb{E}\bar{\mathcal{M}}(\gamma_1)\bar{\mathcal{M}}(\gamma_2) = \mathbb{E}U_i^2(\exp(W_i'\gamma_1) - \mathbb{E}\exp(W_i'\gamma_1))(\exp(W_i'\gamma_2) - \mathbb{E}\exp(W_i'\gamma_2))$.

In the presence of pre-estimates in the test statistic, the multiplier bootstrap in (6) is not valid. To develop valid inference, we now describe how to modify the multiplier bootstrap by exploiting the influence function ζ_{ni}. To ease notation, define

$$\widehat{G}_{2i} := G_2\left(X_i, \left(\theta_{1,0}, \widehat{\theta}_2\right)\right)$$
$$\overline{W}_{ni}(\gamma) := \exp(W_i'\gamma) - \frac{1}{n}\sum_{j=1}^n \exp(W_j'\gamma)$$

Let

$$\begin{aligned} \widehat{M}_{n,*}(\gamma) &:= \frac{1}{n}\sum_{i=1}^n \eta_i^* \left\{ \widehat{U}_i \overline{W}_{ni}(\gamma) + \widehat{\zeta}_{ni}' \frac{1}{n}\sum_{j=1}^n \widehat{G}_{2j}\overline{W}_{nj}(\gamma) \right\}, \\ \widehat{s}^2_{n,*}(\gamma) &:= \frac{1}{n}\sum_{i=1}^n \left[\eta_i^* \left\{ \widehat{U}_i \overline{W}_{ni}(\gamma) + \widehat{\zeta}_{ni}' \frac{1}{n}\sum_{j=1}^n \widehat{G}_{2j}\overline{W}_{nj}(\gamma) \right\} \right]^2, \end{aligned} \tag{23}$$

where $\widehat{\zeta}_{ni}$ denotes a consistent estimator of ζ_{ni}. Then, we proceed with these modified quantities, as in Section 3. That is, to implement the bootstrap, we need to obtain an explicit plug-in formula for the influence function $\zeta_{ni}(\theta_1)$ in Assumption 3, similar to $g_1(X_i, \theta_1) -$

$\frac{1}{n}\sum_{i=1}^{n} g_1(X_i, \theta_1)$ from (20).[10]

6.1 Choice of Penalty

We start with a sequence of local alternatives $\theta_{1n} = \theta_{1,0} + B/\sqrt{n}$. Then, expressing the hypothesized value of θ_{1n} explicitly, we write the corresponding statistics by

$$\begin{aligned}
\widehat{M}_n(\theta_{1n}, \gamma) &:= \frac{1}{n}\sum_{i=1}^{n} \widehat{U}_i\left(\theta_{1n}\right) \overline{W}_{ni}(\gamma), \\
\widehat{s}_n^2(\theta_{1n}, \gamma) &:= \frac{1}{n}\sum_{i=1}^{n} \left[\widehat{U}_i\left(\theta_{1n}\right) \overline{W}_{ni}(\gamma)\right]^2, \\
\widehat{Q}_n(\theta_{1n}, \gamma) &:= \sqrt{n}\frac{|\widehat{M}_n(\theta_{1n}, \gamma)|}{\widehat{s}_n(\theta_{1n}, \gamma)},
\end{aligned}$$

and the test statistic

$$\widehat{T}_n\left(\theta_{1n}\right) := \sup_{\gamma\in\Gamma}\left[\widehat{Q}_n(\theta_{1n}, \gamma) - \lambda_n\|\gamma\|_1\right]. \tag{24}$$

Define $g_i(\theta_1, \theta_2) := g[X_i, \{\theta_1, \theta_2\}]$ and partition $G(X_i, \{\theta_1, \theta_2\}) = [G_{1i}(\theta_1, \theta_2)', G_{2i}(\theta_1, \theta_2)']'$ as before. The limit of the test statistic $\widehat{T}_n\left(\theta_{1n}\right)$ can be easily obtained by modifying the proof of Theorem 5. Specifically, we note that under the additional assumption that $\zeta_{ni}(\theta_1)$ is

[10]Instead, one could implement alternative bootstrap without using the explicit formula of the influence function as in Chen, Linton, and Van Keilegom (2003). We opted not to consider this in this paper.

differentiable with respect to θ_1,

$$\begin{aligned}
\sqrt{n}\widehat{M}_n(\theta_{1n},\gamma) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\left(\theta_{1n},\theta_2(\theta_{1n})\right)\overline{W}_{ni}(\gamma) \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_{ni}(\theta_{1n})'\frac{1}{n}\sum_{j=1}^{n} G_{2j}\left(\theta_{1n},\theta_2(\theta_{1n})\right)\overline{W}_{nj}(\gamma) + o_p\left(1\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\left(\theta_0\right)\overline{W}_{ni}(\gamma) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_{ni}'\frac{1}{n}\sum_{j=1}^{n} G_{2j}\left(\theta_0\right)\overline{W}_{nj}(\gamma) \\
&\quad + B'\frac{1}{n}\sum_{i=1}^{n}\left[G_{1i}\left(\theta_0\right) + \frac{\partial\theta_2\left(\theta_{1,0}\right)'}{\partial\theta_1}G_{2i}\left(\theta_0\right)\right]\overline{W}_{ni}(\gamma) + o_p\left(1\right),
\end{aligned}$$

using the fact that $n^{-1}\sum_{i=1}^{n}\frac{\partial\zeta_{ni}(\theta_{1,0})}{\partial\theta_1} = o_p(1)$. Thus, the noncentrality term is determined by the probability limit of $B'\mathbb{E}\left[\omega_i(\gamma)\right]$, where

$$\omega_i(\gamma) := \left[G_{1i}\left(\theta_0\right) + \frac{\partial\theta_2\left(\theta_{1,0}\right)'}{\partial\theta_1}G_{2i}\left(\theta_0\right)\right]\overline{W}_{ni}(\gamma).$$

As shorthand notation, let $G_{1i} := G_1\left(X_i,\theta_0\right)$ and $G_{2i} := G_2\left(X_i,\theta_0\right)$. We now adjust (15) in Section 4.3 as follows: let

$$\begin{aligned}
&\mathcal{T}\left(\lambda,B\right) \\
&= 1\left\{\sup_{\gamma\in\Gamma}\left|\frac{\bar{\mathcal{M}}\left(\gamma\right) + Z'\mathrm{cov}\left[G_{2i},\exp\left(W_i'\gamma\right)\right] + B'\mathbb{E}[\omega_i(\gamma)]}{s\left(\gamma\right)}\right| - \lambda\left\|\gamma\right\|_1 > c_\alpha(\lambda)\right\}
\end{aligned} \tag{25}$$

for a critical value $c_\alpha(\lambda)$ and $\mathcal{R}\left(\lambda,B\right) = \mathbb{E}[\mathcal{T}\left(\lambda,B\right)]$. Then, as before, choose λ by solving (16). To implement this procedure, we modify the steps in Section 4.3 with

$$\begin{aligned}
\widehat{M}_{n,*,B}(\gamma) &:= \frac{1}{n}\sum_{i=1}^{n}\left[\eta_i^*\left\{\widehat{U}_i\overline{W}_{ni}(\gamma) + \widehat{\zeta}_{ni}'\frac{1}{n}\sum_{j=1}^{n}\widehat{G}_{2j}\overline{W}_{nj}(\gamma)\right\} + \frac{B}{\sqrt{n}}\widehat{\omega}_i(\gamma)\right], \\
\widehat{s}^2_{n,*,B}(\gamma) &:= \frac{1}{n}\sum_{i=1}^{n}\left[\eta_i^*\left\{\widehat{U}_i\overline{W}_{ni}(\gamma) + \widehat{\zeta}_{ni}'\frac{1}{n}\sum_{j=1}^{n}\widehat{G}_{2j}\overline{W}_{nj}(\gamma)\right\} + \frac{B}{\sqrt{n}}\widehat{\omega}_i(\gamma)\right]^2,
\end{aligned}$$

where $\widehat{\omega}_i(\gamma) := \left(G_{1i} + \frac{\partial \widehat{\theta}_2(\theta_{1,0})}{\partial \theta_1}' \widehat{G}_{2i}\right) \overline{W}_{ni}(\gamma)$. Then, the remaining steps are identical to those in Section 4.3.

7 Inferring the Elasticity of Intertemporal Substitution

In this section, we revisit Yogo (2004) and conduct inference on the elasticity of intertemporal substitution (EIS). We investigate the case of the annual US series (1891–1995) used in Yogo (2004), focusing on $U_t(\theta) = \Delta c_t - \theta_2 - \theta_1 r_t$, where Δc_t represents the consumption growth in year t and r_t denotes the real interest rate. The parameter of interest is EIS, denoted by θ_1. The instruments W_t consist of the two-period lags of the nominal interest rate, inflation, consumption growth, and the log dividend-price ratio. Before applying our method, we studentized each instrument and then applied the transformation $\tan^{-1}(\cdot)$. The transformed instruments are denoted by $\widetilde{W}_t$. This ensures that each component of $\widetilde{W}_t$ is bounded and comparable in scale. The data consist of $\{(\Delta c_t, r_t, W_t) : t = 1, \ldots, n\}$, where the time span is $n = 105$.

To perform inference on θ_1 in the presence of θ_2, we use the demeaned version of the generalized residuals, defined as follows:

$$\widehat{U}_t(\theta_1) = \left(\Delta c_t - \frac{1}{n}\sum_{t=1}^{n} \Delta c_t\right) - \theta_1 \left(r_t - \frac{1}{n}\sum_{t=1}^{n} r_t\right).$$

Following the notation in Section 6, we have in this example,

$$\begin{aligned}
\theta_2(\theta_1) &= \mathbb{E}\left[\Delta c_t\right] - \theta_1 \mathbb{E}\left[r_t\right], \\
\widehat{\theta}_2(\theta_1) &= \frac{1}{n}\sum_{t=1}^{n} \Delta c_t - \theta_1 \frac{1}{n}\sum_{t=1}^{n} r_t, \\
\eta_{nt}(\theta_1) &= \left(\Delta c_t - \mathbb{E}[\Delta c_t]\right) - \theta_1 \left(r_t - \mathbb{E}[r_t]\right), \\
\widehat{\eta}_{nt}(\theta_1) &= \widehat{U}_t(\theta_1).
\end{aligned}$$

Then, because $G_{2t} = -1$, adopting (23) yields the following multiplier bootstrap:

$$\widehat{M}_{n,*}(\gamma) = \frac{1}{n}\sum_{t=1}^{n} \eta_t^* \widehat{U}_t(\theta_1)\left\{\exp(\widetilde{W}_t'\gamma) - \frac{1}{n}\sum_{t=1}^{n}\exp(\widetilde{W}_t'\gamma)\right\},$$
$$\widehat{s}^2_{n,*}(\gamma) = \frac{1}{n}\sum_{t=1}^{n}\left[\eta_t^* \widehat{U}_t(\theta_1)\left\{\exp(\widetilde{W}_t'\gamma) - \frac{1}{n}\sum_{t=1}^{n}\exp(\widetilde{W}_t'\gamma)\right\}\right]^2. \tag{26}$$

Furthermore, since $G_{1t} = -r_t$, we use the following to calibrate the optimal λ:

$$\widehat{M}_{n,*,B}(\gamma) = \frac{1}{n}\sum_{t=1}^{n}\left\{\eta_t^* \widehat{U}_t(\theta_1) - \frac{B}{\sqrt{n}}(r_t - \bar{r})\right\}\left\{\exp(\widetilde{W}_t'\gamma) - \frac{1}{n}\sum_{t=1}^{n}\exp(\widetilde{W}_t'\gamma)\right\},$$
$$\widehat{s}^2_{n,*,B}(\gamma) = \frac{1}{n}\sum_{t=1}^{n}\left[\left\{\eta_t^* \widehat{U}_t(\theta_1) - \frac{B}{\sqrt{n}}(r_t - \bar{r})\right\}\left\{\exp(\widetilde{W}_t'\gamma) - \frac{1}{n}\sum_{t=1}^{n}\exp(\widetilde{W}_t'\gamma)\right\}\right]^2,$$

where $\bar{r} = n^{-1}\sum_{t=1}^{n} r_t$.

Computation of the Test Statistic In Table 1, we report the values of the max statistic

$$T(\lambda, a) = \sup_{\gamma\in[-a,a]^4}\left[\widehat{Q}_n(\bar{\theta}_1, \gamma) - \lambda\|\gamma\|_1\right],$$

evaluated at $\bar{\theta}_1 = -0.028$, while varying the penalty level $\lambda \in \{.3, .2, .1, 0\}$ and the domain constant $a \in \{1, 2, 3, 4, 5\}$. This value of $\bar{\theta}_1$ is chosen as the 2SLS estimate from the data. To ensure the accuracy of the algorithm, the max statistic is computed with a swarm size of 5000, which spans the search space $[-a, a]^4$ uniformly at random.

Computation Time The swarm size of 5000, used to compute $T(\lambda, a)$ across (λ, a)-values in Table 1, enables highly accurate optimization. However, it may be inappropriate for practical applications. Therefore, we examine the impact of λ and a on computation time based on a more realistic swarm size of 200. In light of the inherent randomness in swarm generation, we regenerate the swarm randomly and recompute the max statistic until it achieves at least 95% of the maximum values reported in Table 1. That is, we measure the computation times under the constraint that the optimization error does not exceed 5%. In Table 2, we report the means and standard deviations of the computation times across various (λ, a)-values.

Tables 1 and 2 reveal three important observations. First, the average computation time tends to increase as λ decreases across all a values. This aligns with the intuition that introducing the penalty shrinks the feasible space for γ, thereby leading to shorter search times for the maximum. Second, the computation times are numerically most stable when $\lambda = 0.3$ is applied. The impact of a on numerical instability tends to increase with a and is greatest when λ is set to 0. Lastly, the value of the unpenalized statistic is more sensitive to the choice of a than it is with penalization. This suggests the importance of implicit search space selection in Bierens' max test with $\lambda = 0$, as it may significantly impact the outcomes of the analysis. Our method is motivated by tuning a scalar parameter λ rather than $\Gamma = [-a, a]^4$, which makes it suitable to optimize in a data-driven manner.

Calibration of Optimal Penalty The optimal λ in $\Lambda = \{0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0\}$ is calibrated as in (16), based on a single local alternative $\mathcal{B} = \{2\}$. The local alternative can be set to a singleton since $\eta_t^* \sim N(0, 1)$ is symmetrically distributed about zero, leading to an increase in power with the absolute value of B. We set α to 0.1 and use the grid $\Theta_1 = \{-0.6, -0.4, -0.2, \ldots, 0.6\}$ for hypothesized values of $\theta_{1,0}$. For each $\lambda \in \Lambda$ and $\theta_{1,0} \in \Theta_1$, the local power in (19) is simulated following the steps outlined in Section 6.1.

We use two algorithms for computing the max statistic to assess the sensitivity to the choice of the algorithm. The PSO algorithm searches for the global maximum in the entire space $\Gamma = [-5, 5]^4$. In each computation, a swarm with a size of 200 is generated uniformly at random over Γ. The grid search (GS) method calculates the maximum over a discretized grid $\Gamma' \subset [-5, 5]^4$, chosen to minimize potential power loss while leveraging a vectorized algorithm to avoid loops used in the PSO algorithm and other optimization methods. The tuning parameter in the grid search corresponds to the choice of the grid Γ'. We opted for the equi-spaced grid $\Gamma' = \{0.5j : -10 \leq j \leq 10\}^4 = \{-5, -4.5, \ldots, 5\}^4$, which consists of $21^4 = 194481$ points. The numbers of bootstrap replications are $R_{\text{PSO}} = 1000$ and $R_{\text{GS}} = 5000$ for the PSO algorithm and the GS method, respectively.

Figure 2 shows the local powers simulated using these algorithms. Panel A displays the average local powers over Θ_1 across different values of λ, while Panel B exhibits the local powers at a specific $\theta_{1,0} = 0$. From Panel A, we note that the PSO algorithm yields overall higher powers than the GS method, which operates on the discrete Γ'. In Panel B, for both algorithms, we observe a rise in the local power within a small range around $\lambda = 0$, followed by leveling off at approximately $\lambda = 0.2$ or $\lambda = 0.3$. Since this pattern also prevails for the other values in Θ_1, we choose the optimal λ as the maximizer of the average local powers in Panel A. The selected λ are 0.3 for the PSO algorithm and 0.2 for the GS method, respectively. These values are used for the optimally penalized test of $H_0 : \theta_1 = \theta_{1,0}$ uniformly across the hypothesized values.

Table 3 presents the confidence intervals based on our optimally penalized and unpenalized tests using each algorithm. These confidence intervals are constructed by inverting the testing of $H_0 : \theta_1 = \theta_{1,0}$, where the hypothesized values of $\theta_{1,0}$ range over $\{0.01j : -100 \leq j \leq 100\}$. It shows that the optimally penalized test yields significantly narrower confidence intervals than the unpenalized test, irrespective of the chosen algorithm.

Yogo (2004) commented that "there appears to be identification failure for the annual U.S. series." Indeed, the 95% confidence interval from the Anderson-Rubin (AR) test was $[-0.49, 0.46]$ and those from the Lagrange multiplier (LM) test and the conditional likelihood ratio (LR) tests were $[-\infty, \infty]$ (see Table 3 of Yogo, 2004).[11] Our penalized test

[11] If each instrument is used separately for the AR test, the resulting 95% confidence intervals are as follows: (i) $[-\infty, \infty]$ with the nominal interest rate as an instrument; (ii) $[-0.29, 0.28]$ with inflation; (iii) $[-\infty, \infty]$ with consumption growth; (iv) $[-\infty, -0.12] \cup [0.02, \infty]$ with log dividend-price ratio, where the instruments are twice lagged in all cases. The confidence interval using inflation is similar to ours, but its length is more than 25%

provides tighter confidence intervals than any of these similar tests based on unconditional moment restrictions, suggesting that conditional moment restrictions can be more informative than an arbitrarily selected set of unconditional moment restrictions.

In a nutshell, we demonstrate that a seemingly uninformative set of instruments can provide an informative inference result if one strengthens unconditional moment restrictions by making them infinite-dimensional conditional moment restrictions with the aid of penalization.

8 Monte Carlo Experiments

To examine the efficacy of our method in an empirically relevant context, we conduct a series of experiments based on the annual U.S. series (1891-1995) used in Yogo (2004). The main purpose of the experiments is to assess the finite-sample size and power properties of our proposed method in comparison to those of existing methods.

The simulated series is denoted by $\{(\Delta c_t^*, r_t^*, W_t^*)\}_{t=1}^n$ with a $*$-superscript, while the original series is denoted by $\{(\Delta c_t, r_t, W_t)\}_{t=1}^n$, as in Section 7. We consider the following data-generating process in our experiments:

$$\begin{aligned} r_t^* &= (1 \quad W_t^{*\prime})\pi_0 + \bar{\pi} f(W_t^*) + v_t^*, \\ \Delta c_t^* &= (1 \quad r_t^*)\theta_0 + u_t^*, \\ \begin{pmatrix} u_t^* \\ v_t^* \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_u\sigma_v\rho \\ \sigma_u\sigma_v\rho & \sigma_v^2 \end{pmatrix}\right]. \end{aligned} \tag{27}$$

The linear coefficients θ_0 and π_0 are taken from their respective estimates in the original series. Specifically, $\pi_0 = (\pi_{0,0}, \dots, \pi_{4,0})'$ is computed from the first-stage OLS regression $r_t = (1 \quad W_t')\pi_0 + \hat{v}_t$, and $\theta_0 = (\theta_{0,0}, \theta_{1,0})'$ is obtained from the 2SLS regression $\Delta c_t =$

larger than our optimal confidence interval $[-0.30, 0.15]$ using the PSO algorithm and more than 15% larger than the optimal confidence interval using the GS method.

$(1 \quad r_t)\theta_0 + \hat{u}_t$, where W_t are used as instruments.

We specify the nonlinear term $f(W_t^*) = \delta_{0,0} + \delta_{1,0}' W_t^* + \delta_{2,0}'(W_t^* \odot W_t^*)$ as a quadratic function in W_t^* without interaction terms, where $\odot$ denotes the elementwise (Hadamard) product. The coefficients $\delta_0 = (\delta_{0,0}, \delta_{1,0}', \delta_{2,0}')'$ are determined by the OLS regression of $\hat{v}_t$ on $(1,\ W_t,\ W_t \odot W_t)$, where $\hat{v}_t$ denote the fitted residuals from the first-stage regression. This choice of δ_0 is intended to replicate the orthogonality between W_t and $f(W_t)$ in the simulated series. The identification of θ_0 based on the given conditional moment restrictions becomes stronger as $|\bar{\pi}|$ increases. However, in linear IV models, the identification strength should remain largely unaffected.[12]

The disturbance terms (u_t^*, v_t^*) follow a bivariate normal distribution, drawn independently across periods. The homoscedastic error variances, σ_u^2 and σ_v^2, are taken from the corresponding sample variances of $\hat{u}_t$ and $\hat{v}_t$, respectively. Their correlation coefficient is set to $\rho = 0.8$, which determines the degree of endogeneity.

The data are generated recursively over periods. Let $(\Delta c_{-1}^*, \Delta c_0^*)$ be the initial values, set to the observed values in the data and held constant across simulated series. In the t-th iteration step, where $t \in \{1, 2, \dots, n\}$, we first define $W_t^* = (\Delta c_{t-2}^*, W_{-1,t})$, where $W_{-1,t}$ comprises all components of W_t except the twice-lagged consumption growth. Next, we generate r_t^* and Δc_t^* sequentially according to the DGP specified in (27). This procedure is iterated until reaching the final period.

We focus on constructing a confidence set for $\theta_{1,0}$ based on the conditional moment restriction

$$\mathbb{E}[\,\Delta c_t^* - \mathbb{E}[\Delta c_t^*] - \theta_{1,0}(r_t^* - \mathbb{E}[r_t^*])|\,W_t^*] = 0. \tag{28}$$

The strength of identification can be adjusted by varying the constant $\bar{\pi}$ associated with the nonlinear term. In the baseline specification, we set $\bar{\pi} = 2$ to make the variation in the linear and nonlinear components comparable to each other.

[12] The comparison between conditional and unconditional moments is not new and is most recently discussed at length in Antoine and Lavergne (2023).

We follow the same procedure as outlined in Section 7, which involves demeaning both residuals and exponential weights. To construct the test statistic, we calculate the demeaned generalized residuals as $\widehat{U}_t(\theta_1) = \Delta c_t^* - \bar{\Delta} c^* - \theta_1(r_t^* - \bar{r}^*)$, where $\bar{\Delta} c^* = \frac{1}{n}\sum_{t=1}^n \Delta c_t^*$ and $\bar{r}^* = \frac{1}{n}\sum_{t=1}^n r_t^*$. The exponential weights are computed based on $\widetilde{W}_t^*$, defined analogously to $\widetilde{W}_t$. A similar demeaning is also applied to the exponential weights.

The test statistic is computed as defined in (22). To compute this maximum, we use the grid search (GS) method, adopting the same grid $\Gamma' = \{-5, -4.5, -4, \ldots, 5\}^4$ over $[-5, 5]^4$ as in Section 7. Despite concerns regarding potential power loss, we opted for the GS method over the PSO algorithm for simulations. This decision was based on its computational efficiency gained by using vectorized code. Some preliminary experiments with more refined grids support that Γ' successfully generates sufficient powers.

The optimal penalty level is selected from $\Lambda = \{0.1j : j = 0, \ldots, 10\} \subset [0, 1]$. To compute the optimal λ, we employ the following multiplier bootstrap:

$$\widehat{M}_{n,*,B}(\gamma) = \frac{1}{n}\sum_{t=1}^n \left\{\eta_t^* \widehat{U}_t(\theta_1) - \frac{B}{\sqrt{n}}(r_t^* - \bar{r}^*)\right\}\left\{\exp(W_t^{*\prime}\gamma) - \frac{1}{n}\sum_{t=1}^n \exp(W_t^{*\prime}\gamma)\right\},$$

$$\widehat{s}^2_{n,*,B}(\gamma) = \frac{1}{n}\sum_{t=1}^n \left[\left\{\eta_t^* \widehat{U}_t(\theta_1) - \frac{B}{\sqrt{n}}(r_t^* - \bar{r}^*)\right\}\left\{\exp(W_t^{*\prime}\gamma) - \frac{1}{n}\sum_{t=1}^n \exp(W_t^{*\prime}\gamma)\right\}\right]^2,$$

where $B/\sqrt{n} = 2/\sqrt{n}$ approximately corresponds to 0.2 on the actual scale of θ_1, and η_t^* are drawn independently from $N(0, 1)$. The optimal penalty is then calibrated for each $\theta_1 \in \Theta_1 := \{\theta_{1,0} + 0.1j : |j| \leq 6\}$ following the procedure specified in Section 6.1. The number of bootstrap replications is set to $R = 5000$, mirroring the empirical application. Summary statistics regarding the distribution of optimal λ are presented in Table 4 across various values of θ_1 in Θ_1.

We assess the size and power properties of our optimal penalized and unpenalized tests in comparison to those of the AR test and the Wald tests based on the 2SLS estimator and that of Domínguez and Lobato (2004). Figure 3 displays the power curves depicting the performance of each method. This indicates that only the AR and our tests maintain a size close to the nominal level of 0.1. In contrast, the Wald tests based on the estimators of θ_1 exhibit significant size distortion. Notably, our optimal test not only achieves nearly accurate size control but also enhances the power compared to both the unpenalized test (with or without size adjustments) and the AR test.[13] This underscores our motivation that ℓ_1-regularization can enhance the power of the test by selecting relevant information from the conditioning variables.

We conducted additional experiments where we included three- and four-period lags of the consumption growth rate as supplementary instruments. This amounts to adding noise to the IVs in our setting. Regardless of the value of $\bar{\pi}$, it holds that

$$\mathbb{E}^*\left[\frac{\partial U_t(\theta_{1,0})}{\partial \theta_1}\middle| W_t^*, \Delta c_{t-3}^*, \Delta c_{t-4}^*\right] = \mathbb{E}^*\left[\frac{\partial U_t(\theta_{1,0})}{\partial \theta_1}\middle| W_t^*\right].$$

Since Δc_{t-3}^* and Δc_{t-4}^* fail to provide additional explanatory power, it is expected that incorporating these variables as instruments would lead to reduced statistical power, compared to using the original set of 4 IVs. Table 5 (panel A) confirms this insight; however, it also reveals that the degree of power reduction is much less severe for the optimally penalized test than for the unpenalized test and the AR test.[14] This implies that the relative advantage of optimal penalization increases as the number of uninformative or less informative instruments becomes large.

We also investigated the size and power of each test under the linear specification by setting $\bar{\pi}$ to 0 while keeping the other coefficients the same. Table 5 (panel B) presents the results from this analysis. There are two noteworthy observations. First, our tests nearly

[13]There is seemingly no formal guidance on how to adjust the undersize of the unpenalized test to fairly compare its power with the other tests. For visual presentation, we depict 'Unpen. w/ shift' by shifting the power curve up parallel to match the nominal size of 0.1 at $\theta_1 - \theta_{1,0} = 0$.

[14]As a caveat, it is unlikely but possible that power reduction might have occurred due to the reduced time span from 105 to 103.

hold the correct size even when IVs are weak both conditionally and unconditionally. This demonstrates the robustness of our method to the presence of weak IVs. Second, excluding the nonlinear term leads to diminished powers of the penalized and unpenalized tests relative to the AR test. This indicates that a substantial portion of the power in our tests stems from an underlying nonlinear relationship between the noncentrality term and IVs. Furthermore, our results suggest that the nonlinearity term, if present, can serve as a valuable source of identification in IV models.

9 Conclusions

We have developed an inference method for a vector of parameters using an ℓ_1-penalized maximum statistic. Our inference procedure is based on the multiplier bootstrap and combines inference with model selection to improve the power of the test. We have recommended solving a data-dependent max-min problem to select the penalization tuning parameter. We have demonstrated the efficacy of our method using two empirical examples.

There are multiple directions to extend our method. First, we may consider a panel data setting where the number of conditioning variables may grow as the time series dimension increases. Second, unknown parameters may include an unknown function (e.g., Chamberlain, 1992; Newey and Powell, 2003; Ai and Chen, 2003; Chen and Pouzo, 2015). In view of results in Breunig and Chen (2020), Bierens-type tests without penalization might not work well when the parameter of interest is a nonparametric function. It would be interesting to study whether and to what extent our penalization method improves power for nonparametric inference. Third, multiple conditional moment restrictions or a continuum of conditional moment restrictions (e.g., conditional independence assumption) might be relevant in some applications. Fourth, it would be interesting to extend our method for empirical industrial organization. For instance, Gandhi and Houde (2019) proposed a set of relevant instruments from conditional moment restrictions to avoid the weak identification problem. It is an intriguing possibility to combine our approach with their insights into Berry, Levinsohn, and Pakes (1995). All of these extensions call for substantial developments in both theory and computation.

References

Ai, C., and X. Chen (2003): “Efficient estimation of models with conditional moment restrictions containing unknown functions,” *Econometrica*, 71(6), 1795–1843.

Andrews, D. W. K. (1997): “A Conditional Kolmogorov Test,” *Econometrica*, 65(5), 1097–1128.

Antoine, B., and P. Lavergne (2023): “Identification-Robust Nonparametric Inference in a Linear IV Model,” *Journal of Econometrics*, 235(1), 1–24.

Benítez-Silva, H., M. Buchinsky, H. M. Chan, S. Cheidvasser, and J. Rust (2004): “How large is the bias in self-reported disability?,” *Journal of Applied Econometrics*, 19(6), 649–670.

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

Bierens, H. J. (1982): “Consistent model specification tests,” *Journal of Econometrics*, 20(1), 105–134.

——— (1990): “A consistent conditional moment test of functional form,” *Econometrica*, 58(6), 1443–1458.

Bierens, H. J., and W. Ploberger (1997): “Asymptotic Theory of Integrated Conditional Moment Tests,” *Econometrica*, 65(5), 1129–1151.

Bierens, H. J., and L. Wang (2012): "Integrated conditional moment tests for parametric conditional distributions," *Econometric Theory*, 28(2), 328–362.

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

Breunig, C., and X. Chen (2020): "Adaptive, Rate-Optimal Testing in Instrumental Variables Models," arXiv:2006.09587 [econ.EM], `https://arxiv.org/abs/2006.09587`.

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34(3), 305–334.

——— (1992): "Efficiency bounds for semiparametric regression," *Econometrica*, 60(3), 567–596.

Chen, X., and Y. Fan (1999): "Consistent hypothesis testing in semiparametric and non-parametric models for econometric time series," *Journal of Econometrics*, 91, 373–401.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Chen, X., and D. Pouzo (2015): "Sieve Wald and QLR Inference on Semi/Nonparametric Conditional Moment Models," *Econometrica*, 83(3), 1013–1079.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Anti-concentration and honest, adaptive confidence bands," *Annals of Statistics*, 42(5), 1787–1818.

——— (2016): "Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings," *Stochastic Processes and their Applications*, 126(12), 3632–3651.

Davidson, J. (1994): *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.

de Jong, R. M. (1996): "The Bierens test under data dependence," *Journal of Econometrics*, 72(1), 1–32.

Domínguez, M. A., and I. N. Lobato (2004): "Consistent estimation of models defined by conditional moment restrictions," *Econometrica*, 72(5), 1601–1615.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003): "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics*, 117(1), 55–93.

Escanciano, J. C. (2006): "A consistent diagnostic test for regression models using projections," *Econometric Theory*, 22(6), 1030–1051.

Fan, Y., and Q. Li (2000): "Consistent model specification tests: Kernel-based tests versus Bierens' ICM tests," *Econometric Theory*, pp. 1016–1041.

Gandhi, A., and J.-F. Houde (2019): "Measuring Substitution Patterns in Differentiated-Products Industries," Working Paper 26375, National Bureau of Economic Research.

Hall, P., and C. C. Heyde (1980): *Martingale limit theory and its application*. Academic press.

Hansen, B. E. (1996): "Stochastic equicontinuity for unbounded dependent heterogeneous arrays," *Econometric Theory*, 12(2), 347–359.

Horowitz, J. L. (2006): "Testing a Parametric Model Against a Nonparametric Alternative with Identification Through Instrumental Variables," *Econometrica*, 74(2), 521–538.

Kennedy, J., and R. Eberhart (1995): "Particle swarm optimization," in *Proceedings of ICNN'95-International Conference on Neural Networks*, vol. 4, pp. 1942–1948.

Kim, J., and D. Pollard (1990): "Cube root asymptotics," *The Annals of Statistics*, pp. 191–219.

Kitamura, Y., G. Tripathi, and H. Ahn (2004): "Empirical likelihood-based inference in conditional moment restriction models," *Econometrica*, 72(6), 1667–1714.

Lavergne, P., and V. Patilea (2008): "Breaking the curse of dimensionality in nonparametric testing," *Journal of Econometrics*, 143, 103–122.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), 1565–1578.

Qu, Z., and D. Tkachenko (2016): "Global Identification in DSGE Models Allowing for Indeterminacy," *Review of Economic Studies*, 84(3), 1306–1345.

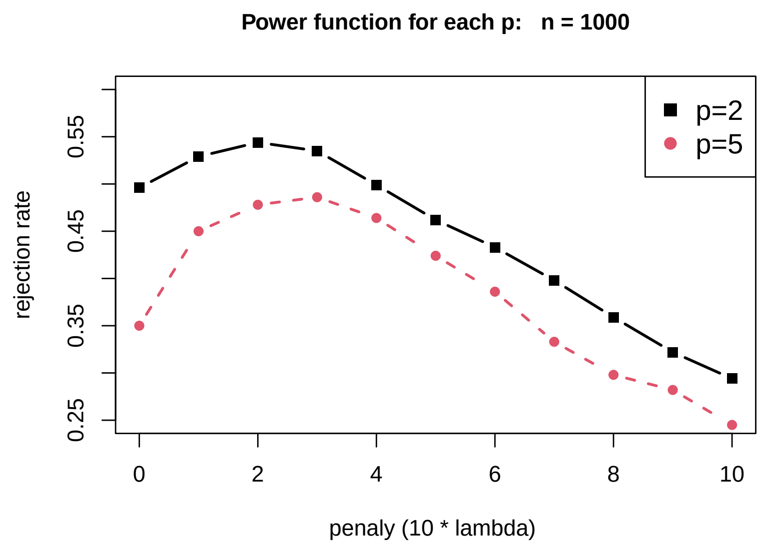
Shao, X., and J. Zhang (2014): "Martingale difference correlation and its use in high-dimensional variable screening," *Journal of the American Statistical Association*, 109(507), 1302–1318.

Stinchcombe, M. B., and H. White (1998): "Consistent Specification Testing with Nuisance Parameters Present Only under the Alternative," *Econometric Theory*, 14(3), 295–325.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer, New York, NY.

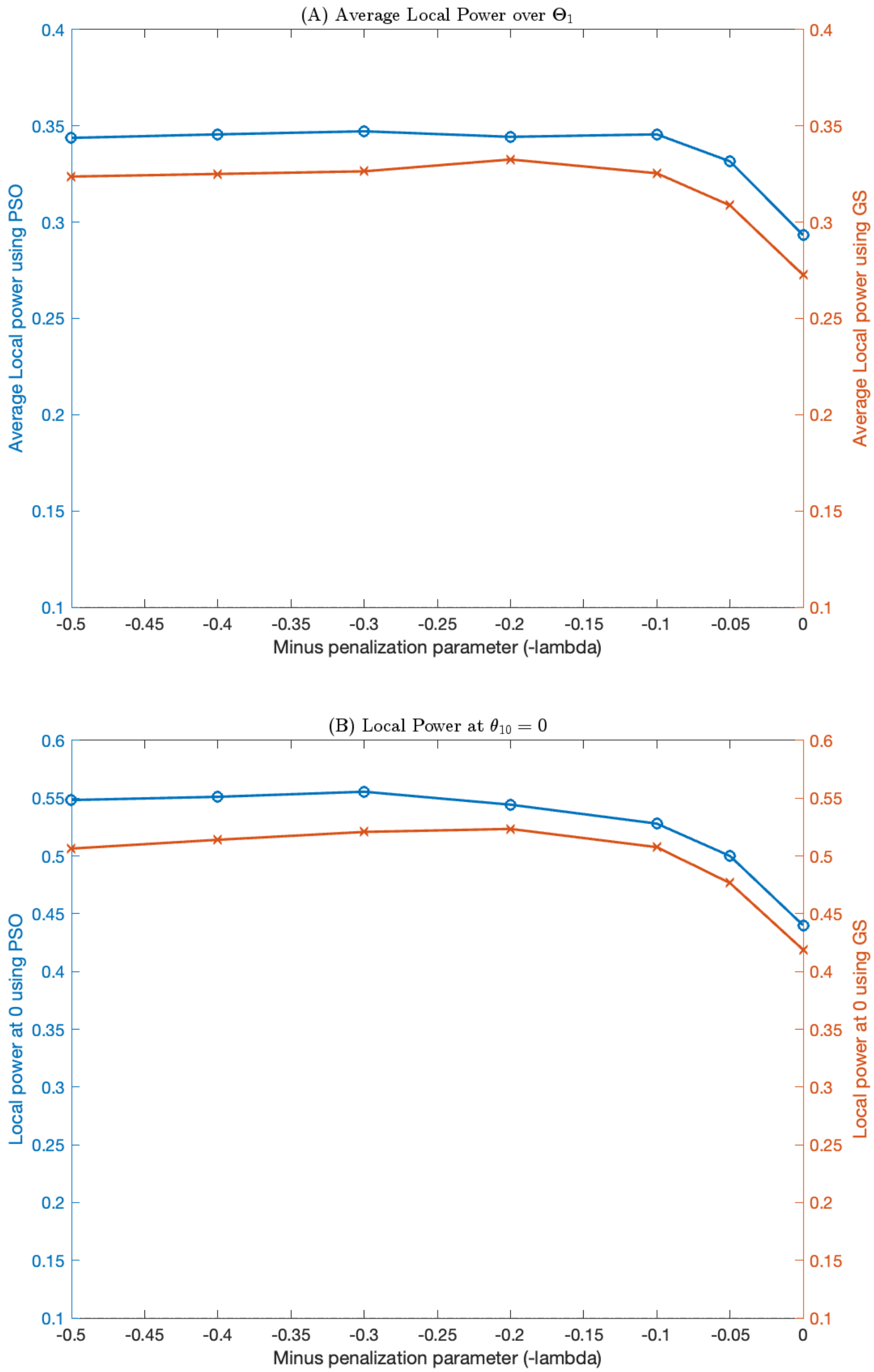
Yogo, M. (2004): "Estimating the elasticity of intertemporal substitution when instruments are weak," *Review of Economics and Statistics*, 86(3), 797–810.

Figure 1: Graphical Representation of Power Improvements via Penalization



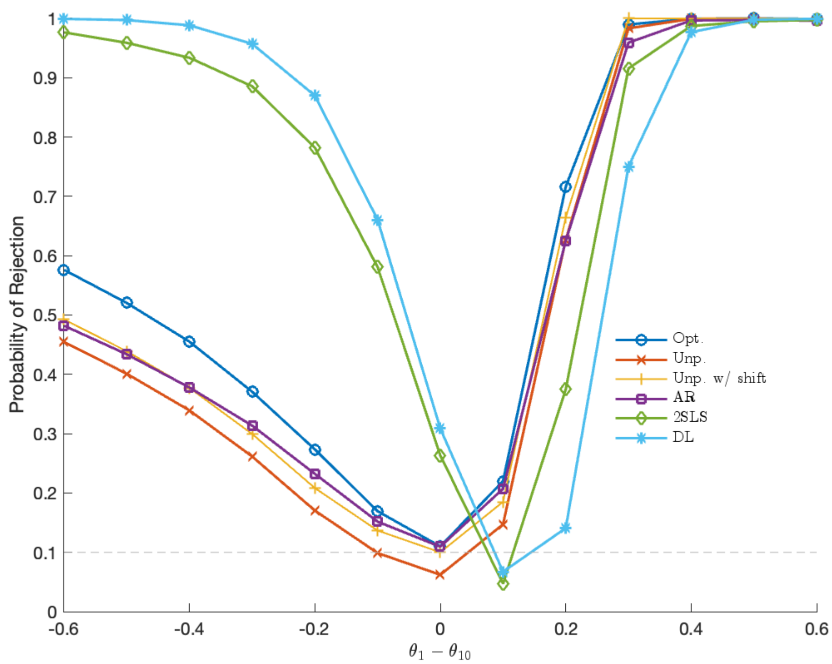
Notes: Figure 1 plots the "theoretical" power functions of our proposed test, where the power curves are obtained via Monte Carlo simulations described as in the main text with $\Gamma = [-5, 5]^p$.

Figure 2: Simulated Local Powers Computed Using PSO and GS



Notes: Figure 2 shows the simulated local powers using both algorithms, computed as per the eq. (15). Panel A of Figure 2 displays the local powers averaged over Θ_1 and Panel B exhibits the local powers evaluated at a specific $\theta_{1,0} = 0$.

Figure 3: Power Curves of Competing Tests under the Baseline Specification



Notes: The nominal size is set to 0.1. The yellow line, labeled as 'Unpen. w/ shift,' is derived by adjusting the size and power of the unpenalized test in parallel to achieve the actual size consistent with the nominal size 0.1. The number of simulation replications used for generating Figure 3 is 5000.

Table 1: Values of Max Statistic $T(\lambda, a)$

a	λ			
	0.30	0.20	0.10	0.00
1	1.807	1.807	1.850	2.120
2	1.807	1.807	1.850	2.176
3	1.807	1.807	1.890	2.325
4	1.807	1.807	1.901	2.410
5	1.807	1.807	1.901	2.448

Notes: The values reported in Table 1 are computed with a swarm size of 5000. The value of $\theta_{1,0}$ is fixed at the value of corresponding 2SLS estimate, namely -0.028.

Table 2: Mean and Standard Deviation of Computation Time for Max Statistic

a	λ			
	0.30	0.20	0.10	0.00
1	3.440 (0.070)	3.408 (0.097)	3.466 (0.114)	3.584 (0.110)
2	3.409 (0.054)	3.389 (0.123)	3.455 (0.121)	3.560 (0.131)
3	3.418 (0.054)	3.424 (0.125)	3.572 (0.224)	3.639 (0.240)
4	3.424 (0.104)	3.466 (0.161)	3.605 (0.223)	3.757 (0.579)
5	3.259 (0.059)	3.299 (0.067)	3.459 (0.245)	3.701 (0.871)

Notes: The unit of measurement is seconds in Table 2. Table 2 presents the average elapsed times for computing the penalized maximum statistic $T(\lambda, a) \equiv \max_{\gamma \in [-a,a]^4}[\widehat{Q}_n(\theta_{1,0}, \gamma) - \lambda \|\gamma\|_1]$ across various a and λ values. The standard deviations are displayed in parentheses. The averages and standard deviations are computed based on 1000 repetitions.

Table 3: Confidence Intervals Using Various Testing Methods

Optim.		Unpen.	
PSO	GS	PSO	GS
$[-0.30, 0.15]$	$[-0.34, 0.15]$	$[-0.43, 0.13]$	$[-0.58, 0.14]$

Notes: The figures within the brackets denote the 95% confidence intervals (CIs) derived from each testing and computation method. The first two columns labeled as 'Optim.' stands for our optimal CIs using the PSO and GS algorithms, respectively. Likewise, the next two columns represent the CIs from our unpenalized test.

Table 4: Selected Optimal λ Values across Different θ_1 Values

stats.	$\theta_1 - \theta_{1,0}$						
	-0.6	-0.4	-0.2	0	0.2	0.4	0.6
mean λ	0.281	0.266	0.237	0.208	0.253	0.353	0.412
pos. prob.	0.959	0.962	0.966	0.968	0.974	0.982	0.987

Notes: The first row represents the averages of the estimated optimal penalty and the second row presents the probability of choosing a strictly positive penalty. Each column represents the optimal λ choice at a given hypothesized value of θ_1. The number of simulation replications is 5000, and for each simulation draw, 5000 bootstrap replications are conducted.

Table 5: Size and Power of Each Test under the Baseline Specification ($\bar{\pi} = 2$) and Linear Specification ($\bar{\pi} = 0$).

Tests	$\theta_{1,0}$		$\theta_{1,0} + 2/\sqrt{n}$		
	4 IVs	6 IVs	4 IVs	6 IVs	Diff.
A. Baseline Specification					
Optimal	0.106	0.124	0.728	0.657	0.071
Unpenalized	0.062	0.041	0.640	0.463	0.177
Unpen. w/ shift	0.100	0.100	0.678	0.522	0.156
2SLS	0.270	0.403	0.391	0.313	0.078
AR	0.118	0.123	0.641	0.568	0.073
DL	0.311	0.882	0.145	0.169	−0.024
B. Linear Specification					
Optimal	0.120	0.129	0.617	0.572	0.045
Unpenalized	0.054	0.039	0.436	0.320	0.116
Unpen. w/ shift	0.100	0.100	0.428	0.381	0.047
2SLS	0.269	0.380	0.438	0.371	0.067
AR	0.113	0.108	0.684	0.633	0.051
DL	0.425	0.908	0.113	0.222	−0.109

Notes: The nominal level is set at 0.1. The first two columns denote the sizes of the tests using 4 and 6 IVs, respectively. The third and fourth columns show the powers of the tests under the local alternative $\theta_{1,0} + \frac{2}{\sqrt{n}} \approx \theta_{1,0} + 0.2$. The final column displays the reductions in power by including less informative IVs. The number of simulation draws is 5000, and for each simulation draw, 5000 bootstrap replications are conducted for the tests in the first three rows in the table.

Online Appendix to "Inference for parameters identified by conditional moment restrictions using a generalized Bierens maximum statistic"

A Proofs

Proof of Theorem 1. First, we show the stochastic equicontinuity of the processes $\sqrt{n}M_n(\gamma)$ and $s_n(\gamma)$ for (8) and (9). Due to the boundedness of Γ and W_i, $U_i \exp(W_i'\gamma)$ is Lipschitz continuous with a bound $K|U_i|\|\gamma_1 - \gamma_2\|$ for some K and for any γ_1 and γ_2. Then, this Lipschitz property and the existence of moment of some $c > d$ implies due to Theorem 2 in Hansen (1996) that the empirical process $\sqrt{n}M_n(\gamma)$ is stochastically equicontinuous. The Lipschitz continuity and the ergodic theorem also imply that $s_n(\gamma)$ is stochastically equicontinuous.

Next, the martingale difference sequence central limit theorem and the ergodic theorem yield the desired finite-dimensional convergence for (8) and (9) under Assumption 1; see e.g. Davidson (1994)'s Section 24.3 and 13.4.

Finally, for the convergence of $T_n(\lambda)$, note that both Λ and Γ are bounded, implying $\lambda\|\gamma\|_1$ is uniformly continuous. Thus, the process $\frac{|\mathcal{M}(\gamma)|}{s(\gamma)} - \lambda\|\gamma\|_1$ converges weakly in $\ell^\infty(\Gamma \times \Lambda)$, the space of bounded functions on $\Gamma \times \Lambda$, and the weak convergence of $T_n(\lambda)$ follows from the continuous mapping theorem since (elementwise) $\sup$ is a continuous operator.

□

Proof of Theorem 2. For the same reason as in the proof of Theorem 1, it is sufficient to verify the conditional finite dimensional convergence. As $\eta_i^* g(X_i, \bar{\theta}) \exp(W_i'\gamma)$ is a martingale difference sequence, we verify the conditions in Hall and Heyde (1980)'s Theorem 3.2, a conditional central limit theorem for martingales. Their first condition that

$$n^{-1/2} \sup_i \left| \eta_i^* g(X_i, \bar{\theta}) \exp(W_i' \gamma) \right| \xrightarrow{p} 0$$

and the last condition $\mathbb{E}\left[\sup_i \eta_i^{*2} g(X_i, \bar{\theta})^2 \exp(2W_i'\gamma)\right] = O(n)$ are straightforward since $\exp(W_i'\gamma)$ is bounded and $|\eta_i^* g(X_i, \bar{\theta})|$ has a finite c moment for $c > 2$. Next,

$$n^{-1} \sum_{i=1}^{n} \eta_i^{*2} g(X_i, \bar{\theta})^2 \exp(2W_i'\gamma) \xrightarrow{p} \mathbb{E}\left[g(X_i, \bar{\theta})^2 \exp(2W_i'\gamma)\right]$$

by the ergodic theorem. This completes the proof. □

Proof of Theorem 3. It follows from Lemma 1 that

$$\frac{\left|\mathbb{E}\left[g\left(X_i, \bar{\theta}\right) \exp(W_i'\gamma)\right]\right|}{\sqrt{\mathbb{E}\left[g^2\left(X_i, \bar{\theta}\right) \exp(2W_i'\gamma)\right]}} > 0$$

for almost every $\gamma \in \Gamma$. Then, the result follows from the ergodic theorem. □

Proof of Lemma 2. Since the derivative of $T(\lambda, b)$ with respect to λ at $\lambda = 0$ is $-\left\|\tilde{\gamma}(b)\right\|_1$ for each $b = 0, B$, the difference between the alternative and null limit experiments $T(0, B) - T(0, 0)$ at $\lambda = 0$ is stochastically dominated by $T(\lambda, B) - T(\lambda, 0)$ at a positive λ. This implies that the rejection probability of the test at any prespecified significance level is bigger at the experiment with the positive λ. Thus, $T(\lambda, 0) - T(0, 0) < T(\lambda, B) - T(0, B) \quad \text{a.s.}$ □

Proof of Theorem 4. We begin with showing that $c_\alpha^*(\lambda) \xrightarrow{p} c_\alpha(\lambda)$ uniformly in Λ. First, recall that the inverse map on the space of the distribution function F that assigns its α-quantile is Hadamard-differentiable at F provided that F is differentiable at $F^{-1}(\alpha)$ with a strictly positive derivative; see e.g. Section 3.9.4.2 in van der Vaart and Wellner (1996). Therefore, for the uniform consistency of the bootstrap, it is sufficient to show that $F_\lambda^*(x) \xrightarrow{p} F_\lambda(x)$ uniformly $(x, \lambda) \in A_\alpha$. However, this is a direct consequence of the conditional stochastic equicontinuity and the convergence of the finite-dimensional distributions established in Theorem 2.

Next, the preceding step implies that $T_{n,*,B}(\lambda) - c_\alpha^*(\lambda)$ converges weakly to

$$\sup_{\gamma \in \Gamma} \left\{ \left| \frac{\mathcal{M}(\gamma) + \mathbb{E}\left[\exp(W_i'\gamma) G_i B\right]}{s(\gamma)} \right| - \lambda \left\|\gamma\right\|_1 \right\} - c_\alpha(\lambda),$$

which is the limit in (15). This in turn yields the uniform convergence of $\mathcal{R}_n(\lambda, B)$ in probability to $\mathcal{R}(\lambda, B)$. Since $\mathcal{R}$ is continuous on a compact set, the standard consistency argument results in that $d(\widehat{\lambda}, \Lambda_0) \xrightarrow{p} 0$.

For the same reason, $T_n(\lambda) - c_\alpha^*(\lambda)$ converges weakly to $T(\lambda) - c_\alpha(\lambda)$ and thus the probability that $T_n(\widehat{\lambda}) \geq c_\alpha^*(\widehat{\lambda})$ converges to α for any sequence $\widehat{\lambda}$ due to the weak convergence.

□

Proof of Theorem 5. Write $g_i(\theta)$ and $G_{2i}(\theta)$ for $g(X_i, \theta)$ and $G_2(X_i, \theta)$, respectively. Note that for $\theta_1 = \theta_{1,0}$,

$$\begin{aligned}\sqrt{n}\widehat{M}_n(\gamma) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i(\theta_0) \left(\exp(W_i'\gamma) - \frac{1}{n} \sum_{j=1}^n \exp(W_j'\gamma) \right) \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_{ni} \frac{1}{n} \sum_{j=1}^n G_{2j}(\theta_0) \left(\exp(W_j'\gamma) - \frac{1}{n} \sum_{l=1}^n \exp(W_l'\gamma) \right) + o_p(1)\end{aligned}$$

due to Assumption 3. Then, $\frac{1}{n}\sum_{j=1}^n G_{2j}(\theta_0)\exp(W_j'\gamma)$ and $\frac{1}{n}\sum_{l=1}^n \exp(W_l'\gamma)$ converge uniformly in probability and $\frac{1}{\sqrt{n}}\sum_{i=1}^n a_1 g_i(\theta_0)\left(\exp(W_i'\gamma) - \mathbb{E}\exp(W_j'\gamma)\right) + a_2\zeta_{ni}$ is P-Donkser for any real a_1 and a_2 for the same reasoning as in the proof of Theorem 1. Similarly, the uniform convergence of $\widehat{s}_n^2(\gamma)$ follows since $g()$ is Lipschitz in θ by (21). □

B Testing Rational Unbiased Reporting of Ability Status

Benítez-Silva, Buchinsky, Chan, Cheidvasser, and Rust (2004, BBCCR hereafter) examine whether a self-reported disability status is a conditionally unbiased indicator of Social

Security Administration (SSA)'s disability award decision. Specifically, they test if $\tilde{U}_i = \tilde{A}_i - \tilde{D}_i$ has mean zero conditional on covariates W_i, where $\tilde{A}_i$ is the SSA disability award decision and $\tilde{D}_i$ is a self-reported disability status indicator. Their null hypothesis is $H_0 : \mathbb{E}[\tilde{A}_i - \tilde{D}_i | W_i] = 0$, which is termed as the hypothesis of *rational unbiased reporting* of ability status (RUR hypothesis).[15] They use a battery of tests, including a modified version of Bierens (1990)'s original test, and conclude that they fail to reject the RUR hypothesis. In fact, their Bierens test has the smallest p-value of 0.09 in their test results (see Table II of their paper). In this section, we revisit this result and apply our testing procedure.

Table B1 shows a two-way table of $\tilde{A}_i$ and $\tilde{D}_i$ and Table B2 reports the summary statistics of $\tilde{A}_i$ and $\tilde{D}_i$ along those of covariates W_i. After removing individuals with missing values in any of covariates, the sample size is $n = 347$ and the number of covariates is $p = 21$.[16]

Table B1: Self-reported disability and SSA award decision

	Self-reported disability ($\tilde{D}$)		Total
SSA award decision ($\tilde{A}$)	0	1	
0	35	51	86
1	61	200	261
Total	96	251	347

As in Section 7, we first studentize each of covariates and transform them by $x \mapsto \tan^{-1}(x)$ componentwise. The space Γ is set as $\Gamma = [-10, 10]^p$. As mentioned before,

[15] In this example, the null hypothesis is simpler than the empirical example in the main text because there is no parameter to estimate. There are alternative tests applicable (e.g., Escanciano, 2006; Shao and Zhang, 2014, among others) but we have not tried to implement them.

[16] According to Table I in BBCCR, the sample size is 393 before removing observations with the missing values; however, there are only 388 observations in the data file archived at the Journal of Applied Econometrics web page. After removing missing values, the size of the sample extract we use is $n = 347$, whereas the originally reported sample size is $n = 356$ in Table 2 in BBCCR.

Table B2: Summary Statistics

Variable	Mean	Stan. Dev.	Min.	Max.	$\widehat{\gamma}$
SSA award decision ($\tilde{A}$)	0.75	0.43	0	1	
Self-reported disability ($\tilde{D}$)	0.72	0.45	0	1	
Covariates					
White	0.57	0.50	0	1	0.07
Married	0.58	0.49	0	1	-0.16
Prof./voc. training	0.36	0.48	0	1	0.17
Male	0.39	0.49	0	1	0.02
Age at application to SSDI	55.97	4.81	33	76	0.33
Respondent income/1000	6.19	10.28	0	52	0.12
Hospitalization	0.88	1.44	0	14	–
Doctor visits	13.12	13.19	0	90	–
Stroke	0.07	0.26	0	1	-0.90
Psych. problems	0.25	0.43	0	1	–
Arthritis	0.40	0.49	0	1	–
Fracture	0.13	0.33	0	1	–
Back problem	0.59	0.49	0	1	-0.13
Problem with walking in room	0.15	0.36	0	1	–
Problem sitting	0.48	0.50	0	1	-0.03
Problem getting up	0.59	0.49	0	1	-0.03
Problem getting out of bed	0.24	0.43	0	1	-0.13
Problem getting up the stairs	0.45	0.50	0	1	–
Problem eating or dressing	0.07	0.26	0	1	–
Prop. worked in $t-1$	0.32	0.41	0	1	1.32
Avg. hours/month worked	2.68	8.85	0	60	–

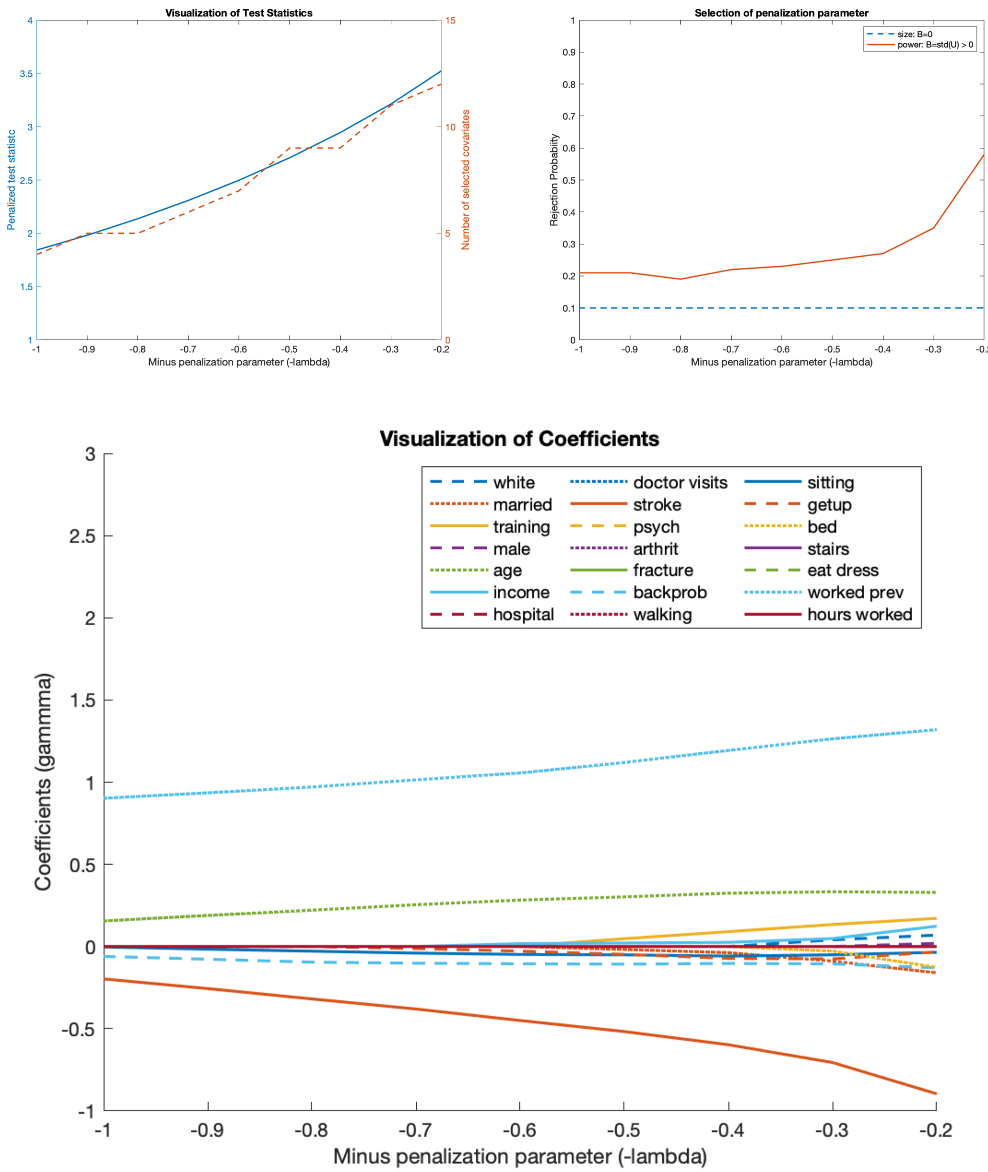
to compute T_n in (4), we use the `particleswarm` solver in `Matlab`. It is computationally easier to obtain T_n in (4) when λ is relatively larger. This is because a relevant space for Γ is smaller with a larger λ.[17]

To choose an optimal λ as described in Section 4.3, we parametrize the null hypothesis $H_0 : \mathbb{E}[\tilde{U}_i|W_i] = 0$ by $g(X_i, \theta) = \tilde{U}_i - \theta$ in (1) with $\theta_0 = 0$. First note that $G_i = 1$ in this example. Therefore, for each $\lambda \in \Lambda$, $\mathcal{R}(\lambda, B)$ is an increasing function of $|B|$. Thus, it suffices to evaluate the smallest value of $|B|$ satisfying $B \in \mathcal{B}$. Here, we take it to the sample standard deviation of $\tilde{U}_i$. For λ, we take $\Lambda = \{1, 0.9, \ldots, 0.2\}$. This range of λ's is chosen by some preliminary analysis. When λ is smaller than 0.2, it is considerably harder to obtain stable solutions; thus, we do not consider smaller values of λ. Since $\lambda \mapsto T_n(\lambda)$ is a decreasing function, we first start with the largest value of λ and then solves sequentially by lowering the value of λ, while checking whether the newly obtained solution indeed is larger than the previous solution. This procedure results in a solution path by λ.

Top-left panel of Figure B1 shows the solution path $\lambda \mapsto T_n(\lambda)$ along with the number of selected covariates, which is defined to be ones whose coefficients are no less than 0.01 in absolute value. For the latter, 4 covariates are selected with $\lambda = 1$, whereas 12 are chosen with $\lambda = 0.2$. Top-right panel displays the rejection probability defined in (19) when $B = 0$ (size) and $B = \widehat{\sigma}(\tilde{U}_i)$, where $\widehat{\sigma}(\tilde{U}_i)$ is the sample standard deviation of $\tilde{U}_i$. The level of the test is 0.1 and there are 100 replications to compute the rejection probability. The power is relatively flat up to $\lambda = 0.4$, increase a bit at $\lambda = 0.3$ and is maximized at $\lambda = 0.2$. The bottom panel visualizes each of 21 coefficients as λ decreases. It can be seen that the proportion of working in $t-1$ (`worked prev` in the legend of the figure) has the largest coefficient (in absolute value) for all values of λ and an indicator of stroke has the second largest coefficient, followed by age at application to Social Security Disability Insurance

[17]The specification of $\Gamma = [-10, 10]^p$ in this section is different from that of $\Gamma = [-5, 5]^p$ in Section 7. Because p is larger in the current example, we chose not to consider smaller values of λ and also opted to use a larger a in $\Gamma = [-a, a]^p$ to make sure that the value of a is not binding in optimization.

Figure B1: Testing Results



(SSDI). The coefficients for selected covariates are given in the last column of Table B2 for $\lambda = 0.2$.

Table B3: Bootstrap Inference

λ	Test statistic	No. of selected cov.s	Bootstrap p-value
0.2	3.525	12	0.021
0.3	3.213	11	0.020

Since the power in the top-right panel of Figure B1 is higher at $\lambda = 0.2$ and 0.3, we report bootstrap test results for $\lambda \in \{0.2, 0.3\}$ in Table B3. There are $R_{\text{PSO}} = 1000$ bootstrap replications to obtain the bootstrap p-values. Interestingly, we are able to reject the RUR hypothesis at the 5 percent level, unlike BBCCR. Furthermore, our analysis suggests that the employment history, captured by the proportion of working previously, stroke, and the age at application to SSDI are the three most indicative covariates that point to the departure from the RUR hypothesis.