ROBUST INFERENCE IN ECONOMETRICS: WEAK INSTRUMENTS, CLUSTERING, AND NON-MONOTONICITY

Luther Yap

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF
ECONOMICS
Advisers: Michal Kolesár and David Lee

May 2025

Abstract

This dissertation consists of three chapters in econometrics. In each chapter, I investigate problems and solutions when assumptions used for valid inference in standard econometric procedures fail.

Chapter 1 considers inference in a linear instrumental variable regression model with many potentially weak instruments and heterogeneous treatment effects. I first show that existing test procedures, including those that are robust to only either weak instruments or heterogeneous treatment effects, can be arbitrarily oversized in this setup. Then, I propose a novel valid inference procedure based on a score statistic and a leave-three-out variance estimator. Within the class of tests that are functions of the leave-one-out analog of a maximal invariant, the score test is asymptotically the uniformly most powerful unbiased test when heterogeneity is imposed. The proposed test also yields a bounded confidence set in empirical applications where existing methods yield unbounded or empty confidence sets.

Chapter 2 proves a new central limit theorem for a sample that exhibits two-way dependence and heterogeneity across clusters. Statistical inference for situations with both two-way dependence and cluster heterogeneity has thus far been an open issue. The existing theory for two-way clustering inference requires identical distributions across clusters (implied by the so-called separate exchangeability assumption). Yet no such homogeneity requirement is needed in the existing theory for one-way clustering. The new result therefore theoretically justifies the view that two-way clustering is a more robust version of one-way clustering, consistent with applied practice. In an application to linear regression, I show that a standard plug-in variance estimator is valid for inference.

Chapter 3 proposes a method to bound policy relevant treatment parameters when the monotonicity assumption that the instrumental variable affects individuals' treatment response in the same direction is weakened. The bounding framework uses the proportion of defiers relative to compliers as a sensitivity parameter, and yields an identified set that is an

interval. The method is illustrated in an empirical application where the same-sex instrument was used to calculate the effect of having a third child on labor force participation. I find that bounds are informative only for small violations in monotonicity.

Acknowledgements

I am grateful for my advisers and committee members Michal Kolesar, David Lee and Ulrich Mueller. They have been very generous with their time, energy, and wisdom in looking through my work, giving thoughtful feedback and guidance. They are outstanding mentors and role models whom I look up to. I am also thankful for other econometricians in the Princeton faculty whom I had helpful conversations with: Bo Honore, Mikkel Plagborg-Moller, Mark Watson, Kevin Dano, and Chris Sims.

I had the privilege of working with outstanding individuals whom I get to call my coauthors: David Lee, Justin McCrary, Marcelo Moreira, Jack Porter, Ruonan Xu, John Gardner, Neil Thakral, Linh To, Jacob Dorn, Andrew Ferdowsian, and Kwok-Hao Lee. They have been tremendously insightful in their work and comments and have often gone out of their way to support me. I have learnt a lot from all of them.

The Princeton Department of Economics and the Industrial Relations Section have given me a safe academic home. I am grateful for the financial support that funded my graduate studies from the Robert W. Ballentine Graduate Scholarship, the Marimar and Cristina Torres Prize, the Harold W. Dodds Honorific Fellowship, and the Clarence J. Hicks Memorial Fellowship. I thank the administrators for answering questions I had, for keeping equipment and finances running, and for ensuring things work in the background.

Friends at Princeton have accompanied me on this journey. I am grateful for fellow econometrics students and my peers in the Industrial Relations Section. I thank the communities at Stone Hill Church and Princeton Christian Fellowship for their unconditional love.

I am grateful for my family: mom, dad, grandparents, and my brother Oliver, for supporting me in my academic pursuit. I also thank my wife Yun for moving across the world for me and for loving me. Thank you, God, for loving me and for creating such a wonderful world that I have the privilege to study.

Contents

List of Tables

List of Figures

Chapter 1

Inference with Many Weak Instruments and Heterogeneity

1.1 Introduction

Many empirical studies in economics involve instrumental variable (IV) models with many instruments. A prominent example is the judge design: several studies argue that judges or case workers are as good as randomly assigned and can affect the treatment status, so they are used as instruments to study the effects of foster care (Doyle, 2007), incarceration (Aizer and Doyle Jr, 2015), detention (Dobbie et al., 2018), disability benefits (Autor et al., 2019), and misdemeanor prosecution (Agan et al., 2023), among others. When the IV is a vector of indicators for judges, the number of instruments can be large relative to the sample size. Another example of many IV is a single instrument interacted with discrete covariates. For instance, when Angrist and Krueger (1991) used the quarter of birth as an instrument to study the returns to education, interacting the quarter of birth with the state of birth can generate 150 instruments.

Despite the pervasiveness and importance of this setting, there does not yet exist an inference procedure that is robust to both heterogeneous treatment effects and weak instru-

ments, which is a gap this paper aims to fill. Weak IV refers to a setting where the first-stage coefficients converge to zero at a rate such that no consistent estimator for the object of interest exists (following from the definition in Mikusheva and Sun (2022) rather than Chao et al. (2012)); and heterogeneous treatment effects refers to a setting where different subsets of the many IV may estimate different local average treatment effects (LATE). There are several recent proposals (Crudu et al., 2021; Mikusheva and Sun, 2022; Matsushita and Otsu, 2022) that are robust to weak IV, but they assume constant treatment effects. A separate literature (Evdokimov and Kolesár, 2018) proposed variance estimators for the Jackknife IV Estimator (JIVE) that are robust to heterogeneous treatment effects, but their t-statistic test is still not robust to many weak IV. While it is clear that weak IV can lead to substantial distortions in inference (e.g., Dufour (1997); Staiger and Stock (1997)), it is less obvious if procedures developed under constant treatment effects that are robust to weak IV are still valid with heterogeneous treatment effects.

In this paper, I first show that neglecting either heterogeneity or weak instruments can result in substantial distortions in inference. Section 1.2 presents a simple simulation that has both weak instruments and heterogeneous treatment effects. For a nominal 5% test, using the procedure from Mikusheva and Sun (2022) (MS22), which is robust to weak instruments but not heterogeneity, can result in 100% rejection under the null, because their test statistic is not centered correctly when there is heterogeneity. This result is attributed to how their test is a joint test of both the parameter value and the null of no heterogeneity. Similarly, the procedure from Evdokimov and Kolesár (2018) (EK18), which is robust to heterogeneity but not weak instruments, can be severely oversized. Additionally, this section documents how an empirically common practice of constructing a "leniency measure" that combines the many instruments and then using weak IV robust procedures from the just-identified IV literature is invalid.

Given the stark simulation results, Section 1.3 proposes a procedure for valid inference. Following the many instruments literature, the JIVE estimand is the object of interest — this estimand can be interpreted as a weighted average of treatment effects when there is heterogeneity (e.g., EK18). Using weak identification asymptotics, I show that the Lagrange Multiplier (LM) (i.e., score) statistic, earlier proposed by Matsushita and Otsu (2022) under constant treatment effects, is mean zero and asymptotically normal even with treatment effect heterogeneity. In fact, I prove a stronger normality result that a set of jackknife statistics that includes the LM is jointly normal, which is the first technical challenge of this paper. This normality result uses an asymptotic environment that nests the asymptotic environments of EK18 and MS22 in that normality holds if either the number of instruments is large or the instruments are strong. This normality implies that, as long as the variance of LM is consistently estimable, a t-statistic can be calculated and critical values from the standard normal distribution are valid for inference. Obtaining a consistent variance estimator is the second technical challenge of the paper, since reduced-form coefficients are not consistently estimable when there are few observations per instrument. Motivated by Anatolyev and Sølvsten (2023) who proposed a method to jointly test the significance of many covariates in OLS, I construct a leave-three-out (L3O) variance estimator for the LM variance and show that it is consistent, even when reduced-form coefficients are not consistently estimable. Due to the generality of the setting considered, beyond its robustness to weak IV and heterogeneity, the procedure proposed in this paper is also robust to heteroskedasticity, and potentially many covariates, so it retains the advantages of existing procedures in the literature.

Section 1.4 argues that the proposed LM procedure is powerful. In the over-identified IV environment with normal homoskedastic errors, Moreira (2009a) showed that, if we are willing to restrict our attention to tests that are invariant to rotations of the instruments, it suffices to consider tests that are functions of three statistics. These three statistics are known as a "maximal invariant". To be robust to non-normality and heteroskedasticity in the

many IV environment, I focus on the leave-one-out (L1O) analog of this maximal invariant. The proposed LM statistic is one of the three statistics in the L1O analog, and I show that the two-sided LM test is asymptotically uniformly most powerful unbiased (UMPU) within the class of tests that are functions of this L1O analog, for the interior of the alternative space (i.e., where heterogeneity is imposed).

Simulation results in Section 1.5 show how the procedure is robust even with a small number of instruments, and it is reasonably powerful even with constant treatment effects. Section 1.6 contains two empirical applications that show how being robust to many weak IV and heterogeneity can change conclusions.[1] In the Angrist and Krueger (1991) quarter of birth application, the Matsushita and Otsu (2022) procedure that are robust to many weak IV but not heterogeneity have unbounded confidence sets while L3O has a bounded confidence set. In the Agan et al. (2023) judge application, MS22 has an empty confidence set, and the length of the L3O confidence interval is more than twice that of EK18 that is not robust to many weak IV.

This paper contributes to the following strands of literature. First, this paper contributes to a growing literature on many weak instruments. There is a strand of literature dealing with many instruments (e.g., Chao et al. (2012)) and another separate strand dealing with weak instruments (e.g., Staiger and Stock (1997); Lee et al. (2023)). While recent procedures accommodate both simultaneously (e.g., Crudu et al. (2021); Mikusheva and Sun (2022); Matsushita and Otsu (2022); Lim et al. (2024)), their focus has been on the linear IV model with constant treatment effects. This paper augments their setup by allowing for heterogeneity in treatment effects, and contributes new results on the limitations of their procedures under heterogeneity. Further, I show how heterogeneity can be understood in a framework analogous to weak instruments.

[1]Implementation code can be found at: https://github.com/lutheryap/mwivhet.

Second, this paper contributes to the literature on heterogeneous treatment effects (e.g., Kolesár (2013); Evdokimov and Kolesár (2018); Blandhol et al. (2022)). The previous papers exploit consistent estimation of the object of interest to conduct inference. In contrast, this paper uses the (more general) weak IV environment where the object of interest may not be consistently estimated. Two recent papers allow weak IV and heterogeneity. Boot and Nibbering (2024) study a single discrete instrument interacted and saturated with many covariates. Their setup is a special case of the environment considered in this paper, so it is unclear if their procedure generalizes to many instruments without covariates (e.g., judges). Kleibergen and Zhan (2025) target a continuous updating (CU) GMM estimator with a fixed number of instruments rather than many instruments.[2]

Third, this paper contributes to a literature on inference when coefficients cannot be consistently estimated. The difficulty in having such a general robust inference procedure lies in consistent variance estimation when the number of coefficients is large. Recent literature that has made substantial progress in a different context. In doing inference in OLS with many covariates, Cattaneo et al. (2018) and Anatolyev and Sølvsten (2023) proposed consistent variance estimators that are robust to heteroskedasticity, which involve inverting a large (n by n, where n is the sample size) matrix and a L3O approach respectively. Boot and Nibbering (2024) adapt the Cattaneo et al. (2018) variance estimator for inference. In contrast, this paper adapts the approach from Anatolyev and Sølvsten (2023) that does not require an inversion of an n by n matrix, and whose L3O implementation is fast when using matrix operations.

Fourth, this paper contributes to a literature on optimal tests. While the UMPU test for just-identified IV has been established since Moreira (2009b), obtaining a UMPU test in the over-identified IV environment has thus far been more challenging. In the over-identified IV

[2]They show that their CU-GMM estimator corresponds to the limited information maximum likelihood (LIML) estimator. However, it is also known that the LIML estimand may not be interpretable as a weighted average of LATE's. (Kolesár, 2013) I am unaware of any paper that allows both weak IV and heterogeneity with a fixed number of instruments and targets a parameter that is a weighted average of LATE's.

environment with constant treatment effects, several statistics are informative of the object of interest. Consequently, there is a large literature that numerically compares various valid tests and characterizes various forms of optimality (e.g., Moreira (2003); Andrews (2016); Andrews et al. (2019); Van de Sijpe and Windmeijer (2023); Lim et al. (2024)). By imposing heterogeneity in the environment, the problem is (somewhat surprisingly) simplified. Since only one statistic in the asymptotic distribution is directly informative of the object of interest, I can obtain a UMPU result.

1.2 Challenges in Conventional Practice

This section explains the challenges faced in conventional practice by considering a simple potential outcomes model without covariates that exhibits weak instruments and heterogeneity in treatment effects. This model is a special case of the general model in Section 1.3. A simulation from the model shows how weak instruments and heterogeneity can lead to substantial distortions in inference for procedures recently proposed in the econometric literature. A common empirical practice of constructing a leave-one-out instrument and then applying inference methods for the instrument as if it is not constructed also has high rejection rates. In contrast, the method proposed in this paper has a rejection rate that is close to the nominal rate.

1.2.1 Setting for Simple Example

The simple example uses the canonical latent variable framework of Heckman and Vytlacil (2005). We are interested in the effect of $X_i \in \{0, 1\}$ (e.g., incarceration) on some outcome Y_i, for $i = 1, \cdots n$ that indexes individuals. To instrument for X_i, we use a vector of judges indicators: Z_i is a $(K + 1)$-dimensional vector of indicators for judges, indexed $k = 1, \cdots, K + 1$, each with $c = 5$ individual cases, so the vector takes value 1 for the kth

component when individual i is matched to judge k, and 0 elsewhere. Then, $n = (K+1)c$. The problem of many instruments arises when c is fixed while K increases. Let $Y_i(0)$ and $Y_i(1)$ denote the untreated and treated potential outcomes respectively, and we observe $Y_i = Y_i(X_i)$. The treatment status given some instrument value z is $X_i(z)$, and we observe $X_i(Z_i)$. The model is:

$$X_i(z) = 1\{z'\lambda > v_i\}, \text{ and } Y_i(x) = xf(v_i) + \varepsilon_i, \tag{1.1}$$

where $1\{\cdot\}$ is an indicator function that takes the value 1 if the argument is true and 0 otherwise. Here, $Z_i'\lambda = \lambda_{k(i)}$, where $k(i)$ is the judge that individual i is matched to. With individual unobservable $v_i \sim U[0,1]$, the probability of treatment (i.e., $X_i = 1$) given judge k is λ_k. I set $\lambda_k = 1/2$ for the base judge, and evenly split all other K judges to take 4 different values of λ_k. Potential outcomes are $Y_i(0) = \varepsilon_i$ and $Y_i(1) = f(v_i) + \varepsilon_i$ so $Y_i(1) - Y_i(0) = f(v_i)$ is the treatment effect. The individual-specific residuals v_i and ε_i are allowed to be arbitrarily correlated. Let β_k denote the local average treatment effect (LATE) when comparing judge k to the base judge: for instance, when $\lambda_k > 1/2$, $\beta_k = \frac{1}{\lambda_k - 1/2}\int_{1/2}^{\lambda_k} f(v)dv$. The values of (λ_k, β_k) for the 4 groups of judges are $(1/2 - s, \beta - h/s)$, $(1/2(1-s), \beta + 2h/s)$, $(1/2(1+s), \beta - 2h/s)$, and $(1/2 + s, \beta + h/s)$. The function $f(v)$ that delivers these parameters and further details of this example are in Section 1.A.2.

The λ_k and β_k values are parameterized by objects s and h, which control the IV strength and heterogeneity in the model respectively. The impact of these parameters are illustrated in Figure 1.1 that plots the point masses for the four groups of judges in reduced-form. Parameter s controls how far $E[X \mid Z]$ are spread across judges, which then affects the instrument strength. Parameter h controls the distance between the mass points and a line with slope β — this slope is the object of interest. If the impact of X on Y is homogeneous, then $h = 0$, and all mass points must lie on a line — this implication is falsifiable by the data.

Figure 1.1: IV Strength and Heterogeneity in Reduced Form

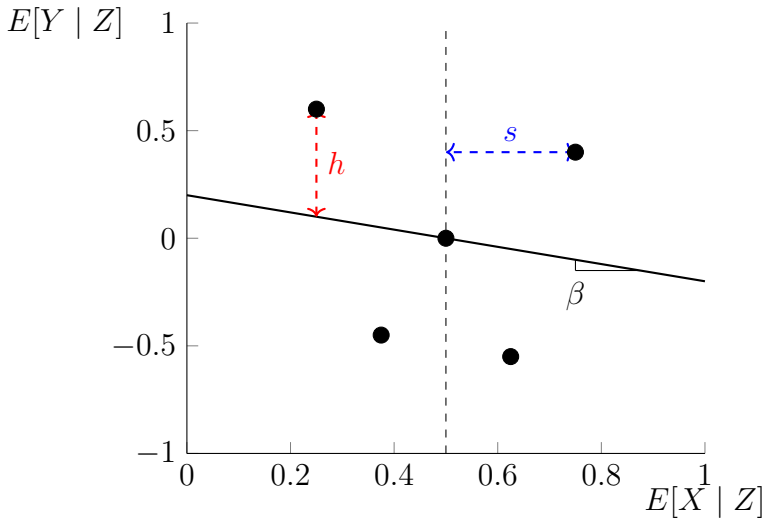


Table 1.1: Rejection rates under the null for nominal size 0.05 test

Designs		Procedures								
$E[T_{AR}]$	$E[T_{FS}]$	TSLS	EK	MS	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O	LMorc	ARorc
$2\sqrt{K}$	$2\sqrt{K}$	0.428	0.066	NaN	0.027	0.041	0.041	0.060	0.048	1.000
	2	0.978	0.040	NaN	0.282	0.114	0.303	0.052	0.049	1.000
	0	0.983	0.025	NaN	0.260	0.054	0.282	0.047	0.055	1.000
2	$2\sqrt{K}$	0.984	0.076	1.000	0.039	0.046	0.049	0.044	0.050	1.000
	2	1.000	0.096	1.000	0.085	0.155	0.149	0.048	0.049	1.000
	0	1.000	0.128	1.000	0.103	0.225	0.177	0.060	0.051	1.000
0	$2\sqrt{K}$	0.994	0.097	0.064	0.064	0.071	0.067	0.059	0.055	0.057
	2	1.000	0.231	0.059	0.047	0.179	0.106	0.049	0.051	0.055
	0	1.000	0.359	0.063	0.041	0.350	0.107	0.048	0.046	0.059

Notes: The table displays rejection rates of various procedures (in columns) for various designs (in rows). Details of the data generating process are in Section 1.A.2. I use $K = 400, c = 5, \beta = 0$ with 1000 simulations. TSLS implements the standard two-stage-least-squares t-test for an over-identified IV system. EK implements the procedure in Evdokimov and Kolesár (2018). MS uses T_{AR} with the cross-fit procedure in Mikusheva and Sun (2022). MO uses the T_{LM} statistic with the variance estimator proposed in Matsushita and Otsu (2022). $\tilde{X}$-t uses a constructed instrument and runs TSLS for a just-identified IV system. $\tilde{X}$-AR uses the Anderson and Rubin (1949) (AR) procedure for a just-identified system using a constructed instrument. L3O uses the variance estimator proposed in this paper. LMorc is the infeasible theoretical benchmark that uses an LM statistic with an oracle variance. ARorc uses the AR statistic with an oracle variance.

The simulation designs vary the values of s and h through the following parameters:

$$E[T_{FS}] = \frac{5}{8}\sqrt{K}(c-1)s^2, \text{ and } E[T_{AR}] = \frac{5}{8}\sqrt{K}(c-1)h^2. \tag{1.2}$$

The statistic T_{FS} is the leave-one-out (L1O) analog of the first-stage "F" statistic in this model, and T_{AR} is similarly the L1O analog of the Anderson-Rubin statistic under the null. These objects are explained in detail in Section 1.3, but it suffices to mention here that, for the given c and K, there is a one-to-one mapping between $(E[T_{FS}], E[T_{AR}])$ and (s, h). Using Staiger and Stock (1997) asymptotics, $E[T_{FS}]$ is the parameter that determines whether there is strong or weak identification. Where C is some positive arbitrary constant, $E[T_{FS}] \to \infty$ is an environment with strong identification where the object of interest can be estimated consistently, and $E[T_{FS}] \to C < \infty$ is an environment with weak identification where no consistent estimator exists.

For every design, I generate data under the null and calculate the frequency that each inference procedure rejects the null of $\beta_0 = 0$. These procedures include the standard TSLS t-test, procedures that are robust to either weak instruments (MO, MS) or heterogeneity (EK), and procedures that use a constructed instrument ($\tilde{X}$). The results are presented in Table 1.1, which I will refer to in the remainder of this section as I explain them.

1.2.2 Issue with Many Weak Instruments

Many IV and weak IV are different but related issues. The many IV problem arises when the number of cases per judge c does not diverge to infinity, so that K is large relative to n. When c is small, the judge-specific λ_k and β_k cannot be consistently estimated and hence inference procedures like the TSLS t-test can be oversized. In Figure 1.1, a small c is attributed to the sample uncertainty surrounding each black circle. The weak IV problem

arises from $E[T_{FS}]$ not diverging: since $E[T_{FS}]$ is a function of K, c, and s, the weak IV issue is related to the many IV issue.

If we simply run the TSLS t-test for an over-identified model, then the estimator can be asymptotically biased and inference is invalid, a fact already known in the literature. This fact is also evident in Table 1.1, where TSLS has 100% rejection in many designs. In TSLS, the first stage regresses X on Z to get a predicted $\hat{X} = Z\hat{\pi}$, where $\hat{\pi}$ is the estimated coefficient; the second stage regresses Y on $\hat{X}$. With constant treatment effects, the asymptotic bias of the TSLS estimator depends on $\sum_i \varepsilon_i \hat{X}_i / \sum_i \hat{X}_i^2$. When every judge only has $c = 5$ cases, the influence of v_i on $\hat{\pi}_{k(i)}$ and hence $\hat{X}_i$ is non-negligible. Since ε_i and v_i can be arbitrarily correlated, the numerator is biased. If the instruments are weak such that the denominator $\sum_i \hat{X}_i^2$ does not diverge sufficiently quickly, then the asymptotic bias can be large. Due to the asymptotic bias, the t-statistic is not centered around β_0 when data is generated under the null, so we observe over-rejection in Table 1.1.

Since the bias in the TSLS estimator arises from using X_i to estimate $\hat{\pi}$, a natural solution to address that bias is to use the JIVE to estimate β. Instead of using $\hat{X}_i = Z_i'\hat{\pi}$ in the second stage, we instead use $\tilde{X}_i = Z_i'\hat{\pi}_{-i}$, where $\hat{\pi}_{-i}$ is the coefficient from the first-stage regression that leaves out observation i. I call $\hat{\pi}_{-i}$ the leave-one-out (L1O) coefficient. With $P = Z\left(Z'Z\right)^{-1}Z'$ denoting the projection matrix, $\tilde{X}_i = Z_i'\hat{\pi}_{-i}$ can be written as $\tilde{X}_i = \sum_{j \neq i} P_{ij} X_j$. Then, the JIVE is:

$$\hat{\beta} = \frac{\sum_i Y_i \left(\sum_{j \neq i} P_{ij} X_j\right)}{\sum_i X_i \left(\sum_{j \neq i} P_{ij} X_j\right)}. \tag{1.3}$$

In the many IV context with constant treatment effects, the asymptotic distribution of the t-statistic of the JIVE is the same as the distribution of the t-statistic of the TSLS estimator in the just-identified environment (Mikusheva and Sun, 2022) — it is a ratio of two normally distributed random variables. It is well-known that, in the just-identified IV context with

weak IV, the rejection rate of the standard t-statistic can be up to 100% for a nominal 5% test (e.g., Dufour (1997)). Hence, like the just-identified IV context, by using a structural model that has sufficiently weak instruments and high covariance, the simulation can deliver high rejection rates.

EK18 have a procedure that is robust to heterogeneity, but not weak instruments, so even if we use their variance estimator for the t-statistic, this problem is not alleviated. This fact is evident in the EK column of Table 1.1, where, with a sufficiently large correlation in the individual unobservables, rejection rates can be large.[3] Hence, ignoring the issue of weak instruments can lead to substantial distortions in inference. In fact, even with strong instruments, there is no guarantee that EK18 achieves the nominal rate, because their variance estimation method requires consistent estimation of the first-stage coefficients $\hat{\pi}$. A condition for consistent variance estimation is that the number of cases per judge is large, which is not $c = 5$.

Remark 1.1. *In the literature, there have been several definitions of weak instruments, which I clarify in this remark. Using Equation (1.2), there are three asymptotic regimes, ordered from the strongest to the weakest: (i)* $\frac{1}{\sqrt{K}}E[T_{FS}] \to \infty$*, (ii)* $E[T_{FS}] \to \infty$*, and (iii)* $E[T_{FS}] \to C < \infty$*. Regime (i) is a necessary condition for the TSLS estimator to be consistent, so* $\frac{1}{\sqrt{K}}E[T_{FS}] \to C < \infty$ *is what Stock and Yogo (2005) would refer to as weak instruments. Regime (ii) is a necessary condition for the JIVE to be consistent (e.g., Chao et al. (2012), EK18). Regime (iii) is where no estimator is consistent (e.g., Mikusheva and Sun (2022)). If K is fixed, then (i) and (ii) are the same asymptotically, and (iii) is the relevant weak identification asymptotic regime. If $K \to \infty$, then there is more ambiguity in what weakness means: Chao et al. (2012) and EK18 who assume (ii) are robust to weak instruments when defined in the Stock and Yogo (2005) sense, because s can converge to 0, albeit at a slower rate than $\sqrt{K}$. In this paper, I follow the Staiger and Stock (1997) standard*

[3]The rejection rate of EK can be 100% under the null in some simulations: one example is given in Table 1.C.1 in Section 1.C.2.

of weak identification where no consistent estimator exists, which corresponds to (iii) that EK18 is not robust to.

1.2.3 Issue with Heterogeneity

Next, consider proposals for inference that are developed for contexts with many weak IV. MS22 (and Crudu et al. (2021)) propose using an Anderson-Rubin (AR) statistic $T_{AR} = \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij} e_i(\beta_0) e_j(\beta_0)$, for $e_i(\beta_0) := Y_i - X_i\beta_0$ where β_0 is the hypothesized null value. With constant treatment effects, $e_i := Y_i - X_i\beta$ is the residual. Hence, if the instrument is orthogonal to the residual, then $E[Z_i e_i] = 0$.[4] Then, T_{AR} is the L1O analog for the quadratic form that tests the moment $E[Z_i e_i] = 0$. Since observations are independent, the critical value for the test is obtained from a mean-zero normal distribution. In this model, $E[T_{AR}] = \sqrt{K}(c-1)h^2$ under the null.[5] Hence, when there are constant treatment effects such that $h = 0$ for all k, the statistic is unbiased. However, in the setup with heterogeneity, the T_{AR} can be biased: in fact, when h does not converge to zero, $E[T_{AR}]$ diverges, resulting in a 100% rejection rate under the null, even if the oracle variance were used. Further, there does not exist any estimand β such that $E[T_{AR}] = 0$, as shown in Lemma 1.1 of Section 1.A.2.

A further problem with the feasible MS procedure is that when there is strong heterogeneity ($E[T_{AR}] = 2\sqrt{K}$) in this simulation, their cross-fit variance estimate is negative for all simulation draws, as the negative heterogeneity terms are larger in magnitude than the positive variances of the residuals. The formal analysis requires more notation from Section 1.3, so details are deferred to Section 1.A.2.

[4] An equivalent way to see how heterogeneity affects inference is through the framework of Hall and Inoue (2003) and Lee (2018): $E[Z_i e_i] = 0$ is a special case of a misspecified over-identified GMM problem. The instruments are individually valid, but every component of the K moments in $E[Z_i e_i] = 0$ identifies a different treatment effect, so there is no parameter that satisfies all moments simultaneously under heterogeneity. Then, while the estimand is still interpretable as a combination of these treatment effects due to how GMM weights these moments, there are additional components in the variance that affect inference.

[5] This result can be obtained as a special case of Theorem 1.1 in Section 1.3 and using the fact that $\sum_i \sum_{j \neq i} P_{ij}^2 = \sum_i \sum_{j \neq i} (1/c^2) = \sum_i \frac{c-1}{c^2} = \sum_k \frac{c-1}{c}$.

Another proposal in the literature that is robust to many weak instruments is Matsushita and Otsu (2022) (MO22) who use the statistic $T_{LM} = \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij} e_i(\beta_0) X_j = \frac{1}{\sqrt{K}} \sum_i e_i(\beta_0) \tilde{X}_i$. This statistic can be interpreted as the LM (or score) statistic that uses the moment $E[e\tilde{X}] = 0$. They propose the following variance estimator $\hat{\Psi}_{MO}$:

$$\hat{\Psi}_{MO} := \frac{1}{K} \sum_i \left(\sum_{j \neq i} P_{ij} X_j \right)^2 e_i(\beta_0)^2 + \frac{1}{K} \sum_i \sum_{j \neq i} P_{ij}^2 X_i e_i(\beta_0) X_j e_j(\beta_0). \tag{1.4}$$

While T_{LM} has zero mean under the null even with heterogeneity, a result shown later in Section 1.3, the MO22 variance estimator was constructed under constant treatment effects, so the variance estimand differs from the true variance. It can be shown that $E\left[\hat{\Psi}_{MO}\right] \neq \text{Var}(T_{LM})$, and $\hat{\Psi}_{MO}$ is inconsistent in general, so when it is used to construct the t-statistic of T_{LM}, the normalized statistic is not distributed $N(0,1)$ asymptotically. Consequently, by constructing a DGP where $\hat{\Psi}_{MO}$ underestimates the variance, it is possible to get over-rejection of the MO22 procedure, as in the cases of Table 1.1 where $E[T_{AR}]$ diverges. As expected, when there is no heterogeneity such that $h = 0$, the rejection rate of MO22 and MS22 are close to the nominal rate.

1.2.4 Issue with a Constructed Instrument

In light of problems with weak identification and heterogeneity, there is a large applied literature that transforms a many IV environment into a just-identified single-IV environment. With a single IV, the Anderson and Rubin (1949) (AR) procedure (among others) is robust to both weak identification and heterogeneity. However, this subsection will argue that such an approach is invalid.

Due to how the JIVE is written, there are several empirical papers that treat $\tilde{X}_i = \sum_{j \neq i} P_{ij} X_j$ as the "instrument" so that $\hat{\beta} = \sum_i Y_i \tilde{X}_i / \sum_i X_i \tilde{X}_i$, and proceed with inference as if $\tilde{X}_i$ is not constructed, but is an observed scalar instrument, usually referred to as a

leniency measure. While the resulting estimator is numerically identical to JIVE, there are distortions in inference because the variance estimators do not account for the variability in constructing $\tilde{X}_i$.

If the TSLS t-statistic inference is used as if $\tilde{X}_i$ is the instrument, then its rejection rates in designs with heterogeneity are usually higher than rejection rates of EK18 that accounts for the variance accurately, by comparing the $\tilde{X}$-t and EK columns in Table 1.1. Consequently, in the cases where EK under-rejects, $\tilde{X}$-t can have close to nominal rejection rates by coincidence.

Even if the weak IV robust AR procedure for just-identified IV were used, there can still be distortion in inference (see $\tilde{X}$-AR in Table 1.1). The AR t-statistic is $t_{\tilde{X}AR} := \sum_i e_i(\beta_0)\tilde{X}_i/\sqrt{\hat{V}}$, where $\hat{V} = \sum_i \tilde{X}_i^2\hat{\varepsilon}_i^2/\left(\sum_i \tilde{X}_i^2\right)^2$ and $\hat{\varepsilon}_i = e_i(\beta_0) - \tilde{X}_i\left(\sum_i e_i(\beta_0)\tilde{X}_i\right)/\left(\sum_i \tilde{X}_i^2\right)$. Even though $t_{\tilde{X}AR}$ is mean zero and asymptotically normal, the variance estimand is inaccurate, much like MO22. In particular, when $\beta_0 = \beta = 0$, the leading term of the variance estimand is $E\left[\sum_i \tilde{X}_i^2 e_i^2\right]$, and it does not converge to the true variance derived in Section 1.3 in general. Hence, using the just-identified AR procedure with a constructed instrument results in over-rejection. There are several papers that cluster standard errors by judges, but this approach faces a similar issue.[6]

As a preview, the L3O procedure proposed in this paper has rejection rates close to the nominal rate while the other procedures can over-reject.

1.3 Valid Inference

In light of how existing procedures are invalid in an environment with many weak instruments and heterogeneity as documented in the previous section, this section describes a novel inference procedure and shows that it is valid. I set up a general model, then show that an

[6]Details of this discussion are relegated to Section 1.C.1.

LM statistic is asymptotically normal and a feasible variance estimator is consistent, which suffices for inference.

1.3.1 Setting: Model and Asymptotic Distribution

The general setup mimics Evdokimov and Kolesár (2018). With an independently drawn sample of individuals $i = 1, \ldots, n$, we observe each individual's scalar outcome Y_i, scalar endogenous variable X_i, instrument Z_i, and covariates W_i, with $\dim(Z_i) = K$.[7] For every instrument value z, there is an associated potential treatment $X_i(z)$, and we observe $X_i = X_i(Z_i)$. Similarly, potential outcomes are denoted $Y_i(x)$, with $Y_i = Y_i(X_i)$. Let $R_i := E[X_i \mid Z_i, W_i]$ and $R_{Yi} := E[Y_i \mid Z_i, W_i]$ be linear in Z_i and W_i. The model, written in the reduced-form and first-stage equations, is:

$$\begin{aligned} Y_i &= R_{Yi} + \zeta_i, \text{ where} & R_{Yi} &= Z_i'\pi_Y + W_i'\gamma_Y, & E[\zeta_i \mid Z_i, W_i] &= 0, \text{ and} \\ X_i &= R_i + \eta_i, \text{ where} & R_i &= Z_i'\pi + W_i'\gamma, & E[\eta_i \mid Z_i, W_i] &= 0. \end{aligned}$$

The setup implicitly conditions on Z_i, W_i, so R_i, R_{Yi} are nonrandom.[8] Linearity in Z and W is not necessarily restrictive when there is full saturation or when K is large.[9]

Define $e_i := Y_i - X_i\beta$, where β is some estimand of interest, and e_i is a linear transformation. Let $e_i(\beta_0) := Y_i - X_i\beta_0$ denote the feasible null-imposed linear transformation. Let $R_{\Delta i} := R_{Yi} - R_i\beta$ and $\nu_i := \zeta_i - \eta_i\beta$. These definitions imply $e_i = R_{\Delta i} + \nu_i$ and $R_{\Delta i} = Z_i'(\pi_Y - \pi\beta) + W_i'(\gamma_Y - \gamma\beta)$. Since $E[\nu_i | Z_i, W_i] = 0$ from the model, $E[e_i | Z_i, W_i] = R_{\Delta i}$,

[7] The endogenous variable X_i can be extended to a vector with some technical modifications and without conceptual complications.

[8] If we are interested in a superpopulation where Z is random, then the estimands would be defined as the probability limit of the conditional objects. Then, it suffices to have regularity conditions to ensure that the conditional object converges to the unconditional object.

[9] Any nonlinear function of the instruments can be arbitrarily well-approximated by a spline with a large number of pieces or a high-order polynomial. Moreover, the arguments in this paper could presumably be extended to a linear approximation of nonlinear functions as long as there are regularity conditions to ensure that higher-order terms are asymptotically negligible.

which need not be zero. For data matrix A, let $H_A = A\left(A'A\right)^{-1}A'$ denote the hat (i.e., projection) matrix and $M_A = I - H_A$ its corresponding annihilator matrix. With Z, W denoting the corresponding data matrices of the instrument and covariates, let $Q = (Z, W)$, $P = H_Q$, and $M = I - P$. C denotes arbitrary constants.

Remark 1.2. *While $E\left[e_i|Z_i, W_i\right] = R_{\Delta i}$ need not be zero under heterogeneous treatment effects, $E\left[e_i|Z_i, W_i\right] = R_{\Delta i} = 0$ under constant treatment effects. Since $R_{\Delta i} = Z_i'(\pi_Y - \pi\beta) + W_i'(\gamma_Y - \gamma\beta)$ for all i, constant treatment effects with $E[Y_i - X_i\beta \mid Z_i, W_i] = 0$ also implies $\pi_Y = \pi\beta$ and $\gamma_Y = \gamma\beta$ outside of edge cases (e.g., when Z_i, W_i are always 0). These R_Δ objects hence capture the impact of having heterogeneous treatment effects in the many instruments model.*

The (conditional) object of interest and its corresponding estimator are:

$$\beta_{JIVE} := \frac{\sum_i \sum_{j\neq i} G_{ij} R_{Yi} R_j}{\sum_i \sum_{j\neq i} G_{ij} R_i R_j}, \text{ and } \quad \hat{\beta}_{JIVE} = \frac{\sum_i \sum_{j\neq i} G_{ij} Y_i X_j}{\sum_i \sum_{j\neq i} G_{ij} X_i X_j},$$

where G is an $n \times n$ matrix that can take several forms. As the leading cases, if there are no covariates, using the projection matrix $G = H_Z = P$ is the standard JIVE, and when there are covariates, I use the unbiased JIVE "UJIVE" (Kolesár, 2013) with $G = \left(I - \text{diag}\left(H_Q\right)\right)^{-1} H_Q - \left(I - \text{diag}\left(H_W\right)\right)^{-1} H_W$. In an environment with a binary instrument and many covariates interacted with the instrument, the saturated estimand "SIVE" (Chao et al., 2023; Boot and Nibbering, 2024) uses $G = P_{BN} - M_Q D_{BN} M_Q$, where $P_{BN} = M_W Z \left(Z' M_W Z\right)^{-1} Z' M_W$ and D_{BN} is defined as a diagonal matrix with elements such that $P_{BN,ii} = \left[M_Q D_{BN} M_Q\right]_{ii}$. With constant treatment effects, the estimand is the same for all the estimators: $R_{Yi} = R_i\beta$ so $\beta_{JIVE} = \beta$. Depending on the application, the estimand is usually interpretable as some weighted average of treatment effects when using JIVE without

covariates or UJIVE with covariates with a saturated regression.[10] (Evdokimov and Kolesár, 2018) The focus of this paper is on inference, so I will not discuss the estimand in detail. The results for valid inference in the paper are established for any G that satisfies properties that will be formally stated in the theorem.

This paper restricts its attention to the following statistics:

$$(T_{AR}, T_{LM}, T_{FS})' := \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} G_{ij} \left(e_i(\beta_0) e_j(\beta_0), e_i(\beta_0) X_j, X_i X_j \right)'. \tag{1.5}$$

It suffices to focus on (T_{AR}, T_{LM}, T_{FS}) for inference as they correspond to a linear transformation of the leave-one-out analog of a maximal invariant — details are in Section 1.4.1. T_{AR} is the (unnormalized) AR statistic used by MS22 for inference, and T_{LM} is the LM (score) statistic used by MO22. T_{FS} corresponds to a first-stage F statistic that can be used as a diagnostic for weak instruments.

The asymptotic behavior depends on the following object:

$$r_n := \sum_i \left(\sum_{j \neq i} G_{ij} R_j \right)^2 + \sum_i \left(\sum_{j \neq i} G_{ij} R_{\Delta j} \right)^2 + \sum_i \sum_{j \neq i} G_{ij}^2. \tag{1.6}$$

Asymptotic theory in this paper uses $r_n / \sqrt{K} \to \infty$, which nests the environments of EK18, MS22, and MO22: as long as one of the three objects in Equation (1.6) diverges at a rate above $\sqrt{K}$, we obtain $r_n / \sqrt{K} \to \infty$. EK18 assume $\sum_i \left(\sum_{j \neq i} G_{ij} R_j \right)^2 / \sqrt{K} \to \infty$, which implies strong identification, but $r_n / \sqrt{K} \to \infty$ can also be achieved if either of the latter terms in r_n diverges. MS22 and MO22 assume $K \to \infty$. Without covariates, $G = P$, so $\sum_i \sum_{j \neq i} G_{ij}^2 = O(K)$, and hence $r_n / \sqrt{K} \to \infty$. Hence, to apply the asymptotic theory in this paper, it suffices to have either strong identification, or $K \to \infty$. The only case ruled

[10] In the judge example without covariates above, we have $G = P$ and $\pi_{Yk} = \beta_k \pi_k$ where β_k is the local average treatment effect (LATE) between judge k and the base judge, so $\beta_{JIVE} = \frac{\sum_k \pi_{Yk} \pi_k}{\sum_k \pi_k^2} = \frac{\sum_k \pi_k^2 \beta_k}{\sum_k \pi_k^2}$ is a weighted average of LATE's.

out is where K is fixed, and there is weak identifcation in that $\sum_i \left(\sum_{j\neq i} G_{ij}R_j\right)^2/\sqrt{K}$ does not diverge.

The following assumption states sufficient conditions for joint asymptotic normality.

Assumption 1.1. *(a) There exists $C<\infty$ such that $E[\eta_i^4]+E[\nu_i^4]\leq C$ for all i.*

(b) $E[\nu_i^2]$ and $E[\eta_i^2]$ are bounded away from 0 and $|\text{corr}(\nu_i,\eta_i)|$ is bounded away from 1.

(c) There exists $\underline{c}>0$ such that for any c_1,c_2,c_3 that are not all 0,

$$\begin{aligned}&\tfrac{1}{r_n}\textstyle\sum_i\left(c_3\sum_{j\neq i}(G_{ij}+G_{ji})R_j+c_2\sum_{j\neq i}G_{ji}R_{\Delta j}\right)^2\\&+\tfrac{1}{r_n}\textstyle\sum_i\left(c_1\sum_{j\neq i}(G_{ij}+G_{ji})R_{\Delta j}+c_2\sum_{j\neq i}G_{ij}R_j\right)^2\\&+\tfrac{1}{r_n}\text{Var}\left(\textstyle\sum_i\sum_{j\neq i}G_{ij}(c_1\nu_i\nu_j+c_2\nu_i\eta_j+c_3\eta_i\eta_j)\right)\geq\underline{c}.\end{aligned}$$

(d) $\tfrac{1}{r_n^2}\sum_i\left(\left(\sum_{j\neq i}G_{ij}R_j\right)^4+\left(\sum_{j\neq i}G_{ij}R_{\Delta j}\right)^4+\left(\sum_{j\neq i}G_{ji}R_j\right)^4+\left(\sum_{j\neq i}G_{ji}R_{\Delta j}\right)^4\right)\to 0.$

(e) $||\tfrac{1}{r_n}G_LG_L'||_F+||\tfrac{1}{r_n}G_UG_U'||_F\to 0$, *where G_L is a lower-triangular matrix with elements $G_{L,ij}=G_{ij}1\{i>j\}$ and G_U is an upper-triangular matrix with elements $G_{U,ij}=G_{ij}1\{i<j\}$.*

Assumption 1.1 states high-level conditions that mimic EK18 so that a central limit theorem (CLT) can be applied. These conditions hence accommodate the G that EK18 consider with covariates. Having bounded moments in (a) is standard. Conditions (b) and (c) are sufficient to ensure that the variance is non-zero asymptotically. In particular, (b) rules out perfect correlation: in the simulation, $\text{corr}(\eta_i,\nu_i)=-1$ is the pathological case that makes the variance zero, but $\text{corr}(\eta_i,\nu_i)=1$ still allows non-zero variance. Conditions (d) and (e) ensure that the weights placed on the individual stochastic terms are not too large. The condition that $r_n/\sqrt{K}\to\infty$ is implied by (e) when $G=P$: due to Lemma B3 of Chao et al. (2012), under weak IV asymptotics where $P_{ii}\leq C<1$, we obtain $||G_LG_L'||_F\leq C\sqrt{K}$.

Mechanically, if there is weak IV and fixed K, then $||\frac{1}{r_n}G_L G_L'||_F = \frac{1}{K}O(\sqrt{K}) \neq o(1)$, so (e) fails when $r_n/\sqrt{K}$ does not diverge. Notably, the conditions do not require $P_{ii} \to 0$ so the π, π_Y coefficients need not be consistently estimated.

Theorem 1.1. *If Assumption 1.1 holds and $\beta = \beta_0$, then $\hat{\beta}_{JIVE} - \beta_{JIVE} = T_{LM}/T_{FS}$, and for $V = Var(T_{AR}, T_{LM}, T_{FS})$,*

$$V^{-1/2}\begin{pmatrix} T_{AR} - \frac{1}{\sqrt{K}}\sum_i \sum_{j\neq i} G_{ij}R_{\Delta i}R_{\Delta j} \\ T_{LM} \\ T_{FS} - \frac{1}{\sqrt{K}}\sum_i \sum_{j\neq i} G_{ij}R_i R_j \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, I_3\right). \tag{1.7}$$

In Theorem 1.1, $E[T_{FS}] = \frac{1}{\sqrt{K}}\sum_i \sum_{j\neq i} G_{ij}R_i R_j$ is the concentration parameter corresponding to the instrument strength. In the model of Section 1.2, the mapping to the reduced-form π can be found in Section 1.A.2, so the concentration parameter is given by $E[T_{FS}] = \frac{1}{\sqrt{K}}\sum_k (c-1)\pi_k^2 = \frac{5}{8}\sqrt{K}(c-1)s^2$.[11] If the instruments are strong, then $E[T_{FS}] \to \infty$, so $\hat{\beta}_{JIVE} - \beta_{JIVE} \xrightarrow{d} 0$. With weak IV, $E[T_{FS}]$ converges to some constant $C < \infty$, so comparing the JIVE t-statistic with the standard normal distribution leads to invalid inference even in large samples.

The asymptotic distribution follows from establishing a quadratic CLT that may be of independent interest: it is proven by rewriting the leave-one-out sums as a martingale difference array, and then applying the martingale CLT. While there are existing quadratic CLT available, they do not fit the context exactly. Chao et al. (2012) Lemma A2 requires G to be symmetric, which works for $G = P$, but G for UJIVE is not symmetric in general. EK18 Lemma D2 is established for scalar random variables, so I extend it to random vectors.

[11] This concentration parameter is comparable to the concentration parameter in just-identified IV. With slight abuse of notation, suppose the just-identified IV model has a first stage equation with $X = Z\pi + v$ where $\pi = s$. Then, omitting variance normalizations, using the notation from Lee et al. (2023), the concentration parameter is $f_0 = \sqrt{n}s$, which determines if the TSLS estimator is consistent. In the L1O asymptotics, $\sqrt{K}(c-1)s^2 \approx ns^2/\sqrt{K}$ by using $n = (K+1)c$ and approximations $\sqrt{K/(K+1)} \approx 1$ and $(c-1)/\sqrt{c} \approx \sqrt{c}$. By comparing the L1O concentration parameter $ns^2/\sqrt{K}$ with the just-identified IV concentration parameter $f_0 = \sqrt{n}s$, I obtain the notions of weak identification in Remark 1.1.

Theorem 1.1 states that T_{LM} is mean zero and asymptotically normal. Hence, if we have access to the oracle variance of T_{LM}, we can simply use the statistic $T_{LM}/\sqrt{\text{Var}(T_{LM})}$ for testing because it has a standard normal distribution under the null. Obtaining a consistent estimator is an issue addressed in the next subsection.

1.3.2 Variance Estimation

To test the null that $H_0 : \beta = \beta_0$, we can calculate T_{LM} using the null-imposed β_0 and an estimator for the variance of $\sqrt{K}T_{LM}$, $\hat{V}_{LM}$, defined later in this section. Then, reject if $KT_{LM}^2/\hat{V}_{LM} \geq \Phi\left(1-\alpha/2\right)^2$ for a size α test where $\Phi(.)$ is the standard normal CDF. This procedure is valid when T_{LM} is asymptotically normal with mean zero as we have established in the previous section, and when $\hat{V}_{LM}$ is consistent.

Before stating the variance estimator, I first decompose the variance expression in the equation below, which follows from substituting $e_i = R_{\Delta i} + \nu_i$ and $X_i = R_i + \eta_i$ into the variance. For $V_{LM} := \text{Var}\left(\sum_i \sum_{j\neq i} G_{ij} e_i X_j\right)$,

$$\begin{aligned} V_{LM} = &\sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\nu_i^2\right] G_{ij} G_{ik} R_j R_k + \sum_i \sum_{j\neq i} G_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] + \sum_i \sum_{j\neq i} G_{ij} G_{ji} E\left[\eta_i \nu_i\right] E\left[\eta_j \nu_j\right] \\ &+ 2\sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\nu_i \eta_i\right] G_{ij} G_{ki} R_j R_{\Delta k} + \sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\eta_i^2\right] G_{ji} G_{ki} R_{\Delta j} R_{\Delta k}. \end{aligned} \tag{1.8}$$

With constant treatment effects, only the first line appears in the variance as $R_\Delta = 0$. With $G = P$, the expression for $\text{Var}\left(\sum_i \sum_{j\neq i} P_{ij} e_i X_j\right)$ matches the expression in EK18 Theorem 5.3, but their variance estimator cannot be used directly as they required consistent estimation of reduced-form coefficients. By adapting the leave-three-out (L3O) approach of Anatolyev and Sølvsten (2023) (AS23), an unbiased and consistent variance estimator can be obtained. Intuitively, just as the own-observation bias in TSLS that involves a single sum can be addressed with L1O, an unbiased estimator for the variance expression that involves

a triple sum can be obtained with L3O. Let $\tau := (\pi', \gamma')'$ and $\tau_\Delta := ((\pi_Y - \pi\beta)', (\gamma_Y - \gamma\beta)')'$ denote the coefficients on Q when running the regression of X and e respectively. The variance estimator is:

$$\hat{V}_{LM} := A_1 + A_2 + A_3 + A_4 + A_5, \tag{1.9}$$

with

$$\begin{aligned}
A_1 &:= \sum_i \sum_{j \neq i} \sum_{k \neq i} G_{ij} X_j G_{ik} X_k e_i(\beta_0) \left(e_i(\beta_0) - Q_i' \hat{\tau}_{\Delta,-ijk}\right), \\
A_2 &:= 2 \sum_i \sum_{j \neq i} \sum_{k \neq i} G_{ij} X_j G_{ki} e_k(\beta_0) e_i(\beta_0) \left(X_i - Q_i' \hat{\tau}_{-ijk}\right), \\
A_3 &:= \sum_i \sum_{j \neq i} \sum_{k \neq i} G_{ji} e_j(\beta_0) G_{ki} e_k(\beta_0) X_i \left(X_i - Q_i' \hat{\tau}_{-ijk}\right), \\
A_4 &:= -\sum_i \sum_{j \neq i} \sum_{k \neq j} G_{ji}^2 X_i \check{M}_{ik,-ij} X_k e_j(\beta_0) \left(e_j(\beta_0) - Q_j' \hat{\tau}_{\Delta,-ijk}\right), \\
A_5 &:= -\sum_i \sum_{j \neq i} \sum_{k \neq j} G_{ij} G_{ji} e_i(\beta_0) \check{M}_{ik,-ij} X_k e_j(\beta_0) \left(X_j - Q_j' \hat{\tau}_{-ijk}\right),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\tau}_{-ijk} &:= \left(\sum_{l \neq i,j,k} Q_l Q_l'\right)^{-1} \sum_{l \neq i,j,k} Q_l X_l, \\
\hat{\tau}_{\Delta,-ijk} &:= \left(\sum_{l \neq i,j,k} Q_l Q_l'\right)^{-1} \sum_{l \neq i,j,k} Q_l e_l(\beta_0), \\
D_{ij} &:= M_{ii} M_{jj} - M_{ij}^2, \text{ and} \\
\check{M}_{ik,-ij} &:= \frac{M_{jj} M_{ik} - M_{ij} M_{jk}}{D_{ij}} = -Q_i' \left(\sum_{l \neq i,j} Q_l Q_l'\right)^{-1} Q_k.
\end{aligned}$$

Following AS23, I make an assumption to ensure that the L3O estimator is well-defined.[12]

[12] If these conditions are not satisfied, then we can follow the modification in AS23 so that the variance estimator is conservative.

Assumption 1.2. *(a)* $\sum_{l\neq i,j,k} Q_l Q_l'$ *is invertible for every* $i,j,k \in \{1,\cdots,n\}$.

(b) $\max_{i\neq j\neq k\neq i} D_{ijk}^{-1} = O_P(1)$, *where* $D_{ijk} := M_{ii}D_{jk} - \left(M_{jj}M_{ik}^2 + M_{kk}M_{ij}^2 - 2M_{jk}M_{ij}M_{ik}\right)$.

Assumption 1.2(a) corresponds to AS23 Assumption 1 and Assumption 1.2(b) corresponds to AS23 Assumption 4. For consistent variance estimation, we additionally require regularity conditions that are stated in Assumption 1.3 of Section 1.A.1. These conditions are satisfied when G is a projection matrix. With these conditions, Theorem 1.2 below claims that the variance estimator is consistent.

Theorem 1.2. *If* $\beta = \beta_0$, *Assumptions 1.1-1.2 hold, and Assumption 1.3 in Section 1.A.1 holds, then* $E\left[\hat{V}_{LM}\right] = V_{LM}$ *and* $\hat{V}_{LM}/V_{LM} \xrightarrow{p} 1$.

With many instruments and potentially many covariates, the reduced-form coefficients $\pi, \pi_Y, \gamma, \gamma_Y$ are not consistently estimable. The usual approach to constructing variance estimators calculates residuals by using the estimated coefficients, but this approach no longer works when these estimated coefficients are inconsistent. To be precise, applying Chebyshev's inequality for any $\epsilon > 0$ yields:

$$\Pr\left(\left|\frac{\hat{V}_{LM} - V_{LM}}{V_{LM}}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\frac{Var\left(\hat{V}_{LM}\right)}{V_{LM}^2} + \frac{1}{\epsilon^2}\frac{\left(E\left[\hat{V}_{LM}\right] - V_{LM}\right)^2}{V_{LM}^2}. \tag{1.10}$$

Without an unbiased estimator and when reduced-form coefficients cannot be consistently estimated, the second term in (1.10) is not necessarily asymptotically negligible. To overcome this problem, I use an unbiased variance estimator so that the second term is exactly zero. Then, it suffices to show that the variance of individual components of the variance are asymptotically small compared to V_{LM}^2, so that the first term in (1.10) is $o(1)$ by applying the Cauchy-Schwarz inequality.

To obtain an unbiased estimator, I use estimators for the reduced-form coefficients $\pi, \pi_Y, \gamma, \gamma_Y$ that are unbiased and independent of objects that they are multiplied with.

The leave-three-out (L3O) approach has this unbiasedness property for linear regressions: when leaving three observations out in the inner-most sum of the A expressions, the estimated coefficient $\hat{\tau}_{-ijk}$ is independent of i, j, k and is unbiased for τ. Then, when taking the expectation through a product of random variables of i, j, k and $\hat{\tau}_{-ijk}$, τ can be used in place of the $\hat{\tau}_{-ijk}$ component, and the expectations of individual components can be isolated. For instance,

$$\begin{aligned} & E\left[\sum_i \sum_{j \neq i} \sum_{k \neq i,j} G_{ij} X_j G_{ik} X_k e_i \left(e_i - Q_i' \hat{\tau}_{\Delta,-ijk}\right)\right] \\ &= \sum_i \sum_{j \neq i} \sum_{k \neq i,j} G_{ij} E\left[X_j\right] G_{ik} E\left[X_k\right] E\left[e_i \left(e_i - Q_i' \hat{\tau}_{\Delta,-ijk}\right)\right] \\ &= \sum_i \sum_{j \neq i} \sum_{k \neq i,j} G_{ij} R_j G_{ik} R_k E\left[\nu_i^2\right], \end{aligned} \tag{1.11}$$

which recovers the triple sums in the V_{LM} expression of (1.8). Without leaving out observations j and k, we would not be able to isolate $E[X_j]$ and $E[X_k]$ in the first equality. Without leaving out observation i, we would not be able to isolate τ_Δ on expectation to obtain $E[\nu_i^2]$ in the second equality. An analogous argument applies to other components of V in (1.7). Assuming that the residuals have zero mean conditional on Q is crucial: if we merely have $E[Q\zeta] = 0$, this argument can no longer be applied.

Remark 1.3. *While the proposed $\hat{V}_{LM}$ is motivated by AS23, the contexts and estimators are different. First, the statistic that we are estimating the variance for is different: AS23 demeaned their $\mathcal{F}$ statistic using $\hat{E}_{\mathcal{F}}$, where $\hat{E}_{\mathcal{F}}$ is estimated using L1O, so they are interested in the variance of $\mathcal{F} - \hat{E}_{\mathcal{F}}$ that is mean zero; I use a mean-zero L1O statistic directly in T_{LM}. Second, the expectation of their variance estimator takes the form of their (9), which is analogous to the sum of A_1 and A_4 using the notation above, so repeated applications of their estimator is insufficient to recover all five terms exactly. Hence, to adjust for the*

A_4 and A_5 terms here, I additionally require another estimator, and its form is similarly motivated by a L3O reasoning.

Inverting the test to obtain a confidence set is straightforward, as the test statistic T_{LM}^2 and variance estimator $\hat{V}_{LM} = B_0 + B_1\beta_0 + B_2\beta_0^2$ are quadratic in β_0 (for some B_0, B_1, B_2 that are functions of the data), so the confidence set is obtained by solving a quadratic inequality.[13]

1.4 Power Properties

This section characterizes power properties of the valid LM procedure. I first argue that we can restrict our attention to three statistics that are jointly normal by extending the argument from Moreira (2009a). Since the covariance matrix can be consistently estimated, the remainder of the section focuses on the 3-variable normal distribution with a known covariance matrix. With this asymptotic distribution, I show that the two-sided LM test is the uniformly most powerful unbiased test within the interior of the parameter space.

1.4.1 Sufficient Statistics and Maximal Invariant

As is standard in the literature, I consider the canonical model without covariates where the reduced-form errors are normal and homoskedastic (e.g., Andrews et al. (2006); Moreira (2009a)). Suppose (η, ζ) in the model of Section 1.3.1 are jointly normal with known variance:

$$\begin{pmatrix} \zeta_i \\ \eta_i \end{pmatrix} \sim N(0, \Omega) = N\left(0, \begin{bmatrix} \omega_{\zeta\zeta} & \omega_{\zeta\eta} \\ \omega_{\zeta\eta} & \omega_{\eta\eta} \end{bmatrix}\right). \tag{1.12}$$

[13] Details are relegated to the appendix.

Define:

$$\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} := \begin{pmatrix} (Z'Z)^{-1/2} Z'Y \\ (Z'Z)^{-1/2} Z'X \end{pmatrix}.$$

I restrict attention to tests that are invariant to rotations of Z, i.e., transformations of the form $Z \to ZF'$ where F is a $K \times K$ orthogonal matrix. In particular, an invariant test $\phi(s_1, s_2)$ is one for which $\phi(Fs_1, Fs_2) = \phi(s_1, s_2)$ for all $K \times K$ orthogonal matrices F. If we focus on invariant tests, then the maximal invariant contains all relevant information from the data for inference.

Due to Moreira (2009a) Proposition 4.1, $(s_1', s_2')'$ are sufficient statistics for $(\pi_Y', \pi')'$. Further, $(s_1's_1, s_1's_2, s_2's_2)$ is a maximal invariant, and

$$\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \sim N\left(\begin{pmatrix} (Z'Z)^{1/2} \pi_Y \\ (Z'Z)^{1/2} \pi \end{pmatrix}, \Omega \otimes I_K \right).$$

The maximal invariant $(s_1's_1, s_1's_2, s_2's_2)$ is jointly normal with a mean that depends on Ω when $K \to \infty$.[14] Extending the argument to allow for heterogeneous treatment effects, the object of interest is $\beta = \frac{\pi' Z' Z \pi_Y}{\pi' Z' Z \pi}$ (following EK18), which is invariant to rotations of the instrument.[15]

To be robust to many instruments, heteroskedasticity, and non-normality, I use the leave-one-out (L1O) analog of the maximal invariant (following MS22; Lim et al. (2024)).[16] Without covariates such that $G = P$, the L1O analog is $\frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij}(Y_iY_j, Y_iX_j, X_iX_j)$, which is a linear transformation of (T_{AR}, T_{LM}, T_{FS}).[17] In the remainder of this section, I focus on

[14] This result is stated in Section 1.C.2.

[15] With rotation matrix F' such that $F'F = I$, observe that $X = ZF'F\pi + \eta$, so if we were to run the regression on ZF' instead of Z, we would obtain coefficients $F\pi$ instead of π. Then, the estimand is $\frac{\pi' F' F Z' Z F' F \pi_Y}{\pi' F' F Z' Z F' F \pi} = \frac{\pi' Z' Z \pi_Y}{\pi' Z' Z \pi}$ as before.

[16] With heteroskedasticity, the variances are not consistently estimable, so we cannot correct for the variances directly. These variances no longer feature in the L1O analog of the maximal invariant.

[17] To see that $\frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij}(Y_iY_j, Y_iX_j, X_iX_j)$ is a linear transformation, use the fact that $e = Y + X\beta$. Then, $\frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij}((e_i + X_i\beta)(e_j + X_j\beta), (e_i + X_i\beta)X_j, X_iX_j) = (T_{AR} + 2T_{LM}\beta + T_{FS}\beta^2, T_{LM} - T_{FS}\beta, T_{FS})$.

testing the null that $\beta_0 = 0$ so $e(\beta_0) = Y$ and the L1O of the maximal invariant is exactly (T_{AR}, T_{LM}, T_{FS}). The results are generalized in the appendix.

The asymptotic problem involving (T_{AR}, T_{LM}, T_{FS}) is:

$$\begin{pmatrix} T_{AR} \\ T_{LM} \\ T_{FS} \end{pmatrix} \sim N(\mu, \Sigma), \mu = \begin{pmatrix} \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij} R_{\Delta i} R_{\Delta j} \\ \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij} R_{\Delta i} R_j \\ \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij} R_i R_j \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \cdot & \sigma_{22} & \sigma_{23} \\ \cdot & \cdot & \sigma_{33} \end{pmatrix}. \tag{1.13}$$

While $\mu_2 = 0$ under the null, μ_2 may not be zero under the alternative. There are several restrictions in the μ vector, which is assumed to be finite. Since P is a projection matrix, $\sum_i \sum_{j \neq i} P_{ij} R_i R_j = \sum_i R_i (\sum_j P_{ij} R_j - P_{ii} R_i) = \sum_i M_{ii} R_i^2$. Since the annihilator matrix M has positive entries on its diagonal, we obtain $\mu_3 \geq 0$ and a similar argument yields $\mu_1 \geq 0$. With $\mu_2 = \sum_i \sum_{j \neq i} P_{ij} R_{\Delta i} R_j = \sum_i M_{ii} R_{\Delta i} R_i$, the Cauchy-Schwarz inequality implies $\mu_2^2 \leq \mu_1 \mu_3$. Constant treatment effects implies $\mu_2^2 = \mu_1 \mu_3$, which is a special case of the environment here. Even with covariates, if the regression is fully saturated with G given by UJIVE, the same inequality restrictions hold.[18] These properties do not contradict the joint normality: even though $\mu_3 \geq 0$, T_{FS} can still be negative when using the L1O statistic. The inequalities $\mu_1, \mu_3 \geq 0$ and $\mu_2^2 \leq \mu_1 \mu_3$ are also the only restrictions on μ, as it can be shown that there exists a structural model where there are no further restrictions.[19]

1.4.2 Optimality Result

With a size α test, the two-sided LM test against the alternative that $\mu_2 \neq 0$ rejects when $T_{LM}^2 / \text{Var}(T_{LM}) > \Phi(1 - \alpha/2)^2$. I consider the benchmark of a uniformly most powerful unbiased test (e.g., Lehmann and Romano (2005); Moreira (2009b)).

[18] See Proposition 1.3 in Section 1.C.2.

[19] Section 1.C.2 establishes that there exists a structural model where Σ is uninformative about μ, and $\mu_1, \mu_3 \geq 0$. Since the model in Section 1.2 is binary, it is insufficient for such a general result, and a continuous X is required. While the result establishes that there exists a structural model where there are no further restrictions, for any given structural model, there can still be further restrictions.

Proposition 1.1. *Consider a restriction of the alternative μ space to the interior i.e., $\mu_1, \mu_3 > 0$ and $\mu_2^2 < \mu_1\mu_3$. Then, within the class of tests that are functions of (T_{AR}, T_{LM}, T_{FS}), the two-sided LM test is the uniformly most powerful unbiased test for testing $H_0 : \mu_2 = 0$ against $H_1 : \mu_2 \neq 0$ in the asymptotic problem of (1.13).*

The argument for optimality applies a standard optimality result from Lehmann and Romano (2005) on the exponential family, which includes the normal distribution. To apply the Lehmann and Romano (2005) result, we require a convex parameter space and the the existence of alternative values above and below the null value. It can be verified that the restricted parameter space is still convex, and the restriction to the interior ensures the latter condition is satisfied. The proposition claims optimality within the class of unbiased tests, and makes no statement about tests that are biased (i.e., where the power somewhere in the alternative space can be lower than the size).

Remark 1.4. *With the characterized asymptotic distribution, there are several other tests that are valid. (1) We can implement a Bonferroni-type correction that constructs a 99% confidence set for both μ_1 and μ_3, then a 97% test for LM. (2) VtF from Lee et al. (2023) can be adapted, because the asymptotic distribution does not rely on homogeneous treatment effects and the JIVE t statistic has the same distribution as the just-identified TSLS t statistic. (3) With a given structural model, the algorithm from Elliott et al. (2015) can also be applied by using a grid on structural parameters.*

Remark 1.5. *Beyond the two-sided UMPU result, we may also consider other power properties. The one-sided LM test is shown to be the most powerful test against a particular subset of the alternative space. Numerically, using a covariance matrix calibrated from an empirical application, the power of the two-sided LM test is also close to that of the nearly optimal test against a weighted average over a grid of alternative values, constructed using the algorithm from Elliott et al. (2015). Details are in Section 1.C.2.*

Studying optimality in the over-identified IV environment has thus far been complicated. With constant treatment effects, both $s_1's_1$ and $s_1's_2$ are informative of the object of interest β, because constant treatment effects implies $\mu_1 = \beta^2\mu_3$ in addition to $\mu_2 = \beta\mu_3$. However, once we impose $\mu_1 > 0$ under the null that $\beta = 0$, we rule out constant treatment effects by focusing on the interior of the alternative space. Then, the statistic associated with μ_1 is no longer directly informative of β. Imposing heterogeneity is hence the key to obtaining this UMPU result.

1.5 Simulations

This section focuses on the simple example from Section 1.2. I report two sets of simulations that assess the size and one that assesses power. One set of size simulations uses a large K while the other a small K. Robustness checks that involve different data generating processes are relegated to the appendix.[20]

Table 1.1 in Section 1.2 reports rejection rates under the null for a relatively large number of judges with $K = 400$, each with a small number of cases at $c = 5$. L3O performs well across various designs, while existing procedures can substantially over-reject in at least one design. The LMorc column is included as an infeasible theoretical benchmark that uses an oracle variance: this should have nominal size when normality holds because the variance is not estimated. The difference between LMorc and L3O is attributed to the variance estimation procedure.

Table 1.2 reports rejection rates under the null for a small number of judges with $K = 4$ and a large number of cases at $c = 200$. Based on the theory in Section 1.3, L3O should be valid when the instrument is strong, i.e., in the cases with $E[T_{FS}] = .5c$, which is what we observe. Notably, even when $E[T_{FS}] = 2$ or $E[T_{FS}] = 0$, the over-rejection for L3O is

[20]There are more simulation results using several different structural models in Section 1.C.4, including settings with continuous treatment X, and with covariates. The results are qualitatively similar in those simulations, suggesting that the numerical findings are not unique to the data-generating process chosen.

Table 1.2: Rejection rates under the null for nominal size 0.05 test

Designs		Procedures								
$E[T_{AR}]$	$E[T_{FS}]$	TSLS	EK	MS	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O	LMorc	ARorc
	.5c	0.210	0.049	1.000	0.212	0.217	0.217	0.048	0.052	1.000
.5c	2	0.642	0.018	1.000	0.806	0.269	0.816	0.043	0.049	1.000
	0	0.498	0.005	1.000	0.881	0.330	0.893	0.062	0.051	1.000
	.5c	0.075	0.064	1.000	0.075	0.073	0.077	0.063	0.061	0.913
2	2	0.462	0.013	0.999	0.436	0.296	0.516	0.095	0.049	0.931
	0	0.440	0.008	1.000	0.448	0.337	0.576	0.088	0.052	0.934
	.5c	0.052	0.048	0.061	0.045	0.050	0.046	0.046	0.048	0.069
0	2	0.376	0.088	0.075	0.044	0.238	0.123	0.101	0.045	0.071
	0	0.590	0.181	0.076	0.029	0.431	0.163	0.075	0.045	0.080

Notes: $K = 4, c = 200$. Designs and procedures are otherwise identical to Table 1.1.

not too severe. EK performs very well in the cases with $E[T_{FS}] = .5c$ as expected in their theory. In contrast, MS and MO can over-reject severely with strong heterogeneity, even when instruments are strong.

Table 1.3 reports rejection rates under the alternative. When $E[T_{FS}] = 0$, the instrument should be completely uninformative about the true parameter, so we should have 0.05 rejection rate for a valid test, which is what we observe for L3O. When $E[T_{FS}] = 2\sqrt{K}$, all procedures, including L3O, are very informative. Considering the designs with $E[T_{AR}] = 0$ is most interesting, because this is an environment where MS and MO are valid, and the theoretical optimality result excludes this case. Looking at the case with $E[T_{AR}] = 0, E[T_{FS}] = 2$, L3O is less powerful than MS and MO in small samples, but the loss is less than 7 percentage points.

Table 1.3: Rejection rates under the alternative for nominal size 0.05 test

Designs		Procedures								
$E[T_{AR}]$	$E[T_{FS}]$	TSLS	EK	MS	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O	LMorc	ARorc
$2\sqrt{K}$	$2\sqrt{K}$	1.000	1.000	NaN	1.000	1.000	1.000	1.000	1.000	1.000
	2	0.310	0.173	NaN	0.539	0.117	0.563	0.240	0.225	1.000
	0	0.722	0.029	NaN	0.291	0.055	0.309	0.048	0.055	1.000
2	$2\sqrt{K}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	2	0.244	0.543	1.000	0.886	0.163	0.907	0.789	0.823	1.000
	0	0.998	0.090	1.000	0.157	0.169	0.221	0.073	0.054	1.000
0	$2\sqrt{K}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	2	0.259	0.662	1.000	0.967	0.230	0.978	0.936	0.961	1.000
	0	1.000	0.373	0.059	0.048	0.333	0.117	0.067	0.055	0.054

Notes: $K = 100, \beta = 0.1, c = 5$. Designs and procedures are otherwise identical to Table 1.1.

1.6 Empirical Applications

1.6.1 Returns to Education

Angrist and Krueger (1991) were interested in the impact of years of education (X) on log weekly wages (Y). They instrument for education using the quarter of birth (QOB). I implement UJIVE using full interaction of QOB with the state of birth and year of birth (resulting in 1530 instruments) without other controls, which is similar to Table VII(2) of Angrist and Krueger (1991) that uses the same set of controls but without full saturation. The implementation here differs from the implementation of MS and MO in that I do not linearly partial out other covariates, but merely saturate on state and year of birth. This implementation is motivated by recent econometric research (e.g., Blandhol et al. (2022); Słoczyński (2020)) that argue that the standard interpretation of estimands as a weighted average of LATE's is only retained with some parametric assumptions or when the specification controls for covariates richly, which can be achieved with full saturation.[21] To ensure that the different

[21] A further advantage of this implementation is that the code is fast: when G is block-diagonal, it suffices to loop over blocks.

procedures are directly comparable, I adapt the MS and MO inference procedures to target UJIVE, so that the estimand is the same across all procedures and differing results can be attributed purely to inference. TSLS is consequently not a meaningful comparison as the estimand is different from the others.

The results are reported in Table 1.4. In addition to the aforementioned procedures, I include results from implementing the procedure in Crudu et al. (2021) (CMS) that uses the T_{AR} statistic like MS22, but uses a plug-in variance estimator like MO22.[22] Being robust to weak and many IV results in L3O having a longer confidence interval than EK. With full saturation, CMS and MO yield unbounded confidence sets, while L3O yields a bounded confidence set, showing how robustness to heterogeneity changes the shape of the confidence set in this context. The shape of the confidence set depends on the coefficient on β_0^2. In particular, for $\Psi_2 := \frac{1}{K}\sum_i \left(\sum_{j\neq i} G_{ij}X_j\right)^2 X_i^2 + \frac{1}{K}\sum_{i\neq j}^n G_{ij}^2 X_i^2 X_j^2$, MO is unbounded when $T_{FS}^2 - q\Psi_2 < 0$ and L3O is unbounded when $T_{FS}^2 - qB_2 < 0$, where q is 3.84 for a 5% test and B_2 is the coefficient on β_0^2 in the expression of $\hat{V}_{LM}$. Consequently, in this application, we can think of $T_{FS}^2/\Psi_2 = 0.102$ and $T_{FS}^2/B_2 = 11.8$ as first-stage statistics for MO and L3O respectively that determine whether the confidence sets are bounded.[23] Analogously, when solving a quartic equation in CMS, an unbounded set occurs as $T_{FS}^2/\left(\frac{2}{K}\sum_{i\neq j}^n G_{ij}^2 X_i^2 X_j^2\right) = 0.0545$, where the denominator is their coefficient on β_0^4 in their variance estimator. In contrast, the MS confidence set is bounded with $T_{FS}^2/\left(\frac{2}{K}\sum_{i\neq j}^n \frac{G_{ij}^2}{M_{ii}M_{jj}+M_{ij}^2} X_i^2 X_j^2\right) = 23.9$.

Due to the $\check{M}$ terms in the L3O expression, it is difficult to compare the estimates directly. However, it is possible to compare the estimands of these coefficients in the judge example

[22] MS22 use a cross-fit variance estimator, while they refer to the CMS variance estimator as the "naive" variance estimator. MS22 argue that their cross-fit variance is more powerful, which corroborates how MS has a bounded confidence set while CMS does not.

[23] These statistics are "F" statistics with different variance estimators, suggesting that the instruments are meaningfully weak. The MS and MO variance estimators converge to the same object under weak identification such that $\frac{1}{K}\sum_i \sum_{j\neq i} G_{ij}R_iR_j \to 0$, which is not imposed by the asymptotic regime in this paper.

Table 1.4: 95% Confidence Sets for Returns to Education

	EK	CMS	MS	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O
LB	0.033	-∞	0.019	-∞	0.027	0.027	0.022
UB	0.173	∞	0.305	∞	0.179	0.189	0.210
Estimate	0.103	0.103	0.103	0.103	0.103	0.103	0.103
CIlength	0.140	∞	0.286	∞	0.152	0.161	0.188

Notes: Estimate reports the UJIVE. CMS implements the procedure from Crudu et al. (2021): use T_{AR} with a plug-in variance. Procedures are otherwise identical to Table 1.1.

without covariates:

$$E\left[K\Psi_2\right] - E\left[B_2\right] = \sum_i M_{ii} R_i^2 \left(R_i^2 - 3\left(1 - 2P_{ii}\right) E\left[\eta_i^2\right]\right). \tag{1.14}$$

If $R_i^2 > 3\left(1 - 2P_{ii}\right) E\left[\eta_i^2\right]$, then MO is more likely unbounded. Intuitively, the MO estimator contains additional products of R that are not present in the true variance, and there are products of R and the error present in the true variance that MO does not account for, motivating the aforementioned difference. We can interpret this condition as MO being more likely unbounded when the signal-to-noise ratio $R_i^2/E\left[\eta_i^2\right]$ is sufficiently large. In this application, by observing that the first-stage statistic of L3O is an order of magnitude larger than that of MO (i.e., B_2 is an order of magnitude smaller), and by comparing the CMS and MO first-stage statistics, there is evidence that $R_i^2/E\left[\eta_i^2\right]$ is large.[24] This result does not depend on heterogeneity, because the coefficient of β_0^2 depends only on how X is combined in the variance estimator.

[24]Comparing the statistics between MO and MS implies $\frac{1}{K}\sum_i \left(\sum_{j\neq i} G_{ij}X_j\right)^2 X_i^2 < \frac{1}{K}\sum_i \sum_{j\neq i} G_{ij}^2 X_i^2 X_j^2$, which can equivalently be written as $\sum_i \sum_{j\neq i} \sum_{k\neq i,j} G_{ij}G_{ik}X_jX_kX_i^2 < 0$. With full saturation, observations i, j have $G_{ij} < 0$ when they are in the same covariate group but have different instrument values. Under the MS and MO asymptotic regimes where $\frac{1}{K}\sum_i \sum_{j\neq i} G_{ij}R_iR_j \to 0$ so $R_i^2/E[\eta_i^2]$ is negligible, we obtain $\sum_i \sum_{j\neq i} G_{ij}^2 E[X_i^2]E[X_j^2] = \sum_i \sum_{j\neq i} G_{ij}^2 \left(R_i^2 + E[\eta_i^2]\right)\left(R_j^2 + E[\eta_j^2]\right) = \sum_i \sum_{j\neq i} G_{ij}^2 E[\eta_i^2]E[\eta_j^2] + o(1)$, and $\frac{1}{K}\sum_i \sum_{j\neq i} \sum_{k\neq i,j} G_{ij}G_{ik}E\left[X_jX_kX_i^2\right] = \frac{1}{K}\sum_i \sum_{j\neq i} \sum_{k\neq i,j} G_{ij}G_{ik}R_jR_k\left(R_i^2 + E[\eta_i^2]\right) = o(1)$ is asymptotically negligible. Since the difference between MS and MO is the same magnitude as the MS statistic, $\frac{1}{K}\sum_i \sum_{j\neq i} \sum_{k\neq i,j} G_{ij}G_{ik}E\left[X_jX_kX_i^2\right]$ is of similar order as $\sum_i \sum_{j\neq i} G_{ij}^2 E[X_i^2]E[X_j^2]$, so $R_i^2/E[\eta_i^2]$ is non-negligible in this application.

Table 1.5: 95% Confidence Sets for Misdemeanor Prosecution

	EK	CMS	MS	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O
LB	-0.151	$\emptyset$	$\emptyset$	-0.220	-0.187	-0.188	-0.201
UB	-0.076	$\emptyset$	$\emptyset$	-0.019	-0.039	-0.038	-0.028
Estimate	-0.113	-0.113	-0.113	-0.113	-0.113	-0.113	-0.113
CIlength	0.075	$\emptyset$	$\emptyset$	0.201	0.148	0.150	0.173

Notes: Procedures are identical to Table 1.4.

1.6.2 Misdemeanor Prosecution

Agan et al. (2023) were interested in the effect of misdemeanor prosecution (X) on criminal complaint in two years (Y). They instrument for misdemeanor prosecution using the assistant district attorneys (ADAs) who decide if a case should be prosecuted in the Suffolk County District Attorney's Office in Massachusetts. As Agan et al. (2023) argued that as-if randomization holds conditional on court-by-time controls and that individual covariates are not required for relevance or exogeneity to hold in this context, the confidence set is constructed using full saturation of court-by-year and court-by-day-of-week fixed effects with no other controls for individual covariates.

As reported in Table 1.5, with full saturation, the UJIVE is -0.11, so not prosecuting decreases the probability of criminal involvement by 11 percentage points.[25] The L3O confidence interval (CI) is more than twice that of EK: unlike Section 1.6.1 where $n/K = 221$, we have $n/K = 11.9$ here, so a variance estimator that is robust to many IV has a larger impact on CI. MS has an empty confidence set while L3O has a bounded set, showing how being robust to heterogeneity can change conclusions.[26] Mechanically, the confidence set for

[25] This result is smaller than -0.36 reported in their Table III(3) that uses TSLS with a leniency measure. The result is more similar to the UJIVE robustness check in their Table A.1(5) of -0.15 with full saturation of the instrument, but their specification includes case/ defendant covariates, which results in a different estimator.

[26] An empty confidence set using the AR procedure also suggests that the model with constant treatment effects is rejected, so there is meaningful heterogeneity. Since the variances of MO and L3O converge to the same object under homogeneity, the difference between MO and L3O confidence sets also suggests that there is heterogeneity.

MS solves a quartic equation, so an empty set can occur, but it is difficult to characterize when this phenomenon occurs in general.

The L3O CI is also shorter than MO, so being robust to heterogeneity decreases the length of the CI. Considering how the length of the MO confidence set is longer than L3O while being oversized in simulations, there is a question of when MO is conservative. While it is difficult to compare the confidence intervals or variance estimators directly, it is possible to compare the null-imposed variance estimands in the judge example without covariates. It can be shown that:

$$E\left[\hat{\Psi}_{MO}\right] - Var\left(\sum_i \sum_{j\neq i} P_{ij} e_i X_j\right) = \sum_i M_{ii} R_{\Delta i}^2 \left(R_i^2 - (1-2P_{ii})E\left[\eta_i^2\right]\right) - 2\sum_i M_{ii}(1-2P_{ii})E\left[\eta_i \nu_i\right] R_i R_{\Delta i}.$$

Then, MO is conservative when: (i) $R_i^2 > (1-2P_{ii})E\left[\eta_i^2\right]$, and (ii) $E\left[\eta_i \nu_i\right]$ is negatively correlated with $R_i R_{\Delta i}$, when $P_{ii} < 1/2$. In (i), $R_{\Delta i}^2$ only affects the magnitude of the difference, and not the sign, so this condition can be interpreted as a condition on the signal-to-noise ratio as before. Condition (ii) results from the $\sum_i M_{ii}(1-2P_{ii})E\left[\eta_i \nu_i\right] R_i R_{\Delta i}$ term that MO does not account for, and covariances can be positive or negative in general.

1.7 Conclusion

This paper has documented how weak instruments and heterogeneity can interact to invalidate existing procedures in the environment of many instruments. Addressing both problems simultaneously, this paper contributes a feasible and robust method for valid inference. The procedure is shown to be valid as the limiting distribution of commonly used statistics, including the LM statistic, in an environment with many weak instruments and heterogeneity, is normal, and a leave-three-out variance estimator is consistent for obtaining the variance

of the LM statistic. Beyond its validity, the LM test is also optimal, as it is the uniformly most powerful unbiased test in the asymptotic distribution for the interior of the alternative space. In light of the broader econometric literature on the value of saturated regressions and how many instruments can arise from them, this paper presents a highly applicable, robust, and powerful inference procedure for IV.

Appendix

1.A Supplement

1.A.1 High-level Assumptions for Inference

Following AS23, to ease notation in the L3O derivations, I define:

$$\check{M}_{il,-ijk} := \frac{M_{il} - M_{ij}\check{M}_{jl,-jk} - M_{ik}\check{M}_{kl,-jk}}{D_{ijk}/D_{jk}},$$

so that $X_i - Q_i'\hat{\tau}_{-ijk} = \sum_{l\neq k} \check{M}_{il,-ijk} X_l$, for instance.

Assumption 1.3 below states high-level conditions for consistency of the variance estimator. To ease notation, let R_{mi} stand for either $R_{\Delta i}$ or R_i. Denote $\tilde{R}_i := \sum_{j\neq i} G_{ij} R_j$ and $\tilde{R}_{\Delta i} := \sum_{j\neq i} G_{ij} R_{\Delta j}$. Let $h_2(i,j)$ be a product of any number of $G_{i_1 i_2}, i_1 \neq i_2$, and $\check{M}_{j_1 j_2}, j_1 \neq j_2$ with $i_1, i_2, j_1, j_2 \in \{i,j\}$. Similarly, $h_3(i,j)$ denotes a product of any number of $G_{i_1 i_2}, i_1 \neq i_2$, and $\check{M}_{j_1 j_2}, j_1 \neq j_2$ with $i_1, i_2, j_1, j_2 \in \{i,j,k\}$ such that every index in $\{i,j,k\}$ occurs at least once as an index of either $G_{i_1 i_2}$ or $\check{M}_{j_1 j_2}$. Let $h_4(i,j,k,l)$ denote a product of any number of $G_{i_1 i_2}, i_1 \neq i_2$ and $\check{M}_{j_1 j_2}, j_1 \neq j_2$ with $i_1, i_2, j_1, j_2 \in \{i,j,k,l\}$ such that every index in $\{i,j,k,l\}$ occurs at least once as an index of either $G_{i_1 i_2}$ or $\check{M}_{j_1 j_2}$, and there is no partition such that $h_4(i_1,i_2,j_1,j_2) = h_2(i_1,i_2)h_2(j_1,j_2)$, where i_1, i_2, j_1, j_2 are all different indices. For instance, $h_4(i,j,k,l)$ could be $G_{ij}\check{M}_{ik,-il}\check{M}_{lj,-ijk}$ but not $G_{ij}\check{M}_{lk,-il}$. Let $\sum_{i\neq j}^{n} = \sum_i \sum_{j\neq i}$ so that sums without the n superscript are still sums of individual indices,

but sums with an n superscript involves the sum over multiple indices. Objects like $\sum_{i\neq j\neq k}^{n}$ and $\sum_{i\neq j\neq k\neq l}^{n}$ are defined similarly. When I refer to the p-sum, I refer to the sum over p non-overlapping indices. For instance, a 3-sum is $\sum_{i\neq j\neq k}^{n}$. Let F stand for either G or G'. $1\{\cdot\}$ is an indicator function that takes the value 1 if the argument is true and 0 otherwise. $I\{\cdot\}$ is a function that takes value 1 if the argument is true and -1 if false.

Assumption 1.3. *For some $C < \infty$,*

(a) $\sum_j F_{ij}^2 \leq C$, $\sum_{j\neq k}^{n}\left(\sum_{i\neq j,k} G_{ij}F_{ik}\right)^2 \leq C\sum_{j\neq k}^{n} G_{jk}^2$, $\sum_{j\neq k}^{n}\left(\sum_{i\neq j,k} G_{ji}G_{ki}\right)^2 \leq C\sum_{j\neq k}^{n} G_{jk}^2$, *and* $|R_{mi}| \leq C$.

(b) $\sum_{i\neq j\neq k}^{n}\left(\sum_{l\neq i,j,k} h_4(i,j,k,l)\, R_{ml}\right)^2 \leq C\sum_i \tilde{R}_{mi}^2$, $\sum_{i\neq j}^{n}\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k} h_4(i,j,k,l)\, R_{ml}\right)^2 \leq C\sum_i \tilde{R}_{mi}^2$, *and* $\sum_i\left(\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} h_4(i,j,k,l)\, R_{ml}\right)^2 \leq C\sum_i \tilde{R}_{mi}^2$.

(c) $\sum_{i\neq j}^{n}\left(\sum_{k\neq i,j} h_3(i,j,k)\, R_{mk}\right)^2 \leq C\sum_i \tilde{R}_{mi}^2$ *and* $\sum_i\left(\sum_{j\neq i}\sum_{k\neq i,j} h_3(i,j,k)\, R_{mk}\right)^2 \leq C\sum_i \tilde{R}_{mi}^2$.

(d) $\sum_i\left(\sum_{j\neq i} h_2(i,j)\, R_{mj}\right)^2 \leq C\sum_i \tilde{R}_{mi}^2$.

The first condition requires the row and column sums of the squares of the G elements to be bounded. Assumption 1.1(e) is insufficient because it does not rule out having $G_{ii} = K$ for some i and 0 elsewhere in the G matrix. These remaining conditions can be interpreted as (approximate) sparsity conditions on M and G as the p-sum of entries of $\check{M}$ and G cannot be too large. The conditions primarily place a restriction on the types of G that can be used: for instance, a G matrix that contains all 1's is excluded. Note that other elements of the covariance matrix can be analogously shown to be consistent using the same strategy by using the lemmas from Section 1.B by using $\tilde{R}_{Yi} := \sum_{j\neq i} G_{ij}R_{Yj}$ instead of $\tilde{R}_{\Delta i}$ where required.

The judges example in Section 1.2 satisfies this assumption when there are no covariates, $G = P$, and R values are bounded. For condition (a), $\sum_j P_{ij}^2 = P_{ii} \leq C$ and, since P is idempotent, $\sum_{j\neq k}^{n}\left(\sum_{i\neq j,k} P_{ij}P_{ik}\right)^2 = \sum_{j\neq k}^{n}(\sum_i P_{ij}P_{ik} - P_{jj}P_{jk} - P_{kk}P_{jk})^2 =$

$\sum_{j\neq k}^{n}\left(P_{jk}-P_{jj}P_{jk}-P_{kk}P_{jk}\right)^2=\sum_{j\neq k}^{n}\left(1-P_{jj}-P_{kk}\right)^2P_{jk}^2\leq\sum_{j\neq k}^{n}P_{jk}^2$. For any $\check{M}_{ij}$ and G_{ij}, these elements are nonzero only when i and j share the same judge p. Further, $R_{mi}=\pi_{mp(i)}$, where π_{mp} can denote π_p or $\pi_{\Delta p}$ in the model. Due to how the h functions are defined, when every judge has at most c cases,

$$\sum_i\left(\sum_{j\neq i}h_2\left(i,j\right)R_{mj}\right)^2=\sum_i\left(\sum_{j\in\mathcal{N}_{p(i)}\backslash\{i\}}h_2\left(i,j\right)R_{mp(i)}\right)^2=\sum_p\sum_{i\in\mathcal{N}_p}\left(\sum_{j\in\mathcal{N}_p\backslash\{i\}}h_2\left(i,j\right)\pi_{mp}\right)^2$$
$$=\sum_p\sum_{i\in\mathcal{N}_p}\left(\sum_{j\in\mathcal{N}_p\backslash\{i\}}h_2\left(i,j\right)\pi_{mp}\right)^2\pi_{mp}^2\leq C\sum_p\sum_{i\in\mathcal{N}_p}\left(c-1\right)^2\pi_{mp}^2=C\sum_i\tilde{R}_{mi}^2.$$

The same argument applies to the other components. For instance,

$$\sum_i\left(\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k}h_4\left(i,j,k,l\right)R_{ml}\right)^2=\sum_p\pi_{mp}^2\sum_{i\in\mathcal{N}_p}\left(\sum_{j\in\mathcal{N}_p\backslash\{i\}}\sum_{k\in\mathcal{N}_p\backslash\{i,j\}}\sum_{l\in\mathcal{N}_p\backslash\{i,j,k\}}h_4\left(i,j,k,l\right)\right)^2$$
$$\leq C\sum_p\sum_{i\in\mathcal{N}_p}\pi_{mp}^2\left(c-1\right)^2\left(c-2\right)^2\left(c-3\right)^2\leq C\sum_i\tilde{R}_{mi}^2.$$

The upper bound is fairly loose because it merely counts the number of nonzero entries in h_4. When every judge has a large number of cases, since h_4 contains only entries from the projection matrix, the inner sum is still bounded and the assumption is satisfied.

1.A.2 Supplement for Section 1.2

Lemma 1.1. *Consider the model of Section 1.2. Suppose $h\neq 0$ and $Ks^2>0$. Then, $E\left[T_{AR}\right]\neq 0$ for all real β_0.*

Data Generating Process. Data is generated from an environment with $E[\varepsilon_i]=0$, and $\int_0^1 f(v)dv=\beta$. To run a regression on judge indicators (without an intercept) in the reduced-form system, I make a transformation $\check{X}=2X-1$ so that the reduced-form equations can

be written as:

$$\check{X}_i = Z_i'\pi + \eta_i, \text{ and } Y_i = Z_i'\pi_Y + \zeta_i,$$

so $\pi_k = \pi_{Yk} = 0$ for the base judge. The reduced-form errors are: $\eta_i = I\left\{\lambda_{k(i)} - v_i \geq 0\right\} - \pi_{k(i)}$ and $\zeta_i = 1\left\{\lambda_{k(i)} - v_i \geq 0\right\} f(v_i) + \varepsilon_i - \pi_{Yk(i)}$ respectively. With $\pi_{\Delta k} = \pi_{Yk} - \pi_k\beta$, the reduced-form parameters for the groups of judges are derived in Table 1.A.1. Since the coefficient of the base judge is normalized to zero, the implementation without covariates in simulations excludes the intercept and uses indicators for all judges, instead of omitting the base judge and having an intercept. This implementation results in a block diagonal projection matrix, which aids computational speed, while retaining the interpretation of π's in the reduced-form model. The $f(v)$ that delivers the parameters in Table 1.A.1 is

$$f(v) = \begin{cases} -s\beta + h & v \in [0, \frac{1}{2} - s] \\ \frac{1}{s}(1-s)\left(-\frac{1}{2}s\beta - h\right) - \frac{1}{s}(1-2s)(-s\beta + h) & v \in (\frac{1}{2} - s, \frac{1}{2} - \frac{1}{2}s] \\ \frac{1}{s}(1-s)\left(\frac{1}{2}s\beta + h\right) & v \in (\frac{1}{2} - \frac{1}{2}s, \frac{1}{2}] \\ \frac{1}{s}(1+s)\left(\frac{1}{2}s\beta - h\right) & v \in (\frac{1}{2}, \frac{1}{2} + \frac{1}{2}s] \\ \frac{1}{s}(1+2s)(s\beta + h) - \frac{1}{s}(1+s)\left(\frac{1}{2}s\beta - h\right) & v \in (\frac{1}{2} + \frac{1}{2}s, \frac{1}{2} + s] \\ \frac{\beta - \left(\frac{1}{2}+s\right)(s\beta+h)}{\frac{1}{2}-s} & v \in (\frac{1}{2} + s, 1] \end{cases}. \quad (1.15)$$

To generate the data in the simulation, I draw $v_i \sim U[0,1]$ as implied by the structural model, then generate $\zeta_i \mid v_i \sim N(\sigma_{\varepsilon v} v_i, \sigma_{\varepsilon\varepsilon})$. Hence, $\sigma_{\varepsilon v}$ and $\sigma_{\varepsilon\varepsilon}$ control the correlation between η_i and ζ_i, with $\sigma_{\varepsilon\varepsilon} = 0$ corresponding to perfect correlation. In the base case, I set $\sigma_{\varepsilon\varepsilon} = 0.1$ and $\sigma_{\varepsilon v} = 0.3$. With the given π_k, π_{Yk}, the observable variables are generated from $\check{X}_i = I\{\pi_{k(i)} > v_i\}$ and $Y_i = \pi_{Yk(i)} + \zeta_i$.

Table 1.A.1: Parameters for Simple Example

λ_k	$\frac{1}{2}-s$	$\frac{1}{2}-\frac{1}{2}s$	$\frac{1}{2}$	$\frac{1}{2}+\frac{1}{2}s$	$\frac{1}{2}+s$
β_k	$\beta-\frac{h}{s}$	$\beta+2\frac{h}{s}$	NA	$\beta-2\frac{h}{s}$	$\beta+\frac{h}{s}$
π_k	$-s$	$-\frac{1}{2}s$	0	$\frac{1}{2}s$	s
π_{Yk}	$-s\beta+h$	$-\frac{1}{2}s\beta-h$	0	$\frac{1}{2}s\beta-h$	$s\beta+h$
$\pi_{\Delta k}$	h	$-h$	0	$-h$	h

MS Variance Estimand. The proposed variance estimator is:

$$\hat{\Phi}_{MS} := \frac{2}{K}\sum_i \sum_{j\neq i} \frac{P_{ij}^2}{M_{ii}M_{jj}+M_{ij}^2}(e_i M_{i\cdot} e)(e_j M_{j\cdot} e).$$

By substituting $e_i = R_{\Delta i} + \nu_i$ and taking expectations,

$$E[e_i M_{i\cdot} e e_j M_{j\cdot} e] = R_{\Delta i}R_{\Delta j}\left(\sum_k M_{ik}M_{jk}E[\nu_k^2]\right) + (M_{ii}M_{jj}+M_{ij}^2)\left(E[\nu_i^2]E[\nu_j^2]\right).$$

This estimand can be positive or negative, but observe that $\sum_i \sum_{j\neq i} R_{\Delta i}R_{\Delta j} = (\sum_i R_{\Delta i})^2 - \sum_i R_{\Delta i}^2 = -\sum_i R_{\Delta i}^2 = -(n-c)h^2$ in the model of Section 1.2.1. Consequently, the negative heterogeneity component can far outweigh the positive components, resulting in a negative estimand when h does not converge to 0.

1.B Main Proofs

A quadratic CLT is used for Theorem 1.1. Let

$$T = \sum_i s_i' v_i + \sum_i \sum_{j\neq i} G_{ij} v_i' A v_j,$$

where v_i is a finite-dimensional random vector independent over $i = 1, \ldots, n$ with bounded 4th moments, s_i is a nonstochastic vector, and A is a conformable matrix.

Lemma 1.2. *Suppose:*

1. $\text{Var}\,(T)^{-1/2}$ *is bounded;*
2. $\sum_i s_{il}^4 \to 0$*; and*
3. $||G_L G_L'||_F + ||G_U G_U'||_F \to 0$*, where* G_L *is a lower-triangular matrix with elements* $G_{L,ij} = G_{ij} 1\{i > j\}$ *and* G_U *is an upper-triangular matrix with elements* $G_{U,ij} = G_{ij} 1\{i < j\}$*.*

Then, $\text{Var}\,(T)^{-1/2} T \xrightarrow{d} N(0,1)$*.*

Proof of Theorem 1.1. By substituting $Y_i = R_{Yi} + \zeta_i$, $X_i = R_i + \eta_i$, $\zeta_i = \nu_i + \beta\eta_i$ and $R_{Yi} = R_{\Delta i} - R_i\beta$ into the expression for $\hat{\beta}_{JIVE}$,

$$\hat{\beta}_{JIVE} - \beta_{JIVE} = \frac{\left(\sum_i \sum_{j\neq i} G_{ij} \left(R_{\Delta i}\eta_j + \nu_i R_j + \nu_i \eta_j\right)\right)}{\sum_i \sum_{j\neq i} G_{ij} R_i R_j + \sum_i \sum_{j\neq i} G_{ij} \left(R_i \eta_j + R_j \eta_i + \eta_i \eta_j\right)}.$$

To see the equivalence with the T objects,

$$\frac{1}{\sqrt{K}} \sum_i \sum_{j\neq i} G_{ij} e_i X_j = \frac{1}{\sqrt{K}} \sum_i \sum_{j\neq i} G_{ij} \left(\nu_i R_j + \nu_i \eta_j + R_{\Delta i} R_j + R_{\Delta i} \eta_j\right), \text{ and}$$

$$\sum_i \sum_{j\neq i} G_{ij} R_{\Delta i} R_j = \sum_i \sum_{j\neq i} G_{ij} R_{Yi} R_j - \sum_i \sum_{j\neq i} G_{ij} R_i R_j \left(\frac{\sum_i \sum_{j\neq i} G_{ij} R_{Yi} R_j}{\sum_i \sum_{j\neq i} G_{ij} R_i R_j}\right) = 0,$$

while T_{FS} is immediate.

Next, I show that the joint distribution of $\sqrt{\frac{K}{r_n}} (T_{AR}, T_{LM}, T_{FS})$ is asymptotically normal. Using the Cramer-Wold device, it suffices to show that $\sqrt{\frac{K}{r_n}}(c_1 T_{AR} + c_2 T_{LM} + c_3 T_{FS})$ is normal for fixed c's, where

$$\sqrt{\frac{K}{r_n}}(c_1 T_{AR} + c_2 T_{LM} + c_3 T_{FS}) = c_1 \frac{1}{\sqrt{r_n}} \sum_i \sum_{j\neq i} G_{ij} \left(\nu_i R_j + \nu_i \nu_j + R_{\Delta i} R_{\Delta j} + R_{\Delta i} \nu_j\right)$$

$$+ c_2 \frac{1}{\sqrt{r_n}} \sum_i \sum_{j\neq i} G_{ij} \left(\nu_i R_j + \nu_i \eta_j + R_{\Delta i} \eta_j\right) + c_3 \frac{1}{\sqrt{r_n}} \sum_i \sum_{j\neq i} G_{ij} \left(\eta_i R_j + \eta_i \eta_j + R_i R_j + R_i \eta_j\right).$$

The object $T = \sqrt{\frac{K}{r_n}}(c_1 T_{AR} + c_2 T_{LM} + c_3 T_{FS}) - c_1 \frac{1}{\sqrt{r_n}} \sum_i \sum_{j \neq i} G_{ij} R_{\Delta i} R_{\Delta j} - c_3 \frac{1}{\sqrt{r_n}} \sum_i \sum_{j \neq i} G_{ij} R_i R_j$ can be written in the CLT form by setting:

$$v_i = (\eta_i, \nu_i)', A = \begin{bmatrix} c_3 & 0 \\ c_2 & c_1 \end{bmatrix}, \text{ and}$$

$$s_i = \begin{bmatrix} c_3 \sum_{j \neq i} (G_{ij} + G_{ji}) R_j + c_2 \sum_{j \neq i} G_{ji} R_{\Delta j} \\ c_1 \sum_{j \neq i} (G_{ij} + G_{ji}) R_{\Delta j} + c_2 \sum_{j \neq i} G_{ij} R_j \end{bmatrix},$$

so that

$$T = \frac{1}{\sqrt{r_n}} \sum_i s_i' v_i + \frac{1}{\sqrt{r_n}} \sum_i \sum_{j \neq i} G_{ij} v_i' A v_j.$$

Bounded 4th moments hold by Assumption 1.1(a). To apply the CLT from Lemma 1.2, I verify the following:

1. $\text{Var}(T)^{-1/2}$ is bounded;
2. $\frac{1}{r_n^2} \sum_i s_{il}^4 \to 0$ for all l; and
3. $||G_L G_L'||_F + ||G_U G_U'||_F \to 0$, where G_L is a lower-triangular matrix with elements $G_{L,ij} = \frac{1}{\sqrt{r_n}} G_{ij} 1\{i > j\}$ and G_U is an upper-triangular matrix with elements $G_{U,ij} = \frac{1}{\sqrt{r_n}} G_{ij} 1\{i < j\}$.

Condition (2) follows from Assumption 1.1(d) and applying the Cauchy-Schwarz inequality. Condition (3) follows from Assumption 1.1(e). For Condition (1), I show that Assumption 1.1(b) and (c) imply that, for any nonstochastic scalars c_1, c_2, c_3 that are finite and not all 0, $\text{Var}(T)^{-1/2}$ is bounded. Since $Cov\left(\sum_i s_i' v_i, \sum_i \sum_{j \neq i} G_{ij} v_i' A v_j\right) = 0$,

$$\text{Var}(T) = \frac{1}{r_n} \text{Var}\left(\sum_i s_i' v_i\right) + \frac{1}{r_n} \text{Var}\left(\sum_i \sum_{j \neq i} G_{ij} v_i' A v_j\right), \tag{1.16}$$

so it suffices to show that either term is bounded below.

The first term can be expanded as follows:

$$\begin{aligned}\text{Var}\left(\sum_i s_i' v_i\right) &= \sum_i s_i' \text{Var}\left(v_i\right) s_i = \sum_i s_{i1}^2 E[\eta_i]^2 + 2 s_{i1} s_{i2} E[\eta_i \nu_i] + s_{i2}^2 E[\nu_i^2] \\ &= \sum_i (1-\rho_i)^2 E[\eta_i^2] s_{i1}^2 + \left(\rho_i s_{i1} \sqrt{E[\eta_i^2]} + s_{i2} \sqrt{E[\nu_i^2]}\right)^2 \geq \sum_i (1-\rho_i)^2 E[\eta_i^2] s_{i1}^2.\end{aligned}$$

A similar argument yields $\text{Var}\left(\sum_i s_i' v_i\right) \geq \sum_i (1-\rho_i)^2 E[\eta_i^2] s_{i2}^2$. Due to Assumption 1.1(c), at least one of the following must hold: (i) $\text{Var}\left(\sum_i \sum_{j \neq i} G_{ij} v_i' A v_j\right) \geq \underline{c}$ (ii) $\frac{1}{r_n} \sum_i s_{i1}^2 \geq \underline{c}$, or (iii) $\frac{1}{r_n} \sum_i s_{i2}^2 \geq \underline{c}$. Hence, $\text{Var}(T)^{-1/2}$ is bounded.

Finally, since ν_i, η_i are mean zero, the expectations are immediate: $E\left[T_{AR}\right] = \sum_i \sum_{j \neq i} G_{ij} R_{\Delta j} R_{\Delta i}$ and $E\left[T_{FS}\right] = \sum_i \sum_{j \neq i} G_{ij} R_j R_i$. □

The proof of Theorem 1.2 involves several lemmas whose proofs are relegated to Section 1.C. The proof strategy is to bound the variances above by components that are in the $h(.)$ form so that Assumption 1.3 inequalities can be applied.

Let $V_{mi} = R_{mi} + v_{mi}$ where R_{mi} denotes the nonstochastic component while v_{mi} denotes the mean zero stochastic component. Following Equation (1.6), $r_n := \sum_i \tilde{R}_i^2 + \sum_i \tilde{R}_{\Delta i}^2 + \sum_i \sum_{j \neq i} G_{ij}^2$. Let C_i, C_{ij}, C_{ijk} denote nonstochastic objects that are non-negative and are bounded above by C. I use $h_4^A(.)$ and $h_4^B(.)$ to denote two different functions that satisfy the definition for h_4.

Lemma 1.3. *Under Assumption 1.3, the following hold:*

(a) $\left|\sum_{i \neq j \neq k}^n C_{ijk} \left(\sum_{l \neq i,j,k} h_4^A\left(i,j,k,l\right) R_{ml}\right) \left(\sum_{l \neq i,j,k} h_4^B\left(i,j,k,l\right) R_{ml}\right)\right| \leq C \sum_i \tilde{R}_{mi}^2$,
$\left|\sum_{i \neq j}^n C_{ij} \left(\sum_{k \neq i,j} \sum_{l \neq i,j,k} h_4^A\left(i,j,k,l\right) R_{ml}\right) \left(\sum_{k \neq i,j} \sum_{l \neq i,j,k} h_4^B\left(i,j,k,l\right) R_{ml}\right)\right| \leq C \sum_i \tilde{R}_{mi}^2$,
and $\left|\sum_i C_i \left(\sum_{j \neq i} \sum_{k \neq i,j} \sum_{l \neq i,j,k} h_4^A\left(i,j,k,l\right) R_{ml}\right) \left(\sum_{j \neq i} \sum_{k \neq i,j} \sum_{l \neq i,j,k} h_4^B\left(i,j,k,l\right) R_{ml}\right)\right| \leq C \sum_i \tilde{R}_{mi}^2$.

(b) $\left|\sum_{i\neq j}^{n} C_{ij}\left(\sum_{k\neq i,j} h_3^A(i,j,k)\,R_{mk}\right)\left(\sum_{k\neq i,j} h_3^B(i,j,k)\,R_{mk}\right)\right| \leq C\sum_i \tilde{R}_{mi}^2$
and $\left|\sum_i C_i\left(\sum_{j\neq i}\sum_{k\neq i,j} h_3^A(i,j,k)\,R_{mk}\right)\left(\sum_{j\neq i}\sum_{k\neq i,j} h_3^B(i,j,k)\,R_{mk}\right)\right| \leq C\sum_i \tilde{R}_{mi}^2$.

(c) $\left|\sum_i C_i\left(\sum_{j\neq i} h_2^A(i,j)\,R_{mj}\right)\left(\sum_{j\neq i} h_2^B(i,j)\,R_{mj}\right)\right| \leq C\sum_i \tilde{R}_{mi}^2$.

Lemma 1.4. *Under Assumption 1.3, the following hold:*

(a) $\mathrm{Var}\left(\sum_{i\neq j}^{n} G_{ij}F_{ij}V_{1i}V_{2i}V_{3j}V_{4j}\right) \leq Cr_n$.

(b) $\mathrm{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}V_{1i}V_{2k}V_{3j}V_{4j}\right) \leq Cr_n$.

(c) $\mathrm{Var}\left(\sum_{i\neq j\neq l}^{n} G_{ij}F_{ij}\check{M}_{jl,-ij}V_{1i}V_{2i}V_{3j}V_{4l}\right) \leq Cr_n$.

(d) $\mathrm{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}V_{1i}\check{M}_{ik,-ij}V_{2k}V_{3j}\check{M}_{jl,-ijk}V_{4l}\right) \leq Cr_n$.

Lemma 1.5. *Under Assumption 1.3, the following hold:*

(a) $\mathrm{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{1j}V_{2k}V_{3i}V_{4i}\right) \leq Cr_n$.

(b) $\mathrm{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{1j}V_{2k}V_{3i}V_{4l}\right) \leq Cr_n$.

Lemma 1.6. *Under Assumption 1.3, the following hold:*

(a) $\mathrm{Var}\left(\sum_{i\neq j}^{n} G_{ji}^2V_{1i}V_{2i}V_{3j}V_{4j}\right) \leq Cr_n$;

(b) $\mathrm{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ji}^2\check{M}_{ik,-ij}V_{1i}V_{2k}V_{3j}V_{4j}\right) \leq Cr_n$;

(c) $\mathrm{Var}\left(\sum_{i\neq j\neq l}^{n} G_{ji}^2\check{M}_{jl,-ij}V_{1i}V_{2i}V_{3j}V_{4l}\right) \leq Cr_n$;

(d) $\mathrm{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ji}^2V_{1i}\check{M}_{ik,-ij}V_{2k}V_{3j}\check{M}_{jl,-ijk}V_{4l}\right) \leq Cr_n$;

(e) $\mathrm{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ji}F_{ki}V_{1j}V_{2k}V_{3i}V_{4i}\right) \leq Cr_n$;

(f) $\mathrm{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ji}F_{ki}\check{M}_{il,-ijk}V_{1j}V_{2k}V_{3i}V_{4l}\right) \leq Cr_n$.

Proof of Theorem 1.2. **Proof of Unbiasedness.** The variance expression can be equivalently be written as:

$$V_{LM} = \sum_i \left(E\left[\nu_i^2\right] \left(\sum_{j \neq i} G_{ij} R_j \right)^2 + 2 \left(\sum_{j \neq i} G_{ij} R_j \right) \left(\sum_{j \neq i} G_{ji} R_{\Delta j} \right) E\left[\nu_i \eta_i\right] + E\left[\eta_i^2\right] \left(\sum_{j \neq i} G_{ji} R_{\Delta j} \right)^2 \right)$$
$$+ \sum_i \sum_{j \neq i} G_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] + \sum_i \sum_{j \neq i} G_{ij} G_{ji} E\left[\eta_i \nu_i\right] E\left[\eta_j \nu_j\right]. \tag{1.17}$$

To ease notation, let:

$$A_{1i} := \sum_{j \neq i} \sum_{k \neq i} G_{ij} X_j G_{ik} X_k e_i \left(e_i - Q_i' \hat{\tau}_{\Delta, -ijk} \right),$$
$$A_{2i} := \sum_{j \neq i} \sum_{k \neq i} G_{ij} X_j G_{ki} e_k e_i \left(X_i - Q_i' \hat{\tau}_{-ijk} \right),$$
$$A_{3i} := \sum_{j \neq i} \sum_{k \neq i} G_{ji} e_j G_{ki} e_k X_i \left(X_i - Q_i' \hat{\tau}_{-ijk} \right),$$
$$A_{4ij} := X_i \sum_{k \neq j} \check{M}_{ik,-ij} X_k e_j \left(e_j - Q_j' \hat{\tau}_{\Delta, -ijk} \right), \text{ and}$$
$$A_{5ij} := e_i \sum_{k \neq j} \check{M}_{ik,-ij} X_k e_j \left(X_j - Q_j' \hat{\tau}_{-ijk} \right).$$

Take expectation of A_1:

$$E\left[\sum_i \sum_{j \neq i} \sum_{k \neq i} G_{ij} X_j G_{ik} X_k e_i \left(e_i - Q_i' \hat{\tau}_{\Delta, -ijk} \right) \right]$$
$$= \sum_i \sum_{j \neq i} \sum_{k \neq i, j} G_{ij} E\left[X_j\right] G_{ik} E\left[X_k\right] E\left[e_i \left(e_i - Q_i' \hat{\tau}_{\Delta, -ijk} \right) \right] + \sum_i \sum_{j \neq i} G_{ij}^2 E\left[X_j^2\right] E\left[e_i \left(e_i - Q_i' \hat{\tau}_{\Delta, -ijk} \right) \right]$$
$$= \sum_i \sum_{j \neq i} \sum_{k \neq i} G_{ij} R_j G_{ik} R_k E\left[\nu_i^2\right] + \sum_i \sum_{j \neq i} G_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right].$$

Similarly,

$$E\left[A_{2i}\right] = \left(\sum_{j\neq i} G_{ij}R_j\right)\left(\sum_{j\neq i} G_{ji}R_{\Delta j}\right) E\left[\nu_i\eta_i\right] + \sum_{j\neq i} G_{ij}G_{ji}E\left[\eta_i\nu_i\right] E\left[\eta_j\nu_j\right], \text{ and}$$
$$E\left[A_{3i}\right] = E\left[\eta_i^2\right]\left(\sum_{j\neq i} G_{ji}R_{\Delta j}\right)^2 + \sum_{j\neq i} G_{ji}^2 E\left[\eta_i^2\right] E\left[\nu_j^2\right].$$

For the A_4 and A_5 terms, observe that:

$$X_i - Q_i'\hat{\tau}_{-ij} = X_i - Q_i'\sum_{k\neq i,j}\left(\sum_{l\neq i,j} Q_l Q_l'\right)^{-1} Q_k X_k = X_i + \sum_{k\neq i,j}\check{M}_{ik,-ij}X_k = \sum_{k\neq j}\check{M}_{ik,-ij}X_k,$$

where the final equality follows from $\check{M}_{ii,-ij} = 1$. Then,

$$\begin{aligned} E\left[A_{4ij}\right] &= E\left[X_i\sum_{k\neq j}\check{M}_{ik,-ij}X_k e_j\left(X_j - Q_j'\hat{\tau}_{\Delta,-ijk}\right)\right] = \sum_{k\neq j} E\left[X_i\check{M}_{ik,-ij}X_k e_j\left(X_j - Q_j'\hat{\tau}_{\Delta,-ijk}\right)\right] \\ &= E\left[X_i\sum_{k\neq j}\check{M}_{ik,-ij}X_k\right] E\left[e_j\left(e_j - Q_j'\hat{\tau}_{\Delta,-ijk}\right)\right] \\ &= E\left[X_i\left(X_i - Q_i'\hat{\tau}_{-ij}\right)\right] E\left[\nu_j^2\right] = E\left[\eta_i^2\right] E\left[\nu_j^2\right]. \end{aligned}$$

Similarly, $E\left[A_{5ij}\right] = E\left[\eta_i\nu_i\right] E\left[\eta_j\nu_j\right]$. Combining these expressions yields the unbiasedness result.

Proof of Consistency. By Chebyshev's inequality,

$$\begin{aligned} &\Pr\left(\left|\frac{\hat{V}_{LM} - \text{Var}\left(\sum_i\sum_{j\neq i} G_{ij}e_iX_j\right)}{\text{Var}\left(\sum_i\sum_{j\neq i} G_{ij}e_iX_j\right)}\right| > \epsilon\right) \\ &\leq \frac{1}{\epsilon^2}\frac{\text{Var}\left(\sum_i\left(A_{1i} + 2A_{2i} + A_{3i}\right) - \sum_i\sum_{j\neq i} G_{ji}^2 A_{4ij} - \sum_i\sum_{j\neq i} G_{ij}G_{ji}A_{5ij}\right)}{\left[\text{Var}\left(\sum_i\sum_{j\neq i} G_{ij}e_iX_j\right)\right]^2} \end{aligned}$$

Observe that the numerator can be written as the variance of the estimator only because $\hat{V}_{LM}$ is unbiased. I first establish the order of the denominator. Let $\tilde{R}_i := \sum_{j \neq i} G_{ij} R_j$, $\tilde{R}_{\Delta i} := \sum_{j \neq i} G_{ji} R_{\Delta j}$ and $\rho_i := corr(\eta_i \nu_i)$.

Since $E[\nu_i^2]$ and $E[\eta_i^2]$ are bounded away from zero and $|\text{corr}(\eta_i \nu_i)|$ is bounded away from one by Assumption 1.1(b), the first line of the V_{LM} expression in Equation (1.17) has order at least $\sum_i \tilde{R}_i^2 + \sum_i \tilde{R}_{\Delta i}^2$, and the second line has order at least $\sum_i \sum_{j \neq i} G_{ij}^2$. To see this, for some $\underline{c} > 0$, the first line is:

$$
\begin{aligned}
&\sum_i E\left[\nu_i^2\right] \tilde{R}_i^2 + 2\tilde{R}_{\Delta i}\tilde{R}_i E\left[\nu_i \eta_i\right] + \tilde{R}_{\Delta i}^2 E\left[\eta_i^2\right] = \sum_i E\left[\nu_i^2\right] \tilde{R}_i^2 + 2\tilde{R}_{\Delta i}\tilde{R}_i \rho_i \sqrt{E\left[\nu_i^2\right] E\left[\eta_i^2\right]} + \tilde{R}_{\Delta i}^2 E\left[\eta_i^2\right] \\
&\geq \sum_i \left(E\left[\nu_i^2\right] \tilde{R}_i^2 + \tilde{R}_{\Delta i}^2 E\left[\eta_i^2\right] \right) (1 - |\rho_i|) + \sum_i |\rho_i| \left(E\left[\nu_i^2\right] \tilde{R}_i^2 + \tilde{R}_{\Delta i}^2 E\left[\eta_i^2\right] - 2\tilde{R}_{\Delta i}\tilde{R}_i \sqrt{E\left[\nu_i^2\right] E\left[\eta_i^2\right]} \right) \\
&= \sum_i \left(E\left[\nu_i^2\right] \tilde{R}_i^2 + \tilde{R}_{\Delta i}^2 E\left[\eta_i^2\right] \right) (1 - |\rho_i|) + \sum_i |\rho_i| \left(\sqrt{E\left[\nu_i^2\right] \tilde{R}_i^2} - \sqrt{\tilde{R}_{\Delta i}^2 E\left[\eta_i^2\right]} \right)^2 \\
&\geq \sum_i \left(E\left[\nu_i^2\right] \tilde{R}_i^2 + \tilde{R}_{\Delta i}^2 E\left[\eta_i^2\right] \right) (1 - |\rho_i|) \geq \underline{c} \sum_i \left(\tilde{R}_i^2 + \tilde{R}_{\Delta i}^2 \right),
\end{aligned}
$$

and the second line is:

$$
\begin{aligned}
&\sum_i \sum_{j \neq i} G_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] + \sum_i \sum_{j \neq i} G_{ij} G_{ji} E\left[\eta_i \nu_i\right] E\left[\eta_j \nu_j\right] \\
&= \frac{1}{2} \sum_i \sum_{j \neq i} G_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] \left(1 - \rho_i^2\right) + \frac{1}{2} \sum_i \sum_{j \neq i} G_{ji}^2 E\left[\nu_j^2\right] E\left[\eta_i^2\right] \left(1 - \rho_j^2\right) \\
&\quad + \frac{1}{2} \sum_i \sum_{j \neq i} E\left[\nu_i^2\right] E\left[\eta_j^2\right] \left(G_{ij}^2 \rho_i^2 + G_{ji}^2 \rho_j^2\right) + \sum_i \sum_{j \neq i} G_{ij} G_{ji} \rho_i \rho_j \sqrt{E\left[\nu_i^2\right] E\left[\eta_j^2\right]} \sqrt{E\left[\nu_j^2\right] E\left[\eta_i^2\right]} \\
&= \frac{1}{2} \sum_i \sum_{j \neq i} G_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] \left(1 - \rho_i^2\right) + \frac{1}{2} \sum_i \sum_{j \neq i} G_{ji}^2 E\left[\nu_j^2\right] E\left[\eta_i^2\right] \left(1 - \rho_j^2\right) \\
&\quad + \frac{1}{2} \sum_i \sum_{j \neq i} \left(G_{ij} \rho_i \sqrt{E\left[\nu_i^2\right] E\left[\eta_j^2\right]} + G_{ji} \rho_j \sqrt{E\left[\nu_j^2\right] E\left[\eta_i^2\right]} \right)^2 \\
&\geq \frac{1}{2} \sum_i \sum_{j \neq i} G_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] \left(1 - \rho_i^2\right) + \frac{1}{2} \sum_i \sum_{j \neq i} G_{ji}^2 E\left[\nu_j^2\right] E\left[\eta_i^2\right] \left(1 - \rho_j^2\right) \geq \underline{c} \sum_i \sum_{j \neq i} G_{ij}^2.
\end{aligned}
$$

Consequently,

$$V_{LM} \succeq \sum_i \tilde{R}_i^2 + \sum_i \tilde{R}_{\Delta i}^2 + \sum_i \sum_{j \neq i} G_{ij}^2 =: r_n. \tag{1.18}$$

Hence, since $r_n \to \infty$ by Assumption 1.1(d), V_{LM} also diverges. By repeated application of the Cauchy-Schwarz inequality, it suffices to show that the variance of each of the 5 A terms above has order at most r_n (i.e., bounded by any of the three terms in Equation (1.18)). If this is true, then since the denominator has order at least r_n^2, the variance estimator is consistent. The A1 and A2 terms have the form:

$$\begin{aligned}
\sum_i \sum_{j \neq i} G_{ij} F_{ik} V_{1j} \sum_{k \neq i} V_{2k} V_{3i} \left(V_{4i} - Q_i' \hat{\tau}_{4,-ijk}\right) &= \sum_i \sum_{j \neq i} \sum_{k \neq i} \sum_{l \neq j,k} G_{ij} F_{ik} V_{1j} V_{2k} V_{3i} \check{M}_{il,-ijk} V_{4l} \\
&= \sum_i \sum_{j \neq i} \sum_{k \neq i,j} \sum_{l \neq i,j,k} G_{ij} F_{ik} \check{M}_{il,-ijk} V_{1j} V_{2k} V_{3i} V_{4l} + \sum_i \sum_{j \neq i} \sum_{k \neq i,j} G_{ij} F_{ik} V_{1j} V_{2k} V_{3i} V_{4i} \\
&\quad + \sum_i \sum_{j \neq i} \sum_{l \neq i,j} G_{ij} F_{ij} \check{M}_{il,-ij} V_{1j} V_{2j} V_{3i} V_{4l} + \sum_i \sum_{j \neq i} G_{ij} F_{ij} V_{1j} V_{2j} V_{3i} V_{4i}.
\end{aligned}$$

In particular, A1 uses $F = G, V_1 = X, V_2 = X, V_3 = e, V_4 = e$, while A2 uses $F = G', V_1 = X, V_2 = e, V_3 = e, V_4 = X$. By applying the Cauchy-Schwarz inequality, it suffices to show that the variance of each of the sums has order at most r_n. The terms $\sum_i \sum_{j \neq i} G_{ij} F_{ij} V_{1j} V_{2j} V_{3i} V_{4i}$ and $\sum_i \sum_{j \neq i} \sum_{l \neq i,j} G_{ij} F_{ij} \check{M}_{il,-ij} V_{1j} V_{2j} V_{3i} V_{4l}$ are identical to the result in Lemma 1.4, with the latter result being obtained by switching the i and j indices. The remaining terms have a variance that has a bounded order by Lemma 1.5. For A3, we can use G_{ji} in place of G_{ij} above, and use $F = G', V_1 = e, V_2 = e, V_3 = X, V_4 = X$ so that

the order is bounded above due to Lemma 1.6. A4 and A5 can be written as:

$$\begin{aligned}
&\sum_i \sum_{j\neq i} G_{ji}F_{ij}V_{1i} \sum_{k\neq j} \check{M}_{ik,-ij}V_{2k}V_{3j}\left(V_{4j} - Q_j'\hat{\tau}_{4,-ijk}\right) \\
&\quad = \sum_i \sum_{j\neq i} \sum_{k\neq j} \sum_{l\neq i,k} G_{ji}F_{ij}V_{1i}\check{M}_{ik,-ij}V_{2k}V_{3j}\check{M}_{jl,-ijk}V_{4l} \\
&\quad = \sum_i \sum_{j\neq i} \sum_{k\neq i,j} \sum_{l\neq i,j,k} G_{ji}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}V_{2k}V_{3j}V_{4l} + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} G_{ji}F_{ij}\check{M}_{ik,-ij}V_{1i}V_{2k}V_{3j}V_{4j} \\
&\quad + \sum_i \sum_{j\neq i} \sum_{l\neq i,j} G_{ji}F_{ij}\check{M}_{jl,-ij}V_{1i}V_{2i}V_{3j}V_{4l} + \sum_i \sum_{j\neq i} G_{ji}F_{ij}V_{1i}V_{2i}V_{3j}V_{4j}.
\end{aligned}$$

In particular, A4 uses $F = G', V_1 = X, V_2 = X, V_3 = e, V_4 = e$, while A5 uses $F = G, V_1 = e, V_2 = X, V_3 = e, V_4 = X$. By applying the Cauchy-Schwarz inequality, it suffices to show that the variance of each of the sums has order at most r_n. This result is immediate from Lemma 1.4 and Lemma 1.6. □

Proof of Proposition 1.1. Let $\mu \in \mathcal{M} = \{\mu : \mu_1 > 0, \mu_3 > 0, \mu_2^2 < \mu_1\mu_3\}$. I first show that $\mathcal{M}$ is convex. For $\lambda \in (0,1)$, it suffices to show, for μ_a and μ_b that satisfy $\mu_{2a}^2 < \mu_{1a}\mu_{3a}$ and $\mu_{2b}^2 < \mu_{1b}\mu_{3b}$, that $(\lambda\mu_{2a} + (1-\lambda)\mu_{2b})^2 < (\lambda\mu_{1a} + (1-\lambda)\mu_{1b})(\lambda\mu_{3a} + (1-\lambda)\mu_{3b})$. This set is intersected with the set that satisfies $\mu_1 > 0$ and $\mu_3 > 0$, which is clearly convex. The following is negative:

$$\begin{aligned}
&(\lambda\mu_{2a} + (1-\lambda)\mu_{2b})^2 - (\lambda\mu_{1a} + (1-\lambda)\mu_{1b})(\lambda\mu_{3a} + (1-\lambda)\mu_{3b}) \\
=&\lambda^2\mu_{2a}^2 + (1-\lambda)^2\mu_{2b}^2 + 2\lambda(1-\lambda)\mu_{2a}\mu_{2b} - \lambda^2\mu_{1a}\mu_{3a} - (1-\lambda)^2\mu_{1b}\mu_{3b} - \lambda(1-\lambda)(\mu_{1b}\mu_{3a} + \mu_{1a}\mu_{3b}) \\
=&\lambda^2\left(\mu_{2a}^2 - \mu_{1a}\mu_{3a}\right) + (1-\lambda)^2\left(\mu_{2b}^2 - \mu_{1b}\mu_{3b}\right) + \lambda(1-\lambda)(2\mu_{2a}\mu_{2b} - \mu_{1b}\mu_{3a} - \mu_{1a}\mu_{3b}) \\
<&\lambda(1-\lambda)(2\sqrt{\mu_{1a}\mu_{1b}\mu_{1b}\mu_{3b}} - \mu_{1b}\mu_{3a} - \mu_{1a}\mu_{3b}) \\
<&-\lambda(1-\lambda)(\sqrt{\mu_{1b}\mu_{3a}} - \sqrt{\mu_{1a}\mu_{3b}})^2 \leq 0.
\end{aligned}$$

The first inequality occurs from applying $\mu_{2a}^2 < \mu_{1a}\mu_{3a}$ and $\mu_{2b}^2 < \mu_{1b}\mu_{3b}$, so $\mathcal{M}$ is convex. Let $m \sim N(\mu, \Sigma)$ denote a statistic drawn from the asymptotic distribution, with m_i being a

component of the vector m, so that m_2 is the LM statistic. Using the linear transformation from Lehmann and Romano (2005) Example 3.9.2 Case 3, we can transform the statistics and parameter such that m_2 is orthogonal to all other components. In particular, consider the following transformation L:

$$L := \begin{pmatrix} \sqrt{\frac{\sigma_{22}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}} & -\frac{\sigma_{12}}{\sigma_{22}}\sqrt{\frac{\sigma_{22}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}} & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & 0 \\ 0 & -\frac{\sigma_{23}}{\sigma_{22}}\sqrt{\frac{\sigma_{22}}{\sigma_{33}\sigma_{22}-\sigma_{23}^2}} & \sqrt{\frac{\sigma_{22}}{\sigma_{33}\sigma_{22}-\sigma_{23}^2}} \end{pmatrix}.$$

Then,

$$Lm \sim N\left(L\mu, \begin{pmatrix} 1 & 0 & \frac{\sigma_{13}\sigma_{22}-\sigma_{12}\sigma_{23}}{\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)\left(\sigma_{33}\sigma_{22}-\sigma_{23}^2\right)} \\ 0 & 1 & 0 \\ \frac{\sigma_{13}\sigma_{22}-\sigma_{12}\sigma_{23}}{\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)\left(\sigma_{33}\sigma_{22}-\sigma_{23}^2\right)} & 0 & 1 \end{pmatrix}\right).$$

The parameter space of $L\mu \in \mathcal{L}$ is also convex because L is a linear transformation: take any $\mu_a, \mu_b \in \mathcal{M}$, then observe that $\lambda L\mu_a + (1-\lambda)L\mu_b = L\left(\lambda\mu_a + (1-\lambda)\mu_b\right)$. Since $\mathcal{M}$ is convex, and every element in $\mathcal{M}$ is linearly transformed into the space on $\mathcal{L}$, we have $\lambda\mu_a + (1-\lambda)\mu_b \in \mathcal{M}$ and hence $L\left(\lambda\mu_a + (1-\lambda)\mu_b\right) \in \mathcal{L}$. Since Lm is normally distributed and $\mathcal{L}$ is convex with rank 3, the problem is in the exponential class, using the definition from Lehmann and Romano (2005) Section 4.4. Since the joint distribution is in the exponential class and the restriction to the interior ensures that there are points in the parameter space that are above and below the null, the uniformly most powerful unbiased test follows the form of Lehmann and Romano (2005) Theorem 4.4.1(iv), by using $U = m_2$ and $T = \left(\sqrt{\frac{\sigma_{22}}{\sigma_{33}\sigma_{22}-\sigma_{23}^2}}m_3 - \frac{\sigma_{23}}{\sqrt{\sigma_{22}\left(\sigma_{33}\sigma_{22}-\sigma_{23}^2\right)}}m_2, \sqrt{\frac{\sigma_{22}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}}m_1 - \frac{\sigma_{12}}{\sqrt{\sigma_{22}\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)}}m_2\right)'$ in their notation. To calculate the critical values of the Lehmann and Romano (2005) Theorem 4.4.1(iv) result, observe that $[Lm]_2$ is orthogonal to $[Lm]_1$ and $[Lm]_3$, so the distribution of

$[Lm]_2$ conditional on $[Lm]_1$ and $[Lm]_3$ is standard normal. Since $[Lm]_2$ is standard normal, it is symmetric around 0 under the null, so the solution to the critical value is ± 1.96 for a 5% test, due to simplification in Lehmann and Romano (2005) Section 4.2. The resulting test is hence identical to the two-sided LM test. □

The full and latest version of the paper, including the online appendix, can be found at https://lutheryap.github.io/files/mwiv_het_wp.pdf.

1.C Supplementary Appendix

1.C.1 Comparing Variance Estimands

Derivations for Constructed Instrument Using the notation for the just-identified IV AR test in Section 1.2.4 when $\beta = \beta_0$,

$$\hat{\varepsilon}_i = e_i - \tilde{X}_i \frac{\sum_i e_i \tilde{X}_i}{\sum_i \tilde{X}_i^2} = \frac{e_i \sum_i \tilde{X}_i^2 - \tilde{X}_i \sum_i e_i \tilde{X}_i}{\sum_i \tilde{X}_i^2}, \text{ and}$$

$$\begin{aligned}\hat{V} &= \frac{\sum_i \tilde{X}_i^2 \hat{\varepsilon}_i^2}{\left(\sum_i \tilde{X}_i^2\right)^2} = \frac{\sum_i \tilde{X}_i^2 \left(e_i \sum_j \tilde{X}_j^2 - \tilde{X}_i \sum_j e_j \tilde{X}_j\right)^2}{\left(\sum_i \tilde{X}_i^2\right)^4} \\ &= \frac{\sum_i \tilde{X}_i^2 e_i^2 \left(\sum_j \tilde{X}_j^2\right)^2 + \sum_i \tilde{X}_i^4 \left(\sum_j e_j \tilde{X}_j\right)^2 - 2\sum_i \tilde{X}_i^3 e_i \left(\sum_j \tilde{X}_j^2\right)\left(\sum_j e_j \tilde{X}_j\right)}{\left(\sum_i \tilde{X}_i^2\right)^4}.\end{aligned}$$

Applying the asymptotic result that $\frac{1}{n}\sum_j e_j \tilde{X}_j \xrightarrow{p} 0$ from Theorem 1.1,

$$\begin{aligned}t_{AR}^2 &= \frac{\frac{\left(\sum_i \tilde{X}_i e_i\right)^2}{\left(\sum_i \tilde{X}_i^2\right)^2}}{\frac{\sum_i \tilde{X}_i^2 e_i^2 \left(\sum_j \tilde{X}_j^2\right)^2 + \sum_i \tilde{X}_i^4 \left(\sum_j e_j \tilde{X}_j\right)^2 - 2\sum_i \tilde{X}_i^3 e_i \left(\sum_j \tilde{X}_j^2\right)\left(\sum_j e_j \tilde{X}_j\right)}{\left(\sum_i \tilde{X}_i^2\right)^4}} \\ &= \frac{\left(\frac{1}{\sqrt{n}}\sum_i \tilde{X}_i e_i\right)^2 \left(\frac{1}{n}\sum_i \tilde{X}_i^2\right)^2}{\frac{1}{n}\sum_i \tilde{X}_i^2 e_i^2 \left(\frac{1}{n}\sum_j \tilde{X}_j^2\right)^2 + \frac{1}{n}\sum_i \tilde{X}_i^4 \left(\frac{1}{n}\sum_j e_j \tilde{X}_j\right)^2 - 2\frac{1}{n}\sum_i \tilde{X}_i^3 e_i \left(\frac{1}{n}\sum_j \tilde{X}_j^2\right)\left(\frac{1}{n}\sum_j e_j \tilde{X}_j\right)} \\ &= \frac{\left(\frac{1}{\sqrt{n}}\sum_i \tilde{X}_i e_i\right)^2 \left(\frac{1}{n}\sum_i \tilde{X}_i^2\right)^2}{\frac{1}{n}\sum_i \tilde{X}_i^2 e_i^2 \left(\frac{1}{n}\sum_j \tilde{X}_j^2\right)^2 + o_P(1)} = \frac{\left(\frac{1}{\sqrt{n}}\sum_i \tilde{X}_i e_i\right)^2}{\frac{1}{n}\sum_i \tilde{X}_i^2 e_i^2} + o_P(1), \text{ and}\end{aligned}$$

$$
\begin{aligned}
E\left[\sum_i \tilde{X}_i^2 e_i^2\right] &= \sum_i E\left[\left(\sum_{j\neq i} P_{ij}\left(R_j+\eta_j\right)\right)^2 \left(R_{\Delta i}+\nu_i\right)^2\right] \\
&= \sum_i E\left[\left(M_{ii}^2 R_i^2 + \left(\sum_{j\neq i} P_{ij}^2 \eta_j^2\right)\right)\left(R_{\Delta i}^2+\nu_i^2\right)\right] \\
&= \sum_i \left(M_{ii}^2 R_i^2 R_{\Delta i}^2 + \sum_{j\neq i} P_{ij}^2 R_{\Delta i}^2 E\left[\eta_j^2\right] + M_{ii}^2 R_i^2 E\left[\nu_i^2\right] + \sum_{j\neq i} P_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right]\right).
\end{aligned}
$$

Clustering by Judges. If we just use the JIVE t-ratio with clustering, weak identification is still a problem, and we should similarly get over-rejection. The just-identified AR with clustered standard errors uses the following estimator:

$$
\hat{V}_{clus} = \frac{\sum_i \sum_{j\in\mathcal{N}_i} \tilde{X}_i \hat{\varepsilon}_i \tilde{X}_j \hat{\varepsilon}_j}{\left(\sum_i \tilde{X}_i^2\right)^2},
$$

where $\mathcal{N}_i$ is the neighborhood of i (i.e., the set of observations that share the same cluster as i). Expanding $\hat{V}_{clus}$ using the same steps as before,

$$
\hat{V}_{clus} = \frac{\sum_i \sum_{j\in\mathcal{N}_i} \tilde{X}_i \tilde{X}_j \left(e_i e_j \left(\sum_k \tilde{X}_k^2\right)^2 - 2\tilde{X}_i e_j \left(\sum_k e_k \tilde{X}_k\right)\left(\sum_k \tilde{X}_k^2\right) + \tilde{X}_i \tilde{X}_j \left(\sum_k e_k \tilde{X}_k\right)^2\right)}{\left(\sum_i \tilde{X}_i^2\right)^4}.
$$

Using the fact that $\frac{1}{n}\sum_k e_k \tilde{X}_k = o_P(1)$, the dominant term is: $\sum_i \sum_{j\in\mathcal{N}_i} \tilde{X}_i \tilde{X}_j e_i e_j$, which is analogous to the previous derivation. The expansion steps are analogous to that required to

derive V_{LM}, so they are omitted. Then,

$$
\begin{aligned}
&E\left[\sum_i \sum_{j\in\mathcal{N}_i} \tilde{X}_i \tilde{X}_j e_i e_j\right] \\
&= E\left[\sum_i \sum_{j\in\mathcal{N}_i} \left(\sum_{k\neq i} P_{ik}\left(R_k+\eta_k\right)\right)\left(\sum_{k\neq j} P_{jk}\left(R_k+\eta_k\right)\right)\left(R_{\Delta i}+\nu_i\right)\left(R_{\Delta j}+\nu_j\right)\right] \\
&= E\left[\sum_i \sum_{j\in\mathcal{N}_i} \left(M_{ii}R_i+\sum_{k\neq i} P_{ik}\eta_k\right)\left(M_{jj}R_j+\sum_{k\neq j} P_{jk}\eta_k\right)\left(R_{\Delta i}R_{\Delta j}+\nu_i R_{\Delta j}+R_{\Delta i}\nu_j+\nu_i\nu_j\right)\right] \\
&= \sum_i \sum_{j\in\mathcal{N}_i}\left(M_{ii}M_{jj}R_iR_jR_{\Delta i}R_{\Delta j}+R_{\Delta i}R_{\Delta j}\sum_{k\neq i,j} P_{ik}P_{jk}E\left[\eta_k^2\right]\right) \\
&\quad + 2\sum_i \sum_{j\in\mathcal{N}_i} M_{ii}R_iR_{\Delta j}P_{ji}E\left[\eta_i\nu_i\right] + \sum_i M_{ii}^2R_i^2E\left[\nu_i^2\right] \\
&\quad + \sum_i \sum_{k\neq i} P_{ik}^2E\left[\nu_i^2\right]E\left[\eta_k^2\right] + \sum_i \sum_{j\in\mathcal{N}_i\setminus\{i\}} P_{ij}^2E\left[\nu_i\eta_i\right]E\left[\nu_j\eta_j\right].
\end{aligned}
$$

By applying the fact that the entries of the projection matrix are nonzero only when the observations share the same judge, the expression simplifies further:

$$
\begin{aligned}
E\left[\sum_i \sum_{j\in\mathcal{N}_i} \tilde{X}_i \tilde{X}_j e_i e_j\right] &= \sum_i \sum_{j\in\mathcal{N}_i} M_{ii}M_{jj}R_iR_jR_{\Delta i}R_{\Delta j} + \sum_i M_{ii}^2R_{\Delta i}^2E\left[\eta_i^2\right] \\
&\quad + 2\sum_i M_{ii}R_iR_{\Delta i}E\left[\eta_i\nu_i\right] + \sum_i M_{ii}^2R_i^2E\left[\nu_i^2\right] \\
&\quad + \sum_i \sum_{j\neq i} P_{ij}^2\left(E\left[\nu_i^2\right]E\left[\eta_j^2\right] + E\left[\nu_i\eta_i\right]E\left[\nu_j\eta_j\right]\right).
\end{aligned}
$$

Compared to the true variance in Equation (1.20), due to the own-observation bias, we have an extra $\sum_i \sum_{j\in\mathcal{N}_i} M_{ii}M_{jj}R_iR_jR_{\Delta i}R_{\Delta j}$ term, and the estimand here has $\sum_i M_{ii}R_iR_{\Delta i}E\left[\eta_i\nu_i\right]$ instead of $\sum_i M_{ii}^2R_iR_{\Delta i}E\left[\eta_i\nu_i\right]$. Even though $\sum_i \sum_{j\in\mathcal{N}_i} M_{ii}M_{jj}R_iR_jR_{\Delta i}R_{\Delta j} \geq 0$, $\sum_i M_{ii}(1-M_{ii})R_iR_{\Delta i}E\left[\eta_i\nu_i\right]$ could be positive or negative, so the clustered variance estimand could either over or underestimate the true variance.

Comparing MO Variance Estimand with L3O. The Matsushita and Otsu (2022) variance estimator presented in Equation (1.4) is biased in general. The model of Section 1.2.1 implies:

$$E\left[\hat{\Psi}_{MO}\right] = \sum_i M_{ii}^2 R_i^2 R_{\Delta i}^2 + \sum_i M_{ii}^2 R_i^2 E\left[\nu_i^2\right] + \sum_i \sum_{j\neq i} P_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] + \sum_i \sum_{j\neq i} P_{ij}^2 R_{\Delta i}^2 E\left[\eta_j^2\right] + \sum_i \sum_{j\neq i} P_{ij}^2 \left(R_i R_{\Delta i} R_j R_{\Delta j} + E\left[\eta_i \nu_i\right] R_j R_{\Delta j} + R_i R_{\Delta i} E\left[\eta_j \nu_j\right] + E\left[\eta_i \nu_i\right] E\left[\eta_j \nu_j\right]\right). \tag{1.19}$$

As a corollary of Equation (1.8), when $G = P$, by observing that P is symmetric, and that since $PR = I$, we have $\sum_{j\neq i} P_{ij} R_j = \sum_{j\neq i} P_{ji} R_j = M_{ii} R_i$, so

$$\operatorname{Var}\left(\sum_i \sum_{j\neq i} P_{ij} e_i X_j\right) = \sum_i E\left[\nu_i^2\right] M_{ii}^2 R_i^2 + \sum_i \sum_{j\neq i} P_{ij}^2 \left(E\left[\nu_i^2\right] E\left[\eta_j^2\right] + E\left[\eta_i \nu_i\right] E\left[\eta_j \nu_j\right]\right) + 2\sum_i E\left[\nu_i \eta_i\right] M_{ii}^2 R_i R_{\Delta i} + \sum_i E\left[\eta_i^2\right] M_{ii}^2 R_{\Delta i}^2. \tag{1.20}$$

If the R_Δ's are zero, then $\hat{\Psi}_{MO}$ is unbiased. Nonetheless, heterogeneity results in many excess terms in the expectation of the variance estimator, generating bias and inconsistency in general. However, $\hat{\Psi}_{MO}$ can be consistent when forcing weak identification and weak heterogeneity. If it is assumed that $\frac{1}{\sqrt{K}} \sum_i M_{ii} R_i^2 \to C_S < \infty$ and $\frac{1}{\sqrt{K}} \sum_i M_{ii} R_{\Delta i}^2 \to C < \infty$ with weak identification and weak heterogeneity, then the excess terms in $\frac{1}{K} E\left[\hat{\Psi}_{MO}\right]$ can be written as $\frac{1}{\sqrt{K}} \frac{1}{\sqrt{K}} \sum_i M_{ii} R_i^2 = \frac{1}{\sqrt{K}} O(1) = o(1)$ and $\frac{1}{\sqrt{K}} \frac{1}{\sqrt{K}} \sum_i M_{ii} R_{\Delta i}^2 = o(1)$. However, when identification or heterogeneity is strong, $\frac{1}{K} \sum_i M_{ii} R_i^2$ or $\frac{1}{K} \sum_i M_{ii} R_{\Delta i}^2$ is nonnegligible and the variance estimator is inconsistent. The variance estimator adapted from MS22 has similar properties. In contrast, the L3O variance estimator is robust regardless of whether the identification is weak or strong.

In general, heterogeneity does not make the MO variance estimator any more or less conservative than L3O. In the simple case with judge instruments and $G = P$, we have:

$$\begin{aligned}
E\left[\hat{\Psi}_{MO}\right] - Var\left(\sum_i \sum_{j\neq i} P_{ij} e_i X_j\right) &= \sum_i M_{ii}^2 R_i^2 R_{\Delta i}^2 + \sum_i \sum_{j\neq i} P_{ij}^2 R_i R_{\Delta i} R_j R_{\Delta j} \\
&+ \sum_i \sum_{j\neq i} P_{ij}^2 R_{\Delta i}^2 E\left[\eta_j^2\right] - \sum_i M_{ii}^2 R_{\Delta i}^2 E\left[\eta_i^2\right] \\
&+ 2\sum_i \sum_{j\neq i} P_{ij}^2 E\left[\eta_i \nu_i\right] R_j R_{\Delta j} - 2\sum_i E\left[\nu_i \eta_i\right] M_{ii}^2 R_i R_{\Delta i} \\
&= \sum_i M_{ii} R_{\Delta i}^2 R_i^2 - \sum_i M_{ii}(1 - 2P_{ii}) R_{\Delta i}^2 E\left[\eta_i^2\right] - 2\sum_i M_{ii}(1 - 2P_{ii}) E\left[\eta_i \nu_i\right] R_i R_{\Delta i} \\
&= \sum_i M_{ii} R_{\Delta i}^2 \left(R_i^2 - (1 - 2P_{ii}) E\left[\eta_i^2\right]\right) - 2\sum_i M_{ii}(1 - 2P_{ii}) E\left[\eta_i \nu_i\right] R_i R_{\Delta i}
\end{aligned}$$

which can be positive or negative. The second equality uses the fact that P and M are non-zero only for observations that share the same judge, and when that occurs, they have the same $R, R_Y, E[\eta_i^2]$, and $E[\zeta_i^2]$, and that $\sum_{j\neq i} P_{ij}^2 = P_{ii} M_{ii}$.

To compare the confidence sets of MO and L3O, observe that the shape of the confidence set depends on the coefficient on β_0^2. In particular, for $\Psi_2 := \sum_i \left(\sum_{j\neq i} P_{ij} X_j\right)^2 X_i^2 + \sum_i \sum_{j\neq i} P_{ij}^2 X_i^2 X_j^2$, MO is unbounded when $\left(\sum_{i\neq j}^n P_{ij} X_i X_j\right)^2 - q\Psi_2 < 0$ and L3O is unbounded when $\left(\sum_{i\neq j}^n P_{ij} X_i X_j\right)^2 - qB_2 < 0$. In the simple judges case without covariates, the expected coefficients can be compared. With

$$E\left[\Psi_2\right] = \sum_i \left(\left(\sum_{j\neq i} P_{ij} R_j\right)^2 + \sum_{j\neq i} P_{ij}^2 E\left[\eta_j^2\right]\right) E[X_i^2] + \sum_i \sum_{j\neq i} P_{ij}^2 E[X_i^2] E[X_j^2],$$

the difference is:

$$\begin{aligned} E\left[\Psi_2\right] - E\left[B_2\right] &= \sum_i \left(\sum_{j\neq i} P_{ij}R_j\right)^2 R_i^2 + \sum_i \sum_{j\neq i} P_{ij}^2 \left(R_i^2 R_j^2 + 3E\left[\eta_i^2\right] R_j^2\right) \\ &\quad - 3\sum_i \left(\sum_{j\neq i} P_{ij}R_j\right)^2 E\left[\eta_i^2\right] \\ &= \sum_i M_{ii} R_i^2 \left(R_i^2 - 3\left(1 - 2P_{ii}\right) E\left[\eta_i^2\right]\right), \end{aligned}$$

where the second equality uses the same trick as before.

Derivation of LM Variance.

Proof of Equation (1.8). Expanding the variance,

$$\begin{aligned} \operatorname{Var}\left(\sum_i \sum_{j\neq i} G_{ij} e_i X_j\right) &= E\left[\left(\sum_i \sum_{j\neq i} G_{ij} e_i X_j\right)^2\right] = E\left[\sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij} e_i X_j G_{kl} e_k X_l\right] \\ &= \sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij} G_{kl} E\left[\nu_i X_j \nu_k X_l\right] + \sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij} G_{kl} E\left[\nu_i X_j R_{\Delta k} X_l\right] \\ &\quad + \sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij} G_{kl} E\left[R_{\Delta i} X_j \nu_k X_l\right] + \sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij} G_{kl} E\left[R_{\Delta i} X_j R_{\Delta k} X_l\right] \end{aligned}$$

The first term is:

$$
\begin{aligned}
&\sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij}G_{kl}E\left[\nu_i X_j \nu_k X_l\right] \\
&\quad = \sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij}G_{kl}E\left[\nu_i R_j \nu_k R_l + \nu_i \eta_j \nu_k R_l + \nu_i R_j \nu_k \eta_l + \nu_i \eta_j \nu_k \eta_l\right] \\
&\quad = \sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij}G_{kl}E\left[\nu_i R_j \nu_k R_l + \nu_i \eta_j \nu_k \eta_l\right] \\
&\quad = \sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\nu_i^2\right] G_{ij}G_{ik}R_j R_k + \sum_i \sum_{j\neq i} \left(\sum_{l\neq i} G_{ij}G_{il}E\left[\nu_i \eta_j \nu_i \eta_l\right] + \sum_{l\neq j} G_{ij}G_{jl}E\left[\nu_i \eta_j \nu_j \eta_l\right] \right) \\
&\qquad + \sum_i \sum_{j\neq i} \left(\sum_{k\neq i,j} G_{ij}G_{ki}E\left[\nu_i \eta_j \nu_k \eta_i\right] + \sum_{k\neq i,j} G_{ij}G_{kj}E\left[\nu_i \eta_j \nu_k \eta_j\right] \right) \\
&\quad = \sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\nu_i^2\right] G_{ij}G_{ik}R_j R_k + \sum_i \sum_{j\neq i} \left(G_{ij}^2 E\left[\nu_i^2 \eta_j^2\right] + G_{ij}G_{ji}E\left[\nu_i \eta_i \eta_j \nu_j\right] \right) \\
&\quad = \sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\nu_i^2\right] G_{ij}G_{ik}R_j R_k + \sum_i \sum_{j\neq i} \left(G_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] + G_{ij}G_{ji}E\left[\nu_i \eta_i\right] E\left[\eta_j \nu_j\right] \right)
\end{aligned}
$$

In the next few terms, the expansion steps are analogous, so intermediate steps are omitted for brevity. The second to fourth terms can be expressed as:

$$
\begin{aligned}
&\sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij}G_{kl}E\left[\nu_i X_j R_{\Delta k} X_l\right] = \sum_i E\left[\nu_i \eta_i\right] \sum_{j\neq i} G_{ij}R_j \sum_{k\neq i} G_{ki}R_{\Delta k}, \\
&\sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij}G_{kl}E\left[R_{\Delta i} X_j \nu_k X_l\right] = \sum_i \sum_{j\neq i} \sum_{l\neq i} G_{ji}G_{il}E\left[\eta_i \nu_i\right] R_{\Delta j} R_l, \text{ and} \\
&\sum_i \sum_{j\neq i} \sum_k \sum_{l\neq k} G_{ij}G_{kl}E\left[R_{\Delta i} X_j R_{\Delta k} X_l\right] = \sum_i E\left[\eta_i^2\right] \sum_{j\neq i} \sum_{k\neq i} G_{ji}G_{ki}R_{\Delta j}R_{\Delta k}.
\end{aligned}
$$

The expression stated in the equation combines these expressions. □

1.C.2 Further Details for Power

Further Analytic Results

Lemma 1.7. *$(s_1', s_2')'$ are sufficient statistics for $(\pi_Y', \pi')'$. Further, for transformations of the form $Z \to ZF'$ where F is a $K \times K$ orthogonal matrix, $(s_1's_1, s_1's_2, s_2's_2)$ is a maximal invariant, and*

$$\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \sim N\left(\begin{pmatrix} (Z'Z)^{1/2}\pi_Y \\ (Z'Z)^{1/2}\pi \end{pmatrix}, \Omega \otimes I_K \right).$$

Proposition 1.2. *With Equation (1.12), $K \to \infty$ and $\frac{1}{\sqrt{K}}(\pi_Y'Z'Z\pi_Y, \pi'Z'Z\pi_Y, \pi'Z'Z\pi) \to (C_{YY}, C_Y, C_S)$,*

$$\frac{1}{\sqrt{K}} \begin{pmatrix} s_1's_1 - K\omega_{\zeta\zeta} - C_{YY} \\ s_1's_2 - K\omega_{\zeta\eta} - C_Y \\ s_2's_2 - K\omega_{\eta\eta} - C_S \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right) \tag{1.21}$$

for some variance matrix Σ. If $C_{YY}, C_Y, C_S < \infty$,

$$\Sigma = \begin{pmatrix} 2\omega_{\zeta\zeta}^2 & 2\omega_{\zeta\eta}\omega_{\zeta\zeta} & 2\omega_{\zeta\eta}^2 \\ 2\omega_{\zeta\eta}\omega_{\zeta\zeta} & \omega_{\zeta\zeta}\omega_{\eta\eta} + \omega_{\zeta\eta}^2 & 2\omega_{\zeta\eta}\omega_{\eta\eta} \\ 2\omega_{\zeta\eta}^2 & 2\omega_{\zeta\eta}\omega_{\eta\eta} & 2\omega_{\eta\eta}^2 \end{pmatrix}.$$

The proof of Proposition 1.2 relies on $K \to \infty$ because objects like $s_1's_1$ can be written as a sum of K objects. With an appropriate representation to obtain independence, a CLT can be applied to yield normality. Compared to MS22, Proposition 1.2 does not require constant treatment effects and characterizes the distribution without orthogonalizing the sufficient statistics. Nonetheless, the form of the covariance matrix is similar to MS22.

Instead of using the maximal invariant, we use the L1O analog

$$(T_{YY}, T_{YX}, T_{XX}) := \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij}(Y_i Y_j, Y_i X_j, X_i X_j),$$

which relates to the JIVE directly as $\hat{\beta}_{JIVE} = T_{YX}/T_{XX}$, so the asymptotic problem is:

$$\begin{pmatrix} T_{YY} \\ T_{YX} \\ T_{XX} \end{pmatrix} \sim N(\mu, \Sigma), \mu = \begin{pmatrix} \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij} R_{Yi} R_{Yj} \\ \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij} R_{Yi} R_j \\ \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} P_{ij} R_i R_j \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \cdot & \sigma_{22} & \sigma_{23} \\ \cdot & \cdot & \sigma_{33} \end{pmatrix}. \tag{1.22}$$

With abuse of notation, μ and Σ refer to the asymptotic mean and variances of (T_{YY}, T_{YX}, T_{XX}) instead of (T_{AR}, T_{LM}, T_{FS}) so that the statistics do not depend on the hypothesized null, but these are identical when $\beta_0 = 0$. Using the same argument as Section 1.4, $\mu_1, \mu_3 \geq 0$ and $\mu_2^2 \leq \mu_1 \mu_3$.

Even with covariates, if the regression is fully saturated with G given by UJIVE, Proposition 1.3 below shows that the same inequality restrictions hold. To be precise, saturation is defined in Section 2 of Evdokimov and Kolesár (2018). All individuals can be partitioned into L covariate groups, so with group index $G_i \in \{1, \cdots, L\}$, we have covariates $W_{i,l} = 1\{G_i = l\}$. We also have an instrument S_i that takes $M + 1$ possible values in each group, and these values for every group l are labeled $s_{l0}, \cdots, s_{lM}$. Then, the vector of instruments has dimension $K = ML$ and $Z_{i,lm} = 1\{S_i = s_{lm}\}$. Adapting Equation (1.22) to the case with covariates,

$$\mu = \left(\frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} G_{ij} R_{Yi} R_{Yj}, \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} G_{ij} R_{Yi} R_j, \frac{1}{\sqrt{K}} \sum_i \sum_{j \neq i} G_{ij} R_i R_j \right)'.$$

Proposition 1.3. *If μ is defined such that $G = (I - \text{diag}(H_Q))^{-1} H_Q - (I - \text{diag}(H_W))^{-1} H_W$, and the regression is fully saturated, then $\mu_1 \geq 0, \mu_3 \geq 0$ and $\mu_2^2 \leq \mu_1 \mu_3$.*

Using the asymptotic problem of Equation (1.22), testing $H_0 : \mu_2/\mu_3 = \beta^*$ is identical to testing $H_0 : \mu_2 - \beta^* \mu_3 = 0$. Since β^* is fixed, and I consider alternatives of the form: $H_A : \mu_2 - \beta^* \mu_3 = h_A$. The LM statistic corresponds to $T_{YX} - \beta^* T_{XX}$, so it can be used to test the null directly. I focus on the most common case of $\beta^* = 0$, and it is analogous to extend the argument for $\beta^* \neq 0$. Let μ^A denote the mean under the alternative and μ^H under the null. The remainder of this section presents theoretical results for power, and numerical results beyond the environment covered by theory are relegated to Section 1.C.2.

The one-sided test is the most powerful test for testing against a particular subset of alternatives $\mathcal{S} := \left\{ \left(\mu_1^A, \mu_2^A, \mu_3^A\right) : \mu_1^A - \frac{\sigma_{12}}{\sigma_{22}} \mu_2^A \geq 0, \mu_3^A - \frac{\sigma_{23}}{\sigma_{22}} \mu_2^A \geq 0 \right\}$. While $\mathcal{S}$ may not be empirically interpretable, this set is constructed so that standard Lehmann and Romano (2005) arguments can be applied to conclude that the one-sided LM test is the most powerful test. The proposition makes no statement about alternative hypotheses that are not in $\mathcal{S}$. A more powerful test can be constructed when μ_2^A is large and covariance σ_{23}, σ_{12} are large.

Proposition 1.4. *The one-sided LM test is the most powerful test for testing any alternative hypothesis $\left(\mu_1^A, \mu_2^A, \mu_3^A\right) \in \mathcal{S}$ in the asymptotic problem of Equation (1.22).*

For a given $\left(\mu_1^A, \mu_2^A, \mu_3^A\right)$ in the alternative space, LM (which just uses the second element) is justified as being most powerful because it is identical to the Neyman-Pearson test when testing against a point null μ^H with $\mu_1^H = \mu_1^A - \frac{\sigma_{12}}{\sigma_{22}} \mu_2^A$, $\mu_2^H = 0$ and $\mu_3^H = \mu_3^A - \frac{\sigma_{23}}{\sigma_{22}} \mu_2^A$. The inequalities in $\mathcal{S}$ are imposed so that $\mu_1^H, \mu_3^H \geq 0$, ensuring that μ^H is in the null space, so LM is the most powerful test. In contrast, if the inequalities fail in the alternative space, then $(\mu_1^A - \frac{\sigma_{12}}{\sigma_{22}} \mu_2^A, 0, \mu_3^A - \frac{\sigma_{23}}{\sigma_{22}} \mu_2^A)$ is not in the null space, and the Lehmann and Romano (2005) argument cannot be applied.

Existence of Structural Model

This section presents a structural model, then argues that any reduced-form model in the form of Equation (1.22) can be justified by this structural model.

Example 1.1. *Consider a linear potential outcomes model with an instrument Z that is a vector of indicators for judges, each with $c = 5$ cases, a continuous endogenous variable X, and outcome Y:*

$$X_i(z) = z'\pi + v_i, \qquad Y_i(x) = x\,(\beta + \xi_i) + \varepsilon_i, \text{ and}$$
$$\begin{pmatrix} \varepsilon_i \\ \xi_i \\ v_i \end{pmatrix} \mid k(i) = k \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\varepsilon\varepsilon} & \sigma_{\varepsilon\xi} & \sigma_{\varepsilon v} \\ \cdot & \sigma_{\xi\xi} & \sigma_{\xi vk} \\ \cdot & \cdot & \sigma_{vv} \end{pmatrix}\right). \tag{1.23}$$

Due to the judge design, $X_i = \pi_{k(i)} + v_i$, where $k(i)$ is the judge that observation i is assigned to. The strength of the instrument is $C_S = \frac{1}{\sqrt{K}}\sum_k (c-1)\pi_k^2$. The π_k's are constructed as such: with $s = \sqrt{C_S/\sqrt{K}/(c-1)}$, set $\pi_k = 0$ for the base judge, $\pi_k = -s$ for half the judges and $\pi_k = s$ for the other half. The heterogeneity covariances $\sigma_{\xi vk}$ are constructed so that $\sum_k \pi_k = 0, \sum_k \sigma_{\xi vk} = 0$, and $\sum_k \pi_k \sigma_{\xi vk} = 0$. With C_H characterizing the heterogeneity in the model, and $h = \sqrt{C_H/\sqrt{K}/(c-1)}$, set $\sigma_{\xi vk} = 0$ of the base judge; among judges with $\pi_k = s$, half of them have $\sigma_{\xi vk} = h$ and the other half $\sigma_{\xi vk} = -h$. The same construction of $\sigma_{\xi vk}$ applies for judges with $\pi_k = -s$.

In this model, the individual treatment effect is $\beta_i = \beta + \xi_i$. We can interpret v_i as the noise associated with the first-stage regression, ε_i as the noise in the intercept of the outcome equation, and ξ_i as the individual-level treatment effect heterogeneity. Further, $\sigma_{\xi vk}$ characterizes the extent of treatment effect heterogeneity. The observed outcome in a model with constant treatment effects is $Y_i(X_i) = X_i\beta + \check{\varepsilon}_i$, with $E[\check{\varepsilon}_i]$=0. When $\sigma_{\xi vk} = 0$, regardless of the values of $\sigma_{\varepsilon\xi}, \sigma_{\xi\xi}$, the observed outcome of Equation (1.23) can be written

as $Y_i(X_i) = X_i\beta + \check{\varepsilon}_i$ where $E[\check{\varepsilon}_i] = E[X_i\xi_i + \varepsilon_i] = E[X_i E[\xi_i \mid X_i]] = 0$, which resembles the constant treatment effect case.

Lemma 1.8. *Consider the model of Example 1.1. If $\sqrt{K}s^2 \to \tilde{C}_S < \infty$ and $\sqrt{K}h^2 \to \tilde{C}_H < \infty$, then*

$$\sigma_{11} = \frac{4}{\sigma_{33}}\left(\sigma_{22} - \frac{\sigma_{23}^2}{2\sigma_{33}}\right)^2 + o(1), \quad \sigma_{12} = 2\frac{\sigma_{23}}{\sigma_{33}}\left(\sigma_{22} - \frac{\sigma_{23}^2}{2\sigma_{33}}\right) + o(1), \quad \sigma_{13} = \frac{\sigma_{23}^2}{\sigma_{33}} + o(1),$$
$$\sigma_{22} = \frac{c-1}{c}\left(\sigma_{vv}\left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\varepsilon v}\beta\right) + \left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)^2\right) + o(1),$$
$$\sigma_{33} = 2\frac{c-1}{c}\sigma_{vv}^2 + o(1), \quad \sigma_{23} = 2\frac{c-1}{c}\sigma_{vv}\left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right) + o(1), \text{ and}$$
$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} \sqrt{K}\left(c-1\right)\left(s^2\beta^2 + h^2\right) \\ \sqrt{K}\left(c-1\right)s^2\beta \\ \sqrt{K}\left(c-1\right)s^2 \end{pmatrix} = \left(c-1\right)\begin{pmatrix} C_S\beta^2 + C_H \\ C_S\beta \\ C_S \end{pmatrix}.$$

Proposition 1.5. *In the model of Example 1.1 with $\sqrt{K}s^2 \to \tilde{C}_S < \infty$ and $\sqrt{K}h^2 \to \tilde{C}_H < \infty$, for any $\sigma_{22}, \sigma_{23}, \sigma_{33}$ such that $\sigma_{22}, \sigma_{33} > 0$, $\sigma_{23}^2 \leq \sigma_{22}\sigma_{33}$ and μ such that $\mu_1 \geq 0, \mu_3 > 0$, $\mu_2^2 \leq \mu_1\mu_3$, the following values of structural parameters:*

$$\tilde{C}_S = \mu_3/\left(c-1\right), \quad \beta = \mu_2/\mu_3, \quad h = \sqrt{\frac{1}{\sqrt{K}}\frac{1}{c-1}\left(\mu_1 - \frac{\mu_2^2}{\mu_3}\right)},$$
$$\Sigma_{SF} = \begin{pmatrix} \sigma_{\varepsilon\varepsilon} & \sigma_{\varepsilon\xi} & \sigma_{\varepsilon v} \\ . & \sigma_{\xi\xi} & \sigma_{\xi vk} \\ . & . & \sigma_{vv} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_{vv}}\frac{c}{c-1}\left(\sigma_{22} - \frac{\sigma_{23}^2}{\sigma_{33}}\right) + \frac{\sigma_{\varepsilon v}^2}{\sigma_{vv}} & 0 & \sigma_{\varepsilon v} \\ . & \frac{h}{\sigma_{vv}} & \pm h \\ . & . & \sigma_{vv} \end{pmatrix},$$
$$\sigma_{vv} = \sqrt{\frac{\sigma_{33}c}{2\left(c-1\right)}}, \text{ and } \quad \sigma_{\varepsilon v} = \frac{1}{\sigma_{vv}}\left(\frac{\sigma_{23}c}{2\left(c-1\right)} - \sigma_{vv}^2\beta\right),$$

satisfy the equations in Lemma 1.8, and $\det\left(\Sigma_{SF}\right)/h \to C_D \geq 0$.

Due to Proposition 1.5, since the principal submatrices of Σ_{SF} are positive semidefinite asymptotically, Σ_{SF} is a symmetric positive semidefinite matrix asymptotically. The propo-

sition thus implies that when the σ's and μ satisfy the conditions, there exists structural parameters that can generate the given μ and Σ asymptotically. Hence, there are no further restrictions on μ from the observed Σ in Example 1.1.

Numerical Results for Power

This section presents numerical results for power in environments not covered by the theory. I first consider one-sided tests beyond the set $\mathcal{S}$ covered by the theory, then weighted average power for two-sided tests rather than the class of unbiased tests.

The power envelope is achieved by a test that is valid across the entire composite null space, and is most powerful for testing against a particular point in the alternative space. To obtain this test, I implement the algorithm from Elliott et al. (2015) (EMW) where all weight on the alternative are placed on a single point while being valid across a composite null. Then, testing against every point in the alternative space requires a different critical value. For the numerical exercises in this subsection, I use a Σ matrix of the form:

$$\Sigma = \begin{pmatrix} 2 & 2\rho & 2\rho^2 \\ \cdot & 1+\rho^2 & 2\rho \\ \cdot & \cdot & 2 \end{pmatrix}, \tag{1.24}$$

which corresponds to the Σ matrix in Proposition 1.2 with $\omega_{\zeta\zeta} = \omega_{\eta\eta} = 1, \omega_{\zeta\eta} = \rho$.

In the numerical exercises, I display the rejection rate across 500 independent draws from $X^* \sim N(\mu, \Sigma)$ at each point on the μ_2 axis, across several μ_1, μ_3 values for a 5% test. The composite null uses a grid of $\mu_1 \in [0, 5], \mu_3 \in [0, 5]$ in 0.5 increments, and assumes the variance is known.

Figure 1.C.1 uses a one-sided LM test, with a large covariance at $\rho = 0.9$. When data is generated from the null, since LM and EMW are valid tests, their rejection rate is at most 0.05. EMW has exact size when testing a weighted average of values in the null space and

is valid across the entire space, so when data is generated from one particular point in the null, EMW can be conservative. Consistent with Proposition 1.4, when μ_2 is small enough for $\mu_1 = 1, \mu_3 = 4$, LM achieves the power envelope, but as μ_2 gets larger, the gap widens substantially. This phenomenon occurs because EMW still uses the same null grid, but now it no longer needs to have correct size for testing against the point $(\mu_1^A - \frac{\sigma_{12}}{\sigma_{22}}\mu_2^A, 0, \mu_3^A - \frac{\sigma_{23}}{\sigma_{22}}\mu_2^A)$, as that point is no longer in the null space.

In Figure 1.C.2, Σ is calibrated by using the Σ matrix calculated from the Angrist and Krueger (1991) application, so after appropriate normalizations, $\rho = 0.37$. With such a low covariance, LM is basically indistinguishable from the EMW bound. Hence, even though there are gains to be made theoretically, in the empirical application considered, the gains are small.

Instead of considering a point alternative, we may be more interested in testing against a composite alternative. Here, the alternative grid for EMW places equal weight on alternatives $(\mu_1^A, \mu_2^A, \mu_3^A) \in [0, 5] \times [-2, 2] \times [0, 5]$ in increments of 0.5 (excluding $\mu_2 = 0$) subject to inequality constraints. Figures 1.C.3 and 1.C.4 present one such possibility by allowing EMW to place equal weight on several points within the alternative space. The resulting test is the nearly optimal test for a weighted average of values the null space against the uniformly weighted average of alternative values. Hence, there is no guarantee that its power is necessarily higher than the LM test at every point in the alternative space. While there are weighted-average power curves that substantially outperform LM, this result is compatible with Proposition 1.1. EMW is a biased test as there are points in the alternative space that are not a part of the grid where LM outperforms EMW. Nonetheless, Figure 1.C.4 suggests that, when using the empirical covariance, LM does not perform substantially worse than EMW.

Figure 1.C.1: One-sided test with $\rho = 0.9$

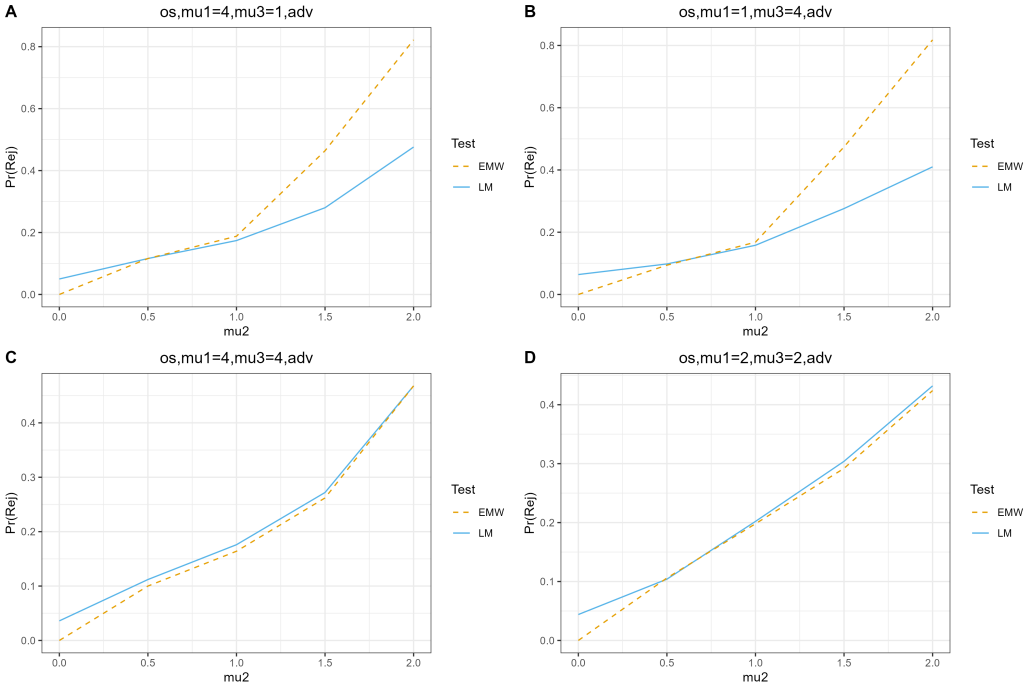


Figure 1.C.2: One-sided test with $\rho = 0.37$

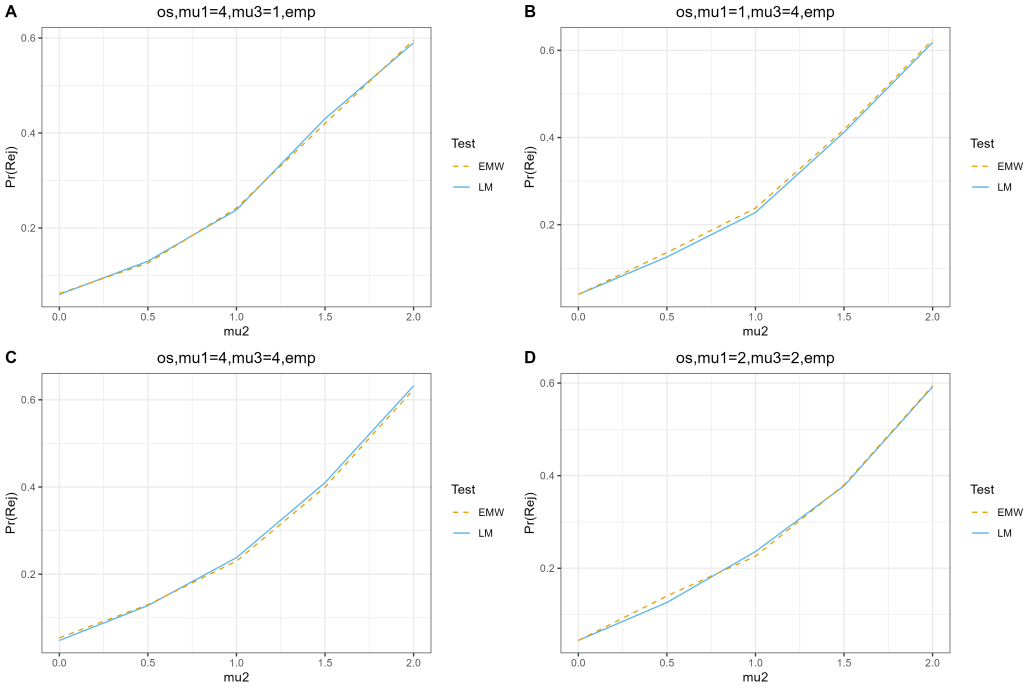


Figure 1.C.3: Uniform Weighting on grid of alternatives with $\rho = 0.9$

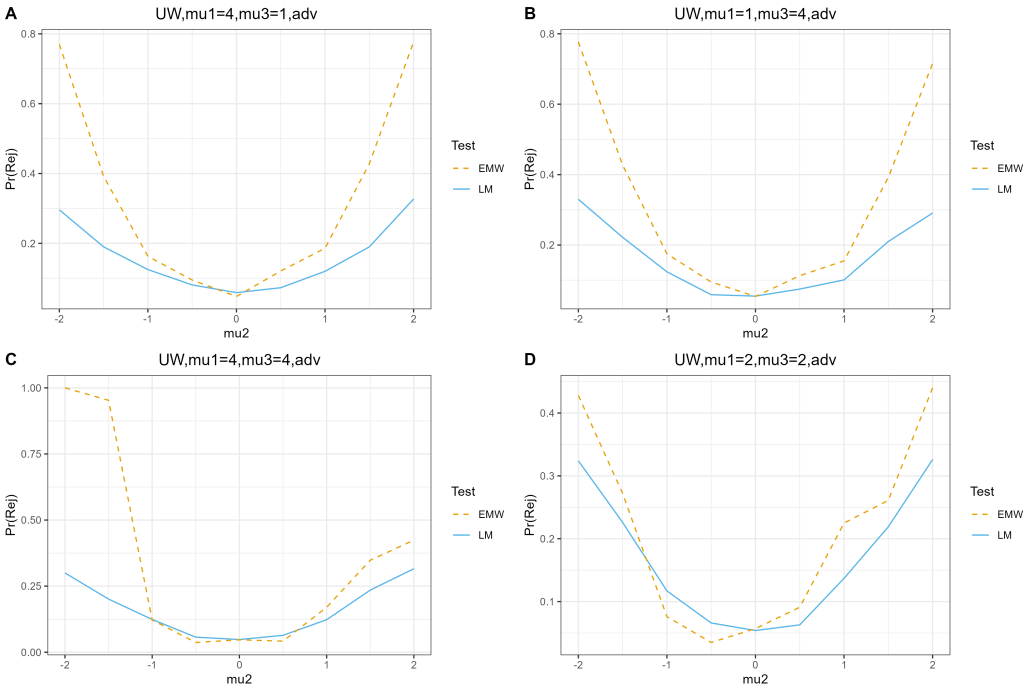
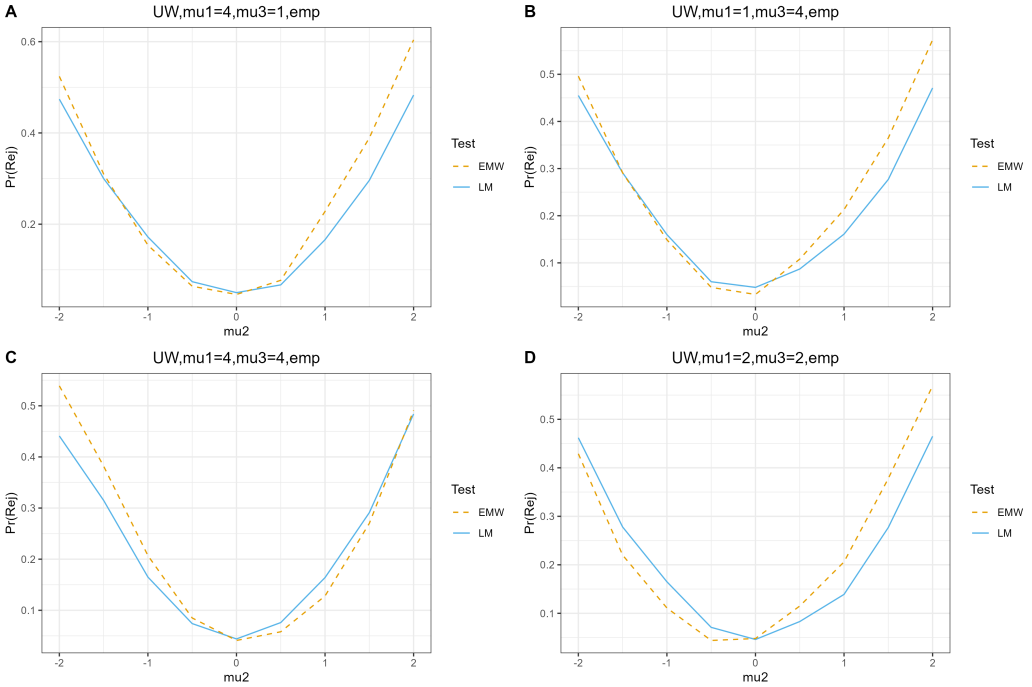


Figure 1.C.4: Uniform Weighting on grid of alternatives with $\rho = 0.37$



1.C.3 Constructing Confidence Sets

Expressions for the test are given in Section 1.3, which can be efficiently implemented using matrix operations. Inverting the test to obtain a confidence set is also straightforward in this procedure, as the bounds of the confidence set are derived in closed-form in this section.

To invert the LM test to obtain a confidence set, use $e_i = Y_i - X_i\beta_0$ and expand the A expressions in Equation (1.9) so that they are written in terms of X and Y. The two-sided test rejects: $\left(\sum_i \sum_{j\neq i} G_{ij} e_i X_j\right)^2 / \hat{V}_{LM} \geq q = \Phi(1-\alpha/2)^2$. Let $T_{YX} := \frac{1}{\sqrt{K}} \sum_i \sum_{j\neq i} G_{ij} Y_i X_j$. Then, $\sum_i \sum_{j\neq i} G_{ij} e_i X_j = \sqrt{K}\,(T_{YX} - T_{XX}\beta_0)$, so squaring it results in a term that is quadratic in β_0^2. With $\hat{V}_{LM} = B_0 + C_1\beta_0 + B_2\beta_0^2$ quadratic in β_0, the analysis for the shape of the confidence intervals is similar to the AR procedure for just-identified IV. Calculations for coefficients is similar to that of the L3O variance.

Proposition 1.6. *The two-sided LM test does not reject β_0 when $(KT_{XX}^2 - qB_2)\,\beta_0^2 - (2KT_{YX}T_{XX} + qB_1)\,\beta_0 + (KT_{YX}^2 - qB_0) \leq 0$. Let*

$$D := (2KT_{YX}T_{XX} + qB_1)^2 - 4\left(KT_{XX}^2 - qB_2\right)\left(KT_{YX}^2 - qB_0\right).$$

If $D \geq 0$ and $KT_{XX}^2 - qB_2 \geq 0$, then the upper and lower bounds of confidence set are:

$$\frac{(2KT_{YX}T_{XX} + qB_1) \pm \sqrt{D}}{2\left(KT_{XX}^2 - qB_2\right)}.$$

If $D < 0$ and $KT_{XX}^2 - qB_2 < 0$, then the confidence set is empty. Otherwise, the confidence set is unbounded.

Due to $+qB_1, -qB_2$ in the expression of the upper and lower bounds, the confidence set is not necessarily centered around $\hat{\beta}_{JIVE} = T_{YX}/T_{XX}$.

1.C.4 Further Simulation Results

This section reports simulation results from several structural models to assess how well various procedures control for size. Since the nominal size is 0.05, and data is generated under the null, the target rejection rate is 0.05. Across the board, the L3O method performs well, and for all existing procedures, there is at least one design where they perform badly.

Continuous Treatment

This subsection reports results for a simulation based on Example 1.1 that has a continuous X. Table 1.C.1 reports results with $K = 500$ and Table 1.C.2 reports results for $K = 40$. The L3O rejection rates are closer to the nominal rate than the existing procedures in the literature, albeit worse with a smaller K. ARorc has high rejection rates with strong heterogeneity and EK has high rejection rates with weak instruments. Notably, with perfect correlation and an irrelevant instrument, EK can achieve 100% rejection in the simulation with $K = 500$. The procedures that use the LM statistic are MO, $\tilde{X}$-AR, L3O and LMorc; they differ only in variance estimation. Hence, while $\tilde{X}$-AR and MO over-reject, the extent of over-rejection is smaller than ARorc and EK in the adversarial cases.

Binary Treatment

This subsection presents a structural model with a binary X. Data is generated from a judge model with $J = K + 1$ judges, each with $c = 5$ cases, and cases are indexed by i. The structural model is:

$$Y_i(x) = x(\beta + \xi_i) + \varepsilon_i, \text{ and}$$
$$X_i(z) = I\left\{z'\pi - v_i \geq 0\right\}.$$

Table 1.C.1: Rejection rates under the null for nominal size 0.05 test for continuous X

	TSLS	EK	ARorc	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O	LMorc
$C_H = C_S = 3\sqrt{K}, \sigma_{\varepsilon v} = 0$	0.061	0.017	1.000	0.061	0.079	0.078	0.042	0.044
$C_H = 2\sqrt{K}, C_S = 2\sqrt{K}$	0.952	0.022	1.000	0.073	0.087	0.084	0.058	0.055
$C_H = 2\sqrt{K}, C_S = 2$	1.000	0.009	1.000	0.096	0.076	0.127	0.053	0.050
$C_H = 2\sqrt{K}, C_S = 0$	1.000	0.006	1.000	0.103	0.061	0.127	0.059	0.052
$C_H = 3, C_S = 3\sqrt{K}$	0.986	0.033	0.109	0.057	0.062	0.064	0.056	0.047
$C_H = 3, C_S = 3$	1.000	0.036	0.168	0.055	0.078	0.087	0.055	0.047
$C_H = 3, C_S = 0$	1.000	0.048	0.184	0.058	0.106	0.088	0.053	0.057
$C_H = 0, C_S = 2\sqrt{K}$	1.000	0.089	0.049	0.063	0.083	0.080	0.061	0.058
$C_H = 0, C_S = 2$	1.000	0.207	0.045	0.054	0.243	0.135	0.057	0.045
$C_H = 0, C_S = 0$	1.000	0.337	0.051	0.042	0.413	0.127	0.045	0.048
$C_H = C_S = 0, \sigma_{\varepsilon v} = 1$	1.000	1.000	0.044	0.042	1.000	0.157	0.052	0.044

Notes: Data generating process corresponds to Example 1.1. Unless mentioned otherwise, simulations use $K = 500, c = 5, \beta = 0, \sigma_{\varepsilon\varepsilon} = \sigma_{vv} = 1, \sigma_{\varepsilon\xi} = 0, = \sigma_{\varepsilon v} = 0.8, \sigma_{\xi\xi} = 1 + h$ for $h^2 < 1$ with 1000 simulations. The table displays rejection rates of various procedures (in columns) for various designs (in rows). $C_H = 0$ uses $\xi_i = 0$ for all i, which uses $\sigma_{\xi\xi} = \sigma_{\varepsilon\xi} = \sigma_{\xi v} = 0$, corresponding to constant treatment effects. Procedures are described in Table 1.1.

Table 1.C.2: Rejection Rates under the null for nominal size 0.05 test for Continuous X with $K = 40$

	TSLS	EK	ARorc	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O	LMorc
$C_H = C_S = 3\sqrt{K}, \sigma_{\varepsilon v} = 0$	0.072	0.022	0.525	0.051	0.074	0.068	0.039	0.055
$C_H = 2\sqrt{K}, C_S = 2\sqrt{K}$	0.238	0.034	0.388	0.051	0.074	0.077	0.055	0.062
$C_H = 2\sqrt{K}, C_S = 2$	0.547	0.033	0.475	0.083	0.096	0.133	0.077	0.053
$C_H = 2\sqrt{K}, C_S = 0$	0.651	0.013	0.511	0.072	0.088	0.102	0.068	0.054
$C_H = 3, C_S = 3\sqrt{K}$	0.213	0.025	0.109	0.048	0.057	0.063	0.055	0.046
$C_H = 3, C_S = 3$	0.658	0.032	0.129	0.045	0.074	0.063	0.064	0.055
$C_H = 3, C_S = 0$	0.849	0.049	0.127	0.063	0.109	0.103	0.087	0.057
$C_H = 0, C_S = 2\sqrt{K}$	0.853	0.105	0.049	0.064	0.068	0.098	0.085	0.056
$C_H = 0, C_S = 2$	0.999	0.152	0.048	0.045	0.201	0.132	0.098	0.037
$C_H = 0, C_S = 0$	1.000	0.342	0.052	0.051	0.439	0.143	0.080	0.049
$C_H = C_S = 0, \sigma_{\varepsilon v} = 1$	1.000	1.000	0.045	0.040	1.000	0.179	0.082	0.045

Note: Designs are identical to Table 1.C.1, but $K = 40$ here.

Table 1.C.3: Rejection Rates under the null for nominal size 0.05 test for binary X

	TSLS	EK	ARorc	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O	LMorc
$C_H = C_S = 3\sqrt{K}, \sigma_{\varepsilon v} = 0$	0.046	0.049	0.059	0.045	0.045	0.045	0.049	0.054
$C_H = 2\sqrt{K}, C_S = 2\sqrt{K}$	0.097	0.047	0.177	0.037	0.038	0.041	0.051	0.052
$C_H = 2\sqrt{K}, C_S = 2$	0.727	0.059	1.000	0.127	0.051	0.143	0.058	0.051
$C_H = 2\sqrt{K}, C_S = 0$	0.891	0.037	1.000	0.204	0.067	0.247	0.059	0.045
$C_H = 3, C_S = 3\sqrt{K}$	0.092	0.060	0.051	0.055	0.057	0.056	0.055	0.047
$C_H = 3, C_S = 3$	0.996	0.089	0.888	0.059	0.086	0.096	0.055	0.048
$C_H = 3, C_S = 0$	1.000	0.124	0.999	0.101	0.289	0.181	0.068	0.052
$C_H = 0, C_S = 2\sqrt{K}$	0.408	0.058	0.055	0.043	0.046	0.046	0.045	0.041
$C_H = 0, C_S = 2$	1.000	0.212	0.052	0.061	0.188	0.108	0.078	0.057
$C_H = 0, C_S = 0$	1.000	0.654	0.046	0.034	0.750	0.149	0.069	0.039
$C_H = C_S = 0, \sigma_{\varepsilon\varepsilon} = 0$	1.000	1.000	0.053	0.057	1.000	0.173	0.076	0.053

Note: The data generating process corresponds to Section 1.C.4. Unless stated otherwise, designs use $K = 100, c = 5, \beta = 0, p = 7/8$, $\sigma_{\varepsilon\varepsilon} = 0.1, \sigma_{\varepsilon v} = 0.5$ with 1000 simulations.

Our unobservables are generated as follows. Draw $v_i \sim U[-1, 1]$, then generate residuals from:

$$\varepsilon_i \mid v_i \sim \begin{cases} N(\sigma_{\varepsilon v}, \sigma_{\varepsilon\varepsilon}) & if \quad v_i \geq 0 \\ N(-\sigma_{\varepsilon v}, \sigma_{\varepsilon\varepsilon}) & if \quad v_i < 0 \end{cases},$$

$$\xi_i \mid v_i \geq 0 = \begin{cases} \sigma_{\xi vk} & w.p. \quad p \\ -\sigma_{\xi vk} & w.p. \quad 1-p \end{cases}, \text{ and } \quad \xi_i \mid v_i < 0 = \begin{cases} \sigma_{\xi vk} & w.p. \quad 1-p \\ -\sigma_{\xi vk} & w.p. \quad p \end{cases}.$$

The process for determining s, h and $\pi_k \in \{0, -s, s\}, \sigma_{\xi vk} \in \{0, -h, h\}$ are identical to Example 1.1, as s controls the strength of the instrument, h the extent of heterogeneity, and β is the object of interest. Then, the problem's variances and covariances are determined by $(p, \sigma_{\varepsilon v}, \sigma_{\varepsilon\varepsilon})$. The JIVE estimand is shown to be β in Section 1.C.6. A simulation is run with $K = 100$, so the sample size is smaller than the normal experiment in Example 1.1.

Results are presented in Table 1.C.3, and are qualitatively similar to Section 1.2. The oracle test consistently obtains rejection rates close to the nominal 5% rate across all designs, in accordance with the normality result, even with heterogeneous treatment effects and non-

normality of errors due to the binary setup. The L3O rejection rate is close to the nominal rate even with a smaller sample size. EK, ARorc and MO continue to have high rejection rates in the adversarial designs.

Incorporating Covariates

This section presents a data-generating process that involves covariates. Instead of judges, consider a model where there are K states. Let $t = 1, \cdots, K$ index the state and let W denote the control vector that is an indicator for states. With a binary exogenous variable (say an indicator for birth being in the fourth quarter) $B \in \{0, 1\}$, the value of the instrument is given by $k = t \times B$. Then, the instrument vector Z is an indicator for all possible values of k. The structural model is:

$$Y_i(x) = x(\beta + \xi_i) + w'\gamma + \varepsilon_i, \text{ and}$$
$$X_i(z) = I\left\{z'\pi + w'\gamma - v_i \geq 0\right\}.$$

In the simulation, every state has 10 observations, of which 5 have $B = 1$ and the other 5 have $B = 0$. The process for generating $(v_i, \varepsilon_i, \xi_i)$, $\pi_k, \sigma_{\xi vk}$, and s, h is identical to the binary case. Hence, $\pi_0 = \sigma_{\xi v0}$ for the base group, which constitutes half the observations. For $k \neq 0$, π_k is the coefficient for observations from state $t = k$ and have $B = 1$, and $\sigma_{\xi vk}$ is the corresponding heterogeneity term. Whenever $\pi_t = s$, set $\gamma_t = g$; whenever $\pi_t = -s$, set $\gamma_t = -g$. In this setup, it can be shown that the UJIVE estimand is β, and the proof is in Section 1.C.6. Table 1.C.4 reports the associated simulation results, which are qualitatively similar to the results described before.

Table 1.C.4: Rejection Rates under the null for nominal size 0.05 test for binary X with covariates

	TSLS	EK	ARorc	MO	$\tilde{X}$-t	$\tilde{X}$-AR	L3O	LMorc
$C_H = C_S = 3\sqrt{K}, \sigma_{\varepsilon v} = 0$	0.048	0.123	0.049	0.052	0.047	0.055	0.054	0.060
$C_H = 2\sqrt{K}, C_S = 2\sqrt{K}$	0.072	0.111	0.052	0.044	0.041	0.046	0.050	0.053
$C_H = 2\sqrt{K}, C_S = 2$	0.171	0.016	0.471	0.083	0.012	0.092	0.060	0.050
$C_H = 2\sqrt{K}, C_S = 0$	0.259	0.002	0.960	0.126	0.008	0.135	0.047	0.058
$C_H = 3, C_S = 3\sqrt{K}$	0.065	0.132	0.048	0.053	0.056	0.054	0.060	0.049
$C_H = 3, C_S = 3$	0.131	0.015	0.108	0.040	0.003	0.042	0.044	0.050
$C_H = 3, C_S = 0$	0.247	0.003	0.300	0.087	0.004	0.091	0.062	0.053
$C_H = 0, C_S = 2\sqrt{K}$	0.084	0.099	0.054	0.041	0.036	0.043	0.048	0.050
$C_H = 0, C_S = 2$	0.178	0.006	0.058	0.043	0.002	0.044	0.052	0.051
$C_H = 0, C_S = 0$	0.246	0.006	0.048	0.063	0.005	0.069	0.081	0.050
$C_H = C_S = 0, \sigma_{\varepsilon\varepsilon} = 0$	1.000	0.497	0.042	0.013	0.147	0.049	0.092	0.035

Note: The data generating process corresponds to Section 1.C.4. Unless stated otherwise, designs use $K = 48, c = 5, \beta = 0, p = 7/8$, $\sigma_{\varepsilon\varepsilon} = 0.5, \sigma_{\varepsilon v} = 0.1$, and $g = 0.1$ with 1000 simulations.

1.C.5 Proofs for Appendix 1.A and 1.B

Proof of Lemma 1.1. Suppose not. Then, for some real β_0,

$$E\left[T_{AR}\right] = \sum_i \sum_{j \neq i} P_{ij} R_{\Delta i} R_{\Delta j} = \sum_i \sum_{j \neq i} P_{ij} \left(R_{Yi} R_{Yj} - R_i R_{Yj} \beta_0 - R_{Yi} R_j \beta_0 + R_i R_j \beta_0^2 \right) = 0.$$

Solving for β_0,

$$\beta_0 = \frac{2 \sum_i \sum_{j \neq i} P_{ij} R_i R_{Yj} \pm \sqrt{4 \left(\sum_i \sum_{j \neq i} P_{ij} R_i R_{Yj} \right)^2 - 4 \left(\sum_i \sum_{j \neq i} P_{ij} R_i R_j \right) \left(\sum_i \sum_{j \neq i} P_{ij} R_{Yi} R_{Yj} \right)}}{2 \left(\sum_i \sum_{j \neq i} P_{ij} R_i R_j \right)}.$$

In our structural model, $R_i = \pi_{k(i)}$ and $R_{Yi} = \pi_{Yk(i)}$. The term in the square root can be written as:

$$D = 4 \left(\sum_k \pi_k \pi_{Yk} \right)^2 - 4 \left(\sum_k \pi_k^2 \right) \left(\sum_k \pi_{Yk}^2 \right)$$

Using Table 1.A.1, $\sum_k \pi_k^2 = \frac{5}{8}s^2K$, $\sum_k \pi_{Yk}^2 = \left(\frac{5}{8}s^2\beta^2 + h^2\right)K$, and $\sum_k \pi_k\pi_{Yk} = \frac{5}{8}s^2\beta K$, we obtain

$$\frac{1}{4}D = \left(\frac{5}{8}s^2\beta K\right)^2 - \left(\frac{5}{8}s^2K\right)\left(\frac{5}{8}s^2\beta^2 + h^2\right)K = -\frac{5}{8}s^2h^2K^2 \leq 0.$$

Since $h \neq 0$ and $Ks^2 > 0$, there are no real roots of β_0, a contradiction. □

Proof of Lemma 1.2. I rewrite the quadratic term to produce a martingale difference array:

$$\begin{aligned}\sum_i \sum_{j\neq i} G_{ij}v_i'Av_j &= \sum_i \sum_{j<i} G_{ij}v_i'Av_j + \sum_i \sum_{j>i} G_{ij}v_i'Av_j \\ &= \sum_i \sum_{j<i} \left(G_{ij}v_i'Av_j + G_{ji}v_j'Av_i\right).\end{aligned}$$

Hence, $\sum_i s_i'v_i + \sum_i \sum_{j\neq i} G_{ij}v_i'Av_j = \sum_i y_i$, where

$$\begin{aligned}y_i &= s_i'v_i + \sum_{j<i}\left(G_{ij}v_i'Av_j + G_{ji}v_j'Av_i\right) = s_i'v_i + v_i'A\left(\sum_{j<i} G_{ij}v_j\right) + \left(\sum_{j<i} G_{ji}v_j'\right)Av_i \\ &= s_i'v_i + v_i'A\left(G_Lv\right)_{i\cdot}' + \left(G_U'v\right)_{i\cdot}Av_i.\end{aligned}$$

Let $\mathcal{F}_i$ denote the filtration of $y_1, \ldots, y_{i-1}$. To apply the martingale CLT, we require:

1. $\sum_i E\left[|y_i|^{2+\epsilon}\right] \to 0$.
2. Conditional variance converges to 1, i.e., $P\left(|\sum_i E\left[B^2y_i^2 \mid \mathcal{F}_i\right] - 1| > \eta\right) \to 0$, where $B = \mathrm{Var}\left(T\right)^{-1/2}$.

The 4th moments of v_i are bounded. With $\epsilon = 2$, we want $\sum_i E\left[y_i^4\right] \to 0$. Using Loeve's c_r inequality, it suffices that, for any element l of the v_i vector,

$$\sum_i s_{il}^4 E\left[v_{il}^4\right] \to 0, \text{ and } \quad \sum_i E\left[v_{il}^4\left(G_Lv\right)_{il}^4\right] \to 0.$$

The first condition is immediate from condition (2). The second condition holds by condition (3) using the proof in EK18. To be precise,

$$\begin{aligned}
\sum_i E\left[v_{il}^4 \left(G_L v\right)_{il}^4\right] &= \sum_i E\left[v_{il}^4\right] E\left[\left(G_L v\right)_{il}^4\right] \preceq \sum_i E\left[\left(G_L v\right)_{il}^4\right] \\
&= \sum_i \sum_j G_{L,ij}^4 E\left[v_{il}^4\right] + 3\sum_i \sum_j \sum_{k\neq j} G_{L,ij}^2 G_{L,ik}^2 E\left[v_{il}^2\right] E\left[v_{jl}^2\right] \\
&\preceq \sum_i \sum_j \sum_k G_{L,ij}^2 G_{L,ik}^2 = \sum_i \left(G_L G_L'\right)_{ii}^2 \\
&\leq \sum_i \sum_j \left(G_L G_L'\right)_{ij}^2 = ||G_L G_L'||_F^2.
\end{aligned}$$

The argument for G_U is analogous. Now, I turn to showing convergence of the conditional variance. With abuse of notation, let $W_i = s_i' v_i$ and $X_i = v_i' A \left(G_L v\right)_{i\cdot}' + v_i' A \left(G_U' v\right)_{i\cdot}'$. Since $\text{Var}\left(BT\right) = B^2 \sum_i E\left[W_i^2\right] + B^2 \sum_i E\left[X_i^2\right] = 1$,

$$\begin{aligned}
\sum_i E\left[B^2 y_i^2 \mid \mathcal{F}_i\right] - 1 = B^2 \sum_i \left(E\left[X_i^2 \mid \mathcal{F}_i\right] - E\left[X_i^2\right]\right) + 2B^2 \sum_i E\left[W_i X_i \mid \mathcal{F}_i\right] \\
+ B^2 \sum_i \left(E\left[W_i^2 \mid \mathcal{F}_i\right] - E\left[W_i^2\right]\right).
\end{aligned}$$

The previous observations in the filtration do not feature, so $E\left[W_i^2 \mid \mathcal{F}_i\right] - E\left[W_i^2\right] = 0$. It suffices to show that the RHS converges to 0. For the $\sum_i E\left[W_i X_i \mid \mathcal{F}_i\right]$ term,

$$\begin{aligned}
B^2 \sum_i E\left[W_i X_i \mid \mathcal{F}_i\right] &= B^2 \sum_i E\left[W_i \left(v_i' A \left(G_L v\right)_{i\cdot}' + v_i' A \left(G_U' v\right)_{i\cdot}'\right) \mid \mathcal{F}_i\right] \\
&= B^2 \sum_i E\left[W_i v_i' A\right] \left(G_L v\right)_{i\cdot}' + B^2 \sum_i E\left[W_i v_i' A\right] \left(G_U' v\right)_{i\cdot}'.
\end{aligned}$$

It suffices to show that the respective squares converge to 0. Due to bounded fourth moments, and applying the Cauchy-Schwarz inequality repeatedly, for some n-vector δ_v with

$||\delta_v||_2 \leq C$,

$$E\left[\left(\sum_i E[W_i v_i'] A\left(G_L v\right)_{i\cdot}'\right)^2\right] \preceq \delta_v' G_L G_L' \delta_v \leq ||\delta_v||_2^2 ||G_L G_L'||_2 \preceq ||G_L G_L'||_F,$$

and the same argument can be applied to the G_U term. Finally,

$$\sum_i \left(E\left[X_i^2 \mid \mathcal{F}_i\right] - E\left[X_i^2\right]\right) = \sum_i \Bigg(E\left[\left(v_i' A\left(G_L v\right)_{i\cdot}' + v_i' A\left(G_U' v\right)_{i\cdot}'\right)^2 \mid \mathcal{F}_i\right] - E\left[\left(v_i' A\left(G_L v\right)_{i\cdot}' + v_i' A\left(G_U' v\right)_{i\cdot}'\right)^2\right]\Bigg).$$

It suffices to consider the G_L term, as the G_U and cross terms are analogous:

$$\sum_i \left(E\left[\left(v_i' A\left(G_L v\right)_{i\cdot}'\right)^2 \mid \mathcal{F}_i\right] - E\left[\left(v_i' A\left(G_L v\right)_{i\cdot}'\right)^2\right]\right)$$
$$= \sum_i \left(\left(G_L v\right)_{i\cdot} A' E\left[v_i v_i'\right] A\left(G_L v\right)_{i\cdot}' - E\left[\left(G_L v\right)_{i\cdot} A' v_i v_i' A\left(G_L v\right)_{i\cdot}'\right]\right).$$

Since $\sum_i \left(G_L v\right)_{i\cdot} A' E\left[v_i v_i'\right] A\left(G_L v\right)_{i\cdot}'$ is demeaned, it suffices to show that its variance converges to 0. Due to bounded moments,

$$\mathrm{Var}\left(\sum_i \left(G_L v\right)_{i\cdot} A' E\left[v_i v_i'\right] A\left(G_L v\right)_{i\cdot}'\right) \preceq \sum_i \sum_j \left(G_L G_L'\right)^2 = ||G_L G_L'||_F^2,$$

which suffices for the result. □

Proof of Lemma 1.3. I begin with part (c). By applying the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\sum_i C_i \left(\sum_{j\neq i} h_2^A(i,j) R_{mj}\right)\left(\sum_{j\neq i} h_2^B(i,j) R_{mj}\right)\right| \\
&\leq \left(\sum_i C_i \left(\sum_{j\neq i} h_2^A(i,j) R_{mj}\right)^2\right)^{1/2} \left(\sum_i C_i \left(\sum_{j\neq i} h_2^B(i,j) R_{mj}\right)^2\right)^{1/2} \\
&\leq \max_i C_i \left(\sum_i \left(\sum_{j\neq i} h_2^A(i,j) R_{mj}\right)^2\right)^{1/2} \left(\sum_i \left(\sum_{j\neq i} h_2^B(i,j) R_{mj}\right)^2\right)^{1/2} \\
&\leq \max_i C_i \left(\sum_i \tilde{R}_{mi}^2\right)^{1/2} \left(\sum_i \tilde{R}_{mi}^2\right)^{1/2} \leq C \sum_i \tilde{R}_{mi}^2.
\end{aligned}$$

The proof of all other parts are entirely analogous. □

Proof of Lemma 1.4. **Proof of Lemma 1.4(a).**

Using the decomposition from AS23,

$$\begin{aligned}
&\mathrm{Var}\left(\sum_i \sum_{j\neq i} G_{ij} F_{ij} V_{1i} V_{2i} V_{3j} V_{4j}\right) \\
&= \sum_{i\neq j}^n G_{ij}^2 F_{ij}^2 \mathrm{Var}\left(V_{1i} V_{2i} V_{3j} V_{4j}\right) + \sum_{i\neq j}^n G_{ij} F_{ij} G_{ji} F_{ji} Cov\left(V_{1i} V_{2i} V_{3j} V_{4j}, V_{1j} V_{2j} V_{3i} V_{4i}\right) \\
&\quad + \sum_{i\neq j\neq k}^n G_{ij} F_{ij} G_{kj} F_{kj} Cov\left(V_{1i} V_{2i} V_{3j} V_{4j}, V_{1k} V_{2k} V_{3j} V_{4j}\right) \\
&\quad + \sum_{i\neq j\neq k}^n G_{ij} F_{ij} G_{jk} F_{jk} Cov\left(V_{1i} V_{2i} V_{3j} V_{4j}, V_{1j} V_{2j} V_{3k} V_{4k}\right) \\
&\quad + \sum_{i\neq j\neq k}^n G_{ij} F_{ij} G_{ik} F_{ik} Cov\left(V_{1i} V_{2i} V_{3j} V_{4j}, V_{1i} V_{2i} V_{3k} V_{4k}\right) \\
&\quad + \sum_{i\neq j\neq k}^n G_{ij} F_{ij} G_{ki} F_{ki} Cov\left(V_{1i} V_{2i} V_{3j} V_{4j}, V_{1k} V_{2k} V_{3i} V_{4i}\right) \\
&\leq 2\left[\max_{i,j} \mathrm{Var}\left(V_{1i} V_{2i} V_{3j} V_{4j}\right)\right] \sum_i \left(\left(\sum_{j\neq i} G_{ij} F_{ij}\right)^2 + \left(\sum_{j\neq i} G_{ij} F_{ij}\right)\left(\sum_{j\neq i} G_{ji} F_{ji}\right)\right).
\end{aligned}$$

Notice that the terms in $\sum_{i\neq j}^{n}$ are absorbed into the sum over k so that the final expression can be written as $\sum_i \sum_{j\neq i} \sum_{k\neq i}$. Then, due to Assumption 1.3(a) and the Cauchy-Schwarz inequality,

$$\sum_i \left(\sum_{j\neq i} G_{ij}F_{ij}\right)^2 \leq \sum_i \left(\sum_{j\neq i} G_{ij}^2\right)\left(\sum_{j\neq i} F_{ij}^2\right) \leq C\sum_i \sum_{j\neq i} G_{ij}^2,$$

and

$$\begin{aligned}\left|\sum_i \left(\sum_{j\neq i} G_{ij}F_{ij}\right)\left(\sum_{j\neq i} G_{ji}F_{ji}\right)\right| &\leq \left(\sum_i \left(\sum_{j\neq i} G_{ij}F_{ij}\right)^2\right)^{1/2}\left(\sum_i \left(\sum_{j\neq i} G_{ji}F_{ji}\right)^2\right)^{1/2} \\ &\leq C\left(\sum_i \sum_{j\neq i} G_{ij}^2\right)^{1/2}\left(\sum_i \sum_{j\neq i} G_{ji}^2\right)^{1/2} = C\sum_i \sum_{j\neq i} G_{ij}^2.\end{aligned}$$

Proof of Lemma 1.4(b). Expand the term:

$$\sum_{i\neq j\neq k}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}V_{1i}V_{2k}V_{3j}V_{4j} = \sum_{i\neq j\neq k}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\left(R_{1i}R_{2k} + v_{1i}R_{2k} + R_{1i}v_{2k} + v_{1i}v_{2k}\right)V_{3j}V_{4j}.$$

Consider the final sum with 4 stochastic terms. The 6-sums have zero covariances due to independent sampling. The 5-sums also have zero covariances, because at least one of v_1 or v_2 needs to have different indices. Within the 4-sum, the covariance is nonzero only for $j_2 \neq j$. We require i_2 to be equal to either i or k and k_2 the other index. Hence, by bounding

covariances above by Cauchy-Schwarz,

$$
\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}v_{1i}v_{2k}V_{3j}V_{4j}\right)\\
&\quad\leq \max_{i,j,k}\operatorname{Var}\left(v_{1i}v_{2k}V_{3j}V_{4j}\right)\sum_{i}\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\left(G_{ij}F_{ij}G_{il}F_{il}\check{M}_{ik,-ij}\check{M}_{ik,-il}+G_{ij}F_{ij}G_{kl}F_{kl}\check{M}_{ik,-ij}\check{M}_{ki,-kl}\right)\\
&\qquad+\max_{i,j,k}\operatorname{Var}\left(v_{1i}v_{2k}V_{3j}V_{4j}\right)3!\sum_{i\neq j\neq k}^{n}G_{ij}^2F_{ij}^2\check{M}_{ik,-ij}^2\\
&\quad\leq \max_{i,j,k}\operatorname{Var}\left(v_{1i}v_{2k}V_{3j}V_{4j}\right)\left(\sum_{i\neq j\neq k\neq l}^{n}G_{ij}^2G_{il}^2\check{M}_{ik,-ij}^2\right)^{1/2}\left(\sum_{i\neq j\neq k\neq l}^{n}F_{ij}^2F_{il}^2\check{M}_{ik,-ij}^2\right)^{1/2}\\
&\qquad+\max_{i,j,k}\operatorname{Var}\left(v_{1i}v_{2k}V_{3j}V_{4j}\right)\left(\sum_{i\neq j\neq k\neq l}^{n}G_{ij}^2G_{kl}^2\check{M}_{ik,-ij}^2\right)^{1/2}\left(\sum_{i\neq j\neq k\neq l}^{n}F_{ij}^2F_{kl}^2\check{M}_{ik,-ij}^2\right)^{1/2}\\
&\qquad+\max_{i,j,k}\operatorname{Var}\left(v_{1i}v_{2k}V_{3j}V_{4j}\right)3!\sum_{i\neq j\neq k}^{n}G_{ij}^2F_{ij}^2\check{M}_{ik,-ij}^2.
\end{aligned}
$$

To obtain the first inequality, observe that once we have fixed 3 indices, there are 3! permutations of the $v_{1i}v_{2k}V_{3j}V_{4j}$ that we can calculate covariances for. They are all bounded above by the variance. In the various combinations, we may have different combinations of G and F, but they are bounded above by the expression. To be precise, the 3-sum is:

$$
\begin{aligned}
&\sum_{i\neq j\neq k}^{n}G_{ij}F_{ij}\check{M}_{ik,-ij}\left(G_{ij}F_{ij}\check{M}_{ik,-ij}+G_{ik}F_{ik}\check{M}_{ij,-ik}+G_{ji}F_{ji}\check{M}_{jk,-ji}\right)\\
&+\sum_{i\neq j\neq k}^{n}G_{ij}F_{ij}\check{M}_{ik,-ij}\left(G_{jk}F_{jk}\check{M}_{ji,-jk}+G_{ki}F_{ki}\check{M}_{kj,-ki}+G_{kj}F_{kj}\check{M}_{ki,-kj}\right).
\end{aligned}
$$

Apply Cauchy-Schwarz to the sum and apply the commutative property of summations to obtain the upper bound. For instance,

$$
\left(\sum_{i\neq j\neq k}^{n}G_{ij}F_{ij}\check{M}_{ik,-ij}G_{jk}F_{jk}\check{M}_{ji,-jk}\right)^2\leq\left(\sum_{i\neq j\neq k}^{n}G_{ij}^2F_{ij}^2\check{M}_{ik,-ij}^2\right)\left(\sum_{i\neq j\neq k}^{n}G_{jk}^2F_{jk}^2\check{M}_{ji,-jk}^2\right).
$$

Then, observe that $\sum_i \sum_{j\neq i} \sum_{k\neq i,j} G_{jk}^2 F_{jk}^2 \check{M}_{ji,-jk}^2 = \sum_j \sum_{k\neq j} \sum_{i\neq j,k} G_{jk}^2 F_{jk}^2 \check{M}_{ji,-jk}^2$ $= \sum_i \sum_{j\neq i} \sum_{k\neq i,j} G_{ij}^2 F_{ij}^2 \check{M}_{ik,-ij}^2$. Due to AS23 Equation (22), $\sum_l \check{M}_{il-ijk}^2 = O(1)$, so $\sum_{i\neq j\neq k}^n G_{ij}^2 F_{ij}^2 \check{M}_{ik,-ij}^2 \leq C \sum_i \sum_{j\neq i} G_{ij}^2 F_{ij}^2 \leq C \sum_i \sum_{j\neq i} G_{ij}^2$. Similarly, $\sum_{i\neq j\neq k\neq l}^n G_{ij}^2 G_{kl}^2 \check{M}_{ik,-ij}^2 = O(1) \sum_{i\neq j\neq k}^n G_{ij}^2 \check{M}_{ik,-ij}^2 = O(1) \sum_{i\neq j}^n G_{ij}^2$, which delivers the order required.

To deal with 3 stochastic terms,

$$
\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq k}^n G_{ij}F_{ij}\check{M}_{ik,-ij}R_{1i}v_{2k}V_{3j}V_{4j}\right) = \operatorname{Var}\left(\sum_{i\neq j}^n v_{2i}V_{3j}V_{4j}\left(\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k}\right)\right)\\
&\quad\leq \sum_{i\neq j}^n \operatorname{Var}\left(v_{2i}V_{3j}V_{4j}\right)\left(\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k}\right)\left[\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k} + \sum_{k\neq i,j} G_{ki}F_{ki}\check{M}_{kj,-ki}R_{1k}\right]\\
&\quad+ \max_{i,j}\operatorname{Var}\left(v_{2i}V_{3j}V_{4j}\right)\sum_{i\neq j}^n\sum_{l\neq i,j}\left(\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k}\right)\left(\sum_{k\neq i,l} G_{kl}F_{kl}\check{M}_{ki,-kl}R_{1k}\right)\\
&\quad\leq \sum_i\sum_{j\neq i}\operatorname{Var}\left(v_{2i}V_{3j}V_{4j}\right)\left(\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k}\right)\\
&\qquad\left[\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k} + \sum_{k\neq i,j} G_{ki}F_{ki}\check{M}_{kj,-ki}R_{1k}\right]\\
&\quad+ \max_{i,j}\operatorname{Var}\left(v_{2i}V_{3j}V_{4j}\right)\sum_i\left(\sum_{j\neq i}\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k}\right)^2\\
&\qquad- \max_{i,j}\operatorname{Var}\left(v_{2i}V_{3j}V_{4j}\right)\sum_i\sum_{j\neq i}\left(\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k}\right)^2\\
&\quad\leq \max_{i,j}\operatorname{Var}\left(v_{2i}V_{3j}V_{4j}\right)\sum_i\left(\sum_{j\neq i}\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k}\right)^2\\
&\quad+ \sum_{i\neq j}^n \operatorname{Var}\left(v_{2i}V_{3j}V_{4j}\right)\left(\sum_{k\neq i,j} G_{kj}F_{kj}\check{M}_{ki,-kj}R_{1k}\right)\left(\sum_{k\neq i,j} G_{ki}F_{ki}\check{M}_{kj,-ki}R_{1k}\right)
\end{aligned}
$$

To get the first inequality, observe that, if for $l \neq i, j$, we have v_{2l} instead of $V_{3l}V_{4l}$, the covariance must be 0. We can then bound the order by using Assumption 1.3 and Lemma 1.3.

Similarly,

$$\begin{aligned}\operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}v_{1i}R_{2k}V_{3j}V_{4j}\right) &= \operatorname{Var}\left(\sum_{i\neq j}^{n} v_{1i}V_{3j}V_{4j}\left(\sum_{k\neq i,j} G_{ij}F_{ij}\check{M}_{ik,-ij}R_{2k}\right)\right)\\ &\leq \max_{i,j}\operatorname{Var}\left(v_{1i}V_{3j}V_{4j}\right)\sum_{i}\left(\sum_{j\neq i}\sum_{k\neq i,j} G_{ij}F_{ij}\check{M}_{ik,-ij}R_{2k}\right)^2\\ &+\sum_{i}\sum_{j\neq i}\operatorname{Var}\left(v_{1i}V_{3j}V_{4j}\right)\left(\sum_{k\neq i,j} G_{ij}F_{ij}\check{M}_{ik,-ij}R_{2k}\right)\left(\sum_{k\neq i,j} G_{ji}F_{ji}\check{M}_{jk,-ij}R_{2k}\right).\end{aligned}$$

since the expansion in the intermediate steps are entirely analogous.

Turning to the sum with two stochastic objects,

$$\begin{aligned}\operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}R_{1i}R_{2k}V_{3j}V_{4j}\right) &= \operatorname{Var}\left(\sum_{i} V_{3i}V_{4i}\left(\sum_{j\neq i}\sum_{k\neq i,j} G_{ji}F_{ji}\check{M}_{jk,-ij}R_{1j}R_{2k}\right)\right)\\ &= \sum_{i}\operatorname{Var}\left(V_{3i}V_{4i}\right)\left(\sum_{j\neq i}\sum_{k\neq i,j} G_{ji}F_{ji}\check{M}_{jk,-ij}R_{1j}R_{2k}\right)^2\\ &\leq \max_{i}\operatorname{Var}\left(V_{3i}V_{4i}\right)\sum_{i}\left(\sum_{j\neq i}\sum_{k\neq i,j} G_{ji}F_{ji}\check{M}_{jk,-ij}R_{1j}R_{2k}\right)^2.\end{aligned}$$

With these inequalities, applying Assumption 1.3 suffices for the result.

Proof of Lemma 1.4(c). Expand the term:

$$\sum_{i\neq j\neq l}^{n} G_{ij}F_{ij}\check{M}_{jl,-ij}V_{1i}V_{2i}V_{3j}V_{4l} = \sum_{i\neq j\neq l}^{n} G_{ij}F_{ij}\check{M}_{jl,-ij}V_{1i}V_{2i}\left(R_{3j}R_{4l}+R_{3j}v_{4l}+v_{3j}R_{4l}+v_{3j}v_{4l}\right).$$

With four stochastic objects,

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq l}^{n} G_{ij}F_{ij}\check{M}_{jl,-ij}V_{1i}V_{2i}v_{3j}v_{4l}\right)\\
&\leq \max_{i,j,k}\operatorname{Var}\left(V_{1i}V_{2i}v_{3j}v_{4l}\right)\sum_{i\neq j\neq l}^{n}\sum_{i_2\neq i,j,l}\left(G_{ij}F_{ij}\check{M}_{jl,-ij}G_{i_2j}F_{i_2j}\check{M}_{jl,-i_2j}+G_{ij}F_{ij}\check{M}_{jl,-ij}G_{i_2l}F_{i_2l}\check{M}_{lj,-i_2l}\right)\\
&\quad+\max_{i,j,k}\operatorname{Var}\left(V_{1i}V_{2i}v_{3j}v_{4l}\right)3!\sum_{i\neq j\neq l}^{n}G_{ij}^2F_{ij}^2\check{M}_{jl,-ij}^2.
\end{aligned}$$

Simplifying the first line,

$$\begin{aligned}
&\sum_{i\neq j\neq l}^{n}\sum_{i_2\neq i,j,l}\left(G_{ij}F_{ij}\check{M}_{jl,-ij}G_{i_2j}F_{i_2j}\check{M}_{jl,-i_2j}+G_{ij}F_{ij}\check{M}_{jl,-ij}G_{i_2l}F_{i_2l}\check{M}_{lj,-i_2l}\right)\\
&\leq\left(\sum_{i\neq j\neq l\neq i_2}^{n}G_{ij}^2G_{i_2j}^2\check{M}_{jl,-ij}^2\right)^{1/2}\left(\sum_{i\neq j\neq l\neq i_2}^{n}F_{ij}^2F_{i_2j}^2\check{M}_{jl,-i_2j}^2\right)^{1/2}\\
&+\left(\sum_{i\neq j\neq l\neq i_2}^{n}G_{ij}^2G_{i_2l}^2\check{M}_{jl,-ij}^2\right)^{1/2}\left(\sum_{i\neq j\neq l\neq i_2}^{n}F_{ij}^2F_{i_2l}^2\check{M}_{lj,-i_2j}^2\right)^{1/2}.
\end{aligned}$$

These terms have the required order due to a proof analogous to Lemma 1.4(b). Next,

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq l}^{n}G_{ij}F_{ij}\check{M}_{jl,-ij}V_{1i}V_{2i}R_{3j}v_{4l}\right)=\operatorname{Var}\left(\sum_{i}\sum_{j\neq i}V_{1i}V_{2i}v_{4j}\left(\sum_{l\neq i,j}G_{il}F_{il}\check{M}_{lj,-il}R_{3l}\right)\right)\\
&\leq\sum_{i}\sum_{j\neq i}\operatorname{Var}\left(V_{1i}V_{2i}v_{4j}\right)\left(\sum_{l\neq i,j}G_{il}F_{il}\check{M}_{lj,-il}R_{3l}\right)\left[\sum_{l\neq i,j}G_{il}F_{il}\check{M}_{lj,-il}R_{3l}+\sum_{l\neq i,j}G_{jl}F_{jl}\check{M}_{li,-jl}R_{3l}\right]\\
&+\max_{i,j}\operatorname{Var}\left(V_{1i}V_{2i}v_{4j}\right)\sum_{i}\sum_{j\neq i}\sum_{i_2\neq i,j}\left(\sum_{l\neq i,j}G_{il}F_{il}\check{M}_{lj,-il}R_{3l}\right)\left(\sum_{k\neq i_2,l}G_{kl}F_{kl}\check{M}_{ki_2,-kl}R_{1k}\right)\\
&\leq\max_{i,j}\operatorname{Var}\left(V_{1i}V_{2i}v_{4j}\right)\sum_{i}\left(\sum_{j\neq i}\sum_{l\neq i,j}G_{il}F_{il}\check{M}_{lj,-il}R_{3l}\right)^2\\
&+\sum_{i}\sum_{j\neq i}\operatorname{Var}\left(V_{1i}V_{2i}v_{4j}\right)\left(\sum_{l\neq i,j}G_{il}F_{il}\check{M}_{lj,-il}R_{3l}\right)\left(\sum_{l\neq i,j}G_{jl}F_{jl}\check{M}_{li,-jl}R_{3l}\right).
\end{aligned}$$

Further, $\text{Var}\left(\sum_{i\neq j\neq l}^{n} G_{ij}F_{ij}\check{M}_{jl,-ij}V_{1i}V_{2i}v_{3j}R_{4l}\right)$ can be bounded by a similar argument. Turning to the sum with two stochastic objects,

$$\text{Var}\left(\sum_{i\neq j\neq l}^{n} G_{ij}F_{ij}\check{M}_{jl,-ij}V_{1i}V_{2i}R_{3j}R_{4l}\right) = \sum_{i}\text{Var}\left(V_{1i}V_{2i}\right)\left(\sum_{j\neq i}\sum_{l\neq i,j} G_{ij}F_{ij}\check{M}_{jl,-ij}R_{3j}R_{4l}\right)^2.$$

These inequalities suffice for the result due to Assumption 1.3.

Proof of Lemma 1.4(d). Expand the term:

$$\begin{aligned}
&\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}V_{2k}V_{3j}V_{4l} \\
&= \sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}R_{2k}\left(R_{3j}R_{4l} + R_{3j}v_{4l} + v_{3j}R_{4l} + v_{3j}v_{4l}\right) \\
&\quad + \sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}v_{2k}\left(R_{3j}R_{4l} + R_{3j}v_{4l} + v_{3j}R_{4l} + v_{3j}v_{4l}\right).
\end{aligned}$$

Consider the v_{2k} line first. We only have the 4-sum to contend with. For 5-sum and above, at least one of the errors can be factored out as a zero expectation. Hence, by using Cauchy-Schwarz and the same argument as above,

$$\begin{aligned}
&\text{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}v_{2k}v_{3j}v_{4l}\right) \\
&\leq \max_{i,j,k,l}\text{Var}\left(V_{1i}v_{2k}v_{3j}v_{4l}\right)4!\sum_{i\neq j\neq k\neq l}^{n} G_{ij}^2F_{ij}^2\check{M}_{ik,-ij}^2\check{M}_{jl,-ijk}^2 \\
&\leq C\sum_{i\neq j\neq k}^{n} G_{ij}^2F_{ij}^2\check{M}_{ik,-ij}^2 \leq C\sum_{i\neq j}^{n} G_{ij}^2F_{ij}^2 \leq C\left(\sum_{i\neq j}^{n} G_{ij}^2\right)^{1/2}\left(\sum_{i\neq j}^{n} F_{ij}^2\right)^{1/2}.
\end{aligned}$$

By using the same expansion step as before,

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}V_{1i}v_{2k}v_{3j}\left(\sum_{l\neq i,j,k}\check{M}_{jl,-ijk}R_{4l}\right)\right)\\
&\leq \max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{3j}\left(\sum_{l\neq i,j,k}\check{M}_{jl,-ijk}R_{4l}\right)\right)\\
&\qquad\sum_{i\neq j\neq k\neq i_2}^{n}\left(G_{ij}F_{ij}G_{i_2j}F_{i_2j}\check{M}_{ik,-ij}\check{M}_{i_2k,-ij}+G_{ij}F_{ij}G_{i_2k}F_{i_2k}\check{M}_{ij,-ik}\check{M}_{i_2j,-ik}\right)\\
&\qquad+\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{3j}\left(\sum_{l\neq i,j,k}\check{M}_{jl,-ijk}R_{4l}\right)\right)3!\sum_{i}\sum_{j\neq i}\sum_{k\neq i,j}G_{ij}^2F_{ij}^2\check{M}_{ik,-ij}^2.
\end{aligned}$$

The $\sum_{i\neq j\neq k\neq i_2}^{n}\left(G_{ij}F_{ij}G_{i_2j}F_{i_2j}\check{M}_{ik,-ij}\check{M}_{i_2k,-ij}+G_{ij}F_{ij}G_{i_2k}F_{i_2k}\check{M}_{ij,-ik}\check{M}_{i_2j,-ik}\right)$ term has the required order due to the same argument as the proof of Lemma 1.4(b). Next,

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i\neq j\neq k\neq l}^{n}G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}v_{2k}R_{3j}v_{4l}\right)=\mathrm{Var}\left(\sum_{i\neq j\neq k}^{n}V_{1i}v_{2k}v_{4j}\left(\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)\right)\\
&\leq\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{4j}\right)\sum_{i\neq j\neq k}^{n}\sum_{i_2\neq i,j,k}\left(\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)\left(\sum_{l\neq i_2,j,k}G_{i_2l}F_{i_2l}\check{M}_{i_2k,-i_2l}\check{M}_{lj,-i_2lk}R_{3l}\right)\\
&\quad+\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{4j}\right)\sum_{i\neq j\neq k}^{n}\sum_{i_2\neq i,j,k}\left(\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)\left(\sum_{l\neq i_2,j,k}G_{i_2l}F_{i_2l}\check{M}_{i_2j,-i_2l}\check{M}_{lk,-i_2lj}R_{3l}\right)\\
&\quad+\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{4j}\right)3!\sum_{i\neq j\neq k}^{n}\check{M}_{ik,-ij}^2\left(\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)^2\\
&\leq\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{4j}\right)\sum_{k}\sum_{j\neq k}\left(\sum_{i\neq k,j}\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)^2\\
&\quad-\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{4j}\right)\sum_{k}\sum_{j\neq k}\sum_{i\neq k,j}\left(\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)^2\\
&\quad+\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{4j}\right)\sum_{k}\sum_{j\neq k}\left(\sum_{i\neq k,j}\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)\left(\sum_{i\neq k,j}\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ij,-il}\check{M}_{lk,-ilj}R_{3l}\right)\\
&\quad-\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{4j}\right)\sum_{k}\sum_{j\neq k}\sum_{i\neq k,j}\left(\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)\left(\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ij,-il}\check{M}_{lk,-ilj}R_{3l}\right)\\
&\quad+\max_{i,j,k}\mathrm{Var}\left(V_{1i}v_{2k}v_{4j}\right)3!\sum_{i\neq j\neq k}^{n}\check{M}_{ik,-ij}^2\left(\sum_{l\neq i,j,k}G_{il}F_{il}\check{M}_{ik,-il}\check{M}_{lj,-ilk}R_{3l}\right)^2.
\end{aligned}$$

The first term in the v_{2k} line is then:

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}v_{2k}R_{3j}R_{4l}\right) = \operatorname{Var}\left(\sum_{i\neq j}^{n} G_{ij}F_{ij}\check{M}_{ij,-ik}V_{1i}v_{2j}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)\\
&\leq \max_{i,j}\operatorname{Var}\left(V_{1i}v_{2j}\right)\sum_{i\neq j}^{n}\left(G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ij,-ik}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)^2\\
&+\max_{i,j}\operatorname{Var}\left(V_{1i}v_{2j}\right)\sum_{i\neq j}^{n}\left(G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ij,-ik}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)\\
&\quad\left(G_{ji}F_{ji}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ji,-jk}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)\\
&+\max_{i,j}\operatorname{Var}\left(V_{1i}v_{2j}\right)\sum_{i\neq j\neq i_2}^{n}\left(G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ij,-ik}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)\\
&\quad\left(G_{i_2j}F_{i_2j}\sum_{k\neq i_2,j}\sum_{l\neq i_2,j,k}\check{M}_{i_2j,-i_2k}\check{M}_{kl,-i_2jk}R_{3k}R_{4l}\right)\\
&\leq \max_{i,j}\operatorname{Var}\left(V_{1i}v_{2j}\right)\sum_{i\neq j}^{n}\left(G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ij,-ik}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)^2\\
&+\max_{i,j}\operatorname{Var}\left(V_{1i}v_{2j}\right)\sum_{i\neq j}^{n}\left(G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ij,-ik}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)\\
&\quad\left(G_{ji}F_{ji}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ji,-jk}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)\\
&+\max_{i,j}\operatorname{Var}\left(V_{1i}v_{2j}\right)\sum_{j}\left(\sum_{i\neq j}\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ij}F_{ij}\check{M}_{ij,-ik}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)^2\\
&\quad-\max_{i,j}\operatorname{Var}\left(V_{1i}v_{2j}\right)\sum_{j}\sum_{i\neq j}\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ij}F_{ij}\check{M}_{ij,-ik}\check{M}_{kl,-ijk}R_{3k}R_{4l}\right)^2.
\end{aligned}$$

Now, we turn back to the R_{2k} expression to complete the proof:

$$\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}R_{2k}\left(R_{3j}R_{4l}+R_{3j}v_{4l}+v_{3j}R_{4l}+v_{3j}v_{4l}\right).$$

Consider the term with three stochastic terms first, and simplify it using the same strategy as before:

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}R_{2k}v_{3j}v_{4l}\right) = \operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ij}V_{1i}v_{3j}v_{4k}\sum_{l\neq i,j,k}\check{M}_{il,-ij}\check{M}_{jk,-ijl}R_{2l}\right)\\
&\leq \max_{i,j,k}\operatorname{Var}\left(V_{1i}v_{3j}v_{4k}\right)\left(\sum_{k\neq j}^{n}\left(\sum_{i\neq k,j}\sum_{l\neq i,j,k}G_{ij}F_{ij}\check{M}_{il,-ij}\check{M}_{jk,-ijl}R_{2l}\right)^2\right.\\
&\quad\left.-\sum_{k\neq j}^{n}\sum_{i\neq k,j}\left(\sum_{l\neq i,j,k}G_{ij}F_{ij}\check{M}_{il,-ij}\check{M}_{jk,-ijl}R_{2l}\right)^2\right)\\
&+\max_{i,j,k}\operatorname{Var}\left(V_{1i}v_{3j}v_{4k}\right)\sum_{k}\sum_{j\neq k}\left(\sum_{i\neq k,j}\sum_{l\neq i,j,k}G_{ij}F_{ij}\check{M}_{il,-ij}\check{M}_{jk,-ijl}R_{2l}\right)\left(\sum_{i\neq k,j}\sum_{l\neq i,j,k}G_{ik}F_{ik}\check{M}_{il,-ik}\check{M}_{kj,-ikl}R_{2l}\right)\\
&-\max_{i,j,k}\operatorname{Var}\left(V_{1i}v_{3j}v_{4k}\right)\sum_{k}\sum_{j\neq k}\sum_{i\neq k,j}\left(\sum_{l\neq i,j,k}G_{ij}F_{ij}\check{M}_{il,-ij}\check{M}_{jk,-ijl}R_{2l}\right)\left(\sum_{l\neq i,j,k}G_{ik}F_{ik}\check{M}_{il,-ik}\check{M}_{kj,-ikl}R_{2l}\right)\\
&+\max_{i,j,k}\operatorname{Var}\left(V_{1i}v_{3j}v_{4k}\right)3!\sum_{i\neq j\neq k}^{n}\left(G_{ij}F_{ij}\sum_{l\neq i,j,k}\check{M}_{il,-ij}\check{M}_{jk,-ijl}R_{2l}\right)^2.
\end{aligned}$$

Next,

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i}\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}R_{2k}v_{3j}R_{4l}\right)\\
&\leq \max_{i,j}\operatorname{Var}\left(V_{1i}v_{3j}\right)\sum_{i\neq j}^{n}\left(G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}R_{2k}R_{4l}\right)^2\\
&+\max_{i,j}\operatorname{Var}\left(V_{1i}v_{3j}\right)\sum_{i\neq j}^{n}\left(G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}R_{2k}R_{4l}\right)\left(G_{ji}F_{ji}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{jk,-ij}\check{M}_{il,-ijk}R_{2k}R_{4l}\right)\\
&+\max_{i,j}\operatorname{Var}\left(V_{1i}v_{3j}\right)\sum_{j}\left(\sum_{i\neq j}G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}R_{2k}R_{4l}\right)^2\\
&-\max_{i,j}\operatorname{Var}\left(V_{1i}v_{3j}\right)\sum_{j\neq i}^{n}\left(G_{ij}F_{ij}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}R_{2k}R_{4l}\right)^2.
\end{aligned}$$

Finally,

$$\begin{aligned}
&\text{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}V_{1i}R_{2k}R_{3j}R_{4l}\right)\\
&\quad=\sum_{i}\text{Var}\left(V_{1i}\right)\left(\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} G_{ij}F_{ij}\check{M}_{ik,-ij}\check{M}_{jl,-ijk}R_{2k}R_{3j}R_{4l}\right)^{2}.
\end{aligned}$$

□

Proof of Lemma 1.5. **Proof of Lemma 1.5(a).** Expand the term:

$$\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{1j}V_{2k}V_{3i}V_{4i}=\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{3i}V_{4i}\left(R_{1j}R_{2k}+R_{1j}v_{2k}+v_{1j}R_{2k}+v_{1j}v_{2k}\right).$$

With four stochastic objects,

$$\begin{aligned}
Var\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{3i}V_{4i}v_{1j}v_{2k}\right)&\leq\max_{i,j,k}\text{Var}\left(V_{3i}V_{4i}v_{1j}v_{2k}\right)\sum_{i\neq j\neq k}^{n}\sum_{i_2\neq i,j,k}\left(G_{ij}F_{ik}G_{i_2j}F_{i_2k}+G_{ij}F_{ik}G_{i_2k}F_{i_2j}\right)\\
&\quad+\max_{i,j,k}\text{Var}\left(V_{1i}V_{2i}v_{3j}v_{4l}\right)3!\sum_{i\neq j\neq k}^{n}G_{ij}^{2}F_{ik}^{2}.
\end{aligned}$$

Observe that, due to Assumption 1.3(a),

$$\begin{aligned}
\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}G_{lj}F_{lk}&=\sum_{j\neq k}^{n}\left(\sum_{i\neq j,k}G_{ij}F_{ik}\right)\left(\sum_{l\neq j,k}G_{lj}F_{lk}-G_{ij}F_{ik}\right)\\
&=\sum_{j\neq k}^{n}\left(\sum_{i\neq j,k}G_{ij}F_{ik}\right)^{2}-\sum_{j\neq k\neq i}^{n}G_{ij}^{2}F_{ik}^{2}
\end{aligned}$$

has the required order, which suffices for the bound. Next,

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{3i}V_{4i}R_{1j}v_{2k}\right)\\
&\quad=\operatorname{Var}\left(\sum_{i}\sum_{j\neq i}F_{ij}V_{3i}V_{4i}v_{2j}\left(\sum_{k\neq i,j}G_{ik}R_{1k}\right)\right)\\
&\quad\leq\sum_{i}\sum_{j\neq i}\operatorname{Var}\left(V_{3i}V_{4i}v_{2j}\right)\left(\sum_{k\neq i,j}F_{ij}G_{ik}R_{1k}\right)\left[\sum_{k\neq i,j}F_{ij}G_{ik}R_{1k}+\sum_{k\neq i,j}F_{ji}G_{jk}R_{1k}\right]\\
&\quad+\max_{i,j}\operatorname{Var}\left(V_{3i}V_{4i}v_{2j}\right)\sum_{i}\sum_{j\neq i}\sum_{i_2\neq i,j}\left(\sum_{k\neq i,j}F_{ij}G_{ik}R_{1k}\right)\left(\sum_{k\neq i_2,l}F_{i_2j}G_{i_2k}R_{1k}\right)\\
&\quad\leq\max_{i,j}\operatorname{Var}\left(V_{3i}V_{4i}v_{2j}\right)\sum_{i}\left(\sum_{j\neq i}\sum_{k\neq i,j}F_{ij}G_{ik}R_{1k}\right)^{2}\\
&\qquad+\sum_{i}\sum_{j\neq i}\operatorname{Var}\left(V_{3i}V_{4i}v_{2j}\right)\left(\sum_{k\neq i,j}F_{ij}G_{ik}R_{1k}\right)\left(\sum_{k\neq i,j}F_{ji}G_{jk}R_{1k}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{3i}V_{4i}v_{1j}R_{2k}\right)=\operatorname{Var}\left(\sum_{i}\sum_{j\neq i}V_{3i}V_{4i}v_{1j}\left(\sum_{k\neq i,j}G_{ij}F_{ik}R_{2k}\right)\right)\\
&\quad\leq\max_{i,j}\operatorname{Var}\left(V_{3i}V_{4i}v_{1j}\right)\sum_{i}\left(\left(\sum_{j\neq i}\sum_{k\neq i,j}G_{ij}F_{ik}R_{2k}\right)^{2}+\sum_{j\neq i}\left(\sum_{k\neq i,j}G_{ij}F_{ik}R_{2k}\right)\left(\sum_{k\neq i,j}G_{ji}F_{jk}R_{2k}\right)\right).
\end{aligned}$$

Turning to the sum with two stochastic objects,

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{3i}V_{4i}R_{1j}R_{2k}\right)=\operatorname{Var}\left(\sum_{i}V_{3i}V_{4i}\left(\sum_{j\neq i}\sum_{k\neq i,j}G_{ij}F_{ik}R_{1j}R_{2k}\right)\right)\\
&\quad\leq\max_{i}\operatorname{Var}\left(V_{3i}V_{4i}\right)\sum_{i}\left(\sum_{j\neq i}\sum_{k\neq i,j}G_{ij}F_{ik}R_{1j}R_{2k}\right)^{2}.
\end{aligned}$$

Proof of Lemma 1.5(b).

Decompose the term:

$$\begin{aligned}
&\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{1j}V_{2k}V_{3i}V_{4l} \\
&= \sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}R_{1j}\left(R_{2k}R_{4l}+R_{2k}v_{4l}+v_{2k}R_{4l}+v_{2k}v_{4l}\right) \\
&\quad + \sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}v_{1j}\left(R_{2k}R_{4l}+R_{2k}v_{4l}+v_{2k}R_{4l}+v_{2k}v_{4l}\right).
\end{aligned}$$

Consider the v_{1j} line first.

$$\operatorname{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}v_{1j}v_{2k}v_{4l}\right) \leq \max_{i,j,k,l}\operatorname{Var}\left(V_{3i}v_{1j}v_{2k}v_{4l}\right)4!\sum_{i\neq j\neq k\neq l}^{n} G_{ij}^2F_{ik}^2\check{M}_{il,-ijk}^2.$$

Next, by using the same expansion and simplification steps as before,

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}v_{1j}v_{2k}R_{4l}\right) = \operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{3i}v_{1j}v_{2k}\sum_{l\neq i,j,k}\check{M}_{il,-ijk}R_{4l}\right) \\
&\leq \max_{i,j,k}\operatorname{Var}\left(V_{3i}v_{1j}v_{2k}\right)\sum_{k}\sum_{j\neq k}\left(\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k} G_{ij}F_{ik}\check{M}_{il,-ijk}R_{4l}\right)^2 - \sum_{i\neq j,k}\left(\sum_{l\neq i,j,k} G_{ij}F_{ik}\check{M}_{il,-ijk}R_{4l}\right)^2\right) \\
&+ \max_{i,j,k}\operatorname{Var}\left(V_{3i}v_{1j}v_{2k}\right)\sum_{k}\sum_{j\neq k}\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k} G_{ij}F_{ik}\check{M}_{il,-ijk}R_{4l}\right)\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k} G_{ik}F_{ij}\check{M}_{il,-ijk}R_{4l}\right) \\
&- \max_{i,j,k}\operatorname{Var}\left(V_{3i}v_{1j}v_{2k}\right)\sum_{k}\sum_{j\neq k}\sum_{i\neq j,k}\left(\sum_{l\neq i,j,k} G_{ij}F_{ik}\check{M}_{il,-ijk}R_{4l}\right)\left(\sum_{l\neq i,j,k} G_{ik}F_{ij}\check{M}_{il,-ijk}R_{4l}\right) \\
&+ \max_{i,j,k}\operatorname{Var}\left(V_{3i}v_{1j}v_{2k}\right)3!\sum_{i\neq j\neq k}^{n} G_{ij}^2F_{ik}^2\left(\sum_{l\neq i,j,k}\check{M}_{il,-ijk}R_{4l}\right)^2
\end{aligned}$$

and

$$\operatorname{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}v_{1j}R_{2k}v_{4l}\right) = \operatorname{Var}\left(\sum_{i\neq j\neq k}^{n} G_{ij}F_{ik}V_{3i}v_{1j}v_{4k}\sum_{l\neq i,j,k}\check{M}_{ik,-ijl}R_{2l}\right)$$

$$\leq \max_{i,j,k}\operatorname{Var}\left(V_{3i}v_{1j}v_{4k}\right)\sum_{k}\sum_{j\neq k}\left(\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k}G_{ij}F_{ik}\check{M}_{ik,-ijl}R_{2l}\right)^{2} - \sum_{i\neq j,k}\left(\sum_{l\neq i,j,k}G_{ij}F_{ik}\check{M}_{ik,-ijl}R_{2l}\right)^{2}\right)$$

$$+ \max_{i,j,k}\operatorname{Var}\left(V_{3i}v_{1j}v_{2k}\right)\sum_{k}\sum_{j\neq k}\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k}G_{ij}F_{ik}\check{M}_{ik,-ijl}R_{2l}\right)\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k}G_{ik}F_{ij}\check{M}_{il,-ijk}R_{2l}\right)$$

$$- \max_{i,j,k}\operatorname{Var}\left(V_{3i}v_{1j}v_{2k}\right)\sum_{k}\sum_{j\neq k}\sum_{i\neq j,k}\left(\sum_{l\neq i,j,k}G_{ij}F_{ik}\check{M}_{ik,-ijl}R_{2l}\right)\left(\sum_{l\neq i,j,k}G_{ik}F_{ij}\check{M}_{il,-ijk}R_{2l}\right)$$

$$+ \max_{i,j,k}\operatorname{Var}\left(V_{3i}v_{1j}v_{4k}\right)3!\sum_{i\neq j\neq k}^{n}G_{ij}^{2}F_{ik}^{2}\left(\sum_{l\neq i,j,k}\check{M}_{ik,-ijl}R_{2l}\right)^{2}$$

with $\left(\sum_{l\neq i,j,k}\check{M}_{ik,-ijl}R_{2l}\right)^{2} \leq C$. Finally,

$$\operatorname{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}v_{1j}R_{2k}R_{4l}\right) = \operatorname{Var}\left(\sum_{i\neq j}^{n} G_{ij}V_{3i}v_{1j}\sum_{k\neq i,j}\sum_{l\neq i,j,k}F_{ik}\check{M}_{il,-ijk}R_{2k}R_{4l}\right)$$

$$\leq \max_{i,j}\operatorname{Var}\left(V_{3i}v_{1j}\right)\sum_{i\neq j}^{n}G_{ij}\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}F_{ik}\check{M}_{il,-ijk}R_{2k}R_{4l}\right)^{2}$$

$$+ \max_{i,j}\operatorname{Var}\left(V_{3i}v_{1j}\right)\sum_{i\neq j}^{n}\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ij}F_{ik}\check{M}_{il,-ijk}R_{2k}R_{4l}\right)\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ji}F_{jk}\check{M}_{jl,-ijk}R_{2k}R_{4l}\right)$$

$$+ \max_{i,j}\operatorname{Var}\left(V_{3i}v_{1j}\right)\sum_{j}\left(\left(\sum_{i\neq j}\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ij}F_{ik}\check{M}_{il,-ijk}R_{2k}R_{4l}\right)^{2} - \sum_{i\neq j}\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ij}F_{ik}\check{M}_{il,-ijk}R_{2k}R_{4l}\right)^{2}\right).$$

Now, return to the R_{1j} line: $\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}R_{1j}\left(R_{2k}R_{4l}+R_{2k}v_{4l}+v_{2k}R_{4l}+v_{2k}v_{4l}\right)$, so

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}R_{1j}v_{2k}v_{4l}\right) = \mathrm{Var}\left(\sum_{i\neq j\neq k}^{n} G_{il}F_{ik}V_{3i}v_{2k}v_{4j}\sum_{l\neq i,j,k}\check{M}_{ij,-ilk}R_{1l}\right)\\
&\leq \max_{i,j,k}\mathrm{Var}\left(V_{3i}v_{2k}v_{4j}\right)\left(\sum_{j}\sum_{k\neq j}\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k}G_{il}F_{ik}\check{M}_{ij,-ilk}R_{1l}\right)^2-\sum_{j}\sum_{k\neq j}\sum_{i\neq j,k}\left(\sum_{l\neq i,j,k}G_{il}F_{ik}\check{M}_{ij,-ilk}R_{1l}\right)^2\right)\\
&+\max_{i,j,k}\mathrm{Var}\left(V_{3i}v_{2k}v_{4j}\right)\sum_{j}\sum_{k\neq j}\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k}G_{il}F_{ik}\check{M}_{ij,-ilk}R_{1l}\right)\left(\sum_{i\neq j,k}\sum_{l\neq i,j,k}G_{il}F_{ij}\check{M}_{ik,-ilj}R_{1l}\right)\\
&-\max_{i,j,k}\mathrm{Var}\left(V_{3i}v_{2k}v_{4j}\right)\sum_{j}\sum_{k\neq j}\sum_{i\neq j,k}\left(\sum_{l\neq i,j,k}G_{il}F_{ik}\check{M}_{ij,-ilk}R_{1l}\right)\left(\sum_{l\neq i,j,k}G_{il}F_{ij}\check{M}_{ik,-ilj}R_{1l}\right)\\
&+\max_{i,j,k}\mathrm{Var}\left(V_{3i}v_{2k}v_{4j}\right)3!\sum_{i\neq j\neq k}^{n}\left(F_{ik}\sum_{l\neq i,j,k}G_{il}\check{M}_{ij,-ilk}R_{1l}\right)^2,
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}R_{1j}v_{2k}R_{4l}\right) = \mathrm{Var}\left(\sum_{i\neq j}^{n} F_{ij}V_{3i}v_{2j}\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ik}\check{M}_{il,-ijk}R_{1k}R_{4l}\right)\\
&\leq \max_{i,j}\mathrm{Var}\left(V_{3i}v_{2j}\right)\sum_{i\neq j}^{n}\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}F_{ij}G_{ik}\check{M}_{il,-ijk}R_{1k}R_{4l}\right)^2\\
&+\max_{i,j}\mathrm{Var}\left(V_{3i}v_{2j}\right)\sum_{i\neq j}^{n}\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}F_{ij}G_{ik}\check{M}_{il,-ijk}R_{1k}R_{4l}\right)\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}F_{ij}G_{jk}\check{M}_{jl,-ijk}R_{1k}R_{4l}\right)\\
&+\max_{i,j}\mathrm{Var}\left(V_{3i}v_{2j}\right)\sum_{j}\left(\left(\sum_{i\neq j}\sum_{k\neq i,j}\sum_{l\neq i,j,k}F_{ij}G_{ik}\check{M}_{il,-ijk}R_{1k}R_{4l}\right)^2-\sum_{i\neq j}\left(\sum_{k\neq i,j}\sum_{l\neq i,j,k}F_{ij}G_{ik}\check{M}_{il,-ijk}R_{1k}R_{4l}\right)^2\right).
\end{aligned}$$

The $\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}R_{1j}R_{2k}v_{4l}$ term is symmetric, because it does not matter which R_m we use. Finally,

$$\mathrm{Var}\left(\sum_{i\neq j\neq k\neq l}^{n} G_{ij}F_{ik}\check{M}_{il,-ijk}V_{3i}R_{1j}R_{2k}R_{4l}\right) = \sum_{i}\mathrm{Var}\left(V_{3i}\right)\left(\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k}G_{ij}F_{ik}\check{M}_{il,-ijk}R_{1j}R_{2k}R_{4l}\right)^2.$$

□

Proof of Lemma 1.6. The proof of Lemma 1.6 is entirely analogous to Lemmas 1.4 and 1.5 just that G_{ji} is used in place of G_{ij}. □

1.C.6 Proofs for Appendix 1.C

Proofs for Propositions in Appendix 1.C

Proof of Proposition 1.2. Let

$$\begin{pmatrix} \Pi_Y \\ \Pi \end{pmatrix} := \begin{pmatrix} (Z'Z)^{1/2}\,\pi_Y \\ (Z'Z)^{1/2}\,\pi \end{pmatrix}.$$

With this definition, $(\pi_Y' Z'Z\pi_Y, \pi'Z'Z\pi_Y, \pi'Z'Z\pi) = (\Pi_Y'\Pi_Y, \Pi_Y'\Pi, \Pi'\Pi)$, and

$$\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \Pi_Y \\ \Pi \end{pmatrix}, \Omega \otimes I_K\right).$$

Split s_1 and s_2 into the Π component and a random normal component: $s_{1k} = \Pi_{Yk} + z_{1k}$ and $s_{2k} = \Pi_k + z_{2k}$. Then, for all k,

$$\begin{pmatrix} z_{1k} \\ z_{2k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \omega_{\zeta\zeta} & \omega_{\zeta\eta} \\ \omega_{\zeta\eta} & \omega_{\eta\eta} \end{bmatrix}\right), \text{ and}$$

$$\begin{pmatrix} s_1's_1 \\ s_1's_2 \\ s_2's_2 \end{pmatrix} = \begin{pmatrix} \sum_k s_{1k}^2 \\ \sum_k s_{1k}s_{2k} \\ \sum_k s_{2k}^2 \end{pmatrix} = \begin{pmatrix} \sum_k (\Pi_{Yk} + z_{1k})^2 \\ \sum_k (\Pi_{Yk} + z_{1k})(\Pi_k + z_{2k}) \\ \sum_k (\Pi_k + z_{2k})^2 \end{pmatrix}$$
$$= \begin{pmatrix} \sum_k \Pi_{Yk}^2 + 2\sum_k \Pi_{Yk}z_{1k} + \sum_k z_{1k}^2 \\ \sum_k \Pi_{Yk}\Pi_k + \sum_k \Pi_{Yk}z_{2k} + \sum_k \Pi_k z_{1k} + \sum_k z_{1k}z_{2k} \\ \sum_k \Pi_k^2 + 2\sum_k \Pi_k z_{2k} + \sum_k z_{2k}^2 \end{pmatrix}.$$

Under the assumption, $\Pi'\Pi/\sqrt{K} \to C_S$, so $\frac{1}{\sqrt{K}}\sum_k \Pi_k^2 \to C_S$. By applying the Lindeberg CLT due to bounded moments,

$$\frac{1}{\sqrt{K}} \begin{pmatrix} \sum_k \Pi_k z_{1k} \\ \sum_k \Pi_{Yk} z_{1k} \\ \sum_k \Pi_{Yk} z_{2k} \\ \sum_k \Pi_k z_{2k} \\ \sum_k z_{1k}z_{2k} \\ \sum_k z_{2k}^2 \\ \sum_k z_{1k}^2 \end{pmatrix} \overset{a}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \sqrt{K}\omega_{\zeta\eta} \\ \sqrt{K}\omega_{\eta\eta} \\ \sqrt{K}\omega_{\zeta\zeta} \end{pmatrix}, V \right),$$

where V is some variance matrix. By assumption, $\frac{1}{\sqrt{K}}\sum_k \Pi_{Yk}\Pi_k \to C_Y$ and $\frac{1}{\sqrt{K}}\sum_k \Pi_{Yk}^2 \to C_{YY}$, so

$$\frac{1}{\sqrt{K}}\begin{pmatrix} s_1's_1 \\ s_1's_2 \\ s_2's_2 \end{pmatrix} = \frac{1}{\sqrt{K}}\begin{pmatrix} \sum_k \Pi_{Yk}^2 + 2\sum_k \Pi_{Yk}z_{1k} + \sum_k z_{1k}^2 \\ \sum_k \Pi_{Yk}\Pi_k + \sum_k \Pi_{Yk}z_{2k} + \sum_k \Pi_k z_{1k} + \sum_k z_{1k}z_{2k} \\ \sum_k \Pi_k^2 + 2\sum_k \Pi_k z_{2k} + \sum_k z_{2k}^2 \end{pmatrix}$$

$$\overset{a}{\sim} \begin{pmatrix} C_{YY} \\ C_Y \\ C \end{pmatrix} + A\frac{1}{\sqrt{K}}\begin{pmatrix} \sum_k \Pi_k z_{1k} \\ \sum_k \Pi_{Yk} z_{1k} \\ \sum_k \Pi_{Yk} z_{2k} \\ \sum_k \Pi_k z_{2k} \\ \sum_k z_{1k}z_{2k} \\ \sum_k z_{2k}^2 \\ \sum_k z_{1k}^2 \end{pmatrix}, \text{ where}$$

$$A = \begin{pmatrix} 0 & 2 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 & 0 \end{pmatrix}.$$

This means:

$$\frac{1}{\sqrt{K}}\begin{pmatrix} s_1's_1 \\ s_1's_2 \\ s_2's_2 \end{pmatrix} \overset{a}{\sim} N\left(\begin{pmatrix} C_{YY} + \sqrt{K}\omega_{\zeta\zeta} \\ C_Y + \sqrt{K}\omega_{\zeta\eta} \\ C + \sqrt{K}\omega_{\eta\eta} \end{pmatrix}, AVA'\right).$$

Let $\Sigma = AVA'$ to obtain the result as stated. To derive Σ explicitly, I derive V by applying the Isserlis' Theroem. As a special case of the Isserlis' Theorem for X's that are multivariate normal and mean zero,

$$E\left[X_1X_2X_3X_4\right] = E\left[X_1X_2\right]E\left[X_3X_4\right] + E\left[X_1X_3\right]E\left[X_2X_4\right] + E\left[X_1X_4\right]E\left[X_2X_3\right].$$

Another corrolary is that if n is odd, then there is no such pairing, so the moment is always zero. Hence,

$$E\left[z_{1k}^2 z_{2k}^2\right] = E\left[z_{1k}^2\right] E\left[z_{2k}^2\right] + 2E\left[z_{1k} z_{2k}\right] E\left[z_{1k} z_{2k}\right] = \omega_{\zeta\zeta}\omega_{\eta\eta} + 2\omega_{\zeta\eta}^2, \text{ and}$$
$$\text{Var}\left(z_{1k} z_{2k}\right) = \omega_{\zeta\zeta}\omega_{\eta\eta} + \omega_{\zeta\eta}^2.$$

Similarly,

$$\begin{aligned}
\text{Var}\left(z_{2k}^2\right) &= E\left[z_{2k}^4\right] - \omega_{\eta\eta}^2 = 3\omega_{\eta\eta}^2 - \omega_{\eta\eta}^2 = 2\omega_{\eta\eta}^2, \\
Cov\left(z_{1k}, z_{1k} z_{2k}\right) &= E\left[z_{1k}^2 z_{2k}\right] - E\left[z_{1k}\right] E\left[z_{1k} z_{2k}\right] = 0, \\
Cov\left(z_{1k}^2, z_{1k} z_{2k}\right) &= E\left[z_{1k}^3 z_{2k}\right] - E\left[z_{1k}^2\right] E\left[z_{1k} z_{2k}\right] \\
&= 3\omega_{\zeta\eta}\omega_{\zeta\zeta} - \omega_{\zeta\zeta}\omega_{\zeta\eta} = 2\omega_{\zeta\eta}\omega_{\zeta\zeta}, \\
Cov\left(z_{1k}^2, z_{2k}^2\right) &= E\left[z_{1k}^2 z_{2k}^2\right] - \omega_{\zeta\zeta}\omega_{\eta\eta} = 2\omega_{\zeta\eta}^2, \text{ and}
\end{aligned}$$

$$V = \begin{bmatrix} V_{11} & 0 \\ 0 & V_{22} \end{bmatrix}, \text{ where}$$

$$V_{11} = \begin{bmatrix} \frac{1}{K}\sum_k \Pi_k^2 \omega_{\zeta\zeta} & \frac{1}{K}\sum_k \Pi_k \Pi_{Yk} \omega_{\zeta\zeta} & \frac{1}{K}\sum_k \Pi_k \Pi_{Yk} \omega_{\zeta\eta} & \frac{1}{K}\sum_k \Pi_k^2 \omega_{\zeta\eta} \\ . & \frac{1}{K}\sum_k \Pi_{Yk}^2 \omega_{\zeta\zeta} & \frac{1}{K}\sum_k \Pi_{Yk}^2 \omega_{\zeta\eta} & \frac{1}{K}\sum_k \Pi_k \Pi_{Yk} \omega_{\zeta\eta} \\ . & . & \frac{1}{K}\sum_k \Pi_{Yk}^2 \omega_{\eta\eta} & \frac{1}{K}\sum_k \Pi_k \Pi_{Yk} \omega_{\eta\eta} \\ . & . & . & \frac{1}{K}\sum_k \Pi_k^2 \omega_{\eta\eta} \end{bmatrix},$$

$$V_{22} = \begin{bmatrix} \omega_{\zeta\zeta}\omega_{\eta\eta} + \omega_{\zeta\eta}^2 & 2\omega_{\zeta\eta}\omega_{\eta\eta} & 2\omega_{\zeta\eta}\omega_{\zeta\zeta} \\ . & 2\omega_{\eta\eta}^2 & 2\omega_{\zeta\eta}^2 \\ . & . & 2\omega_{\zeta\zeta}^2 \end{bmatrix}.$$

If $\frac{1}{K}\sum_k \Pi_k^2 \to 0, \frac{1}{K}\sum_k \Pi_k \Pi_{Yk} \to 0, \frac{1}{K}\sum_k \Pi_{Yk}^2 \to 0$ under weak identification, then we obtain the Σ expression stated in the proposition. □

Proof of Proposition 1.3. Use n_q^Q and n_w^W to denote the number of observations in the instrument and covariate groups respectively, so

$$\begin{aligned}
\mu_3 &= \sum_i \sum_{j\neq i} G_{ij} R_i R_j = \sum_q \frac{n_q^Q}{n_q^Q - 1} \sum_{i\in\mathcal{N}_q^Q} \sum_{j\in\mathcal{N}_q^Q, j\neq i} \frac{1}{n_q^Q} R_i R_j - \sum_w \frac{n_w^W}{n_w^W - 1} \sum_{i\in\mathcal{N}_w^W} \sum_{j\in\mathcal{N}_w^W, j\neq i} \frac{1}{n_w^W} R_i R_j \\
&= \sum_q \frac{1}{n_q^Q - 1} \sum_{i\in\mathcal{N}_q^Q} \sum_{j\in\mathcal{N}_q^Q, j\neq i} R_i R_j - \sum_w \frac{1}{n_w^W - 1} \sum_{i\in\mathcal{N}_w^W} \sum_{j\in\mathcal{N}_w^W, j\neq i} R_i R_j \\
&= \sum_w \left(\sum_{q\in\mathcal{M}_w} \frac{1}{n_q^Q - 1} \sum_{i\in\mathcal{N}_q^Q} \sum_{j\in\mathcal{N}_q^Q, j\neq i} R_i R_j - \frac{1}{n_w^W - 1} \sum_{i\in\mathcal{N}_w^W} \sum_{j\in\mathcal{N}_w^W, j\neq i} R_i R_j \right) \\
&= \sum_w \left(\sum_{q\in\mathcal{M}_w} \frac{n_q^Q \left(n_q^Q - 1\right)}{n_q^Q - 1} \left(\pi_q + \gamma_w\right)^2 - \frac{1}{n_w^W - 1} \sum_{i\in\mathcal{N}_w^W} \sum_{j\in\mathcal{N}_w^W, j\neq i} R_i R_j \right).
\end{aligned}$$

Considering the second term,

$$\begin{aligned}
\sum_{i\in\mathcal{N}_w^W} \sum_{j\in\mathcal{N}_w^W, j\neq i} R_i R_j &= \sum_{i\in\mathcal{N}_w^W} \sum_{j\in\mathcal{N}_w^W, j\neq i} \left(\pi_{q(i)} + \gamma_w\right)\left(\pi_{q(j)} + \gamma_w\right) \\
&= \sum_{q\in\mathcal{M}_w} \sum_{i\in\mathcal{N}_q} \sum_{j\in\mathcal{N}_w^W, j\neq i} \left(\pi_{q(i)} + \gamma_w\right)\left(\pi_{q(j)} + \gamma_w\right) \\
&= \sum_{q\in\mathcal{M}_w} n_q^Q \left(n_q^Q - 1\right)\left(\pi_q + \gamma_w\right)^2 + \sum_{q\in\mathcal{M}_w} \sum_{i\in\mathcal{N}_q} \sum_{j\in\mathcal{N}_w^W, j\neq i} \left(\pi_{q(i)} + \gamma_w\right)\left(\pi_{q(j)} + \gamma_w\right) \\
&= \sum_{q\in\mathcal{M}_w} n_q^Q \left(n_q^Q - 1\right)\left(\pi_q + \gamma_w\right)^2 + \sum_{q\in\mathcal{M}_w} \sum_{q'\in\mathcal{M}_w, q'\neq q} n_q^Q n_{q'}^Q \left(\pi_q + \gamma_w\right)\left(\pi_{q'} + \gamma_w\right).
\end{aligned}$$

Since

$$\sum_{q\in\mathcal{M}_w} n_q^Q \left(\pi_q + \gamma_w\right)^2 - \frac{1}{n_w^W - 1} \sum_{q\in\mathcal{M}_w} n_q^Q \left(n_q^Q - 1\right)\left(\pi_q + \gamma_w\right)^2 = \sum_{q\in\mathcal{M}_w} n_q^Q \left(\frac{n_w^W - n_q^Q}{n_w^W - 1}\right)\left(\pi_q + \gamma_w\right)^2,$$

and $n_w^W = \sum_{q \in \mathcal{M}_w} n_q^Q$, we obtain

$$\begin{aligned}
\mu_3 &= \sum_w \left(\sum_{q \in \mathcal{M}_w} n_q^Q \left(\frac{n_w^W - n_q^Q}{n_w^W - 1} \right) (\pi_q + \gamma_w)^2 - \frac{1}{n_w^W - 1} \sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' \neq q} n_q^Q n_{q'}^Q (\pi_q + \gamma_w)(\pi_{q'} + \gamma_w) \right) \\
&= \sum_w \frac{1}{n_w^W - 1} \left(\sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' \neq q} n_q^Q n_{q'}^Q (\pi_q + \gamma_w)^2 - \sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' \neq q} n_q^Q n_{q'}^Q (\pi_q + \gamma_w)(\pi_{q'} + \gamma_w) \right) \\
&= \sum_w \frac{1}{n_w^W - 1} \left(\sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' \neq q} n_q^Q n_{q'}^Q (\pi_q + \gamma_w)(\pi_q - \pi_{q'}) \right).
\end{aligned}$$

Then, observe that

$$\begin{aligned}
&\sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' \neq q} n_q^Q n_{q'}^Q (\pi_q + \gamma)(\pi_q - \pi_{q'}) \\
&= \sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' < q} n_q^Q n_{q'}^Q (\pi_q + \gamma)(\pi_q - \pi_{q'}) + \sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' > q} n_q^Q n_{q'}^Q (\pi_q + \gamma)(\pi_q - \pi_{q'}) \\
&= \sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' < q} n_q^Q n_{q'}^Q (\pi_q + \gamma)(\pi_q - \pi_{q'}) - \sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' < q} n_q^Q n_{q'}^Q (\pi_{q'} + \gamma)(\pi_q - \pi_{q'}) \\
&= \sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' < q} n_q^Q n_{q'}^Q (\pi_q - \pi_{q'})(\pi_q + \gamma - \pi_{q'} - \gamma) = \sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' < q} n_q^Q n_{q'}^Q (\pi_q - \pi_{q'})^2,
\end{aligned}$$

where the second equality switches the indices of q and q' in the second element. Hence,

$$\mu_3 = \sum_w \frac{1}{n_w^W - 1} \left(\sum_{q \in \mathcal{M}_w} \sum_{q' \in \mathcal{M}_w, q' < q} n_q^Q n_{q'}^Q (\pi_q - \pi_{q'})^2 \right) \geq 0.$$

Analogously,

$$\begin{aligned}
\mu_2 &= \sum_w \left(\sum_{q\in\mathcal{M}_w} \frac{1}{n_q^Q - 1} \sum_{i\in\mathcal{N}_q^Q} \sum_{j\in\mathcal{N}_q^Q, j\neq i} R_i R_{Yj} - \frac{1}{n_w^W - 1} \sum_{i\in\mathcal{N}_w^W} \sum_{j\in\mathcal{N}_w^W, j\neq i} R_i R_{Yj} \right) \\
&= \sum_w \left(\sum_{q\in\mathcal{M}_w} n_q^Q \left(\pi_q + \gamma_w\right)\left(\pi_{Yq} + \gamma_w\right) - \frac{1}{n_w^W - 1} \sum_{q\in\mathcal{M}_w} n_q^Q \left(n_q^Q - 1\right)\left(\pi_q + \gamma_w\right)\left(\pi_{Yq} + \gamma_w\right) \right) \\
&\quad - \sum_w \frac{1}{n_w^W - 1} \sum_{q\in\mathcal{M}_w} \sum_{q'\in\mathcal{M}_w, q'\neq q} n_q^Q n_{q'}^Q \left(\pi_q + \gamma_w\right)\left(\pi_{Yq'} + \gamma_w\right) \\
&= \sum_w \left(\sum_{q\in\mathcal{M}_w} n_q^Q \left(\frac{n_w^W - n_q^Q}{n_w^W - 1} \right) \left(\pi_q + \gamma_w\right)\left(\pi_{Yq} + \gamma_w\right) \right. \\
&\quad \left. - \frac{1}{n_w^W - 1} \sum_{q\in\mathcal{M}_w} \sum_{q'\in\mathcal{M}_w, q'\neq q} n_q^Q n_{q'}^Q \left(\pi_q + \gamma_w\right)\left(\pi_{Yq'} + \gamma_w\right) \right) \\
&= \sum_w \frac{1}{n_w^W - 1} \left(\sum_{q\in\mathcal{M}_w} \sum_{q'\in\mathcal{M}_w, q'\neq q} n_q^Q n_{q'}^Q \left(\pi_q + \gamma_w\right)\left(\pi_{Yq} - \pi_{Yq'}\right) \right) \\
&= \sum_w \frac{1}{n_w^W - 1} \left(\sum_{q\in\mathcal{M}_w} \sum_{q'\in\mathcal{M}_w, q'<q} n_q^Q n_{q'}^Q \left(\pi_q - \pi_{q'}\right)\left(\pi_{Yq} - \pi_{Yq'}\right) \right),
\end{aligned}$$

and

$$\mu_1 = \sum_w \frac{1}{n_w^W - 1} \left(\sum_{q\in\mathcal{M}_w} \sum_{q'\in\mathcal{M}_w, q'<q} n_q^Q n_{q'}^Q \left(\pi_{Yq} - \pi_{Yq'}\right)^2 \right) \geq 0.$$

□

Proof of Proposition 1.4. Fix any alternative $\left(\pi^A, \pi_Y^A\right) \in \mathcal{S}$ with a corresponding $\left(\mu_1^A, \mu_2^A, \mu_3^A\right)$. Due to the restriction in $\mathcal{S}$,

$$\begin{pmatrix} \mu_1^H \\ \mu_2^H \\ \mu_3^H \end{pmatrix} = \begin{pmatrix} \mu_1^A - \frac{\sigma_{12}}{\sigma_{22}} \mu_2^A \\ 0 \\ \mu_3^A - \frac{\sigma_{23}}{\sigma_{22}} \mu_2^A \end{pmatrix}$$

is in the null space. The Neyman-Pearson test for μ^H against μ^A rejects large values of:

$$\log \frac{dN\left(\mu^A, \Sigma\right)}{dN\left(\mu^H, \Sigma\right)} = \frac{\mu_2^A}{\sigma_{22}} X_2 - \frac{1}{2} \frac{\left(\mu_2^A\right)^2}{\sigma_{22}}.$$

Hence, the most powerful test rejects large values of X_2, which is what LM does. By Lehmann and Romano (2005) Theorem 3.8.1(i), since LM is valid for any distribution in the null space (by Theorem 1.1) and it is most powerful for some distribution in the null space, LM is most powerful for testing the composite null against the given alternative $\left(\pi^A, \pi_Y^A\right)$. □

Proof of Proposition 1.5. The first two are straightforward: $C_S = \mu_3 / (c-1)$ and $\beta = \mu_2/\mu_3$ imply $\mu_3 = (c-1) C_S$ and $\mu_2 = (c-1) C_S \beta$. For μ_1, observe that:

$$h = \sqrt{\frac{1}{\sqrt{K}} \frac{1}{c-1} \left(\mu_1 - \frac{\mu_2^2}{\mu_3}\right)} = \sqrt{\frac{1}{\sqrt{K}} \left(\mu_1 - C_S \beta^2\right)}, \text{ and}$$

$$C_H = \sqrt{K} h^2 = \mu_1 / (c-1) - C_S \beta^2, \text{ so}$$

$$(c-1)\left(C_S \beta^2 + C_H\right) = (c-1)\left(C_S \beta^2 + \mu_1 / (c-1) - C_S \beta^2\right) = \mu_1$$

as required. Next, since $\sigma_{vv} = \sqrt{\frac{\sigma_{33} c}{2(c-1)}}$, $\sigma_{33} = 2\frac{c-1}{c}\sigma_{vv}^2$ is immediate. Similarly, with $\sigma_{\varepsilon v} = \frac{1}{\sigma_{vv}}\left(\frac{\sigma_{23} c}{2(c-1)} - \sigma_{vv}^2 \beta\right)$, $\sigma_{23} = 2\frac{c-1}{c}\sigma_{vv}\left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)$. From these two expressions, we can observe that:

$$\left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)^2 = \frac{c}{2(c-1)} \frac{\sigma_{23}^2}{\sigma_{33}}.$$

To obtain an expression for σ_{22}, rearrange $\sigma_{\varepsilon\varepsilon} = \frac{1}{\sigma_{vv}} \frac{c}{c-1}\left(\sigma_{22} - \frac{\sigma_{23}^2}{\sigma_{33}}\right) + \frac{\sigma_{\varepsilon v}^2}{\sigma_{vv}} \geq 0$:

$$\begin{aligned}
\sigma_{22} &= \frac{\sigma_{23}^2}{\sigma_{33}} + \frac{c-1}{c}\left(\sigma_{\varepsilon\varepsilon}\sigma_{vv} - \sigma_{\varepsilon v}^2\right) \\
&= \frac{c-1}{c}\left(\sigma_{vv}\left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\varepsilon v}\beta\right) + \left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)^2\right) + o(1),
\end{aligned}$$

where the final step uses $\sigma_{\xi\xi} = h/\sigma_{vv}$. This expression for σ_{22} is of the form required in Lemma 1.8. Then,

$$\begin{aligned} det\left(\Sigma_{SF}\right) &= \sigma_{\varepsilon\varepsilon}\sigma_{\xi\xi}\sigma_{vv} - \sigma_{\varepsilon\varepsilon}h^2 - \sigma_{\varepsilon\xi}^2\sigma_{vv} + 2\sigma_{\varepsilon\xi}\sigma_{\varepsilon v}h - \sigma_{\xi\xi}\sigma_{\varepsilon v}^2 \\ &= \sigma_{\varepsilon\varepsilon}\sigma_{\xi\xi}\sigma_{vv} - \sigma_{\varepsilon\varepsilon}h^2 - \sigma_{\xi\xi}\sigma_{\varepsilon v}^2 = \sigma_{\varepsilon\varepsilon}h - \sigma_{\varepsilon\varepsilon}h^2 - h\frac{\sigma_{\varepsilon v}^2}{\sigma_{vv}}; \text{ and} \\ det\left(\Sigma_{SF}\right)/h &= \sigma_{\varepsilon\varepsilon} - \frac{\sigma_{\varepsilon v}^2}{\sigma_{vv}} - \sigma_{\varepsilon\varepsilon}h = \sigma_{\varepsilon\varepsilon} - \frac{\sigma_{\varepsilon v}^2}{\sigma_{vv}} + o(1). \end{aligned}$$

An analogous argument holds for $\sigma_{\xi vk} = -h$. From the σ_{22} equation, $\sigma_{\varepsilon\varepsilon} - \frac{\sigma_{\varepsilon v}^2}{\sigma_{vv}} = \frac{c}{c-1}\left(\sigma_{22} - \frac{\sigma_{23}^2}{\sigma_{33}}\right) \geq 0$, which delivers the result that $det\left(\Sigma_{SF}\right)/h \to C_D \geq 0$. □

Proof of Proposition 1.6. The A expressions can be written as:

$$\begin{aligned} A_1 &= \sum_i \sum_{j\neq i} \sum_{k\neq i} \sum_{l\neq k} \check{M}_{il,-ijk} G_{ij} X_j G_{ik} X_k \left(Y_i Y_l - X_i Y_l \beta_0 - Y_i X_l \beta_0 + X_i X_l \beta_0^2\right); \\ A_2 &= \sum_i \sum_{j\neq i} \sum_{k\neq i} \sum_{l\neq k} \check{M}_{il,-ijk} G_{ij} X_j G_{ki} X_l \left(Y_i Y_k - X_i Y_k \beta_0 - Y_i X_k \beta_0 + X_i X_k \beta_0^2\right); \\ A_3 &= \sum_i \sum_{j\neq i} \sum_{k\neq i} \sum_{l\neq k} \check{M}_{il,-ijk} X_l G_{ji} G_{ki} X_i \left(Y_j Y_k - X_j Y_k \beta_0 - Y_j X_k \beta_0 + X_j X_k \beta_0^2\right); \\ A_4 &= -\sum_i \sum_{j\neq i} \sum_{k\neq j} \sum_{l\neq i,k} \check{M}_{jl,-ijk} \check{M}_{ik,-ij} G_{ji}^2 X_i X_k \left(Y_j Y_l - X_j Y_l \beta_0 - Y_j X_l \beta_0 + X_j X_l \beta_0^2\right); \text{ and} \\ A_5 &= -\sum_i \sum_{j\neq i} \sum_{k\neq j} \sum_{l\neq i,k} \check{M}_{ik,-ij} \check{M}_{jl,-ijk} G_{ij} G_{ji} X_k X_l \left(Y_i Y_j - X_i Y_j \beta_0 - Y_i X_j \beta_0 + X_i X_j \beta_0^2\right). \end{aligned}$$

Since these terms have a quadratic form, the variance estimator is also quadratic in β_0, i.e.,

$$\hat{V}_{LM} = B_0 + B_1\beta_0 + B_2\beta_0^2,$$

where the B's can be worked out by collecting the expressions above. For instance,

$$B_0 = \sum_i \sum_{j\neq i} \sum_{k\neq i} \sum_{l\neq k} \check{M}_{il,-ijk} G_{ij} X_j G_{ik} X_k Y_i Y_l + 2 \sum_i \sum_{j\neq i} \sum_{k\neq i} \sum_{l\neq k} \check{M}_{il,-ijk} G_{ij} X_j G_{ki} X_l Y_i Y_k$$
$$+ \sum_i \sum_{j\neq i} \sum_{k\neq i} \sum_{l\neq k} \check{M}_{il,-ijk} X_l G_{ji} G_{ki} X_i Y_j Y_k$$
$$- \sum_i \sum_{j\neq i} \sum_{k\neq j} \sum_{l\neq i,k} \check{M}_{jl,-ijk} \check{M}_{ik,-ij} G_{ji}^2 X_i X_k Y_j Y_l - \sum_i \sum_{j\neq i} \sum_{k\neq j} \sum_{l\neq i,k} \check{M}_{ik,-ij} \check{M}_{jl,-ijk} G_{ij} G_{ji} X_k X_l Y_i Y_j$$

B_1 and B_2 are analogous by collecting the coefficients on β_0, β_0^2 from expressions A_1 to A_5.

The test does not reject:

$$\frac{\left(KT_{YX} - KT_{XX}\beta_0\right)^2}{B_0 + B_1\beta_0 + B_2\beta_0^2} \leq q \Leftrightarrow \left(KT_{XX}^2 - qB_2\right)\beta_0^2 - \left(2KT_{YX}T_{XX} + qB_1\right)\beta_0 + \left(KT_{YX}^2 - qB_0\right) \leq 0.$$

Solutions exist when:

$$D := \left(2KT_{YX}T_{XX} + qB_1\right)^2 - 4\left(KT_{XX}^2 - qB_2\right)\left(KT_{YX}^2 - qB_0\right) \geq 0.$$

The rest of the lemma is immediate from properties of solving quadratic inequalities. □

Proofs for Lemmas in Appendix 1.C

Proof of Lemma 1.7. The joint distribution of $(Y', X')'$ is:

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} Z\pi_Y \\ Z\pi \end{bmatrix}, \begin{bmatrix} I_n\omega_{\zeta\zeta} & I_n\omega_{\zeta\eta} \\ I_n\omega_{\zeta\eta} & I_n\omega_{\eta\eta} \end{bmatrix} \right).$$

Stack them together with their predicted values $PY = Z\left(Z'Z\right)^{-1}Z'Y$ and $PX = Z\left(Z'Z\right)^{-1}Z'X$:

$$\begin{bmatrix} Y \\ X \\ P'Y \\ P'X \end{bmatrix} \sim N\left(\begin{bmatrix} Z\pi_Y \\ Z\pi \\ Z\pi_Y \\ Z\pi \end{bmatrix}, \begin{bmatrix} I_n\omega_{\zeta\zeta} & I_n\omega_{\zeta\eta} & \omega_{\zeta\zeta}P' & \omega_{\zeta\eta}P' \\ I_n\omega_{\zeta\eta} & I_n\omega_{\eta\eta} & \omega_{\zeta\eta}P' & \omega_{\eta\eta}P' \\ \omega_{\zeta\zeta}P' & \omega_{\zeta\eta}P' & \omega_{\zeta\zeta}P' & \omega_{\zeta\eta}P' \\ \omega_{\zeta\eta}P' & \omega_{\eta\eta}P' & \omega_{\zeta\eta}P' & \omega_{\eta\eta}P' \end{bmatrix}\right).$$

Then, the conditional normal distribution is:

$$\begin{aligned} \begin{bmatrix} Y \\ X \end{bmatrix} \mid \begin{bmatrix} Z\left(Z'Z\right)^{-1}Z'Y \\ Z\left(Z'Z\right)^{-1}Z'X \end{bmatrix} &\sim N\left(\begin{bmatrix} Z\pi_Y \\ Z\pi \end{bmatrix} + \begin{bmatrix} Z\left(Z'Z\right)^{-1}Z'Y - Z\pi_Y \\ Z\left(Z'Z\right)^{-1}Z'X - Z\pi \end{bmatrix}, V\right) \\ &= N\left(\begin{bmatrix} Z\left(Z'Z\right)^{-1}Z'Y \\ Z\left(Z'Z\right)^{-1}Z'X \end{bmatrix}, V\right) = N\left(\begin{bmatrix} PY \\ PX \end{bmatrix}, V\right) \end{aligned}$$

Hence, PX and PY (i.e, $Z'X$, $Z'Y$) are sufficient statistics for π_Y, π.

To show that $(s_1's_1, s_1's_2, s_2's_2)$ is a maximal invariant, let F be some conformable orthogonal matrix so $F'F = I$. For invariance, let $s_1^* = Fs_1$. Then, $s_1^{*\prime}s_1^* = s_1'F'Fs_1 = s_1's_1$. Invariance of $(s_1's_2, s_2's_2)$ is analogous. Maximality states that if $s_1^{*\prime}s_1^* = s_1's_1$, then $s_1^* = Fs_1$ for some F. Suppose not. This means $s_1^* = Gs_1$, and G is not an orthogonal matrix but yet $s_1^{*\prime}s_1^* = s_1's_1$. Since G is not an orthogonal matrix, $G'G \neq I$. Hence, $s_1^{*\prime}s_1^* = s_1'G'Gs_1 \neq s_1's_1$, a contradiction. To obtain the distribution,

$$\begin{bmatrix} s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} \left(Z'Z\right)^{-1/2}Z'\left(Z\pi_Y + \zeta\right) \\ \left(Z'Z\right)^{-1/2}Z'\left(Z\pi + \eta\right) \end{bmatrix} = \begin{bmatrix} \left(Z'Z\right)^{1/2}\pi_Y \\ \left(Z'Z\right)^{1/2}\pi \end{bmatrix} + \begin{bmatrix} \left(Z'Z\right)^{-1/2}Z'\zeta \\ \left(Z'Z\right)^{-1/2}Z'\eta \end{bmatrix}.$$

Since $\mathrm{Var}\left(\left(Z'Z\right)^{-1/2}Z'\eta\right) = \left(Z'Z\right)^{-1/2}Z'\omega_{\eta\eta}Z\left(Z'Z\right)^{-1/2} = I_K\omega_{\eta\eta}$,

$$\begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \sim N\left(\begin{pmatrix} \left(Z'Z\right)^{1/2}\pi_Y \\ \left(Z'Z\right)^{1/2}\pi \end{pmatrix}, \Omega \otimes I_K\right).$$

□

Proof of Lemma 1.8. I work out the μ's first. Using the judge structure, $\sum_i M_{ii}^2 = \sum_k \frac{(c-1)^2}{c}, \sum_i \sum_{j\neq i} P_{ij} = \sum_k \frac{c-1}{c}$. We have also chosen $\pi_k, \sigma_{\xi vk}$ such that $\sum_k \pi_k = 0, \sum_k \sigma_{\xi vk} = 0, \sum_k \pi_k\sigma_{\xi vk} = 0$. Then, we get the result for means:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{K}}\sum_k (c-1)\left(\pi_k^2\beta^2 + 2\pi_k\beta\sigma_{\xi vk} + \sigma_{\xi vk}^2\right) \\ \frac{1}{\sqrt{K}}\sum_k (c-1)\left(\pi_k^2\beta + \pi_k\sigma_{\xi vk}\right) \\ \frac{1}{\sqrt{K}}\sum_k (c-1)\pi_k^2 \end{pmatrix} = \begin{pmatrix} \sqrt{K}(c-1)\left(s^2\beta^2 + h^2\right) \\ \sqrt{K}(c-1)s^2\beta \\ \sqrt{K}(c-1)s^2 \end{pmatrix}.$$

Using a derivation similar to that of the lemma for V_{LM} expression,

$$
\begin{aligned}
K\sigma_{22} &= \sum_i \sum_{j\neq i} \sum_{k\neq i} \left(G_{ji}G_{ki}E\left[\zeta_i^2\right] R_j R_k + 2G_{ij}G_{ki}E\left[\eta_i\zeta_i\right] R_{Yj}R_k + G_{ij}G_{ik}E\left[\eta_i^2\right] R_{Yj}R_{Yk} \right) \\
&\quad + \sum_i \sum_{j\neq i} \left(G_{ij}^2 E\left[\eta_i^2\right] E\left[\zeta_j^2\right] + G_{ij}G_{ji}E\left[\eta_i\zeta_i\right] E\left[\eta_j\zeta_j\right] \right); \\
K\sigma_{11} &= \sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\zeta_i^2\right] R_{Yj}R_{Yk} \left(G_{ji}G_{ki} + 2G_{ij}G_{ki} + G_{ij}G_{ik} \right) \\
&\quad + \sum_i \sum_{j\neq i} E\left[\zeta_i^2\right] E\left[\zeta_j^2\right] \left(G_{ij}^2 + G_{ij}G_{ji} \right); \\
K\sigma_{33} &= \sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\eta_i^2\right] R_j R_k \left(G_{ji}G_{ki} + 2G_{ij}G_{ki} + G_{ij}G_{ik} \right) \\
&\quad + \sum_i \sum_{j\neq i} E\left[\eta_i^2\right] E\left[\eta_j^2\right] \left(G_{ij}^2 + G_{ij}G_{ji} \right); \\
K\sigma_{12} &= \sum_i \sum_{j\neq i} \sum_{k\neq i} \left(G_{ji}G_{ki}E\left[\zeta_i^2\right] R_j R_{Yk} + 2G_{ij}G_{ki}E\left[\zeta_i^2\right] R_{Yj}R_k + G_{ij}G_{ik}E\left[\eta_i\zeta_i\right] R_{Yj}R_{Yk} \right) \\
&\quad + \sum_i \sum_{j\neq i} E\left[\eta_i\zeta_i\right] E\left[\zeta_j^2\right] \left(G_{ij}^2 + G_{ij}G_{ji} \right); \\
K\sigma_{23} &= \sum_i \sum_{j\neq i} \sum_{k\neq i} \left(G_{ji}G_{ki}E\left[\eta_i^2\right] R_{Yj}R_k + 2G_{ij}G_{ki}E\left[\eta_i^2\right] R_j R_{Yk} + G_{ij}G_{ik}E\left[\eta_i\zeta_i\right] R_j R_k \right) \\
&\quad + \sum_i \sum_{j\neq i} E\left[\eta_i\zeta_i\right] E\left[\eta_j^2\right] \left(G_{ij}^2 + G_{ij}G_{ji} \right); \text{ and} \\
K\sigma_{13} &= \sum_i \sum_{j\neq i} \sum_{k\neq i} E\left[\eta_i\zeta_i\right] R_{Yj}R_k \left(G_{ji}G_{ki} + 2G_{ij}G_{ki} + G_{ij}G_{ik} \right) \\
&\quad + \sum_i \sum_{j\neq i} E\left[\eta_i\zeta_i\right] E\left[\eta_j\zeta_j\right] \left(G_{ij}^2 + G_{ij}G_{ji} \right).
\end{aligned}
$$

The equalities hold regardless of whether identification is strong or weak and whether heterogeneity converges or not. Without covariates, $G = P$ is symmetric and the above expressions simplify. For instance,

$$
K\sigma_{22} = \sum_k \frac{(c-1)^2}{c} \left(\omega_{\zeta\zeta k}\pi_k^2 + 2\omega_{\zeta\eta k}\pi_k\pi_{Yk} + \omega_{\eta\eta k}\pi_{Yk}^2 \right) + \sum_k \frac{c-1}{c} \left(\omega_{\eta\eta k}\omega_{\zeta\zeta k} + \omega_{\zeta\eta k}^2 \right).
$$

Evaluate the terms in the expression. For higher moments of π_k, $\sum_k \pi_k^2 = Ks^2$, $\sum_k \pi_k^3 = 0$, and $\sum_k \pi_k^4 = Ks^4$. Similarly, $\sum_k \pi_k^3 \sigma_{\xi v} = 0$. Treating the heterogeneity in the same way, $\sum_k \sigma_{\xi v}^2 = Kh^2$. Then,

$$\begin{aligned}
\sum_k \omega_{\zeta\zeta k}\pi_k^2 &= \sum_k \left(\pi_k^2\sigma_{\xi\xi} + 2\pi_k\beta\sigma_{\xi vk} + 2\pi_k\sigma_{\varepsilon\xi} + \sigma_{\varepsilon\varepsilon} - \sigma_{\xi vk}^2 + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right)\pi_k^2 \\
&= s^2K\left(s^2\sigma_{\xi\xi} + \sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + h^2 + 2\sigma_{\varepsilon v}\beta\right); \text{ and} \\
\sum_k \omega_{\zeta\eta k}\pi_k\pi_{Yk} &= \sum_k \left(\pi_k\sigma_{\xi vk} + \sigma_{vv}\beta + \sigma_{\varepsilon v}\right)\pi_k\left(\pi_k\beta + \sigma_{\xi vk}\right) \\
&= \sum_k \left(\sigma_{vv}\beta^2\pi_k^2 + \sigma_{\varepsilon v}\pi_k^2\beta + \pi_k^2\sigma_{\xi vk}^2\right) = s^2K\left(\sigma_{vv}\beta^2 + \sigma_{\varepsilon v}\beta + h^2\right).
\end{aligned}$$

Now, for the P_{ij}^2 part,

$$\begin{aligned}
\sum_k \omega_{\eta\eta k}\omega_{\zeta\zeta k} &= \sum_k \sigma_{vv}\left(\pi_k^2\sigma_{\xi\xi} + 2\pi_k\beta\sigma_{\xi vk} + 2\pi_k\sigma_{\varepsilon\xi} + \sigma_{\varepsilon\varepsilon} - \sigma_{\xi vk}^2 + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right) \\
&= \sum_k \sigma_{vv}\left(\pi_k^2\sigma_{\xi\xi} + \sigma_{\varepsilon\varepsilon} - \sigma_{\xi vk}^2 + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right) \\
&= K\sigma_{vv}\left(s^2\sigma_{\xi\xi} + \sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + h^2 + 2\sigma_{\varepsilon v}\beta\right); \text{ and} \\
\sum_k \omega_{\zeta\eta k}^2 &= \sum_k \left(\pi_k\sigma_{\xi vk}\pi_k\sigma_{\xi vk} + \sigma_{vv}\beta\pi_k\sigma_{\xi vk} + \sigma_{\varepsilon v}\pi_k\sigma_{\xi vk} + \pi_k\sigma_{\xi vk}\sigma_{vv}\beta + \sigma_{vv}\beta\sigma_{vv}\beta + \sigma_{\varepsilon v}\sigma_{vv}\beta\right) \\
&\quad + \sum_k \left(\pi_k\sigma_{\xi vk}\sigma_{\varepsilon v} + \sigma_{vv}\beta\sigma_{\varepsilon v} + \sigma_{\varepsilon v}^2\right) \\
&= \sum_k \left(\pi_k^2\sigma_{\xi vk}^2 + \sigma_{vv}^2\beta^2 + \sigma_{\varepsilon v}\sigma_{vv}\beta + \sigma_{vv}\beta\sigma_{\varepsilon v} + \sigma_{\varepsilon v}^2\right) = K\left(s^2h^2 + \left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)^2\right).
\end{aligned}$$

Combine the expressions for σ_{22} and impose asymptotics where $s \to 0$ and $h \to 0$:

$$\begin{aligned}
\sigma_{22} &= \frac{1}{K}\sum_k \frac{(c-1)^2}{c} h^2 \\
&\quad + \frac{1}{K}\sum_k \frac{c-1}{c}\left(\sigma_{vv}\left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + h^2 + 2\sigma_{\varepsilon v}\beta\right) + \left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)^2\right) + o(1) \\
&= \frac{c-1}{c}\left(\sigma_{vv}\left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\varepsilon v}\beta\right) + \left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)^2\right) + o(1).
\end{aligned}$$

Next, evaluate a few more sums that feature in the other σ expressions:

$$\begin{aligned}
\sum_k \omega_{\zeta\zeta}\pi_{Yk}^2 &= \sum_k \left(\pi_k^2\sigma_{\xi\xi} + 2\pi_k\beta\sigma_{\xi vk} + 2\pi_k\sigma_{\varepsilon\xi} + \sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + \sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right) \\
&\qquad \left(\pi_k^2\beta^2 + 2\pi_k\sigma_{\xi vk} + \sigma_{\xi v}^2\right) \\
\frac{1}{K}\sum_k \omega_{\zeta\zeta}\pi_{Yk}^2 &= \frac{1}{K}\sum_k \sigma_{\xi v}^2\left(\pi_k^2\sigma_{\xi\xi} + 2\pi_k\beta\sigma_{\xi vk} + 2\pi_k\sigma_{\varepsilon\xi} + \sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + \sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right) \\
&= h^2\left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + h^2 + 2\sigma_{\varepsilon v}\beta\right) = o(1); \\
\frac{1}{K}\sum_k \omega_{\zeta\zeta}^2 &= \frac{1}{K}\sum_k \left(\pi_k^2\sigma_{\xi\xi} + 2\pi_k\beta\sigma_{\xi vk} + 2\pi_k\sigma_{\varepsilon\xi} + \sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + \sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right)^2 \\
&= \frac{1}{K}\sum_k \left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + \sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right)^2 = \left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\varepsilon v}\beta\right)^2; \\
\frac{1}{K}\sum_k \omega_{\zeta\eta}\pi_{Yk}^2 &= \frac{1}{K}\sum_k \left(\pi_k\sigma_{\xi vk} + \sigma_{vv}\beta + \sigma_{\varepsilon v}\right)\left(\pi_k^2\beta^2 + 2\pi_k\sigma_{\xi vk} + \sigma_{\xi v}^2\right) \\
&= h^2\left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right) = o(1); \text{ and} \\
\frac{1}{K}\sum_k \omega_{\zeta\eta}\omega_{\zeta\zeta} &= \frac{1}{K}\sum_k \left(\pi_k\sigma_{\xi vk} + \sigma_{vv}\beta + \sigma_{\varepsilon v}\right) \\
&\qquad \left(\pi_k^2\sigma_{\xi\xi} + 2\pi_k\beta\sigma_{\xi vk} + 2\pi_k\sigma_{\varepsilon\xi} + \sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + \sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right) \\
&= \frac{1}{K}\sum_k \left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)\left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + \sigma_{\xi vk}^2 + 2\sigma_{\varepsilon v}\beta\right) \\
&= \left(\sigma_{vv}\beta + \sigma_{\varepsilon v}\right)\left(\sigma_{\varepsilon\varepsilon} + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\varepsilon v}\beta\right) + o(1).
\end{aligned}$$

Using these results,

$$\begin{aligned}
\sigma_{22} &= \frac{c-1}{c}\left(\sigma_{vv}\left(\sigma_{\varepsilon\varepsilon}+\sigma_{vv}\beta^2+\sigma_{vv}\sigma_{\xi\xi}+2\sigma_{\varepsilon v}\beta\right)+\left(\sigma_{vv}\beta+\sigma_{\varepsilon v}\right)^2\right)+o(1);\\
\sigma_{11} &= 2\frac{c-1}{c}\left(\sigma_{\varepsilon\varepsilon}+\sigma_{vv}\beta^2+\sigma_{vv}\sigma_{\xi\xi}+2\sigma_{\varepsilon v}\beta\right)^2+o(1);\\
\sigma_{33} &= 2\frac{c-1}{c}\sigma_{vv}^2+o(1);\\
\sigma_{12} &= 2\frac{c-1}{c}\left(\sigma_{vv}\beta+\sigma_{\varepsilon v}\right)\left(\sigma_{\varepsilon\varepsilon}+\sigma_{vv}\beta^2+\sigma_{vv}\sigma_{\xi\xi}+2\sigma_{\varepsilon v}\beta\right)+o(1);\\
\sigma_{23} &= 2\frac{c-1}{c}\sigma_{vv}\left(\sigma_{vv}\beta+\sigma_{\varepsilon v}\right)+o(1);\ \text{and}\\
\sigma_{13} &= 2\frac{c-1}{c}\left(\sigma_{vv}\beta+\sigma_{\varepsilon v}\right)^2+o(1).
\end{aligned}$$

Hence, $\sigma_{13}=\sigma_{23}^2/\sigma_{33}+o(1)$ is immediate. Further, for σ_{12},

$$\begin{aligned}
2\frac{\sigma_{23}}{\sigma_{33}}\left(\sigma_{22}-\frac{\sigma_{23}^2}{2\sigma_{33}}\right) &= 2\frac{\sigma_{vv}\beta+\sigma_{\varepsilon v}}{\sigma_{vv}}\left(\sigma_{22}-\frac{\left(2\frac{c-1}{c}\sigma_{vv}\left(\sigma_{vv}\beta+\sigma_{\varepsilon v}\right)\right)^2}{2\times 2\frac{c-1}{c}\sigma_{vv}^2}\right)+o(1)\\
&= 2\frac{c-1}{c}\left(\sigma_{vv}\beta+\sigma_{\varepsilon v}\right)\left(\sigma_{\varepsilon\varepsilon}+\sigma_{vv}\beta^2+\sigma_{vv}\sigma_{\xi\xi}+2\sigma_{\varepsilon v}\beta\right)+o(1)=\sigma_{12}+o(1).
\end{aligned}$$

Finally, the σ_{11} can be obtained:

$$\begin{aligned}
\frac{4}{\sigma_{33}}\left(\sigma_{22}-\frac{\sigma_{23}^2}{2\sigma_{33}}\right)^2 &= \frac{2}{\frac{c-1}{c}\sigma_{vv}^2}\left(\frac{c-1}{c}\left(\sigma_{vv}\left(\sigma_{\varepsilon\varepsilon}+\sigma_{vv}\beta^2+\sigma_{vv}\sigma_{\xi\xi}+h^2+2\sigma_{\varepsilon v}\beta\right)\right)\right)^2+o(1)\\
&= 2\frac{c-1}{c}\left(\sigma_{\varepsilon\varepsilon}+\sigma_{vv}\beta^2+\sigma_{vv}\sigma_{\xi\xi}+2\sigma_{\varepsilon v}\beta\right)^2+o(1)=\sigma_{11}+o(1).
\end{aligned}$$

□

Derivations for Simulations

Derivation for continuous setup without covariates.

This subsection derives expressions for objects in the reduced-form model. Comparing the first-stage equations, $\eta_i = v_i$. As a corollary, for all i, $E\left[\eta_i^2\right]=\sigma_{vv}$. Then, $\zeta_i =$

$Z_i'(\pi\beta_i - \pi_Y) + v_i\beta_i + \varepsilon_i$. Define π_Y using $E[\zeta_i] = 0$ and $E[v_i\beta_i] = E[v_i(\beta + \xi_i)] = \sigma_{\xi v k(i)}$, which implies $\pi_{Yk} = \pi_k\beta + \sigma_{\xi v k}$. Hence, we can rewrite ζ_i as:

$$\zeta_i = \pi_{k(i)}\xi_i - \sigma_{\xi v k(i)} + v_i\beta + v_i\xi_i + \varepsilon_i.$$

By substituting the expression for ζ_i, the covariance is $E[\eta_i\zeta_i \mid k] = \pi_k\sigma_{\xi v k} + \sigma_{vv}\beta + E[v_i^2\xi_i] + \sigma_{\varepsilon v}$. By Isserlis' theorem, $E[v_i^2\xi_i] = 0$, so $E[\eta_i\zeta_i \mid k] = \pi_k\sigma_{\xi v k} + \sigma_{vv}\beta + \sigma_{\varepsilon v}$. The variance of ζ_i can be derived analogously. Since $E[v_i^2\beta_i^2] = \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\xi v k}^2$ by applying Isserlis' theorem, with $\omega_{\eta\eta k} := E[\eta_i^2 \mid k(i) = k]$, $\omega_{\zeta\eta k} := E[\zeta_i\eta_i \mid k(i) = k]$, and $\omega_{\zeta\zeta k} := E[\zeta_i^2 \mid k(i) = k]$, we obtain:

$$\begin{aligned}
&\omega_{\eta\eta k} = \sigma_{vv}^2, \\
&\omega_{\zeta\eta k} = \pi_k\sigma_{\xi v k} + \sigma_{vv}\beta + \sigma_{\varepsilon v}, \text{ and} \\
&\omega_{\zeta\zeta k} = \pi_k^2\sigma_{\xi\xi} + 2\pi_k\beta\sigma_{\xi v k} + 2\pi_k\sigma_{\varepsilon\xi} + \sigma_{\varepsilon\varepsilon} + \sigma_{\xi v k}^2 + \sigma_{vv}\beta^2 + \sigma_{vv}\sigma_{\xi\xi} + 2\sigma_{\varepsilon v}\beta.
\end{aligned} \tag{1.25}$$

In this model, the local average treatment effect (LATE) of judge k relative to the base judge 0 is:

$$LATE_k = \frac{\pi_{Yk}}{\pi_k} = \beta + \frac{\sigma_{\xi v k}}{\pi_k}. \tag{1.26}$$

Derivation for binary setup without covariates.

The reduced-form residuals are given by:

$$\eta_i \mid v_i = \begin{cases} 1 - \pi_k & if \quad v_i \leq \pi_k \\ -\pi_k & if \quad v_i > \pi_k \end{cases}, \text{ and } \quad \zeta_i = \pi_{k(i)}\beta_i - \pi_{Yk(i)} + \eta_i\beta_i + \varepsilon_i.$$

Imposing $E[\zeta_i] = 0$, $\pi_{Yk(i)} = \pi_{k(i)}\beta + E[\eta_i\beta_i]$, where $E[\eta_i\beta_i] = -(1-s)(2p-1)\sigma_{\xi v k}$. Hence,

$$\pi_{Yk} = \pi_k\beta - (1-s)(2p-1)\sigma_{\xi v k}.$$

Due to the judge setup, the estimand is:

$$\frac{\sum_k \pi_{Yk}\pi_k}{\sum_k \pi_k^2} = \frac{\sum_k \left(\pi_k\beta - (1-s)\,(2p-1)\,\sigma_{\xi vk}\right)\pi_k}{\sum_k \pi_k^2} = \beta$$

because $\sum_k \sigma_{\xi vk}\pi_k = 0$ by construction.

Derivation for binary setup with covariates.

Consider the structural model:

$$\begin{aligned} Y_i(x) &= x(\beta + \xi_i) + w'\gamma + \varepsilon_i, \text{ and} \\ X_i(z) &= I\left\{z'\pi + w'\gamma - v_i \geq 0\right\}. \end{aligned}$$

Let $\mathcal{N}_t$ denote the set of observations in state t. Then, using the G that corresponds to UJIVE,

$$\begin{aligned} \sum_{i\in\mathcal{N}_t}\sum_{j\in\mathcal{N}_t\setminus i} G_{ij}R_{Yi}R_j &= \sum_{i\in\mathcal{N}_t}\sum_{j\in\mathcal{N}_t\setminus i} G_{ij}\left(\pi_{Yk(i)} + \gamma_{t(i)}\right)\left(\pi_{k(j)} + \gamma_{t(j)}\right) \\ &= \sum_{i\in\mathcal{N}_t}\sum_{j\in\mathcal{N}_t\setminus i} G_{ij}\left(\pi_{Yk(i)}\pi_{k(j)} + \gamma_{t(i)}\pi_{k(j)} + \pi_{Yk(i)}\gamma_{t(j)} + \gamma_{t(i)}\gamma_{t(j)}\right) \\ &= \frac{1}{1-1/5}\sum_{k\in\{0,t\}} 5\times 4\times\frac{1}{5}\left(\pi_{Yk}\pi_k + \gamma_t\pi_k + \pi_{Yk}\gamma_t + \gamma_t^2\right) \\ &\quad - \frac{1}{1-1/10}\sum_{i\in\mathcal{N}_t}\sum_{j\in\mathcal{N}_t\setminus i}\frac{1}{10}\left(\pi_{Yk(i)}\pi_{k(j)} + \gamma_t\pi_{k(j)} + \pi_{Yk(i)}\gamma_t + \gamma_t^2\right) \\ &= \sum_{k\in\{0,t\}} 5\left(\pi_{Yk}\pi_k + \gamma_t\pi_k + \pi_{Yk}\gamma_t + \gamma_t^2\right) - \frac{1}{9}\sum_{k\in\{0,t\}} 5\times 4\left(\pi_{Yk}\pi_k + \gamma_{Yt}\pi_k + \pi_{Yk}\gamma_{Xt} + \gamma_t^2\right) \\ &\quad - \frac{1}{9}5\times 5\left(\pi_{Yt}\pi_0 + \gamma_t\pi_0 + \pi_{Yt}\gamma_t + \gamma_t^2\right) - \frac{1}{9}5\times 5\left(\pi_{Y0}\pi_t + \gamma_t\pi_t + \pi_{Y0}\gamma_t + \gamma_t^2\right) \\ &= 5\left(\frac{5}{9}\right)\left(\pi_{Y0}\pi_0 + \pi_{Yt}\pi_t - \pi_{Yt}\pi_0 - \pi_{Y0}\pi_t\right). \end{aligned}$$

Using the result that $\pi_{Yk} = \pi_k\beta - (1-s)(2p-1)\sigma_{\xi vk}$,

$$\sum_{i\in\mathcal{N}_t}\sum_{j\in\mathcal{N}_t\setminus i} G_{ij}R_{Yi}R_j = 5\left(\frac{5}{9}\right)(\pi_{Y0}\pi_0 + \pi_{Yt}\pi_t - \pi_{Yt}\pi_0 - \pi_{Y0}\pi_t) = \frac{25}{9}\pi_{Yt}\pi_t.$$

Analogously, $\sum_{i\in\mathcal{N}_t}\sum_{j\in\mathcal{N}_t\setminus i} G_{ij}R_iR_j = \frac{25}{9}\pi_t^2$. Hence, as long as $\sum_t \sigma_{\xi vt}\pi_t = 0$, which is the case for the construction in the main text, we still recover β as our estimand:

$$\begin{aligned}\frac{\sum_i\sum_{j\neq i} G_{ij}R_{Yi}R_j}{\sum_i\sum_{j\neq i} G_{ij}R_iR_j} &= \frac{\sum_t \pi_{Yt}\pi_t}{\sum_t \pi_t^2} = \frac{\sum_t (\pi_t\beta - (1-s)(2p-1)\sigma_{\xi vt})\pi_t}{\sum_t \pi_t^2} \\ &= \beta - \frac{\sum_t (1-s)(2p-1)\sigma_{\xi vt}\pi_t}{\sum_t \pi_t^2} = \beta,\end{aligned}$$

regardless of γ_t.

Derivations for Variance Estimands

Proof of Equation (1.19).

$$
\begin{aligned}
E\left[\hat{\Psi}_{MO}\right] &= E\left[\sum_i \left(\sum_{j\neq i} P_{ij}\left(R_j+\eta_j\right)\right)^2 \left(R_{\Delta i}+\nu_i\right)^2 + \sum_i \sum_{j\neq i} P_{ij}^2 \left(R_i+\eta_i\right)\left(R_{\Delta i}+\nu_i\right)\left(R_j+\eta_j\right)\left(R_{\Delta j}+\nu_j\right)\right] \\
&= E\left[\sum_i \left(\left(\sum_{j\neq i} P_{ij} R_j\right)^2 + \left(\sum_{j\neq i} P_{ij}\eta_j\right)^2\right)\left(R_{\Delta i}+\nu_i\right)^2\right] \\
&\quad + E\left[\sum_i \sum_{j\neq i} P_{ij}^2 \left(R_i R_{\Delta i} + \eta_i R_{\Delta i} + R_i\nu_i + \eta_i\nu_i\right)\left(R_j R_{\Delta j} + \eta_j R_{\Delta j} + R_j\nu_j + \eta_j\nu_j\right)\right] \\
&= \sum_i M_{ii}^2 R_i^2 \left(R_{\Delta i}^2 + E\left[\nu_i^2\right]\right) + \sum_i R_{\Delta i}^2 E\left[\left(\sum_{j\neq i} P_{ij}\eta_j\right)^2\right] + \sum_i E\left[\nu_i^2\right] E\left[\left(\sum_{j\neq i} P_{ij}\eta_j\right)^2\right] \\
&\quad + \sum_i \sum_{j\neq i} P_{ij}^2 \left(R_i R_{\Delta i} + E\left[\eta_i\nu_i\right]\right)\left(R_j R_{\Delta j} + E\left[\eta_j\nu_j\right]\right) \\
&= \sum_i M_{ii}^2 R_i^2 \left(R_{\Delta i}^2 + E\left[\nu_i^2\right]\right) + \sum_i \sum_{j\neq i} P_{ij}^2 E\left[\eta_j^2\left(R_{\Delta i}^2 + \nu_i^2\right)\right] \\
&\quad + \sum_i \sum_{j\neq i} P_{ij}^2 \left(R_i R_{\Delta i} + E\left[\eta_i\nu_i\right]\right)\left(R_j R_{\Delta j} + E\left[\eta_j\nu_j\right]\right) \\
&= \sum_i M_{ii}^2 R_i^2 R_{\Delta i}^2 + \sum_i M_{ii}^2 R_i^2 E\left[\nu_i^2\right] + \sum_i \sum_{j\neq i} P_{ij}^2 E\left[\nu_i^2\right] E\left[\eta_j^2\right] + \sum_i \sum_{j\neq i} P_{ij}^2 R_{\Delta i}^2 E\left[\eta_j^2\right] \\
&\quad + \sum_i \sum_{j\neq i} P_{ij}^2 \left(R_i R_{\Delta i} R_j R_{\Delta j} + E\left[\eta_i\nu_i\right] R_j R_{\Delta j} + R_i R_{\Delta i} E\left[\eta_j\nu_j\right] + E\left[\eta_i\nu_i\right] E\left[\eta_j\nu_j\right]\right)
\end{aligned}
$$

□

Chapter 2

Asymptotic Theory for Two-Way Clustering[1]

2.1 Introduction

Clustering standard errors on multiple dimensions is common and attractive in applied econometrics because it allows observations to be dependent whenever they share a cluster on any dimension. Though more broadly applicable, a common instance of two-way clustering is in linear regressions, where a researcher wants to do inference on the coefficient of interest when the residual is two-way clustered. The variance estimator proposed by Cameron et al. (2011) (henceforth CGM) has thus been widely applied to contexts with such two-way dependence.[2] For instance, Nunn and Wantchekon (2011) clustered on ethnic group and district when studying the effect of slave trade on trust; Michalopoulos and Papaioannou (2013) clustered on country and ethnolinguistic family when studying the effect of pre-colonial institutions on development; Jackson (2018) clustered on teacher and student when studying the effect of the teacher on students' skill; Neumark et al. (2019) clustered on resume and

[1] This chapter is published in the *Journal of Econometrics* (see Yap (2025)), and presented at the Econometric Society European Meeting in 2023 (Barcelona).

[2] CGM has 3886 citations on Google Scholar at the time of writing.

job ad when studying the effect of age on getting a call-back. The existing justification for the asymptotic validity of the CGM estimator and other inference procedures in two-way clustering (e.g., MacKinnon et al. (2021); Davezies et al. (2021); Menzel (2021)) relies on separate exchangeability, which implies homogeneity of clusters, a restriction that is not required in one-way clustering. This paper provides sufficient general conditions for valid inference in two-way clustering by proving that, even with cluster heterogeneity, a central limit theorem holds, and the CGM variance estimator is consistent.

An environment with two-way clustering permits dependence whenever observations share at least one cluster. To fix ideas, consider Jackson (2018): observations of the same student or of the same teacher are plausibly correlated, but two observations of different students and different teachers are assumed to be independent.[3] The CGM variance estimator accommodates such dependence, and a subsequent literature provided a theoretical basis for its validity: MacKinnon et al. (2021) obtained sufficient conditions for validity of the CGM estimator in regression models; Davezies et al. (2021) obtained analogous results for empirical processes. Menzel (2021) also showed the validity of a bootstrap procedure for two-way clustering that is robust to asymptotic non-normalities.[4] The theoretical basis for inference thus far relies on separate exchangeability, the assumption that random variables are exchangeable on either clustering dimension, though not necessarily both.

However, separate exchangeability implies identical marginal distributions. Separate exchangeability in the student-teacher example thus implies the random variables for all students must be drawn from the same distribution, including students of different cohorts over time. As Wooldridge (2010, p. 146) notes in the discussion of pooled data in his graduate

[3]This setting permits more general dependence structures than one-way clustering. If there is one-way clustering by student, then two observations from different students are automatically independent. In two-way clustering, two observations from different students are not necessarily independent because they may share the same teacher.

[4]Menzel (2021) pointed out that a purely interactive data generating processes unique to two-way dependence has an asymptotic distribution that is not normal. Section 2.2 will consider this process and show how the assumptions of this paper rule it out.

textbook, distributions of variables tend to change over time, so the identical distribution assumption is not usually valid. In other examples, separate exchangeability implies that countries (Michalopoulos and Papaioannou, 2013) and jobs (Neumark et al., 2019) are identically distributed. Applied researchers surely would want size to be controlled in such heterogeneous environments, but the existing theories that rely on separate exchangeability do not imply this result. Further, in linear regressions with regressor X_i and residual u_i, asymptotic theory is applied to $X_i u_i$. Separate exchangeability of the product implies that the regressors must also be separately exchangeable, which is not plausible when the regressors include a time trend, say.

In contrast, existing asymptotic theory on one-way clustering (e.g., Hansen and Lee (2019); Djogbenou et al. (2019)) allows the distribution of the random variable to be heterogeneous over clusters. Since the only available conditions for the validity of two-way clustering require separate exchangeability, the literature lacks conditions for two-way clustering that generalize one-way clustering and permit heterogeneity over clusters. This paper fills the gap, and thus justifies two-way clustering as a more robust version of one-way clustering.

Example 2.1. *To illustrate separate exchangeability, consider an additive random effects model. Individual i who belongs to cluster $g(i)$ on the G dimension and cluster $h(i)$ on the H dimension is characterized by a random variable W_i generated from $W_i = \alpha_{g(i)} + \gamma_{h(i)} + \varepsilon_i$, where cluster-specific $\alpha_1, \ldots, \alpha_g, \ldots, \alpha_G, \gamma_1, \ldots, \gamma_h, \ldots, \gamma_H$ and individual-specific $\varepsilon_1, \ldots, \varepsilon_i, \ldots, \varepsilon_n$ are mutually independent. If we assume separate exchangeability, then α_g, γ_h, and ε_i are iid.*[5] *In contrast, under one-way cluster asymptotics, the cluster-specific error α_g need not be identically distributed. The general conditions provided in this paper permit valid inference even when $\alpha_g, \gamma_h, \varepsilon_i$ are not identically distributed in this model.*

[5]To see this, for individuals i and j where $g(i) \neq g(j)$, $h(i) = h(j) = h$, separate exchangeability implies $\alpha_{g(i)} + \gamma_h + \varepsilon_i \overset{d}{=} \alpha_{g(j)} + \gamma_h + \varepsilon_j$. Since α_g, γ_h and ε_i are independent, $\varepsilon_i \overset{d}{=} \varepsilon_j$ and $\alpha_g \overset{d}{=} \alpha_{g'}$.

The main result is a central limit theorem for two-way clustering with heterogeneous cluster sizes and distributions. This result is proven using Stein's method. It adapts the strategy from Ross (2011) Theorem 3.6: I first derive an upper bound on the distance between the distribution of a pivotal statistic and the standard normal, then show that this distance converges to zero asymptotically. This proof strategy hence yields intermediate results on non-asymptotic Berry-Esseen type bounds that provide worst-case bounds on the quality of approximation between the pivotal statistic and the standard normal, which may be of independent interest. I apply the theorem to a simple setting of a linear regression, but it is more broadly applicable to many other econometric procedures that exhibit a similar clustering structure.

This paper contributes to the literature on multi-way clustering and Stein's method. This paper differs from the existing literature on multi-way clustering (e.g., MacKinnon et al. (2021); Davezies et al. (2021); Menzel (2021); Chiang and Sasaki (2023); Chiang et al. (2024)) in that it does not rely on separate exchangeability. Stein's method has been applied to other contexts such as two-way fixed effects (Verdier, 2020), spillover effects (e.g., Chin (2018), Leung (2022) and Braun and Verdier (2023)), and network formation (e.g., Chandrasekhar and Jackson (2016)). Unlike the aforementioned papers, this paper speaks directly to multi-way clustering, and it makes a modification to the proof of Ross (2011) Theorem 3.6 to obtain the result instead of applying the theorem directly.

2.2 Setting and Main Result

2.2.1 Setup

Consider a setup with two-way clustering on dimensions G and H for random vectors $\{W_i\}_{i=1}^n$, where $W_i := (W_{i1}, W_{i2}, \cdots, W_{iK})' \in \mathbb{R}^K$ and $i = 1, \ldots, n$ is the unit of observation. For example, G could denote states and H could denote industries. Clustering in more than

two dimensions is possible, and derivations are entirely analogous. This section establishes a central limit theorem (CLT) for $\sum_i W_i$, as $n \to \infty$. Here and in the following, sums are over (subsets of) $\{1, 2, \ldots, n\}$. For $C \in \{G, H\}$, let $\mathcal{N}_c^C$ denote the set of observations in cluster c on dimension C — this setup partitions the sample on the C dimension.

Let $g(i)$ and $h(i)$ denote the cluster that observation i belongs to on the G and H dimensions respectively. These cluster identities are nonstochastic and observed. Let $N_c^C := |\mathcal{N}_c^C|$ denote the size of cluster c on dimension $C \in \{G, H\}$ and $N_{gh} := |\mathcal{N}_g^G \cap \mathcal{N}_h^H|$. These cluster sizes are allowed to be heterogeneous in a way that will be formalized in the assumptions below. W_i is assumed to be independent of the joint distribution of $\{W_j\}$ for $j \notin \mathcal{N}_{g(i)}^G \cup \mathcal{N}_{h(i)}^H =: \mathcal{N}_i$, i.e., when i and j do not share a cluster on either dimension. Hence, $\mathcal{N}_i$ is the set of observations that are arbitrarily dependent with i. This environment is stated as Assumption 2.1.

Assumption 2.1. *With* $\mathcal{N}_i = \mathcal{N}_{g(i)}^G \cup \mathcal{N}_{h(i)}^H$,

(a) $W_i \perp\!\!\!\perp \{W_j\}_{j \notin \mathcal{N}_i}$ *for all* i.

(b) For observations i, j *and* $k \in \mathcal{N}_i, l \in \mathcal{N}_j$ *and all nonstochastic* $\mu \in \mathbb{R}^K$, *if* $j, l \notin (\mathcal{N}_i \cup \mathcal{N}_k)$, *then* $Cov(\mu' W_i W_k' \mu, \mu' W_j W_l' \mu) = 0$.

While the dependence structure is implicitly described in the setup of many clustering papers (e.g., Hansen and Lee (2019); Menzel (2021)), Assumption 2.1 makes the dependence structure explicit. Assumption 2.1(a) is a dissociation assumption similar to Definition 3.5 of Ross (2011) required to apply Stein's method. Assumption 2.1(b) is required because, for a scalar W_i, a crucial step of the proof requires $E[W_i W_j W_k W_l] = E[W_i W_k] E[W_j W_l]$ when j, l do not share any cluster with i, k. Even when $W_i \perp\!\!\!\perp (W_j, W_l)$ and $W_k \perp\!\!\!\perp (W_j, W_l)$, we cannot conclude that $E[W_i W_j W_k W_l] = E[W_i W_k] E[W_j W_l]$ in general, because independence of marginal distributions does not imply independence of the joint distribution. Assumption 2.1(b) hence makes an assumption on the joint distribution. It can alternatively be stated as $(W_i, W_k) \perp\!\!\!\perp (W_j, W_l)$, which is stronger but more interpretable than the zero-covariance

assumption. I further discuss the relationship between Assumption 2.1 and the existing literature in Section 2.2.3.

Assumption 2.1 is agnostic about the dependence structure between W_i and W_j when i and j share at least one cluster. It also allows the data generating process to be arbitrarily heterogeneous across different clusters, mimicking the heterogeneity permitted in one-way clustering (e.g., Hansen and Lee (2019); Djogbenou et al. (2019)). Since one-way clustering is a special case of two-way clustering where the H cluster consists of single observations, the result here generalizes the existing results in one-way clustering. In contrast, the existing literature on two-way clustering assumes separate exchangeability that additionally imposes identical distribution over clusters, so it does not generalize the results on one-way clustering.

For positive definite matrix Q, let $\lambda_{\min}(Q)$ denote the smallest eigenvalue of Q. Then, let $Q_n := Var\left(\sum_i W_i\right)$ denote the variance of the sum and $\lambda_n := \lambda_{\min}(Q_n)$ denote its smallest eigenvalue. For example, when $K = 1$, W_i is a scalar and $\lambda_n = Q_n = Var(\sum_i W_i)$. K_0 is used throughout the paper to denote an arbitrary constant.

Assumption 2.2. *For $C \in \{G, H\}$, and $k \in \{1, 2, \cdots, K\}$, there exists $K_0 < \infty$ such that:*

(a) $E[W_{ik}^4] \leq K_0$ *for all* i.

(b) $\frac{1}{\lambda_n} \max_c (N_c^C)^2 \to 0$.

(c) $\frac{1}{\lambda_n} \sum_c (N_c^C)^2 \leq K_0$.

Since the objective of this paper is to prove a CLT, Assumption 2.2 imposes restrictions that rule out data generating processes that are asymptotically non-Gaussian. One such example is explained later in Remark 2.1. Nonetheless, as reflected in Table 1 of Chiang and Sasaki (2023), such a non-Gaussian regime is an exception rather than the norm when considering a generic separately exchangeable process.

Assumption 2.2(a) requires the fourth moment to be bounded, which is stronger than the moment condition in one-way clustering (e.g., Equation (7) of Hansen and Lee (2019)

and Assumption 1 of Djogbenou et al. (2019)). The proof in one-way clustering usually verifies a Lindeberg condition then applies the Lindeberg CLT because blocks of observations are independent of each other. With two-way dependence, we no longer have independent blocks because each cluster can have observations that are dependent on observations from a different cluster when these observations share a cluster on a different dimension. Hence, a different proof strategy is required. The proof in this paper uses Stein's method, which requires stronger moment restrictions, but provides a non-asymptotic bound on the approximation error — details are in Subsection 2.2.4. By using this strategy, a bounded fourth moment is required.

Assumption 2.2(b) requires the size of the largest cluster to be small relative to the total variance. This condition mimics the sparsity condition in the networks literature (e.g., Graham (2020)). Intuitively, this condition is required so that the removal of a cluster does not change the variance substantively. This assumption allows the ratio of any two cluster sizes to diverge to infinity. It is identical to Equation (12) of Hansen and Lee (2019) and Assumption 3 of Djogbenou et al. (2019) for one-way clustering. Assumption 2.2(b) also rules out having components that are perfectly correlated: if the components of the vector were perfectly correlated (i.e., $\mu' W_i = 0$ for some $\mu \neq (0, \ldots, 0)'$), then $\lambda_n = 0$. If cluster sizes are uniformly bounded, and $\lambda_n \to \infty$, then Assumption 2.2(b) is satisfied.[6]

Assumption 2.2(c) is a summability condition that requires λ_n not to be too small, and requires λ_n to be the same order as $\sum_c (N_c^C)^2$, i.e., $\lambda_n \asymp \sum_c (N_c^C)^2$, $C \in \{G, H\}$.[7] With strictly positive covariance within clusters, $\lambda_n \asymp \sum_c (N_c^C)^2$ is satisfied. However, if the researcher were conservative and clustered on C when the data is indeed iid, then $\lambda_n \asymp n$, which then requires $\sum_c (N_c^C)^2 \asymp n$ for the condition to hold. The assumption that

[6] Assumption 2.2(b) is hence a more general version of sparsity than having the size of the dependency neighborhood (i.e., the number of observations plausibly correlated with some observation i) being bounded above. The conditions are also comparable with Verdier (2020) in the two-way fixed effects literature: when the neighborhood size is bounded, $\lambda_n \asymp n$, which matches his assumption 2(c).

[7] For sequences a_n and b_n, $a_n \asymp b_n$ if and only if there exists $K_0 < \infty$ such that $a_n/b_n, b_n/a_n \in [-K_0, K_0]$ for all elements in the sequence.

$(1/\lambda_n)\sum_c (N_c^C)^2 \leq K_0$ matches Equation (11) of Hansen and Lee (2019) and Assumption 2 of Djogbenou et al. (2019).

In general, the structure of dependence affects λ_n while the structure of clustering affects $\sum_c (N_c^C)^2$. For example, using the common shocks model of Example 1, $\lambda_n \asymp \sum_c (N_c^C)^2$ when the variances of common shocks α_g and γ_h are non-zero, but if the variances of α_g and γ_h are zero, then $\lambda_n \asymp n$. With a balanced clustering structure where $g \in \{1, \cdots, M\}$, $h \in \{1, \cdots, M\}$ and $N_{gh} = 1$, we have $n = M^2$ and $\sum_c (N_c^C)^2 = M^3$. However, if we have one large cluster, say when all observations are the only observation in their H cluster, i.e., $h(i) = i$, and on the G dimension, the first cluster has size $N_1^G = n^{1/4}$, while all other clusters have size 1, then, $\sum_c (N_c^C)^2 \asymp n^{1/2} + (n - n^{1/4}) \asymp n$.

Remark 2.1. *Assumptions 2.2(b) and 2.2(c) rule out the following purely interactive model. For $g \in \{1, \cdots, M\}$, $h \in \{1, \cdots, M\}$ and $N_{gh} = 1$, we observe $W_{gh} = \alpha_g \gamma_h$, where α_g and γ_h are iid with mean zero and variances σ_α^2 and σ_γ^2 respectively, so there are M^2 observations. As pointed out by Menzel (2021) Example 1.7, this model has an asymptotic distribution that is non-normal, with no analog in one-way clustering. To see this, $\sum_{g,h} W_{gh}/M = \left(\sum_g \alpha_g/\sqrt{M}\right)\left(\sum_h \gamma_h/\sqrt{M}\right) \xrightarrow{d} Z_1 Z_2$, where Z_1 and Z_2 are independent standard normal random variables. This limiting distribution is also known as Gaussian chaos. Since $\max_g (N_g^G)^2/\lambda_n = M^2/(M^2\sigma_\alpha^2\sigma_\gamma^2) = 1/(\sigma_\alpha^2\sigma_\gamma^2)$ does not converge to 0, Assumption 2.2(b) fails. Further, $\sum_g (N_g^G)^2/\lambda_n = M^3/(M^2\sigma_\alpha^2\sigma_\gamma^2) = M/\sigma_\alpha^2\sigma_\gamma^2 \to \infty$ violates Assumption 2.2(c).*

Remark 2.2. *Assumptions 2(b) and 2(c) mimic the Lindeberg condition as they divide by the variance of the sum. Nonetheless, if we are willing to make stronger assumptions on variances, we can rewrite the assumptions in terms of primitives. Consider the simple case where W_i is a scalar. If we assume that $E[W_i W_j] \geq c > 0$ for all i and $j \in \mathcal{N}_i$, then Assumption 2(c) is satisfied as $\lambda_n \geq c\left(\sum_g \left(N_g^G\right)^2 + \sum_h \left(N_h^H\right)^2 - \sum_{g,h}\left(N_{(g,h)}^{G\cap H}\right)^2\right) \geq c\sum_g \left(N_g^G\right)^2$ and $\lambda_n \geq c\sum_h (N_h^H)^2$. Then, as long as the largest cluster is small relative to*

$\sum_c \left(N_c^C\right)^2$, *i.e.,* $\max_c (N_c^C)^2 / \sum_c \left(N_c^C\right)^2 \to 0$, *(b) is satisfied. Consequently, a stronger way to state (b) and (c) is that* $\max_c (N_c^C)^2 / \sum_c \left(N_c^C\right)^2 \to 0$ *and* $E[W_i W_j] \geq c > 0$ *for all* i *and* $j \in \mathcal{N}_i$.

2.2.2 Main Result

The main result is that the sum of a sequence of two-way clustered random variables is asymptotically normal. Further, the plug-in variance estimator originally proposed by CGM, $\hat{Q}_n := \sum_i \sum_{j \in \mathcal{N}_i} W_i W_j'$, is consistent. This plug-in expression matches Equation (2.8) of CGM, where W is used here in place of their $\hat{u}$.

Theorem 2.1. *Under Assumptions 2.1 and 2.2,* $Q_n^{-1/2} \sum_i (W_i - E[W_i]) \xrightarrow{d} N(0, I_K)$. *Further, if* $E[W_i] = 0 \ \forall i$, *then* $Q_n^{-1/2} \hat{Q}_n Q_n^{-1/2} \xrightarrow{p} I_K$.

One-way clustering is a special case of this theorem when one dimension is weakly nested within the other: examples include $G = H$ so both dimensions are identical, and clustering by county and state (as counties are nested in states). A sufficient condition for consistent variance estimation is $E[W_i] = 0$, similar to Theorem 3 of Hansen and Lee (2019). This assumption is sufficient in many applications: for example, linear regressions considered in Section 2.3 are identified by requiring the expectation of the residual term to be zero. If $E[W_i] = \mu$ for all i as in Theorem 4 of Hansen and Lee (2019), consistency can be obtained under the same assumptions.[8]

Remark 2.3. *A double array of random vectors, where the random vector* W_{in} *is indexed by* n, *can be accommodated. In this setup, with* $K = 1$ *for simplicity, we can define* $\mathcal{W}_n$ *as the class of distributions of* n *random variables* $\{W_{in}\}_{i=1}^n$ *that satisfy Assumptions 2.1, 2.2(a), 2.2(c), and that for* $C \in \{G, H\}$, *there exists* $K_0 < \infty$ *and* $\epsilon > 0$ *such that* $\frac{1}{\lambda_n} \max_c (N_c^C)^2 \leq K_0 n^{-\epsilon}$ *(which is a modification of Assumption 2.2(b)). Then, for*

[8] Since $(1/n) \sum_i W_i$ consistently estimates μ, the result follows by using $\tilde{W}_i = W_i - \mu$ in place of W_i.

$R_n := Q_n^{-1/2} \sum_i (W_{in} - E[W_{in}])$, d_W *denoting the Wasserstein distance*[9], *and* Z *denoting the standard normal random variable, we have* $\sup_{\{W_{in}\}_{i=1}^n \in \mathcal{W}_n} d_W(R_n, Z) \to 0$ *as* $n \to 0$. *Consequently, normality holds for a double array uniformly over distributions in* $\mathcal{W}_n$. *The proof of such a result is the same as the proof of Theorem 2.1. In the double array, Assumption 2.2(c) rules out a balanced setting where component variances are of order smaller than one: there are* $O(M^3)$ *variance and covariance objects in* λ_n, *so when they are of order* r_M, $\lambda_n = O(M^3 r_M)$ *while* $\sum_c (N_c^C)^2 = M^3$. *Then, any* r_M *that decays at any order of* M *violates Assumption 2.2(c).*[10]

Remark 2.4. *While the CGM variance estimator is valid in this environment without separate exchangeability, we must be more careful with bootstrap methods that were developed under separate exchangeability (e.g., Menzel (2021), MacKinnon et al. (2021)). Bootstrap methods often resample cluster-specific means, such as* $\hat{\alpha}_g = (1/N_g^G) \sum_{i \in \mathcal{N}_g^G} W_i - (1/n) \sum_i W_i$. *Consider a data generating process where, with* $\alpha_g = (1/N_g^G) \sum_{i \in \mathcal{N}_g^G} [W_i] - (1/n) \sum_i E[W_i]$, *odd-numbered* g *clusters have* $\alpha_g = -1$ *and even-numbered* g *clusters have* $\alpha_g = 2$, *and there are twice as many units in odd-numbered clusters as even-numbered clusters. Such a process is not exchangeable. Resampling* $\hat{\alpha}_g$*'s with equal probability results in a positive mean, which invalidates naive bootstrap procedures.*

The following two subsections discuss technicalities on the dependence structure and the proof sketch. A general-interest audience may wish to proceed immediately to Section 2.3.

[9] See details in Section 2.2.4.

[10] These assumptions are primarily used in Lemmas 2.6 and 2.7 of the appendix, so an alternative way to proceed with the proof of normality is to assume their conclusions $\frac{1}{\lambda_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E\left[|W_i| W_j W_k\right] = o(1)$ and $\frac{1}{\lambda_n^4} Var\left(\sum_i \sum_{j \in \mathcal{N}_i} W_i W_j\right) = o(1)$ directly. In the balanced design where the second, third and fourth moments decay at the same rate r_M, $\sum_i \sum_{j,k \in \mathcal{N}_i} E\left[|W_i| W_j W_k\right] = O\left(M^4 r_M\right)$ and $Var\left(\sum_i \sum_{j \in \mathcal{N}_i} W_i W_j\right) = O\left(M^5 r_M\right)$. Then, $\frac{1}{\sigma_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E\left[|W_i| W_j W_k\right] = O\left(M^{-1/2} r_M^{-1/2}\right)$ and $\frac{1}{\sigma_n^4} Var\left(\sum_i \sum_{j \in \mathcal{N}_i} W_i W_j\right) = O\left(M^{-1} r_M^{-1}\right)$. Hence, the conclusions can still hold if these moments decay at a rate slower than M: for instance, if $r_M = M^{-1/2}$, then $O\left(M^{-1} r_M^{-1}\right) = O\left(M^{-1/2}\right) = o(1)$.

2.2.3 Discussion of Dependence Structure

To compare the setup used in Assumption 2.1 to the existing literature, I carefully define a few terms used in Menzel (2021), whose setup uses a dissociated separately exchangeable array. Let Y_{gh} denote an infinite array of observations in cluster g on the G dimension and cluster h on the H dimension. Y_{gh} is a separately exchangeable array if, for any integers $\tilde{G}, \tilde{H}$ and permutations $\pi_1 : \{1, \ldots, \tilde{G}\} \to \{1, \ldots, \tilde{G}\}$ and $\pi_2 : \{1, \ldots, \tilde{H}\} \to \{1, \ldots, \tilde{H}\}$, we have:

$$(Y_{\pi_1(g)\pi_2(h)})_{g,h} \overset{d}{=} (Y_{gh})_{g,h},$$

where $\overset{d}{=}$ denotes equality in distribution.[11] Such an array is dissociated if, for any $G_0, H_0 \geq 1$, $(Y_{gh})_{g=1,h=1}^{g=G_0,h=H_0}$ is independent of $(Y_{gh})_{g>G_0,h>H_0}$. Dissociation is how the existing literature formally incorporates the multi-way clustering structure. Separate exchangeability implies that the cluster indices are not meaningful, and it is stronger than having identical distributions across clusters. This environment is a special case of Assumption 2.1, as the following proposition claims.

Proposition 2.1. *A dissociated separately exchangeable array satisfies Assumption 2.1.*

One formal generalization of separate exchangeability is relative exchangeability in Crane and Towsner (2018), where exchangeability need not hold for the full sample, but only within each stratum (i.e., relative to some structure), such as within cohorts of students. However, such a generalization is insufficient in finite-population settings with two-way clustered sampling. Suppose there is a finite superpopulation of outcomes $\{Y_i\}_{i=1}^n$ that is nonstochastic,

[11] Due to Kallenberg (2005), $\{Y_{gh}\}_{g\geq 1,h\geq 1}$ is separately exchangeable if and only if there exists a representation $Y_{gh} = f(\alpha_g, \gamma_h, \varepsilon_{gh})$, where $(\alpha_g, \gamma_h, \varepsilon_{gh}) \overset{iid}{\sim} U[0,1]$. The setup in this paper does not require $(\alpha_g, \gamma_h, \varepsilon_{gh}) \overset{iid}{\sim} U[0,1]$, which allows some data generating processes ruled out by separate exchangeability. For example, suppose there is some $Y_{gh} = -Y_{gh'}$. These random variables are allowed to be perfectly correlated under Assumption 2.1 since they share a cluster. However, the representation $f(.)$ implies $E[Y_{gh}|\alpha_g] \perp\!\!\!\perp E[Y_{gh'}|\alpha_g]$, so no such representation exists.

and two-way clustered sampling by ethnic groups and district (e.g., Nunn and Wantchekon (2011)): a subset of districts are independently sampled, a subset of ethnic groups are independently sampled, and units are sampled from the intersections of districts and ethnic groups that are both sampled. We are interested in the mean of Y in the finite superpopulation. With R_i denoting the indicator for whether individual i is sampled and hence observed, the observed random variable is $W_i = R_i Y_i$. Even though R_i is separately exchangeable, $R_i Y_i$ is neither separately exchangeable nor relatively exchangeable due to conditioning on $\{Y_i\}_{i=1}^n$, but Assumption 2.1 is still satisfied.

2.2.4 Proof Sketch

The proof of Theorem 2.1 proceeds by first proving a CLT for a scalar random variable, then applying the Cramer-Wold device to obtain the multivariate CLT. The scalar CLT is proven using Stein's method. I adapt the proof strategy from Ross (2011) Theorem 3.6 to obtain an upper bound on the Wasserstein distance between a pivotal statistic and the standard normal random variable. By exploiting the two-way clustering structure, the upper bound on the distance can be shown to converge to zero. All details are in Appendix 2.A.

For ease of exposition, consider a simpler environment where $K = 1$, and $E[W_i] = 0$. Let $\sigma_n^2 := Q_n$, $R = \sum_i W_i/\sigma_n$, and $Z \sim N(0,1)$. Lemma 2.4 in Appendix 2.A provides an explicit bound on the Wasserstein distance between R and Z. With $d_W(.)$ denoting the Wasserstein distance, and $d_K(.)$ denoting the Kolmogorov distance, Proposition 1.2 from Ross (2011) implies that $d_K(R,Z) \leq (2/\pi)^{1/4}\sqrt{d_W(R,Z)}$.[12] The Kolmogorov distance is the maximal distance between two CDF's, so it is informative of the maximum distance between

[12]For completeness, I define both distance metrics using the notation in Ross (2011). For two probability measures μ and ν, and family of test functions $\mathcal{H}$, distances are defined as:

$$d_{\mathcal{H}}(\mu,\nu) = \sup_{h\in\mathcal{H}} \left| \int h(x)d\mu(x) - \int h(x)d\nu(x) \right|.$$

As special cases, the Kolmogorov distance uses $\mathcal{H} = \{1[\cdot \leq x] : x \in \mathbb{R}\}$ and the Wasserstein distance uses $\mathcal{H} = \{h : \mathbb{R} \to \mathbb{R} : |h(x) - h(y)| \leq |x - y|\}$.

the distribution of the pivotal statistic and the standard normal. If $d_W(R, Z) \to 0$, then $d_K(R, Z) \to 0$, so the statistic R is asymptotically normal. By using Assumption 2.1 to adapt the proof of Theorem 3.6 in Ross (2011),

$$d_W(R, Z) \leq \frac{1}{\sigma_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E[|W_i| W_j W_k] + \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2} \sqrt{Var\left(\sum_i \sum_{j \in \mathcal{N}_i} W_i W_j \right)}. \tag{2.1}$$

This inequality is informative of the quality of the normal approximation. This bound on the Wasserstein distance (and hence the Kolmogorov distance) is non-asymptotic, and of the Berry-Esseen type, thereby giving a worst-case bound on the distance between the pivotal statistic and the standard normal. Ross (2011) Theorem 3.6 is a corollary of (2.1): the term with the third moment is immediate, while the term with the fourth moment results from the last line of their proof.

At this point, my proof departs from the proofs in the existing statistical literature that employ Stein's method (e.g., Chen and Shao (2004); Janisch and Lehéricy (2024)). Let $N_i := |\mathcal{N}_i|$. Hölder's inequality is employed on objects such as $\sum_i \sum_{j,k \in \mathcal{N}_i} E[|W_i| W_j W_k]$. The existing literature uses the L^1 norm of moments $E[|W_i|^3]$ and the L^∞ norm of N_i, resulting in $(\max_m N_m)^2 \sum_i E[|W_i|^3]$. In contrast, my proof uses the L^∞ norm of $E[|W_i|^3]$ and the L^1 norm of N_i, resulting in $\max_m E[|W_m|^3] \sum_i N_i^2$. Hence,

$$\frac{1}{\sigma_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E[|W_i| W_j W_k] \leq \frac{1}{\sigma_n^3} \max_m E[|W_m|^3] \sum_i N_i^2.$$

Since $\max_m E[|W_m|^3]$ is bounded by Assumption 2.2(a), it suffices to show that $\sum_i N_i^2/\sigma_n^3 \to 0$. Due to Assumption 2.1(a), $N_i \leq N_{g(i)}^G + N_{h(i)}^H$, so

$$\begin{aligned}\frac{1}{\sigma_n^3}\sum_i N_i^2 \leq \frac{1}{\sigma_n^3}\sum_i (N_{g(i)}^G + N_{h(i)}^H)^2 &\leq \frac{1}{\sigma_n^3}\max_{g,h}(N_g^G + N_h^H)\sum_i (N_{g(i)} + N_{h(i)}) \\ &\leq \left[\frac{1}{\sigma_n}\max_{g,h}(N_g^G + N_h^H)\right]\frac{1}{\sigma_n^2}\left(\sum_g (N_g^G)^2 + \sum_h (N_h^H)^2\right).\end{aligned}$$

Since $\lambda_n = \sigma_n^2$ when $K = 1$, $\max_{g,h}(N_g^G + N_h^H)/\sigma_n \to 0$ by Assumption 2.2(b) and the final term $\left(\sum_g (N_g^G)^2 + \sum_h (N_h^H)^2\right)/\sigma_n^2$ is bounded by Assumption 2.2(c). Hence, the term is $o(1)$.

A similar argument is made for the fourth moment that features in $Var\left(\sum_i \sum_{j\in\mathcal{N}_i} W_i W_j\right)$. To complete the proof for variance estimation, observe that since the fourth moments exist, the consistency of the plug-in variance estimator can be proven by using Chebyshev's inequality and the existing intermediate results.

Remark 2.5. *By modifying the proof of Theorem 3.6 in Ross (2011), the conditions in this paper permit some forms of heterogeneity in cluster sizes that Theorem 3.6 of Ross (2011) does not. The following is one such example. All observations are the only observation in their H cluster, i.e., $h(i) = i$. On the G dimension, the first cluster has size $N_1^G = n^{1/4}$, while all other clusters have size 1. Then, with positive correlation for units within each cluster such that $\lambda_n \asymp \sum_c (N_c^C)^2$, we have $\lambda_n \asymp n^{1/2} + (n - n^{1/4}) \asymp n$ and $(N_1^G)^2/\lambda_n \asymp n^{1/2}/n = o(1)$, so the conditions of Theorem 2.1 are satisfied. However, Theorem 3.6 of Ross (2011) bounds the Wasserstein distance by $\left(N_1^2/\lambda_n^{3/2}\right)\sum_i E|W_i|^3$ and a term that involves the fourth moment. We have $N_1^2/\lambda_n^{3/2}\sum_i E|W_i|^3 \asymp n^{-1}\sum_i E|W_i|^3 \neq o(1)$, so we may not obtain convergence. This example similarly rules out using results from Janisch and Lehéricy (2024) directly.*

Remark 2.6. *There are several early papers in probability theory that deliver similar results, but are insufficient for Theorem 2.1. For instance, Theorem 2 of Janson (1988) is a central*

limit theorem that uses the condition (with $m = 3$):

$$\left(\frac{n}{\max_i N_i}\right)^{1/3} \frac{(\max_i N_i)\max_i |W_i|}{\sigma_n} = \left(\frac{n}{\sigma_n^3}\left(\max_i N_i\right)^2\right)^{1/3} \max_i |W_i| \to 0.$$

In this proof sketch, I have shown that $\sum_i N_i^2/\sigma_n^3 \to 0$, but $n(\max_i N_i)^2/\sigma_n^3 \geq \sum_i N_i^2/\sigma_n^3$, so the Janson (1988) condition need not hold in this environment.

2.3 Theory for Least Squares Regression

This section applies Theorem 2.1 to linear regressions, showing that using the normal approximation with the CGM variance estimator is valid. Consider a linear model where the scalar outcome Y_i is generated by:

$$Y_i = X_i'\beta + u_i.$$

with $X_i \in \mathbb{R}^K$. We are interested in estimating β. Suppose $E[X_i u_i] = 0$ for all i, and (X_i', u_i) is allowed to be two-way clustered. The standard OLS estimator is:

$$\hat{\beta} = \left(\sum_i X_i X_i'\right)^{-1}\left(\sum_i X_i Y_i\right) = \beta + \left(\sum_i X_i X_i'\right)^{-1}\left(\sum_i X_i u_i\right).$$

This object is assumed to be well-defined in that $\sum_i X_i X_i'$ is invertible. Define $S_n := \sum_i E[X_i X_i']$ and $Q_n := Var\left(\sum_i X_i u_i\right)$, and denote their sample analogs as $\hat{S}_n = \sum_i X_i X_i'$ and $\hat{Q}_n := \sum_i \sum_{j \in \mathcal{N}_i} \hat{u}_i \hat{u}_j X_i X_j'$. Let the smallest eigenvalue of Q_n be $\lambda_n := \lambda_{\min}(Q_n)$. The asymptotic variance of $\hat{\beta}$ and its sample analog are $V(\hat{\beta}) := S_n^{-1} Q_n S_n^{-1}$ and $\hat{V}(\hat{\beta}) := \hat{S}_n^{-1}\hat{Q}_n\hat{S}_n^{-1}$ respectively.

Assumption 2.3 provides sufficient conditions for the estimator $\hat{\beta}$ to be asymptotically normal and for the CGM variance estimator to be consistent. The conditions mimic Assump-

tion 2.2 so that Theorem 2.1 is applicable to the random vector $X_i u_i$. The new condition is a weak regularity condition that $\lambda_{\min}(S_n/n) \geq K_1 > 0$, mimicking the rank condition in OLS.

Assumption 2.3. *For $C \in \{G, H\}$, and $k \in \{1, 2, \cdots, K\}$, there exists $K_0 < \infty$ and $K_1 > 0$ such that:*

(a) $E[u_i^4 | X_i] \leq K_0$, $E[X_{ik}^4] \leq K_0$, $E[X_i u_i] = 0$ *for all* i.

(b) $\frac{1}{\lambda_n} \max_c (N_c^C)^2 \to 0$.

(c) $\frac{1}{\lambda_n} \sum_c (N_c^C)^2 \leq K_0$.

(d) $(X_i', u_i)' \perp\!\!\!\perp \{(X_j', u_j)'\}_{j \notin \mathcal{N}_i}$. *For observations i, j and $k \in \mathcal{N}_i, l \in \mathcal{N}_j$ and all non-stochastic $\mu \in \mathbb{R}^K$, if $j, l \notin (\mathcal{N}_i \cup \mathcal{N}_k)$, then $(X_i', u_i, X_k', u_k)' \perp\!\!\!\perp (X_j', u_j, X_l', u_l)'$.*

(e) $\lambda_{\min}\left(\frac{1}{n} S_n\right) \geq K_1$.

Proposition 2.2. *Under Assumption 2.3, $Q_n^{-1/2} S_n (\hat{\beta} - \beta) \xrightarrow{d} N(0, I_K)$, and $[S_n^{-1} Q_n S_n^{-1}]^{-1} [\hat{S}_n^{-1} \hat{Q}_n \hat{S}_n^{-1}] \xrightarrow{p} I_K$.*

Proposition 2.2 is useful for performing F tests on a subvector of β. The proof of Proposition 2.2 proceeds by applying Theorem 2.1 to $\sum_i X_i u_i$, then showing that $S_n^{-1} \hat{S}_n \xrightarrow{p} I_K$, which uses the rank condition of Assumption 2.3(e). It then remains to show that the remainder terms are asymptotically negligible.

The practitioner's takeaway from Proposition 2.2 is that the existing CGM variance estimator can be used for valid inference with two-way clustering. The result provides the formal theoretical guarantee for using the estimator, under conditions that permit heterogeneity across clusters.

Besides the application mentioned, Theorem 2.1 also has implications on the conditions required for valid inference when the random variable is two-way clustered in many other econometric models, including design-based settings and instrumental variables models. This

theory is especially relevant for design-based settings where the researcher conditions on potential outcomes, so the random variable cannot be separately exchangeable by construction — see Xu and Yap (2024), for instance. Inference for estimators based on moment conditions can be done by straightforward application of Theorem 2.1 as in linear regression. Practically, this paper has shown that the popular CGM estimator is robust in an environment without separate exchangeability, but practitioners should exercise caution when applying bootstrap methods to environments that are not separately exchangeable. While the results are presented for two-way clustering, they can be easily extended to clustering on three or more dimensions.

Appendix

2.A Proof of Theorem 2.1

The proof strategy is as follows. I first prove Lemma 2.1, which is a central limit theorem (CLT) for scalars. The proof of Lemma 2.1 relies on Lemmas 2.2 to 2.7. Lemmas 2 to 4 derive an upper bound on the Wasserstein distance between a pivotal statistic and standard normal Z. Lemmas 5 to 7 then show that the derived upper bound is $o(1)$. With Lemma 2.1, the multivariate CLT of Theorem 2.1 is obtained by using the Cramer-Wold device. The remainder of the proof proceeds in the following order: (i) introduce definitions and notation, (ii) state Lemma 2.1, (iii) state and prove Lemmas 2.2 to 2.7, (iv) prove Lemma 2.1, then (v) complete the proof of Theorem 2.1.

The following definitions and notations are used throughout the proof. Let $d_W(X, Y)$ denote the Wasserstein distance between random variables X and Y, so $d_W(X, Y) = 0$ if and only if the distributions of X and Y are identical. The norms of functions are defined as the sup norm i.e., $||f|| = \sup_{x \in D} |f(x)|$. For vector a, $||a|| = (a'a)^{1/2}$ is the Euclidean norm, and for positive semi-definite matrix A and $\lambda_{\max}(A)$ denoting the largest eigenvalue, $||A|| = \sqrt{\lambda_{\max}(A'A)}$ denotes the spectral norm, and $A^{1/2}$ denotes the symmetric matrix such that $A^{1/2}A^{1/2} = A$. $\sum_{i \in \mathcal{N}_g^G} \sum_{j \in \mathcal{N}_g^G}$ is abbreviated as $\sum_{i,j \in \mathcal{N}_g^G}$. The dependency neighborhood of i, $\mathcal{N}_i \subseteq \{1, \cdots, n\}$, is defined as the set of observations where $i \in \mathcal{N}_i$ and X_i is independent of $\{X_j\}_{j \neq \mathcal{N}_i}$, and $N_i := |\mathcal{N}_i|$ is the number of observations in i's dependency neighborhood. $1[A]$ is an indicator function that takes value 1 if A is true and 0 otherwise. In the rest of

this proof, X_i denotes a scalar random variable while $W_i \in \mathbb{R}^K$ as stated in the main text is a random vector. Denote the variance of the sum of the scalar random variable X_i as $\sigma_n^2 := Var\left(\sum_i X_i\right)$. We are interested in the asymptotic distribution of $(1/\sigma_n)\sum_i X_i$.

Assumption 2.4. *For $C \in \{G, H\}$, there exists $K_0 < \infty$ such that:*

(a) $E[X_i] = 0$ *and* $E[X_i^4] \leq K_0 < \infty$ *for all* i*;*

(b) $\frac{1}{\sigma_n^2}\max_c \left(N_c^C\right)^2 \to 0$*;*

(c) $\frac{1}{\sigma_n^2}\sum_c \left(N_c^C\right)^2 \leq K_0 < \infty$*;*

(d) $X_i \perp\!\!\!\perp \{X_j\}_{j \notin \mathcal{N}_i}$*; and*

(e) for observations $i, j, k \in \mathcal{N}_i, l \in \mathcal{N}_j$*, if* $(\mathcal{N}_i \cup \mathcal{N}_k) \cap (\mathcal{N}_j \cup \mathcal{N}_l) = \emptyset$*, then* $Cov(X_i X_k, X_j X_l) = 0$*.*

Lemma 2.1. *Under Assumption 2.4,* $(1/\sigma_n)\sum_i X_i \xrightarrow{d} N(0,1)$*, where* $\sigma_n^2 := Var\left(\sum_i X_i\right)$*. Further, using feasible estimator* $\hat{\sigma}_n^2 := \sum_i \sum_{j \in \mathcal{N}_i} X_i X_j$*,* $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$*.*

Lemma 2.2. *(Theorem 3.1 of Ross (2011)) If R is a random variable, Z has a standard normal distribution, and we define the family of functions* $\mathcal{F} = \{f : ||f||, ||f''|| \leq 2, ||f'|| \leq \sqrt{2\pi}\}$*, then* $d_W(R, Z) \leq \sup_{f \in \mathcal{F}} |E[f'(R) - Rf(R)]|$*.*

The proofs of Lemmas 2.3 and 2.4 follow Ross (2011) Theorem 3.6 up to Equations (3.11) and (3.12).

Lemma 2.3. *Let $X_1, \cdots, X_n$ be random variables such that* $E[X_i] = 0, \sigma_n^2 = Var(\sum_i X_i)$*, and define* $R = \sum_i X_i/\sigma_n$*. If* $R_i := \sum_{j \notin \mathcal{N}_i} X_j/\sigma_n$*, then, for all* $f \in \mathcal{F}$*,*

$$E[Rf(R)] = E\left[\frac{1}{\sigma_n}\sum_i X_i(f(R) - f(R_i) - (R - R_i)f'(R))\right] + E\left[\frac{1}{\sigma_n}\sum_i X_i(R - R_i)f'(R)\right].$$

Proof. Start from right-hand side:

$$
\begin{aligned}
&E\left[\frac{1}{\sigma_n}\sum_i X_i(f(R)-f(R_i)-(R-R_i)f'(R))\right]+E\left[\frac{1}{\sigma_n}\sum_i X_i(R-R_i)f'(R)\right]\\
&=E\left[\frac{1}{\sigma_n}\sum_i X_i(f(R)-f(R_i))\right]=E\left[\frac{1}{\sigma_n}\sum_i X_i f(R)\right]-E\left[\frac{1}{\sigma_n}\sum_i X_i f(R_i)\right]\\
&=E\left[\frac{1}{\sigma_n}\sum_i X_i f(R)\right]=E[Rf(R)].
\end{aligned}
$$

The first equality in the final line comes from the fact that R_i is independent of X_i based on how dependency neighborhoods are defined. Hence, $E[X_i f(R_i)] = 0$. □

Lemma 2.4. *Let $X_1, \cdots, X_n$ be random variables such that, $E[X_i] = 0, \sigma_n^2 = Var(\sum_i X_i)$, and define $R = \sum_i X_i/\sigma_n$. Let the collection $(X_1, \cdots, X_n)$ have dependency neighborhoods $\mathcal{N}_i$, $i = 1, \cdots, n$. Then for Z a standard normal random variable,*

$$
d_W(R,Z) \le \frac{1}{\sigma_n^3}\sum_i\sum_{j,k\in\mathcal{N}_i} E\left[|X_i|X_jX_k\right] + \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2}\sqrt{Var\left(\sum_i\sum_{j\in\mathcal{N}_i} X_iX_j\right)}. \tag{2.2}
$$

Proof. Due to Lemma 2.2, to bound $d_W(R,Z)$ from above, it is sufficient to bound $|E[f'(R) - Rf(R)]|$, where $||f||, ||f''|| \le 2, ||f'|| \le \sqrt{2/\pi}$. Define $R_i := \sum_{j\notin\mathcal{N}_i} X_j/\sigma_n$, so X_i is independent of R_i. Then,

$$
\begin{aligned}
&|E[f'(R)-Rf(R)]| = |E[f'(R)]-E[Rf(R)]|\\
&\le \left|E[f'(R)]-E\left[\frac{1}{\sigma_n}\sum_i X_i(f(R)-f(R_i)-(R-R_i)f'(R))\right]-E\left[\frac{1}{\sigma_n}\sum_i X_i(R-R_i)f'(R)\right]\right|\\
&\le \left|E\left[\frac{1}{\sigma_n}\sum_i X_i(f(R)-f(R_i)-(R-R_i)f'(R))\right]\right| + \left|E\left[f'(R)\left(1-\frac{1}{\sigma_n}\sum_i X_i(R-R_i)\right)\right]\right|.
\end{aligned}
$$

The first inequality applies Lemma 2.3, and the second inequality applies the triangle inequality. Consequently, it is sufficient to show that the first term is bounded by the corre-

sponding first term of Equation (2.2), and the second term is bounded by the corresponding second term.

Consider the first term. By Taylor expansion of $f(R_i)$ around $f(R)$, and the triangle inequality, the term that generates the third moment is:

$$\left| E\left[\frac{1}{\sigma_n} \sum_i X_i (f(R) - f(R_i) - (R - R_i) f'(R)) \right] \right| \leq \frac{||f''||}{2\sigma_n} \left| \sum_i E[|X_i|(R - R_i)^2] \right|$$
$$\leq \frac{1}{\sigma_n^3} \sum_i E\left[|X_i| \left(\sum_{j \in \mathcal{N}_i} X_j \right)^2 \right] = \frac{1}{\sigma_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E[|X_i| X_j X_k].$$

Turning now to the second term,

$$\left| E\left[f'(R) \left(1 - \frac{1}{\sigma_n} \sum_i X_i (R - R_i) \right) \right] \right|$$
$$\leq \frac{||f'||}{\sigma_n^2} E \left| \sigma_n^2 - \sum_i X_i \left(\sum_{j \in N_i} X_j \right) \right| \leq \frac{||f'||}{\sigma_n^2} E \left[\left(\sigma_n^2 - \sum_i X_i \left(\sum_{j \in N_i} X_j \right) \right)^2 \right]^{1/2} 1^{1/2}$$
$$\leq \frac{\sqrt{2}}{\sqrt{\pi} \sigma_n^2} \sqrt{Var\left(\sum_i \sum_{j \in N_i} X_i X_j \right)}.$$

□

Lemma 2.5. *$E[|X_i X_j X_k|] \leq \max_m E[|X_m|^3]$, $E[|X_i X_j X_k X_l|] \leq \max_m E[|X_m|^4]$, and $|E[X_i X_k] E[X_j X_l]| \leq \max_m E[|X_m|^4]$.*

Proof. By the arithmetic mean — geometric mean (AM-GM) inequality,

$$E|X_i X_j X_k| \leq \frac{1}{3} \left(E|X_i|^3 + E|X_j|^3 + E|X_k|^3 \right) \leq \max_m E[|X_m|^3].$$

A similar argument yields $E[|X_iX_jX_kX_l|] \leq \max_m E[|X_m|^4]$. For the final result, first observe that $E[X_iX_k]^2 \pm 2E[X_iX_k]E[X_jX_l] + E[X_jX_l]^2 = (E[X_iX_k] \pm E[X_jX_l])^2 \geq 0$. Hence,

$$\begin{aligned}|E[X_iX_k]E[X_jX_l]| &\leq \frac{1}{2}(E[X_iX_k]^2 + E[X_jX_l]^2) \leq \frac{1}{2}(E[X_i^2X_k^2] + E[X_j^2X_l^2]) \\ &\leq \frac{1}{4}(E[X_i^4] + E[X_j^4] + E[X_k^4] + E[X_l^4]) \leq \max_m E[X_m^4].\end{aligned}$$

□

Lemma 2.6. *Under Assumption 2.4,* $\frac{1}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} E\left[|X_i|X_jX_k\right] = o(1)$.

Proof. Using Lemma 2.5,

$$\begin{aligned}\frac{1}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} E\left[|X_i|X_jX_k\right] &\leq \frac{1}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} E\left[||X_i|X_jX_k|\right] \\ &\leq \frac{\max_m E[|X_m|^3]}{\sigma_n^3}\sum_i \sum_{j,k\in\mathcal{N}_i} 1 = \frac{\max_m E[|X_m|^3]}{\sigma_n^3}\sum_i N_i^2.\end{aligned}$$

Observe $\max_m E[|X_m|^3] \leq K_0$ since the 4th moment exists, so it remains to show that the remaining terms are $o(1)$. Due to Assumption 2.1, $N_i \leq N_{g(i)}^G + N_{h(i)}^H$, so

$$\begin{aligned}\frac{1}{\sigma_n^3}\sum_i N_i^2 \leq \frac{1}{\sigma_n^3}\sum_i (N_{g(i)}^G + N_{h(i)}^H)^2 &\leq \frac{1}{\sigma_n^3}\max_{g,h}(N_g^G + N_h^H)\sum_i (N_{g(i)} + N_{h(i)}) \\ &\leq \left[\frac{1}{\sigma_n}\max_{g,h}(N_g^G + N_h^H)\right]\frac{1}{\sigma_n^2}\left(\sum_g (N_g^G)^2 + \sum_h (N_h^H)^2\right).\end{aligned}$$

$\max_{g,h}(N_g^G + N_h^H)/\sigma_n \to 0$ by Assumption 2.2(b) and the final term $\left(\sum_g (N_g^G)^2 + \sum_h (N_h^H)^2\right)/\sigma_n^2$ is bounded by Assumption 2.2(c). Hence, the term is $o(1)$. □

Lemma 2.7. *Under Assumption 2.4,* $\frac{1}{\sigma_n^4} Var\left(\sum_i \sum_{j\in\mathcal{N}_i} X_iX_j\right) = o(1)$.

Proof. Observe that:

$$\frac{1}{\sigma_n^4} Var\left(\sum_i \sum_{j \in \mathcal{N}_i} X_i X_j\right) = \frac{1}{\sigma_n^4} E\left[\left(\sum_i \sum_{j \in \mathcal{N}_i} X_i X_j\right)^2\right] - \frac{1}{\sigma_n^4}\left(\sum_i \sum_{j \in \mathcal{N}_i} E[X_i X_j]\right)^2$$
$$= \frac{1}{\sigma_n^4} \sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} (E[X_i X_j X_k X_l] - E[X_i X_k] E[X_j X_l]).$$

Due to Assumption 2.1(b), when j, l do not share any cluster with i, k, $E[X_i X_j X_k X_l] = E[X_i X_k] E[X_j X_l]$. Hence, we only have to consider terms where there is at least one pair that shares a cluster. Let $A_{ij} := 1[j \in \mathcal{N}_i]$. With finite 4th moment and Lemma 2.5, using the same argument as the proof of Lemma 2.6, it is sufficient to show

$$\frac{1}{\sigma_n^4} \sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} (A_{ij} + A_{il} + A_{kj} + A_{kl}) = o(1).$$

It is sufficient to consider the A_{ij} term because the other terms are symmetric. In particular,

$$\sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} A_{il} = \sum_i \sum_{k \in \mathcal{N}_i} \sum_l \sum_{j \in \mathcal{N}_l} A_{il} = \sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} A_{ij},$$
$$\sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} A_{kj} = \sum_k \sum_{i \in \mathcal{N}_k} \sum_j \sum_{l \in \mathcal{N}_j} A_{kj} = \sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} A_{ij}, \text{ and}$$
$$\sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} A_{kl} = \sum_k \sum_l \sum_{i \in \mathcal{N}_k} \sum_{j \in \mathcal{N}_l} A_{kl} = \sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} A_{ij}.$$

Considering the A_{ij} term,

$$\sum_i \sum_j \sum_{k \in \mathcal{N}_i} \sum_{l \in \mathcal{N}_j} A_{ij} \leq \sum_i \left(\sum_{j \in \mathcal{N}^G_{g(i)}} + \sum_{j \in \mathcal{N}^H_{h(i)}}\right)\left(\sum_{k \in \mathcal{N}^G_{g(i)}} + \sum_{k \in \mathcal{N}^H_{h(i)}}\right)\left(\sum_{l \in \mathcal{N}^G_{g(j)}} + \sum_{l \in \mathcal{N}^H_{h(j)}}\right) A_{ij}.$$

The first and last terms of the summation take the form:

$$\sum_i \sum_{j \in \mathcal{N}^G_{g(i)}} \sum_{k \in \mathcal{N}^G_{g(i)}} \sum_{l \in \mathcal{N}^G_{g(j)}} A_{ij} = \sum_g \sum_{i,j,k,l \in \mathcal{N}^G_g} A_{ij} = \sum_g (N^G_g)^4.$$

The first equality in the equation above follows from how $\sum_i = \sum_g \sum_{i \in \mathcal{N}^G_g}$ and that if $j \in \mathcal{N}^G_{g(i)}$, then i and j share the same g and hence $\sum_{l \in \mathcal{N}^G_{g(j)}} = \sum_{l \in \mathcal{N}^G_{g(i)}}$. The second equality occurs as $A_{ij} = 1$ when i and j share the same g cluster. With this equality, observe that $\sum_g (N^G_g)^4 = \left(\max_g (N^G_g)^2\right) \sum_g (N^G_g)^2$. Since $\frac{1}{\sigma_n^2} \max_g \left(N^G_g\right)^2 = o(1)$ and $\frac{1}{\sigma_n^2} \sum_g \sum_{i,j \in \mathcal{N}^G_g} A_{ij} \leq \frac{1}{\sigma_n^2} \sum_g (N^G_g)^2 < \infty$ by Assumption 2.4, these terms are $o(1)$ when divided by σ_n^4.

An upper bound can similarly be derived for the interactive terms. To explain the steps carefully, I label the equalities and inequalities (i) to (iv):

$$\begin{aligned}
&\sum_i \sum_{j \in \mathcal{N}^G_{g(i)}} \sum_{k \in \mathcal{N}^G_{g(i)}} \sum_{l \in \mathcal{N}^H_{h(j)}} A_{ij} \\
&\quad \overset{(i)}{=} \sum_{i,j,k} \sum_g 1[i \in \mathcal{N}^G_g] 1[j \in \mathcal{N}^G_g] 1[k \in \mathcal{N}^G_g] \sum_l \sum_h 1[j \in \mathcal{N}^H_h] 1[l \in \mathcal{N}^H_h] A_{ij} \\
&\quad \overset{(ii)}{=} \sum_g \sum_{i,j,k} 1[i \in \mathcal{N}^G_g] 1[j \in \mathcal{N}^G_g] 1[k \in \mathcal{N}^G_g] A_{ij} \sum_h \sum_l 1[j \in \mathcal{N}^H_h] 1[l \in \mathcal{N}^H_h] \\
&\quad \overset{(iii)}{\leq} \left(\max_j \sum_h \sum_l 1[j \in \mathcal{N}^H_h] 1[l \in \mathcal{N}^H_h] \right) \left(\sum_g \sum_{i,j,k \in \mathcal{N}^G_g} A_{ij} \right) \\
&\quad \overset{(iv)}{\leq} \left(\max_h \sum_{l \in \mathcal{N}^H_h} 1 \right) \left(\max_g \sum_{k \in \mathcal{N}^G_g} 1 \right) \left(\sum_g \sum_{i,j \in \mathcal{N}^G_g} A_{ij} \right) = \left(\max_h N^H_h \right) \left(\max_g N^G_g \right) \left(\sum_g (N^G_g)^2 \right).
\end{aligned}$$

The equality in (i) is obtained by transforming the conditional sums into sums over products of indicators. The equality in (ii) is obtained from commutative and associative properties of additional and multiplication. In step (iii), the inequality is obtained by using the upper bound on the innermost sum over h and l. In step (iv), to see how $\max_j \sum_h \sum_l 1[j \in \mathcal{N}^H_h] 1[l \in \mathcal{N}^H_h] = \max_h \sum_{l \in \mathcal{N}^H_h} 1$, observe that once we choose the index

j, the indicator $1[j \in \mathcal{N}_h^H]$ can only take value 1 for one particular h, so the maximum occurs when we choose a corresponding h that results in the largest $\sum_{l \in \mathcal{N}_h^H} 1$. The inequality in (iv) is due to extracting $\left(\max_g \sum_{k \in \mathcal{N}_g^G} 1\right)$ from $\left(\sum_g \sum_{i,j,k \in \mathcal{N}_g^G} A_{ij}\right)$. Since $\sum_g (N_g^G)^2/\sigma_n^2 \leq K_0$ and $\max_g N_g^G/\sigma_n = o(1)$,

$$\frac{1}{\sigma_n^4} \sum_i \sum_{j \in \mathcal{N}_{g(i)}^G} \sum_{k \in \mathcal{N}_{g(i)}^G} \sum_{l \in \mathcal{N}_{h(j)}^H} A_{ij} \leq \left(\frac{1}{\sigma_n} \max_h N_h^H\right) \left(\frac{1}{\sigma_n} \max_g N_g^G\right) \left(\frac{1}{\sigma_n^2} \sum_g (N_g^G)^2\right) = o(1).$$

□

Proof of Lemma 2.1. Apply Lemma 2.4 to obtain:

$$d_W(R, Z) \leq \frac{1}{\sigma_n^3} \sum_i \sum_{j,k \in \mathcal{N}_i} E[|X_i| X_j X_k] + \frac{\sqrt{2}}{\sqrt{\pi}\sigma_n^2} \sqrt{Var\left(\sum_i \sum_{j \in \mathcal{N}_i} X_i X_j\right)}.$$

Applying Lemma 2.6 and 2.7 on each of the two terms, $d_W(R, Z) = o(1)$. Proof for consistency of the variance estimator is equivalent to proving that $(\hat{\sigma}_n^2 - \sigma_n^2)/\sigma_n^2 = o_P(1)$. By Chebyshev's inequality,

$$P\left(\frac{\hat{\sigma}_n^2 - \sigma_n^2}{\sigma_n^2} > \epsilon\right) \leq \frac{1}{\epsilon^2} \frac{1}{\sigma_n^4} E[(\hat{\sigma}_n^2 - \sigma_n^2)^2] = \frac{Var\left(\sum_i \sum_{j \in \mathcal{N}_i} X_i X_j\right)}{\epsilon^2 \sigma_n^4} = o_P(1).$$

The convergence in the last step occurs by Lemma 2.7. □

Proof of Theorem 2.1. To show that $Q_n^{-1/2} \sum_i (W_i - E[W_i]) \xrightarrow{d} N(0, I_K)$, due to the Cramer-Wold device, it suffices to show that $\forall \mu \in \mathbb{R}^K$, $\mu' Q_n^{-1/2} \sum_i (W_i - E[W_i]) \xrightarrow{d} \mu' N(0, I_K)$. If μ is a vector of zeroes, then $\mu' Q_n^{-1/2} \sum_i (W_i - E[W_i]) \xrightarrow{d} \mu' N(0, I_K)$ is immediate. For $||\mu|| > 0$, it suffices to show $(1/||\mu||)\mu' Q_n^{-1/2} \sum_i (W_i - E[W_i]) \xrightarrow{d} (1/||\mu||)\mu' N(0, I_K) = N(0, 1)$. Without loss of generality, we can set $||\mu|| = 1$. For all nonstochastic $\mu \in \mathbb{R}^K \backslash \{0\}$, let $\sigma_n^2(\mu) := Var\left(\sum_i \mu' \left(Q_n/\lambda_n\right)^{-1/2} (W_i - E[W_i])\right)$, so the following hold:

1. $E\left[\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right)\right] = 0$ and $E\left[\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right)^4\right] \leq K_0$ for all i.

2. $\frac{1}{\sigma_n^2(\mu)}\max_c\left(N_c^C\right)^2 \to 0$.

3. $\frac{1}{\sigma_n^2(\mu)}\sum_c (N_c^C)^2 \leq K_0$.

4. $\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}(W_i - E[W_i])\right) \perp\!\!\!\perp \left\{\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_j\right)\right\}_{j\notin\mathcal{N}_i}$.

5. For observations $i, j, k \in \mathcal{N}_i, l \in \mathcal{N}_j$, if $(\mathcal{N}_i \cup \mathcal{N}_k) \cap (\mathcal{N}_j \cup \mathcal{N}_l) = \emptyset$, then

$$Cov\left(\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_i\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_k, \mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_j\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}W_l\right) = 0.$$

For item 1, since $\lambda_n := \lambda_{\min}(Q_n)$, all eigenvalues of Q_n/λ_n must be at least 1. Hence, all eigenvalues of $(Q_n/\lambda_n)^{-1/2}$ are bounded above by 1, which implies $|\mu'(Q_n/\lambda_n)^{-1/2}| \leq K_1$ for some arbitrary constant $K_1 < \infty$. Item 1 then follows from Assumption 2.2(a). Observe that $\sigma_n^2(\mu) = \mu'(Q_n/\lambda_n)^{-1/2}Q_n(Q_n/\lambda_n)^{-1/2}\mu = \lambda_n$. Then, Assumption 2.2(b) yields item 2, and Assumption 2.2(c) yields item 3. Item 4 is immediate from Assumption 2.1(a), and item 5 from Assumption 2.1(b). By applying Lemma 2.1, $(1/\sigma_n(\mu))\mu'(Q_n/\lambda_n)^{-1/2}\sum_i(W_i - E[W_i]) \xrightarrow{d} N(0,1)$. By using $\sigma_n^2(\mu) = \lambda_n$, this result is equivalent to $\mu'Q_n^{-1/2}\sum_i(W_i - E[W_i]) \xrightarrow{d} N(0,1)$ as required.

Turning to consistent variance estimation, I first show that $(1/\lambda_n)(\hat{Q}_n - Q_n) \xrightarrow{p} 0_{K\times K}$, where $0_{K\times K}$ is a $K \times K$ matrix of zeroes. Since $\hat{Q}_n - Q_n = \sum_i\sum_{j\in\mathcal{N}_i} W_iW_j' - E[W_iW_j']$, it suffices to show convergence elementwise. Let X_i and Y_i denote scalar components of W_i,

i.e., $X_i = W_{im}, Y_i = W_{ip}$, where $m, p \in \{1, 2, \cdots, K\}$. Then,

$$
\begin{aligned}
P\left(\frac{1}{\lambda_n}\sum_i\sum_{j\in\mathcal{N}_i}(X_iY_j - E[X_iY_j]) > \epsilon\right) &\leq \frac{1}{\epsilon^2}\frac{1}{\lambda_n^2}Var\left(\sum_i\sum_{j\in\mathcal{N}_i}X_iY_j\right) \\
&\leq \frac{1}{\epsilon^2\lambda_n^2}\sum_i\sum_j\sum_{k\in\mathcal{N}_i}\sum_{l\in\mathcal{N}_j}|E[X_iX_jY_kY_l] - E[X_iY_k]E[X_jY_l]| \\
&\leq \frac{K_0}{\lambda_n^2}\sum_i\sum_j\sum_{k\in\mathcal{N}_i}\sum_{l\in\mathcal{N}_j}(A_{ij} + A_{il} + A_{kj} + A_{kl}) = o(1).
\end{aligned}
$$

The inequality in the last line is obtained due to Hölder's inequality and finite moments. An argument similar to that of Lemma 2.7 yields the $o(1)$ equality. Then,

$$
\mu'(Q_n^{-1/2}(\hat{Q}_n - Q_n)Q_n^{-1/2})\mu = \mu_0'\frac{1}{\lambda_n}(\hat{Q}_n - Q_n)\mu_0 \xrightarrow{p} 0.
$$

where μ_0 is a vector whose entries are all bounded above by some arbitrary constant $K_1 < \infty$ by a similar argument as before. Convergence occurs because $(1/\lambda_n)(\hat{Q}_n - Q_n) \xrightarrow{p} 0_{K\times K}$.

□

2.B Proof of Propositions

Proof of Proposition 2.1. For (a), take any observation i and its associated clusters $g(i), h(i)$. Use the permutation function $\pi_1(g(i)) = 1$ and $\pi_2(h(i)) = 1$ so the array has the same distribution as before due to separate exchangeability. Since the array is dissociated, by setting $G_0 = H_0 = 1$, W_i is independent of all observations that are not in $g(i)$ or $h(i)$, verifying (a).

For (b), take any i and $k \in \mathcal{N}_i$. Without loss of generality, suppose that $g(i) = g(k)$. Consider the case where $h(i) \neq h(k)$. Use the permutation function $\pi_1(g(i)) = 1$ and $\pi_2(h(i)) = 1, \pi_2(h(k)) = 2$ to get another array that has the same distribution. Since the array is dissociated, by setting $G_0 = 1, H_0 = 2$, (W_i, W_k) is independent of all observations

that are not in $(\mathcal{N}_i \cup \mathcal{N}_k)$. Since $j, l \notin (\mathcal{N}_i \cup \mathcal{N}_k)$, $(W_i, W_k) \perp\!\!\!\perp (W_j, W_l)$, which yields (b). If $h(k) = h(i)$, set $\pi_2(h(k)) = 1$ and $G_0 = 1, H_0 = 1$. The same argument applies. □

For Proposition 2, I first prove a consistency result.

Lemma 2.8. *Under Assumptions 2.1, 2.2(a) and 2.2(b), and* $E[W_i] = 0\ \forall i$, $||(1/n\sum_i(W_iW_i' - E[W_iW_i']))|| \xrightarrow{p} 0$.

Proof. It suffices to show convergence elementwise. Let X_i and Y_i denote scalar components of W_i, i.e., $X_i = W_{im}, Y_i = W_{ip}$, where $m, p \in \{1, 2, \cdots, K\}$. By Chebyshev's inequality, and Assumption 2.2(a) that $\max_{m,k} E[W_{mk}^4] < K_0$,

$$P\left(\frac{1}{n}\sum_i (X_iY_i - E[X_iY_i]) > \epsilon\right)$$
$$\leq \frac{1}{\epsilon^2}\frac{1}{n^2}E\left(\sum_i\sum_{j\in\mathcal{N}_i}(X_iY_i - E[X_iY_i])(X_jY_j - E[X_jY_j])\right) \leq \frac{K_0}{\epsilon^2 n^2}\sum_i\sum_{j\in\mathcal{N}_i} 1.$$

Hence, it suffices to show $(\sum_i \sum_{j\in\mathcal{N}_i} 1)/n^2 = o(1)$. Observe

$$\frac{\sum_i\sum_{j\in\mathcal{N}_i} 1}{n^2} \leq \frac{\max_i N_i}{n}\frac{(\sum_i 1)}{n},$$

so it suffices to show $\max_i N_i/n = o(1)$. Since

$$\lambda_n \leq \sum_i\sum_{j\in\mathcal{N}_i}\max_m E[W_{mk}^2] \leq n^2 \max_m E[W_{mk}^2],$$

we have:

$$\frac{(\max_i N_i)^2}{n^2} = \frac{(\max_i N_i)^2 \max_m E[W_{mk}^2]}{n^2 \max_m E[W_{mk}^2]} \leq \max_m E[W_{mk}^2]\frac{(\max_i N_i)^2}{\lambda_n} = o(1).$$

Convergence occurs due to Assumption 2.2(b) and $\max_m E[W_{mk}^2] < K_0$. □

Proof of Proposition 2.2. Since $E[u_i^4|X_i] \leq K_0$, $E[u_i^4 X_{ik}^4] = E[E[u_i^4|X_i]X_{ik}^4] \leq K_0 E[X_{ik}^4] \leq K_0^2$ is bounded. By Theorem 2.1, $Q_n^{-1/2}\sum_i X_i u_i \xrightarrow{d} N(0, I_K)$.

To complete the normality result, I show that $S_n^{-1}\hat{S}_n \xrightarrow{p} I_K$, which is the same as showing that $||S_n^{-1}(\hat{S}_n - S_n)|| \xrightarrow{p} 0$. By applying Lemma 2.8, $(1/n)(\hat{S}_n - S_n) = (1/n)\sum_i (X_i X_i' - E[X_i X_i']) = o_P(1)$. Hence, it suffices that $(S_n/n)^{-1}$ has bounded eigenvalues, i.e., $\lambda_{\min}(S_n/n) \geq K_1 > 0$, which is true by Assumption 2.3(e). Since $\hat{\beta} - \beta = \hat{S}_n^{-1}\sum_i X_i u_i$, by Slutsky's lemma, $Q_n^{-1/2} S_n(\hat{\beta} - \beta) \xrightarrow{d} N(0, I_K)$.

Next, proceed to consistent variance estimation. Showing that $||Q_n^{-1}\hat{Q}_n - I_K|| = o_P(1)$ is equivalent to showing that, $\forall \mu \in \mathbb{R}^K$, $\mu'\left(Q_n^{-1/2}(\hat{Q}_n - Q_n)Q_n^{-1/2}\right)\mu = o_P(1)$. Expanding $\hat{Q}_n$,

$$\begin{aligned}\hat{Q}_n &:= \sum_i \sum_{j\in\mathcal{N}_i} \hat{u}_i\hat{u}_j X_i X_j' = \sum_i \sum_{j\in\mathcal{N}_i} (u_i - X_i'(\hat{\beta}-\beta))(u_j - X_j'(\hat{\beta}-\beta))X_i X_j' \\ &= \sum_i \sum_{j\in\mathcal{N}_i} u_i u_j X_i X_j' - 2\left(\sum_i \sum_{j\in\mathcal{N}_i} u_i X_j'(\hat{\beta}-\beta)X_i X_j'\right) + \left(\sum_i \sum_{j\in\mathcal{N}_i} X_i'(\hat{\beta}-\beta)X_j'(\hat{\beta}-\beta)X_i X_j'\right).\end{aligned}$$

By Theorem 2.1, $\mu' Q_n^{-1/2}(\sum_i \sum_{j\in\mathcal{N}_i} u_i u_j X_i X_j' - Q_n)Q_n^{-1/2}\mu = o_P(1)$. Hence, it remains to show:

$$\left|\left| Q_n^{-1/2}\left[-2\left(\sum_i \sum_{j\in\mathcal{N}_i} u_i X_j'(\hat{\beta}-\beta)X_i X_j'\right) + \left(\sum_i \sum_{j\in\mathcal{N}_i} X_i'(\hat{\beta}-\beta)X_j'(\hat{\beta}-\beta)X_i X_j'\right)\right] Q_n^{-1/2}\right|\right| = o_P(1).$$

Observe that $X_i'(\hat{\beta}-\beta) = \left(X_i' S_n^{-1} Q_n^{1/2}\right)\left(Q_n^{-1/2} S_n(\hat{\beta}-\beta)\right) = \left(X_i' S_n^{-1} Q_n^{1/2}\right)(Z_K + 1_K o_P(1))$, where 1_K is a K-vector of ones and $Z_K \sim N(0, I_K)$. Hence, addressing the second

term,

$$\begin{aligned} X_i'(\hat{\beta}-\beta)X_j'(\hat{\beta}-\beta) &= \left(X_i'S_n^{-1}Q_n^{1/2}\right)(Z_K+1_K o_P(1))(Z_K+1_K o_P(1))'\left(X_j'S_n^{-1}Q_n^{1/2}\right)' \\ &= \left(X_i'S_n^{-1}Q_n^{1/2}\right)(I_K O_P(1)+o_P(1))\left(X_j'S_n^{-1}Q_n^{1/2}\right)' \\ &= X_i'S_n^{-1}Q_nS_n^{-1}X_jO_P(1). \end{aligned}$$

This equality implies:

$$\begin{aligned} &Q_n^{-1/2}\left(\sum_i\sum_{j\in\mathcal{N}_i}X_i'(\hat{\beta}-\beta)X_j'(\hat{\beta}-\beta)X_iX_j'\right)Q_n^{-1/2} \\ &= Q_n^{-1/2}\left(\sum_i\sum_{j\in\mathcal{N}_i}\left(X_i'S_n^{-1}Q_nS_n^{-1}X_j\right)X_iX_j'\right)Q_n^{-1/2}O_P(1) \\ &= \frac{1}{n^2}\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\left(\sum_i\sum_{j\in\mathcal{N}_i}\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)\left(\frac{1}{n}S_n\right)^{-1}X_j\right)X_iX_j'\right)\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}O_P(1). \end{aligned}$$

The eigenvalues of (Q_n/λ_n) are bounded. To see this, it suffices to show that there exists $K_0<\infty$ such that $\lambda_{\max}(Q_n)/\lambda_n\leq K_0$. Due to finite moments, $Q_n:=Var(\sum_i X_i)\leq K_0 1_{K\times K}\sum_c(N_c^C)^2$. Since $(\sum_c(N_c^C)^2)/\lambda_n\leq K_0$ by Assumption 2.3, $\lambda_nK_0\geq\sum_c(N_c^C)^2$, which implies $\lambda_n\geq(\sum_c(N_c^C)^2)/K_0$. Hence,

$$\frac{\lambda_{\max}(Q_n)}{\lambda_n}\leq\frac{\sum_c(N_c^C)^2K_0}{\sum_c(N_c^C)^2\frac{1}{K_0}}=K_0^2.$$

Recall that $(S_n/n)^{-1}$ has bounded eigenvalues. The proof of Theorem 2.1 also showed that $(Q_n/\lambda_n)^{-1}$ has bounded eigenvalues. By using Markov and Minkowski inequalities, and the

same argument as the proof of Theorem 2.1 for $\mu \in \mathbb{R}^K, ||\mu|| = 1$,

$$
\begin{aligned}
&P\left(\frac{1}{n^2}\left|\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\left(\sum_i\sum_{j\in\mathcal{N}_i}\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)\left(\frac{1}{n}S_n\right)^{-1}X_j\right)X_iX_j'\right)\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\mu\right| > \epsilon\right)\\
&\leq \frac{1}{n^2\epsilon}E\left[\left|\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\left(\sum_i\sum_{j\in\mathcal{N}_i}\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)\left(\frac{1}{n}S_n\right)^{-1}X_j\right)X_iX_j'\right)\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\mu\right|\right]\\
&\leq \frac{1}{n^2\epsilon}\sum_i N_i \max_{m,k} E[X_{mk}^4]K_0 \leq \frac{\max_i N_i}{n}\frac{n}{n}K_0 \to 0,
\end{aligned}
$$

where $K_0 \in \mathbb{R}$ is an arbitrary (finite) constant. Convergence occurs due to Assumption 2.3(b), which implies $\max_i N_i/n \to 0$, since $\max_i \sum_{j\in\mathcal{N}_i} N_i/n = o(1)$ in the proof of Lemma 2.8.

Going back to the first term,

$$
\begin{aligned}
Q_n^{-1/2}\sum_i\sum_{j\in\mathcal{N}_i}u_iX_j'(\hat{\beta}-\beta)X_iX_j'Q_n^{-1/2} &= Q_n^{-1/2}\sum_i\sum_{j\in\mathcal{N}_i}u_i\left(X_i'S_n^{-1}Q_n^{1/2}\right)(Z_K + 1_K o_P(1))X_iX_j'Q_n^{-1/2}\\
&= \frac{1}{n\sqrt{\lambda_n}}\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\sum_i\sum_{j\in\mathcal{N}_i}u_i\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)^{1/2}\right)X_iX_j'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}O_P(1).
\end{aligned}
$$

By using Markov and Minkowski inequalities,

$$
\begin{aligned}
&P\left(\frac{1}{n\sqrt{\lambda_n}}\left|\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\sum_i\sum_{j\in\mathcal{N}_i}u_i\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)^{1/2}\right)X_iX_j'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\mu\right| > \epsilon\right)\\
&\leq \frac{1}{n\sqrt{\lambda_n}\epsilon}E\left[\left|\mu'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\sum_i\sum_{j\in\mathcal{N}_i}u_i\left(X_i'\left(\frac{1}{n}S_n\right)^{-1}\left(\frac{1}{\lambda_n}Q_n\right)^{1/2}\right)X_iX_j'\left(\frac{1}{\lambda_n}Q_n\right)^{-1/2}\mu\right|\right]\\
&\leq \frac{1}{n\sqrt{\lambda_n}\epsilon}\sum_i\sum_{j\in\mathcal{N}_i}\max_{m_1,m_2,k}E\left[\left|X_{m_1k}u_{m_1}X_{m_2}^2\right|\right]K_0\\
&\leq \frac{1}{n\sqrt{\lambda_n}\epsilon}\sum_i N_i \max_{m_1,m_2,k}E\left[|X_{m_1k}u_{m_1}|^2\right]^{1/2}E\left[|X_{m_2}^2|^2\right]^{1/2}K_0\\
&\leq \frac{\max_i N_i}{\sqrt{\lambda_n}}\frac{1}{\epsilon}\max_{m_1,m_2,k}E[X_{m_1k}^2u_{m_1}^2]^{1/2}E[X_{m_2}^4]^{1/2}K_0 = o(1).
\end{aligned}
$$

The penultimate inequality occurs due to Hölder's inequality. Observe that $\max_i N_i/\sqrt{\lambda_n} = o(1)$ if and only if $\max_c (N_c^C)^2/\lambda_n = o(1)$, which is given by Assumption 2.3(b). Convergence in the last step occurs because $\max_i N_i/\sqrt{\lambda_n} = o(1)$, and the moments are finite.

Hence, it has been shown that $Q_n^{-1}\hat{Q}_n \xrightarrow{p} I_K$. Then, $[S_n^{-1}Q_nS_n^{-1}]^{-1}[\hat{S}_n^{-1}\hat{Q}_n\hat{S}_n^{-1}] \xrightarrow{p} I_K$ by the continuous mapping theorem. □

Chapter 3

Sensitivity of Policy Relevant Treatment Parameters to Violations of Monotonicity[1]

3.1 Introduction

Since the seminal work of Heckman and Vytlacil (2005), there has been a large literature that is concerned with identification and inference of policy relevant treatment parameters (PRTP) in instrumental variable (IV) settings with heterogeneous treatment effects (TE). PRTP is a general class of objects that includes the local average treatment effect (LATE) and various TE in counterfactual environments. Existing methods that target the general class of PRTP rely on the monotonicity assumption that the instrument affects all individuals' treatment response in the same direction, which is usually imposed through an additively separable treatment selection equation (e.g., Mogstad et al. (2018)). However, monotonicity may not be realistic in many applications. Consider the Angrist and Evans (1998) study that was interested in the effect of having a third child on the mother's labor supply. They

[1]This chapter is accepted at the *Journal of Applied Econometrics*.

used an indicator for whether the first two kids are of the same sex as an instrument for the third child. Since parents have a preference for gender balance among their children, families with two boys or two girls are more likely to have a third child. But some parents may want two sons or two daughters, so they would violate monotonicity, which rules out families who would have a third child if their first two children are of the same sex, and would not have a third child if their first two children are of different sex. Further examples of monotonicity failure are considered in De Chaisemartin (2017). This observation raises the question of how much bounds on PRTP would change when monotonicity fails. This paper explicitly places a bound on the extent that monotonicity fails, which nests approaches that either impose or drop monotonicity as special cases.

The goal is to place bounds on PRTP while accommodating limited violations of monotonicity. Sensitivity restrictions characterize these violations: I use a sensitivity parameter that places an upper bound on the proportion of defiers relative to compliers. To obtain bounds on PRTP, I adapt the setup and linear program in Mogstad et al. (2018) to accommodate defiers. PRTP can be written as linear combinations of conditional means of potential outcomes for subgroups defined by their treatment response to the instrument. Hence, with appropriate assumptions, the linear program can be retained. The baseline specification of the constraint set uses mean compatibility restrictions across conditional outcome distributions, but the method is amenable to additional restrictions researchers may wish to impose. This procedure yields an identified set that is an interval, and can be modified to incorporate covariates. Providing this tool for sensitivity analysis of PRTP is the main contribution of the paper.

As an application of the general theoretical results, I detail a particular type of PRTP — the treatment effect for compliers under a counterfactual policy environment, which I call the LATE*. In the Angrist and Evans (1998) study, the estimated effect of a third child on the mother's employment status from the IV regression is specific to the policy environment

surrounding childcare in the dataset. Would we still have the same conclusion when the government gives a subsidy for childcare? What would the effect of a third child be for compliers in this counterfactual environment? These are questions answered by LATE*, which nests LATE as a special case (i.e., when there is no extrapolation). The LATE* is one way to think about external validity of a study's conclusions, which researchers are often interested in (e.g., Muralidharan et al. (2019); Ito et al. (2021)).[2]

In the counterfactual environment described, the treatment propensity for the entire population changes while the instrument values are the same. To obtain the LATE*, it suffices to characterize the mass of various treatment response groups in the original environment becoming compliers in the counterfactual environment. At a high level, (partial) identification of the LATE* is possible because the data places some restrictions on the means of potential outcomes, and objects of interest merely reweight these potential outcome means. If we are willing to put bounds on the fraction of people who respond to the instrument in the counterfactual environment relative to the original, meaningful bounds can be obtained. The same logic applies to other PRTP.

The procedure is implemented in the Angrist and Evans (1998) example. An instrument is used because it is believed that the OLS estimand is downward-biased: due to unobserved factors, women who are less likely to work are also those who are more likely to have a third kid. Hence, when the lower bound of the IV estimand reaches the OLS estimand, the bounds are no longer informative. I find that the bounds are informative only for small violations of monotonicity. Consider a counterfactual environment where a childcare subsidy is available. When the mass of defiers is more than 20% the mass of compliers, the lower bound for the

[2]When calculating policy effects in counterfactual environments, parametric models of Brinch et al. (2017) and Kline and Walters (2019) are often used. However, these approaches are less useful when thinking of LATE* as a means to check external validity: since identification of heterogeneous treatment effects are often done without a parametric model, it seems desirable to avoid parametric models when evaluating the robustness of results.

LATE* falls from -0.103 under monotonicity to below the OLS benchmark of -0.134. Hence, the informativeness of the counterfactual estimates depends crucially on monotonicity.

This paper relates to several strands of literature. First, it is related to a literature on the failure of monotonicity in IV settings. Some papers that address violation of monotonicity include reinterpreting the estimand for the LATE (De Chaisemartin, 2017), using weaker monotonicity assumptions (Small et al., 2017; Heckman and Pinto, 2018; Kamat, 2018; Dahl et al., 2023) or alternative assumptions (Klein, 2010), and testing if it is indeed a concern (Kitagawa, 2015). Another common approach is the put bounds on the ATE or the LATE either using worst-case bounds or through some form of sensitivity analysis (Manski, 1989; Balke and Pearl, 1997; Horowitz and Manski, 2000; Noack, 2021; Kitagawa, 2021). There is also a literature that place bounds on further populations (e.g., compliers, defiers, never takers and always takers) (Richardson and Robins, 2010; Huber and Mellace, 2015; Huber et al., 2017; Ding and Lu, 2017). By targeting the PRTP, this paper not only covers bounds on these subpopulations, but also contributes bounds on extrapolated objects in counterfactual environments without monotonicity. Nonetheless, the approach in this paper does not have sharpness guarantees or closed-form solutions like in much of the existing literature.

Second, this paper is related to the literature on extrapolation and external validity in IV settings. In counterfactual environments, parametric models are often used (Brinch et al., 2017; Kline and Walters, 2019). Papers that target PRTP without a parametric model rely on a separable selection equation (Heckman and Vytlacil, 2005; Mogstad et al., 2018). The approach used in this paper neither uses a parametric model nor a separable selection equation — the latter cannot hold by construction when allowing for defiers. In light of the numerical equivalence between selection equations and the group primitives (Heckman and Vytlacil, 2005; Kline and Walters, 2019), this paper additionally contributes an example of how group primitives map to some nonseparable equation that permits extrapolation when monotonicity fails.

The rest of this paper discusses the proposed method and its applications. Section 3.2 explains the general framework in forming bounds for PRTP; Section 3.3 applies the framework to LATE*. Section 3.4 applies the procedure to the Angrist and Evans (1998) example. Section 3.5 concludes.

3.2 Framework for Identification without Monotonicity

3.2.1 Setting

We observe random variables (T, Z, Y), denoting treatment, instrument, and outcome respectively. We are interested in the effect of the endogenous T on Y in a counterfactual environment. Outcome Y can be discrete or continuous; instrument $Z \in \mathcal{Z} = \{0, 1, \cdots, k-1\}$ takes one of $k < \infty$ discrete values, and treatment $T \in \{0, 1\}$ is binary. Although the setup can be adapted to multivalued T, I focus on the binary case for simplicity. Let $T(z)$ denote the potential treatment when given instrument z, and let $Y(t)$ denote the potential outcome when given treatment t, which assumes that Y is not affected by Z directly. Let $T^*(z^*)$ denote the potential treatment when given instrument $z^* \in \mathcal{Z}^*$ in the counterfactual environment, where $\mathcal{Z}^*$ is the set of values that the instrument can take in the counterfactual environment. Without loss of generality, the instrument values are ordered such that $\Pr(T(z) = 1)$ is increasing in z.[3] Then, the observed T and Y are $Y = Y(T)$ and $T = T(Z)$.

Treatment response groups $g \in \mathcal{G}$ are characterized by the vector of potential treatments, i.e., $((T(z))_{z\in\mathcal{Z}}, (T^*(z^*))_{z^*\in\mathcal{Z}^*})$. $\mathcal{G}$ is the set of all possible combinations of $((T(z))_{z\in\mathcal{Z}}, (T^*(z^*))_{z^*\in\mathcal{Z}^*})$: with a binary treatment, k instrument values, and $\mathcal{Z}^* = \mathcal{Z}$, we have $|\mathcal{G}| = 2^{2k}$. Without extrapolation, the counterfactual environment is the original

[3]There is a bijection from any set $\mathcal{Z}'$ with k discrete values to $\mathcal{Z}$ such that for any $z, z' \in \mathcal{Z}$ such that $z > z'$, $\Pr(T(z) = 1) \geq \Pr(T(z') = 1)$. Hence, beyond having k discrete values for the instrument, assumptions on $\mathcal{Z} \subset \mathbb{N}$ and the ordering of the values are without loss of generality.

environment. Then, $\mathcal{Z}^* = \mathcal{Z}$ and $T(z) = T^*(z)$, $\forall z \in \mathcal{Z}$. In general, the mass of each group in the population is

$$q_g := \Pr(g).$$

Let q denote a vector that stacks all q_g values that are nonzero. In applications, there may be groups with $q_g = 0$. Hence, the dimension of q, d_q, is defined as the number of groups with nonzero mass, so $d_q \leq |\mathcal{G}|$.

For example, consider an environment with binary treatment, $k = 2$ instrument values and $\mathcal{Z} = \mathcal{Z}^*$. Using terminology in the literature (e.g, Angrist et al. (1996)), the 4 response groups in the original environment are always-takers (A) with $T(0) = T(1) = 1$, compliers (C) with $T(0) = 0$ and $T(1) = 1$, defiers (D) with $T(0) = 1$ and $T(1) = 0$ and never-takers (N) with $T(0) = T(1) = 0$. Then, $|\{((T(z))_{z\in\mathcal{Z}}, (T^*(z^*))_{z^*\in\mathcal{Z}^*})\}| = 2^{2\times 2} = 16$. If we are not interested in the extrapolated environment, then $T(z) = T^*(z)$, so we only have $d_q = 4$ groups.

Define the conditional mean for each group as follows:

$$\mu_{gt} := E[Y(t)|g].$$

Similarly, let μ be the vector that stacks the μ_{gt} values, and let $d_\mu := dim(\mu)$ denote the dimension of μ. It is implicitly assumed that these μ_{gt} objects are well-defined. When treatment is binary, $d_\mu = 2d_q$.

Following Huber et al. (2017), it suffices to have mean independence of the potential outcomes across groups instead of full independence:

Assumption 3.1. $E[Y(t)|g, z] = E[Y(t)|g]$ *and* $\Pr(g|z) = \Pr(g)$ *for all* g, z.

These groups are the primitives of the setup. Random assignment of the instrument Z satisfies Assumption 1. In addition to Assumption 1, following Angrist and Imbens (1994),

many papers also assume monotonicity, the assumption that the instrument weakly affects treatment in the same direction for all individuals.

Assumption 3.2. *For all* $z_1, z_2 \in \mathcal{Z}$ *either* $\Pr(T(z_1) \geq T(z_2)) = 1$ *or* $\Pr(T(z_1) \leq T(z_2)) = 1$. *For* $z_1^*, z_2^* \in \mathcal{Z}^*$, *either* $\Pr(T^*(z_1^*) \geq T^*(z_2^*)) = 1$ *or* $\Pr(T^*(z_1^*) \leq T^*(z_2^*)) = 1$.

Assumption 2 implies there are particular groups g with $q_g = 0$, which, in the environment without extrapolation, reduces number of treatment response types with nonzero mass from 2^k to $k+1$. This paper conducts sensitivity analysis for the failure of this assumption, so it relaxes Assumption 2. Since this assumption is a statement about the potential treatment response, sensitivity analysis involves careful consideration of the masses q_g of various groups.

The object of interest is the PRTP, defined as any estimand that can be written as:

$$\beta = \sum_{g,t} c_{gt}(q)\mu_{gt} = c(q)'\mu. \tag{3.1}$$

where $c_{gt}(q)$'s denote the weights on each of the μ_{gt}'s, and these coefficients can depend on q. The equality requires the object of interest to be linear in μ. $c(q)$ is the coefficient vector, with $c : [0,1]^{d_q} \to \mathbb{R}^{d_\mu}$ transforming the vector of proportions into weights on the conditional expectations. Once q is known, $c(q)$ is known. Objects of interest like the LATE and the average treatment effect (ATE) can be written in this form. For example, the ATE uses $c(q) = q \otimes (1,-1)'$, the average treatment effect on the treated (ATT) uses $c(q) = (q_A, -q_A, q_C, -q_C, q_D, -q_D, 0, 0)'/(q_A+q_C+q_D)$, and the LATE* is explained in Section 3.3. This β can be viewed as a discretized version of the PRTP defined in Mogstad et al. (2018).

The relationship between β and the target object in Mogstad et al. (2018) warrants further discussion. Mogstad et al. (2018) assumed monotonicity, so treatment can be written as $T = 1[\tilde{\nu}(Z) \geq u]$, for unobserved $u \sim U[0,1]$. The primitives of their model are marginal treatment responses $E[Y(t) \mid u]$, and their target parameter integrates a weighted average of

$E[Y(t) \mid u]$ over u. In the monotonic setting, u has a natural interpretation as a treatment propensity, where high values of u correspond to N, middle values to C, and and low values to A for a binary instrument. However, when monotonicity fails, the treatment equation becomes nonseparable with $T = 1[\nu(Z, u) \geq 0]$. Then, the interpretation of u is unclear unless a researcher has a particular $\nu(Z, u)$ in mind. Nonetheless, the groups remain well-defined in general. The unobserved u is meaningful in the target object insofar as it defines the groups that we are interested in. Hence, this paper uses the unobserved groups g to characterize conditional means, and characterizes the target object in terms of $E[Y(t) \mid g]$ instead of $E[Y(t) \mid u]$. The relationship between this group characterization and a nonseparable selection equation will be further clarified in Section 3.3.2 through an example.

3.2.2 Constraints on mu and q

The method places bounds on objects of interest by using the researcher's input for a sensitivity parameter. To explain this method, I first explain the constraints on μ implied by Assumption 1 in Section 3.2.2, where it is assumed that the vector q is known. Then, Section 3.2.2 shows how a single sensitivity parameter that affects q captures the extent that monotonicity is violated.

Constraint Set for μ

$\mathcal{M}(q)$ denotes the set of μ that satisfies defined equality and inequality constraints. These constraints may depend on q, and may include ex ante restrictions and features of the data. The researcher can specify what these constraints are, but I require these constraints to be linear in μ and the set $\mathcal{M}(q)$ to be convex.

One example of $\mathcal{M}(q)$ is a set of mean compatibility constraints implied by Assumption 1. In $Y|T = t, Z = z$, the mean of the various structural μ_{gt} such that $T(z) = t$, weighted by their proportions, is equal to the reduced-form mean $E[Y|T = t, Z = z]$. Hence, where

$p_{tz} := \Pr(T = t|Z = z)$, for all z, t, these constraints take the form:

$$\sum_{g:T(z)=t} q_g \mu_{gt} = p_{tz} E[Y|T = t, Z = z]. \tag{3.2}$$

Observe that (3.2) is a function of the q vector, so q parameterizes the constraint set $\mathcal{M}(q)$. Without ex ante restrictions, the set of μ that satisfies mean compatibility is in Equation (3.3). This set is denoted $\mathcal{M}_m(q)$ to avoid confusion with the general constraint set $\mathcal{M}(q)$:

$$\mathcal{M}_m(q) := \left\{ \mu \in \mathbb{R}^{d_\mu} : \sum_{g:T(z)=t} q_g \mu_{gt} = p_{tz} E[Y|T = t, Z = z] \qquad \forall z \in \mathcal{Z}, t \in \{0, 1\} \right\}. \tag{3.3}$$

The constraints in set $\mathcal{M}_m(q)$ do not exploit all distributional information, but nonetheless make the problem tractable, so $\mathcal{M}_m(q)$ can be used as a default. With binary outcomes, $\mu_{gt} \in [0, 1]$ should be used as a constraint. Without binary outcomes, we may consider additional constraints implied by Assumption 1, such as the trimming bounds of Lee (2009). Additional restrictions that the researcher may impose include selection into treatment (e.g., Roy (1951)).

Sensitivity Parameter

To form a sensitivity parameter for violation of Assumption 2, I first define compliers and defiers. For $z > z'$, define sets of defiers and compliers respectively as:

$$S^d_{(z,z')} := \{g : T(z) < T(z')\}, \text{ and}$$
$$S^c_{(z,z')} := \{g : T(z) > T(z')\}.$$

Since $\Pr(T(z) = 1)$ is increasing in z, $\Pr(g \in S^d_{(z,z')}) \leq \Pr(g \in S^c_{(z,z')})$. Assumption 2 is equivalent to having no defiers, so the sensitivity parameter should control the proportion

of defiers, which then affects the q vector. Hence, the sensitivity parameter λ imposes the restriction that, for all pairs (z, z'),

$$\sum_{g \in S^d_{(z,z')}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z')}} q_{g'}. \tag{3.4}$$

I refer to the inequality restriction (3.4) imposed by λ as a "sensitivity restriction".[4] I also place an analogous sensitivity restriction on the counterfactual environment with $T^*(.)$. In particular, for $S^{d*}_{(z,z')} := \{g : T^*(z) < T^*(z')\}$ and $S^{c*}_{(z,z')} := \{g : T^*(z) > T^*(z')\}$, the sensitivity restriction is $\sum_{g \in S^{d*}_{(z,z')}} q_g \leq \lambda \sum_{g' \in S^{c*}_{(z,z')}} q_{g'}$. It is possible to have a different sensitivity parameter for every pair (z, z') in nonbinary settings — this does not change the method, but increases the number of sensitivity parameters. To keep the exposition simple, I work with a single sensitivity parameter λ. When $\lambda = 0$, there is no pair of instrument values for which there are defiers.

With $\mathcal{Q}(\lambda)$ denoting the general constraint set, the proportion vector satisfies $q \in \mathcal{Q}(\lambda)$. As in the treatment of $\mathcal{M}(q)$, the researcher can specify additional restrictions, but I propose the minimal set of restrictions. Namely, the proportions chosen must be compatible with the observed p_{tz}. Assumption 1 implies $\forall t, z$,

$$\sum_{g: T(z)=t} q_g = p_{tz}. \tag{3.5}$$

The set $\mathcal{Q}(\lambda)$ may be empty for some choices of λ. Due to Proposition 1 of Noack (2021), there are bounds imposed on q_D by the data, so if λ is too small, the set will be empty. Notably, Noack (2021) assumes full independence rather than mean independence that is assumed in this paper, so if we assume full independence, tighter bounds on q_D

[4]This sensitivity parameter was earlier proposed in Ding and Lu (2017) for the case with a binary instrument and binary treatment when targeting subpopulations in the sample.

can be obtained from her method.[5] The sensitivity restriction thus describes monotonicity violations that are not detectable by the data. Even if $q_D = 0$ is rejected by the data, the existing tests can construct a confidence interval for q_D that can feature as a restriction in $\mathcal{Q}(\lambda)$, which can still be used to bound the PRTP.

The minimal constraint set satisfies (3.4) and (3.5), so it takes the form $\mathcal{Q}(\lambda) = \mathcal{Q}_m(\lambda)$:

$$\mathcal{Q}_m(\lambda) := \left\{ q \in [0,1]^{d_q} : \sum_{g \in S^d_{(z',z'')}} q_g \leq \lambda \sum_{g' \in S^c_{(z',z'')}} q_{g'}, \sum_{g \in S^{d*}_{(z',z'')}} q_g \leq \lambda \sum_{g' \in S^{c*}_{(z',z'')}} q_{g'}, \sum_{g:T(z)=t} q_g = p_{tz}, \forall (z',z''), t, z \right\}. \tag{3.6}$$

Observe that $\sum_g q_g = 1$ is implied by the condition that $\sum_{g:T(z)=t} q_g = p_{tz}, \forall t, z$. Following Mogstad et al. (2018), define our identified set for PRTP:

$$\mathcal{B}_\lambda = \{b \in \mathbb{R} : b = c(q)'\mu \text{ for some } \mu \in \mathcal{M}(q), q \in \mathcal{Q}(\lambda)\}. \tag{3.7}$$

More precisely, $\mathcal{B}_\lambda$ is the set identified by constraints in $\mathcal{M}$ and $\mathcal{Q}$.

Remark 3.1. *Due to the generality of the framework, several extensions can be accommodated. First, we can extend the analysis to multivalued treatments. With $|\mathcal{T}|$ treatment values, we can analogously define $\mathcal{G}$ so that $|\mathcal{G}| = |\mathcal{T}|^{2|\mathcal{Z}|}$. The object of interest remains as a linear combination of group-specific average potential outcomes. Second, we can extend the analysis to multiple binary instruments. With b binary variables, $|\mathcal{Z}| = 2$, and we have $|\mathcal{G}| = 2^{2|\mathcal{Z}|^b}$ groups. Then, we may conduct sensitivity analysis with respect to partial monotonicity (Mogstad et al., 2021) or limited monotonicity (van't Hoff et al., 2023) by imposing inequality (3.4) only with respect to their affected groups rather than all pairs.*

[5] Bounds on q_D are obtained from implications on the outcome distribution. With full independence, the entire outcome distribution can be used to obtain the bounds, but with mean independence, we can only use the conditional means of the outcome distribution.

3.2.3 Theoretical Properties

This subsection presents the main identification result of the paper, that the identified set is an interval. The method for finding bounds on PRTP solves an optimization problem in light of the constraints on μ and q from the previous subsection. Since obtaining the upper and lower bounds of the interval involves optimizing over μ and q, it is helpful to break the optimization problem into an inner problem that optimizes over μ given q and an outer problem that optimizes over q. Write the inner optimization problem as:

$$\underline{R}(q) := \min_{\mu \in \mathcal{M}(q)} c(q)'\mu, \text{ and } \qquad \overline{R}(q) := \max_{\mu \in \mathcal{M}(q)} c(q)'\mu. \tag{3.8}$$

These inner optimization problems are linear programs by assumption, given q. This rewriting is convenient because linear programs are computationally cheap. The linearity of the general program conditional on q is similar to the generic framework presented in Mogstad et al. (2018), which did not allow for monotonicity violations. Assumption 3.3 below provides sufficient conditions for the identified set to be an interval.

Assumption 3.3. *For a given $\lambda \in [0, 1)$, the following hold:*

(a) For all $g \in \mathcal{G}$, if $q_g > 0$, then μ_{gt} is well-defined and finite $\forall t \in \{0, 1\}$.

(b) Restrictions in $\mathcal{M}(q)$ can be written as a system of linear inequalities in μ such that $\mathcal{M}(q) = \{\mu : A(q)\mu \leq b(q)\}$ is continuous in $(A(q), b(q))$, and $\mathcal{M}(q)$ is convex.[6] Hyperparameters $A(q)$ and $b(q)$ of the linear program are continuous in q.

(c) $c(q)$ is continuous in q.

(d) $\mathcal{Q}(\lambda)$ is a nonempty convex set.

[6]The set $\mathcal{M}$ is continuous in (A, b) if it is lower and upper hemi-continuous in (A, b). In general, $\mathcal{M}$ is not lower hemi-continuous. For a counterexample, consider $t = (A, b)$ and $K(t) = \{x : Ax \leq b, x \geq 0\}$. The sequence $t_\nu = (A = \nu^{-1}, b = \nu^{-1})$ converges to $t^* = (0, 0)$. Observe that $K(t_\nu) = [0, 1]$. The point $2 \in K(t^*)$ cannot be reached by any sequence $\{x_\nu, \nu = 1, \cdots\}$ with $x_\nu \in [0, 1]$.

Theorem 3.1. *(Identified Set). Suppose Assumption 1 and 3 hold for some λ. Then, either $\mathcal{M}(q)$ is empty for all $q \in \mathcal{Q}(\lambda)$ and hence $\mathcal{B}_\lambda$ is empty, or the closure of $\mathcal{B}_\lambda$ is equal to the interval $[\underline{\beta}_\lambda, \overline{\beta}_\lambda]$, where*

$$\underline{\beta}_\lambda = \min_{q \in \mathcal{Q}(\lambda)} \underline{R}(q), \text{ and} \qquad \overline{\beta}_\lambda = \max_{q \in \mathcal{Q}(\lambda)} \overline{R}(q). \tag{3.9}$$

The theorem claims that the identified set is an interval, so every point in the interval is achievable by some $\mu \in \mathcal{M}(q), q \in \mathcal{Q}(\lambda)$. This property is not immediately obvious when optimizing over (q, μ): when we optimize over q, the objective function is potentially nonconvex, since $c(q)$ is nonlinear in q. Continuity of functions and convexity of sets are hence required for the result. Proof details are in Appendix 3.D. Notably, even if Assumption 3 fails, (3.9) still yields valid bounds, albeit conservative.

Generally, when using $\mathcal{M}_m(q)$ and $\mathcal{Q}_m(\lambda)$, the bounds are not sharp in that the (q, μ) pair that solves the problem need not be compatible with the data. The non-sharpness arises from two problems. The first problem is that not all $q \in \mathcal{Q}_m(\lambda)$ is compatible: for instance, it is known in the literature that there are tests for monotonicity (e.g., Richardson and Robins (2010); Kitagawa (2015); Huber et al. (2017); Noack (2021)), so $q_D = 0$ need not be compatible with the data. The second problem occurs because we have only used information on the means across distributions, and we have not yet exploited all distributional information. If outcomes are discrete, sharp bounds can be obtained by parameterizing the entire joint distribution of $(Y(0), Y(1), g)$, which is the approach taken by Balke and Pearl (1997). If the outcome is binary and all $q \in \mathcal{Q}_m$ are compatible, then we have sharp bounds, since $\mathcal{M}_m(q) \cup [0,1]^{d_\mu}$ contains all distributional information.

The sensitivity parameter also has a nice feature stated in Theorem 3.2, which reduces the number of inequality constraints.

Theorem 3.2. *Let* $z_0, z, z' \in \mathcal{Z}$. *If* $\sum_{g \in S^d_{(z_0,z_0+1)}} q_g \leq \lambda \sum_{g' \in S^c_{(z_0,z_0+1)}} q_{g'}$ *for all* $z_0 \in \mathcal{Z} \backslash \{k-1\}$, *then* $\sum_{g \in S^d_{(z,z')}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z')}} q_{g'}$ *for any instrument value pair* (z, z').

This theorem implies that we do not need to consider all instrument pairs — it suffices to consider adjacent instrument pairs. For intuition, when there are no defiers at both the $(z, z+1)$ and $(z+1, z+2)$ margins, it must be that there are no defiers at the $(z, z+2)$ margin, because the defiers at the $(z, z+2)$ margin must switch at either margin. In light of this result, we only have to check $k-1$ instead of $\binom{k}{2}$ constraints.

Remark 3.2. *The property in Theorem 3.2 is a feature of defining the sensitivity parameter in this way. If we had instead defined the sensitivity parameter as an upper bound on the proportion of defiers as done in Noack (2021), we no longer have this property. To see this, suppose we have three discrete instrument values* $\{0, 1, 2\}$. *Sensitivity parameter* η *is such that* $q_{(1,0,0)} + q_{(1,0,1)} \leq \eta^*$ *and* $q_{(0,1,0)} + q_{(1,1,0)} \leq \eta^*$ *at the* $(0, 1)$ *and* $(1, 2)$ *margin of the instrument respectively. In the worst case, we will have* $q_{(1,0,0)} = \eta^*$ *and* $q_{(1,1,0)} = \eta^*$. *Then, at the* $(0, 2)$ *margin,* $q_{(1,0,0)} + q_{(1,1,0)} = 2\eta^*$, *which is not bounded above by* η^*.

Remark 3.3. *Constructing the sensitivity restriction as* $q_D/q_C \leq \lambda$ *makes* λ *interpretable across applications. Suppose we have* $q_D = 0.01$ — *if* $q_C = 0.5$, *then the violation of monotonicity is relatively small; but if* $q_C = 0.02$, *the violation would be rather large.* λ *reflects the difference, despite having the same* q_D. *Nonetheless, if making an assumption on* q_D *directly instead of* q_D/q_C *is more interpretable in a particular application, a constraint of the form* $q_D \leq \lambda_D$ *can be used in* $\mathcal{Q}$ *without loss.*

3.2.4 Implementation

To implement the procedure proposed in the paper, we can simply use the sample analog. We observe data (Y_i, T_i, Z_i) for $i = 1, \cdots, n$. An implementable algorithm is:

1. Estimate probability objects p_{tz} by $\hat{p}_{tz} = \frac{\sum_{i=1}^n 1[T_i=t,Z_i=z]}{\sum_{i=1}^n 1[Z_i=z]}$. Use sample analog $\hat{E}[Y|T = t, Z = z] = \frac{1}{n_{tz}} \sum_{i:T_i=t,Z_i=z} Y_i$ for $E[Y|T = t, Z = z]$, $n_{tz} := \sum_{i=1}^n 1[T_i = t, Z_i = z]$.

2. For given $q \in \mathcal{Q}(\lambda)$,

   (a) Plug in $\hat{E}[Y|T = t, Z = z]$ and q into (3.2).

   (b) Set up the objective function and solve the linear program in (3.8). Output the value of the objective function $R(q)$.

3. For given λ, optimize output of Step 2 over q in the outer loop as in (3.9) using the sample analog.

Denote the estimators obtained from the sample $(\hat{\underline{\beta}}_\lambda, \hat{\overline{\beta}}_\lambda)$ for the lower and upper bounds respectively for the problem in (3.9). These estimators can be shown to be consistent by applying the Glivenko-Cantelli theorem to iid data, for instance, and applying the continuous mapping theorem after proving continuity in the program. Inference can be done by the projection method, and details are in Appendix 3.B. In empirical applications, the instrument may be valid only conditional on covariates, so Appendix 3.C extends the procedure to incorporate covariates.

While the above procedure suffices for the numerical results in this paper, as Section 3.3.1 shows how Step 3 can be reduced to a one-dimensional optimization problem, Step 3 may be unwieldy in general as the dimension of q can be large. To address this concern, Steps 2 and 3 can be combined into a bilinear program so we jointly optimize over (q, μ). Most objective functions considered can be written as linear fractionals of q, i.e., $c(q)'\mu = q'A\mu/d'q$, for some conformable matrix A and vector d, with $d'q > 0$ and linear constraints on (μ, q), say $Bq \leq b, C\mu \leq c$. Applying the Charnes-Cooper transformation by defining $t := 1/d'q, r := q/t$, the program is equivalent to optimizing $r'A\mu$ over (r, t, μ) such that $d'r = 1, Br \leq Bt, C\mu \leq c$. Then, standard algorithms for bilinear programs (Dutz et al., 2021; Shea, 2022) can be applied.

3.3 Identification of LATE*

The method in Section 3.2 is general, and allows partial identification of any combination of treatment response groups. Nonetheless, researchers often care about compliers. Hence, this section discusses and interprets LATE*, which is defined as the TE on compliers in counterfactual policy environments.[7]

For ease of exposition, I consider a binary instrument using the (A,C,D,N) notation as discussed in Section 3.2. The response groups in the counterfactual environment $\{A^*, C^*, D^*, N^*\}$ can be defined on $T^*(z)$ analogously. Using $G \in \{A, C, D, N\}$ and $G^* \in \{A^*, C^*, D^*, N^*\} =: \mathcal{G}_{cf}$ to denote response in the original and counterfactual environments respectively, $q_{GG^*} = \Pr(G, G^*)$ denotes the proportion who were G in the original environment and G^* in the new environment. Conditional probabilities are denoted $q_{G^*|G} := \Pr(G^*|G) = q_{GG^*}/(\sum_{H^* \in \mathcal{G}_{cf}} q_{GH^*})$. Using the definition that the LATE* is the TE for the counterfactual compliers, and $\mu_{GG^*t} := E[Y(t) \mid G, G^*]$,

$$LATE^* = \frac{\sum_G q_{GC^*} \left(\mu_{GC^*1} - \mu_{GC^*0}\right)}{\sum_G q_{GC^*}}.$$

The LATE* is useful for several reasons. First, the counterfactual environment could differ in place or time. Since the Angrist and Evans (1998) used US data, if we believe that the Canadian population is similar to the US, and its only difference is that it has better childcare, then the LATE* is what the LATE in Canada would be. For extrapolation over time, the study used 1990 data, but the current policy environment has changed since then, so the LATE* tells us what the LATE is now. Second, the LATE* is as useful to the policy

[7] LATE in Angrist and Imbens (1994) is defined under monotonicity as the TE for the subpopulation who respond (i.e., change their treatment status) to the instrument, which is equivalent to the TE on compliers (TEC). In the presence of defiers, the TE on the marginal population (TEM) and TEC are no longer equivalent. Since LATE was defined on a subpopulation with a particular treatment response status, it is sensible to define it as the TEC when there are defiers present. Hence, I define LATE* in the rest of this paper as the TEC in the counterfactual environment. We could also instead calculate TEM*, but I focus on LATE* to be concrete.

maker as the LATE. If LATE features in the policy function, then so must the LATE* once the policy is implemented because the environment would have changed. For example, if the policy maker wishes to give a $2000 subsidy in two tranches, once the first $1000 has been rolled out, the "LATE" would have changed, and we cannot expect the second $1000 to yield the same effect. This occurs because people no longer stick to their original groups. Such a setting is relevant when policy makers only have old studies or surveys available to inform current policy implementation. Third, the LATE* is useful in calibration. Parameter values in a model may be calibrated by using estimates from other studies. Then, the approach in this paper gives an explicit way of thinking about how the study at hand differs from the original study that the parameter value was calibrated from, and consequently the appropriate bounds on these values. Fourth, even though the LATE* is not point-identified, it is useful in policy choice when the social planner has a min/max objective function. The policy-maker can then choose policy rules by using the worst-case bounds obtained. Finally, since LATE* identifies the TE for a subpopulation, it is useful for assessing the robustness of conclusions on TE.

Since the object of interest is the LATE*, when considering policy changes that do not change the potential outcomes and unobservables, it suffices to characterize the proportions of original groups becoming C^* in the counterfactual environment. Hence, the counterfactual policy environment is characterized by the four extrapolation parameters $q_{C^*|G}$, denoting the proportion of the original groups switching into our group C^* of interest. Using this setup, LATE and ATE are special cases of the LATE*: LATE is the LATE* without extrapolation, and ATE is the LATE* when everyone switches into C^*.[8] Nonetheless, in many counterfactual policies of interest, such as increasing the instrument strength or increasing

[8]Observe that there is no gain in using sensitivity analysis for ATE, as observed by Kitagawa (2021), because the bounds are the widest when the proportion of defiers is the smallest.

treatment propensity, only $q_{C^*|N}$ and $q_{C^*|C}$ matter, as these counterfactual environments imply $q_{C^*|A} = 0$ and $q_{C^*|D} = 0$.[9] I provide two examples.

Example 3.1. *(Changing Instrument Value). In Duflo and Saez (2003), people were randomly given a letter that gave them \$20 if they attended the meeting, but they could have been given \$30 instead. This counterfactual corresponds to changing the instrument value (Z), say from 1 to 2. Researchers were interested in the effect of the meeting (T) on taking up a pension plan (Y). Here, $T^*(0) = T(0)$. The counterfactual compliers are those with $T^*(2) = 1, T^*(0) = 0$. Groups with $T(0) = 0$ are the original compliers and never-takers, so only C and N can become the counterfactual C^* group.*

Example 3.2. *(Changing Treatment Propensity). A policy may subsidize childcare in the Angrist and Evans (1998) context: regardless of a couple's gender preference, the probability of having a third child increases, i.e., $T^*(z) \geq T(z)$. Researchers were interested in the effect of a third child (T) on labor force participation (Y), and T is instrumented by first two kids having the same sex (Z). The counterfactual compliers are those with $T^*(1) = 1, T^*(0) = 0$. Since the policy weakly incentivizes treatment, individuals in C^* must have had $T(0) = 0$ in the original environment, which can only include the original C and N groups.*

While we have not seen people respond to the counterfactual incentives, we have seen people respond to other incentives. If we put bounds on the fraction of people who respond to the counterfactual environment but not the original, we can make progress. To bound such fractions, some economic reasoning is required for how the environment maps to the fraction: in Duflo and Saez (2003), we require a mapping from the financial incentive to fraction of

[9] Recent literature that deal with counterfactual environments as in Carneiro et al. (2010), Carneiro et al. (2011) and Mogstad et al. (2018) consider three counterfactual policies. These policy counterfactuals are in the class considered by Heckman and Vytlacil (2005), which involves policies that do not affect the marginal treatment response of T on Y. Their policy counterfactuals include (i) Additive α change in propensity score with the same instrument value (ii) Proportional $1+\alpha$ change in propensity score with same instrument (iii) Additive α shift of the jth component of Z, so $Z^* = Z + \alpha e_j$ and $p^*(x, z) = p(x, z)$. Changing the value of the instrument corresponds to policy type (iii) and monotonically changing the probability of being treated corresponds to (i) and (ii), so I group the first two together.

people changing their behavior; in Angrist and Evans (1998), we require a mapping from the subsidy amount to the fraction.

Remark 3.4. *(Relation to extrapolation in Marginal Treatment Effects framework). Without monotonicity, $T = 1[\nu(Z,u) \geq 0]$ for some $\nu(.)$. The counterfactual policies map to (1) Change the value of the instrument so $T^* = 1[\nu(Z^*,u) \geq 0]$; and (2) Change the threshold for everyone so $T^* = 1[\nu(Z,u) \geq -\alpha]$, increasing treatment propensity. I defer details to Section 3.3.2.*

The objective is hence $LATE^* = E[Y(1) - Y(0) | g \in \{CC^*, NC^*\}]$, which can be written as a linear function of μ_{GG^*}. The sensitivity restrictions may be constructed analogously, where λ restricts the proportion in both the original and counterfactual environments. For instance, when increasing the treatment propensity, the defier restrictions are:

$$\begin{aligned} q_{DD^*} + q_{DA^*} &\leq \lambda(q_{CC^*} + q_{CA^*}), \text{ and} \\ q_{DD^*} + q_{ND^*} &\leq \lambda(q_{CC^*} + q_{NC^*}). \end{aligned} \tag{3.10}$$

This problem can then be written in the form of the linear program in Section 3.2, which uses an inner linear program $R(q)$ that is cheap, and an outer problem that optimizes over q. The implementation for the threshold crossing counterfactual is explained in the next subsection; the implementation for changing the instrument value is analogous, and is explained in Appendix 3.A.1.

3.3.1 Treatment Propensity Implementation

To show how the framework of Section 3.2 applies, it suffices to specify the following: (i) the objective function (ii) what the groups g are (iii) linear restrictions for μ in the inner problem (iv) the constraint set for q in the outer optimization problem. Item (i) is LATE*, so the rest of this subsection explains the other items.

In our policy counterfactual, A will still be A^*. C can remain C^*, or they can become A^* when the policy is strong enough to shift their $Z = 0$ treatment to $T = 1$. The same argument applies to D. Finally, consider the N group. If the policy is weak, they would remain N^*. The policy may affect the outcome for only either $Z = 0$ or $Z = 1$, which changes their response behavior to D^* or C^*. The policy may also be strong enough to get the N group to $T = 1$ regardless of the instrument. Then, N can change their behavior to N^*, C^*, D^*, or A^*.

Although there are 9 response types, if the researcher does not wish to impose restrictions on q_{GG^*} that affects the sensitivity inequalities, we can essentially deal with 6 response groups $(A, CA^*, CC^*, D, NC^*, NC'^*)$, where NC'^* denotes the set of groups that switch from N to anything but C^* in the counterfactual policy environment, and D is the cell that collects all types who were defiers in the original environment. To be precise, define the following objects when there are 9 treatment response groups:

$$LATE^* = \frac{q_{CC^*}(\mu_{CC^*1} - \mu_{CC^*0}) + q_{NC^*}(\mu_{NC^*1} - \mu_{NC^*0})}{q_{CC^*} + q_{NC^*}},$$

$$q = (q_A, q_{CA^*}, q_{CC^*}, q_{DA^*}, q_{DA}, q_{NA^*}, q_{NC^*}, q_{ND^*}, q_{NN^*})',$$

$$\overline{R}^{TC}(q) := \max_{\mu \in \mathcal{M}_m^{TC}(q)} LATE^*, \text{ and}$$

$$\mathcal{M}_m^{TC}(q) := \left\{ \mu \in [0,1]^{18} : \sum_{g:T(z)=t} q_g \mu_{gt} = p_{tz} E[Y|T = t, Z = z] \qquad \forall z \in \{0,1\}, t \in \{0,1\} \right\}.$$

When there are 6 treatment response groups,

$$\tilde{R}(\tilde{q}) := \max_{\tilde{\mu} \in \tilde{\mathcal{M}}_m(\tilde{q})} LATE^*,$$

$$\tilde{q} := (q_A, q_{CC^*}, q_{CC'^*}, q_D, q_{NC^*}, q_{NC'^*})',$$

$$\tilde{\mu} := (\mu_{A1}, \mu_{A0}, \mu_{CC^*1}, \mu_{CC^*0}, \mu_{CC'^*1}, \mu_{CC'^*0}, \mu_{D1}, \mu_{D0}, \mu_{NC^*1}, \mu_{NC^*0}, \mu_{NC'^*1}, \mu_{NC'^*0})', \text{ and}$$

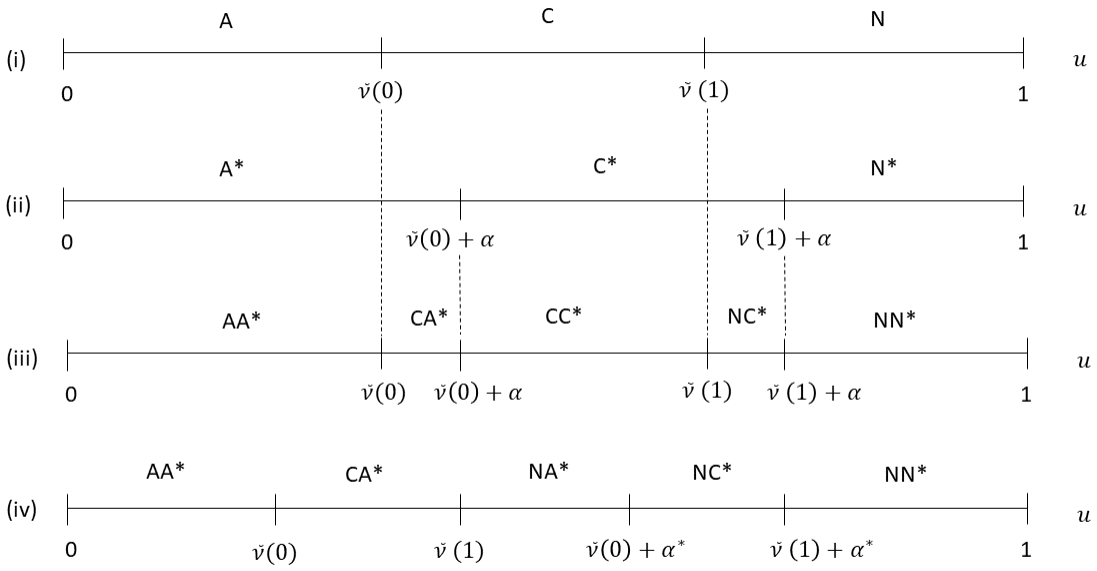
$$\tilde{\mathcal{M}}_m(\tilde{q}) = \left\{ \mu \in [0,1]^{12} : \sum_{g:T(z)=t} \tilde{q}_g \mu_{gt} = p_{tz} E[Y|T=t, Z=z] \qquad \forall z \in \{0,1\}, t \in \{0,1\} \right\}.$$

Proposition 3.1. *Consider* $q = (q_A, q_{CA^*}, q_{CC^*}, q_{DA^*}, q_{DA}, q_{NA^*}, q_{NC^*}, q_{ND^*}, q_{NN^*})'$. *If* $q_{CC'^*} = q_{CA^*}$, $q_D = q_{DA^*} + q_{DA}$, *and* $q_{NC'^*} = q_{NA^*} + q_{ND^*} + q_{NN^*}$, *then* $\overline{R}^{TC}(q) = \tilde{R}(\tilde{q})$.

Proposition 3.1 tells us that the bound for our object of interest does not change when we solve the 6 response group problem instead of the 9 response group problem, as long as we use the minimal constraint set $\mathcal{M}_m^{TC}(q)$ for μ. The proof proceeds by using the observation that $LATE^*$ is a function of $(q_{CC^*}, q_{NC^*}, \mu_{CC^*1}, \mu_{CC^*0}, \mu_{NC^*1}, \mu_{NC^*0})$. Then, it remains to argue that both optimization problems place the same restrictions on those parameters.

Finally, we can consider the constraint set on q. There are two restrictions in the form of (3.5); two restrictions based on the chosen $q_{C^*|C}, q_{C^*|N}$ as $q_{C^*|G} = q_{GC^*}/(\sum_{H^* \in \mathcal{G}_{cf}} q_{GH^*})$; and probabilities must sum to one. In addition to the five linear equality restrictions, sensitivity restrictions (3.10) must be satisfied. By using the linear equality restrictions, we only need to optimize over a single parameter in the outer problem with q. To see this result, there are 5 linearly independent restrictions involving q, and we can also write $q_D = q_D$ as a trivial relationship. Hence, using a system of 6 equations and 6 unknowns in q, for a given environment, once we know q_D, we know the rest of the q vector. Details are in Appendix 3.A.2. Consequently, bounds on LATE* can be obtained by solving a cheap linear program in μ in the inner loop with a one-dimensional optimization over q_D in the outer loop.

Figure 3.3.1: Separable Selection Equation

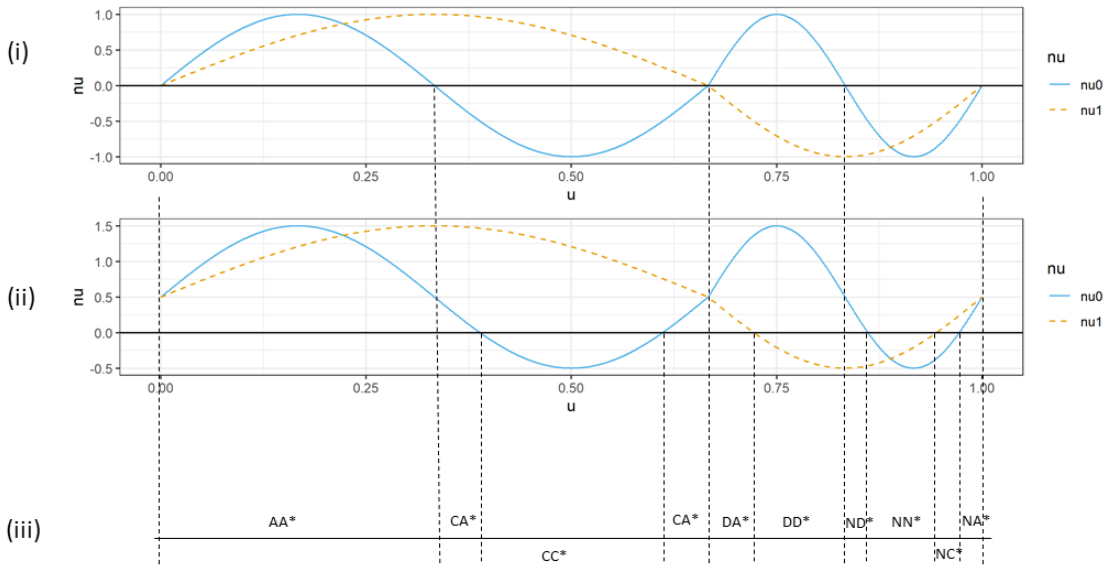


The next subsection gives examples of selection equations that justify treatment response groups. It can be skipped without loss of continuity.

3.3.2 Example of Selection Equations

In counterfactual environments, we could augment $\nu(Z,u)$ in Mogstad et al. (2018) to account for defiers, but it is difficult to do so without more structure on how defiers feature in the selection equation. Since characterizing the counterfactual environment based on groups is new, it is instructive to consider how this approach relates to selection equations. In particular, I show how selection equations under monotonicity with a binary instrument maps to groups in the counterfactual environment. I then use that intuition to explain what happens with a nonseparable selection equation. To begin, I consider the case without defiers, so the selection equation is given by $T = 1[\tilde{\nu}(Z) \geq u]$, where $u \sim U[0,1]$. Since Z is binary, $\tilde{\nu}(Z)$ can only take two values, and the environment is illustrated in Figure 3.3.1.

Figure 3.3.2: Nonseparable Selection Equation



In Figure 3.3.1, panel (i) illustrates the original environment, so low values of u are always-takers, those with middle values of u are compliers and those with high values of u are never-takers. Since u is uniformly distributed, $q_A = \tilde{\nu}(0), q_C = \tilde{\nu}(1) - \tilde{\nu}(0), q_N = 1 - \tilde{\nu}(1)$. In panel (ii), we have a counterfactual environment where the threshold is shifted by α such that $T^* = 1[\tilde{\nu}(Z) + \alpha \geq u]$. Consequently, the A^*, C^*, N^* groups are defined by the new cutoffs at $\tilde{\nu}(0) + \alpha$ and $\tilde{\nu}(1) + \alpha$. Panel (iii) combines the groups from panels (i) and (ii): for instance, the CA^* group are observations with $u \in [\tilde{\nu}(0), \tilde{\nu}(0) + \alpha]$, as they would have been compliers in the original environment, but always-takers in the new environment. With monotonicity, α has a natural interpretation in that propensity for treatment is increased by α. With the existing illustration, there is no NA^* group, because α is small. When α is large enough, we will have a scenario like panel (iv), where, by doing a similar analysis as before, an NA^* group exists, but we no longer have a CC^* group. A corollary is that,

under monotonicity, we can only either have NA^* or CC^*, but not both. In the empirical application, I have a relatively small α, so I have the CC^* group.

When monotonicity fails, we have a nonseparable selection equation $T = 1[\nu(Z,u) \geq 0]$, $u \sim U[0,1]$. One possible $\nu(.)$ function that can generate a nontrivial proportion of defiers (though not unique or interpretable) is as follows:

$$\nu(Z,u) = 1\left[u \leq \frac{2}{3}\right] sin\left(3u\pi - \frac{3}{2}Zu\pi\right) + 1\left[u > \frac{2}{3}\right] sin\left(6u\pi - 6\pi - Z\left(3u\pi - 3\pi\right)\right). \tag{3.11}$$

The $\nu(Z,u)$ is application-specific. The goal here is not to argue for the empirical relevance of any particular $\nu(Z,u)$, but to show that there exists such a function that rationalizes the group formulation. This function is more clearly illustrated in Figure 3.3.2 in panel (i). The solid nu0 line plots $\nu(0,u)$ while the dashed nu1 line plots $\nu(1,u)$. For $u < 1/3$, both the solid and dashed lines are above 0, so they form the A group. For $u \in [1/3, 2/3]$, only the dashed line is above zero, so they would be treated when $Z = 1$ and untreated when $Z = 0$, so they are the C group. By doing the same analysis, $u \in [2/3, 5/6]$ are the defiers and $u \in [5/6, 1]$ are the never-takers. Panel (ii) illustrates the counterfactual environment where $T^* = 1[\nu(Z,u) + \alpha \geq 0]$, which shifts the ν function up by $\alpha = 0.5$, but the shape remains unchanged. In this non-monotonic environment, α is less interpretable. By looking at the regions where the dashed and solid lines are above or below 0, we can work out the new A^*, C^*, D^*, N^* groups. Panel (iii) combines the old and new groups from the previous panels to illustrate the region of u values that form the 9 treatment response groups. Unlike the separable case, it is possible to generate all 9 groups simultaneously.

If the researcher has a selection function ν in mind, such as (3.11), then it is possible to analytically derive the intercepts of the relevant curves and hence the q vector. With q known, bounds can be obtained conveniently using the linear program. Instead of estimating $\nu(Z,u)$ or imposing additional assumptions on ν, the approach in this paper transparently makes assumptions on the q vector by using $q_{C^*|C}, q_{C^*|N}$ as extrapolation parameters.

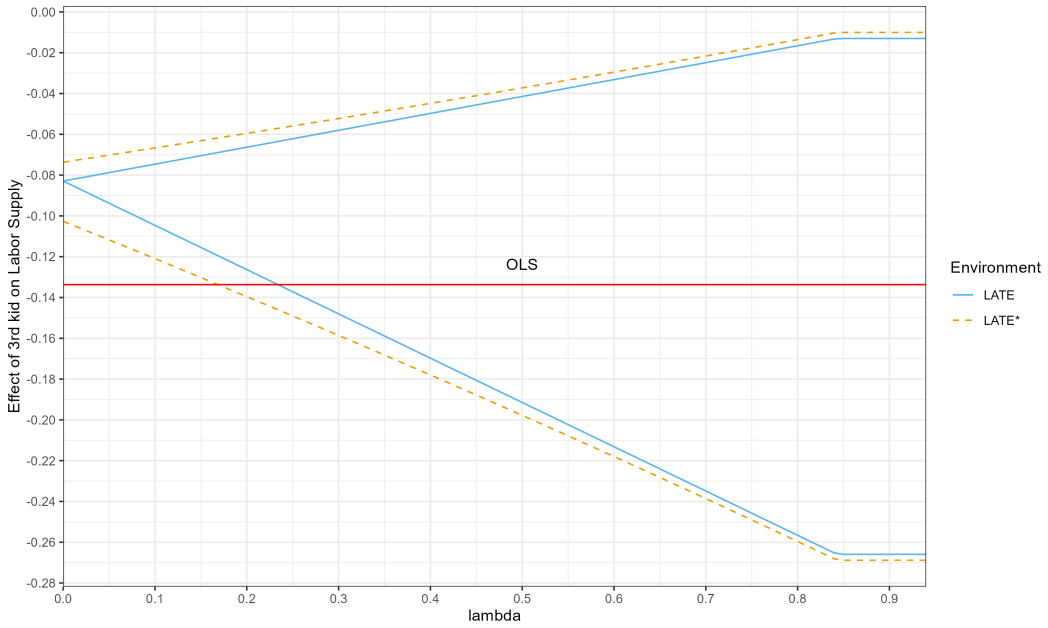
3.4 Empirical Application

In the Angrist and Evans (1998) problem, we are interested in the effect of a third child (T) on women's labor force participation (Y), and the instrument is whether the first two kids are of the same sex (Z). All variables are binary. Defiers are parents who have a preference for either two boys or two girls. Following Angrist and Evans (1998), I focus on the 1990 IPUMS data for mothers.[10] This empirical application illustrates how sensitivity analysis bridges the two extremes of monotonicity and worst-case bounds for LATE and LATE*, giving bounds at intermediate values of λ. The bounds vary continuously with λ, and are sensitive to failures of monotonicity. As a benchmark, De Chaisemartin (2017) argues that 5% of defiers is a conservative upper bound, which translates to $\lambda = 0.44$. Further, $q_D = 0$ is in the confidence interval constructed by Noack (2021).

We have $n = 380007$ observations and the proportions are given by $\widehat{\Pr}(Z = 1) = 0.504$, $\widehat{\Pr}(T = 1|Z = 1) = 0.402$ and $\widehat{\Pr}(T = 1|Z = 0) = 0.339$. Hence, the first stage is 0.063. Suppose we are interested in a counterfactual environment where there is a childcare subsidy that has a marginal effect on the probability of a third child. Here, $LATE^* = E[Y(1)-Y(0)|C^*]$ is the TE for people who used to be C and remain C^*, and people who were N but become C^* when there is a childcare subsidy. The units in the CC^* group have very strong preference for gender balance, and are hence unmoved by the subsidy, and the units in the NC^* group may be interpreted as those with weak preference for gender balance, but need a sufficient financial incentive to have a third child. Since existing papers (e.g., Carneiro et al. (2010)) calculate counterfactual effects at the margin i.e., $q_{C^*|C} \to 1, q_{C^*|N} \to 0$, I use $q_{C^*|C} = 0.99, q_{C^*|N} = 0.01$ to mimic their approach. This environment can also be interpreted as a 1% change in the relevant proportions.

[10] The baseline implementation follows their Table 5 where no additional covariates were included. The implementation with covariates follows their Table 8(2).

Figure 3.4.1: Plot of $LATE^* = E[Y(1) - Y(0)|C^*]$ bounds against λ without covariates



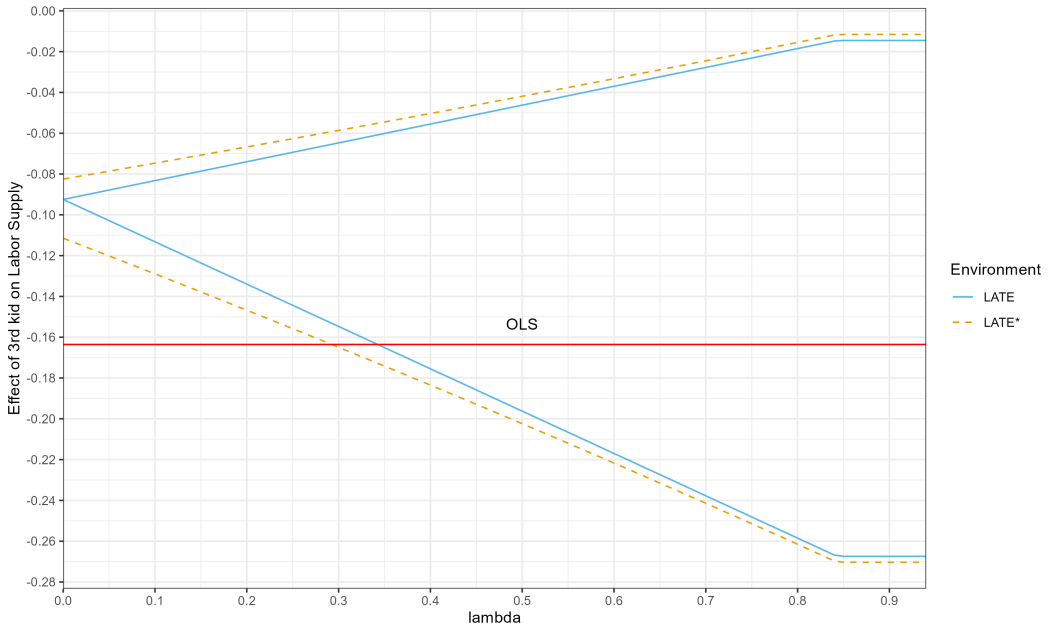
Impose $-0.3 \leq \mu_{g1} - \mu_{g0} \leq 0$ for all g. LATE has $q_{C^*|C} = 1, q_{C^*|N} = 0$ so there is no extrapolation; LATE* has $q_{C^*|C} = 0.99, q_{C^*|N} = 0.01$. The red horizontal line is the OLS benchmark of -0.134.

Figure 3.4.1 presents the main result for sensitivity analysis. I impose the condition that $-0.3 \leq \mu_{g1} - \mu_{g0} \leq 0$ for all g, which is reasonable when the researcher believes that the TE for all groups is negative, and the data informs us how negative this TE is. With the OLS benchmark of -0.134, TE of -0.3 (which is more than twice the OLS benchmark) is a conservative a priori lower bound. Since the intersection of convex sets is convex, the additional a priori restriction of $\mu_{g1} - \mu_{g0} \in [-0.3, 0]$ satisfies the conditions of Theorem 3.1.[11] Only estimated bounds are presented, and issues on inference are omitted.

The OLS estimate of -0.134 is a benchmark for how informative the bounds are. Instruments are used in this context because we believe that OLS is downward biased: there are unobservable characteristics where people who are more likely to have a third child are also those who are less likely to work. Since IV is used to correct this downward bias,

[11] $\mu_{g1} - \mu_{g0} \in [-0.3, 0]$ is the intersection of half planes, which is convex.

Figure 3.4.2: Plot of $LATE^* = E[Y(1) - Y(0)|C^*]$ bounds against λ with covariates



Impose $-0.3 \leq \mu_{g1} - \mu_{g0} \leq 0$ for all g. LATE has $q_{C^*|C} = 1, q_{C^*|N} = 0$ so there is no extrapolation; LATE* has $q_{C^*|C} = 0.99, q_{C^*|N} = 0.01$. The red horizontal line is the OLS benchmark of -0.164. Covariates include age of mother, age at first birth, gender of the first two kids, and race indicators for white, black, and hispanic.

when the lower bound of the identified set hits the OLS estimate, the procedure is no longer informative about correcting the downward bias.

The curve labeled LATE is the original policy environment (i.e., no extrapolation). At $\lambda = 0$, there is point identification, resulting in the original LATE of -0.083. It is evident here that, even without extrapolating, bounds can be very wide (and uninformative) when monotonicity does not hold, but sensitivity analysis allows us to obtain the intermediate points. The lower bound of the LATE is above OLS for $\lambda \leq 0.2$, but it becomes uninformative for $\lambda \geq 0.25$. Hence, we can conclude that the LATE bounds are informative only for small values of λ. In the special case where the minimal constraint set $\mathcal{M}_m$ is imposed, the LATE bounds are linear in λ, a result in Noack (2021).

The counterfactual environment labeled LATE* with $q_{A^*|C} = 0.01, q_{C^*|N} = 0.01$ incentivizes both groups into treatment. When q_{NC^*} is nonzero, the worst-case bounds of $\{-0.3, 0\}$

are imposed for $\mu_{NC^*1} - \mu_{NC^*0}$, and the bounds are no longer linear in λ. When monotonicity holds, the identified set is $[-0.103, -0.0737]$, which is informative; the bounds become uninformative for $\lambda \geq 0.2$. When we relax the sensitivity parameter, the upper bound eventually gets close to the trivial upper bound of 0. The numerical bounds on LATE* depend on the extrapolated environment: if we had extrapolated more, the LATE* at $\lambda = 0$ can be much wider than LATE at $\lambda = 0.1$. Hence, the bounds are informative only for small violations of monotonicity, and a counterfactual environment that differs locally.

The curve in Figure 3.4.2 uses the same set of covariates as in Angrist and Evans (1998). Implementing the procedure in Section 3.C yields the curve in Figure 3.4.2. The result is qualitatively similar to Figure 3.4.1, but the magnitudes differ when controls are included. When there is no extrapolation and monotonicity holds, we point identify the TSLS estimand from the original study. As we extrapolate the environment and allow λ to increase, we obtain bounds on the LATE* that widen. The λ required before the result is uninformative is also higher than the setting without covariates.

3.5 Conclusion

This paper shows how policy relevant treatment parameters, including LATE and LATE*, can be partially identified with a sensitivity parameter that controls the extent monotonicity fails. Identification uses assumptions on proportions of the population that have a particular response to the instrument instead of assumptions on the outcome function. This paper impacts empirical practice by providing a novel tool: sensitivity analysis of PRTP to failures of monotonicity, even for various treatment effects in extrapolated environments and when covariates are present. Depending on the empirical application, it may be more sensible to construct some structural model on selection $\nu(Z, u)$ (e.g., Chan et al. (2022)) instead of parameterizing the problem based on $E[Y(d)|g]$. Having a structural model is application-specific and left for future work.

Appendix

3.A Details on LATE*

3.A.1 Changing Instrument Value

Suppose we have a counterfactual instrument value Z^*. For illustration, I extrapolate the instrument rightward: in the original study, we have $Z \in \{0, 1\}$, but now we have $Z^* = 2$. The reasoning is similar if we wish to interpolate the instrument, or extrapolate leftward.[12] For every original group $G \in \{A, C, D, N\}$, individuals can have two possible responses at $Z^* = 2$, resulting in 8 treatment response groups, given by $(T(0), T(1), T^*(2))$.

LATE* is the TE for the compliers in the counterfactual environment. Since LATE is defined on an instrument pair, we have to consider which instrument pair the researcher is referring to. In the right-extrapolation exercise, one of the instrument values is $Z = 2$ that we do not have data for, so the LATE* can be defined either at the $(0, 2)$ pair or the $(1, 2)$ pair. In the Duflo and Saez (2003) running example, we ask what the LATE of the experiment would have been if we had given people \$30 instead of \$20. This corresponds to the $(0, 2)$ pair, because the control group still did not receive any financial incentive, and the treatment group simply received a larger incentive. Hence, $T^*(0) = T(0)$. In the right-extrapolation setup, compliers are those who switch their treatment status from

[12] Interpolation in Duflo and Saez (2003) with a \$20 incentive would ask what the LATE is if \$10 had been offered.

0 to 1 at the 0-2 instrument margin. This would be groups (0,0,1) and (0,1,1). Thus, $LATE^* = E[Y(1)-Y(0)|g \in \{(0,0,1),(0,1,1)\}]$. We can write this using the $c(q)'\mu$ notation for the objective function:[13]

$$LATE^* = E[Y(1) - Y(0)|g \in \{(0,0,1),(0,1,1)\}] = \sum_{g,t} c_{gt}(q)\mu_{gt} = E[Y(1) - Y(0)|C^*].$$

The four observed distributions will now each be a mixture of four extrapolated groups. Namely, in $T = 0, Z = 0$, we originally had N and C. When extrapolating rightward, the original N consists of $(0,0,0) = NN^*$ and $(0,0,1) = NC^*$ while the original C consists of $(0,1,1) = CC^*$ and $(0,1,0) = CN^*$. Groups such as NA^* cannot exist in this environment. Hence, the distribution $Y|T = 0, Z = 0$ now contains a mixture of four groups (0,0,0), (0,0,1), (0,1,1) and (0,1,0). Given q, $\mathcal{M}(q)$ is well defined by mean compatibility as before, specialized to our binary context, and using only the information from the original environment that we have data for:

$$\mathcal{M}_m^{Ex}(q) := \left\{ \mu \in [0,1]^{16} : \sum_{g:T(z)=t} q_g \mu_{gt} = p_{tz} E[Y|T = t, Z = z] \qquad \forall z \in \{0,1\}, t \in \{0,1\} \right\}. \tag{3.A.1}$$

It remains to consider what $\mathcal{Q}(\lambda)$ is with extrapolation. For the extrapolation parameter, observe that $q_{C^*|A} = q_{C^*|D} = 0$ by construction, so we only have to consider $q_{C^*|C}$ and $q_{C^*|N}$. To make the extrapolated environment comparable with and without monotonicity, we can set $q_{N^*|C} = 0$, so $q_{C^*|C} = 1$. This rules out the $(0,1,0)$ group, which has defiers at the 1-2 margin. Hence, the only extrapolation parameter is $q_{C^*|N} = \Pr((0,0,1)|N)$. At the 0-1 and

[13]The interpretation for the LATE* is the TE of the meeting for people who are somewhat sensitive to financial incentives: this group includes (0,0,1) who are less sensitive to the incentive than the original compliers who are (0,1,1). The coefficient takes the form:

$$c_{gt}(q) = \begin{cases} \frac{(-1)^{1-t} q_{(0,0,1)}}{q_{(0,0,1)} + q_{(0,1,1)}} & \text{if } g = (0,0,1) \\ \frac{(-1)^{1-t} q_{(0,1,1)}}{q_{(0,0,1)} + q_{(0,1,1)}} & \text{if } g = (0,1,1) \\ 0 & \text{otherwise} \end{cases}.$$

1-2 instrument margins, sensitivity restrictions are:

$$\begin{aligned} q_{(1,0,0)} + q_{(1,0,1)} &\leq \lambda(q_{(0,1,1)} + q_{(0,1,0)}), \text{ and} \\ q_{(0,1,0)} + q_{(1,1,0)} &\leq \lambda(q_{(0,0,1)} + q_{(1,0,1)}). \end{aligned} \tag{3.A.2}$$

Hence, the constraint set is:

$$\mathcal{Q}_m^{Ex}(\lambda; q_{C^*|N}) = \left\{ q : \text{Eq. (3.5) and (3.A.2)}, \frac{q_{(0,0,1)}}{q_{(0,0,1)} + q_{(0,0,0)}} = q_{C^*|N}, q_{(0,1,0)} = 0 \right\}. \tag{3.A.3}$$

Corollary 3.1. *Suppose μ_{gt} is finite for all g, t. Then, using $\mathcal{M}(q) = \mathcal{M}_m^*(q)$ and $\mathcal{Q}(\lambda) = \mathcal{Q}_m^{Ex}(\lambda)$, the identified set for the LATE* is an interval.*

Extrapolation is characterized by q, so the analysis here does not depend on the value of Z. Regardless of whether the counterfactual Z is 1.1 or 100, the same argument from extrapolating rightward applies. Instead, the approach parameterizes the extent of extrapolation by the q vector. Namely, $Z = 1.1$ is an environment that is very similar to the original policy, so we expect $q_{C^*|N}$ close to zero. In contrast, with $Z = 100$, or a very different propensity score, it is analogous to a large extrapolation with $q_{C^*|N}$ close to 1. For instance, this could be a monetary incentive, so having a large incentive would move all N into taking up treatment.

Remark 3.5. *When LATE* is defined as the TE on some subpopulation, we can use the same method to obtain TE on other subpopulations that are potentially more interesting. For instance, (1,0,1) is the group that are defiers at the \$0-\$20 margin, and compliers at the \$20-\$30 margin in Duflo and Saez (2003). Behavioral studies on fund raisers in Gneezy and Rustichini (2000) show such behavior exist, where giving a bit of financial incentive disincentivises intrinsic effort, but offering a large financial incentive increases their effort. LATE* answers: For people with such behavioral responses, what is their take-up rate of a pension plan?*

Remark 3.6. *(Overpowering Experiments). Having a large incentive, say \$100 in the Duflo and Saez (2003) experiment, can incentivize many people into treatment (the meeting). But this also includes people who go just for the money rather than because they are interested in the pension plan. If the incentive were \$5 instead, the LATE of information on taking up the pension plan is likely larger, since this excludes the people who are not interested in the plan. Exercise in extrapolation places bounds on what the results of the experiment would have been if it had been designed differently.*

3.A.2 Unified Econometric Approach

This section explains how the multi-dimensional optimization in the outer loop over the vector q in the two different counterfactual environments can be simplified into a one-dimensional optimization problem in the outer loop. The inner loop is then a function of this one-dimensional parameter, and solves a linear program. Hence, estimation of the bounds is tractable. The treatment propensity counterfactual is explained in Section 3.3.1.

There is an analogous result in the counterfactual that changes the instrument value. For right extrapolation, with $\mathcal{M}_m^{Ex}(q)$ as defined in (3.A.1), define:

$$\overline{R}^{Ex}(q) = \max_{\mu \in \mathcal{M}_m^{Ex}(q)} LATE^* = \max_{\mu \in \mathcal{M}_m^{Ex}(q)} \frac{q_{(0,0,1)}(\mu_{(0,0,1),1} - \mu_{(0,0,1),0}) + q_{(0,1,1)}(\mu_{(0,1,1),1} - \mu_{(0,1,1),0})}{q_{(0,0,1)} + q_{(0,1,1)}}.$$

Lemma 3.1. *Consider* $q = (q_{(0,0,0)}, q_{(0,0,1)}, q_{(0,1,0)}, q_{(0,1,1)}, q_{(1,0,0)}, q_{(1,0,1)}, q_{(1,1,0)}, q_{(1,1,1)})$. *If* $q_A = q_{(1,1,0)} + q_{(1,1,1)}$, $q_{CC^*} = q_{(0,1,1)}$, $q_{CC'^*} = q_{(0,1,0)}$, $q_D = q_{(1,0,0)} + q_{(1,0,1)}$, $q_{NC^*} = q_{(0,0,1)}$, *and* $q_{NC'^*} = q_{(0,0,0)}$, *then* $\overline{R}^{Ex}(q) = \tilde{R}(\tilde{q})$.

With Proposition 3.1 and Lemma 3.1 telling us that the inner loop of the two counterfactual programs can be solved using a 6-parameter problem, the main result of this section is:

Theorem 3.3. *With scalar q_D, there exists an invertible matrix J and vector $v(q_D)$ that are known functions of $(q_D, p_{tz}, q_{C^*|G})$ such that:*

$$\max_{q \in \mathcal{Q}_m^{TC}(\lambda)} \overline{R}^{TC}(q) = \max_{q_D \in \mathcal{Q}_d^{TC}(\lambda)} \tilde{R}(J^{-1}v(q_D)), \text{ and} \tag{3.A.4}$$

$$\max_{q \in \mathcal{Q}_m^{Ex}(\lambda)} \overline{R}^{Ex}(q) = \max_{q_D \in \mathcal{Q}_d^{Ex}(\lambda)} \tilde{R}(J^{-1}v(q_D)), \tag{3.A.5}$$

where

$$\mathcal{Q}_d^{TC}(\lambda) = \left\{ q_D \in [0,1] : q_D \leq \frac{\lambda(p_{00} + p_{11} - 1)}{1 - \lambda} \right\}, \text{ and} \tag{3.A.6}$$

$$\mathcal{Q}_d^{Ex}(\lambda) = \mathcal{Q}_d^{TC}(\lambda) \cap \left\{ q_D : \frac{1 - (p_{00} + p_{11} - q_D)(1 - q_{C^*|C}) - 2q_{C^*|C}}{-2 + q_{C^*|C}} \leq \lambda \left(q_D + \frac{-1 + p_{11} + q_{C^*|N} + q_D}{-2 + q_{C^*|N}} \right) \right\}. \tag{3.A.7}$$

The result for the minimum is analogous.

The upshot of Theorem 3.3 is that bounds on the object of interest such as $\max_{q \in \mathcal{Q}_m^{TC}(\lambda)} \overline{R}^{TC}(q)$ can be obtained by solving a one-dimensional optimization problem in q_D instead of a multi-dimensional problem. Observe that we are using the same $\tilde{R}$, J, and $v(q_D)$ in both problems, so the inner problem is econometrically identical. Further, $\tilde{R}(\tilde{q})$ is a linear program in $\tilde{\mu}$, so it can be solved efficiently. To prove this result, first use the previous two lemmas to obtain equivalence in the inner program. Then, observe that there are 5 linearly independent equality constraints in the $\tilde{q}$ problem, so once q_D is known, $\tilde{q} = J^{-1}v(q_D)$ is known. Their expressions are provided in the proof in Appendix 3.D. The remaining constraint set for q_D comes from sensitivity restrictions that have been set up differently.

3.B Inference

Using Theorem 3.3, there are six groups when using $\mathcal{M}_m$ and $\mathcal{Q}_m$: $(A, CA^*, CC^*, D, NC^*, NC'^*)$. Proportion restrictions on p_{tz} yield:

$$\begin{aligned} E[(q_{CC^*} + q_{CA^*} + q_{NC^*} + q_{NC^{*\prime}} - (1-T))(1-Z)] = 0, \text{ and} \\ E[(q_{CC^*} + q_{CA^*} + q_A - T)Z] = 0. \end{aligned} \tag{3.B.8}$$

Mean compatibility constraints are:

$$\begin{aligned} E\left[\left(\frac{q_{NC^*}\mu_{NC^*0} + q_{NC'^*}\mu_{NC'^*0} + q_{CC^*}\mu_{CC^*0} + q_{CA^*}\mu_{CA^*0}}{q_{NC^*} + q_{NC'^*} + q_{CC^*} + q_{CA^*}} - Y\right)(1-T)(1-Z)\right] = 0, \\ E\left[\left(\frac{q_A\mu_{A1} + q_D\mu_{D1}}{q_A + q_D} - Y\right)T(1-Z)\right] = 0, \\ E\left[\left(\frac{q_{NC^*}\mu_{NC^*0} + q_{NC'^*}\mu_{NC'^*0} + q_D\mu_{D0}}{q_{NC^*} + q_{NC'^*} + q_D} - Y\right)(1-T)Z\right] = 0, \text{ and} \\ E\left[\left(\frac{q_A\mu_{A1} + q_{CA^*}\mu_{CA^*1} + q_{CC^*}\mu_{CC^*1}}{q_A + q_{CA^*} + q_{CC^*}} - Y\right)TZ\right] = 0. \end{aligned} \tag{3.B.9}$$

Finally, there are inequality constraints imposed by a binary outcome, and further restrictions on the q's imposed by the sensitivity parameter:

$$\begin{aligned} 0 \leq \mu_{gt} \leq 1, \qquad 0 \leq q_g \leq 1, \qquad q_D \leq \lambda(q_{CC^*} + q_{CA^*}), \qquad \sum_g q_g = 1, \\ \frac{q_{CC^*}}{q_{CC^*} + q_{CA^*}} = q_{C^*|C}, \text{ and} \qquad \frac{q_{NC^*}}{q_{NC^*} + q_{NC'^*}} = q_{C^*|N}. \end{aligned} \tag{3.B.10}$$

In general, with moment equalities and inequalities, algorithms such as Andrews and Soares (2010) can be applied. In this application, uncertainty from the data only features in moment equalities of (3.B.8) and (3.B.9), so I proceed only with moment equalities.

Parameters are denoted $\theta := (q', \mu')'$. Let $m(\theta) = 0$ denote the moment conditions of (3.B.8) and (3.B.9), where $m(\theta)$ is the vector of expectations, and let $\hat{m}(\theta)$ be the sample ana-

log. Under standard CLT assumptions, $\sqrt{n}(\hat{m}(\theta)-m(\theta)) \xrightarrow{d} N(0,\Omega)$, where Ω is the variance covariance matrix for the moment conditions. Since $m(\theta)=0$, $T(\theta) := n\hat{m}(\theta)'\Omega^{-1}\hat{m}(\theta) \xrightarrow{d} \chi^2_6$ for the test statistic $T(\theta)$. The χ^2 distribution has 6 degrees of freedom because there are 6 moment conditions. We do not reject θ if $T(\theta) \leq \chi^2_6(1-\alpha) =: c_\alpha$ for a size α test, where c_α denotes the critical value. Since Ω can be consistently estimated, plug in the sample analog $\hat{\Omega}$ to use feasible test statistic $\hat{T}(\theta) := n\hat{m}(\theta)'\hat{\Omega}^{-1}\hat{m}(\theta) \xrightarrow{d} \chi^2_6$ for inference.

Finally, to calculate the upper bound for the confidence interval, solve the following problem:

$$\max_{\theta:=(q',\mu')'} c(q)'\mu \qquad s.t. \qquad \hat{T}(\theta) \leq c_\alpha, \text{ and } \theta \text{ satisfies Eq. (3.B.10).} \tag{3.B.11}$$

Calculating the lower bound is analogous. This problem corresponds to having a partially identified θ that is in confidence set C_θ, and we are interested in the confidence set (CS) of $g(\theta) = c(q)'\mu$, and in particular the extremum of the CS of $g(\theta)$. The procedure described here is identical to the projection method described in Dufour (1997) Section 5.2 for obtaining a CS for $g(\theta)$. These optimization problems can be implemented using canned packages.

3.C Extension to Incorporate Covariates

In many situations, the instrument is valid only conditional on covariates, and hence researchers may wish to incorporate covariates into their model. Covariates W feature in the model through Assumption 1, which would be: $E[Y(t)|g,z,W] = E[Y(t)|g,W]$ and $\Pr(g|z,W) = \Pr(g \mid W)$ for all g,z,W. There are at least two ways that they can be incorporated. One way mimics Noack (2021, appendix A3): we can run the aforementioned procedure at every covariate level w, then reweigh the bounds by the covariate masses. While this procedure yields more restrictions and hence tighter bounds, it is computationally intensive, requires the researcher to make an assumption on defier bounds and extrapolation

parameters for every covariate value, and does not nest the two-stage least squares (TSLS) estimand in general. It is also cumbersome when W is continuous. Without further assumptions, this is the only procedure available to the best of my knowledge.

Instead, I propose a second approach for the LATE*. With covariates and without extrapolation, researchers run the TSLS regression as a standard practice. Hence, a goal of the procedure is to nest TSLS with covariates as a special case without extrapolation and when monotonicity holds, and I provide conditions under which such a procedure is reasonable. This procedure allows some dependence of μ_{gt} and q_g on W, and augments the existing linear program.

Without extrapolation and with monotonicity, parametric assumptions are already required to interpret TSLS with heterogeneous treatment effects when there are covariates (e.g., Blandhol et al. (2022)). In particular, theory has developed around interpreting TSLS as some weighted average of LATE's (i.e., weighted average of treatment effect of compliers at different covariate values), but it is often not obvious why that particular weighting is the most interesting. To circumvent the issue of which weighted average of LATE's should be targeted, I consider the environment where the treatment effect for compliers is the same at all covariate values, motivating the assumption below.

To be clear on notation, linear regressions are run with a constant, and W does not include the intercept term. T and Z are binary. Assume the following:

Assumption 3.4. *(a) For $g \in \{CA^*, CC^*, NC^*, D\}$, $q_g = Pr(g|W_1) = Pr(g|W_2)$ for all W_1, W_2, while for $g \in \{NC'^*\}$, $q_g = \alpha_g^{int} + \alpha_g' W$. q_A can depend on W flexibly. $q \in [0,1]^{d_q}$.*

(b) $\mu_{Dt}(W) = \eta_{Dt} + \xi_D(W)$; for $g \in \{CA^*, CC^*, NC^*\}$, $\mu_{gt}(W) = \eta_{gt} + \xi_g' W$; for $g \in \{NC'^*\}$ $\mu_{gt}(W) = \eta_{gt} + \xi_{gt}' W$. Finally, $\mu_{At}(W)$ can depend on W flexibly.

(c) $E[Z|W] = \tilde{\xi}^{int} + \tilde{\xi}' W \in [0,1]$.

There are three parts to the assumption. Part (a) makes restrictions on the q vector; part (b) makes restriction on the μ vector; part (c) ensures that linear projections are interpretable as conditional expectations.

In Assumption 3.4(a), we cannot have q_{CA^*}, q_{CC^*}, q_D depend on W so that the TSLS estimand does not depend on W. When we are interested in the LATE*, we additionally cannot have q_{NC^*} depend on W so that the target object LATE* does not depend on W. $q_{NC'^*}$ is linear in W so that the conditional expectation of $Y|Z=0,T=0,W$ is quadratic in W. Other parametric forms may be possible, but the expression of the conditional expectation has to match accordingly. q_A is allowed to depend on W flexibly, as it is differenced out in the procedure.

The TE for $g \in \{CA^*, CC^*, NC^*, D\}$ must be constant for all covariate values so that the LATE* and the TSLS estimand do not depend on W. This requirement is denoted in Assumption 3.4(b) as having the same ξ_g for treated and untreated potential outcomes so that the treatment effect $\eta_{g1} - \eta_{g0}$ does not depend on W. Having the same TE is required even without extrapolation and with monotonicity so that TSLS identifies the unique LATE. The functional form in $\mu(W)$ is required in this paper's framework so that we can match coefficients and obtain a linear program. $\xi_D(W)$ can depend flexibly on W because it is not used in coefficient matching. In contrast, for $g \in \{CA^*, CC^*, NC^*\}$, $\mu_{gt}(W)$ is linear in W so that the conditional expectation of $Y|Z=0,T=0,W$ is quadratic in W. For $g \in \{NC'^*\}$, $\mu_{gt}(W)$ allows ξ_{gt} to vary by potential treatments, and its linearity is required for coefficient matching. No restriction is required for $\mu_{At}(W)$.

Once q_D and the extrapolation parameters $q_{C^*|C}$ and $q_{C^*|N}$ are fixed, the rest of the q vector and α's are point-identified. Details are in Appendix 3.D.3. With α and q point identified, and $p_{00}(W) := Pr[T=0|Z=0,W]$, the assumption implies:

$$p_{00}(W) = \alpha^{int}_{NC'^*} + \alpha'_{NC'^*} W + q_{NC^*} + q_{CC^*} + q_{CA^*},$$

and hence

$$p_{00}(W)E\left[Y|Z=0,T=0,W\right]=\alpha_{NC'*}^{int}\eta_{NC'*0}+q_{NC*}\eta_{NC*0}+q_{CC*}\eta_{CC*0}+q_{CA*}\eta_{CA*0}$$
$$+\left(\eta_{NC'*0}\alpha'_{NC'*}+\alpha_{NC'*}^{int}\xi'_{NC'*0}+q_{NC*}\xi'_{NC*}+q_{CC*}\xi'_{CC*}+q_{CA*}\xi'_{CA*}\right)W+\alpha'_{NC'*}W\xi'_{NC'*0}W. \tag{3.C.12}$$

The object of interest can then be written as:

$$LATE^{*}=\frac{1}{q_{CC*}+q_{NC*}}\left(q_{CC*}\eta_{CC*1}+q_{NC*}\eta_{NC*1}-q_{CC*}\eta_{CC*0}-q_{NC*}\eta_{NC*0}\right).$$

Then, the proposed algorithm for finding LATE* uses the following steps (S):

S1. Run TSLS regression with the full set of controls W to obtain the TSLS estimand β.

S2. Calculate the sample analogs of q and α based on the identification argument of Appendix 3.D.3 to construct $p_{00}(W)$. Using the partition on $T=0,Z=0$, run the regression of $p_{00}(W)Y$ on 1, W, and $(\alpha'_{NC'*}W)W$. Denote the intercept as γ_0.

S3. Set up the linear program, whose objective is LATE*, optimizing over parameters η and appropriate a priori linear restrictions. Additionally, use the following linear restrictions:

(a) $\beta\left(q_{CA*}+q_{CC*}-q_{D}\right)=q_{CC*}\left(\eta_{CC*1}-\eta_{CC*0}\right)+q_{CA*}\left(\eta_{CA*1}-\eta_{CA*0}\right)+q_{D}\left(\eta_{D0}-\eta_{D1}\right)$, and

(b) $\gamma_0=\alpha_{NC'*}^{int}\eta_{NC'*0}+q_{NC*}\eta_{NC*0}+q_{CC*}\eta_{CC*0}+q_{CA*}\eta_{CA*0}$.

To see how this procedure is reasonable, first observe that S1 and S2 merely calculates objects used in S3, so it suffices to motivate S3. When there is no extrapolation, the TSLS estimand without covariates is given by $[q_C(\eta_{C1}-\eta_{C0})+q_D(\eta_{D0}-\eta_{D1})]/(q_C-q_D)$. Since the assumptions are constructed such that the TSLS estimand does not depend on W, S3(a) uses an analogous expression for the TSLS that accommodates the counterfactual environment.

S3(b) is motivated by (3.C.12). By regressing the left-hand side on W and a quadratic term, the intercept term γ_0 must match the structural objects described.

Due to the constraint in S3(a), the proposed algorithm collapses exactly to TSLS without extrapolation and under monotonicity. We also use covariate information through S3(b).

3.D Proof of Results

3.D.1 Proofs for Section 3.2

Let b denote the target object, so for a given q, the set of feasible values in the inner problem is:

$$\mathcal{B}(q) = \{b \in \mathbb{R} : b = c(q)'\mu \text{ for some } \mu \in \mathcal{M}(q)\}. \tag{3.D.1}$$

Lemma 3.2. *Under Assumption 1, suppose that $\mathcal{M}(q)$ is convex for some fixed q. Then, either $\mathcal{M}(q)$ is empty and hence $\mathcal{B}(q)$ is empty, or the closure of $\mathcal{B}(q)$ is equal to the interval $[\underline{R}(q), \overline{R}(q)]$, defined in (3.8). Further, if $\mathcal{M}(q)$ can be written as a system of linear inequalities in μ, both optimization problems are linear programs.*

Proof of Lemma 3.2. Convex $\mathcal{M}(q)$ is either empty or nonempty. If $\mathcal{M}(q)$ is empty, then by definition $\mathcal{B}(q) = \emptyset$. Next, consider a nonempty $\mathcal{M}(q)$. Since a linear mapping of a convex set also yields a convex set, and $c(q)'\mu$ is a linear map of μ, it follows that $\mathcal{B}(q)$ is a convex set. Thus, any $b \in [\underline{R}(q), \overline{R}(q)]$ must also be in $\mathcal{B}(q)$. Proving that optimization problems are indeed linear programs is straightforward from its construction. The constraints are linear in μ and the objective function is a linear function of μ. □

Proof of Theorem 3.1. Proof for an empty $\mathcal{B}_\lambda$ is identical to the proof in Lemma 3.2. Only consider nonempty $\mathcal{M}(q)$. The objective is to show that any $b \in [\underline{\beta}_\lambda, \overline{\beta}_\lambda]$ is achievable for some $q \in \mathcal{Q}(\lambda)$. For this, I first show first show that $\overline{R}(q)$ and $\underline{R}(q)$ are continuous in q. Continuity of these objects can then be used to complete the argument.

Apply Theorem 2 from Wets (1985) that the objective value of a linear program is continuous in its hyperparameters. The sufficient condition for the theorem is that the feasible set in both the primal and dual linear programs are continuous in the hyperparameters. For the dual problem, it is assumed that $\mathcal{M}(q)$ is bounded by Assumption 3(a), so Corollary 11 from Wets (1985) implies that the feasible set of the dual problem is continuous in the hyperparameters. Turning to the primal problem, continuity in the hyperparameters is given by Assumption 3(b). The conditions for the Wets (1985) theorem is hence satisfied. Then, using Theorem 2 from Wets (1985), and the fact that the composition of continuous functions is continuous, with $c(q)$ continuous in q due to Assumption 3(c), $\underline{R}(q)$ and $\overline{R}(q)$ are continuous in q.

It remains to show that any $b \in [\underline{\beta}_\lambda, \overline{\beta}_\lambda]$ is achievable for some $q \in \mathcal{Q}(\lambda)$. Pick a point $q^0 \in \mathcal{Q}(\lambda)$ such that $\mathcal{M}(q^0)$ is nonempty. This is guaranteed to exist because because we work in the environment where $\exists q \in \mathcal{Q}(\lambda)$ s.t. $\mathcal{M}(q)$ is nonempty. Using Lemma 3.2, any $r \in [\underline{R}(q^0), \overline{R}(q^0)]$ can be satisfied by some $\mu \in \mathcal{M}(q^0)$. Since an analogous argument can be made for $[\underline{\beta}_\lambda, \underline{R}(q^0)]$, it suffices to show that for all $b \in [\overline{R}(q^0), \overline{\beta}_\lambda]$, there exists some $q \in \mathcal{Q}(\lambda)$ such that $b = \overline{R}(q)$. If $\overline{R}(q^0) = \overline{\beta}_\lambda$, the desired conclusion is immediate, so I focus on $\overline{R}(q^0) < \overline{\beta}_\lambda$.

Let $\overline{q}$ be the q that achieves $\overline{\beta}_\lambda$ i.e., $\overline{\beta}_\lambda = \overline{R}(\overline{q})$. With slight abuse of notation, let $[q^0, \overline{q}]$ denote the set of convex combinations on $\mathbb{R}^{d_q}$ between q^0 and $\overline{q}$, so it is a convex set. Since by Assumption 3(d) $\mathcal{Q}(\lambda)$ is convex, any $q \in [q^0, \overline{q}]$ must also lie in $\mathcal{Q}(\lambda)$ and is hence feasible. Since convex sets are connected, $[q^0, \overline{q}]$ is connected. Using the fact that the image of a connected set is connected for a continuous mapping, $\overline{R}([q^0, \overline{q}])$ is a connected set. Since $\overline{R}(q^0)$ and $\overline{R}(\overline{q})$ are both feasible, and $\overline{R}(.) \in \mathbb{R}$, $[\overline{R}(q^0), \overline{R}(\overline{q})] \subseteq \overline{R}([q^0, \overline{q}])$. Hence, $\exists q \in [q^0, \overline{q}] \subseteq \mathcal{Q}(\lambda)$ such that $\overline{R}(q) \in [\overline{R}(q^0), \overline{R}(\overline{q})]$.

□

The following lemma is used to prove Theorem 3.2.

Lemma 3.3. *Suppose that for any* $z, z' \in \mathcal{Z} \subset \mathbb{N}$, $z > z'$ *implies* $\Pr(T(z) = 1) \geq \Pr(T(z') = 1)$. *For any* $n_1, n_2 \in \mathbb{Z}_+$ *with* $n_2 > n_1$ *and* $z + n_2 \in \mathcal{Z}$, *if* $\sum_{g \in S^d_{(z,z+n_1)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,+n_1)}} q_{g'}$ *and* $\sum_{g \in S^d_{(z+n_1,z+n_2)}} q_g \leq \lambda \sum_{g' \in S^c_{(z+n_1,z+n_2)}} q_{g'}$, *then* $\sum_{g \in S^d_{(z,z+n_2)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,+n_2)}} q_{g'}$.

Proof of Lemma 3.3. The defiers at the $(z, z + n_2)$ margin switch exactly once: either at $(z, z + n_1)$ or $(z + n_1, z + n_2)$. Individuals who switch twice are either always takers or never takers when looking at the $(z, z + n_2)$ margin. It also means that they will be compliers at either one of the two margins and defiers at the other margin. This implies

$$S_2 := S^d_{(z,z+n_2)} \subset S^d_{(z,z+n_1)} \cup S^d_{(z+n_1,z+n_2)} =: S_1.$$

To be precise, S_2 consists of defiers who switch exactly once, and $S_1 \backslash S_2$ consists of defiers who switch twice, resulting in their being compliers at one margin.

Let $q(.)$ be the probability measure on sets. By assumption, $q(S^d_{(z,z+n_1)}) \leq \lambda q(S^c_{(z,z+n_1)})$ and $q(S^d_{(z+n_1,z+n_2)}) \leq \lambda q(S^c_{(z+1,z+n_2)})$. Due to binary treatment, the sets $S^d_{(z,z+n_1)}, S^d_{(z+n_1,z+n_2)}$ are disjoint. Similarly, the sets $S^c_{(z,z+n_1)}, S^c_{(z+n_1,z+n_2)}$ are also disjoint. Summing the inequalities,

$$q(S_1) = q(S^d_{(z,z+n_1)}) + q(S^d_{(z+n_1,z+n_2)}) \leq \lambda(q(S^c_{(z,z+n_1)}) + q(S^c_{(z+n_1,z+n_2)})).$$

Consider the set $S^c_{(z,z+n_1)} \cup S^c_{(z+n_1,z+n_2)}$. This set consists of compliers at the $(z, z + n_2)$ margin (which implies $S^c_{(z,z+n_2)}$ is a subset), and $S_1 \backslash S_2$. Namely, $S^c_{(z,z+n_1)} \cup S^c_{(z+n_1,z+n_2)} = (S_1 \backslash S_2) \cup S^c_{(z,z+n_2)}$. Observe that $S^c_{(z,z+n_2)}$ is the set of compliers who switch their treatment

status exactly once in the correct direction. Then, the summed inequality is:

$$\begin{aligned} q(S_2) + (q(S_1) - q(S_2)) &\leq \lambda(q(S^c_{(z,z+n_2)}) + q(S_1) - q(S_2)) \\ \Rightarrow q(S_2) &\leq \lambda q(S^c_{(z,z+n_2)}) - (1-\lambda)(q(S_1) - q(S_2)) \\ \Rightarrow q(S^d_{(z,z+n_2)}) &\leq \lambda q(S^c_{(z,z+n_2)}). \end{aligned}$$

□

Proof of Theorem 3.2. The condition of Lemma 3.3 is satisfied due to how $\mathcal{Z}$ is defined. For $z' > z$, we can write $z' = z + l$ with $l > 0$. Thus, it is sufficient to show that $\sum_{g \in S^d_{(z,z+l)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+l)}} q_{g'}$ for any $l \in \mathbb{Z}_+$.

Prove by induction. Apply Lemma 3.3, using $n_1 = 1, n_2 = 2$. Since $\sum_{g \in S^d_{(z,z+1)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+1)}} q_{g'}$ and $\sum_{g \in S^d_{(z+1,z+2)}} q_g \leq \lambda \sum_{g' \in S^c_{(z+1,z+2)}} q_{g'}$, obtain $\sum_{g \in S^d_{(z,z+2)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+2)}} q_{g'}$. Suppose $\sum_{g \in S^d_{(z,z+l)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+l)}} q_{g'}$, and we want to show $\sum_{g \in S^d_{(z,z+l+1)}} q_g \leq \lambda \sum_{g' \in S^c_{(z,z+l+1)}} q_{g'}$ due to adjacency. Apply Lemma 3.3 with $n_1 = l, n_2 = l+1$ to obtain the result. □

Proof of Proposition 3.1. The objective is to show $\max_{\mu \in \mathcal{M}^{TC}_m(q)} LATE^* = \max_{\tilde{\mu} \in \tilde{\mathcal{M}}_m(\tilde{q})} LATE^*$, where $LATE^*$ is a function of $(q_{CC^*}, q_{NC^*}, \mu_{CC^*1}, \mu_{CC^*0}, \mu_{NC^*1}, \mu_{NC^*0})$.

Let $h(\mu) = \mu_4 := (\mu_{CC^*1}, \mu_{CC^*0}, \mu_{NC^*1}, \mu_{NC^*0})'$ denote the function that extracts the subvector μ_4 from a higher-dimensional vector $\mu \in \mathbb{R}^{18}$. Then, since $LATE^*$ only contains μ_4, $\max_{\mu \in \mathcal{M}^{TC}_m(q)} LATE^* = \max_{\mu_4 \in \mathcal{M}^{TC}_4(q)} LATE^*$, where

$$\mathcal{M}^{TC}_4(q) = \left\{ \mu_4 : \mu_4 = h(\mu), \mu \in \mathcal{M}^{TC}_m(q) \right\}.$$

Let $\tilde{h}(.)$ similarly extract μ_4 from $\tilde{\mu} \in \mathbb{R}^{12}$. Then, $\max_{\mu \in \tilde{\mathcal{M}}_m(\tilde{q})} LATE^* = \max_{\mu_4 \in \tilde{\mathcal{M}}_4(\tilde{q})} LATE^*$:

$$\tilde{\mathcal{M}}_4(\tilde{q}) = \left\{ \mu_4 : \mu_4 = \tilde{h}(\mu), \mu \in \tilde{\mathcal{M}}_m(\tilde{q}) \right\}.$$

Hence, it is sufficient to show that $\tilde{\mathcal{M}}_4(\tilde{q}) = \mathcal{M}_4^{TC}(q)$ to obtain the result. Do change of variables for $\mathcal{M}_m^{TC}(q)$, with the given substitution for q. Then, the respective μ's can be redefined:

$$\begin{aligned}
\mu_{NC'^*t} &= \frac{1}{q_{NC'^*}}[q_{NA^*}\mu_{NA^*t} + q_{ND^*}\mu_{ND^*t} + q_{NN^*}\mu_{NN^*t}], \\
\mu_{Dt} &= \frac{1}{q_D}[q_{DA^*}\mu_{DA^*t} + q_{DD^*}\mu_{DD^*t}], \text{ and} \\
\mu_{CC'^*t} &= \mu_{CA^*t}.
\end{aligned}$$

Then, equality constraints characterized by $\sum_{G:T(z)=t} q_G \mu_{Gt} = p_{tz} E[Y|T = t, Z = z]$ are identical to those of $\tilde{\mathcal{M}}_m(\tilde{q})$. Since the counterfactual μ's are weighted averages of the original μ's, the counterfactual μ's in $\tilde{\mu}$ must also lie in $[0, 1]$, so $\mathcal{M}_4^{TC}(q) \subseteq \tilde{\mathcal{M}}_4(\tilde{q})$. Then, it is sufficient to show $\tilde{\mathcal{M}}_4(\tilde{q}) \backslash \mathcal{M}_4^{TC}(q) = \emptyset$. The set $\tilde{\mathcal{M}}_4(\tilde{q}) \backslash \mathcal{M}_4^{TC}(q)$ contains values of μ_4 where the μ's in $\tilde{\mu}$ are in $[0, 1]$, but the individual components that construct the averages, such as μ_{DD^*t} need not be in $[0, 1]$. However, restrictions on μ_4 only occur through the averages in the equality constraints, in addition to $\mu_4 \in [0, 1]^4$. Thus, since the averages in $\tilde{\mathcal{M}}_4(\tilde{q})$ and in $\mathcal{M}_4^{TC}(q)$ face the same constraints, μ_4 face the same constraints in both sets. Hence, $\tilde{\mathcal{M}}_4(\tilde{q}) \backslash \mathcal{M}_4^{TC}(q) = \emptyset$, which then implies $\tilde{\mathcal{M}}_4(\tilde{q}) = \mathcal{M}_4^{TC}(q)$. □

3.D.2 Proofs for Appendix 3.A

Proof of Corollary 3.1. The condition satisfies Assumption 3(a). It is sufficient to check other conditions of Assumption 3, then apply Theorem 3.1. Continuity of $c(q)$ is immediate. $\mathcal{M}_m^{Ex}(q)$ is convex because it is intersection of linear subspaces. To see that $\mathcal{Q}(\lambda) = \mathcal{Q}_m^{Ex}(\lambda)$ satisfies convexity, take any two elements $q^0, q^1 \in \mathcal{Q}_m^{Ex}(\lambda)$ with $q^0 \neq q^1$. Form convex combination $q^* = \alpha q^0 + (1 - \alpha) q^1$, with $\alpha \in (0, 1)$. Taking the weighted sums of the constraints on q^0 and q^1, $q^* \in \mathcal{Q}_m^{Ex}(\lambda)$ is immediate. □

Proof of Lemma 3.1. By redefining groups as stated, the proof is analogous to Proposition 3.1. □

Proof of Theorem 3.3. By defining $\tilde{q}$ appropriately and applying Proposition 3.1 and Lemma 3.1, $\overline{R}^{TC}(q) = \tilde{R}(\tilde{q})$ and $\overline{R}^{Ex}(q) = \tilde{R}(\tilde{q})$.

Then, consider equality restrictions in $\mathcal{Q}_m^{TC}(\lambda)$ and $\mathcal{Q}_m^{Ex}(\lambda)$. In both constraint sets, there are 5 linearly independent restrictions, with 2 from p_{11} and p_{00}, 1 from the fact that probabilities sum to 1, and 2 from the extrapolation parameters. We can also write $q_D = q_D$ as a trivial relationship. Writing these 6 equations in matrix form, we have $J\tilde{q} = v(q_D)$, where

$$J = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & q_{C^*|C} - 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & q_{C^*|N} - 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \text{ and}$$

$$v(q_D) = (p_{00}, p_{11}, 1, q_{C^*|C}, q_{C^*|N}, q_D)'.$$

Note that $det(J) = 4 - 2q_{C^*|C} - 2q_{C^*|N} + q_{C^*|C}q_{C^*|N} = (2 - q_{C^*|N})(2 - q_{C^*|C}) \neq 0$. Since J is invertible, $\tilde{q} = J^{-1}v(q_D)$.

It remains to consider the inequality restrictions imposed by sensitivity parameters. In $\mathcal{M}_m^{Ex}(q)$, the specific choice of q_{DD^*} and q_{ND^*} makes restrictions on q_{CC^*} and q_{NC^*}. Since q_{CC^*} and q_{NC^*} are arguments in the optimization problem, the optimum is found by using the least restrictive setting for q_{CC^*} and q_{NC^*}. Due to Proposition 3.1, setting $q_{DD^*} = q_{ND^*} = 0$ is innocuous. Then, $q_{DD^*} + q_{ND^*} \leq \lambda(q_{CC^*} + q_{NC^*})$ is automatically satisfied. The only relevant sensitivity restriction is $q_D = q_{DD^*} + q_{DA^*} \leq \lambda(q_{CC^*} + q_{CA^*})$. Using the substitution

in Proposition 3.1, and $\tilde{q} = J^{-1}v(q_D)$, the constraint set results. This constraint set will give us the optimum, because the objective value has to perform weakly better with one fewer constraint.

Turning now to $\mathcal{Q}_m^{Ex}(\lambda)$, use the least restrictive values as before. One can set $q_{(1,0,0)} = 0$ so $q_D = q_{(1,0,1)}$, and upper bound at the 0-2 margin is the largest possible, while the inner problem does not change. Then, the constraint at the 0-1 margin will be $q_D \leq \frac{\lambda(p_{00}+p_{11}-1)}{1-\lambda}$ by using the relevant substitutions and $\tilde{q} = J^{-1}v(q_D)$. With the relevant substitutions, and setting $q_{(1,1,0)} = 0$ (to create the most flexible constraint), the constraint at the 1-2 margin is $q_{CC'^*} \leq \lambda(q_{NC^*} + q_D)$. Finally, substitute $\tilde{q} = J^{-1}v(q_D)$ to obtain the required inequality. □

3.D.3 Derivations for Appendix 3.C

$$
\begin{aligned}
&p_{00}(W)E\left[Y|Z=0,T=0,W\right] \\
&\quad = \left(\alpha_{NC'^*}^{int} + \alpha'_{NC'^*}W\right)\left(\eta_{NC'^*0} + \xi'_{NC'^*0}W\right) + q_{NC^*}\left(\eta_{NC^*0} + \xi'_{NC^*}W\right) \\
&\qquad + q_{CC^*}\left(\eta_{CC^*0} + \xi'_{CC^*}W\right) + q_{CA^*}\left(\eta_{CA^*0} + \xi'_{CA^*}W\right) \\
&\quad = \alpha_{NC'^*}^{int}\eta_{NC'^*0} + q_{NC^*}\eta_{NC^*0} + q_{CC^*}\eta_{CC^*0} + q_{CA^*}\eta_{CA^*0} \\
&\qquad + \left(\eta_{NC'^*0}\alpha'_{NC'^*} + \alpha_{NC'^*}^{int}\xi'_{NC'^*0} + q_{NC^*}\xi'_{NC^*} + q_{CC^*}\xi'_{CC^*} + q_{CA^*}\xi'_{CA^*}\right)W \\
&\qquad + \alpha'_{NC'^*}W\xi'_{NC'^*0}W
\end{aligned}
$$

First Stage Identification. Let the first-stage regression be $T = \pi Z + \theta_0 + \theta' W + v$. Then, the first stage estimand is:

$$
\begin{aligned}
\pi &= \frac{E\left[T\left(Z - E^*\left[Z|W\right]\right)\right]}{E\left[\left(Z - E^*\left[Z|W\right]\right)^2\right]} \\
&= \frac{E\left[E\left[Z|W\right]\left(1 - E\left[Z|W\right]\right)\left(E\left[T|Z=1,W\right] - E\left[T|Z=0,W\right]\right)\right]}{E\left[E\left[Z|W\right]\left(1 - E\left[Z|W\right]\right)\right]}.
\end{aligned}
$$

Using Assumption 3.4(a),

$$E[T|Z=1,W]-E[T|Z=0,W]=q_A(W)+q_{CA^*}+q_{CC^*}-q_D-q_A(W)$$
$$=q_{CA^*}+q_{CC^*}-q_D=\pi.$$

For a given q_D and an extrapolation parameter $q_{C^*|C}$, since π is identified, q_{CA^*}, q_{CC^*}, q_D are all identified. Since $E[T|Z=0,W]-q_D=q_A(W)=\alpha_A^{int}+\alpha_A'W$, by regressing $T-q_D$ in the partition with $Z=0$ on W, α_A^{int} and α_A are identified. Conversely,

$$1-E[T|Z=1,W]=q_{NC^*}+q_{NC'^*}(W)+q_D$$
$$1-E[T|Z=1,W]-q_{NC^*}-q_D, \text{ and } =q_{NC'^*}(W)=\alpha_{NC'^*}^{int}+\alpha_{NC'^*}'W.$$

Observe that $1-E[T|Z=1]=q_{NC^*}+E[q_{NC'^*}(W)]+q_D$. Since q_D and $q_{C^*|N}$ are known, q_{NC^*} is identified. Then, by regressing $1-T-q_{NC^*}-q_D$ in the partition with $Z=1$ on W, $\alpha_{NC'^*}^{int}$ and $\alpha_{NC'^*}$ are identified.

TSLS Estimand. Due to Assumption 3.4(c), the TSLS estimand is given by

$$\beta=\frac{E[Y(Z-E^*[Z|W])]}{E[T(Z-E^*[Z|W])]}$$
$$=\frac{E[E[Z|W](1-E[Z|W])(E[Y|Z=1,W]-E[Y|Z=0,W])]}{E[E[Z|W](1-E[Z|W])(E[T|Z=1,W]-E[T|Z=0,W])]}.$$

Then, due to Assumptions 3.4(a) and 3.4(b),

$$
\begin{aligned}
&E\left[Y|Z=1,W\right]-E\left[Y|Z=0,W\right] \\
&\quad=q_{CC^*}\left(\mu_{CC^*1}(W)-\mu_{CC^*0}(W)\right)+q_{CA^*}\left(\mu_{CA^*1}(W)-\mu_{CA^*0}(W)\right) \\
&\qquad+q_D\left(\mu_{D0}(W)-\mu_{D1}(W)\right) \\
&\quad=q_{CC^*}\left(\eta_{CC^*1}-\eta_{CC^*0}\right)+q_{CA^*}\left(\eta_{CA^*1}-\eta_{CA^*0}\right)+q_D\left(\eta_{D0}-\eta_{D1}\right), \text{ and} \\
\beta&=\frac{q_{CC^*}\left(\eta_{CC^*1}-\eta_{CC^*0}\right)+q_{CA^*}\left(\eta_{CA^*1}-\eta_{CA^*0}\right)+q_D\left(\eta_{D0}-\eta_{D1}\right)}{q_{CA^*}+q_{CC^*}-q_D}.
\end{aligned}
$$

Bibliography

Agan, Amanda, Jennifer L Doleac, and Anna Harvey, "Misdemeanor prosecution," *The Quarterly Journal of Economics*, 2023, *138* (3), 1453–1505.

Aizer, Anna and Joseph J Doyle Jr, "Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges," *The Quarterly Journal of Economics*, 2015, *130* (2), 759–803.

Anatolyev, Stanislav and Mikkel Sølvsten, "Testing many restrictions under heteroskedasticity," *Journal of Econometrics*, 2023, *236* (1), 105473.

Anderson, Theodore W and Herman Rubin, "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of mathematical statistics*, 1949, *20* (1), 46–63.

Andrews, Donald WK and Gustavo Soares, "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 2010, *78* (1), 119–157.

_ **, Marcelo J Moreira, and James H Stock**, "Optimal two-sided invariant similar tests for instrumental variables regression," *Econometrica*, 2006, *74* (3), 715–752.

_ **, Vadim Marmer, and Zhengfei Yu**, "On optimal inference in the linear IV model," *Quantitative Economics*, 2019, *10* (2), 457–485.

Andrews, Isaiah, "Conditional linear combination tests for weakly identified models," *Econometrica*, 2016, *84* (6), 2155–2182.

Angrist, Joshua D and Alan B Krueger, "Does compulsory school attendance affect schooling and earnings?," *The Quarterly Journal of Economics*, 1991, *106* (4), 979–1014.

_ **and Guido W Imbens**, "Identification and estimation of local average treatment effects," *Econometrica*, 1994, *62* (2), 467–475.

_ **and William N Evans**, "Children and their parents' labor supply: Evidence from exogenous variation in family size," *American Economic Review*, 1998, pp. 450–477.

_ **, Guido W Imbens, and Donald B Rubin**, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 1996, *91* (434), 444–455.

Autor, David, Andreas Kostøl, Magne Mogstad, and Bradley Setzler, "Disability benefits, consumption insurance, and household labor supply," *American Economic Review*, 2019, *109* (7), 2613–2654.

Balke, Alexander and Judea Pearl, "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 1997, *92* (439), 1171–1176.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky, "When is TSLS actually late?," Technical Report, National Bureau of Economic Research 2022.

Boot, Tom and Didier Nibbering, "Inference on LATEs with covariates," *arXiv preprint arXiv:2402.12607*, 2024.

Braun, Martin and Valentin Verdier, "Estimation of spillover effects with matched data or longitudinal network data," *Journal of Econometrics*, 2023, *233* (2), 689–714.

Brinch, Christian N, Magne Mogstad, and Matthew Wiswall, "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 2017, *125* (4), 985–1039.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller, "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 2011, *29* (2), 238–249.

Carneiro, Pedro, James J Heckman, and Edward J Vytlacil, "Estimating marginal returns to education," *American Economic Review*, 2011, *101* (6), 2754–81.

_ , _ , and Edward Vytlacil, "Evaluating marginal policy changes and the average effect of treatment for individuals at the margin," *Econometrica*, 2010, *78* (1), 377–394.

Cattaneo, Matias D, Michael Jansson, and Whitney K Newey, "Inference in linear regression models with many covariates and heteroscedasticity," *Journal of the American Statistical Association*, 2018, *113* (523), 1350–1361.

Chaisemartin, Clement De, "Tolerating defiance? Local average treatment effects without monotonicity," *Quantitative Economics*, 2017, *8* (2), 367–396.

Chan, David C, Matthew Gentzkow, and Chuan Yu, "Selection with variation in diagnostic skill: Evidence from radiologists," *The Quarterly Journal of Economics*, 2022, *137* (2), 729–783.

Chandrasekhar, Arun G and Matthew O Jackson, "A network formation model based on subgraphs," *arXiv preprint arXiv:1611.07658*, 2016.

Chao, John C, Norman R Swanson, and Tiemen Woutersen, "Jackknife estimation of a cluster-sample IV regression model with many weak instruments," *Journal of Econometrics*, 2023, *235* (2), 1747–1769.

_ , _ , **Jerry A Hausman, Whitney K Newey, and Tiemen Woutersen**, "Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments," *Econometric Theory*, 2012, *28* (1), 42–86.

Chen, Louis HY and Qi-Man Shao, "Normal approximation under local dependence," *The Annals of Probability*, 2004, *32* (3), 1985–2028.

Chiang, Harold D and Yuya Sasaki, "On Using The Two-Way Cluster-Robust Standard Errors," *arXiv preprint arXiv:2301.13775*, 2023.

_ , **Bruce E Hansen, and Yuya Sasaki**, "Standard errors for two-way clustering with serially correlated time effects," *Review of Economics and Statistics*, 2024, pp. 1–40.

Chin, Alex, "Central limit theorems via Stein's method for randomized experiments under interference," *arXiv preprint arXiv:1804.03105*, 2018.

Crane, Harry and Henry Towsner, "Relatively exchangeable structures," *The Journal of Symbolic Logic*, 2018, *83* (2), 416–442.

Crudu, Federico, Giovanni Mellace, and Zsolt Sándor, "Inference in instrumental variable models with heteroskedasticity and many instruments," *Econometric Theory*, 2021, *37* (2), 281–310.

Dahl, Christian M, Martin Huber, and Giovanni Mellace, "It is never too LATE: a new look at local average treatment effects with or without defiers," *The Econometrics Journal*, 2023, *26* (3), 378–404.

Davezies, Laurent, Xavier D'Haultfœuille, and Yannick Guyonvarch, "Empirical process results for exchangeable arrays," *The Annals of Statistics*, 2021, *49* (2), 845–862.

de Sijpe, Nicolas Van and Frank Windmeijer, "On the power of the conditional likelihood ratio and related tests for weak-instrument robust inference," *Journal of Econometrics*, 2023, *235* (1), 82–104.

Ding, Peng and Jiannan Lu, "Principal stratification analysis using principal scores," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2017, *79* (3), 757–777.

Djogbenou, Antoine A, James G MacKinnon, and Morten Ørregaard Nielsen, "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 2019, *212* (2), 393–412.

Dobbie, Will, Jacob Goldin, and Crystal S Yang, "The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges," *American Economic Review*, 2018, *108* (2), 201–40.

Doyle, Joseph J, "Child protection and child outcomes: Measuring the effects of foster care," *American Economic Review*, 2007, *97* (5), 1583–1610.

Duflo, Esther and Emmanuel Saez, "The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment," *The Quarterly Journal of Economics*, 2003, *118* (3), 815–842.

Dufour, Jean-Marie, "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica: Journal of the Econometric Society*, 1997, pp. 1365–1387.

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie Van Dijk, "Selection in surveys: Using randomized incentives to detect and account for nonresponse bias," Technical Report, National Bureau of Economic Research 2021.

Elliott, Graham, Ulrich K Müller, and Mark W Watson, "Nearly optimal tests when a nuisance parameter is present under the null hypothesis," *Econometrica*, 2015, *83* (2), 771–811.

Evdokimov, Kirill S and Michal Kolesár, "Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects," *Working Paper*, 2018.

Gneezy, Uri and Aldo Rustichini, "Pay enough or don't pay at all," *The Quarterly Journal of Economics*, 2000, *115* (3), 791–810.

Graham, Bryan S, "Sparse network asymptotics for logistic regression," *arXiv preprint arXiv:2010.04703*, 2020.

Hall, Alastair R and Atsushi Inoue, "The large sample behaviour of the generalized method of moments estimator in misspecified models," *Journal of Econometrics*, 2003, *114* (2), 361–394.

Hansen, Bruce E and Seojeong Lee, "Asymptotic theory for clustered samples," *Journal of econometrics*, 2019, *210* (2), 268–290.

Heckman, James J and Edward Vytlacil, "Structural equations, treatment effects, and econometric policy evaluation 1," *Econometrica*, 2005, *73* (3), 669–738.

_ **and Rodrigo Pinto**, "Unordered monotonicity," *Econometrica*, 2018, *86* (1), 1–35.

Horowitz, Joel L and Charles F Manski, "Nonparametric analysis of randomized experiments with missing covariate and outcome data," *Journal of the American Statistical Association*, 2000, *95* (449), 77–84.

Huber, Martin and Giovanni Mellace, "Testing instrument validity for LATE identification based on inequality moment constraints," *Review of Economics and Statistics*, 2015, *97* (2), 398–411.

_ , **Lukas Laffers, and Giovanni Mellace**, "Sharp IV bounds on average treatment effects on the treated and other populations under endogeneity and noncompliance," *Journal of Applied Econometrics*, 2017, *32* (1), 56–79.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka, "Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice," Technical Report, National Bureau of Economic Research 2021.

Jackson, C Kirabo, "What do test scores miss? The importance of teacher effects on non–test score outcomes," *Journal of Political Economy*, 2018, *126* (5), 2072–2107.

Janisch, Maximilian and Thomas Lehéricy, "Berry–Esseen-Type Estimates for Random Variables with a Sparse Dependency Graph," *Journal of Theoretical Probability*, 2024, *37* (4), 3627–3653.

Janson, Svante, "Normal convergence by higher semiinvariants with applications to sums of dependent random variables and random graphs," *The Annals of Probability*, 1988, pp. 305–312.

Kallenberg, Olav, *Probabilistic symmetries and invariance principles*, Vol. 9, Springer, 2005.

Kamat, Vishal, "On the identifying content of instrument monotonicity," *arXiv preprint arXiv:1807.01661*, 2018.

Kitagawa, Toru, "A test for instrument validity," *Econometrica*, 2015, *83* (5), 2043–2063.

_ , "The identification region of the potential outcome distributions under instrument independence," *Journal of Econometrics*, 2021.

Kleibergen, Frank and Zhaoguo Zhan, "Double robust inference for continuous updating GMM," *Quantitative Economics*, 2025, *16* (1), 295–327.

Klein, Tobias J, "Heterogeneous treatment effects: Instrumental variables without monotonicity?," *Journal of Econometrics*, 2010, *155* (2), 99–116.

Kline, Patrick and Christopher R Walters, "On Heckits, LATE, and numerical equivalence," *Econometrica*, 2019, *87* (2), 677–696.

Kolesár, Michal, "Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity," Working Papers 2013-2, Princeton University. Economics Department. November 2013.

Lee, David S, "Training, wages, and sample selection: Estimating sharp bounds on treatment effects," *The Review of Economic Studies*, 2009, *76* (3), 1071–1102.

_ , **Justin McCrary, Marcelo J Moreira, Jack R Porter, and Luther Yap**, "What to do when you can't use '1.96' Confidence Intervals for IV," Working Paper 31893, National Bureau of Economic Research November 2023.

Lee, Seojeong, "A consistent variance estimator for 2SLS when instruments identify different LATEs," *Journal of Business & Economic Statistics*, 2018, *36* (3), 400–410.

Lehmann, Erich Leo and Joseph P Romano, *Testing statistical hypotheses*, Vol. 3, Springer, 2005.

Leung, Michael P, "Rate-optimal cluster-randomized designs for spatial interference," *The Annals of Statistics*, 2022, *50* (5), 3064–3087.

Lim, Dennis, Wenjie Wang, and Yichong Zhang, "A conditional linear combination test with many weak instruments," *Journal of Econometrics*, 2024, *238* (2), 105602.

MacKinnon, James G, Morten Ørregaard Nielsen, and Matthew D Webb, "Wild bootstrap and asymptotic inference with multiway clustering," *Journal of Business & Economic Statistics*, 2021, *39* (2), 505–519.

Manski, Charles F, "Anatomy of the selection problem," *Journal of Human Resources*, 1989, pp. 343–360.

Matsushita, Yukitoshi and Taisuke Otsu, "A jackknife Lagrange multiplier test with many weak instruments," *Econometric Theory*, 2022, pp. 1–24.

Menzel, Konrad, "Bootstrap With Cluster-Dependence in Two or More Dimensions," *Econometrica*, 2021, *89* (5), 2143–2188.

Michalopoulos, Stelios and Elias Papaioannou, "Pre-colonial ethnic institutions and contemporary African development," *Econometrica*, 2013, *81* (1), 113–152.

Mikusheva, Anna and Liyang Sun, "Inference with many weak instruments," *The Review of Economic Studies*, 2022, *89* (5), 2663–2686.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R Walters, "The causal interpretation of two-stage least squares with multiple instrumental variables," *American Economic Review*, 2021, *111* (11), 3663–3698.

_ , Andres Santos, and Alexander Torgovitsky, "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 2018, *86* (5), 1589–1619.

Moreira, Marcelo J, "A conditional likelihood ratio test for structural models," *Econometrica*, 2003, *71* (4), 1027–1048.

_ , "A Maximum Likelihood Method for the Incidental Parameter Problem," *The Annals of Statistics*, 2009, *37* (6A), 3660–3696.

_ , "Tests with correct size when instruments can be arbitrarily weak," *Journal of Econometrics*, 2009, *152* (2), 131–140.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian, "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 2019, *109* (4), 1426–60.

Neumark, David, Ian Burn, and Patrick Button, "Is it harder for older workers to find jobs? New and improved evidence from a field experiment," *Journal of Political Economy*, 2019, *127* (2), 922–970.

Noack, Claudia, "Sensitivity of LATE Estimates to Violations of the Monotonicity Assumption," *arXiv preprint arXiv:2106.06421*, 2021.

Nunn, Nathan and Leonard Wantchekon, "The slave trade and the origins of mistrust in Africa," *American Economic Review*, 2011, *101* (7), 3221–52.

Richardson, Thomas S and James M Robins, "Analysis of the binary instrumental variable model," *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 2010, *25*, 415–444.

Ross, Nathan, "Fundamentals of Stein's method," *Probability Surveys*, 2011, *8*, 210–293.

Roy, A. D., "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 06 1951, *3* (2), 135–146.

Shea, Joshua, "Testing for Racial Bias in Police Traffic Searches," *University of Illinois, Champaign Urbana, USA*, 2022.

Słoczyński, Tymon, "When should we (not) interpret linear iv estimands as late?," *arXiv preprint arXiv:2011.06695*, 2020.

Small, Dylan S, Zhiqiang Tan, Roland R Ramsahai, Scott A Lorch, and M Alan Brookhart, "Instrumental variable estimation with a stochastic monotonicity assumption," *Statistical Science*, 2017, *32* (4), 561–579.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 1997, *65*, 557–586.

Stock, James H. and Motohiro Yogo, "Testing for Weak Instruments in Linear IV Regression," in Donald W.K. Andrews and James H. Stock, eds., *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, Cambridge University Press, 2005, chapter 5, pp. 80–108.

van't Hoff, Nadja, Arthur Lewbel, and Giovanni Mellace, *Limited Monotonicity and the Combined Compliers LATE*, University of Southern Denmark, Faculty of Business and Social Sciences., 2023.

Verdier, Valentin, "Estimation and inference for linear models with two-way fixed effects and sparsely matched data," *Review of Economics and Statistics*, 2020, *102* (1), 1–16.

Wets, Roger J-B, "On the continuity of the value of a linear program and of related polyhedral-valued multifunctions," in "Mathematical Programming Essays in Honor of George B. Dantzig Part I," Springer, 1985, pp. 14–29.

Wooldridge, Jeffrey M, *Econometric analysis of cross section and panel data*, MIT press, 2010.

Xu, Ruonan and Luther Yap, "Clustering with Potential Multidimensionality: Inference and Practice," *arXiv preprint arXiv:2411.13372*, 2024.

Yap, Luther, "Asymptotic theory for two-way clustering," *Journal of Econometrics*, 2025, *249*, 106001.