

Extrapolating LATE with Weak IVs ^{*}

Muyang Ren

University of Tennessee, Knoxville[†]

October 12, 2026

Abstract

To evaluate the effectiveness of a counterfactual policy, it is often necessary to extrapolate treatment effects on compliers to broader populations. This extrapolation relies on exogenous variation in instruments, which is often weak in practice. This limited variation leads to invalid confidence intervals that are typically too short and cannot be accurately detected by classical methods. For instance, the F -test may falsely conclude that the instruments are strong. Consequently, I develop inference results that are valid even with limited variation in the instruments. These results lead to asymptotically valid confidence sets for various linear functionals of marginal treatment effects, including LATE, ATE, ATT, and policy-relevant treatment effects, regardless of identification strength. Finally, I illustrate my results using data from [Agan, Doleac, and Harvey \(2023\)](#) to analyze counterfactual policies of changing prosecutors' leniency and their effects on reducing recidivism.

Keywords: Identification-robust Inference, Subvector Inference, Marginal Treatment Effects, Examiner Design

JEL classification: C12, C21, C26.

^{*}I am very grateful to my advisor, Matt Masten, for his invaluable guidance and encouragement. I also thank my committee members, Federico Bugni, Adam Rosen, and Arnaud Maurel, for their generous advice and feedback. I am grateful to Jason Baron, Patrick Bayer, Xinyue Bei, Anna Bykhovskaya, Désiré Kédagni, Michael Pollmann, and Daniel Yi Xu, as well as seminar participants at the University of Hong Kong, UC Santa Barbara, Rice University, Southern Methodist University, Aarhus University, the University of Tennessee, Queen's University, Texas A&M University, and participants at various conferences, for their helpful comments and suggestions.

[†]Address: Muyang Ren, Department of Economics, University of Tennessee, Knoxville, 2200 Sutherland Avenue, Knoxville, TN 37919; Email: mren7@utk.edu

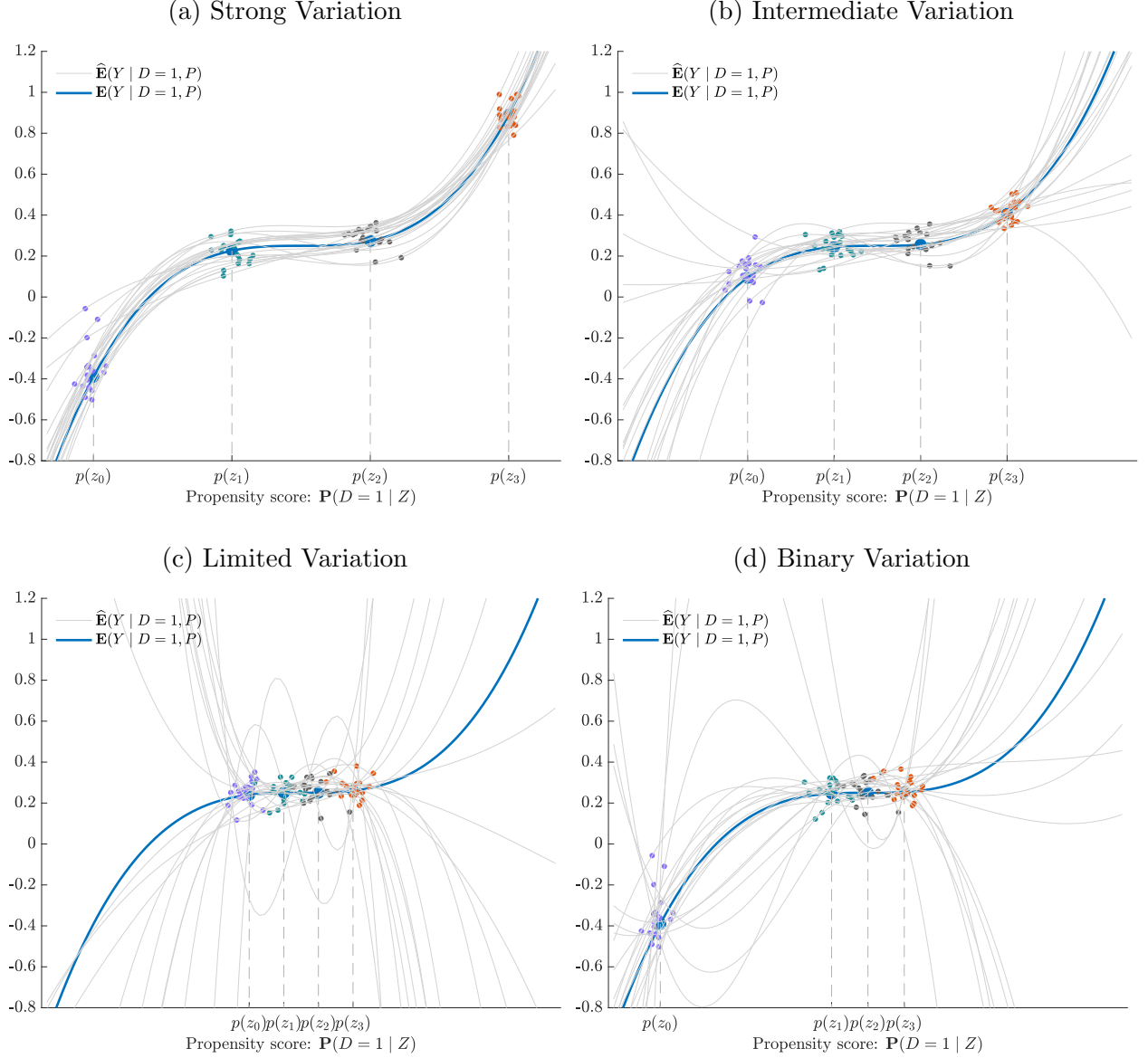
1 Introduction

This paper provides the first formal treatment of weak identification analysis in the marginal treatment effect (MTE) model. Originally developed in the seminal works of [Björklund and Moffitt \(1987\)](#) and [Heckman and Vytlacil \(1999, 2001, 2005\)](#), the MTE model has been widely adopted in various studies for extrapolating treatment effects, such as returns to schooling ([Carneiro, Heckman, and Vytlacil, 2011](#)), analysis of recidivism effects ([Agan et al., 2023](#)), and evaluation of social insurance programs ([Maestas, Mullen, and Strand, 2013](#)) (see Table 6 of [Mogstad and Torgovitsky \(2024\)](#) for a broad survey of MTE applications). Despite its widespread use in applied economic research, traditional confidence sets for the extrapolated causal effects within this model are often invalid when the probability of receiving treatment (i.e., the propensity score) exhibits limited variation across the instrument’s support. Moreover, this variation can be very weak even if the first-stage F -statistic is very large.

To demonstrate the consequences of weak IVs in MTE models and to illustrate why the F -statistic fails, I plot regression estimates of the outcome on propensity scores for treated units in Figure 1. These estimators are computed from independent simulations based on a cubic MTE model with a discrete IV that takes four values. Since the MTE can be directly recovered from this conditional regression and the one for untreated sample (Section 2), Figures 1a through 1c show the weak IV problem: as propensity score variation diminishes, the regression estimates (shown in gray) become increasingly volatile and diverge from the true population curve (shown in blue). The MTE estimator thus loses consistency, and the traditional Wald confidence interval built on these estimates fails to achieve its desired coverage probability. These problems arise not only when propensity scores cluster around one value but also when they exhibit binary variation (Figure 1d), since two distinct propensity scores cannot identify a cubic MTE. In this case, the F -statistic can be misleadingly large because it detects the deviation from the null where all propensity scores are equal.

To achieve valid coverage, this paper develops inference methods that are robust against weak IV variation in MTE models. I consider a class of MTE models with a discrete instrument as in [Brinch, Mogstad, and Wiswall \(2017\)](#) and [Kline and Walters \(2019\)](#). By imposing a parametric structure on the marginal treatment response (MTR) functions, the MTE is fully characterized by a finite-dimensional parameter, which can be point identified using discrete variation from the instrument and estimated by running separate regressions for

Figure 1: Estimators under Different Types of IV Strengths



Note: The figure shows the estimated expected outcomes for treated units conditional on propensity scores under different designs of propensity score variation, based on 20 independent simulations. All designs and simulations use a cubic specification for the MTE curve with a discrete IV that takes four values in its support. The sample size is 2,000.

the treated and control groups. The proposed procedures deliver confidence sets for causal parameters that are linear functionals of the MTE function, which cover a broad range of parameters including the MTE function itself, the average treatment effect (ATE), the average treatment effect on the treated (ATT), the local average treatment effect (LATE), and the policy-relevant treatment effect (PRTE). The asymptotic coverage rate of these confidence sets is no smaller than the nominal level, regardless of identification strength.

The first contribution of this paper is a simple inference method for linear MTE models (Brinch et al., 2017). This linear structure offers a novel moment condition for direct estimation of treatment effects of interest, bypassing the necessity to estimate the full model. Building on this moment condition, I construct a conditional Wald test whose critical values are simulated from a conditional distribution of the test statistic. The test is asymptotically similar, so the inverted confidence set has exact asymptotic coverage and is not conservative, whether the model is weakly identified or not.

For a general class of MTE models, causal effects of interest cannot be estimated directly without relying on other primitive parameters in the MTE function, and these weakly identified primitive parameters hamper valid inference on linear functions of them. The second contribution is a modified linear combination (MLC) test that builds on the improved projection approach of I. Andrews (2018) and achieves uniform validity regardless of identification strength. The inverted confidence set is straightforward to compute but potentially conservative under weak identification. Under strong identification, it is approximately as efficient as the Wald confidence set (Online Appendix E).

In a Monte Carlo study based on a quadratic MTE model, the MLC test is size-correct whether the model is strongly identified, weakly identified, or partially identified. In contrast, the classical Wald test over-rejects the null value at 14% frequency versus the 5% significance level under partial identification and exhibits trivial power under weak identification.

Finally, I illustrate the proposed methods by using data from Agan et al. (2023), who show that not prosecuting defendants with misdemeanor offenses reduces recidivism. While their analysis uses LATE estimates and ATE/ATT estimates under additive separability, I complement their findings by constructing confidence sets for the reduction in recidivism under two counterfactual prosecution policies: (1) a homogeneous marginal increase in non-prosecution rates across prosecutors, and (2) a minimum nonprosecution rate threshold. The results highlight that weak identification is a significant concern in higher-order polynomial models, and empirical conclusions differ substantially between robust and classical inference methods. In this application, I condition on covariates instead of imposing additive separability. Appendix B.1 explains why additive separability is a strong assumption, and Appendix B.2 shows how to implement the MLC test when researchers maintain it.

Related literature

This paper contributes to three strands of the literature: inference in MTE models, subvector inference in weakly identified models, and weak instruments in linear IV models.

Although the MTE model has been used to extrapolate treatment effects in a variety of fields, there are relatively few studies on the theory of estimation and inference in MTE models. In a semiparametric MTE model with continuous propensity scores, [Heckman, Urzua, and Vytlačil \(2006\)](#) use bootstrap methods to construct confidence bands for MTE functions. This approach is now common practice¹, but its theoretical validity has not yet been established. [Carneiro and Lee \(2009\)](#) analyze the pointwise limit distributions of MTE estimators using a separate regression approach. Building on this work, [Sasaki and Ura \(2023\)](#) further derive asymptotic theory for the PRTE using an orthogonalized score for double debiased estimation. Both papers assume a semiparametric MTE model with additive separability and strong IV/covariate variation to achieve consistent estimation. [Mogstad, Santos, and Torgovitsky \(2017, 2018\)](#) propose a bootstrap procedure for conducting inference on the PRTE when the MTE is partially identified. None of these papers study weak identification problems. Thus, compared to this existing literature, the inference approach I propose is the first to achieve robustness against weak instruments in MTE models.

The inference problem in this paper is also related to the literature on subvector inference in weakly identified models, where a subvector is a subset of the structural parameters (see [Andrews and Guggenberger \(2019, Section 12 in supplementary appendix\)](#) for thorough reviews). [Chaudhuri and Zivot \(2011\)](#) modify the robust Lagrange multiplier statistic of [Kleibergen \(2005\)](#) such that it is locally equivalent to asymptotically efficient subvector tests under strong identification. I. [Andrews \(2018\)](#) generalizes their results to inference on functions of parameters, which include subvectors as a special case. He proposes a robust confidence set that achieves *sequential* validity under a set of high-level conditions imposed on a sequence of data generating processes (DGPs), and concludes that “conditions for *uniform* asymptotic validity are an interesting open question” (Section V, page 347). D. [Andrews \(2017\)](#) improves the Bonferroni method of [Chaudhuri and Zivot \(2011\)](#) and shows that the refined tests are asymptotically non-conservative and uniformly valid. However, his method focuses on subvectors and does not directly address inference on a linear function of parameters, and D. [Andrews \(2017, page 2\)](#) acknowledges six key limitations, including computational challenges and the need for additional tuning parameters to categorize the identification strength. [Andrews and Cheng \(2012\)](#) and [Han and McCloskey \(2019\)](#) propose uniformly valid inference on functions of weakly identified parameters for models whose identification status is determined by whether a single parameter vector approaches zero. Moreover, all of these procedures rarely achieve asymptotic similarity when the nuisance parameters are weakly identified.

¹The Stata implementation of MTE estimation by [Brave and Walstrum \(2014\)](#) and [Andresen \(2018\)](#) uses bootstrap methods to produce confidence sets.

This paper differs from these works in two ways. First, for linear MTE models, the conditional Wald test of Section 3 circumvents the weakly identified nuisance parameters in the MTE. This delivers both asymptotic similarity and computational simplicity that are highly desirable for subvector inference. Existing subvector procedures with weakly identified nuisance parameters generally achieve neither property, except when there exists a reparametrization that separates a strongly identified nuisance vector from a weakly identified scalar parameter, as in the low-dimensional factor models of Cox (2026). Second, for general MTE models, I establish the *uniform* validity of the MLC test under primitive conditions on the parameter space via a modification to the test statistic. In contrast to Chaudhuri and Zivot (2011) and I. Andrews (2018), whose high-level conditions are tailored to weak identification of the local-to-zero sequence, as in linear IV regressions, the MLC test also addresses weak identification of the *local-to-singularity* sequence: when the propensity scores exhibit limited variation, the Jacobian of the moment functions approaches a singular but nonzero matrix (Section 4.2). Unlike Andrews and Cheng (2012) and Han and McCloskey (2019), where identification fails only as a single parameter vector approaches zero, identification in the MTE model can fail along multiple directions as the propensity scores approach a set rather than a point, as illustrated in Figure 1(c)-(d). Compared to D. Andrews (2017), the MLC test targets a linear function rather than a subvector and avoids the computational challenges and additional tuning parameters noted above. His test could in principle be applied to the MTE setting after reparametrizing the model, but finding a universal reparametrization rule that works for all linear hypotheses of interest while keeping the asymptotic analysis tractable is challenging.

Finally, this paper relates to the literature on weak instruments in linear IV models (see Andrews, Stock, and Sun (2019) for a review). This literature mostly considers inference on the IV estimand, which can be interpreted as a weighted average of causal effects under suitable assumptions (Imbens and Angrist, 1994). In contrast, the MTE model can directly target the policy counterfactuals of researchers' interest, such as the PRTE (Heckman and Vytlacil, 2001; Carneiro, Heckman, and Vytlacil, 2010), which are usually not recovered by linear IV regressions. The procedures in this paper allow robust inference on these policy counterfactuals.

2 MTE Model with Discrete IVs

In this section, I describe the MTE model and the related identification result following [Brinch et al. \(2017\)](#). Based on this result, the weakness of IVs can be characterized by limited variation of propensity scores. Then I introduce the relevant parameter space on which we achieve uniform validity of the proposed inference procedures.

2.1 Setup

Let Y_1 be the potential outcome of an individual who receives a binary treatment ($D = 1$) and Y_0 denote her potential outcome in the untreated state ($D = 0$). The observed outcome Y is realized through

$$Y = (1 - D)Y_0 + DY_1. \quad (1)$$

We further specify

$$Y_d = \mu_d + V_d, \quad d = 0, 1,$$

where $\mu_d \equiv \mathbb{E}[Y_d]$ is the mean potential outcome. For simplicity, I leave out additional covariates and discuss them in [Appendix B](#). The treatment is determined by a weakly separable selection equation

$$D = \mathbb{1}[U \leq \nu(Z)], \quad (2)$$

where $\nu(\cdot)$ is an unknown function, U is a continuous random variable representing the unobserved cost of selection into treatment, and $Z \in \text{supp}(Z) = \{z_0, z_1, \dots, z_K\}$ is the excluded discrete instrument. Researchers observe the outcome Y , the binary treatment D , and the excluded instrument Z from data, while the unobservables are the potential outcomes (Y_0, Y_1) and the variable U in the selection equation. The MTE model allows individuals to be selected into treatment based on their information on potential outcomes, which leads to potential dependence between (V_1, V_0) and U .

The key identifying assumption from [Brinch et al. \(2017, proposition 1.ii\)](#) is as follows:

Assumption 1 (MTE model with discrete IVs).

1. $Z \perp\!\!\!\perp U$.
2. $\mathbb{E}[Y_d | Z, U] = \mathbb{E}[Y_d | U]$ and $\mathbb{E}|Y_d| < \infty$ for $d \in \{0, 1\}$.
3. U is continuously distributed.
4. $0 < \mathbb{P}(D = 1 | Z = z) < 1$ for all $z \in \text{supp}(Z)$.

5. Let $\{h_m(\cdot)\}_{m=1}^M$ be known continuous functions on $[0, 1]$ with $\int_0^1 h_m(u) du = 0$ for $m = 1, \dots, M$. For $d \in \{0, 1\}$, the MTR function is given by

$$\mathbb{E}[Y_d \mid F_U(U) = u] = \mu_d + \sum_{m=1}^M \rho_{dm} h_m(u) \quad \text{for } u \in (0, 1),$$

where $F_U(\cdot)$ denotes the distribution function of U .

6. Let $\lambda_{00}(\cdot) = \lambda_{10}(\cdot) \equiv 1$, and define

$$\lambda_{1m}(p) \equiv \frac{1}{p} \int_0^p h_m(u) du \quad \text{and} \quad \lambda_{0m}(p) \equiv \frac{1}{1-p} \int_p^1 h_m(u) du \quad \text{for } m = 1, \dots, M.$$

for $p \in [0, 1]$. Then $\{\lambda_{1m}(\cdot)\}_{m=0}^M$ and $\{\lambda_{0m}(\cdot)\}_{m=0}^M$ are unisolvent² on $(0, 1)$.

7. $\overline{K} \equiv |\{\mathbb{P}(D = 1 \mid Z = z) : z = z_0, z_1, \dots, z_K\}| \geq M + 1$.

Assumptions 1.1 and 1.2 require the excluded instrument Z to be exogenous to both the selection and outcome processes. Assumption 1.3 allows us to normalize the marginal distribution of U to be uniformly distributed over $[0, 1]$. That is, we can transform U to a uniformly distributed variable $\tilde{U} = F_U(U)$. Under the exogeneity of Z in Assumption 1.1, the function $F_U(\nu(Z))$ can then be interpreted as propensity score:

$$\begin{aligned} p(z) &\equiv \mathbb{P}(D = 1 \mid Z = z) \\ &= \mathbb{P}(U \leq \nu(Z) \mid Z = z) \\ &= \mathbb{P}(U \leq \nu(z)) \\ &= F_U(\nu(z)) \end{aligned}$$

where the third line uses the Assumption 1.1. Define $\tilde{\nu}(z) \equiv F_U(\nu(z))$, and then we can work with $(\tilde{U}, \tilde{\nu}(z))$ in place of $(U, \nu(z))$ without affecting the empirical content of the selection model. For simplicity, we drop the tilde and assume U is uniformly distributed throughout our analysis, and therefore $\nu(z) = p(z)$. Assumption 1.4 validates the overlap condition, so that we observe both treated and untreated individuals for each group defined by the value of the instrument. Assumption 1.5 imposes a parametric restriction on the MTR function $\mathbb{E}[Y_d \mid F_U(U) = u]$ so that the MTE can be extrapolated outside the discrete support of the instrument. Assumption 1.6 is a weak condition that rules out redundant specifications in

²A set of n functions $f_1, f_2, \dots, f_n$ is unisolvent on domain Ω if the matrix $F \in \mathbb{R}^{n \times n}$ with entries $f_i(x_j)$ has nonzero determinant for any choice of n distinct points $x_1, x_2, \dots, x_n$ in Ω .

$\{h_m(\cdot)\}_{m=0}^M$ that cause multicollinearity in $\{\lambda_{dm}(\cdot)\}_{m=0}^M$. In particular, the usual polynomial specification $h_m(u) = u^m - \frac{1}{m+1}$ satisfies this condition. Assumption 1.7 requires sufficient variation of the exogenous instrument to point identify the structural parameters. While $\bar{K} \leq K+1$, with strict inequality when multiple instrument values yield identical propensity scores, point identification requires the number of distinct propensity scores, not instrument values, to exceed the order of the MTE model.

Define $\theta_d \equiv (\mu_d, \rho_{d1}, \dots, \rho_{dM})'$ for $d = 0, 1$, and let $\theta \equiv (\theta'_1, \theta'_0)'$. Then we have the following identification result:

Theorem 2.1 (Identification). Suppose Assumption 1 holds. Then θ is point identified.

Proof. Based on the model (1)–(2) and Assumption 1, we have

$$\begin{aligned}
\mathbb{E}[Y \mid D = 1, Z = z] &= \mathbb{E}[Y_1 \mid U \leq p(z), Z = z] \\
&= \mathbb{E}[Y_1 \mid U \leq p(z)] \\
&= \frac{1}{p(z)} \int_0^{p(z)} \mathbb{E}[Y_1 \mid U = u] du \\
&= \mu_1 + \sum_{m=1}^M \frac{\rho_{1m}}{p(z)} \int_0^{p(z)} h_m(u) du. \\
&= (\lambda_{10}(p(z)), \lambda_{11}(p(z)), \dots, \lambda_{1M}(p(z)))\theta_1.
\end{aligned} \tag{3}$$

The first line holds by equations (1), (2), and the normalization that $\nu(z) = p(z)$, which is jointly implied by Assumption 1.1 and 1.3. The second line holds by the exogeneity condition in Assumption 1.2. The third line follows by the overlap condition in Assumption 1.4 and the normalization that U is uniformly distributed over $[0, 1]$. The fourth line holds by parametric restriction in Assumption 1.5. Likewise, we have

$$\begin{aligned}
\mathbb{E}[Y \mid D = 0, Z = z] &= \mu_0 + \sum_{m=1}^M \frac{\rho_{0m}}{1 - p(z)} \int_{p(z)}^1 h_m(u) du \\
&= (\lambda_{00}(p(z)), \lambda_{01}(p(z)), \dots, \lambda_{0M}(p(z)))\theta_0.
\end{aligned} \tag{4}$$

Taking $z = z_0, z_1, \dots, z_K$ in equations (3) and (4) then yields two matrix equalities:

$$\beta_d = A_d \theta_d \quad \text{for } d = 0, 1, \tag{5}$$

where

$$A_d = \begin{bmatrix} \lambda_{d0}(p(z_0)) & \lambda_{d1}(p(z_0)) & \dots & \lambda_{dM}(p(z_0)) \\ \lambda_{d0}(p(z_1)) & \lambda_{d1}(p(z_1)) & \dots & \lambda_{dM}(p(z_1)) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{d0}(p(z_K)) & \lambda_{d1}(p(z_K)) & \dots & \lambda_{dM}(p(z_K)) \end{bmatrix}$$

and

$$\beta_d = (\mathbb{E}[Y \mid D = d, Z = z_0], \dots, \mathbb{E}[Y \mid D = d, Z = z_K])'.$$

Note that Assumptions 1.6 and 1.7 guarantee that there exists a full-rank submatrix of A_d with $\overline{K}$ rows and $M + 1$ columns for each $d = 0, 1$. Therefore, both A_1 and A_0 have full column rank. Then θ_1 and θ_0 are point identified by equation (5). $\square$

Remark 2.1. This identification result closely aligns with the proof of Brinch et al. (2017, Proposition 1.ii), who show that Assumption 1.7 is a necessary condition for identifying θ in an MTE model satisfying Assumption 1.1–1.5. Building on their result, I add Assumption 1.6 to establish sufficient conditions for point identification. $\square$

Based on the parametric restriction in Assumption 1.5, treatment effects can often be written as linear functions of the primitive parameter θ . For example, $\text{ATE} = \mu_1 - \mu_0$ and

$$\begin{aligned} \text{MTE}(u) &= \mathbb{E}[Y_1 - Y_0 \mid U = u] \\ &= c(u)' \theta_1 - c(u)' \theta_0, \end{aligned}$$

where $c(u) = (1, h_1(u), \dots, h_M(u))$. Moreover, suppose we are interested in the parameters that inform potential policy decisions such as average treatment effects on the (un)treated groups or policy-relevant treatment effects. In that case, the weight c is often unknown and depends on the underlying DGP (see Table 1 and 2 in Mogstad and Torgovitsky (2018)). If we write $c = (c'_1, c'_0)'$ with $c_d = (c_{d,0}, \dots, c_{d,M})'$ for $d = 0, 1$ and denote $h_0(u) \equiv 1$, Table 1 summarizes the weights of various treatment effects under Assumption 1. In this paper, I consider inference on treatment effects of the form $c'\theta$ where $c_1 = -c_0$ since many causal effects of interest including those in Table 1 have symmetric weights: $c_{1,m} = -c_{0,m}$ for all $m = 1, \dots, M$.

Define $q(z_\ell) = \mathbb{P}(Z = z_\ell)$, i.e., the probability mass function of the discrete IV. For the commonly used polynomial MTE model in empirical studies where $h_m = u^m - \frac{1}{m+1}$, the

Table 1: Weights for Treatment Effects

Target parameter	Expression	Weights
ATE	$\mathbb{E}[Y_1 - Y_0]$	$c_{1,m} = \mathbb{1}[m = 0]$
MTE	$\mathbb{E}[Y_1 - Y_0 \mid U = u]$	$c_{1,m} = h_m(u)$
ATT	$\mathbb{E}[Y_1 - Y_0 \mid D = 1]$	$c_{1,m} = \frac{\mathbb{E}(\int_0^{p(Z)} h_m(u) du)}{\mathbb{P}(D=1)}$
ATU	$\mathbb{E}[Y_1 - Y_0 \mid D = 0]$	$c_{1,m} = \frac{\mathbb{E}(\int_{p(Z)}^1 h_m(u) du)}{\mathbb{P}(D=0)}$
LATE	$\mathbb{E}[Y_1 - Y_0 \mid p(z_0) < U < p(z_k)]$	$c_{1,m} = \frac{\int_{p(z_0)}^{p(z_k)} h_m(u) du}{p(z_k) - p(z_0)}$
Additive PRTE	PRTE with $p^\epsilon(z) = p(z) + \epsilon$	$c_{1,m} = \frac{\mathbb{E}[\int_{p(Z)}^{p^\epsilon(Z)} h_m(u) du]}{\mathbb{E}[p^\epsilon(Z) - p(Z)]}$
Proportional PRTE	PRTE with $p^\epsilon(z) = (1 + \epsilon)p(z)$	
Quota	PRTE with $p^\epsilon(z) = \max\{p(z), \epsilon\}$	

weights of ATT, LATE, additive PRTE, and proportional PRTE become

$$\begin{aligned}
c_{1,m}^{\text{ATT}} &= \frac{1}{m+1} \left(\frac{\sum_{\ell=0}^K p(z_\ell)^{m+1} q(z_\ell)}{\sum_{\ell=0}^K p(z_\ell) q(z_\ell)} - 1 \right) \\
c_{1,m}^{\text{LATE}} &= \frac{1}{m+1} \left(\sum_{j=0}^m p(z_k)^j p(z_0)^{m-j} - 1 \right) \\
c_{1,m}^{\text{A-PRTE}} &= \frac{1}{m+1} \left(\sum_{\ell=0}^K q(z_\ell) \sum_{j=0}^m [p(z_\ell)^j (p(z_\ell) + \epsilon)^{m-j}] - 1 \right) \\
c_{1,m}^{\text{P-PRTE}} &= \frac{1}{m+1} \left[\frac{(1 + \epsilon)^{m+1} - 1}{\epsilon} \cdot \frac{\sum_{\ell=0}^K p(z_\ell)^{m+1} q(z_\ell)}{\sum_{\ell=0}^K p(z_\ell) q(z_\ell)} - 1 \right]
\end{aligned}$$

for $m = 1, \dots, M$, and the first element $c_{1,0} = 1$ for all the above causal effects.

2.2 Weak identification

The identification strategy relies on the conditions in Assumption 1, particularly the existence of sufficiently many distinct propensity scores to point identify the structural parameter θ through Assumption 1.7. However, variation in propensity scores is often limited. This limited variation may not be enough to guarantee correct asymptotic approximation in the construction of classical confidence sets. In section 3 and 4, I develop robust inference procedures for treatment effects of the form $c'\theta$ without requiring the potentially restrictive Assumption 1.7. My results have two implications. Along a sequence of DGPs: (1) When Assumption 1.7 holds but is close to failing, the parameters are point identified, and the robust confidence set achieves correct coverage for the causal parameter of interest; (2)

When Assumption 1.7 fails, the parameters are partially identified, and the proposed robust confidence set maintains correct coverage of the causal parameter $c'\theta$, where θ lies in the set defined by the linear system (5). Additionally, this uniform validity result does not rely on Assumption 1.6, though this assumption is typically satisfied under researchers' common specifications of the MTR functions.

2.3 Parameter space restriction

This section introduces the parameter space for the joint distribution of (Y, D, Z) , on which the proposed robust inference procedures are uniformly valid. First, consider the usual i.i.d. sampling distribution of the observables.

Assumption 2. The random vectors (Y_i, D_i, Z_i) , $i = 1, \dots, n$, are i.i.d. with common distribution F .

Next, I introduce a set of regularity conditions on the joint distribution F .

Definition 2.1 (Parameter Space). For some $\delta, \zeta > 0$ and $\epsilon \in (0, 1/2)$, define the parameter space $\mathcal{P}$ as the set of pairs (θ, F) satisfying the following properties:

1. Equation (5) is satisfied with $K \geq M$, where $\theta = (\theta'_1, \theta'_0)' \in \text{int}(\Theta) \subseteq \mathbb{R}^{2(M+1)}$ for some compact set Θ with nonempty interior;
2. $\sup_{d=0,1} \sup_{z \in \text{supp}(Z)} \mathbb{E}_F[|Y|^{2+\delta} \mid D = d, Z = z] \leq \zeta$;
3. $\epsilon \leq \inf_{z \in \text{supp}(Z)} \mathbb{P}_F(D = 1 \mid Z = z) \leq \sup_{z \in \text{supp}(Z)} \mathbb{P}_F(D = 1 \mid Z = z) \leq 1 - \epsilon$;
4. $\epsilon \leq \inf_{z \in \text{supp}(Z)} \mathbb{P}_F(Z = z) \leq \sup_{z \in \text{supp}(Z)} \mathbb{P}_F(Z = z) \leq 1 - \epsilon$;
5. $\epsilon \leq \inf_{d=0,1} \inf_{z \in \text{supp}(Z)} \text{var}_F(Y \mid D = d, Z = z)$.

The first condition assumes the correct model specification derived from Assumptions 1.1-1.5, and the order of the MTE model does not exceed the support of the instrument.³ The second condition is a mild restriction on the existence of moments of Y conditional on D and Z , which is essential for establishing the uniform version of the law of large numbers and central limit theorems. The third and fourth conditions, also known as the strong overlap conditions, require observing units for each value of the treatment and the instrument.

³The condition $K \geq M$ is necessary but not sufficient for point identification, particularly when Assumption 1.7 does not hold.

Together with the final condition regarding sufficient variation in outcomes, these conditions rule out the singularity or near-singularity of the asymptotic variance of the moment equation (5). Since our parameter of interest has the form of $c'\theta$, let $\mathcal{P}_0 = \{(\lambda, F) : \lambda = c'\theta, (\theta, F) \in \mathcal{P}\}$ denote the null parameter space for this linear function, in which the weight c could possibly depend on underlying DGP F . It is important to emphasize that our parameter space does not require the relevance Assumptions 1.6 and 1.7, the latter of which is likely to fail. For a given $\lambda \in \mathbb{R}$, our goal is to develop uniformly valid tests for assessing the hypothesis $H_0 : c'\theta = \lambda$ that is robust to the limited variation of propensity scores.

2.4 Notation and preliminaries

The asymptotic behavior of test statistics to be proposed will be unified by the asymptotic distribution of estimators on marginal distribution of instrument $q(z_\ell) \equiv \mathbb{P}(Z = z_\ell)$, propensity score $p(z_\ell) = \mathbb{P}(D = 1 \mid Z = z_\ell)$, and the expected outcome conditional on the treatment and instrument $\beta_{d\ell} = \mathbb{E}[Y \mid D = d, Z = z_\ell]$ for $d = 0, 1$ and $\ell = 0, 1, \dots, K$. Under the parameter space restriction outlined in section 2.3, we have the following sample-analog estimators for these quantities:

$$\begin{aligned}\hat{q}(z_\ell) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}[Z_i = z_\ell] \\ \hat{p}(z_\ell) &= \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}[D_i = 1, Z_i = z_\ell]}{\hat{q}(z_\ell)} \\ \hat{\beta}_{d\ell} &= \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{\hat{q}(d, z_\ell)},\end{aligned}$$

where $\hat{q}(1, z_\ell) = \hat{p}(z_\ell)\hat{q}(z_\ell)$ and $\hat{q}(0, z_\ell) = (1 - \hat{p}(z_\ell))\hat{q}(z_\ell)$.

Collecting these estimators into vectors, we have $\hat{p} = (\hat{p}(z_0), \dots, \hat{p}(z_K))'$, $\hat{q} = (\hat{q}(z_0), \dots, \hat{q}(z_K))'$, $\hat{\beta}_d = (\hat{\beta}_{d0}, \dots, \hat{\beta}_{dK})'$ for $d = 0, 1$, and $\hat{\beta} = (\hat{\beta}'_1, \hat{\beta}'_0)'$. Lemma D.1(a) provides the asymptotic distribution of the estimators $\hat{p}$, $\hat{q}$, and $\hat{\beta}$ under a drifting sequence in the parameter space $\mathcal{P}$. Moreover, Lemma D.1(b) also shows that their asymptotic variances can be consistently estimated by $\hat{\Sigma}_p$, $\hat{\Sigma}_q$, and $\hat{\Sigma}_\beta \equiv \text{diag}\{\hat{\Sigma}_{\beta_1}, \hat{\Sigma}_{\beta_0}\}$, defined in the lemma.

Throughout the paper, I use $\partial_x f \in \mathbb{R}^k$ to denote the gradient of a scalar function f with respect to its argument $x \in \mathbb{R}^k$. If f maps to a vector in $\mathbb{R}^l$, then $\partial_x f \in \mathbb{R}^{l \times k}$ denotes the Jacobian of f with respect to $x \in \mathbb{R}^k$. Let $P_A = A(A'A)^{-1}A'$ denote the projection matrix onto the column spaces of matrix A and define $M_A = I - P_A$ as the corresponding annihilator

matrix. Let $\text{supp}(X)$ denote the support of a random element X . I use subscripts F and F_n to denote population quantities that depend on distribution F and on a sequence of distributions $\{F_n\}_{n \geq 1}$, respectively. I make this dependence explicit if necessary.

3 Robust Inference in Linear MTE Models

In this section, I start with the simplest functional specification and impose linearity on the MTR function.

Assumption 3 (Linear MTE model). Assumption 1.5 holds with

$$h_m(u) = u - \frac{1}{2}$$

and $M = 1$.

This assumption was first introduced by [Olsen \(1980\)](#) to characterize sample selection bias and later generalized by [Brinch et al. \(2017\)](#) to model MTE functions. This linear MTE specification was recently adopted by [Kowalski \(2023\)](#) to extrapolate treatment effects between two experimental studies. The parameters $\theta = (\mu_1, \rho_{11}, \mu_0, \rho_{01})$ can be identified using any pair of propensity scores $(p(z_0), p(z_k))$ for $k = 1, \dots, K$ that differ from each other. Point identification fails if and only if all propensity scores $p(z_0), \dots, p(z_K)$ are identical, i.e., when there is no variation in propensity scores.

The linear MTE specification allows us to derive closed-form estimands for θ . In particular, equation (5) implies for each $k \in \{1, \dots, K\}$, we have

$$\begin{pmatrix} \mu_1 \\ \rho_{11} \end{pmatrix} = \frac{1}{p(z_k) - p(z_0)} \begin{pmatrix} p(z_k) - 1 & -(p(z_0) - 1) \\ -2 & 2 \end{pmatrix} \begin{pmatrix} \mathbb{E}[Y \mid D = 1, Z = z_0] \\ \mathbb{E}[Y \mid D = 1, Z = z_k] \end{pmatrix}$$

and

$$\begin{pmatrix} \mu_0 \\ \rho_{01} \end{pmatrix} = \frac{1}{p(z_k) - p(z_0)} \begin{pmatrix} p(z_k) & -p(z_0) \\ -2 & 2 \end{pmatrix} \begin{pmatrix} \mathbb{E}[Y \mid D = 0, Z = z_0] \\ \mathbb{E}[Y \mid D = 0, Z = z_k] \end{pmatrix}.$$

Recall $\beta_{dk} = \mathbb{E}[Y \mid D = d, Z = z_k]$. Then the ATE and the slope of MTE can be written as

$$\begin{pmatrix} \mu_1 - \mu_0 \\ \rho_{11} - \rho_{01} \end{pmatrix} = \frac{1}{p(z_k) - p(z_0)} \begin{pmatrix} p(z_k) [\beta_{10} - \beta_{00}] - p(z_0) [\beta_{1k} - \beta_{0k}] + \beta_{1k} - \beta_{10} \\ 2(\beta_{00} - \beta_{10} + \beta_{1k} - \beta_{0k}) \end{pmatrix}.$$

Denote the numerator of the right-hand-side equation by

$$\Delta_\mu(z_0, z_k) \equiv p(z_k) [\beta_{10} - \beta_{00}] - p(z_0) [\beta_{1k} - \beta_{0k}] + \beta_{1k} - \beta_{10}$$

and

$$\Delta_\rho(z_0, z_k) \equiv 2(\beta_{00} - \beta_{10} + \beta_{1k} - \beta_{0k}).$$

Since these quantities can be directly identified from data, we can express the treatment effects parameters $\lambda = c'\theta$ with $c = (c_\mu, c_\rho, -c_\mu, -c_\rho)'$ as the solution of the following moment function:

$$g_k(\lambda) = [p(z_k) - p(z_0)] \lambda - c_\mu \Delta_\mu(z_0, z_k) - c_\rho \Delta_\rho(z_0, z_k) = 0. \quad (6)$$

If the instrument Z is binary, then $c'\theta$ is just identified by this linear moment restriction, otherwise, we can construct a vector of moment functions:

$$g(\lambda) = (g_1(\lambda), \dots, g_K(\lambda))' = 0_{K \times 1} \quad (7)$$

to improve the efficiency for inference on $\lambda = c'\theta$. It is worth noting that other components of primitive parameter θ do not enter into (7). As a result, their nonstandard asymptotic behavior (under weak identification) is irrelevant in this context. To this end, this linear moment function $g(\lambda)$ will be used to construct asymptotically similar tests for treatment effects in linear MTE models.

To fix ideas, I first consider inference for the parameter $c'\theta$ with a known weight c and exploit the binary variation $\{z_0, z_k\}$ for $k = 1, \dots, K$. This known-weight setting allows for inference on parameters such as the ATE and the MTE function. Online Appendix F discusses the extension to cases where c needs to be estimated.

Equation (6) shows that $g_k(\lambda)$ suffices to estimate λ , the causal effects of interest. Let $\hat{g}_k(\lambda)$ be the sample analog of this moment function after plugging in the estimators $\{\hat{p}(z_\ell), \hat{\beta}_{0\ell}, \hat{\beta}_{1\ell}\}_{\ell=0}^K$, i.e.,

$$\hat{g}_k(\lambda) = [\hat{p}(z_k) - \hat{p}(z_0)] \lambda - c_\mu \hat{\Delta}_\mu(z_0, z_k) - c_\rho \hat{\Delta}_\rho(z_0, z_k)$$

where

$$\begin{aligned} \hat{\Delta}_\mu(z_0, z_k) &= \hat{p}(z_k) [\hat{\beta}_{10} - \hat{\beta}_{00}] - \hat{p}(z_0) [\hat{\beta}_{1k} - \hat{\beta}_{0k}] + \hat{\beta}_{1k} - \hat{\beta}_{10} \\ \hat{\Delta}_\rho(z_0, z_k) &= 2(\hat{\beta}_{00} - \hat{\beta}_{10} + \hat{\beta}_{1k} - \hat{\beta}_{0k}). \end{aligned}$$

Based on this moment condition, one might consider an Anderson-Rubin (AR) test statistic

to assess the null hypothesis $H_0 : c'\theta = \lambda$ as follows

$$\text{AR}_{n,k}(\lambda) = \left| \frac{\sqrt{n}\hat{g}_k(\lambda)}{\hat{s}_k(\lambda)} \right|^2 \quad (8)$$

where $\hat{s}_k^2(\lambda)$ consistently estimates the asymptotic variance of the sample moment $\hat{g}_k(\lambda)$. To construct this variance estimator, note that an asymptotic linear expansion of $\hat{g}_k(\lambda) - g_k(\lambda)$ gives

$$\hat{g}_k(\lambda) - g_k(\lambda) = \begin{pmatrix} \partial_{p'} g_k(\lambda)[\hat{p} - p] \\ + \partial_{\beta_1'} g_k(\lambda)[\hat{\beta}_1 - \beta_1] \\ + \partial_{\beta_0'} g_k(\lambda)[\hat{\beta}_0 - \beta_0] \end{pmatrix} + o_p(n^{-1/2}).$$

with coefficients $\partial_p g_k(\lambda)$, $\partial_{\beta_1} g_k(\lambda)$, and $\partial_{\beta_0} g_k(\lambda)$ defined as the gradient of $g_k(\lambda)$ with respect to p , β_1 , and β_0 , respectively. It is natural to estimate these coefficients by plugging in sample analog estimators below:

$$\begin{aligned} \partial_p \hat{g}_k(\lambda) &= (-\lambda + (\hat{\beta}_{1k} - \hat{\beta}_{0k})c_\mu, 0, \dots, 0, \lambda - (\hat{\beta}_{10} - \hat{\beta}_{00})c_\mu, 0, \dots, 0)' \\ \partial_{\beta_1} \hat{g}_k(\lambda) &= ((1 - \hat{p}(z_k))c_\mu + 2c_\rho, 0, \dots, 0, -(1 - \hat{p}(z_0))c_\mu - 2c_\rho, 0, \dots, 0)' \\ \partial_{\beta_0} \hat{g}_k(\lambda) &= (\hat{p}(z_k)c_\mu - 2c_\rho, 0, \dots, 0, -\hat{p}(z_0)c_\mu + 2c_\rho, 0, \dots, 0)', \end{aligned} \quad (9)$$

where nonzero elements appear on the first and the $(k+1)$ 'th elements of vectors.

This leads to a consistent variance estimator

$$\hat{s}_k^2(\lambda) = \partial_{p'} \hat{g}_k(\lambda) \hat{\Sigma}_p \partial_p \hat{g}_k(\lambda) + \partial_{\beta_1'} \hat{g}_k(\lambda) \hat{\Sigma}_{\beta_1} \partial_{\beta_1} \hat{g}_k(\lambda) + \partial_{\beta_0'} \hat{g}_k(\lambda) \hat{\Sigma}_{\beta_0} \partial_{\beta_0} \hat{g}_k(\lambda).$$

Let $\alpha \in (0, 1)$ be the significance level. The proposition below establishes the uniform validity and asymptotic similarity of the AR test for $H_0 : c'\theta = \lambda$ with $\lambda \in \mathbb{R}$.

Proposition 3.1. Let Assumption 2 and 3 hold, and suppose that the weight $c = (c_\mu, c_\rho, -c_\mu, -c_\rho)'$ is a nonzero fixed vector, then

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left(\text{AR}_{n,k}(\lambda) > q_{\chi_1^2}(1 - \alpha) \right) = \limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left(\text{AR}_{n,k}(\lambda) > q_{\chi_1^2}(1 - \alpha) \right) = \alpha,$$

where $q_{\chi_1^2}(1 - \alpha)$ denotes the $(1 - \alpha)$ -quantile of the χ_1^2 distribution.

Remark 3.1. The above result shows that the proposed AR test is valid and (uniformly) asymptotically similar in the sense of Andrews, Cheng, and Guggenberger (2020, eq.(2.6)). Constructing an asymptotically valid subvector test is already challenging in models with

weakly identified nuisance parameters⁴, because the asymptotic distributions of many test statistics depend on those unknown nuisance parameters that cannot be consistently estimated (Andrews and Mikusheva, 2016b, page 1595). However, I introduce a novel moment function that isolates the causal effects of interest while being independent of θ , thereby making the simple AR test feasible in this context. Unlike many other existing approaches to subvector inference with weak identification, this method achieves asymptotic similarity, ensuring the inverted confidence set maintains $1 - \alpha$ coverage asymptotically under both strong and weak identification. $\square$

If Z is a binary instrument taking values in $\{z_0, z_1\}$, there exists a unique AR test statistic $\text{AR}_{n,1}(\lambda)$ for robust inference on the causal parameter $c'\theta$. When the instrument is discrete with multiple values $Z \in \{z_0, \dots, z_K\}$ where $K > 1$, we obtain multiple AR statistics $\{\text{AR}_{n,k}(\lambda)\}_{k=1}^K$. The informativeness of each statistic depends on the variation in propensity scores $p(z_0) - p(z_k)$. In this case, combining tests based on different propensity score pairs yields greater statistical power. Let

$$\hat{\pi} = (\hat{p}(z_1) - \hat{p}(z_0), \dots, \hat{p}(z_K) - \hat{p}(z_0))'$$

and

$$\hat{\gamma} = (c_\mu \hat{\Delta}_\mu(z_0, z_1) + c_\rho \hat{\Delta}_\rho(z_0, z_1), \dots, c_\mu \hat{\Delta}_\mu(z_0, z_K) + c_\rho \hat{\Delta}_\rho(z_0, z_K))'.$$

Then our goal is to obtain a valid inference procedure based on a vector of linear moment functions:

$$\hat{g}(\lambda) = \hat{\pi}\lambda - \hat{\gamma} = (\hat{g}_1(\lambda), \dots, \hat{g}_K(\lambda))' \in \mathbb{R}^K.$$

Under strong identification where $\hat{\pi}$ converges to a nonzero limit, a Wald statistic based on efficient weighting matrix $\hat{S}(\lambda)^{-1}$ for testing $H_0 : c'\theta = \lambda$ can be constructed below:

$$W_n(\lambda) = \frac{n\hat{g}(\lambda)'\hat{S}(\lambda)^{-1}\hat{\pi}'\hat{S}(\lambda)^{-1}\hat{g}(\lambda)}{\hat{\pi}'\hat{S}(\lambda)^{-1}\hat{\pi}},$$

where

$$\hat{S}(\lambda) = \partial_p \hat{g}(\lambda) \hat{\Sigma}_p \partial_p' \hat{g}(\lambda) + \partial_{\beta_1} \hat{g}(\lambda) \hat{\Sigma}_{\beta_1} \partial_{\beta_1}' \hat{g}(\lambda) + \partial_{\beta_0} \hat{g}(\lambda) \hat{\Sigma}_{\beta_0} \partial_{\beta_0}' \hat{g}(\lambda) \quad (10)$$

is a consistent estimator of the asymptotic covariance matrix of $\hat{\pi}\lambda - \hat{\gamma}$ under null hypothesis.

⁴Since $c'\theta$ is the parameter of interest rather than θ itself, the vector of primitive parameters $\theta = (\mu_1, \rho_{11}, \mu_0, \rho_{01})$ becomes the weakly identified nuisance parameter under limited variation of propensity scores. Although we can show some functions of θ are strongly identified using the reparametrization technique proposed by Han and McCloskey (2019), two parameters (ρ_{11}, ρ_{01}) remain weakly identified in the transformed model.

Under weak identification, the propensity score differences $\hat{\pi}$ may converge in probability to a zero vector, resulting in a nonstandard asymptotic distribution for $W_n(\lambda)$. To illustrate this problem, consider a sequence of DGPs where $\sqrt{n}\pi$ converges to a constant vector π_0 . In this case, $\sqrt{n}\hat{\pi}$ converges in distribution to a multivariate normal distribution with mean π_0 , which cannot be consistently estimated from the data. Consequently, the asymptotic distribution of $W_n(\lambda)$ contains the non-estimable nuisance parameter π_0 , making it infeasible to consistently estimate the unconditional quantiles of $W_n(\lambda)$'s limiting distribution. However, similar to the arguments of [Moreira \(2003\)](#) for linear IV models and [Andrews and Mikushcheva \(2016b\)](#) for quasi-likelihood ratio tests in GMM models, the asymptotic distribution of $W_n(\lambda)$ becomes independent of π_0 when conditioned on a sufficient statistic, making it feasible to approximate the conditional distribution of $W_n(\lambda)$.

Define

$$\hat{h}(\lambda) = \sqrt{n}\hat{\pi} - [\partial_p \pi] \hat{\Sigma}_p [\partial_{p'} \hat{g}(\lambda)] \hat{S}(\lambda)^{-1} \sqrt{n}\hat{g}(\lambda).$$

The statistic $\hat{h}(\lambda)$ is asymptotically independent of $\sqrt{n}\hat{g}(\lambda)$ and contains sufficient information about π_0 such that the (asymptotic) distribution of $W_n(\lambda)$ conditional on $\hat{h}(\lambda)$ is free of π_0 . To simulate this conditional distribution, I construct a simulated counterpart of $\sqrt{n}\hat{\pi}$, denoted as π_s , given by

$$\pi_s = \hat{h}(\lambda) + [\partial_p \pi] \hat{\Sigma}_p [\partial_{p'} \hat{g}(\lambda)] \hat{S}(\lambda)^{-1/2} \eta^*$$

where $\eta^* \sim \mathcal{N}(0_{K \times 1}, I_{K \times K})$ is drawn independently of the data. We then approximate the distribution of $W_n(\lambda)$ by replacing $\sqrt{n}\hat{\pi}$ with π_s and $\sqrt{n}\hat{g}(\lambda)$ with $\hat{S}(\lambda)^{1/2} \eta^*$, which yields

$$W_n^*(\lambda) = \frac{(\eta^*)' \hat{S}(\lambda)^{-1/2} \pi_s \pi_s' \hat{S}(\lambda)^{-1/2} (\eta^*)}{\pi_s' \hat{S}(\lambda)^{-1} \pi_s}.$$

It follows that $W_n^*(\lambda)$ has the same asymptotic distribution as $W_n(\lambda)$ conditional on the realization of $\hat{h}(\lambda)$. Let $\hat{q}_{W^*}(1 - \alpha)$ denote the $(1 - \alpha)$ -quantile of the distribution of $W_n^*(\lambda)$ conditional on the data. This critical value can be used to construct valid conditional Wald test:

Theorem 3.1. Let Assumption 2 and 3 hold, and suppose that the weight $c = (c_\mu, c_\rho, -c_\mu, -c_\rho)'$ is a nonzero fixed vector, then

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(W_n(\lambda) > \hat{q}_{W^*}(1 - \alpha)) = \limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(W_n(\lambda) > \hat{q}_{W^*}(1 - \alpha)) = \alpha.$$

Remark 3.2. The statistic $W_n(\lambda)$ can be interpreted as a linear combination of AR statistics

in equation (8), with weight $\hat{S}^{-1}\hat{\pi}$. Under strong identification, this weighting vector can be shown to be optimal among all linear combinations of AR statistics, yielding the highest local asymptotic power. $\square$

Remark 3.3. In Lemma D.2, I show that the asymptotic variance of the moment function is uniformly positive definite as long as the weight $c = (c_\mu, c_\rho, -c_\mu, -c_\rho)$ is nonzero. If we incorporate additional moment conditions induced by the difference of propensity scores of the form $p(z_k) - p(z_j)$ for $k, j \neq 0$, then the asymptotic variance may become singular for some nonzero weight c . For example, if we set $c_\mu = 0$ and $c_\rho = 1$, it follows that

$$[\hat{p}(z_k) - \hat{p}(z_j)]\lambda - c_\rho \hat{\Delta}_\rho(z_j, z_k) = \hat{g}_k(\lambda) - \hat{g}_j(\lambda),$$

implying that the moment condition constructed with the variation between z_k and z_j (on the left-hand side) is the difference of the moments constructed by using (z_k, z_0) and (z_j, z_0) (on the right-hand side). $\square$

Remark 3.4. An alternative statistic one could consider is the quasi-likelihood ratio statistic

$$\text{QLR}_n(\lambda) = n \left[\hat{g}(\lambda)' \hat{S}(\lambda)^{-1} \hat{g}(\lambda) - \inf_{\lambda \in \mathbb{R}} \hat{g}(\lambda)' \hat{S}(\lambda)^{-1} \hat{g}(\lambda) \right]$$

and its corresponding conditional approach discussed in Andrews and Mikusheva (2016b). Here I focus on the conditional Wald approach because it is simpler to implement in practice and is first-order asymptotically equivalent to QLR test under strong identification. $\square$

4 Robust Inference in General MTE Models

In this section, I relax the linearity Assumption 3 and extend the analysis to allow for any functional forms specified in Assumption 1.5. This includes the linear MTE model and the polynomial specification $h_m(\cdot) = u^m - \frac{1}{m+1}$ that is commonly used in empirical studies.

For this broader class of models, constructing moment conditions that only involve the target causal parameter is less obvious. Therefore, I develop an alternative *improved projection* approach that builds on I. Andrews (2018) to achieve valid inference. When compared to the conditional Wald test in section 3, this new approach applies to a wider range of MTE models but is more computationally intensive and is potentially more conservative under weak identification. On the theoretical front, I argue that the high-level conditions imposed by I. Andrews (2018) cannot be directly verified in the MTE setting. Therefore, I

extend his sequential validity result (that based on high-level conditions) by establishing the uniform validity under primitive conditions on the parameter space outlined in Definition 2.1. This allows the correct coverage of resulting confidence set in large sample, even when the propensity scores are close to one another or the coefficients are partially identified.

4.1 The MLC test and uniform validity

Following the separate regression approach in Section 2, we conduct inference based on the linear system of moment equations:

$$A\theta = \beta \tag{11}$$

where

$$A = \begin{pmatrix} A_1 & 0_{(K+1) \times (M+1)} \\ 0_{(K+1) \times (M+1)} & A_0 \end{pmatrix} \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_0 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix},$$

as defined below equation (5). Note that rows of A depend on the propensity scores $p(z) = \mathbb{P}(D = 1 \mid Z = z)$ and β collects $\mathbb{E}[Y \mid D = d, Z = z]$, both of which can be consistently estimated by their sample analogs $\hat{A}$ and $\hat{\beta}$ if we plug in the sample propensity scores $\hat{p}(z)$ for the construction of $\hat{A}$.

Let $\hat{m}(\theta) = \hat{A}\theta - \hat{\beta}$ denote the sample moment function, and let $\hat{\Omega}(\theta)$ denote an estimator of the asymptotic variance of $\sqrt{n} \hat{m}(\theta)$, whose explicit form is given in Section 4.3. Under weak identification, the matrix $\hat{A}$ is close to singular, so standard inference procedures that treat it as being bounded away from full rank can be badly distorted. Therefore, I replace $\hat{A}$ with a modified Jacobian that remains well behaved under weak identification. The construction follows by two steps.

First, following Kleibergen (2005), I orthogonalize each column $\hat{a}_j$ of $\hat{A}$ so that it is asymptotically uncorrelated with the sample moment:

$$\hat{d}_j(\theta) = \hat{a}_j - \hat{\Gamma}_j(\theta) \hat{\Omega}(\theta)^{-1} \hat{m}(\theta), \quad j = 1, \dots, 2(M+1), \tag{12}$$

where $\hat{\Gamma}_j(\theta)$ estimates the asymptotic covariance between $\sqrt{n}(\hat{a}_j - a_j)$ and $\sqrt{n} \hat{m}(\theta)$. Let $\hat{D}(\theta) = (\hat{d}_1(\theta), \dots, \hat{d}_{2(M+1)}(\theta))$ collect the orthogonalized columns.

Second, I perturb the orthogonalized Jacobian by a small random term of order $n^{-1/2}$:

$$\tilde{D}(\theta) = \hat{D}(\theta) + \kappa n^{-1/2} \xi, \tag{13}$$

where $\kappa > 0$ is a fixed constant and $\xi \in \mathbb{R}^{2(K+1) \times 2(M+1)}$ is a random matrix with i.i.d. standard normal entries, drawn independently of the data. This perturbation ensures that $\tilde{D}(\theta)$, after suitable rescaling and rotation, has full column rank almost surely in the limit along every weakly identified sequence in the parameter space. The order $n^{-1/2}$ is chosen such that (1) In weakly identified directions, the singular values of $\hat{D}(\theta)$ are themselves of order $n^{-1/2}$, so a perturbation of the same order is needed to rule out a rank-deficient limit, and (2) Under strong identification, by contrast, $\hat{D}(\theta)$ converges to a full-rank matrix and the perturbation vanishes asymptotically, so the local efficiency of the procedure remains unaffected. Following D. Andrews (2017), I set $\kappa = 10^{-6}$ in the simulations and in the empirical application.

Using the modified Jacobian $\tilde{D}(\theta)$, I define the modified robust Lagrange multiplier (MRLM) statistic

$$\text{MRLM}_n(\theta) = n \hat{m}(\theta)' \hat{\Omega}(\theta)^{-1/2} P_{\hat{Q}(\theta)} \hat{\Omega}(\theta)^{-1/2} \hat{m}(\theta), \quad (14)$$

where $P_{\hat{Q}(\theta)}$ denotes the orthogonal projection onto the span of

$$\hat{Q}(\theta) = \hat{\Omega}(\theta)^{-1/2} \tilde{D}(\theta) (\tilde{D}(\theta)' \hat{\Omega}(\theta)^{-1} \tilde{D}(\theta))^{-1} c \in \mathbb{R}^{2(K+1)},$$

and c is the vector that defines the causal parameter of interest $c'\theta$ (e.g., Table 1). The vector $\hat{Q}(\theta)$ selects the efficient direction for testing hypotheses on $c'\theta$. This MRLM statistic differs from the robust LM statistics of Kleibergen (2005), Chaudhuri and Zivot (2011), and Andrews (2018) since the MRLM statistic uses the perturbed Jacobian $\tilde{D}(\theta)$ in place of the orthogonalized Jacobian $\hat{D}(\theta)$. This construction is the key that delivers uniform validity across all identification strengths.

I also define the Anderson–Rubin (AR) statistic

$$\text{AR}_n(\theta) = n \hat{m}(\theta)' \hat{\Omega}(\theta)^{-1} \hat{m}(\theta), \quad (15)$$

and combine the two into the modified linear combination (MLC) statistic

$$\text{MLC}_n(\theta) = \text{MRLM}_n(\theta) + a \cdot \text{AR}_n(\theta), \quad (16)$$

where $a > 0$ is a tuning parameter that sets the weight on the AR statistic. The two components play complementary roles. Under strong identification, $\text{AR}_n(\theta)$ diverges for every θ whose distance from the true value θ^* is of larger order than $n^{-1/2}$. The AR term therefore forces the test to have power against all such ‘distant’ alternatives. Within $n^{-1/2}$ -

neighborhoods of θ^* , however, the MRLM term dominates when the weight a is small. The projection test based on MLC_n is then approximately equivalent to the standard Wald test for $c'\theta$ under strong identification. The choice of a thus involves a trade-off. A small value keeps the procedure close to the efficient Wald test under strong identification, while a larger value places more weight on the AR statistic and on power against distant alternatives. I set $a = 0.05$ in the simulations and the empirical application. This choice is close to the weight suggested by I. Andrews (2018, page 343) for the dimensions used in the empirical application of Section 6.

To test $H_0: c'\theta = \lambda$ at significance level $\alpha \in (0, 1)$, I use the projection test

$$\hat{\phi}_{\text{MLC}}(\lambda) = \mathbb{1} \left[\inf_{\theta \in \Theta: c'\theta = \lambda} \text{MLC}_n(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2}^2(1 - \alpha) \right], \quad (17)$$

where $q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2}^2(1 - \alpha)$ denotes the $(1 - \alpha)$ -quantile of $(1 + a)\chi_1^2 + a\chi_{2K+1}^2$, with χ_1^2 and χ_{2K+1}^2 independent. For $\lambda \notin \{c'\theta : \theta \in \Theta\}$, I set $\hat{\phi}_{\text{MLC}}(\lambda) = 1$ by convention. Inverting the test yields the confidence set

$$\mathcal{C}_{\text{MLC}} = \{\lambda \in \mathbb{R} : \hat{\phi}_{\text{MLC}}(\lambda) = 0\}. \quad (18)$$

In practice, I report the convex hull of $\mathcal{C}_{\text{MLC}}$. The convex hull is an interval and contains $\mathcal{C}_{\text{MLC}}$, so it maintains the coverage guarantee below.

Theorem 4.1. Let Assumption 2 hold, and suppose that the weight c is a nonzero fixed vector. Then we have

$$\limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{E}_F[\hat{\phi}_{\text{MLC}}(\lambda)] \leq \alpha.$$

Theorem 4.1 shows that the MLC test controls asymptotic size uniformly over the null parameter space $\mathcal{P}_0$ defined in Definition 2.1. Consequently, $\mathcal{C}_{\text{MLC}}$ has asymptotic coverage of at least $1 - \alpha$ even when Assumption 1.7 fails or is close to failing. When θ is only partially identified, the coverage guarantee holds for every value of $c'\theta$ in the identified set. To achieve this robustness, the resulting test may be conservative. Because the test profiles over the directions of θ not pinned down by $c'\theta = \lambda$, its null rejection probability can be strictly below α , particularly under weak identification. This conservativeness is not substantive when identification is strong. In Online Appendix E, I show that the MLC test is consistent for distant (or fixed) alternatives and has comparable local power to the asymptotically efficient Wald test under strong identification when the AR weight a is sufficiently small. These findings highlight the usefulness of the MLC test for researchers seeking robust inference against weak identification.

The critical value $q_{(1+a)\chi_1^2+a\chi_{2K+1}^2}(1-\alpha)$ follows from a decomposition of the MLC statistic at the true value θ^* . The AR statistic measures the deviation of $2(K+1)$ sample moments from zero, relative to their sampling variability. The MRLM statistic keeps only the part of this deviation along $\hat{Q}(\theta^*)$, the direction most informative about $c'\theta$. The remaining $2K+1$ directions account for the difference $\text{AR}_n(\theta^*) - \text{MRLM}_n(\theta^*)$. Rewriting the MLC statistic in terms of these two components gives

$$\text{MLC}_n(\theta^*) = (1+a) \text{MRLM}_n(\theta^*) + a [\text{AR}_n(\theta^*) - \text{MRLM}_n(\theta^*)].$$

The two components, $\text{MRLM}_n(\theta^*)$ and $\text{AR}_n(\theta^*) - \text{MRLM}_n(\theta^*)$, are asymptotically independent χ^2 variables with one and $2K+1$ degrees of freedom, respectively. Hence, $\text{MLC}_n(\theta^*)$ is asymptotically distributed as the weighted sum $(1+a)\chi_1^2 + a\chi_{2K+1}^2$.

4.2 Relation to Current Literature

The MLC test builds on the linear combination (LC) test of I. Andrews (2018), which combines a subvector robust LM statistic with the AR statistic for inference on functions of weakly identified parameters. The main difference from I. Andrews (2018) lies in the type of weak identification. See Appendix C for technical details. His validity result is sequential and hinges on high-level conditions (in Assumption 4) that require the orthogonalized Jacobian, after suitable normalization, to converge to a Gaussian matrix of full rank. These conditions can be trivially satisfied when weak identification is of the local-to-zero type, that is, when the Jacobian (or a block of its columns) shrinks to zero at rate $n^{-1/2}$, as in Stock and Wright (2000), Kleibergen (2005), and Chaudhuri and Zivot (2011). In the MTE model, however, weak identification can happen whenever the propensity scores approach a set on which Assumption 1.7 fails, so that A converges to a *singular* but *nonzero* matrix. Along such local-to-singularity sequences, the orthogonalized Jacobian, after normalization, may have a rank-deficient limit, and the high-level conditions of I. Andrews (2018) need not hold. I address this problem in two steps. First, I normalize the orthogonalized Jacobian through a singular value decomposition (SVD) of A , which extends the argument of Andrews and Guggenberger (2017) from the full parameter vector to linear functionals $c'\theta$. Second, I perturb the orthogonalized Jacobian as in (13) by drawing insights from the AR/LM test constructed by D. Andrews (2017), so that its normalized limit has full rank almost surely. These two steps deliver uniform validity established by Theorem 4.1 under the primitive conditions in Definition 2.1, which addresses, for the inference in a nearly singular linear system, the question of uniform validity that I. Andrews (2018, page 347) leaves open.

The closest related work on inference under near singularity is [Han and McCloskey \(2019\)](#), who study GMM models with a nearly singular Jacobian. In their framework, the singularity arises only through a *single* source as a parameter vector measuring identification strength approaches zero. In contrast, the higher-order MTE model with $M \geq 2$ does not fit this structure, because Assumption 1.7 does not hold whenever the propensity scores take fewer than $M + 1$ distinct values. This implies identification can fail along *multiple* directions as the propensity scores approach a set, rather than a point. Another related approach is the refined Bonferroni test of D. [Andrews \(2017\)](#), which is uniformly valid for subvector inference. However, it requires additional tuning parameters to categorize identification strength and does not directly address inference on a linear function $c'\theta$, which is the focus of this paper. In principle, his test could be applied to $c'\theta$ after reparametrizing the model, but finding a universal reparametrization rule that works for all linear hypotheses of interest while keeping the asymptotic analysis tractable remains challenging. Like the procedures of [Chaudhuri and Zivot \(2011\)](#), D. [Andrews \(2017\)](#), and I. [Andrews \(2018\)](#), the MLC test is not asymptotically similar under weak identification, as it profiles over the nuisance directions of θ . In this respect the MLC test complements the conditional Wald test of Section 3, which is asymptotically similar but limited to linear MTE models. Finally, it is crucial to note that existing inference results for MTE models ([Carneiro and Lee, 2009](#); [Mogstad et al., 2018](#); [Sasaki and Ura, 2023](#)) do not consider weak identification.

4.3 Implementation of the MLC test

In this section, I describe the implementation of the projection test based on the MLC statistic.

1. Construct the key quantities for the test statistics.

- (a) Construct the estimators $\hat{p}$ and $\hat{\beta}$, as well as their asymptotic variance estimators $\hat{\Sigma}_p$ and $\hat{\Sigma}_\beta$, as in section 2.4. Plugging $\hat{p}$ into A then gives the estimator $\hat{A}$.
- (b) Estimate the asymptotic variance of $\sqrt{n} \hat{m}(\theta)$ by

$$\hat{\Omega}(\theta) = H(\hat{p}, \theta) \hat{\Sigma}_p H(\hat{p}, \theta)' + \hat{\Sigma}_\beta, \quad (19)$$

where

$$H(p, \theta) = \begin{bmatrix} \text{diag} \left\{ \sum_{m=0}^M \theta_{1m} \lambda'_{1m}(p(z_\ell)) : \ell = 0, 1, \dots, K \right\} \\ \text{diag} \left\{ \sum_{m=0}^M \theta_{0m} \lambda'_{0m}(p(z_\ell)) : \ell = 0, 1, \dots, K \right\} \end{bmatrix}. \quad (20)$$

- (c) Compute the orthogonalized columns $\hat{d}_j(\theta)$, $j = 1, \dots, 2(M+1)$, defined in (12), using the covariance estimator

$$\hat{\Gamma}_j(\theta) = M_j(\hat{p})\hat{\Sigma}_p H(\hat{p}, \theta)', \quad (21)$$

where

$$M_j(\hat{p}) = \begin{cases} \begin{bmatrix} L_{j-1}(\hat{p}) \\ 0_{(K+1) \times (K+1)} \end{bmatrix} & \text{if } j \leq M+1, \\ \begin{bmatrix} 0_{(K+1) \times (K+1)} \\ R_{j-M-2}(\hat{p}) \end{bmatrix} & \text{if } j > M+1, \end{cases} \quad (22)$$

with, for $m = 0, 1, \dots, M$,

$$\begin{aligned} L_m(\hat{p}) &= \text{diag}\{\lambda'_{1m}(\hat{p}(z_0)), \dots, \lambda'_{1m}(\hat{p}(z_K))\}, \\ R_m(\hat{p}) &= \text{diag}\{\lambda'_{0m}(\hat{p}(z_0)), \dots, \lambda'_{0m}(\hat{p}(z_K))\}. \end{aligned}$$

- (d) Collect the columns into $\hat{D}(\theta)$ and compute the perturbed Jacobian $\tilde{D}(\theta)$ in (13), with $\kappa = 10^{-6}$ following D. Andrews (2017).
2. For each $\theta \in \Theta$, compute the MRLM statistic (14), the AR statistic (15), and the MLC statistic (16). For the empirical application and simulation studies, I set the AR weight to $a = 0.05$, which is close to the suggested weight $a(\gamma)$ of I. Andrews (2018, page 343) with $\alpha = 5\%$, $\gamma = 10\%$, and $K = 14$, consistent with the application in section 6.
 3. For a significance level $\alpha \in (0, 1)$, profile the MLC statistic over the linear manifold $\{\theta \in \Theta : c'\theta = \lambda\}$ and reject $H_0: c'\theta = \lambda$ according to the projection test (17), whose critical value is the $(1 - \alpha)$ -quantile of $(1 + a)\chi_1^2 + a\chi_{2K+1}^2$ with $\chi_1^2 \perp \chi_{2K+1}^2$.

Inverting the test gives the confidence set $\mathcal{C}_{\text{MLC}}$ in (18), and I report its convex hull.

5 Monte Carlo Simulation

In this section, I compare the finite-sample performance of the proposed MLC test with the classical Wald test based on moment conditions (11) using simulated data from a simple quadratic MTE model with a three-valued instrument. The power comparison between

conditional Wald tests, MLC tests, and classical Wald tests in a linear MTE model can be found in Online Appendix [G](#).

Consider a DGP specified as follows: For each $d = 0, 1$,

$$\begin{aligned} Y_d &= \mu_d + V_d \\ D &= \mathbb{1}[U \leq p(Z)] \\ V_d &= \rho_{d1} \left(U - \frac{1}{2} \right) + \rho_{d2} \left(U^2 - \frac{1}{3} \right) + e_d, \end{aligned}$$

where Z is uniformly distributed over three points $\{z_0, z_1, z_2\}$ and is independent of (U, e_1, e_0) ; U is uniformly distributed over $[0, 1]$ and is independent of the error term (e_1, e_0) , which follow a joint normal distribution with zero mean and covariance matrix $\Sigma_e = 0.5 \cdot I_{2 \times 2}$. For simplicity, I set $\mu_1 = \mu_0 = 0$ and impose strong endogeneity by specifying

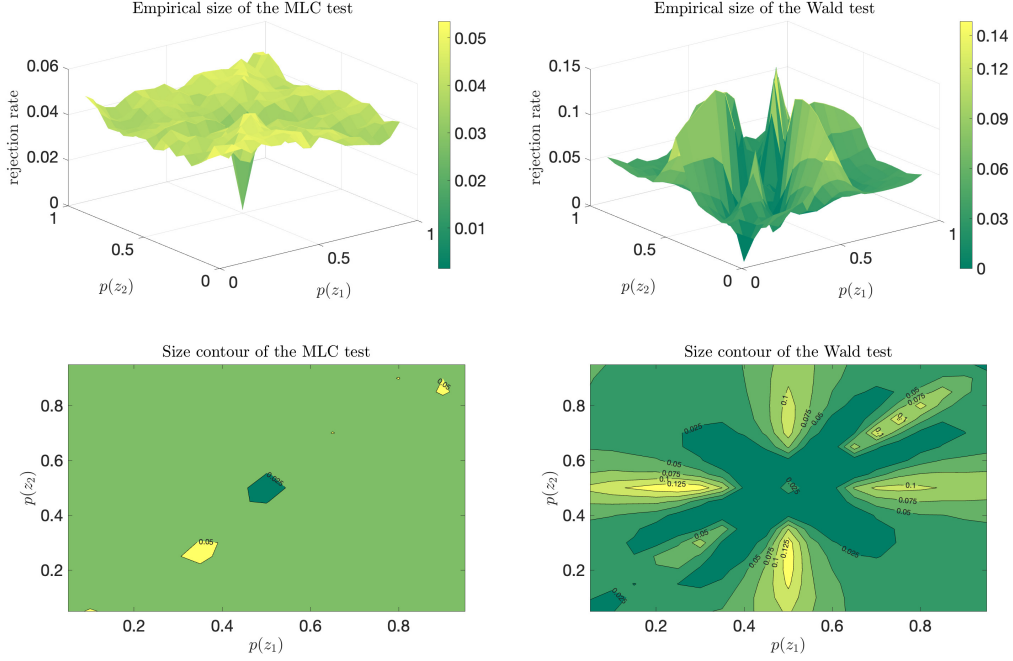
$$\rho_{11} = \rho_{12} = -5 \quad \text{and} \quad \rho_{01} = \rho_{02} = 5.$$

Regarding the specification of the propensity score $p(z)$ for $z \in \{z_0, z_1, z_2\}$, I fix $p(z_0) = 0.5$ and let $(p(z_1), p(z_2))$ vary across $(0.05, 0.95)^2$ to generate a variety of degrees/directions of weak identification. By drawing 2,000 i.i.d. simulation samples, I implement the Wald test and the MLC test with AR weight $a = 0.05$ to assess the null hypothesis on testing ATE $H_0 : \mu_1 - \mu_0 = 0$. The average null rejection rates based on 5% significance level are displayed in Figure [2](#).

The null rejection rates reveal distinct patterns in tests' performance across propensity score variation. The conventional Wald test exhibits under-rejection when all three propensity scores are close, indicating trivial power in this weakly identified scenario to be shown in Figure [3\(c\)](#). When the propensity scores cluster at two points (i.e., when either $p(z_1)$ or $p(z_2)$ equals $p(z_0) = 0.5$), the ATE parameter becomes partially identified, and the Wald test's rejection rates reach 14%, substantially exceeding the nominal 5% significance level. This demonstrates the Wald test's invalidity under identification failure. In contrast, the proposed MLC test maintains proper size control across all propensity score combinations, demonstrating its robustness against weak identification.

Figure [3](#) compares the power of the MLC and Wald tests across different identification scenarios. In Panel (a), where propensity scores are well-separated (strong identification), the Wald test achieves asymptotic efficiency, and the MLC test delivers comparable rejection power. In Panel (b), where the propensity score takes only two distinct values (partial identification), the Wald test exhibits not only size distortion at the null but also power

Figure 2: 3D Plot of Empirical Size for MLC and Wald Tests



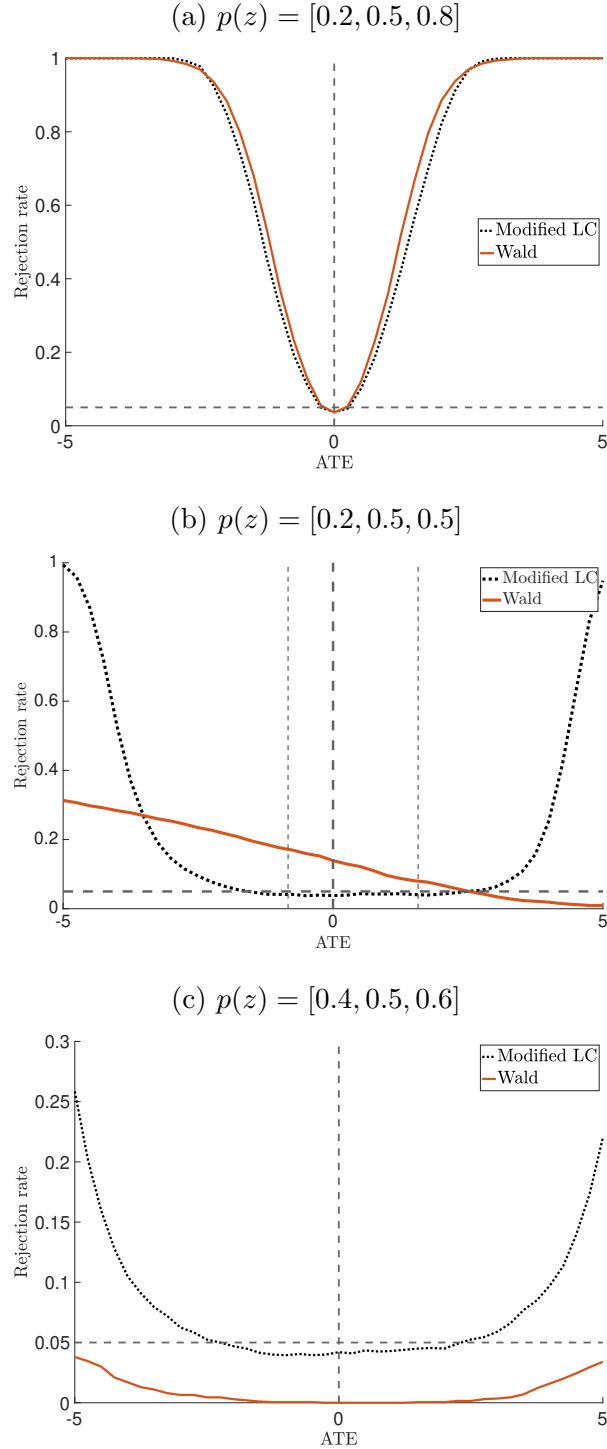
Note: The plots show empirical rejection rates for testing the null hypothesis $H_0 : \text{ATE} = 0$ at the 5% significance level. The upper panel presents three-dimensional surfaces of rejection rates, while the lower panel shows the corresponding contour plots. Each plot varies $p(z_1)$ and $p(z_2)$ across (0.05, 0.95) with $p(z_0)$ fixed at 0.5. The contour lines correspond to rejection rates of 2.5%, 5%, 7.5%, 10%, and 12.5%, with regions between these lines representing intermediate rejection rates. Darker regions indicate lower rejection rates.

loss at alternatives. In contrast, the MLC test maintains correct size and almost surely rejects distant alternatives outside the identified set. In Panel (c), where propensity scores are clustered together (weak identification), the Wald test's rejection rates fall consistently below the nominal level, indicating negligible power, while the MLC test retains nontrivial power at distant alternatives.

6 Empirical Application to Misdemeanor Prosecution

In this section, I revisit the empirical analysis by [Agan et al. \(2023\)](#), who examined the causal effects of misdemeanor prosecution on defendants' subsequent criminal activity. Based on the quasi-randomized assignment of nonviolent misdemeanor cases to assistant district attorneys (ADAs), their LATE and MTE estimates demonstrate that nonprosecution leads to a large reduction in the likelihood of defendants' future criminal involvement over the next two

Figure 3: Power Curves of the MLC and Wald Test



Note: Testing ATE at values on $[-5, 5]$ with the true effects fixed at zero. The significance level is set at 5%. The sample size equals 2,000 and the average rejection rates are computed with 2,000 independent Monte-Carlo simulations. The two dashed lines around zero in subfigure (b) characterize the identified set of ATE by imposing a parameter space $[-10, 10]$ for each element of θ .

years. To study the policy effects of increasing nonprosecution, they analyzed the impact of imposing a presumption of misdemeanor nonprosecution by relying on an existing policy change issued by a new district attorney. In this section, I use their data to answer some policy-relevant questions concerning the exogenous change in ADA nonprosecution rates. My approach does not require the additional data or information related to an actual policy that is already implemented, but instead extrapolates causal effects to address this issue.

Similar to their analysis, I use the MTE framework to extrapolate treatment effects outside compliers. However, my goal is to analyze causal effects of implementing several counterfactual policies on ADA leniency, which were not explored in their empirical studies. To adapt their data into the framework considered in this paper, I implement the proposed inference procedure conditional on each court and use ADAs' identity as the discrete instrument since ADAs are randomly assigned conditional on courts and time. To avoid bias introduced by many instruments (see the references in [Mikusheva and Sun \(2022\)](#) for further discussions), I combine ADAs into a total of 15 groups for each court, excluding the smaller courts (BRI, CHE, and CHA) due to insufficient ADA counts.

Let the treatment D be the nonprosecution status of defendants (that takes value 1 if the defendant is not prosecuted) and the outcome Y be the indicator of subsequent criminal complaints within two years postarrest. To validate the strong overlap condition on propensity scores, I focus on ADAs that handle more than 50 cases and have nonprosecution rate at least 0.025 within each court. The summary statistics of ADAs' nonprosecution rates conditional on each court are provided in Table 2 below. From this table we can see that the range of propensity scores varies from 0.21 to 0.50 across different courts. When employing a high-order polynomial MTE model such as the cubic specification in Figure IV of [Agan et al. \(2023\)](#) in this setup, the finite-sample performance of confidence sets may be too poor to guarantee valid coverage. Specifically, in Online Appendix H, I show the evidence of weak instruments in cubic and quartic MTE models, while the classical F test (with threshold 10) does not deliver the same conclusion. The reason is that F statistic aims to detect deviations from the null where all propensity scores are equal to each other. However, this is not sufficient to strongly identify MTE models with flexible structure that require three or more propensity score to be well separated from each other (see Figure 1).

For all the counterfactual experiments, I consider an exogenous change to the ADA nonprosecution rates. Let $p^\epsilon(z)$ be the counterfactual propensity score of not prosecuting defendants, where ϵ is a nonnegative scalar (or vector) denoting the degree of deviation from

Table 2: Summary Statistics of Nonprosecution Rates across Courts

Court	Number of ADA	Nonprosecution Rate			Sample Size
		Min	Mean	Max	
SBO	16	0.04	0.24	0.54	3,921
EBOS	22	0.08	0.32	0.53	7,566
WROX	43	0.05	0.37	0.55	8,905
BMC	65	0.04	0.17	0.4	9,593
ROX	66	0.06	0.16	0.33	13,333
DOR	73	0.04	0.13	0.43	13,523
CHA	4	0.09	0.24	0.42	362
CHE	12	0.07	0.16	0.3	872
BRI	5	0.14	0.29	0.48	885

Notes: Courts include South Boston (SBO), East Boston (EBOS), West Roxbury (WROX), Central (BMC), Roxbury (ROX), Dorchester (DOR), Charlestown (CHA), Chelsea (CHE), and Brighton (BRI).

status quo. I compare the induced outcome Y^ϵ , defined as

$$Y^\epsilon = Y_1 \mathbb{1}[U \leq p^\epsilon(Z)] + Y_0 \mathbb{1}[U > p^\epsilon(Z)]$$

with the observed outcome Y in the data.

Similar to the policy invariance assumption imposed by [Heckman and Vytlacil \(2005\)](#), suppose that this counterfactual change on propensity score does not shift the distribution of potential outcomes and unobserved cost U of being selected into treatment, therefore omitting any “general equilibrium” effects that may arise after the change of prosecution rates. I consider three different comparisons between the counterfactual outcome Y^ϵ and the observed outcome Y :

1. Non-normalized policy effects:

$$\alpha(\epsilon) \equiv \mathbb{E}[Y^\epsilon - Y]. \quad (23)$$

This criterion follows from [Heckman and Vytlacil \(2001\)](#). The term “non-normalized” distinguishes these effects from the policy-relevant treatment effects discussed below.

2. Policy-relevant treatment effects (PRTE)

$$\bar{\alpha}(\epsilon) \equiv \frac{\mathbb{E}[Y^\epsilon - Y]}{\mathbb{E}[p^\epsilon(Z) - p(Z)]}. \quad (24)$$

This criterion follows from [Heckman and Vytlacil \(2005\)](#). This quantity can be inter-

preted as the treatment effect for units shifted into treatment via the counterfactual policy. Note that the additive PRTEs suggested by [Carneiro et al. \(2010, 2011\)](#) and [Mogstad and Torgovitsky \(2018\)](#) are special cases of $\bar{\alpha}(\epsilon)$ for particular choices of p^ϵ , as described in equations (26) below.

3. Marginal policy-relevant treatment effects (MPRTE)

$$\alpha_+(0) \equiv \lim_{\epsilon \searrow 0} \frac{\mathbb{E}[Y^\epsilon - Y]}{\mathbb{E}[p^\epsilon(Z) - p(Z)]}. \quad (25)$$

This criterion follows from [Carneiro et al. \(2010\)](#). When $p^\epsilon(z) \geq p(z)$ for all $z \in \text{supp}(Z)$, this parameter can be interpreted as the average treatment effects for marginal defendants at the edge of being prosecuted.

Next, I discuss the counterfactual propensity scores of interest and present confidence sets for these policy effects.

6.1 Marginal policy relevant treatment effects

First, suppose policymakers increase the ADAs’ nonprosecution rates by a positive constant ϵ , where $\epsilon > 0$. Consider the additive marginal PRTE $\alpha_+(0)$ under the new policy:

$$\text{Additive change: } p^\epsilon(z) = p(z) + \epsilon, \quad (26)$$

Table 3 collects the 95% and 90% confidence sets on marginal policy effects $\alpha_+(0)$ for the additive leniency increase in equation (26) by letting ϵ approach zero. The “unconditional” confidence sets are obtained without controlling for court identities, whereas the “average” confidence sets are weighted averages of court-specific confidence sets weighted by the size of courts. The differences between these two sets of results highlight the importance of adjusting for court identity as a confounding variable. Incorporating court identity generally increases the uncertainty of causal effect predictions.

Comparing average Wald confidence sets with average MLC confidence sets, the results show that weak identification is a potential concern for cubic and quartic specifications. In such cases, robust confidence sets may not be informative about the sign of causal effects due to the inability to estimate a highly flexible model by using limited variation of propensity scores. Although the ranges of both confidence sets are similar under the quadratic specification, Wald confidence sets remain negatively significant, whereas MLC confidence sets

Table 3: 95% (top) and 90% (bottom) Confidence Sets for Additive MP RTE $\alpha_+(0)$

	MTE polynomial	Wald	MLC
Uncond.	Linear	[-0.19, -0.11]	[-0.17, -0.13]
Uncond.	Quadratic	[-0.19, -0.11]	[-0.18, -0.13]
Uncond.	Cubic	[-0.17, -0.06]	[-0.15, -0.08]
Uncond.	Quartic	[-0.18, -0.05]	[-0.17, -0.06]
Average	Linear	[-0.24, -0.03]	[-0.23, -0.03]
Average	Quadratic	[-0.29, -0.01]	[-0.33, 0.09]
Average	Cubic	[-0.32, 0.03]	[-0.79, 0.59]
Average	Quartic	[-0.39, 0.01]	[-0.76, 0.61]
Uncond.	Linear	[-0.18, -0.12]	[-0.16, -0.14]
Uncond.	Quadratic	[-0.19, -0.12]	[-0.17, -0.14]
Uncond.	Cubic	[-0.17, -0.07]	[-0.14, -0.09]
Uncond.	Quartic	[-0.17, -0.06]	[-0.13, -0.08]
Average	Linear	[-0.22, -0.04]	[-0.21, -0.06]
Average	Quadratic	[-0.27, -0.04]	[-0.28, 0.01]
Average	Cubic	[-0.29, -0.00]	[-0.60, 0.34]
Average	Quartic	[-0.35, -0.02]	[-0.63, 0.38]

Note: Results for the classical Wald and MLC confidence sets under polynomial specifications up to the fourth order. The counterfactual policy considered is a uniformly additive increase in nonprosecution rates by a marginal amount. The top panel reports the 95% confidence sets, and the bottom panel reports the 90% confidence sets. The “Uncond.” confidence sets are obtained without controlling for court identities, whereas the “Average” confidence sets are weighted averages of the court-specific confidence sets, with weights proportional to court size.

lose such significance due to weak identification. In Online Appendix H (Table 5), I evaluate the strength of identification for the marginal PRTEs based on the additive change in (26). The findings further confirm the presence of weak identification for MTE models with cubic or quartic orders.

6.2 Quota

Instead of mandating all ADAs to increase their nonprosecution rates simultaneously, it might be more convenient for policymakers to set up lower and upper bounds for ADA nonprosecution rates. For example, let $\epsilon = (\underline{\epsilon}, \bar{\epsilon}) \in [0, 1]^2$ with $\underline{\epsilon} \leq \bar{\epsilon}$ denoting the lower and upper bounds of ADAs’ nonprosecution rate, respectively. Then the counterfactual propensity score becomes

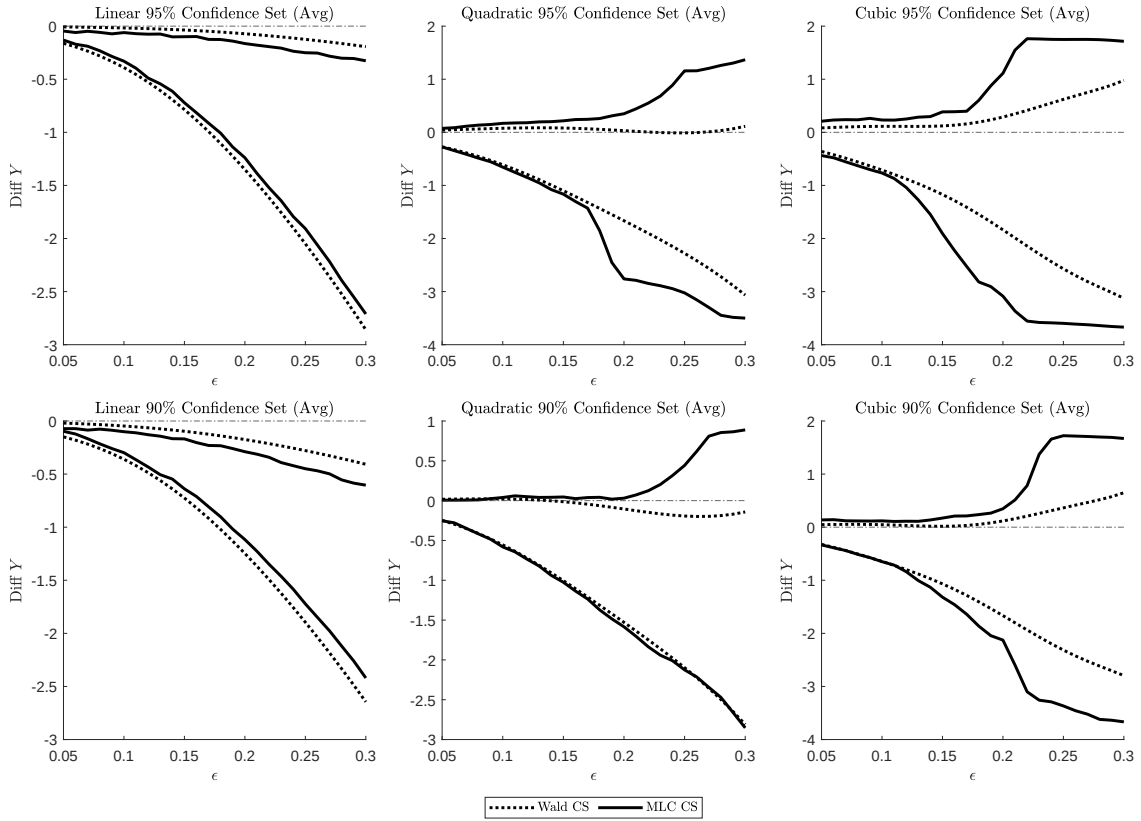
$$p^\epsilon(z) = \min(\max\{\underline{\epsilon}, p(z)\}, \bar{\epsilon})$$

Since this counterfactual propensity score is non-smooth in $p(z)$ which may invalidate the inferential results, I employ a smooth approximation as follows:

$$p^\epsilon(z; \phi) = -\frac{1}{\phi} \log \left(\frac{1}{e^{\phi p(z)} + e^{\phi \underline{\epsilon}}} + \frac{1}{e^{\phi \bar{\epsilon}}} \right).$$

One can show that $p^\epsilon(z; \phi) \rightarrow p^\epsilon(z)$ as $\phi \rightarrow \infty$. In practice, I set $\phi = 30$ to approximate the counterfactual propensity score $p^\epsilon(z)$. I analyze the (non-)normalized policy effects when policymakers implement a lower bound $\underline{\epsilon} \in [0.05, 0.3]$ while maintaining $\bar{\epsilon} = 1$ for each court. For notational consistency, I use ϵ to denote the value of this lower bound instead of $\underline{\epsilon}$.

Figure 4: 95% (top) and 90% (bottom) Average Confidence Sets for Non-normalized Policy Effects $\alpha(\epsilon)$ when Setting a Lower Bound $\underline{\epsilon}$ for Nonprosecution Rate



Note: Results for the classical Wald and MLC confidence sets under polynomial specifications up to the third order. The counterfactual policy under consideration is a universal lower bound on the nonprosecution rates shown on the x -axis. The y -axis depicts changes in recidivism rates (in percentage points). The top row presents the 95% confidence sets, and the bottom row presents the 90% confidence sets. All confidence sets are computed as weighted averages of the court-specific confidence sets, with weights proportional to court size.

Figure 4 displays the average confidence sets for the non-normalized policy effects $\alpha(\epsilon)$

of implementing a lower bound on nonprosecution rates. Under linear extrapolation, the results suggest that imposing a lower quota on nonprosecution is beneficial. However, this evidence becomes less conclusive with quadratic and cubic specifications as the lower bound ϵ increases, reflecting limitations of available exogenous variation at higher extrapolation levels. The Wald confidence sets appear “spuriously precise” by failing to account for this limited IV variation. Consequently, policy recommendations derived from the robust MLC approach may differ from those based on the classical Wald approach. For instance, based on the 90% confidence sets from the quadratic MTE model, policymakers concerned with robustness to weak IVs may prefer setting the minimum nonprosecution rate below 20%, since the reductions in recidivism become insignificant above this threshold under robust confidence sets. This sign reversal could potentially be driven by the fact that defendants with greater risk of recidivism are more likely to be released as the lower bound of nonprosecution rate increases.

7 Conclusion

In this paper, I propose two robust inference procedures for a class of causal effects identified by MTEs with discrete instruments. For the linear MTE model, I introduce a conditional Wald test, which is simple to implement and asymptotically similar regardless of identification strength. For a broader class of MTE models that nest polynomial specifications, I propose a modified linear combination test that achieves uniform validity against weak identification and has satisfactory power properties under strong identification. Finally, I use the proposed methods to investigate the counterfactual effects of manipulating ADAs’ leniency in the study of misdemeanor prosecution by [Agan et al. \(2023\)](#).

There are several avenues for future research. First, while this paper focuses on using discrete variation from instruments, extending the weak IV analysis to a semiparametric MTE model with continuous propensity scores would be empirically relevant. In addition, the MLC test requires the choice of tuning parameter on the weight of AR statistics, which trades off the power of test under different identification strengths (see more discussion in Online Appendix E). It would be interesting to investigate the optimal choice of this tuning parameter following the regret analysis of [Andrews \(2016\)](#). Finally, extending robust inference techniques to accommodate many weak instruments would be valuable, especially considering the typically large number of judges involved in empirical studies ([Jochmans, 2023](#)).

A Proofs of Main Results

This Appendix provides proofs of results in the main text: Proposition 3.1, Theorem 3.1, and Theorem 4.1. These proofs rely on Proposition A.1 and Proposition A.2, which are stated in this Appendix and proved in the Online Appendix D.

Throughout the appendix, we employ the notation in Table 4, which was not necessarily introduced in the main text. In particular, the subscript F emphasizes dependence on the distribution F , and the numerical subscripts (such as d , ℓ , and k) following the subscript F refer to the corresponding index of the parameter vector.

Table 4: Important notation

$q_F(z_\ell)$	$\mathbb{P}_F(Z = z_\ell)$
$p_F(z_\ell)$	$\mathbb{P}_F(D = 1 \mid Z = z_\ell)$
p_F	$(p_F(z_0), p_F(z_1), \dots, p_F(z_K))'$
A_F	The matrix of propensity scores defined in equations (5) and (11)
$\beta_{F,d\ell}$	$\mathbb{E}_F[Y \mid D = d, Z = z_\ell]$
$\beta_{F,d}$	$(\beta_{F,d0}, \dots, \beta_{F,dK})'$
β_F	$(\beta'_{F,1}, \beta'_{F,0})'$
$\sigma_{F,d\ell}^2$	$\text{var}_F(Y \mid D = d, Z = z_\ell)$
B_F, C_F	The SVD of matrix A_F defined in (27)
$\{\tau_{j,F}\}_{j=1}^{2(M+1)}$	The singular values of A_F (in a descending order)

A.1 Proof of Proposition 3.1 and Theorem 3.1

The proof of asymptotic similarity of the AR and conditional tests in Proposition 3.1 and Theorem 3.1 uses the sub-sequencing techniques from Andrews et al. (2020). Specifically, I verify their Assumption B* in the following Proposition A.1.

Proposition A.1. For any subsequence $\{p_n\}$ of $\{n\}$ and any sequence $\{(\lambda_{p_n}, F_{p_n}) \in \mathcal{P}_0\}$ for which

1. $\theta_{F_{p_n}} \rightarrow \theta_\infty \in \Theta$, where $\theta_{F_{p_n}}$ is a sequence of parameters such that $(\theta_{F_{p_n}}, F_{p_n}) \in \mathcal{P}$ and $\lambda_{p_n} = c'\theta_{F_{p_n}}$;
2. $\beta_{F_{p_n},d\ell} \rightarrow \beta_{\infty,d\ell}$ for $d = 0, 1$ and $\ell = 0, 1, \dots, K$;
3. $\sigma_{F_{p_n},d\ell}^2 \rightarrow \sigma_{\infty,d\ell}^2$ for $d = 0, 1$ and $\ell = 0, 1, \dots, K$;

4. $p_{F_{p_n}}(z_\ell) \rightarrow p_\infty(z_\ell)$ for all $\ell = 0, 1, \dots, K$;

5. $q_{F_{p_n}}(z_\ell) \rightarrow q_\infty(z_\ell)$ for all $\ell = 0, 1, \dots, K$;

The following result holds:

(a) the convergence of the type-I error rate of AR test:

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} \left(\text{AR}_{p_n, k}(\lambda_{p_n}) > q_{\chi_1^2}(1 - \alpha) \right) = \alpha, \text{ for } k = 1, \dots, K,$$

(b) Suppose further that

$$\sqrt{p_n} \|\pi_{F_{p_n}}\| \rightarrow \psi_\infty \in [0, \infty], \text{ and } \pi_{F_{p_n}} / \|\pi_{F_{p_n}}\| \rightarrow \iota_\infty,$$

where ι_∞ is a unit vector⁵, then we have the convergence of the type-I error rate of the conditional test

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} (W_{p_n}(\lambda_{p_n}) > q_{W^*}(1 - \alpha)) = \alpha.$$

Proof of Proposition 3.1. This proposition is a consequence from Corollary 2.1(c) of [Andrews et al. \(2020\)](#) after replacing “CP (Coverage Probability)” with “RP (Null Rejection Probability)” as noted in Comment 4 below their Theorem 2.1. Assumption B* in their paper is verified by Proposition A.1(a) by taking their parameter λ to be (λ, F) in our notation, and by letting their function $h(\lambda)$ correspond to $(\theta, \beta, \sigma_{d\ell, \ell}^2, p, q)$, where θ is an element of Θ satisfying $c'\theta = \lambda$ and $(\theta, F) \in \mathcal{P}$. (By the definition of $\mathcal{P}_0$, such a θ exists.) $\square$

Proof of Theorem 3.1. This theorem is a consequence from Corollary 2.1(c) of [Andrews et al. \(2020\)](#) after replacing “CP (Coverage Probability)” with “RP (Null Rejection Probability)” as noted in Comment 4 below their Theorem 2.1. Assumption B* in their paper is verified by Proposition A.1(b) by taking their parameter λ as (λ, F) in our notation, and their function $h_n(\lambda)$ to correspond to the vector

$$(\theta, \beta, \{\sigma_{d\ell}^2\}_{d, \ell}, p, q, \sqrt{n}\|\pi\|, \pi/\|\pi\|),$$

where θ is an element of Θ satisfying $c'\theta = \lambda$ and $(\theta, F) \in \mathcal{P}$. $\square$

⁵Define $\pi/\|\pi\| = 1_{K \times 1}/\sqrt{K}$ if $\|\pi\| = 0$.

A.2 Proof of Theorem 4.1

The proof of uniform size control in Theorem 4.1 uses the subsequencing technique of Andrews et al. (2020). I verify part of their Assumption B in Proposition A.2 below. The proposition is stated along drifting sequences of distributions, and it characterizes the strength of identification through the SVD of A_F .

For each distribution F , let

$$A_F = C_F \Pi_F B_F' \quad (27)$$

be a SVD of $A_F \in \mathbb{R}^{2(K+1) \times 2(M+1)}$. Here $C_F \in \mathbb{R}^{2(K+1) \times 2(K+1)}$ and $B_F \in \mathbb{R}^{2(M+1) \times 2(M+1)}$ are orthogonal matrices whose columns are the left and right singular vectors of A_F , respectively, and

$$\Pi_F = \begin{bmatrix} \text{diag}\{\tau_{1,F}, \dots, \tau_{2(M+1),F}\} \\ 0_{2(K-M) \times 2(M+1)} \end{bmatrix}, \quad \tau_{1,F} \geq \tau_{2,F} \geq \dots \geq \tau_{2(M+1),F} \geq 0.$$

The matrix Π_F has this form because $K \geq M$ in Definition 2.1. The singular values $\tau_{j,F}$ are unique, but B_F and C_F are not. For each F , fix one such pair (B_F, C_F) , and the results below hold for every choice. The test statistics do not depend on B_F or C_F , which enter only the proofs.

Proposition A.2. Let $\{p_n\}$ be any subsequence of $\{n\}$, and let $\{(\lambda_{p_n}, F_{p_n})\}$ be any sequence in $\mathcal{P}_0$ that satisfies the following conditions:

1. $\theta_{F_{p_n}} \rightarrow \theta_\infty \in \Theta$, where $(\theta_{F_{p_n}}, F_{p_n}) \in \mathcal{P}$ and $\lambda_{p_n} = c' \theta_{F_{p_n}}$;
2. $\beta_{F_{p_n}, d\ell} \rightarrow \beta_{\infty, d\ell}$ for $d = 0, 1$ and $\ell = 0, 1, \dots, K$;
3. $\sigma_{F_{p_n}, d\ell}^2 \rightarrow \sigma_{\infty, d\ell}^2$ for $d = 0, 1$ and $\ell = 0, 1, \dots, K$;
4. $p_{F_{p_n}}(z_\ell) \rightarrow p_\infty(z_\ell)$ for $\ell = 0, 1, \dots, K$;
5. $q_{F_{p_n}}(z_\ell) \rightarrow q_\infty(z_\ell)$ for $\ell = 0, 1, \dots, K$;
6. $B_{F_{p_n}} \rightarrow B_\infty$ and $C_{F_{p_n}} \rightarrow C_\infty$, where B_F and C_F are defined in (27);
7. $\sqrt{p_n} \tau_{j, F_{p_n}} \rightarrow t_j \in [0, \infty]$ for $j = 1, \dots, 2(M+1)$.

Then

$$\limsup_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} \left(\inf_{c' \theta = \lambda_{p_n}} \text{MLC}_{p_n}(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) \leq \alpha.$$

Proof of Theorem 4.1. This theorem is a consequence from Corollary 2.1(a) of [Andrews et al. \(2020\)](#) after replacing “CP (Coverage Probability)” with “RP (Null Rejection Probability)” as noted in Comment 4 below their Theorem 2.1. Assumption B in their paper is verified by Proposition A.2 by taking their parameter λ as (λ, F) in our notation, and their function $h_n(\lambda)$ to correspond to the vector of

$$(\theta, \beta, \{\sigma_{d\ell}^2\}_{d,\ell}, p, q, B, C, \{\sqrt{n}\tau_j\}_{j=1}^{2(M+1)}),$$

where θ is an element of Θ satisfying $c'\theta = \lambda$ and $(\theta, F) \in \mathcal{P}$. □

B Inference under Additive Separability

Empirical studies of MTE models usually include covariates $W \in \mathbb{R}^L$ and assume that the MTR functions are additively separable in the covariates and the selection unobservable ([Brinch et al., 2017](#), Assumption 2). In the empirical application of Section 6, I do not impose this assumption. Instead, I condition on covariates and apply the robust procedures within each covariate value. In this appendix, I first explain why additive separability is a strong assumption. I then show how to implement the MLC test of Section 4 when researchers still want to maintain it.

B.1 Two concerns with additive separability

For a coefficient vector $\tau_d \in \mathbb{R}^L$, let $V_d \equiv Y_d - \mu_d - W'\tau_d$ for $d = 0, 1$.

Assumption 4 (Additive separability). For $d = 0, 1$, $\mathbb{E}[Y_d | W, U] = \mu_d + W'\tau_d + \mathbb{E}[V_d | U]$.

Assumption 4 rules out any interaction between W and U in the MTR functions. It allows the MTE to be estimated on the unconditional support of the propensity score, because variation in W is pooled with variation in the instrument ([Carneiro et al., 2011](#), Section I.B). There are two concerns with this assumption.

The first concern is misspecification. In an earlier version of this paper ([Ren, 2025](#), Section 5), I generalize the specification of [Kline and Walters \(2019, equation \(12\)\)](#) by letting the coefficient on the selection unobservable vary linearly with W , which violates Assumption 4. Theorem 5.1 of that version shows that imposing Assumption 4 on this DGP can significantly bias the ATE, the conditional ATE, and the slope of the MTE. The bias is driven by the

heterogeneous effects of W that vary with U and by the imbalance of W between the treated and untreated groups. It persists even when W is uncorrelated with the control functions of the propensity score up to second moments within each treatment group.

The second concern is more fundamental. Assumption 4 can point identify the model without any excluded instrument when the outcome equations are linear in W and the propensity score is nonlinear in W (Pan, Wang, Zhang, and Zhou, 2024). In that case, the identifying power comes from the functional form and from covariates that enter both the selection equation and the outcome equations, not from exogenous variation in the instrument. As a consequence, the instrument may be weak while the model appears strongly identified, and the source of that strength is mainly driven by functional form and endogenous covariates. For these two reasons, I do not rely on Assumption 4 in the empirical application.

B.2 A model with covariates

Consider the model

$$Y_d = \mu_d + W'\tau_d + V_d, \quad D = \mathbb{1}[U \leq W'\pi + \pi_Z(Z)], \quad U \perp\!\!\!\perp (W, Z), \quad U \sim F_U, \quad (28)$$

where Z is a discrete instrument, $\pi_Z: \text{supp}(Z) \rightarrow \mathbb{R}$, and F_U is a known CDF, such as the logistic or the standard normal CDF. The propensity score is $p(w, z) = F_U(w'\pi + \pi_Z(z))$. Under Assumption 4, I maintain $\mathbb{E}[V_d \mid W, Z, F_U(U) = u] = \sum_{m=1}^M \rho_{dm} h_m(u)$ for $d = 0, 1$. The MTE is then

$$\text{MTE}(w, u) = \mu_1 - \mu_0 + w'(\tau_1 - \tau_0) + \sum_{m=1}^M (\rho_{1m} - \rho_{0m}) h_m(u).$$

For the PRTE in (24) with a counterfactual propensity score $p^\epsilon(w, z)$ as in Section 6,

$$\bar{\alpha}(\epsilon) = \mu_1 - \mu_0 + \bar{w}'_\epsilon(\tau_1 - \tau_0) + \sum_{m=1}^M c_{1,m}(\rho_{1m} - \rho_{0m}), \quad \bar{w}_\epsilon = \frac{\mathbb{E}[(p^\epsilon(W, Z) - p(W, Z))W]}{\mathbb{E}[p^\epsilon(W, Z) - p(W, Z)]}, \quad (29)$$

where $c_{1,m}$ is the PRTE weight in Table 1 with $p(Z)$ replaced by $p(W, Z)$.

B.3 Implementation for the PRTE

Weak identification affects only the coefficients $\rho = (\rho'_1, \rho'_0)'$ on the control functions. For a fixed ρ , the coefficients (μ_d, τ_d) can be recovered from a linear regression on $(1, W')$ and

are strongly identified when $\tau_{\min}(\text{var}(W \mid D = d))$ is bounded away from zero. I therefore concentrate out (μ_d, τ_d) and apply the MLC test to moment functions on ρ .

Algorithm B.1 (MLC inference on the PRTE under additive separability).

1. In the first stage equation, estimate (π, π_Z) by maximum likelihood with known F_U , and compute $\hat{p}_i = F_U(W_i' \hat{\pi} + \hat{\pi}_Z(Z_i))$ and $\hat{p}_i^\epsilon$.
2. For $d = 0, 1$, let $\tilde{Y}_{d,i}$ and $\tilde{\lambda}_{d,i}$ be the residuals of Y_i and of the control functions $\lambda_d(\hat{p}_i) = (\lambda_{d1}(\hat{p}_i), \dots, \lambda_{dM}(\hat{p}_i))'$ from their regressions on $(1, W_i')$ in the subsample $\{i : D_i = d\}$. Construct the moments

$$\hat{m}_d(\rho_d) = \frac{1}{n} \sum_{i=1}^n f_d(W_i, Z_i) \mathbb{1}[D_i = d] (\tilde{Y}_{d,i} - \tilde{\lambda}_d(\hat{p}_i)' \rho_d),$$

where $f_d(w, z) \in \mathbb{R}^p$ with $p \geq M$ is chosen so that the asymptotic variance of the moments is nonsingular. The stacked moment $\hat{m}(\rho)$ is affine in ρ , so it has the form $\hat{A}\rho - \hat{\beta}$ of (11) with ρ in place of θ .

3. Replacing (μ_d, τ_d) with their concentrated OLS estimators and $(\bar{w}_\epsilon, c_{1,m})$ with sample analogs in (29) writes the PRTE as $\hat{c}_0 + \hat{c}'_\rho \rho$, with weights estimated as in Online Appendix F.
4. Construct the MLC statistic (16) as in section 4.3, with variance estimators that account for the first stage in step 1. Reject $H_0: \bar{\alpha}(\epsilon) = \bar{\alpha}_{\text{null}}$ if the MLC statistic profiled over $\{\rho : \hat{c}_0 + \hat{c}'_\rho \rho = \bar{\alpha}_{\text{null}}\}$ exceeds the $(1 - \alpha)$ -quantile of $(1 + a)\chi_1^2 + a\chi_{2p-1}^2$, and invert the test over λ .

I do not establish the uniform validity of Algorithm B.1 in this appendix. Its validity also relies on Assumption 4 and on the parametric first stage, which are subject to the concerns in section B.1.

C Technical Details on Comparison with Andrews (2018)

Section 4.2 briefly explains why the high-level conditions of I. Andrews (2018) need not hold along local-to-singularity sequences in MTE models. This appendix provides more details through two examples. The linear MTE model motivates the SVD normalization, and the just-identified quadratic MTE model shows why normalization alone does not deliver

uniform validity. Along a sequence $(\theta_{F_n}, F_n) \in \mathcal{P}$, I evaluate all quantities at θ_{F_n} and write $\hat{D} \equiv \hat{D}(\theta_{F_n})$.

Why the SVD normalization. Assumption 4.1 of I. Andrews (2018) requires full-rank, nonrandom matrices $\Lambda_{1,n}$ with entries of order $O(\sqrt{n})$ such that $\hat{D}\Lambda_{1,n}$ converges to a Gaussian matrix of full column rank almost surely. In Stock and Wright (2000) and Chaudhuri and Zivot (2011), $\Lambda_{1,n}$ rescales the local-to-zero columns of the Jacobian, which correspond to a known set of weakly identified parameters. MTE models do not have this structure. Let $K = M = 1$ and $h_1(u) = u - \frac{1}{2}$, so that $\lambda_{11}(x) = \frac{x-1}{2}$ and $\lambda_{01}(x) = \frac{x}{2}$, and set $p_{F_n}(z_0) = p \in (0, 1)$ and $p_{F_n}(z_1) = p + v_n$ with $v_n \rightarrow 0$. Each block A_{d,F_n} has determinant $v_n/2$, so $A_{F_n} = \text{diag}\{A_{1,F_n}, A_{0,F_n}\}$ converges to a singular but nonzero matrix. The data strongly identify the conditional mean of the outcome, $\mu_d + \rho_{d1}\lambda_{d1}(p)$, but neither μ_d nor ρ_{d1} separately. The strong and weak directions, which are the right singular vectors of A_{F_n} with nonvanishing and vanishing singular values, lead to strongly identified and weakly identified *linear combinations* of the parameters. The choice $\Lambda_{1,n} = \sqrt{n}I$ of Kleibergen (2005) fails too because $\sqrt{n}A_{F_n}$ diverges. The SVD normalization $\sqrt{n}\hat{D}B_{F_n}\Psi_{F_n}$ in (27) and (43) handles this by rotating $\hat{D}$ onto the singular directions of A_{F_n} and rescaling each one. It enters only the proof, since the statistic in (14) is invariant to it. The remaining question is whether the normalized limit $\mathcal{D}$ of Lemma D.4 has full column rank. The rest of this appendix shows that it does in the linear model but can fail in the quadratic model, which is what the second modification, the random perturbation, addresses.

The limit of the weak columns. Consider sequences along which all propensity scores collapse to a common p faster than $n^{-1/2}$:

$$\max_{0 \leq \ell \leq K} |p_{F_n}(z_\ell) - p| = o(n^{-1/2}). \quad (30)$$

For $h_m(u) = u^m - \frac{1}{m+1}$, each block A_{d,F_n} then converges to a rank-one matrix whose rows all equal $(1, \lambda_{d1}(p), \dots, \lambda_{dM}(p))$. So each block has one strong direction and M weak directions, which span the null space of this row. Let $\bar{B}_d = (\bar{b}_{d,1}, \dots, \bar{b}_{d,M})$ collect them. Along the weak directions, $\sqrt{n}\tau_{j,F_n} \rightarrow 0$, so the weak columns of $\mathcal{D}$ have no deterministic part and consist only of the sampling noise $\text{diag}\{\mathcal{Z}_{D,1}\bar{B}_1, \mathcal{Z}_{D,0}\bar{B}_0\}$, where $\mathcal{Z}_{D,d}$ is the limit of $\sqrt{n}(\hat{D}_d - A_{d,F_n})$. Hence $\mathcal{D}$ has full column rank only if this noise does.

To compute it, let $s_d(x; b) \equiv \sum_{m=0}^M b_m \lambda_{dm}(x)$ for $b \in \mathbb{R}^{M+1}$, the change in the conditional mean when θ_d moves along b . The delta method gives $\sqrt{n}(\hat{A}_d - A_{d,F_n})b \xrightarrow{d} s'_d(p; b)\mathcal{Z}_p$, where $\mathcal{Z}_p$ is the limit of $\sqrt{n}(\hat{p} - p_{F_n})$. The orthogonalization in (12) replaces $\mathcal{Z}_p$ by its residual $\mathcal{R} \equiv \mathcal{Z}_p - \text{cov}(\mathcal{Z}_p, \mathcal{Z}_m) \text{var}(\mathcal{Z}_m)^{-1} \mathcal{Z}_m$, whose variance is positive definite under Definition 2.1.

Therefore

$$\mathcal{Z}_{D,d}\bar{B}_d = \mathcal{R}(s'_d(p; \bar{b}_{d,1}), \dots, s'_d(p; \bar{b}_{d,M})). \quad (31)$$

This outer product has rank at most one. Because every slope is evaluated at the common limit p , all columns of (31) are multiples of the same vector $\mathcal{R}$. Therefore, whether $\mathcal{D}$ has full rank depends on the number of weak directions per block, M .

Linear MTE model. For $K = M = 1$, each block has a single weak direction,

$$\bar{B}_1 = \frac{(1-p, 2)'}{\sqrt{(1-p)^2 + 4}}, \quad \bar{B}_0 = \frac{(-p, 2)'}{\sqrt{p^2 + 4}}.$$

Under (30), Lemma D.4 and (31) give, up to column order and signs,

$$\mathcal{D} = \text{diag} \left\{ \left(\frac{1_{2 \times 1}}{\sqrt{2}}, \frac{\mathcal{R}}{\sqrt{(1-p)^2 + 4}} \right), \left(\frac{1_{2 \times 1}}{\sqrt{2}}, \frac{\mathcal{R}}{\sqrt{p^2 + 4}} \right) \right\},$$

where $1_{2 \times 1}$ is the two-vector of ones. Since $\mathcal{R}$ is nondegenerate Gaussian, it lies outside the span of $1_{2 \times 1}$ almost surely. So $\mathcal{D}$ has full column rank, and Assumption 4 holds along this sequence.

Just-identified quadratic MTE model. For $K = M = 2$, each block has two weak directions, but by (31) their 3×2 noise matrix $\mathcal{Z}_{D,d}\bar{B}_d$ has rank at most one. Hence $\text{rank}(\mathcal{D}) \leq 2 + 2 = 4 < 6$ almost surely. The same argument gives $\text{rank}(\mathcal{D}) \leq 4 < 2(M+1)$ for any polynomial MTE model with $M \geq 2$ along sequences satisfying (30), so (30) with $M \geq 2$ is sufficient for a rank-deficient normalized limit.

Consequences. The rank deficiency is not an artifact of the SVD normalization. In the quadratic example, no other normalization $\Lambda_{1,n}$ allowed by I. Andrews (2018) restores full column rank. Any such convergent normalization has, along a subsequence, a limit $\mathcal{D}N$ for a nonrandom N , which remains singular.⁶ So Assumption 4 fails along a sequence in the parameter space. Without the perturbation, the normalized limit of $\hat{D}'\hat{\Omega}^{-1}\hat{D}$ is singular, so the full-rank conditioning argument behind the χ_1^2 limit is unavailable, even though the finite-sample inverse may exist. The perturbation $\kappa n^{-1/2}\xi$ in (13) adds $\kappa\xi B_{2(M+1)-q,\infty}$ to the weak columns, where $B_{2(M+1)-q,\infty}$ collects the limiting weak right singular vectors of A_{F_n} and q counts the directions with $\sqrt{n}\tau_{j,F_n} \rightarrow \infty$. This independent Gaussian term makes the

⁶Write $\hat{D}\Lambda_{1,n} = (\sqrt{n}\hat{D}B_{F_n}\Psi_{F_n})N_n$, where $N_n = n^{-1/2}\Psi_{F_n}^{-1}B_{F_n}'\Lambda_{1,n}$. A row j of N_n with $\sqrt{n}\tau_{j,F_n}$ bounded is bounded because $\Lambda_{1,n} = O(\sqrt{n})$. A row j with $\sqrt{n}\tau_{j,F_n} \rightarrow \infty$ equals $c'_{j,F_n}A_{F_n}\Lambda_{1,n}$, where c_{j,F_n} is a left singular vector, and is bounded because $\hat{D}\Lambda_{1,n}$ is tight and $\sqrt{n}(\hat{D} - A_{F_n}) = O_p(1)$. Hence N_n is bounded and has a convergent subsequence. Denote the limit as N .

normalized limit full rank almost surely and independent of the moment limit along every subsequence in Lemma D.4, as Lemma D.5 establishes. The linear example does not need it, but it is necessary for uniform argument over the general parameter space.

References

- AGAN, A., J. L. DOLEAC, AND A. HARVEY (2023): “Misdemeanor prosecution,” *The Quarterly Journal of Economics*, 138, 1453–1505.
- ANDRESEN, M. E. (2018): “Exploring marginal treatment effects: Flexible estimation using Stata,” *The Stata Journal*, 18, 118–158.
- ANDREWS, D. W. (2017): “Identification-robust subvector inference,” *Available at SSRN 3032675*.
- ANDREWS, D. W. AND X. CHENG (2012): “Estimation and inference with weak, semi-strong, and strong identification,” *Econometrica*, 80, 2153–2211.
- ANDREWS, D. W., X. CHENG, AND P. GUGGENBERGER (2020): “Generic results for establishing the asymptotic size of confidence sets and tests,” *Journal of Econometrics*, 218, 496–531.
- ANDREWS, D. W. AND P. GUGGENBERGER (2017): “Asymptotic size of Kleibergen’s LM and conditional LR tests for moment condition models,” *Econometric Theory*, 33, 1046–1080.
- (2019): “Identification-and singularity-robust inference for moment condition models,” *Quantitative Economics*, 10, 1703–1746.
- ANDREWS, I. (2016): “Conditional linear combination tests for weakly identified models,” *Econometrica*, 84, 2155–2182.
- (2018): “Valid two-step identification-robust confidence sets for GMM,” *Review of Economics and Statistics*, 100, 337–348.
- ANDREWS, I. AND A. MIKUSHEVA (2016b): “Conditional inference with a functional nuisance parameter,” *Econometrica*, 84, 1571–1612.
- ANDREWS, I., J. H. STOCK, AND L. SUN (2019): “Weak instruments in instrumental variables regression: Theory and practice,” *Annual Review of Economics*, 11, 727–753.
- BJÖRKLUND, A. AND R. MOFFITT (1987): “The estimation of wage gains and welfare gains in self-selection models,” *The Review of Economics and Statistics*, 42–49.
- BRAVE, S. AND T. WALSTRUM (2014): “Estimating marginal treatment effects using parametric and semiparametric methods,” *The Stata Journal*, 14, 191–217.

- BRINCH, C. N., M. MOGSTAD, AND M. WISWALL (2017): “Beyond LATE with a discrete instrument,” *Journal of Political Economy*, 125, 985–1039.
- CARNEIRO, P., J. J. HECKMAN, AND E. J. VYTLACIL (2010): “Evaluating marginal policy changes and the average effect of treatment for individuals at the margin,” *Econometrica*, 78, 377–394.
- (2011): “Estimating marginal returns to education,” *American Economic Review*, 101, 2754–81.
- CARNEIRO, P. AND S. LEE (2009): “Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality,” *Journal of Econometrics*, 149, 191–208.
- CHAUDHURI, S. AND E. ZIVOT (2011): “A new method of projection-based inference in GMM with weakly identified nuisance parameters,” *Journal of Econometrics*, 164, 239–251.
- COX, G. F. (2026): “Weak identification in low-dimensional factor models with one or two factors,” *Review of Economics and Statistics*, 108, 1109–1123.
- HAN, S. AND A. MCCLOSKEY (2019): “Estimation and inference with a (nearly) singular Jacobian,” *Quantitative Economics*, 10, 1019–1068.
- HECKMAN, J., S. URZUA, AND E. VYTLACIL (2006): “Estimation of treatment effects under essential heterogeneity,” *Working Paper*.
- HECKMAN, J. J. AND E. VYTLACIL (2001): “Policy-relevant treatment effects,” *American Economic Review*, 91, 107–111.
- (2005): “Structural equations, treatment effects, and econometric policy evaluation,” *Econometrica*, 73, 669–738.
- HECKMAN, J. J. AND E. J. VYTLACIL (1999): “Local instrumental variables and latent variable models for identifying and bounding treatment effects,” *Proceedings of the national Academy of Sciences*, 96, 4730–4734.
- IMBENS, G. W. AND J. D. ANGRIST (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475.
- JOCHMANS, K. (2023): “Many (Weak) Judges in Judge-Leniency Designs,” *TSE Working Paper*.
- KLEIBERGEN, F. (2005): “Testing parameters in GMM without assuming that they are identified,” *Econometrica*, 73, 1103–1123.
- KLINE, P. AND C. R. WALTERS (2019): “On Heckits, LATE, and numerical equivalence,” *Econometrica*, 87, 677–696.

- KOWALSKI, A. E. (2023): “Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform,” *Review of Economics and Statistics*, 105, 646–664.
- MAESTAS, N., K. J. MULLEN, AND A. STRAND (2013): “Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt,” *American Economic Review*, 103, 1797–1829.
- MIKUSHEVA, A. AND L. SUN (2022): “Inference with Many Weak Instruments,” *Review of Economic Studies*, 89, 2663–2686.
- MOGSTAD, M., A. SANTOS, AND A. TORGOVITSKY (2017): “Using instrumental variables for inference about policy relevant treatment parameters,” Tech. rep., National Bureau of Economic Research.
- (2018): “Using instrumental variables for inference about policy relevant treatment parameters,” *Econometrica*, 86, 1589–1619.
- MOGSTAD, M. AND A. TORGOVITSKY (2018): “Identification and extrapolation of causal effects with instrumental variables,” *Annual Review of Economics*, 10, 577–613.
- (2024): “Instrumental Variables with Unobserved Heterogeneity in Treatment Effects,” Tech. rep., NBER.
- MOREIRA, M. J. (2003): “A conditional likelihood ratio test for structural models,” *Econometrica*, 71, 1027–1048.
- OLSEN, R. J. (1980): “A least squares correction for selectivity bias,” *Econometrica: Journal of the Econometric Society*, 1815–1820.
- PAN, Z., Z. WANG, J. ZHANG, AND Y. ZHOU (2024): “Marginal treatment effects in the absence of instrumental variables,” *arXiv preprint arXiv:2401.17595*.
- REN, M. (2025): “Extrapolating LATE with Weak IVs,” *arXiv preprint arXiv:2512.23854*.
- SASAKI, Y. AND T. URA (2023): “Estimation and inference for policy relevant treatment effects,” *Journal of Econometrics*, 234, 394–450.
- STOCK, J. H. AND J. H. WRIGHT (2000): “GMM with weak identification,” *Econometrica*, 68, 1055–1096.

Online appendices for “Extrapolating LATE with Weak IVs”

Muyang Ren[†]

October 12, 2026

Abstract

This document contains a set of results that accompany the main text. Appendix [D](#) provides proofs of the omitted results stated in Appendix [A](#). Appendix [E](#) establishes consistency and conducts local power analysis for the MLC test introduced in section [4](#). Appendix [F](#) shows how to extend the proposed methods to settings with estimated weights. Appendix [G](#) provides simulation evidence on the performance of the conditional Wald test introduced in section [3](#). Appendix [H](#) discusses issues related to testing IV strength in the empirical application.

D Proof of Omitted Results in Appendix [A](#)

This appendix provides the proofs of Propositions [A.1](#) and [A.2](#), which are omitted from Appendix [A](#). I use the following notation and abbreviations throughout. Let $\mathbb{S}_{++}^k$ denote the set of $k \times k$ positive definite matrices. I abbreviate “positive definite” and “positive semi-definite” as PD and PSD. For triangular arrays, I abbreviate the weak law of large numbers as WLLN and the central limit theorem as CLT; Lemmas [D.6](#) and [D.7](#) state the versions used here. For scalars $a_0, \dots, a_K$, $\text{diag}\{a_k\}_{k=0}^K$ denotes the $(K+1) \times (K+1)$ diagonal matrix with diagonal elements $a_0, \dots, a_K$. For matrices A and B , $\text{diag}(A, B)$ denotes the block diagonal matrix with A in the upper block and B in the lower block. For a vector, $\|\cdot\|$ denotes the Euclidean norm; for a matrix, it denotes the spectral norm, i.e., the largest singular value. Finally, $\tau_{\min}(\cdot)$ and $\tau_{\max}(\cdot)$ denote the smallest and largest singular values of a matrix.

D.1 Proof of Proposition [A.1](#)

Proof of Proposition [A.1](#). For simplicity of notations, the proof is shown for the full sequence $\{n\}$. Then, we note that the same proof goes through with p_n in place of n .

Part (a): Note that $g(\lambda)$ is a continuously differentiable function of λ , p , β_1 , and β_0 . Applying uniform delta method and Lemma [D.1\(a\)](#) to the sample moment $\hat{g}(\lambda)$ yields

$$\begin{aligned}\sqrt{n}\hat{g}(\lambda_n) &= [\partial_p g(\lambda_n)]\hat{Z}_p + [\partial_{\beta_1} g(\lambda_n)]\hat{Z}_{\beta_1} + [\partial_{\beta_0} g(\lambda_n)]\hat{Z}_{\beta_0} + o_p(1) \\ &\xrightarrow{d} \partial_p g(\lambda_\infty)\mathcal{Z}_p + \partial_{\beta_1} g(\lambda_\infty)\mathcal{Z}_{\beta_1} + \partial_{\beta_0} g(\lambda_\infty)\mathcal{Z}_{\beta_0} \equiv \mathcal{Z}_g\end{aligned}$$

[†]Department of Economics, University of Tennessee, mren7@utk.edu

Then it follows that $\mathcal{Z}_{\mathbf{g}} \sim \mathcal{N}(0_{K \times 1}, S_{\infty})$ where

$$S_{\infty} \equiv \partial_p g(\lambda_{\infty}) \Sigma_{p,\infty} \partial_{p'} g(\lambda_{\infty}) + \partial_{\beta_1} g(\lambda_{\infty}) \Sigma_{\beta_1,\infty} \partial_{\beta_1'} g(\lambda_{\infty}) + \partial_{\beta_0} g(\lambda_{\infty}) \Sigma_{\beta_0,\infty} \partial_{\beta_0'} g(\lambda_{\infty}).$$

Here λ_{∞} is the limit of $\lambda_n = c' \theta_n$ as n goes to infinity, and $g(\lambda_{\infty})$ equals $g(\lambda)$ by sending λ , p , β_1 , and β_0 to their limiting values.

Again by Lemma D.1(b), $\Sigma_{p,\infty}$, $\Sigma_{\beta_1,\infty}$, and $\Sigma_{\beta_0,\infty}$ can be consistently estimated by their sample analogs $\hat{\Sigma}_p$, $\hat{\Sigma}_{\beta_1}$, and $\hat{\Sigma}_{\beta_0}$ along the drifting sequence, respectively. The continuous mapping theorem then implies that $\hat{S}(\lambda_n) \xrightarrow{p} S_{\infty}$, where $\hat{S}(\lambda)$ is the asymptotic variance estimator defined in (10). Note that Lemma D.2 implies S_{∞} is PD. By Slutsky's theorem, we have

$$\sqrt{n} \hat{S}(\lambda_n)^{-1/2} \hat{g}(\lambda_n) \xrightarrow{d} \mathcal{N}(0, I_{K \times K}).$$

In particular, for k -th element of $\hat{g}(\lambda_n)$, applying the same arguments yields $\sqrt{n} \hat{s}_k^{-1} \hat{g}_k(\lambda_n) \xrightarrow{d} N(0, 1)$, where $\hat{s}_k^2$ is the (k, k) -th element of $\hat{S}$. Since $\text{AR}_{n,k}(\lambda_n)$ is the square of this component, we have $\text{AR}_{n,k}(\lambda_n) \xrightarrow{d} \chi_1^2$, as desired.

Part (b): We divide the discussion into two cases.

Case 1: $\psi_{\infty} < \infty$.

In this case, we have

$$\begin{aligned} \sqrt{n} \hat{\pi} &= \sqrt{n} \pi_{F_n} + [\partial_p \pi] \hat{Z}_p \\ &\xrightarrow{d} \psi_{\infty} \iota_{\infty} + [\partial_p \pi] \mathcal{Z}_p \equiv \mathcal{Z}_{\pi} \end{aligned}$$

and

$$\begin{aligned} \hat{h}(\lambda_n) &= \sqrt{n} \hat{\pi} - [\partial_p \pi] \hat{\Sigma}_p [\partial_{p'} \hat{g}(\lambda_n)] \hat{S}(\lambda_n)^{-1} \sqrt{n} \hat{g}(\lambda_n) \\ &\xrightarrow{d} \psi_{\infty} \iota_{\infty} + [\partial_p \pi] \mathcal{Z}_p - [\partial_p \pi] \Sigma_{p,\infty} [\partial_{p'} g(\lambda_{\infty})] S_{\infty}^{-1} \mathcal{Z}_{\mathbf{g}} \equiv \mathcal{Z}_h \end{aligned} \tag{32}$$

From the independence $(\mathcal{Z}_{\beta_1}, \mathcal{Z}_{\beta_0}) \perp \mathcal{Z}_p$, it follows that $\mathcal{Z}_h \perp \mathcal{Z}_{\mathbf{g}}$ since they are jointly normal and uncorrelated.

By continuous mapping theorem, we obtain the asymptotic distribution of the Wald statistic under the null:

$$\begin{aligned} W_n(\lambda_n) &= \frac{n \hat{g}(\lambda_n)' \hat{S}(\lambda_n)^{-1} \hat{\pi} \hat{\pi}' \hat{S}(\lambda_n)^{-1} \hat{g}(\lambda_n)}{\hat{\pi}' \hat{S}(\lambda_n)^{-1} \hat{\pi}} \\ &\xrightarrow{d} \frac{\mathcal{Z}_{\mathbf{g}}' S_{\infty}^{-1} \mathcal{Z}_{\pi} \mathcal{Z}_{\pi}' S_{\infty}^{-1} \mathcal{Z}_{\mathbf{g}}}{\mathcal{Z}_{\pi}' S_{\infty}^{-1} \mathcal{Z}_{\pi}}. \end{aligned}$$

On the other hand, we note that the simulated statistic $W_n^*(\lambda_n)$ can be written as a function of standard normal draw η^* and a conditioning statistic

$$\hat{\Upsilon} \equiv (\hat{h}(\lambda_n), \hat{S}(\lambda_n), \hat{\Sigma}_p, \partial_p \hat{g}(\lambda_n)). \tag{33}$$

Specifically, we write

$$W_n^*(\lambda_n) = W(\eta^*, \hat{\Upsilon}) \equiv \frac{(\eta^*)' \hat{S}(\lambda_n)^{-1/2} \pi_s(\eta^*, \hat{\Upsilon}) \pi_s(\eta^*, \hat{\Upsilon})' \hat{S}(\lambda_n)^{-1/2} \eta^*}{\pi_s(\eta^*, \hat{\Upsilon})' \hat{S}(\lambda_n)^{-1} \pi_s(\eta^*, \hat{\Upsilon})} \quad (34)$$

where

$$\pi_s(\eta^*, \hat{\Upsilon}) \equiv \hat{h}(\lambda_n) + [\partial_p \pi] \hat{\Sigma}_p [\partial_{p'} \hat{g}(\lambda_n)] \hat{S}(\lambda_n)^{-1/2} \eta^*.$$

Therefore, $\hat{\Upsilon}$ is the source of sampling uncertainty on the simulated statistic $W_n^*(\lambda_n)$.

Let $G(\cdot \mid \hat{\Upsilon})$ denote the CDF of the simulated Wald statistic $W_n^*(\lambda_n)$ conditional on data, i.e.,

$$\begin{aligned} G(x \mid \hat{\Upsilon}) &\equiv \mathbb{P}(W_n^*(\lambda_n) \leq x \mid \{Y_i, D_i, Z_i\}_{i=1}^n) \\ &= \mathbb{P}(W(\eta^*, \hat{\Upsilon}) \leq x \mid \hat{\Upsilon}). \end{aligned}$$

Define $q(1 - \alpha, \Upsilon)$ as the $(1 - \alpha)$ -quantile of $G(\cdot \mid \Upsilon)$ for a quartet of elements Υ such as $\hat{\Upsilon}$ in (33). By this definition, the critical value can be expressed as

$$\hat{q}_{W^*}(1 - \alpha) = q(1 - \alpha, \hat{\Upsilon}).$$

Next we show that $q(1 - \alpha, \Upsilon)$ is a continuous function of Υ on the set $\mathcal{U}$, where

$$\begin{aligned} \mathcal{U} = \{ &(u_1, u_2, u_3, u_4) : u_1 + [\partial_p \pi] u_3 u_4' u_2^{-1/2} \eta^* \neq 0_{K \times 1} \text{ a.s.}, \\ &u_1 \in \mathbb{R}^K, u_2 \in \mathbb{S}_{++}^K, u_3 \in \mathbb{S}_{++}^{K+1}, u_4 \in \mathbb{R}^{K \times (K+1)} \}. \end{aligned} \quad (35)$$

Let $\{\Upsilon_n\}$ be a sequence in $\mathcal{U}$ such that $\Upsilon_n \rightarrow \Upsilon \in \mathcal{U}$ as $n \rightarrow \infty$. Given zero probability of discontinuity in the limit by the definition of $\mathcal{U}$, $\{W(\eta^*, \Upsilon_n)\}_{n \geq 1}$ converges almost surely to $W(\eta^*, \Upsilon)$. This implies the convergence in distribution for any continuity point x of $G(\cdot \mid \Upsilon)$:

$$G(x \mid \Upsilon_n) = \mathbb{P}_{\eta^*}(W(\eta^*, \Upsilon_n) \leq x) \rightarrow \mathbb{P}_{\eta^*}(W(\eta^*, \Upsilon) \leq x) = G(x \mid \Upsilon).$$

The distribution function $G(\cdot \mid \Upsilon)$ is increasing at its $(1 - \alpha)$ -quantile $q(1 - \alpha, \Upsilon)$ because the random variable $W(\eta^*, \Upsilon)$ is continuously distributed. By [Andrews and Guggenberger \(2010, Lemma 5\)](#), it follows that $q(1 - \alpha, \Upsilon_n) \rightarrow q(1 - \alpha, \Upsilon)$. This establishes continuity of quantile function on the set $\mathcal{U}$.

The consistency of $(\hat{S}(\lambda_n), \hat{\Sigma}_p, \partial_p \hat{g}(\lambda_n))$ established in Part(a) and (32) give

$$\hat{\Upsilon} \xrightarrow{d} \Upsilon_\infty \equiv (\mathcal{Z}_h, S_\infty, \Sigma_{p,\infty}, \partial_p g(\lambda_\infty)),$$

and by [Lemma D.3](#), we have $\mathbb{P}(\Upsilon_\infty \in \mathcal{U}) = 1$. From the continuity of $q(1 - \alpha, \Upsilon)$, it follows by continuous mapping theorem that

$$W_n(\lambda_n) - q(1 - \alpha, \hat{\Upsilon}) \xrightarrow{d} \frac{\mathcal{Z}'_g S_\infty^{-1} \mathcal{Z}_\pi \mathcal{Z}'_\pi S_\infty^{-1} \mathcal{Z}_g}{\mathcal{Z}'_\pi S_\infty^{-1} \mathcal{Z}_\pi} - q(1 - \alpha, \Upsilon_\infty),$$

which implies

$$\mathbb{P}_{F_n} \left(W_n(\lambda_n) > q(1 - \alpha, \hat{\Upsilon}) \right) \rightarrow \mathbb{P} \left(\frac{\mathcal{Z}'_{\mathbf{g}} S_{\infty}^{-1} \mathcal{Z}_{\pi} \mathcal{Z}'_{\pi} S_{\infty}^{-1} \mathcal{Z}_{\mathbf{g}}}{\mathcal{Z}'_{\pi} S_{\infty}^{-1} \mathcal{Z}_{\pi}} > q(1 - \alpha, \Upsilon_{\infty}) \right).$$

Next, we show that the limit probability on the right-hand side equals α . We first examine the conditional probability as follows.

$$\begin{aligned} & \mathbb{P} \left(\frac{\mathcal{Z}'_{\mathbf{g}} S_{\infty}^{-1} \mathcal{Z}_{\pi} \mathcal{Z}'_{\pi} S_{\infty}^{-1} \mathcal{Z}_{\mathbf{g}}}{\mathcal{Z}'_{\pi} S_{\infty}^{-1} \mathcal{Z}_{\pi}} > q(1 - \alpha, \Upsilon_{\infty}) \mid \Upsilon_{\infty} \right) \\ &= \mathbb{P} \left(\frac{(\eta^*)' S_{\infty}^{-1/2} \pi_s(\eta^*, \Upsilon_{\infty}) \pi_s(\eta^*, \Upsilon_{\infty})' S_{\infty}^{-1/2} \eta^*}{\pi_s(\eta^*, \Upsilon_{\infty})' S_{\infty}^{-1} \pi_s(\eta^*, \Upsilon_{\infty})} > q(1 - \alpha, \Upsilon_{\infty}) \mid \Upsilon_{\infty} \right) \\ &= \mathbb{P}(W(\eta^*, \Upsilon_{\infty}) > q(1 - \alpha, \Upsilon_{\infty}) \mid \Upsilon_{\infty}) \\ &= \alpha \quad \text{a.s.} \end{aligned}$$

The second line holds by the observation that η^* and $S_{\infty}^{-1/2} \mathcal{Z}_{\mathbf{g}}$ are both independent of Υ_{∞} , and that $\mathcal{Z}_{\pi} = \pi_s(S_{\infty}^{-1/2} \mathcal{Z}_{\mathbf{g}}, \Upsilon_{\infty})$, which has the same distribution as $\pi_s(\eta^*, \Upsilon_{\infty})$. The third line holds by the definition of $W(\eta^*, \Upsilon_{\infty})$. The last line holds by the definition of $q(1 - \alpha, \Upsilon)$, and the fact that $W(\eta^*, \Upsilon)$ is continuously distributed for any $\Upsilon \in \mathcal{U}$. By taking the law of total probability, it follows that the unconditional rejection probability equals α as well. This completes the proof for the case $\psi_{\infty} < \infty$.

Case 2: $\psi_{\infty} = \infty$.

In this case, by the assumption $\sqrt{n} \|\pi_{F_n}\| \rightarrow \infty$, we have

$$\frac{\hat{\pi}}{\|\pi_{F_n}\|} = \frac{\pi_{F_n}}{\|\pi_{F_n}\|} + O_p(n^{-1/2} \|\pi_{F_n}\|^{-1}) = \iota_{\infty} + o_p(1). \quad (36)$$

and

$$\frac{\hat{h}(\lambda_n)}{\sqrt{n} \|\pi_{F_n}\|} = \frac{\hat{\pi}}{\|\pi_{F_n}\|} - [\partial_p \pi] \hat{\Sigma}_p [\partial_p' \hat{g}(\lambda_n)] \hat{S}^{-1} \sqrt{n} \hat{g}(\lambda_n) \|\sqrt{n} \pi_{F_n}\|^{-1} = \iota_{\infty} + o_p(1). \quad (37)$$

From the condition $\iota_{\infty} \neq 0_{K \times 1}$, the convergence results established in Part(a), and (36),

$$W_n(\lambda_n) = \frac{n \hat{g}(\lambda_n)' \hat{S}(\lambda_n)^{-1} (\hat{\pi} / \|\pi_{F_n}\|) (\hat{\pi}' / \|\pi_{F_n}\|) \hat{S}(\lambda_n)^{-1} \hat{g}(\lambda_n)}{(\hat{\pi}' / \|\pi_{F_n}\|) \hat{S}(\lambda_n)^{-1} (\hat{\pi} / \|\pi_{F_n}\|)} \xrightarrow{d} \chi_1^2. \quad (38)$$

Now we examine the stochastic behavior of critical value $\hat{q}_{W^*}(1 - \alpha)$, defined as $(1 - \alpha)$ -quantile of $W_n^*(\lambda_n)$ conditional on data. First, define a normalized conditioning statistic in the construction of $W_n^*(\lambda_n)$:

$$\bar{\Upsilon} \equiv \left(\frac{\hat{h}(\lambda_n)}{\sqrt{n} \|\pi_{F_n}\|}, \hat{S}(\lambda_n), \hat{\Sigma}_p, \frac{\partial_p \hat{g}(\lambda_n)}{\sqrt{n} \|\pi_{F_n}\|} \right).$$

By the definition of $q(1 - \alpha, \Upsilon)$ introduced in Case 1 and the fact that

$$W_n^*(\lambda_n) \equiv W_n(\eta^*, \hat{\Upsilon}) = W_n(\eta^*, \bar{\Upsilon}),$$

we have

$$\hat{q}_{W^*}(1 - \alpha) = q(1 - \alpha, \bar{\Upsilon}).$$

It therefore suffices to obtain the limit of $\bar{\Upsilon}$ as the sample size diverges. By the consistency of $(\hat{S}(\lambda_n), \hat{\Sigma}_p, \partial_p \hat{g}(\lambda_n))$, (37), and the condition that $\sqrt{n} \|\pi_{F_n}\| \rightarrow \infty$, we obtain

$$\bar{\Upsilon} \xrightarrow{p} \bar{\Upsilon}_\infty \equiv (\iota_\infty, S_\infty, \Sigma_{p,\infty}, 0_{K \times (K+1)}) \in \mathcal{U},$$

where the inclusion in $\mathcal{U}$ follows from $\iota_\infty \neq 0_{K \times 1}$. This and the continuity of the quantile function $q(1 - \alpha, \cdot)$ give $q(1 - \alpha, \bar{\Upsilon}) \xrightarrow{p} q(1 - \alpha, \bar{\Upsilon}_\infty)$. So we have shown the $(1 - \alpha)$ -conditional quantile of $W_n^*(\lambda_n)$ converges in probability to $q(1 - \alpha, \bar{\Upsilon}_\infty)$, which equals the $(1 - \alpha)$ -quantile of χ_1^2 distribution by replacing $\hat{\Upsilon}$ with $\bar{\Upsilon}_\infty$ in (34). Hence, we conclude that

$$\hat{q}_{W^*}(1 - \alpha) \xrightarrow{p} q_{\chi_1^2}(1 - \alpha). \quad (39)$$

By (38) and (39), we have $W_n(\lambda_n) - \hat{q}_{W^*}(1 - \alpha) \xrightarrow{d} \chi_1^2 - q_{\chi_1^2}(1 - \alpha)$. Then the desired conclusion follows by the definition of convergence in distribution. $\square$

D.2 Proof of Proposition A.2

Proof of Proposition A.2. For simplicity of notations, the proof is shown for the full sequence $\{n\}$. Then, we note that the same proof goes through with p_n in place of n .

Since $A_{F_n} \theta_{F_n} = \beta_{F_n}$, the uniform delta method gives

$$\begin{aligned} \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) &= \sqrt{n}(\hat{A} - A_{F_n})\theta_{F_n} - \sqrt{n}(\hat{\beta} - \beta_{F_n}) \\ &= H(p_{F_n}, \theta_{F_n})\hat{Z}_p - \hat{Z}_\beta + o_p(1), \\ \sqrt{n}(\hat{a}_j - a_{j,F_n}) &= M_j(p_{F_n})\hat{Z}_p + o_p(1), \quad \text{for } j = 1, \dots, 2(M+1), \end{aligned}$$

where $H(p, \theta)$ and $M_j(p)$ are defined in (20) and (22). Both expansions are linear in $(\hat{Z}_p, \hat{Z}_\beta)$, which converges jointly to the Gaussian vector $(\mathcal{Z}_p, \mathcal{Z}_\beta)$ by Lemma D.1(a). Because H and M_j are continuous and $(p_{F_n}, \theta_{F_n}) \rightarrow (p_\infty, \theta_\infty)$, Slutsky's theorem and the continuous mapping theorem imply that the two left-hand sides converge jointly to $M_j(p_\infty)\mathcal{Z}_p$ and

$$\mathcal{Z}_m \equiv H(p_\infty, \theta_\infty)\mathcal{Z}_p - \mathcal{Z}_\beta.$$

By the same lemma and continuity,

$$\begin{aligned} \hat{\Omega}(\theta_{F_n}) &= H(\hat{p}, \theta_{F_n})\hat{\Sigma}_p H(\hat{p}, \theta_{F_n})' + \hat{\Sigma}_\beta \\ &\xrightarrow{p} H(p_\infty, \theta_\infty)\Sigma_{p,\infty} H(p_\infty, \theta_\infty)' + \Sigma_{\beta,\infty} = \text{var}(\mathcal{Z}_m), \end{aligned} \quad (40)$$

and

$$\begin{aligned}\hat{\Gamma}_j(\theta_{F_n}) &= M_j(\hat{p})\hat{\Sigma}_p H(\hat{p}, \theta_{F_n})' \\ &\xrightarrow{p} M_j(p_\infty)\Sigma_{p,\infty} H(p_\infty, \theta_\infty)' = \text{cov}(M_j(p_\infty)\mathcal{Z}_p, \mathcal{Z}_m).\end{aligned}\quad (41)$$

The asymptotic variance $\text{var}(\mathcal{Z}_m)$ is PD because it is a sum of two PSD matrices and $\Sigma_{\beta,\infty}$ is PD by the restrictions on $\mathcal{P}$. Combining these results,

$$\begin{aligned}\sqrt{n}(\hat{d}_j(\theta_{F_n}) - a_{j,F_n}) &= \sqrt{n}(\hat{a}_j - a_{j,F_n}) - \hat{\Gamma}_j(\theta_{F_n})\hat{\Omega}(\theta_{F_n})^{-1}\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) \\ &\xrightarrow{d} M_j(p_\infty)\mathcal{Z}_p - \text{cov}(M_j(p_\infty)\mathcal{Z}_p, \mathcal{Z}_m) \text{var}(\mathcal{Z}_m)^{-1}\mathcal{Z}_m \equiv \mathcal{Z}_{d_j},\end{aligned}$$

jointly over j and with $\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) \xrightarrow{d} \mathcal{Z}_m$. By construction $\text{cov}(\mathcal{Z}_{d_j}, \mathcal{Z}_m) = 0$ for every j . Since each $\mathcal{Z}_{d_j}$ and $\mathcal{Z}_m$ are linear functions of the Gaussian vector $(\mathcal{Z}_p, \mathcal{Z}_\beta)$, they are jointly Gaussian, so zero covariance implies that $\mathcal{Z}_D = (\mathcal{Z}_{d_1}, \dots, \mathcal{Z}_{d_{2(M+1)}})$ is independent of $\mathcal{Z}_m$. This shows

$$\sqrt{n} \begin{pmatrix} \hat{A}\theta_{F_n} - \hat{\beta} \\ \text{vec}(\hat{D}(\theta_{F_n})) - \text{vec}(A_{F_n}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{Z}_m \\ \text{vec}(\mathcal{Z}_D) \end{pmatrix}. \quad (42)$$

From the SVD of A_{F_n} in (27), construct a normalization matrix Ψ_{F_n} with positive diagonal elements:

$$\Psi_{F_n} \equiv \text{diag}\{(\sqrt{n}\tau_{1,F_n})^{-1}, \dots, (\sqrt{n}\tau_{q,F_n})^{-1}, 1, \dots, 1\} \in \mathbb{R}^{2(M+1) \times 2(M+1)}, \quad (43)$$

where $q \equiv \max\{j : t_j = \infty\}$. For all large n , Ψ_{F_n} is nonsingular, and since B_{F_n} is orthogonal and $c \neq 0$, the vector

$$\iota_n \equiv \frac{\Psi_{F_n} B_{F_n}' c}{\|\Psi_{F_n} B_{F_n}' c\|}$$

is well defined and has unit length. I first assume that $\iota_n \rightarrow \iota_\infty$, and I remove this assumption in the last step.

By (42) and Lemma D.5, jointly with $\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) \xrightarrow{d} \mathcal{Z}_m$,

$$\hat{G}_n \equiv \sqrt{n} \tilde{D}(\theta_{F_n}) B_{F_n} \Psi_{F_n} \xrightarrow{d} \mathcal{D}_\xi,$$

where $\mathcal{D}_\xi$ has full column rank almost surely and is independent of $\mathcal{Z}_m$. Write $\hat{\Omega}_n \equiv \hat{\Omega}(\theta_{F_n})$. Substituting $\tilde{D}(\theta_{F_n}) = n^{-1/2} \hat{G}_n \Psi_{F_n}^{-1} B_{F_n}'$ into $\hat{Q}(\theta_{F_n})$ gives

$$\hat{Q}(\theta_{F_n}) = \sqrt{n} \|\Psi_{F_n} B_{F_n}' c\| \cdot \hat{Q}_{\text{norm}}, \quad \hat{Q}_{\text{norm}} \equiv \hat{\Omega}_n^{-1/2} \hat{G}_n (\hat{G}_n' \hat{\Omega}_n^{-1} \hat{G}_n)^{-1} \iota_n.$$

A projection is invariant to positive rescaling of the vector it projects onto, so $P_{\hat{Q}(\theta_{F_n})} = P_{\hat{Q}_{\text{norm}}}$. By (40) and the continuous mapping theorem, we can show $\hat{Q}(\theta_{F_n})$ has a limit after normalizing by $(\sqrt{n} \|\Psi_{F_n} B_{F_n}' c\|)^{-1}$:

$$\hat{Q}_{\text{norm}} \xrightarrow{d} \mathcal{Q} \equiv \text{var}(\mathcal{Z}_m)^{-1/2} \mathcal{D}_\xi (\mathcal{D}_\xi' \text{var}(\mathcal{Z}_m)^{-1} \mathcal{D}_\xi)^{-1} \iota_\infty,$$

jointly with $\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})$. Since $\mathcal{D}_\xi$ has full column rank almost surely and $\|\iota_\infty\| = 1$, the limit $\mathcal{Q} \neq 0$ almost surely. Moreover, $\mathcal{Q}$ is independent of $\mathcal{Z}_m$, since $\mathcal{D}_\xi \perp \mathcal{Z}_m$ from Lemma D.5.

Let $\mathcal{V} \equiv \text{var}(\mathcal{Z}_m)^{-1/2} \mathcal{Z}_m \sim \mathcal{N}(0, I_{2(K+1)})$ and $M_Q \equiv I - P_Q$. The map $Q \mapsto P_Q$ is continuous at every $Q \neq 0$. The continuous mapping theorem applied to (14) and (15) therefore gives

$$(\text{MRLM}_n(\theta_{F_n}), \text{AR}_n(\theta_{F_n}) - \text{MRLM}_n(\theta_{F_n})) \xrightarrow{d} (\mathcal{V}' P_Q \mathcal{V}, \mathcal{V}' M_Q \mathcal{V}).$$

Conditional on $\mathcal{Q}$, note that $\mathcal{V} \sim \mathcal{N}(0, I_{2(K+1)})$ and P_Q is a rank-one orthogonal projection. By Cochran's theorem, $\mathcal{V}' P_Q \mathcal{V}$ and $\mathcal{V}' M_Q \mathcal{V}$ are then independent χ_1^2 and χ_{2K+1}^2 variables. This conditional law does not depend on $\mathcal{Q}$, so it is also the unconditional law. Hence, by (16),

$$\text{MLC}_n(\theta_{F_n}) = (1 + a) \text{MRLM}_n(\theta_{F_n}) + a(\text{AR}_n(\theta_{F_n}) - \text{MRLM}_n(\theta_{F_n})) \xrightarrow{d} (1 + a)\chi_1^2 + a\chi_{2K+1}^2,$$

with the two chi-square variables independent. This limit law is continuous. Since $\theta_{F_n} \in \Theta$ and $c'\theta_{F_n} = \lambda_n$, the infimum in the test is at most $\text{MLC}_n(\theta_{F_n})$, so

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \mathbb{P}_{F_n} \left(\inf_{c'\theta = \lambda_n} \text{MLC}_n(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) \\ & \leq \lim_{n \rightarrow \infty} \mathbb{P}_{F_n} \left(\text{MLC}_n(\theta_{F_n}) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) = \alpha. \end{aligned}$$

Finally, I remove the assumption that ι_n converges. Suppose the conclusion of the proposition fails along $\{n\}$. Then there exist $\varepsilon > 0$ and a subsequence $\{n_k\}$ along which the null rejection probability exceeds $\alpha + \varepsilon$ for all k . Because ι_{n_k} lies in the compact unit sphere, $\{n_k\}$ has a further subsequence along which $\iota_{n_k} \rightarrow \iota_\infty$ for some unit vector ι_∞ . Conditions 1–7 continue to hold along this further subsequence with the same limits. The argument above then shows that the limit superior of the null rejection probability is at most α , which is a contradiction. $\square$

D.3 Supporting Lemmas

In this section, I provide supporting lemmas used in Proposition A.1 and Proposition A.2, followed by the proofs of these results.

Lemma D.1. Let $\{(\theta_{F_n}, F_n)\}_{n \geq 1} \subset \mathcal{P}$ satisfy $q_{F_n} \rightarrow q_\infty$, $p_{F_n} \rightarrow p_\infty$, $\beta_{F_n} \rightarrow \beta_\infty$, and $\sigma_{F_n, d\ell}^2 \rightarrow \sigma_{\infty, d\ell}^2$ for $d = 0, 1$ and $\ell = 0, 1, \dots, K$. Then we have

(a) The estimators are jointly asymptotically normal:

$$\begin{pmatrix} \hat{Z}_p \\ \hat{Z}_{\beta_1} \\ \hat{Z}_{\beta_0} \\ \hat{Z}_q \end{pmatrix} \equiv \sqrt{n} \begin{pmatrix} \hat{p} - p_{F_n} \\ \hat{\beta}_1 - \beta_{F_n, 1} \\ \hat{\beta}_0 - \beta_{F_n, 0} \\ \hat{q} - q_{F_n} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{Z}_p \\ \mathcal{Z}_{\beta_1} \\ \mathcal{Z}_{\beta_0} \\ \mathcal{Z}_q \end{pmatrix} \sim \mathcal{N}(0_{4(K+1) \times 1}, \text{diag}\{\Sigma_{p, \infty}, \Sigma_{\beta_1, \infty}, \Sigma_{\beta_0, \infty}, \Sigma_{q, \infty}\}), \quad (44)$$

where

$$\begin{aligned}\Sigma_{p,\infty} &\equiv \text{diag} \left\{ \frac{p_\infty(z_\ell)(1-p_\infty(z_\ell))}{q_\infty(z_\ell)} \right\}_{\ell=0}^K, & \Sigma_{\beta_1,\infty} &\equiv \text{diag} \left\{ \frac{\sigma_{\infty,1\ell}^2}{q_\infty(z_\ell)p_\infty(z_\ell)} \right\}_{\ell=0}^K, \\ \Sigma_{\beta_0,\infty} &\equiv \text{diag} \left\{ \frac{\sigma_{\infty,0\ell}^2}{q_\infty(z_\ell)(1-p_\infty(z_\ell))} \right\}_{\ell=0}^K, & \Sigma_{q,\infty} &\equiv \text{diag}\{q_\infty\} - q_\infty q'_\infty.\end{aligned}$$

(b) Let $\hat{\Sigma}_p$, $\hat{\Sigma}_{\beta_d}$ and $\hat{\Sigma}_q$ be the sample analogs of $\Sigma_{p,\infty}$, $\Sigma_{\beta_d,\infty}$ and $\Sigma_{q,\infty}$. They replace $(p_\infty, q_\infty, q_\infty(d, z_\ell), \sigma_{\infty,d\ell}^2)$ with $(\hat{p}, \hat{q}, \hat{q}(d, z_\ell), \hat{\sigma}_{d\ell}^2)$, where

$$\hat{\sigma}_{d\ell}^2 \equiv \frac{1}{n} \sum_{i=1}^n \frac{(Y_i - \hat{\beta}_{d\ell})^2 \mathbb{1}[D_i = d, Z_i = z_\ell]}{\hat{q}(d, z_\ell)}.$$

Then $\hat{\Sigma}_p \xrightarrow{p} \Sigma_{p,\infty}$, $\hat{\Sigma}_\beta \equiv \text{diag}\{\hat{\Sigma}_{\beta_1}, \hat{\Sigma}_{\beta_0}\} \xrightarrow{p} \Sigma_{\beta,\infty} \equiv \text{diag}\{\Sigma_{\beta_1,\infty}, \Sigma_{\beta_0,\infty}\}$, and $\hat{\Sigma}_q \xrightarrow{p} \Sigma_{q,\infty}$.

Lemma D.2. Suppose $c_\mu \neq 0$ or $c_\rho \neq 0$ are fixed weights and let $\lambda = c'\theta$. For any $(\theta, F) \in \mathcal{P}$, the asymptotic variance matrix

$$S \equiv [\partial_p g(\lambda)] \Sigma_p [\partial_{p'} g(\lambda)] + [\partial_{\beta_1} g(\lambda)] \Sigma_{\beta_1} [\partial_{\beta'_1} g(\lambda)] + [\partial_{\beta_0} g(\lambda)] \Sigma_{\beta_0} [\partial_{\beta'_0} g(\lambda)]$$

is uniformly bounded and positive definite. That is,

$$\inf_{(\theta, F) \in \mathcal{P}} \tau_{\min}(S) > 0, \quad (45)$$

and

$$\sup_{(\theta, F) \in \mathcal{P}} \tau_{\max}(S) < \infty, \quad (46)$$

where $\tau_{\min}(S)$ and $\tau_{\max}(S)$ denote the smallest and largest eigenvalues (or singular value) of S , respectively.

Remark D.1. This lemma shows that any choice of nonzero linear weights c_μ and c_ρ would not entail an asymptotic singular covariance matrix for AR test, which is essential for establishing uniform validity. $\square$

Remark D.2. Indeed, we can relax the restriction that $\lambda = c'\theta$ and show this result for all λ that belongs to a compact subset on the real line, which will be useful to establish the power property of the test. Since the above statement is sufficient to prove the uniform validity of the conditional Wald test, I do not pursue this extension here. $\square$

Lemma D.3. Suppose the conditions in Proposition A.1 hold for a sequence (λ_n, F_n) for which $\sqrt{n}\|\pi_{F_n}\| \rightarrow \psi_\infty < \infty$. Then $\Upsilon_\infty = (\mathcal{Z}_h, S_\infty, \Sigma_{p,\infty}, \partial_p g(\lambda_\infty))$ belongs to the set $\mathcal{U}$ defined in (35) almost surely. That is,

$$\mathbb{P}_{\eta^*} \left(\mathcal{Z}_h + [\partial_p \pi] \Sigma_{p,\infty} \partial_{p'} g(\lambda_\infty) S_\infty^{-1/2} \eta^* \neq 0_{K \times 1} \mid \mathcal{Z}_h \right) = 1 \quad \text{a.s.}$$

Lemma D.4. [Andrews and Guggenberger (2017, Lemma 10.3)] Assume the following conditions hold under a sequence of DGPs $\{(\theta_{F_n}, F_n)\}_{n \geq 1}$:

1. $\sqrt{n}\tau_{j,F} \rightarrow \infty$ for $j \leq q$ and $\sqrt{n}\tau_{j,F_n} \rightarrow t_j \in [0, \infty)$ for $j > q$;
2. C_{F_n} and B_{F_n} converge to their limits C_∞ and B_∞ , respectively;
3. The moment condition and the normalized $\hat{D}(\theta_{F_n})$ converge jointly:

$$\sqrt{n} \begin{pmatrix} \hat{A}\theta_{F_n} - \hat{\beta} \\ \text{vec}(\hat{D}(\theta_{F_n})) - \text{vec}(A_{F_n}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{Z}_m \\ \text{vec}(\mathcal{Z}_D) \end{pmatrix}$$

where $\mathcal{Z}_D = (\mathcal{Z}_{d_1}, \dots, \mathcal{Z}_{d_{2(M+1)}})$ is independent of $\mathcal{Z}_m$.

Then we have

$$n^{1/2} \hat{D}(\theta_{F_n}) B_{F_n} \Psi_{F_n} \xrightarrow{d} \mathcal{D} = O_p(1)$$

Moreover, $\mathcal{D}$ is independent of $\mathcal{Z}_m$.

Lemma D.5. Suppose the assumptions in Lemma D.4 hold. Let $\xi \in \mathbb{R}^{2(K+1) \times 2(M+1)}$ be a matrix of i.i.d. standard normal random variables, and let $\kappa > 0$. Then, jointly with $\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) \xrightarrow{d} \mathcal{Z}_m$,

$$n^{1/2}(\hat{D}(\theta_{F_n}) + \kappa n^{-1/2}\xi) B_{F_n} \Psi_{F_n} \xrightarrow{d} \mathcal{D}_\xi$$

where $\mathcal{D}_\xi$ has full column rank with probability one and is independent of $\mathcal{Z}_m$.

Lemma D.6 (WLLN for $L^{1+\delta}$ triangular array). Let $\{X_{ni}\}$ be a row-wise i.i.d. triangular array of random variables. Suppose

$$\sup_n \mathbb{E}|X_{ni}|^{1+\delta} < \infty \tag{47}$$

for some $\delta > 0$. Then

$$\frac{1}{n} \sum_{i=1}^n X_{ni} - \mathbb{E}[X_{ni}] \xrightarrow{p} 0.$$

Lemma D.7 (Multivariate Lyapunov CLT). Let $\{X_{ni}\}$ be a row-wise i.i.d. triangular array of random vectors in $\mathbb{R}^k$ with expectations $\mathbb{E}[X_{ni}] = 0$ and covariance matrices $\Sigma_n = \mathbb{E}[X_{ni}X'_{ni}]$. Suppose

$$\Sigma_n \rightarrow \Sigma \geq 0 \tag{48}$$

and for some $\delta > 0$

$$\sup_n \mathbb{E}\|X_{ni}\|^{2+\delta} < \infty. \tag{49}$$

Then as $n \rightarrow \infty$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_{ni} \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Proof of Lemma D.1. Part (a): Let $q_F(1, z_\ell) \equiv q_F(z_\ell)p_F(z_\ell)$ and $q_F(0, z_\ell) \equiv q_F(z_\ell)(1 - p_F(z_\ell))$. Under Definition 2.1, $q_F(d, z_\ell) \geq \epsilon^2$ for all $(\theta, F) \in \mathcal{P}$. Define the vector of cell sums

$$W_i \equiv (\mathbb{1}[Z_i = z_\ell], \mathbb{1}[D_i = 1, Z_i = z_\ell], Y_i \mathbb{1}[D_i = 1, Z_i = z_\ell], Y_i \mathbb{1}[D_i = 0, Z_i = z_\ell])_{\ell=0}^K,$$

with mean $\mu_F \equiv (q_F(z_\ell), q_F(1, z_\ell), q_F(1, z_\ell)\beta_{F,1\ell}, q_F(0, z_\ell)\beta_{F,0\ell})_{\ell=0}^K$. Condition 2 of Definition 2.1 gives $\sup_n \mathbb{E}_{F_n} \|W_i\|^{2+\delta} < \infty$. The covariance matrix of W_i is a continuous function of $(q_F, p_F, \beta_F, \sigma_F^2)$, so it converges along $\{F_n\}$. By the CLT (Lemma D.7), $\sqrt{n}(\bar{W}_n - \mu_{F_n})$ converges in distribution to a mean-zero normal vector, where $\bar{W}_n$ denotes the sample average of $\{W_i\}_{i=1}^n$.

Note that estimators are a smooth function of the cell means: $(\hat{p}, \hat{\beta}_1, \hat{\beta}_0, \hat{q}) = \varphi(\bar{W}_n)$, where

$$\varphi((a_\ell, b_\ell, c_{1\ell}, c_{0\ell})_\ell) \equiv ((b_\ell/a_\ell)_\ell, (c_{1\ell}/b_\ell)_\ell, (c_{0\ell}/(a_\ell - b_\ell))_\ell, (a_\ell)_\ell).$$

By construction, $\varphi(\mu_F) = (p_F, \beta_{F,1}, \beta_{F,0}, q_F)$.

Conditions 2–4 of Definition 2.1 give $a_\ell \geq \epsilon$, $b_\ell \geq \epsilon^2$ and $a_\ell - b_\ell \geq \epsilon^2$ for $\mu_F = (a_\ell, b_\ell, c_{1\ell}, c_{0\ell})_\ell$ and every $(\theta, F) \in \mathcal{P}$. These closed bounds also hold for $\mu_\infty \equiv \lim_n \mu_{F_n}$, which exists under the conditions of the lemma. Every component of φ is a ratio with denominator a_ℓ , b_ℓ or $a_\ell - b_\ell$, so φ is continuously differentiable on a neighborhood of μ_∞ . Moreover, $\bar{W}_n$ lies in that neighborhood with probability approaching one, since $\bar{W}_n - \mu_{F_n} = O_p(n^{-1/2})$. The uniform delta method (Van der Vaart, 2000, Theorem 3.8) gives

$$\sqrt{n}(\varphi(\bar{W}_n) - \varphi(\mu_{F_n})) = \partial\varphi(\mu_\infty)\sqrt{n}(\bar{W}_n - \mu_{F_n}) + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_{F_n}(Y_i, D_i, Z_i) + o_p(1),$$

where $\phi_F(Y_i, D_i, Z_i) \equiv \partial\varphi(\mu_F)(W_i - \mu_F)$, and the last equality uses $\partial\varphi(\mu_{F_n}) \rightarrow \partial\varphi(\mu_\infty)$. Partition ϕ_F conformably with $(\hat{p}, \hat{\beta}_1, \hat{\beta}_0, \hat{q})$ as

$$\phi_F = (\phi'_{p,F}, \phi'_{\beta_1,F}, \phi'_{\beta_0,F}, \phi'_{q,F})', \quad \phi_{x,F} = (\phi_{x,F,0}, \dots, \phi_{x,F,K})' \quad \text{for } x \in \{p, \beta_1, \beta_0, q\},$$

Each block of ϕ_F corresponds to one of the vectors p_F , $\beta_{F,1}$, $\beta_{F,0}$ and q_F , and its ℓ -th element corresponds to that quantity evaluated at $z = z_\ell$. That is, $\phi_{p,F,\ell}$, $\phi_{\beta_d,F,\ell}$ and $\phi_{q,F,\ell}$ are the influence functions of $\hat{p}(z_\ell)$, $\hat{\beta}_{d\ell}$ and $\hat{q}(z_\ell)$, respectively. The ℓ -th element of each block of φ depends only on the ℓ -th cell $(a_\ell, b_\ell, c_{1\ell}, c_{0\ell})$ of its argument. The corresponding cell of $W_i - \mu_F$ is

$$(\mathbb{1}[Z_i = z_\ell] - q_F(z_\ell), \mathbb{1}[D_i = 1, Z_i = z_\ell] - q_F(1, z_\ell), (Y_i \mathbb{1}[D_i = d, Z_i = z_\ell] - q_F(d, z_\ell)\beta_{F,d\ell})_{d=1,0}).$$

Differentiating each ratio in φ at μ_F and applying it to this cell gives, for $\ell = 0, 1, \dots, K$ and $d = 0, 1$,

$$\begin{aligned} \phi_{p,F,\ell}(Y, D, Z) &= \frac{(D - p_F(z_\ell))\mathbb{1}[Z = z_\ell]}{q_F(z_\ell)}, \\ \phi_{\beta_d,F,\ell}(Y, D, Z) &= \frac{(Y - \beta_{F,d\ell})\mathbb{1}[D = d, Z = z_\ell]}{q_F(d, z_\ell)}, \\ \phi_{q,F,\ell}(Y, D, Z) &= \mathbb{1}[Z = z_\ell] - q_F(z_\ell). \end{aligned} \tag{50}$$

By (50), $\mathbb{E}_F[\phi_{p,F,\ell} \mid Z] = 0$ and $\mathbb{E}_F[\phi_{\beta_d,F,\ell} \mid D, Z] = 0$, so the four blocks of ϕ_F are mutually uncorrelated. Within each of the blocks $\phi_{p,F}$, $\phi_{\beta_1,F}$ and $\phi_{\beta_0,F}$, distinct ℓ index disjoint cells, so each block has a diagonal covariance matrix. Its diagonal entries are

$$\text{var}_F(\phi_{p,F,\ell}) = \frac{p_F(z_\ell)(1 - p_F(z_\ell))}{q_F(z_\ell)}, \quad \text{var}_F(\phi_{\beta_d,F,\ell}) = \frac{\sigma_{F,d\ell}^2}{q_F(d, z_\ell)}, \quad \text{and} \quad \text{var}_F(\phi_{q,F}) = \text{diag}\{q_F\} - q_F q_F'.$$

These variances converge to the matrices in (44), and the CLT (Lemma D.7) applied to ϕ_{F_n} then establishes (44).

Part (b): The matrices $\hat{\Sigma}_p$ and $\hat{\Sigma}_q$ are continuous functions of $(\hat{p}, \hat{q})$. Their consistency follows from part (a) and the continuous mapping theorem. For $\hat{\sigma}_{d\ell}^2$, expanding the square and using $n^{-1} \sum_i Y_i \mathbb{1}[D_i = d, Z_i = z_\ell] = \hat{q}(d, z_\ell) \hat{\beta}_{d\ell}$ gives

$$\hat{\sigma}_{d\ell}^2 = \frac{1}{\hat{q}(d, z_\ell)} \cdot \frac{1}{n} \sum_{i=1}^n Y_i^2 \mathbb{1}[D_i = d, Z_i = z_\ell] - \hat{\beta}_{d\ell}^2.$$

The summands have $\sup_n \mathbb{E}_{F_n} (|Y_i^2 \mathbb{1}[D_i = d, Z_i = z_\ell]|^{1+\delta/2}) \leq \zeta$. By the WLLN (Lemma D.6),

$$\frac{1}{n} \sum_{i=1}^n Y_i^2 \mathbb{1}[D_i = d, Z_i = z_\ell] = q_{F_n}(d, z_\ell)(\sigma_{F_n, d\ell}^2 + \beta_{F_n, d\ell}^2) + o_p(1).$$

Together with part (a), this gives $\hat{\sigma}_{d\ell}^2 - \sigma_{F_n, d\ell}^2 \xrightarrow{p} 0$. Since $\sigma_{F_n, d\ell}^2 \rightarrow \sigma_{\infty, d\ell}^2$, the continuous mapping theorem gives $\hat{\Sigma}_\beta \xrightarrow{p} \Sigma_{\beta, \infty}$. $\square$

Proof of Lemma D.2. I first show (45). Let $a_k \equiv (1 - p(z_k))c_\mu + 2c_\rho$ and $b_k \equiv p(z_k)c_\mu - 2c_\rho$ for $k = 0, 1, \dots, K$. By (9), the k -th row of $\partial_{\beta_1} g(\lambda)$ has a_k in position 0 and $-a_0$ in position k . Likewise, the k -th row of $\partial_{\beta_0} g(\lambda)$ has b_k in position 0 and $-b_0$ in position k . Hence $\partial_{\beta_1} g(\lambda) = (\nu_1, -a_0 I_K)$ and $\partial_{\beta_0} g(\lambda) = (\nu_0, -b_0 I_K)$, where $\nu_1 \equiv (a_1, \dots, a_K)'$ and $\nu_0 \equiv (b_1, \dots, b_K)'$. Since Σ_{β_d} is diagonal,

$$[\partial_{\beta_1} g(\lambda)] \Sigma_{\beta_1} [\partial_{\beta_1'} g(\lambda)] = \Sigma_{\beta_1} [0, 0] \nu_1 \nu_1' + a_0^2 \text{diag}\{\Sigma_{\beta_1}[k, k]\}_{k=1}^K,$$

and the same holds for $[\partial_{\beta_0} g(\lambda)] \Sigma_{\beta_0} [\partial_{\beta_0'} g(\lambda)]$ with (ν_0, b_0) in place of (ν_1, a_0) . Therefore, we can decompose S into the sum of PSD matrices:

$$\begin{aligned} S &= \Sigma_{\beta_1} [0, 0] \nu_1 \nu_1' + a_0^2 \text{diag}\{\Sigma_{\beta_1}[k, k]\}_{k=1}^K \\ &\quad + \Sigma_{\beta_0} [0, 0] \nu_0 \nu_0' + b_0^2 \text{diag}\{\Sigma_{\beta_0}[k, k]\}_{k=1}^K + \partial_p g(\lambda) \Sigma_p \partial_{p'} g(\lambda). \end{aligned}$$

By Weyl's inequality, $\tau_{\min}(S) \geq \min_{1 \leq k \leq K} (\Sigma_{\beta_1}[k, k] a_0^2 + \Sigma_{\beta_0}[k, k] b_0^2)$. Under Definition 2.1, $\Sigma_{\beta_d}[k, k] = \sigma_{dk}^2 / q(d, z_k) \geq \epsilon / (1 - \epsilon)^2 \equiv \underline{\delta}(\epsilon)$, so $\tau_{\min}(S) \geq \underline{\delta}(\epsilon)(a_0^2 + b_0^2)$. Moreover, $(a_0, b_0)' = \mathbf{M}(c_\mu, c_\rho)'$ with

$$\mathbf{M} \equiv \begin{pmatrix} 1 - p(z_0) & 2 \\ p(z_0) & -2 \end{pmatrix}, \quad |\det \mathbf{M}| = 2, \quad \|\mathbf{M}\| \leq \|\mathbf{M}\|_F = ((1 - p(z_0))^2 + p(z_0)^2 + 8)^{1/2} \leq 3,$$

where $\|\mathbf{M}\|$ is the spectral norm (the largest singular value) and $\|\mathbf{M}\|_F$ is the Frobenius norm. Since $|\det \mathbf{M}|$ is the product of the two singular values of $\mathbf{M}$, the smallest singular value equals $|\det \mathbf{M}| / \|\mathbf{M}\| \geq 2/3$. Hence $a_0^2 + b_0^2 = \|\mathbf{M}(c_\mu, c_\rho)'\|^2 \geq \frac{4}{9}(c_\mu^2 + c_\rho^2)$. Therefore,

$$\inf_{(\theta, F) \in \mathcal{P}} \tau_{\min}(S) \geq \frac{4}{9} \underline{\delta}(\epsilon) (c_\mu^2 + c_\rho^2) > 0,$$

since $c_\mu \neq 0$ or $c_\rho \neq 0$.

Next, I show (46). For $x \in \{p, \beta_1, \beta_0\}$, $\tau_{\max}(\partial_x g \Sigma_x \partial_x' g) \leq \|\partial_x g(\lambda)\|^2 \tau_{\max}(\Sigma_x)$. Under Definition 2.1, $\tau_{\max}(\Sigma_p) \leq 1/(4\epsilon)$ and $\tau_{\max}(\Sigma_{\beta_d}) \leq \zeta^{2/(2+\delta)}/\epsilon^2$. The entries of $\partial_x g(\lambda)$ are polynomials in $(\lambda, p, \beta, c_\mu, c_\rho)$, where $\lambda = c'\theta$ with θ in the compact set Θ , $p \in [0, 1]^{K+1}$, and $|\beta_{d\ell}| \leq \zeta^{1/(2+\delta)}$. Hence $\sup_{(\theta, F) \in \mathcal{P}} \|\partial_x g(\lambda)\| < \infty$. The subadditivity of $\tau_{\max}$ over PSD matrices completes the proof. $\square$

Proof of Lemma D.3. Let $\mathbf{H} \equiv [\partial_p \pi] \Sigma_{p,\infty} [\partial_p \pi]'$ and $\mathcal{V} \equiv \mathcal{Z}_h + \mathbf{H} S_\infty^{-1/2} \eta^*$, where $\eta^* \sim \mathcal{N}(0, I_K)$ is independent of $\mathcal{Z}_h$. The vector $\mathcal{Z}_h - \psi_\infty \iota_\infty$ is the residual from the projection of $[\partial_p \pi] \mathcal{Z}_p$ on $\mathcal{Z}_g$, and $\text{cov}([\partial_p \pi] \mathcal{Z}_p, \mathcal{Z}_g) = \mathbf{H}$. Hence

$$\text{var}(\mathcal{V}) = ([\partial_p \pi] \Sigma_{p,\infty} [\partial_p \pi]' - \mathbf{H} S_\infty^{-1} \mathbf{H}') + \mathbf{H} S_\infty^{-1} \mathbf{H}' = [\partial_p \pi] \Sigma_{p,\infty} [\partial_p \pi]',$$

which is PD because $\partial_p \pi$ has full row rank and $\Sigma_{p,\infty}$ is PD. So $\mathcal{V}$ is a normal vector with a PD covariance matrix, and $\mathbb{P}(\mathcal{V} = 0_{K \times 1}) = 0$. Since $\mathbb{E}[\mathbb{P}(\mathcal{V} = 0 \mid \mathcal{Z}_h)] = \mathbb{P}(\mathcal{V} = 0) = 0$, it follows that $\mathbb{P}(\mathcal{V} = 0 \mid \mathcal{Z}_h) = 0$ almost surely. This implies that $\Upsilon_\infty \in \mathcal{U}$ almost surely. $\square$

Proof of Lemma D.4. This lemma is a simplified version of Lemma 10.3(d) in Andrews and Guggenberger (2017). For completeness, I give a proof that follows their argument. Let $d_\theta \equiv 2(M+1)$ and $d_m \equiv 2(K+1)$ denote the numbers of columns and rows of $\hat{D}(\theta)$. Partition $B_{F_n} = (B_q, B_{d_\theta-q})$ and $C_{F_n} = (C_q, C_{d_m-q})$, where B_q and C_q contain the first q columns. Let $B_{q,\infty}$, $B_{d_\theta-q,\infty}$, $C_{q,\infty}$ and $C_{d_m-q,\infty}$ denote the corresponding blocks of B_∞ and C_∞ . By (43), $\Psi_{F_n} = \text{diag}\{\Psi_q, I_{d_\theta-q}\}$ with $\Psi_q \equiv \text{diag}\{(\sqrt{n}\tau_{1,F_n})^{-1}, \dots, (\sqrt{n}\tau_{q,F_n})^{-1}\}$, and $\Psi_q \rightarrow 0_{q \times q}$ by condition 1. Hence

$$\sqrt{n} \hat{D}(\theta_{F_n}) B_{F_n} \Psi_{F_n} = (\sqrt{n} \hat{D}(\theta_{F_n}) B_q \Psi_q, \sqrt{n} \hat{D}(\theta_{F_n}) B_{d_\theta-q}),$$

and I derive the limit of each block in turn.

For the first block,

$$\begin{aligned} \sqrt{n} \hat{D}(\theta_{F_n}) B_q \Psi_q &= \sqrt{n} A_{F_n} B_q \Psi_q + \sqrt{n} (\hat{D}(\theta_{F_n}) - A_{F_n}) B_q \Psi_q \\ &\xrightarrow{p} C_{q,\infty}. \end{aligned}$$

By (27) and the orthogonality of B_{F_n} , $\sqrt{n} A_{F_n} B_q \Psi_q = C_q$, which converges to $C_{q,\infty}$ by condition 2. The second term is $o_p(1)$, because $\sqrt{n} (\hat{D}(\theta_{F_n}) - A_{F_n}) = O_p(1)$ by condition 3 and $\Psi_q \rightarrow 0_{q \times q}$.

For the second block, define

$$T_\infty \equiv \begin{bmatrix} 0_{q \times (d_\theta-q)} \\ \text{diag}\{t_{q+1}, \dots, t_{d_\theta}\} \\ 0_{(d_m-d_\theta) \times (d_\theta-q)} \end{bmatrix} \in \mathbb{R}^{d_m \times (d_\theta-q)}.$$

Then

$$\begin{aligned} \sqrt{n} \hat{D}(\theta_{F_n}) B_{d_\theta-q} &= \sqrt{n} A_{F_n} B_{d_\theta-q} + \sqrt{n} (\hat{D}(\theta_{F_n}) - A_{F_n}) B_{d_\theta-q} \\ &\xrightarrow{d} C_\infty T_\infty + \mathcal{Z}_D B_{d_\theta-q,\infty}. \end{aligned}$$

The first term converges because $\sqrt{n}A_{F_n}B_{d_\theta-q} = C_{F_n}(\sqrt{n}\Pi_{F_n})B'_{F_n}B_{d_\theta-q} \rightarrow C_\infty T_\infty$ by conditions 1 and 2. The second term converges to $\mathcal{Z}_D B_{d_\theta-q,\infty}$, jointly with $\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})$, by conditions 2 and 3 and the continuous mapping theorem.

Combining the two blocks gives

$$\sqrt{n}\hat{D}(\theta_{F_n})B_{F_n}\Psi_{F_n} \xrightarrow{d} \mathcal{D} \equiv (C_{q,\infty}, C_\infty T_\infty + \mathcal{Z}_D B_{d_\theta-q,\infty}).$$

The limit $\mathcal{D}$ depends on the data only through $\mathcal{Z}_D$, so it is independent of $\mathcal{Z}_m$. $\square$

Proof of Lemma D.5. Again, let $B_{q,\infty}$, $B_{d_\theta-q,\infty}$, $C_{q,\infty}$ and $C_{d_m-q,\infty}$ denote the corresponding blocks of B_∞ and C_∞ . Let T_∞ be the matrix defined in the proof of Lemma D.4.

Step 1: establish the limit. The perturbation contributes $\sqrt{n} \cdot \kappa n^{-1/2} \xi B_{F_n} \Psi_{F_n} = \kappa \xi B_{F_n} \Psi_{F_n}$. Since $(\sqrt{n}\tau_{j,F_n})^{-1} \rightarrow 0$ for $j \leq q$, $\Psi_{F_n} \rightarrow \text{diag}\{0_{q \times q}, I_{d_\theta-q}\}$. Together with $B_{F_n} \rightarrow B_\infty$, this gives

$$\kappa \xi B_{F_n} \Psi_{F_n} \rightarrow \kappa \xi B_\infty \text{diag}\{0_{q \times q}, I_{d_\theta-q}\} = (0_{d_m \times q}, \kappa \xi B_{d_\theta-q,\infty})$$

for every realization of ξ . Because ξ is independent of the data, Lemma D.4 implies

$$(\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}), \sqrt{n}\hat{D}(\theta_{F_n})B_{F_n}\Psi_{F_n}, \xi) \xrightarrow{d} (\mathcal{Z}_m, \mathcal{D}, \xi),$$

with ξ independent of $(\mathcal{Z}_m, \mathcal{Z}_D)$. The continuous mapping theorem then gives, jointly with $\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) \xrightarrow{d} \mathcal{Z}_m$,

$$\sqrt{n}(\hat{D}(\theta_{F_n}) + \kappa n^{-1/2} \xi)B_{F_n}\Psi_{F_n} \xrightarrow{d} \mathcal{D}_\xi \equiv (C_{q,\infty}, C_\infty T_\infty + (\mathcal{Z}_D + \kappa \xi)B_{d_\theta-q,\infty}).$$

The limit $\mathcal{D}_\xi$ is a function of $(\mathcal{Z}_D, \xi)$. Since $\mathcal{Z}_D$ is independent of $\mathcal{Z}_m$ and ξ is independent of $(\mathcal{Z}_m, \mathcal{Z}_D)$, $(\mathcal{Z}_D, \xi)$ is independent of $\mathcal{Z}_m$, which implies the independence between $\mathcal{Z}_m$ and $\mathcal{D}_\xi$.

Step 2: full column rank when $q = d_\theta$. In this case the second block of $\mathcal{D}_\xi$ is empty, and $\mathcal{D}_\xi = C_{q,\infty}$ consists of the first d_θ columns of the orthogonal matrix C_∞ . So $\mathcal{D}_\xi$ has full column rank, and it equals $\mathcal{D}$. Under strong identification, the perturbation therefore has no effect on the limit.

Step 3: full column rank when $q < d_\theta$. Since C_∞ is orthogonal,

$$C'_\infty \mathcal{D}_\xi = \begin{bmatrix} I_q & C'_{q,\infty}(C_\infty T_\infty + (\mathcal{Z}_D + \kappa \xi)B_{d_\theta-q,\infty}) \\ 0_{(d_m-q) \times q} & E \end{bmatrix},$$

where

$$E \equiv C'_{d_m-q,\infty} C_\infty T_\infty + C'_{d_m-q,\infty} \mathcal{Z}_D B_{d_\theta-q,\infty} + \kappa C'_{d_m-q,\infty} \xi B_{d_\theta-q,\infty} \in \mathbb{R}^{(d_m-q) \times (d_\theta-q)},$$

The matrix $C'_\infty \mathcal{D}_\xi$ is block upper triangular with an identity block, and C_∞ is nonsingular. So $\text{rank}(\mathcal{D}_\xi) = \text{rank}(C'_\infty \mathcal{D}_\xi) = q + \text{rank}(E)$, and $\mathcal{D}_\xi$ has full column rank d_θ if and only if E has full column rank $d_\theta - q$.

It remains to show that E has full column rank almost surely. Let $G \equiv C'_{d_m-q,\infty} \xi B_{d_\theta-q,\infty}$. Since $\text{vec}(G) = (B'_{d_\theta-q,\infty} \otimes C'_{d_m-q,\infty}) \text{vec}(\xi)$ and the columns of $C_{d_m-q,\infty}$ and $B_{d_\theta-q,\infty}$ are orthonormal,

$$\text{var}(\text{vec}(G)) = (B'_{d_\theta-q,\infty} B_{d_\theta-q,\infty}) \otimes (C'_{d_m-q,\infty} C_{d_m-q,\infty}) = I_{(d_m-q)(d_\theta-q)}.$$

Thus G is a matrix of i.i.d. standard normal random variables. Moreover, G is independent of the other two terms in E , because ξ is independent of $\mathcal{Z}_D$. Since $\mathcal{Z}_D$ is Gaussian, $\text{vec}(E)$ is Gaussian with variance

$$\text{var}(\text{vec}(E)) = (B'_{d_\theta-q,\infty} \otimes C'_{d_m-q,\infty}) \text{var}(\text{vec}(\mathcal{Z}_D)) (B_{d_\theta-q,\infty} \otimes C_{d_m-q,\infty}) + \kappa^2 I_{(d_m-q)(d_\theta-q)},$$

which is PD because the first term is PSD and $\kappa > 0$. By Comment(i) of [Andrews and Guggenberger \(2017, Lemma 16.1\)](#), E has full column rank almost surely, which establishes the desired conclusion. $\square$

Proof of Lemma D.6. Let $\delta' \equiv \min\{\delta, 1\}$ and $M \equiv \sup_n \mathbb{E}|X_{ni}|^{1+\delta'}$, which is finite by Lyapunov's inequality. I apply the weak law for triangular arrays in [Durrett \(2019, Theorem 2.2.11\)](#) with $b_n = n$. By Markov's inequality, $n\mathbb{P}(|X_{ni}| > n) \leq Mn^{-\delta'} \rightarrow 0$. Since $X_{ni}^2 \leq |X_{ni}|^{1+\delta'} n^{1-\delta'}$ on $\{|X_{ni}| \leq n\}$, also $n^{-1}\mathbb{E}[X_{ni}^2 \mathbb{1}(|X_{ni}| \leq n)] \leq Mn^{-\delta'} \rightarrow 0$. The theorem then gives $n^{-1} \sum_{i=1}^n X_{ni} - \mathbb{E}[X_{ni} \mathbb{1}(|X_{ni}| \leq n)] \xrightarrow{P} 0$. Finally, $|\mathbb{E}[X_{ni} \mathbb{1}(|X_{ni}| > n)]| \leq Mn^{-\delta'} \rightarrow 0$, which proves the desired result. $\square$

Proof of Lemma D.7. This central limit theorem for a triangular array of random vectors is a special case of heterogeneous CLT in [Hansen \(2022, Theorem 9.4\)](#) by restricting observations in each row to be identically distributed. $\square$

E Power Analysis of MLC Tests

In this section, I analyze the power of MLC test under strong identification. To define the sequences of DGPs that are strongly identified, I impose the following additional restrictions on the space $\mathcal{P}$.

Definition E.1 (Parameter space under strong identification). For some $\delta, \zeta > 0$ and $\epsilon \in (0, 1/2)$, define the set $\mathcal{P}_s$ of pairs (θ, F) that satisfy the following conditions:

1. $(\theta, F) \in \mathcal{P}$. That is, (θ, F) satisfies the conditions imposed in Definition 2.1 with the given $\delta, \zeta > 0$, and $\epsilon \in (0, 1/2)$.
2. Equation (5) holds with functions $\{h_m(\cdot)\}_{m=1}^M$ satisfying Assumption 1.6.
3. There exists an index set $\mathcal{S} \subseteq \{0, 1, \dots, K\}$ with $|\mathcal{S}| = M + 1$ such that

$$\min_{j, k \in \mathcal{S}, j \neq k} |p(z_j) - p(z_k)| \geq \epsilon.$$

Specifically, the second condition brings back the unisolvent property on the specified functions, and the third condition assumes that there is a set of isolated propensity scores sufficient to point identify primitive parameters in the MTE model.

Lemma E.1. With $\mathcal{P}_s$ defined in Definition E.1, we have

$$\inf_{(\theta, F) \in \mathcal{P}_s} \tau_{\min}(A_F) > 0.$$

where A_F is defined below equation (11) for some distribution F .

The parameter space $\mathcal{P}_s$ guarantees that the smallest singular value of A_F , denoted as $\tau_{F, 2(M+1)}$, is uniformly bounded away from zero. Since A_F has full rank for each $(\theta, F) \in \mathcal{P}_s$, $\theta = (A_F' A_F)^{-1} A_F' \beta_F$ is uniquely determined by the data distribution F , as is $\lambda = c' \theta$. Therefore, $\mathcal{P}_s$ can be regarded as a collection of distribution F , and we denote $\theta_F \equiv (A_F' A_F)^{-1} A_F' \beta_F$ and $\lambda_F \equiv c' \theta_F$ for each $F \in \mathcal{P}_s$. To define the classical Wald test under strong identification, let $\hat{\theta}^{\text{eff}} \in \arg\min_{\theta \in \Theta} n(\hat{A}\theta - \hat{\beta})' \hat{\Omega}(\theta)^{-1} (\hat{A}\theta - \hat{\beta})$ be the efficient minimum distance estimator based on $\hat{A}\theta = \hat{\beta}$. The Wald test of $H_0: c' \theta = \lambda$ is

$$\hat{\phi}_{\text{Wald}}(\lambda) \equiv \mathbb{1}[\text{Wald}_n(\lambda) > q_{\chi_1^2}(1 - \alpha)], \quad \text{Wald}_n(\lambda) \equiv \frac{n(c' \hat{\theta}^{\text{eff}} - \lambda)^2}{c' (\hat{A}' \hat{\Omega}(\hat{\theta}^{\text{eff}})^{-1} \hat{A})^{-1} c}.$$

The next section establishes the consistency and local power property of the MLC test under strong identification, when compared to this Wald test.

E.1 Main results

Proposition E.1. Suppose $\{F_n : n \geq 1\} \subseteq \mathcal{P}_s$, where for some $F_0 \in \mathcal{P}_s$, assume

1. $p_{F_n}(z_\ell) \rightarrow p_{F_0}(z_\ell)$ for all $\ell = 0, 1, \dots, K$,

2. $\beta_{F_n} \rightarrow \beta_{F_0}$,
3. $\sigma_{F_n, d\ell}^2 \rightarrow \sigma_{F_0, d\ell}^2$ for all $d = 0, 1$ and $\ell = 0, 1, \dots, K$,
4. $q_{F_n}(z_\ell) \rightarrow q_{F_0}(z_\ell)$ for all $\ell = 0, 1, \dots, K$.

Consider a sequence of null values $\lambda_n^* = \lambda_{F_n} + bn^{-r}$ with $b \in \mathbb{R}$ and $b \neq 0$, then the following conclusion holds

(a) If $r \in [0, 1/2)$, then we have

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_n}(\hat{\phi}_{\text{MLC}}(\lambda_n^*) = 1) = 1. \quad (51)$$

(b) If $r = 1/2$, then we have

$$\lim_{a \searrow 0} \limsup_{n \rightarrow \infty} \mathbb{E}_{F_n} [\hat{\phi}_{\text{Wald}}(\lambda_n^*) - \hat{\phi}_{\text{MLC}}(\lambda_n^*)] = 0,$$

It is worth noting that the proposed MLC test is approximately as powerful as the asymptotic efficient Wald test under strong identification if we set the weight assigned to AR statistic sufficiently small. Andrews (2018, Theorem 3) establishes a related result by showing that a $(1 - \alpha - \gamma)$ LC confidence set is contained by a $(1 - \alpha)$ Wald confidence set with probability approaching one for all $\gamma > 0$ based on a specific choice of weight function $a(\gamma)$ such that the critical values of the two tests are equal: $q_{(1+a(\gamma))\chi_1^2 + a(\gamma)\chi_{2K+1}^2}(1 - \alpha - \gamma) = q_{\chi_1^2}(1 - \alpha)$. Nevertheless, it is less clear from his result that the MLC test would have similar local power as the Wald test at the *same significance level, while not targeting at a specific AR weight*. This result is now formally established in Proposition E.1(b).

It is not recommended to set $a = 0$ for the MLC test. On the one hand, the AR statistic helps direct the optimizer of the profiled MLC statistic to converge within a small neighborhood of the true parameter θ under strong identification. Within this neighborhood, the MRLM statistic is first-order equivalent to the Wald statistic. However, the MRLM statistic alone cannot detect deviations from the true parameter except in the direction of the target parameter $c'\theta$. On the other hand, assigning a larger weight to the AR statistic increases power under weak identification, as the MRLM statistic might be small for distant alternatives due to the near-singularity of $\hat{A}$ (Kleibergen, 2005). While Andrews (2016) discusses the optimal choice of a for full vector inference problems, the optimal choice of a for inference of parameter functions remains an open question for future research.

E.2 Lemmas for local power analysis

The following lemmas are used to establish Proposition E.1.

Lemma E.2. Under the same assumptions in Proposition E.1(b), the local power of Wald statistic equals:

$$\lim_{n \rightarrow \infty} \mathbb{E}_{F_n}[\hat{\phi}_{\text{Wald}}(\lambda_n^*)] = \mathbb{P}((\mathcal{Z} + \nu)'P_\nu(\mathcal{Z} + \nu) > q_{\chi_1^2}(1 - \alpha)),$$

where $\mathcal{Z} \sim \mathcal{N}(0_{2(K+1) \times 1}, I_{2(K+1)})$ and ν is defined in equation (60).

For each population distribution F , let

$$\begin{aligned}\Omega_F(\theta) &\equiv H(p_F, \theta) \Sigma_{p,F} H(p_F, \theta)' + \Sigma_{\beta,F} \\ \Gamma_{j,F}(\theta) &\equiv M_j(p_F) \Sigma_{p,F} H(p_F, \theta)',\end{aligned}$$

where $H(p, \theta)$ and $M_j(p)$ are defined in section 4.3.

Lemma E.3. Suppose $\{F_n\}_{n \geq 1} \subseteq \mathcal{P}_s$ is a sequence of distributions satisfying the conditions in Proposition E.1. The following results hold:

$$0 < \inf_{\theta \in \Theta} \tau_{\min}(\Omega_{F_0}(\theta)) \leq \sup_{\theta \in \Theta} \tau_{\max}(\Omega_{F_0}(\theta)) < \infty \quad (52)$$

$$\sup_{\theta \in \Theta} \|\hat{\Omega}(\theta) - \Omega_{F_0}(\theta)\| = o_p(1) \quad (53)$$

$$\sup_{\theta \in \Theta} \|\hat{\Gamma}_j(\theta) - \Gamma_{j,F_0}(\theta)\| = o_p(1). \quad (54)$$

E.3 Proofs in Appendix E

Proof of Lemma E.1. Since $A_F = \text{diag}\{A_{1F}, A_{0F}\}$, it suffices to bound the smallest singular value $\tau_{\min}(A_{dF})$ away from zero uniformly over $\mathcal{P}_s$ for each $d = 0, 1$. For $(p_0, \dots, p_M) \in (0, 1)^{M+1}$, let $\mathbf{A}(p_0, \dots, p_M)$ be the $(M+1) \times (M+1)$ matrix with (ℓ, m) -th element $\lambda_{dm}(p_\ell)$. Let

$$f(p_0, \dots, p_M) \equiv \tau_{\min}(\mathbf{A}(p_0, \dots, p_M)).$$

By Assumption 1.6, f is continuous, and by unisolvence $\det \mathbb{A} \neq 0$, so $f > 0$ on the compact set

$$\mathcal{E} \equiv \left\{ (p_0, \dots, p_M) : p_\ell \in [\epsilon, 1 - \epsilon] \text{ for all } \ell, \min_{j \neq k} |p_j - p_k| \geq \epsilon \right\}. \quad (55)$$

Hence f has a positive minimum over $\mathcal{E}$. For each $(\theta, F) \in \mathcal{P}_s$, Definitions 2.1 and E.1 give an index set $\mathcal{S}$ with $p_{\mathcal{S}} \equiv (p_F(z_\ell))_{\ell \in \mathcal{S}} \in \mathcal{E}$. The matrix $\mathbb{A}(p_{\mathcal{S}})$ is the submatrix of A_{dF} formed by the rows in $\mathcal{S}$. Therefore

$$\tau_{\min}(A_{dF}) = \min_{\|x\|=1} \|A_{dF}x\| \geq \min_{\|x\|=1} \|\mathbf{A}(p_{\mathcal{S}})x\| \geq \min_{p_{\mathcal{S}} \in \mathcal{E}} f(p_{\mathcal{S}}) > 0,$$

as desired. □

Proof of Proposition E.1. Throughout the proof, let

$$\hat{m}(\theta) \equiv \hat{A}\theta - \hat{\beta}, \quad \Delta_n(\theta) \equiv \sqrt{n}(\theta - \theta_{F_n}), \quad \bar{\Omega} \equiv \Omega_{F_0}(\theta_{F_0}), \quad \bar{V} \equiv (A'_{F_0} \bar{\Omega}^{-1} A_{F_0})^{-1},$$

and $\tau_{\min}(\cdot)$ and $\tau_{\max}(\cdot)$ for the smallest and largest singular values of a matrix. Under $\mathcal{P}_s$, $\theta_F = (A'_F A_F)^{-1} A'_F \beta_F$ is continuous in (p_F, β_F) , so $\theta_{F_n} \rightarrow \theta_{F_0}$. Since we assumed convergence of $(p_{F_n}, \beta_{F_n}, \sigma_{F_n}^2, q_{F_n})$ and $\theta_{F_n} \rightarrow \theta_{F_0}$, which are conditions 1–5 of Proposition A.2, similar delta-method argument at the beginning of the proof of Proposition A.2 gives

$$\sqrt{n}\hat{m}(\theta_{F_n}) \xrightarrow{d} \mathcal{Z}_m \sim \mathcal{N}(0_{2(K+1) \times 1}, \Omega_{F_0}(\theta_{F_0})), \quad (56)$$

where the variance follows from (40) with $(p_\infty, \theta_\infty) = (p_{F_0}, \theta_{F_0})$. Moreover, $\hat{A}$ is a continuous function of $\hat{p}$, and singular values are continuous functions of a matrix. Together with (56) and Lemmas D.1 and E.1, this gives

$$\hat{A} \xrightarrow{p} A_{F_0}, \quad \tau_{\min}(\hat{A}) \xrightarrow{p} \tau_{\min}(A_{F_0}) > 0, \quad \sqrt{n}\hat{m}(\theta_{F_n}) = O_p(1). \quad (57)$$

By (52) and (53), $\bar{\tau}_n \equiv \sup_{\theta \in \Theta} \tau_{\max}(\hat{\Omega}(\theta)) = O_p(1)$. Below I use two inequalities that hold for every $\theta \in \Theta$. First, since $\hat{m}(\theta) = \hat{m}(\theta_{F_n}) + \hat{A}(\theta - \theta_{F_n})$, and for each $x \in \mathbb{R}^{2(M+1)}$, $\|\hat{A}x\| \geq \tau_{\min}(\hat{A})\|x\|$, and $\|x\| \geq |c'x|/\|c\|$, we have

$$\begin{aligned} \|\sqrt{n}\hat{m}(\theta)\| &\geq \left(\|\hat{A}\Delta_n(\theta)\| - \|\sqrt{n}\hat{m}(\theta_{F_n})\| \right)_+ \\ &\geq \left(\tau_{\min}(\hat{A})\|\Delta_n(\theta)\| - \|\sqrt{n}\hat{m}(\theta_{F_n})\| \right)_+ \\ &\geq \left(\tau_{\min}(\hat{A}) \frac{\sqrt{n}|c'\theta - \lambda_{F_n}|}{\|c\|} - \|\sqrt{n}\hat{m}(\theta_{F_n})\| \right)_+, \end{aligned} \quad (58)$$

where $(x)_+ \equiv \max\{x, 0\}$, which is nondecreasing in x . Second, by the definition of $\text{AR}_n(\theta)$ in (15),

$$\text{AR}_n(\theta) \geq \bar{\tau}_n^{-1} \|\sqrt{n}\hat{m}(\theta)\|^2. \quad (59)$$

Part (a). If no $\theta \in \Theta$ satisfies $c'\theta = \lambda_n^*$, the test rejects by construction. Otherwise, every θ with $c'\theta = \lambda_n^*$ satisfies $\sqrt{n}|c'\theta - \lambda_{F_n}| = |b|n^{1/2-r}$, which diverges because $r < 1/2$. By (57)–(59) and $\text{MRLM}_n \geq 0$,

$$\inf_{c'\theta = \lambda_n^*} \text{MLC}_n(\theta) \geq a \bar{\tau}_n^{-1} \left(\tau_{\min}(\hat{A}) \frac{|b|n^{1/2-r}}{\|c\|} - \|\sqrt{n}\hat{m}(\theta_{F_n})\| \right)_+^2 \xrightarrow{p} \infty,$$

which establishes (51), as desired.

Part (b). Let

$$\hat{Z}_n \equiv \sqrt{n} \bar{\Omega}^{-1/2} \hat{m}(\theta_{F_n}) \xrightarrow{d} \mathcal{Z} \sim \mathcal{N}(0_{2(K+1) \times 1}, I_{2(K+1)}),$$

which follows from (56), and let

$$\nu \equiv \frac{\bar{\Omega}^{-1/2} A_{F_0} \bar{V} c}{c' \bar{V} c} \cdot b, \quad \|\nu\| = \frac{|b|}{\sqrt{c' \bar{V} c}}. \quad (60)$$

Step 1: a local expansion along data-dependent sequences. Let $\{\tilde{\theta}_n\}$ be any sequence in Θ , possibly depending on the data, such that

$$c'\tilde{\theta}_n = \lambda_n^* \quad \text{and} \quad \Delta_n(\tilde{\theta}_n) = O_p(1). \quad (61)$$

Next I show that

$$\begin{aligned} \text{MRLM}_n(\tilde{\theta}_n) &= (\hat{Z}_n + \nu)' P_\nu (\hat{Z}_n + \nu) + o_p(1), \\ \text{AR}_n(\tilde{\theta}_n) &= \left\| \hat{Z}_n + \bar{\Omega}^{-1/2} A_{F_0} \Delta_n(\tilde{\theta}_n) \right\|^2 + o_p(1). \end{aligned} \quad (62)$$

First, (61) gives $\tilde{\theta}_n - \theta_{F_n} = O_p(n^{-1/2})$, so $\tilde{\theta}_n \xrightarrow{p} \theta_{F_0}$. Moreover,

$$\sqrt{n}\hat{m}(\tilde{\theta}_n) = \sqrt{n}\hat{m}(\theta_{F_n}) + \hat{A}\Delta_n(\tilde{\theta}_n) = O_p(1). \quad (63)$$

Second, by the triangle inequality,

$$\|\hat{\Omega}(\tilde{\theta}_n) - \bar{\Omega}\| \leq \sup_{\theta \in \Theta} \|\hat{\Omega}(\theta) - \Omega_{F_0}(\theta)\| + \|\Omega_{F_0}(\tilde{\theta}_n) - \Omega_{F_0}(\theta_{F_0})\|.$$

The first term is $o_p(1)$ by (53). The second term is $o_p(1)$ by the continuity of $\Omega_{F_0}(\cdot)$ and the continuous mapping theorem. The same argument with (54) applies to $\hat{\Gamma}_j$, so

$$\hat{\Omega}(\tilde{\theta}_n) \xrightarrow{p} \bar{\Omega}, \quad \hat{\Gamma}_j(\tilde{\theta}_n) \xrightarrow{p} \Gamma_{j,F_0}(\theta_{F_0}) \quad \text{for } j = 1, \dots, 2(M+1). \quad (64)$$

Third, the j -th column of $\tilde{D}(\tilde{\theta}_n)$ equals $\hat{a}_j - \hat{\Gamma}_j(\tilde{\theta}_n)\hat{\Omega}(\tilde{\theta}_n)^{-1}\hat{m}(\tilde{\theta}_n) + \kappa n^{-1/2}\xi_j$. By (57), (63) and (64), the last two terms are $O_p(n^{-1/2})$, so $\tilde{D}(\tilde{\theta}_n) \xrightarrow{p} A_{F_0}$. Since A_{F_0} has full column rank, the continuous mapping theorem gives

$$\hat{Q}(\tilde{\theta}_n) \xrightarrow{p} \bar{\Omega}^{-1/2} A_{F_0} \bar{V} c = \frac{c' \bar{V} c}{b} \nu.$$

The limit is a nonzero multiple of ν , and the projection $Q \mapsto P_Q$ is continuous at every $Q \neq 0$. Hence

$$P_{\hat{Q}(\tilde{\theta}_n)} = P_\nu + o_p(1). \quad (65)$$

Fourth, (57), (61), and (63) give

$$\hat{\Omega}(\tilde{\theta}_n)^{-1/2} \sqrt{n}\hat{m}(\tilde{\theta}_n) = \hat{Z}_n + \bar{\Omega}^{-1/2} A_{F_0} \Delta_n(\tilde{\theta}_n) + o_p(1), \quad (66)$$

which yields the AR part of (62). For the MRLM part, $c' \Delta_n(\tilde{\theta}_n) = \sqrt{n}(\lambda_n^* - \lambda_{F_n}) = b$ by (61). Hence

$$\nu' \bar{\Omega}^{-1/2} A_{F_0} \Delta_n(\tilde{\theta}_n) = \frac{b c' \bar{V} A_{F_0}' \bar{\Omega}^{-1} A_{F_0} \Delta_n(\tilde{\theta}_n)}{c' \bar{V} c} = \frac{b^2}{c' \bar{V} c} = \|\nu\|^2,$$

so that

$$P_\nu \bar{\Omega}^{-1/2} A_{F_0} \Delta_n(\tilde{\theta}_n) = \nu. \quad (67)$$

The right-hand side of (66) is $O_p(1)$. Combining (65), (66) and (67) with the definition of MRLM $_n$ in (14),

$$\text{MRLM}_n(\tilde{\theta}_n) = \|P_\nu \hat{Z}_n + \nu\|^2 + o_p(1) = (\hat{Z}_n + \nu)' P_\nu (\hat{Z}_n + \nu) + o_p(1).$$

This proves (62).

Step 2: an upper bound. Let

$$\bar{\theta}_n \equiv \theta_{F_n} + \frac{b}{\sqrt{n}} \cdot \frac{\bar{V} c}{c' \bar{V} c}.$$

Then $c' \bar{\theta}_n = \lambda_n^*$, and $\bar{\theta}_n \in \Theta$ for large n because $\theta_{F_0} \in \text{int}(\Theta)$. The sequence $\bar{\theta}_n$ satisfies (61), and

$\bar{\Omega}^{-1/2} A_{F_0} \Delta_n(\bar{\theta}_n) = \nu$. By (62) and $\text{MLC}_n = \text{MRLM}_n + a \text{AR}_n$,

$$\text{MLC}_n(\bar{\theta}_n) \xrightarrow{d} (1+a)(\mathcal{Z} + \nu)' P_\nu(\mathcal{Z} + \nu) + a(\mathcal{Z} + \nu)' M_\nu(\mathcal{Z} + \nu). \quad (68)$$

Step 3: a lower bound. Let θ_n^* satisfy $c' \theta_n^* = \lambda_n^*$ and

$$\text{MLC}_n(\theta_n^*) \leq \inf_{c' \theta = \lambda_n^*} \text{MLC}_n(\theta) + n^{-1}.$$

By the definition of infimum, such sequence $\{\theta_n^*\}_{n \geq 1}$ always exists. Since $\text{MRLM}_n \geq 0$, (68) gives $a \text{AR}_n(\theta_n^*) \leq \text{MLC}_n(\bar{\theta}_n) + n^{-1} = O_p(1)$. By (59) and $\bar{\tau}_n = O_p(1)$,

$$\|\sqrt{n} \hat{m}(\theta_n^*)\|^2 \leq \bar{\tau}_n \text{AR}_n(\theta_n^*) = O_p(1). \quad (69)$$

Next, the second line in (58), evaluated at $\theta = \theta_n^*$, and $(x)_+ \geq x$ give

$$\tau_{\min}(\hat{A}) \|\Delta_n(\theta_n^*)\| \leq \|\sqrt{n} \hat{m}(\theta_n^*)\| + \|\sqrt{n} \hat{m}(\theta_{F_n})\|.$$

Both terms on the right are $O_p(1)$, the first by (69) and the second by (57). By (57), $\tau_{\min}(\hat{A}) \xrightarrow{p} \tau_{\min}(A_{F_0}) > 0$, so $\Delta_n(\theta_n^*) = O_p(1)$. Therefore, θ_n^* satisfies (61), and (62) yields

$$\text{MRLM}_n(\theta_n^*) \xrightarrow{d} (\mathcal{Z} + \nu)' P_\nu(\mathcal{Z} + \nu). \quad (70)$$

Step 4: conclusion. Let $q(a)$ denote the $(1 - \alpha)$ -quantile of $(1 + a)\chi_1^2 + a\chi_{2K+1}^2$. Since $\text{MRLM}_n(\theta_n^*) - n^{-1} \leq \inf_{c' \theta = \lambda_n^*} \text{MLC}_n(\theta) \leq \text{MLC}_n(\bar{\theta}_n)$,

$$\mathbb{P}_{F_n}(\text{MRLM}_n(\theta_n^*) > q(a) + n^{-1}) \leq \mathbb{E}_{F_n}[\hat{\phi}_{\text{MLC}}(\lambda_n^*)] \leq \mathbb{P}_{F_n}(\text{MLC}_n(\bar{\theta}_n) > q(a)).$$

The limit laws in (68) and (70) are continuous. Hence

$$L(a) \leq \liminf_{n \rightarrow \infty} \mathbb{E}_{F_n}[\hat{\phi}_{\text{MLC}}(\lambda_n^*)] \leq \limsup_{n \rightarrow \infty} \mathbb{E}_{F_n}[\hat{\phi}_{\text{MLC}}(\lambda_n^*)] \leq U(a),$$

where

$$L(a) \equiv \mathbb{P}((\mathcal{Z} + \nu)' P_\nu(\mathcal{Z} + \nu) > q(a)),$$

$$U(a) \equiv \mathbb{P}((1+a)(\mathcal{Z} + \nu)' P_\nu(\mathcal{Z} + \nu) + a(\mathcal{Z} + \nu)' M_\nu(\mathcal{Z} + \nu) > q(a)).$$

As $a \searrow 0$, $q(a) \rightarrow q_{\chi_1^2}(1 - \alpha)$, and both $L(a)$ and $U(a)$ converge to $\mathbb{P}((\mathcal{Z} + \nu)' P_\nu(\mathcal{Z} + \nu) > q_{\chi_1^2}(1 - \alpha))$. By Lemma E.2, this limit equals $\lim_{n \rightarrow \infty} \mathbb{E}_{F_n}[\hat{\phi}_{\text{Wald}}(\lambda_n^*)]$. $\square$

Proof of Lemma E.2. I use the notation of the proof of Proposition E.1. Under $\mathcal{P}_s$, the efficient estimator $\hat{\theta}^{\text{eff}}$ satisfies $\sqrt{n}(\hat{\theta}^{\text{eff}} - \theta_{F_n}) = -\bar{V} A'_{F_0} \bar{\Omega}^{-1/2} \hat{Z}_n + o_p(1)$, and $c'(\hat{A}' \hat{\Omega}(\hat{\theta}^{\text{eff}})^{-1} \hat{A})^{-1} c \xrightarrow{p} c' \bar{V} c$ (Newey and McFadden, 1994, Theorem 5.2). Since $\lambda_n^* = \lambda_{F_n} + b n^{-1/2}$ and $c' \bar{V} A'_{F_0} \bar{\Omega}^{-1/2} = b^{-1}(c' \bar{V} c) \nu'$,

$$\frac{\sqrt{n}(c' \hat{\theta}^{\text{eff}} - \lambda_n^*)}{\sqrt{c' \bar{V} c}} = -\frac{\sqrt{c' \bar{V} c}}{b} \nu' \hat{Z}_n - \frac{b}{\sqrt{c' \bar{V} c}} + o_p(1) = -\text{sign}(b) \left(\frac{\nu' \hat{Z}_n}{\|\nu\|} + \|\nu\| \right) + o_p(1).$$

Squaring and applying Slutsky's theorem gives

$$\text{Wald}_n(\lambda_n^*) \xrightarrow{d} \left(\frac{\nu' \mathcal{Z}}{\|\nu\|} + \|\nu\| \right)^2 = (\mathcal{Z} + \nu)' P_\nu (\mathcal{Z} + \nu).$$

This limit distribution is continuous, which gives the result. $\square$

Proof of Lemma E.3. Let $u \equiv (p, \Sigma_p, \Sigma_\beta)$ and define

$$\Omega(\theta; u) \equiv H(p, \theta) \Sigma_p H(p, \theta)' + \Sigma_\beta, \quad \Gamma_j(\theta; u) \equiv M_j(p) \Sigma_p H(p, \theta)'.$$

Let $\hat{u} \equiv (\hat{p}, \hat{\Sigma}_p, \hat{\Sigma}_\beta)$ and $u_0 \equiv (p_{F_0}, \Sigma_{p, F_0}, \Sigma_{\beta, F_0})$. By (19) and (21),

$$\hat{\Omega}(\theta) = \Omega(\theta; \hat{u}), \quad \Omega_{F_0}(\theta) = \Omega(\theta; u_0), \quad \hat{\Gamma}_j(\theta) = \Gamma_j(\theta; \hat{u}), \quad \Gamma_{j, F_0}(\theta) = \Gamma_j(\theta; u_0). \quad (71)$$

Both maps are continuous in (θ, u) , because the entries of $H(p, \theta)$ and $M_j(p)$ are continuous functions of (p, θ) built from $\lambda'_{dm}(\cdot)$, which is continuous on $(0, 1)$.

Proof of (52). The matrix $H(p_{F_0}, \theta) \Sigma_{p, F_0} H(p_{F_0}, \theta)'$ is PSD. By Weyl's inequality and Definition 2.1, for every $\theta \in \Theta$,

$$\tau_{\min}(\Omega_{F_0}(\theta)) \geq \tau_{\min}(\Sigma_{\beta, F_0}) = \min_{d, \ell} \frac{\sigma_{F_0, d\ell}^2}{q_{F_0}(d, z_\ell)} \geq \frac{\epsilon}{(1 - \epsilon)^2}. \quad (72)$$

By the subadditivity of $\tau_{\max}$ over PSD matrices and the submultiplicativity of the spectral norm,

$$\sup_{\theta \in \Theta} \tau_{\max}(\Omega_{F_0}(\theta)) \leq \tau_{\max}(\Sigma_{p, F_0}) \sup_{\theta \in \Theta} \|H(p_{F_0}, \theta)\|^2 + \tau_{\max}(\Sigma_{\beta, F_0}) < \infty. \quad (73)$$

The supremum is finite because $H(p_{F_0}, \cdot)$ is continuous on the compact set Θ . Together, (72) and (73) give (52).

Proof of (53) and (54). Let $\mathcal{N}$ be a compact neighborhood of u_0 on which $p \in [\epsilon/2, 1 - \epsilon/2]^{K+1}$. Such a neighborhood exists because $p_{F_0} \in [\epsilon, 1 - \epsilon]^{K+1}$ by Definition 2.1. The maps Ω and Γ_j are uniformly continuous on the compact set $\Theta \times \mathcal{N}$. Hence the modulus of continuity

$$w(\eta) \equiv \sup \left\{ \|\Omega(\theta; u) - \Omega(\theta; u')\| + \max_j \|\Gamma_j(\theta; u) - \Gamma_j(\theta; u')\| : \theta \in \Theta, u, u' \in \mathcal{N}, \|u - u'\| \leq \eta \right\}$$

satisfies $w(\eta) \rightarrow 0$ as $\eta \rightarrow 0$. Under the conditions of Proposition E.1, Lemma D.1 gives

$$\hat{u} \xrightarrow{p} u_0. \quad (74)$$

So the event $\{\hat{u} \in \mathcal{N}\}$ has probability approaching one. On this event, (71) gives

$$\sup_{\theta \in \Theta} \|\hat{\Omega}(\theta) - \Omega_{F_0}(\theta)\| + \max_j \sup_{\theta \in \Theta} \|\hat{\Gamma}_j(\theta) - \Gamma_{j, F_0}(\theta)\| \leq w(\|\hat{u} - u_0\|).$$

The right-hand side is $o_p(1)$ by (74) and $w(\eta) \rightarrow 0$. This proves (53) and (54). $\square$

F Inference with Estimated Weights

Sometimes the weight c needs to be estimated if researchers' interests focus on causal parameters such as ATT and LATE. In this case, the weight usually depends on the joint distribution of (D, Z) as in Table 1, therefore I make the following assumption:

Assumption 5. The weight $c = c(p, q)$ is a function of propensity scores $p = (p(z_0), \dots, p(z_K))'$ and marginal distribution of instrument $q = (q(z_0), \dots, q(z_K))'$, and this function is continuously differentiable.

Under this assumption, I develop inferential theory for causal parameter that depends on estimated weights. In section F.1, I generalize the conditional Wald test proposed in section 3 for linear MTE models. In section F.2 I extend the theory for MLC test proposed in section 4 under a general class of MTE models that include the polynomial structure.

F.1 Linear MTE models

When the weight depends on (p, q) , I replace c_μ and c_ρ in the sample moment function with the plug-in estimators $\hat{c}_\mu \equiv c_\mu(\hat{p}, \hat{q})$ and $\hat{c}_\rho \equiv c_\rho(\hat{p}, \hat{q})$. That is, throughout this subsection I *redefine*

$$\hat{g}_k(\lambda) \equiv [\hat{p}(z_k) - \hat{p}(z_0)] \lambda - \hat{c}_\mu \hat{\Delta}_\mu(z_0, z_k) - \hat{c}_\rho \hat{\Delta}_\rho(z_0, z_k), \quad k = 1, \dots, K, \quad (75)$$

and $\hat{g}(\lambda) \equiv (\hat{g}_1(\lambda), \dots, \hat{g}_K(\lambda))'$. This is a new definition, but I keep the notation of Section 3. The same convention applies to $\hat{S}(\lambda)$, $\hat{h}(\lambda)$, π_s , $W_n(\lambda)$ and $W_n^*(\lambda)$ below.

Because $\hat{c}$ is estimated, its sampling error enters the asymptotic distribution of $\hat{g}(\lambda)$. Let $\Delta_\mu \equiv (\Delta_\mu(z_0, z_1), \dots, \Delta_\mu(z_0, z_K))'$ and $\Delta_\rho \equiv (\Delta_\rho(z_0, z_1), \dots, \Delta_\rho(z_0, z_K))'$, and let $\hat{\Delta}_\mu$ and $\hat{\Delta}_\rho$ be their sample analogs. Let $\partial_{p'} c_\mu$ and $\partial_{q'} c_\mu$ denote the $1 \times (K+1)$ gradients of $c_\mu(p, q)$, and similarly for c_ρ . Define the $K \times (K+1)$ matrices

$$J_p(\lambda) \equiv \partial_p g(\lambda) - \Delta_\mu \partial_{p'} c_\mu - \Delta_\rho \partial_{p'} c_\rho, \quad J_q \equiv -\Delta_\mu \partial_{q'} c_\mu - \Delta_\rho \partial_{q'} c_\rho, \quad (76)$$

which are the derivatives of $g(\lambda)$ with respect to p and q when $c = c(p, q)$. By the delta method,

$$\sqrt{n}(\hat{g}(\lambda) - g(\lambda)) = J_p(\lambda) \hat{Z}_p + J_q \hat{Z}_q + \partial_{\beta_1} g(\lambda) \hat{Z}_{\beta_1} + \partial_{\beta_0} g(\lambda) \hat{Z}_{\beta_0} + o_p(1).$$

Since $(\mathcal{Z}_p, \mathcal{Z}_{\beta_1}, \mathcal{Z}_{\beta_0}, \mathcal{Z}_q)$ has a block-diagonal covariance matrix by Lemma D.1, the asymptotic variance of $\sqrt{n}\hat{g}(\lambda)$ is estimated by

$$\hat{S}(\lambda) \equiv \hat{J}_p(\lambda) \hat{\Sigma}_p \hat{J}_p(\lambda)' + \hat{J}_q \hat{\Sigma}_q \hat{J}_q' + \partial_{\beta_1} \hat{g}(\lambda) \hat{\Sigma}_{\beta_1} \partial_{\beta_1}' \hat{g}(\lambda) + \partial_{\beta_0} \hat{g}(\lambda) \hat{\Sigma}_{\beta_0} \partial_{\beta_0}' \hat{g}(\lambda), \quad (77)$$

where $\hat{J}_p(\lambda)$, $\hat{J}_q$ and $\partial_x \hat{g}(\lambda)$ replace (p, q, β, c) with $(\hat{p}, \hat{q}, \hat{\beta}, \hat{c})$, and replace the gradients of c_μ and c_ρ with their values at $(\hat{p}, \hat{q})$.

The vector $\sqrt{n}\hat{\pi}$ depends on the data only through $\hat{p}$. Its asymptotic covariance with $\sqrt{n}\hat{g}(\lambda)$ is therefore $[\partial_p \pi] \Sigma_p J_p(\lambda)'$. Accordingly, I redefine the conditioning statistic and its simulated

counterpart as

$$\begin{aligned}\hat{h}(\lambda) &\equiv \sqrt{n}\hat{\pi} - [\partial_p \pi] \hat{\Sigma}_p \hat{J}_p(\lambda)' \hat{S}(\lambda)^{-1} \sqrt{n}\hat{g}(\lambda), \\ \pi_s &\equiv \hat{h}(\lambda) + [\partial_p \pi] \hat{\Sigma}_p \hat{J}_p(\lambda)' \hat{S}(\lambda)^{-1/2} \eta^*,\end{aligned}\tag{78}$$

where $\eta^* \sim \mathcal{N}(0_{K \times 1}, I_K)$ is drawn independently of the data. With these definitions, $W_n(\lambda)$, $W_n^*(\lambda)$ and the critical value $\hat{q}_{W^*}(1 - \alpha)$ are constructed exactly as in Section 3. When c does not depend on (p, q) , $\hat{J}_p(\lambda) = \partial_p \hat{g}(\lambda)$ and $\hat{J}_q = 0$, so all definitions reduce to those of Section 3.

Corollary F.1. Let Assumptions 2, 3 and 5 hold, and suppose that $\inf_{(\theta, F) \in \mathcal{P}} \|c(p_F, q_F)\| > 0$. With $\hat{g}(\lambda)$, $\hat{S}(\lambda)$, $\hat{h}(\lambda)$ and π_s redefined in (75)–(78), the conclusion of Theorem 3.1 holds.

Remark F.1. The AR test in Proposition 3.1 remains asymptotically valid with $\hat{g}_k(\lambda)$ redefined in (75) and $\hat{s}_k^2(\lambda)$ redefined as the k -th diagonal element of $\hat{S}(\lambda)$ in (77). $\square$

F.2 General MTE models

In this section, I maintain Assumption 5 and consider modified linear combination test for the treatment effects parameters $\lambda = c(p, q)' \theta$. Naively plugging in the estimator $\hat{c} = c(\hat{p}, \hat{q})$ into the test statistic can introduce estimation errors in the constraint $\hat{c}' \theta = \lambda$, potentially leading to asymptotic bias in the limiting distribution of the test statistic. To address this, I modify the test statistic to account for the effects of these estimation errors.

One possible solution involves reparameterizing the model so that the target parameter $\hat{c}' \theta$ becomes an element of the reparameterized model parameter, e.g., see D. Andrews (2017). However, given the broad class of causal parameters of interest, finding a universal reparameterization rule that works for all these parameters is challenging, and such reparameterization would greatly complicate the asymptotic analysis.

As an alternative, I impose a normalization on the weights of the treatment effect parameters. This assumption is broadly applicable to various causal effects, including ATE, ATT, LATE, and the normalized counterfactual policy effects $\bar{\alpha}(\epsilon)$ considered in this paper.

Assumption 6. The first and $(M + 2)$ -th elements of $c(p, q)$ are 1 and -1 , respectively, and the weight is symmetric: $c = (c'_1, c'_0)'$ with $c_1 = -c_0$.

Under Assumption 6, I construct sequences that satisfy the estimated constraint $\hat{c}' \theta = \lambda_n$, where $\lambda_n = c(p_{F_n}, q_{F_n})' \theta_{F_n}$ is the null value. Let $\hat{e}_n \equiv [\hat{c} - c(p_{F_n}, q_{F_n})]' \theta_{F_n}$, and define

$$\begin{aligned}\delta_{n,1} &\equiv (-\hat{e}_n, \mathbf{0}_{1 \times (2M+1)})', \\ \delta_{n,0} &\equiv (\mathbf{0}_{1 \times (M+1)}, \hat{e}_n, \mathbf{0}_{1 \times M})' .\end{aligned}$$

Since the first element of $\hat{c}$ equals 1 and its $(M + 2)$ -th element equals -1 , both $\theta_{F_n} + \delta_{n,1}$ and $\theta_{F_n} + \delta_{n,0}$ satisfy $\hat{c}' \theta = \lambda_n$. For any $\omega \in [0, 1]$, define

$$\theta_{F_n}^{(\omega)} \equiv \omega(\theta_{F_n} + \delta_{n,1}) + (1 - \omega)(\theta_{F_n} + \delta_{n,0}),\tag{79}$$

which also satisfies the estimated constraint. The sequence $\{\theta_{F_n}^{(\omega)}\}$ plays the role that $\{\theta_{F_n}\}$ plays in the proof of Theorem 4.1.

Proposition F.1. Let Assumptions 2, 5 and 6 hold, and we further restrict $\theta \in \Theta_0$ in condition 1 of Definition 2.1, where Θ_0 is a compact subset of $\text{int}(\Theta)$. Fix $\omega \in [0, 1]$ and let $\varpi \equiv (-\omega \mathbf{1}_{1 \times (K+1)}, (1 - \omega) \mathbf{1}_{1 \times (K+1)})'$, where $\mathbf{1}_{1 \times (K+1)}$ is a $1 \times (K+1)$ vector of ones. In the MLC test of Section 4.3, replace c with $\hat{c} = c(\hat{p}, \hat{q})$ in the constraint set, and replace $\hat{\Omega}(\theta)$ and $\hat{\Gamma}_j(\theta)$ with

$$\begin{aligned}\hat{\Omega}(\theta; \omega) &\equiv (H(\hat{p}, \theta) + \varpi \theta' [\partial_p \hat{c}]) \hat{\Sigma}_p (H(\hat{p}, \theta) + \varpi \theta' [\partial_p \hat{c}])' \\ &\quad + (\varpi \theta' [\partial_q \hat{c}]) \hat{\Sigma}_q (\varpi \theta' [\partial_q \hat{c}])' + \hat{\Sigma}_\beta, \\ \hat{\Gamma}_j(\theta; \omega) &\equiv M_j(\hat{p}) \hat{\Sigma}_p (H(\hat{p}, \theta) + \varpi \theta' [\partial_p \hat{c}])',\end{aligned}$$

where $\partial_p \hat{c}$ and $\partial_q \hat{c}$ are the $2(M+1) \times (K+1)$ Jacobians of $c(p, q)$ evaluated at $(\hat{p}, \hat{q})$. In addition, replace c in the construction of $\hat{Q}(\theta)$ with the orthogonalized weight

$$\begin{aligned}\hat{c}_\perp(\theta) &\equiv \hat{c} - \hat{\Gamma}_c(\theta; \omega) \hat{\Omega}(\theta; \omega)^{-1} \hat{m}(\theta), \\ \hat{\Gamma}_c(\theta; \omega) &\equiv [\partial_p \hat{c}] \hat{\Sigma}_p (H(\hat{p}, \theta) + \varpi \theta' [\partial_p \hat{c}])' + [\partial_q \hat{c}] \hat{\Sigma}_q (\varpi \theta' [\partial_q \hat{c}])',\end{aligned}$$

where $\hat{\Gamma}_c(\theta; \omega)$ estimates the asymptotic covariance between $\sqrt{n}(\hat{c} - c)$ and $\sqrt{n} \hat{m}(\theta)$. Then the resulting test $\hat{\phi}_{\text{MLC}}(\lambda)$ satisfies

$$\limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{E}_F[\hat{\phi}_{\text{MLC}}(\lambda)] \leq \alpha,$$

where $\mathcal{P}_0 = \{(\lambda, F) : \lambda = c(p_F, q_F)' \theta, (\theta, F) \in \mathcal{P}\}$.

F.3 Proofs in Appendix F

Proof of Corollary F.1. It suffices to verify Proposition A.1(b) for the redefined statistics; the rest is as in the proof of Theorem 3.1. By Assumption 5, Lemma D.1 and the uniform delta method, $\sqrt{n} \hat{g}(\lambda_n) \xrightarrow{d} \mathcal{N}(0_{K \times 1}, S_\infty)$ and $\hat{S}(\lambda_n) \xrightarrow{p} S_\infty$, where S_∞ is the probability limit of (77). Its last two terms are bounded below as in the proof of Lemma D.2, a bound that holds for any weight, so $\tau_{\min}(S_\infty) \geq \frac{2}{9} \underline{\delta}(\epsilon) \inf_{(\theta, F) \in \mathcal{P}} \|c(p_F, q_F)\|^2 > 0$. Since $\hat{\pi}$ depends only on $\hat{p}$ and $\mathcal{Z}_p$ is independent of $(\mathcal{Z}_{\beta_1}, \mathcal{Z}_{\beta_0}, \mathcal{Z}_q)$, the asymptotic covariance of $\sqrt{n} \hat{\pi}$ and $\sqrt{n} \hat{g}(\lambda_n)$ is $[\partial_p \pi] \Sigma_{p, \infty} J'_{p, \infty}$, where $J_{p, \infty}$ is the limit of $\hat{J}_p(\lambda_n)$. These are exactly the properties used in the proofs of Proposition A.1(b) and Lemma D.3, with $\partial_p \hat{g}(\lambda_n)$ and $\partial_p g(\lambda_\infty)$ replaced by $\hat{J}_p(\lambda_n)$ and $J_{p, \infty}$.

Repeating the arguments in those proofs with these replacements gives $\mathbb{P}_{F_n}(W_n(\lambda_n) > \hat{q}_{W^*}(1 - \alpha)) \rightarrow \alpha$, which verifies Proposition A.1(b) for the redefined statistics. The corollary then follows from Corollary 2.1(c) of Andrews et al. (2020), as in the proof of Theorem 3.1. $\square$

Proof of Proposition F.1. As in the proofs of Theorem 4.1 and Proposition A.2, it suffices to show that, along every sequence satisfying conditions 1–7 of Proposition A.2 with $p_n = n$,

$$\limsup_{n \rightarrow \infty} \mathbb{P}_{F_n} \left(\inf_{\theta \in \Theta: \hat{c}' \theta = \lambda_n} \text{MLC}_n(\theta) > q_\alpha \right) \leq \alpha, \quad q_\alpha \equiv q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2}(1 - \alpha). \quad (80)$$

Step 1: the moment condition at $\theta_{F_n}^{(\omega)}$. By Assumption 5, Lemma D.1(a) and the uniform delta method,

$$\sqrt{n}(\hat{c} - c_n) = [\partial_p c] \hat{Z}_p + [\partial_q c] \hat{Z}_q + o_p(1) \xrightarrow{d} \mathcal{Z}_c \equiv [\partial_p c] \mathcal{Z}_p + [\partial_q c] \mathcal{Z}_q, \quad (81)$$

where $c_n \equiv c(p_{F_n}, q_{F_n})$ and the Jacobians are evaluated at (p_∞, q_∞) . Hence $\theta_{F_n}^{(\omega)} - \theta_{F_n} = O_p(n^{-1/2})$, and $\mathbb{P}_{F_n}(\theta_{F_n}^{(\omega)} \notin \Theta) \rightarrow 0$ because θ_{F_n} lies in Θ_0 , a compact subset of $\text{int}(\Theta)$. Moreover, the first columns of A_1 and A_0 are vectors of ones by (5), so $\hat{A}(\theta_{F_n}^{(\omega)} - \theta_{F_n}) = \varpi \hat{c}_n$. The delta-method arguments in the proof of Proposition A.2 and (81) then give

$$\sqrt{n} \hat{m}(\theta_{F_n}^{(\omega)}) = \sqrt{n} \hat{m}(\theta_{F_n}) + \varpi \sqrt{n} \hat{c}_n \xrightarrow{d} \mathcal{Z}_m^{(\omega)} \equiv (H(p_\infty, \theta_\infty) + \varpi \theta'_\infty [\partial_p c]) \mathcal{Z}_p + \varpi \theta'_\infty [\partial_q c] \mathcal{Z}_q - \mathcal{Z}_\beta,$$

whose variance is PD because $\Sigma_{\beta, \infty}$ is PD. By Lemma D.1(b) and continuity, $\hat{\Omega}$, $\hat{\Gamma}_j$ and $\hat{\Gamma}_c$ evaluated at $\theta_{F_n}^{(\omega)}$ are consistent for $\text{var}(\mathcal{Z}_m^{(\omega)})$, $\text{cov}(M_j(p_\infty) \mathcal{Z}_p, \mathcal{Z}_m^{(\omega)})$ and $\text{cov}(\mathcal{Z}_c, \mathcal{Z}_m^{(\omega)})$, respectively.

Step 2: orthogonalization. As in the proof of Proposition A.2, Step 1 implies that (42) holds with $(\theta_{F_n}, \mathcal{Z}_m, \mathcal{Z}_D)$ replaced by $(\theta_{F_n}^{(\omega)}, \mathcal{Z}_m^{(\omega)}, \mathcal{Z}_D^{(\omega)})$ and that, jointly,

$$\sqrt{n}(\hat{c}_\perp(\theta_{F_n}^{(\omega)}) - c_n) \xrightarrow{d} \mathcal{Z}_{c_\perp} \equiv \mathcal{Z}_c - \text{cov}(\mathcal{Z}_c, \mathcal{Z}_m^{(\omega)}) \text{var}(\mathcal{Z}_m^{(\omega)})^{-1} \mathcal{Z}_m^{(\omega)}. \quad (82)$$

Since $\mathcal{Z}_D^{(\omega)}$, $\mathcal{Z}_{c_\perp}$ and $\mathcal{Z}_m^{(\omega)}$ are jointly normal and the first two are uncorrelated with the third, $(\mathcal{Z}_D^{(\omega)}, \mathcal{Z}_{c_\perp})$ is independent of $\mathcal{Z}_m^{(\omega)}$.

Step 3: the direction of $\hat{Q}$. In the proof of Proposition A.2, $\hat{Q}$ is a positive multiple of $\hat{Q}_{\text{norm}}$, which depends on c only through ι_n , and the limit ι_∞ of ι_n is used only through two properties: it is nonzero, and it is independent of the limit of the moment condition. Here ι_∞ is random, so I verify both. With $r_n \equiv 1/\|\Psi_{F_n} B'_{F_n} c_n\|$, the same substitution as in that proof gives $P_{\hat{Q}(\theta_{F_n}^{(\omega)})} = P_{\hat{Q}_{\text{norm}}}$ with

$$\iota_n \equiv r_n \Psi_{F_n} B'_{F_n} \hat{c}_\perp(\theta_{F_n}^{(\omega)}) = r_n \Psi_{F_n} B'_{F_n} c_n + r_n n^{-1/2} \Psi_{F_n} B'_{F_n} \sqrt{n}(\hat{c}_\perp(\theta_{F_n}^{(\omega)}) - c_n).$$

The first term is the deterministic unit vector that appears in the proof of Proposition A.2. The diagonal elements of Ψ_{F_n} are at least $\min\{1, (\sqrt{n} \tau_{1, F_n})^{-1}\}$, $\tau_{1, F}$ is bounded over $\mathcal{P}$, and $\|c_n\| \geq 1$ by Assumption 6. Hence $\sqrt{n} \|\Psi_{F_n} B'_{F_n} c_n\|$ is bounded away from zero, and $r_n n^{-1/2}$ is bounded. By the subsequence argument at the end of the proof of Proposition A.2, I can take $r_n \Psi_{F_n} B'_{F_n} c_n$ and $r_n n^{-1/2}$ to converge, to a unit vector $\bar{\iota}$ and some $\bar{r} \in [0, \infty)$. Since $\Psi_{F_n} \rightarrow \text{diag}\{0_{q \times q}, I_{2(M+1)-q}\}$ and $B_{F_n} \rightarrow B_\infty$, (82) gives

$$\iota_n \xrightarrow{d} \iota_\infty \equiv \bar{\iota} + \bar{r} \text{diag}\{0_{q \times q}, I_{2(M+1)-q}\} B'_\infty \mathcal{Z}_{c_\perp},$$

which is independent of $\mathcal{Z}_m^{(\omega)}$ by Step 2. The first q elements of ι_∞ are those of $\bar{\iota}$; if they are all zero, the remaining elements form a normal vector with a nonzero mean. Either way, $\iota_\infty \neq 0$ almost surely.

Step 4: conclusion. By Steps 2 and 3, the limit $\mathcal{Q}$ of $\hat{Q}_{\text{norm}}$ is nonzero almost surely and is a function of $(\mathcal{Z}_D^{(\omega)}, \mathcal{Z}_{c_\perp}, \xi)$, so it is independent of $\mathcal{Z}_m^{(\omega)}$. The remaining arguments in the proof of Proposition A.2 then apply without change and give $\text{MLC}_n(\theta_{F_n}^{(\omega)}) \xrightarrow{d} (1+a)\chi_1^2 + a\chi_{2K+1}^2$. Since $\theta_{F_n}^{(\omega)}$ is feasible for the infimum in (80) whenever it lies in Θ , the probability in (80) is at most

$\mathbb{P}_{F_n}(\text{MLC}_n(\theta_{F_n}^{(\omega)}) > q_\alpha) + \mathbb{P}_{F_n}(\theta_{F_n}^{(\omega)} \notin \Theta)$, which converges to α by Step 1. The proposition then follows from Corollary 2.1(a) of [Andrews et al. \(2020\)](#), as in the proof of Theorem 4.1. $\square$

G Power Comparison in Linear MTE Models

In this appendix, I compare the power of the proposed two approaches, namely, the conditional Wald test proposed in section 3 and modified linear combination test proposed in section 4, and contrast them with the conventional Wald test based on moment conditions (11) in a linear MTE model. The linear MTE model is generated by the DGP below: For each $d = 0, 1$,

$$\begin{aligned} Y_d &= \mu_d + V_d \\ D &= \mathbb{1}[U \leq p(Z)] \\ V_d &= \rho_d \left(U - \frac{1}{2} \right) + e_d, \end{aligned}$$

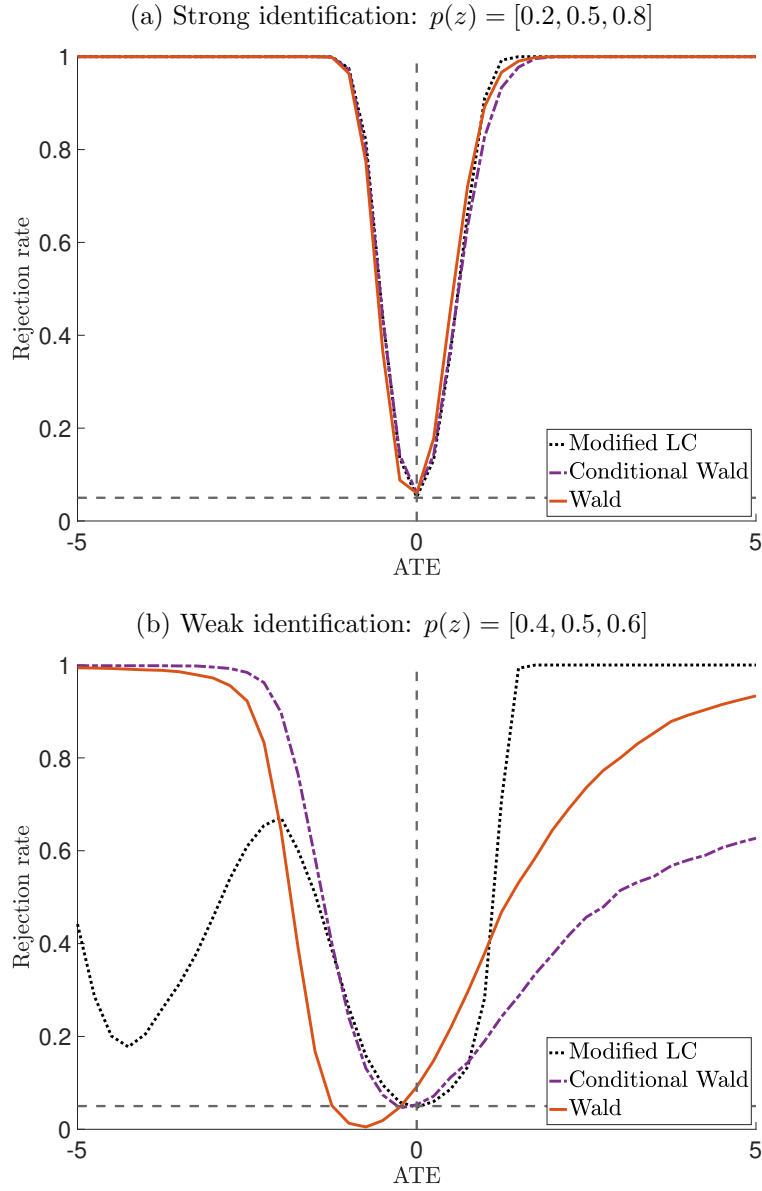
where U is uniformly distributed over a unit interval $[0, 1]$, Z has the same distribution as the one used in section 5, i.e., uniformly distributed over $\{z_0, z_1, z_2\}$ and independent of (U, e_1, e_0) , and (e_1, e_0) follows the joint normal distribution with zero mean and covariance matrix $\Sigma_e = 0.5 \cdot I_{2 \times 2}$. Since instrument has ternary support, the linear MTE model is over-identified and is weakly identified if all propensity scores converge to a single point. I consider the following specification of the parameters in the Monte-Carlo simulation:

- Mean potential outcomes: $\mu_1 = \mu_0 = 0$,
- Slope of MTR functions: $\rho_0 = \rho_1 = 5$
- Propensity scores:
 - (i) Strong identification: $p^s(z) = [0.2, 0.5, 0.8]$
 - (ii) Weak identification: $p^w(z) = [0.4, 0.5, 0.6]$

The data is generated with $n = 500$ units to evaluate the power of the conditional Wald test, modified linear combination test, and the classical Wald test. I repeat the Monte-Carlo experiments 2,000 times to compute the average rejection rates for testing ATE $H_0 : \mu_1 - \mu_0 = \delta_\mu$ for $\delta_\mu \in [-5, 5]$.

Figure 5 plots the power curves under strong and weak identification. When the propensity score varies strongly, all tests control size at the true ATE and have similar power against fixed alternatives. With a small AR weight ($a = 0.05$), the MLC test does not sacrifice much power under strong identification, which matches my local power analysis. Under weak identification, the MLC and conditional Wald tests control size at the true ATE of zero, whereas the non-robust Wald test over-rejects. The MLC power curve is non-monotonic, like that of the robust LM test of [Kleibergen \(2005\)](#) for full-vector inference, so the test can lack power at some distant alternatives. The conditional Wald test has power against both positive and negative values of the ATE, so I recommend it for linear MTE models.

Figure 5: Power Curves of the Conditional Wald, MLC, and Wald Tests



Note: Testing ATE at values on $[-5, 5]$ with the true effects fixed at zero. The significance level is 5%. The sample size equals 500. The average rejection rates are computed with 2,000 independent Monte-Carlo simulations.

H Testing IV Strength

Several studies have developed tests for IV strength (Stock and Yogo, 2005; Montiel Olea and Pflueger, 2013; Lewis and Mertens, 2022). These tests typically focus on the bias of IV estimators relative to OLS estimators and the size distortion of conventional Wald and t-tests under their null hypotheses. This problem is well-studied in linear IV regressions with homoskedastic error structure (Andrews et al., 2019, Section 4). However, little is known about nonlinear models with a focus on inference for functions of parameters. In this section, I evaluate the identification strength by using the first-stage diagnostic proposed by I. Andrews (2018).

Following I. Andrews (2018), we regard the target parameter as being weakly identified if the Wald confidence set fails to include a locally equivalent robust confidence set with a strictly smaller coverage level. Let $\gamma \in [0, 1 - \alpha]$. We choose the weight $a = a(\gamma)$ in the MLC test such that $a(\gamma)$ solves

$$q_{(1+a(\gamma))\chi_1^2 + a(\gamma)\chi_{2K+1}^2}(1 - \alpha - \gamma) = q_{\chi_1^2}(1 - \alpha).$$

With the choice of weight $a(\gamma)$ and the critical value $q_{\chi_1^2}(1 - \alpha)$, the following set $\mathcal{C}_P(\gamma)$ achieves asymptotic coverage rate $1 - \alpha - \gamma$ under weak identification:

$$\mathcal{C}_P(\gamma) = \left\{ \lambda \in \mathbb{R} : \inf_{c' \theta = \lambda} \text{MRLM}_n(\theta) + a(\gamma) \cdot \text{AR}_n(\theta) < q_{\chi_1^2}(1 - \alpha) \right\}.$$

As shown by Andrews (2018, Theorem 3), this set is contained by a non-robust $(1 - \alpha)$ Wald confidence set $\mathcal{C}_W$ with probability approaching one under *strong identification*. In other words, the failure of this containment relationship suggests evidence of weak identification. This leads to the following test of weak identification:

$$\phi_{ICS}(\gamma) = \mathbb{1}[\mathcal{C}_P(\gamma) \not\subseteq \mathcal{C}_W].$$

The tuning parameter γ can be regarded as the maximum amount of size distortion on Wald confidence set that researchers can tolerate. If $\phi_{ICS}(\gamma) = 1$, then researchers are willing to use the robust confidence set with a smaller coverage level $(1 - \alpha - \gamma)$ since it is always valid while the classical $(1 - \alpha)$ Wald confidence set is unreliable if the test rejects. However, this type of test may not have sufficient power to detect weak identification, which implies the model might still be weakly identified even if we did not reject the test.

For the example of additive MP RTE conditional on courts, I report the Andrews' pretesting result in Table 5 by specifying $\gamma = 10\%$ and $\alpha = 5\%$ (i.e., the researchers can tolerate at most 10% size distortion). The result indicates that cubic and quartic MTE models lead to potential concerns of weak identification conditional on most courts, while linear MTE models are strongly identified. Quadratic MTE models are strongly identified for several courts, but not all of them. This result reveals the fact that the identification strength depends on both the variation of instruments and the flexibility of the model specified by researchers, while this information is not reflected by the conventional rule-of-thumb applied to F -statistics, since the latter rule only works under homoskedastic linear IV models with a scalar endogenous coefficient (Andrews et al., 2019).

Table 5: Testing Weak Identification of Additive MPRTE at $\gamma = 10\%$ and $\alpha = 5\%$

Court	MTE Poly.	Valid Test	Wald (95%)	MLC (85%)	F-stat	RoT
SBO	Linear MTE	Possibly Strong	[-0.17, -0.02]	[-0.14, -0.06]	77.90	Strong
SBO	Quadratic MTE	Weak	[-0.17, 0.18]	[-0.18, 0.20]		
SBO	Cubic MTE	Weak	[-0.40, 0.45]	[-0.67, 0.55]		
SBO	Quartic MTE	Weak	[-0.64, 0.46]	[-0.76, 0.75]		
EBOS	Linear MTE	Possibly Strong	[-0.24, -0.10]	[-0.23, -0.12]	50.01	Strong
EBOS	Quadratic MTE	Possibly Strong	[-0.22, -0.06]	[-0.20, -0.08]		
EBOS	Cubic MTE	Weak	[-0.34, -0.04]	[-0.45, 0.03]		
EBOS	Quartic MTE	Weak	[-0.36, 0.04]	[-0.63, 0.37]		
WROX	Linear MTE	Possibly Strong	[-0.32, -0.14]	[-0.30, -0.16]	28.15	Strong
WROX	Quadratic MTE	Possibly Strong	[-0.33, -0.12]	[-0.31, -0.13]		
WROX	Cubic MTE	Weak	[-0.33, -0.13]	[-0.40, 0.05]		
WROX	Quartic MTE	Weak	[-0.33, -0.12]	[-0.89, 0.44]		
BMC	Linear MTE	Possibly Strong	[-0.24, 0.00]	[-0.16, -0.07]	31.47	Strong
BMC	Quadratic MTE	Weak	[-0.28, -0.02]	[-0.21, 0.04]		
BMC	Cubic MTE	Weak	[-0.28, 0.00]	[-0.35, 0.12]		
BMC	Quartic MTE	Weak	[-0.37, -0.07]	[-0.30, 0.18]		
ROX	Linear MTE	Possibly Strong	[-0.03, 0.23]	[0.06, 0.17]	34.68	Strong
ROX	Quadratic MTE	Possibly Strong	[-0.02, 0.28]	[0.08, 0.23]		
ROX	Cubic MTE	Weak	[-0.02, 0.34]	[0.05, 0.38]		
ROX	Quartic MTE	Possibly Strong	[-0.10, 0.31]	[0.13, 0.24]		
DOR	Linear MTE	Possibly Strong	[-0.42, -0.18]	[-0.37, -0.25]	43.60	Strong
DOR	Quadratic MTE	Possibly Strong	[-0.60, -0.26]	[-0.56, -0.36]		
DOR	Cubic MTE	Weak	[-0.60, -0.26]	[-0.56, -0.16]		
DOR	Quartic MTE	Weak	[-0.67, -0.29]	[-0.79, -0.08]		

Note: Valid Test is based on [I.Andrews \(2018\)](#) first-stage pretesting procedure under 10% size distortion. F -statistics are computed by regressing treatment on a set of saturated instruments, and the Rule of Thumb (RoT) suggests strong identification if the F -statistics exceed 10.

References

- ANDREWS, D. W. (2017): “Identification-robust subvector inference,” *Available at SSRN 3032675*.
- ANDREWS, D. W., X. CHENG, AND P. GUGGENBERGER (2020): “Generic results for establishing the asymptotic size of confidence sets and tests,” *Journal of Econometrics*, 218, 496–531.
- ANDREWS, D. W. AND P. GUGGENBERGER (2010): “Asymptotic size and a problem with subsampling and with the m out of n bootstrap,” *Econometric Theory*, 26, 426–468.
- (2017): “Asymptotic size of Kleibergen’s LM and conditional LR tests for moment condition models,” *Econometric Theory*, 33, 1046–1080.
- ANDREWS, I. (2016): “Conditional linear combination tests for weakly identified models,” *Econometrica*, 84, 2155–2182.
- (2018): “Valid two-step identification-robust confidence sets for GMM,” *Review of Economics and Statistics*, 100, 337–348.
- ANDREWS, I., J. H. STOCK, AND L. SUN (2019): “Weak instruments in instrumental variables regression: Theory and practice,” *Annual Review of Economics*, 11, 727–753.
- DURRETT, R. (2019): *Probability: theory and examples*, vol. 49, Cambridge university press.
- HANSEN, B. (2022): *Probability and Statistics for Economists*, Princeton University Press.
- KLEIBERGEN, F. (2005): “Testing parameters in GMM without assuming that they are identified,” *Econometrica*, 73, 1103–1123.
- LEWIS, D. J. AND K. MERTENS (2022): “A robust test for weak instruments with multiple endogenous regressors,” *Working Paper*.
- MONTIEL OLEA, J. L. AND C. PFLUEGER (2013): “A robust test for weak instruments,” *Journal of Business & Economic Statistics*, 31, 358–369.
- NEWKEY, W. K. AND D. MCFADDEN (1994): “Large sample estimation and hypothesis testing,” *Handbook of Econometrics*, 4, 2111–2245.
- STOCK, J. H. AND M. YOGO (2005): “Testing for weak instruments in linear IV regression,” *Identification and Inference for Econometric Models: Essays in honor of Thomas Rothenberg*, 80–108.
- VAN DER VAART, A. W. (2000): *Asymptotic Statistics*, vol. 3, Cambridge university press.