Title

Meditations on Econometric Modeling

Author

Navjeevan, Manvendu

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Meditations on Econometric Modeling

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Economics

by

Manvendu Navjeevan

2024

ABSTRACT OF THE DISSERTATION

Meditations on Econometric Modeling

by

Manvendu Navjeevan

Doctor of Philosophy in Economics

University of California, Los Angeles, 2024

Professor Denis Nikolaye Chetverikov, Chair

The contents of this dissertation are split into three chapters, each of which covers a distinct problem in econometrics.

The first chapter considers the problem of hypothesis testing in a weakly identified instrumental variables models with a potentially large number of instrumental variables. Instrumental variables strategies, where a causal effect is identified by exploiting exogenous variation in the explanatory variable induced by changes in instrumental variables, are one of the most common quasi-experimental research designs used in economics. In recent years, there has been interest in using a large number of instruments in combination with some regularized method, such as LASSO, in order to flexibly model the relationship between the instrumental and explanatory variable. However, in these setting there has been little work on testing hypotheses about the structural parameter when this first-stage relationship is weak. In this chapter I propose a new test for the structural parameter in a instrumental variables models that has correct asymptotic size even when the number of instruments is potentially much larger than the sample size and identification is arbitrarily weak. The limiting distribution of the test statistic is derived through a novel direct Gaussian approximation argument and is combined with the sup-score test of Belloni et al. (2012a) in order to improve power against certain alternatives. In both empirical data and simulation study the proposed methods are shown to have favorable size control and power properties compared to existing methods.

In the second chapter, coauthored with Adam Baybutt, we consider inference on the conditional average treatment effect (CATE) under first stage model misspecification. Plausible identification of CATEs can rely on controlling for a large number of variables to account for confounding factors. In these high-dimensional settings, estimation of the CATE requires estimating first-stage models whose consistency relies on correctly specifying their parametric forms. While doubly-robust estimators of the CATE exist, inference procedures based on the second stage CATE estimator are not doubly robust. Using the popular augmented inverse propensity weighting signal, we propose an estimator for the CATE whose resulting Wald-type confidence intervals are doubly robust. We assume a logistic model for the propensity score and a linear model for the outcome regression, and estimate the parameters of these models using an ℓ_1 (Lasso) penalty to address the high dimensional covariates. Our proposed estimator remains consistent at the nonparametric rate and our proposed pointwise and uniform confidence intervals remain asymptotically valid even if one of the logistic propensity score or linear outcome regression models are misspecified.

The final chapter, coauthored with Prof. Rodrigo Pinto, investigates the relationship among monotonicity conditions in IV models with multiple choices and categorical instruments. The comparison between monotonicity conditions of ordered and unordered choice models is central to our analysis. We show that these seemingly unrelated conditions exhibit non-trivial symmetries that can be traced back to a weaker condition called Minimal Monotonicity. This novel condition captures an essential property for identifying causal parameters while being necessary for ascribing causal interpretation to Two-Stage Least Squares (2SLS) estimands. We show that minimal monotonicity naturally arises from a notion of rationality in revealed preference analysis. The condition enables to describe non-standard choice behaviors and serves as a building block for a wide range of economically-justified monotonicity conditions that do not fit the narrative dictated by either ordered or unordered choice models.

The dissertation of Manvendu Navjeevan is approved.

Rodrigo Ribeiro Antunes Pinto

Zhipeng Liao

Andres Santos

Denis Nikolaye Chetverikov, Committee Chair

University of California, Los Angeles

2024

To my lovely mother, whom I am always laughing with, my dedicated father, who has cared for me in a thousand ways, and my sister who has permanently brightened my life.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would first like to thank my committe chair, Denis Chetverikov, for his guidance and insight throughout my years at UCLA. While there were many times I entered his office discouraged and confused, I always exited optimistic and with a more fundamental understanding of econometrics. It is my sincere hope that I am half as helpful to students in the future. I am thankful to Andres Santos for always making time for my questions and encouraging me to think deeper about a problem. Even when I had not accomplished much, Andres always made my thinking seem valuable. I am grateful to Rodrigo Pinto for his support and friendship during the Ph.D and to Zhipeng Liao for his acuity and approachability. I am thankful to my fellow econometric cohort mate, Danny Ober-Reynolds, for always being willing to help when I would get stuck. Most broadly, I am grateful to everyone involved in econometric theory during my time at UCLA. I have found it a very welcoming and exciting place to be over the last six years.

I am grateful for every one of my twenty-five fellow Ph.D cohort members, who have made life in grad school better in innumerable ways. Calvin, who always believed in me, and Daniel, with whom I shared many adventures in LA, made unbeatable roommates. Nicole, Adam, and Danny were excellent running partners who accomodated my frequent tardiness. Domenico and Akira were officemates who I was always excited to see for lunch.

I am grateful to my childhood friends, Andy Kidder and Arjun Venkatesh, whose support of me in this endeavor, and all others, has meant the world.

VITA

Education

University of California, Los Angeles

M.A. in Economics 2018-2020

Carnegie Mellon University

B.S. in Economics and Mathematical Sciences 2014-2018

Relevant Positions

University of California Los Angeles

Teaching Assistant 2019-2024

Summer Instructor 2020-2023

Fellowships, Honors, and Awards

Perfect Score (170) on Math GRE. 2017

Dissertation Year Fellowship, Graduate Division, UCLA. 2023

Best Proseminar in Econometrics Award, UCLA. 2023

Chapter 1

An Identification-and Dimensionality-Robust Test for Instrumental Variables Models

1.1. Introduction

Consider a linear instrumental variables (IV) model

$$y_i = x_i'\beta + z_{1i}'\Gamma + \epsilon_i, \quad \mathbb{E}[\epsilon_i|z_i] = 0 \tag{1.1.1}$$

where $y_i \in \mathbb{R}$ is an outcome of interest and $x_i \in \mathbb{R}^{d_x}$ is a vector of endogenous variables that may be correlated with the structural error $\epsilon_i \in \mathbb{R}$. The variable $z_i = (z_{1i}, z_{2i})' \in \mathbb{R}^{d_c} \times \mathbb{R}^{d_z}$ represents a vector of instrumental variables, of which a subvector of fixed dimension, $z_{1i}' \in \mathbb{R}^{d_c}$, is included in the structural equation (1.1.1) as exogenous control. I assume that the researcher has access to n independent observations of $(y_i, x_i', z_i')'$. In this setting, I propose a new test for a two-sided restriction on the structural parameter; $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$. The proposed test has exact asymptotic size even when instruments are potentially high-dimensional ($d_z \gg n$) and arbitrarily weak.

When instruments are suspected to be weak, researchers may want to test hypotheses about

structural parameters using testing procedures that are robust to identification strength. These identification-robust testing procedures all require some conditions on the rate of growth of the number of instruments, d_z, in relation to the sample size, n. The testing procedure considered in this paper seeks to fill two perceived gaps in the literature. The first is for the cases where the number of instruments is high-dimensional ($d_z \gg n$), which can occur if the researcher chooses to enhance an initial set of instruments via polynomial or other transformations in order to flexibly model the first stage relationship between the endogenous variables and the instruments.

In these settings, Belloni et al. (2012a) show that, when identification is strong, LASSO, post-LASSO, or other machine learning based estimators can be used in the first stage without affecting the asymptotic normality of resulting second stage estimators. This is possible because the conditional moment restriction in (1.1.1) implies a certain orthogonality that, under strong identification, allows the researcher to ignore estimation error in the first stage. However, when identification is sufficiently weak, the signal from the instruments can be on a similar or lesser order to the first-stage estimation error and the limiting behavior of the first-stage estimate must be explicitly accounted for (Mikusheva, 2023). This is problematic in high-dimensional settings as the exact limiting behavior of machine learning based estimators is typically not known. As such, there has been limited work on identification robust testing in the high dimensional setting and existing identification robust tests that allow $d_z \gg n$ (Belloni et al., 2012a; Gautier and Rose, 2021; Mikusheva, 2023) either fail to incorporate first-stage information or rely on sample splitting, both of which may reduce power in overidentified models.

An second gap in the literature is for cases with a moderate number of instruments. The initial identification robust tests of Anderson and Rubin (1949), Staiger and Stock (1997), Moreira (2001, 2003), and Kleibergen (2002, 2005) are shown by Andrews and Stock (2007) to control size under heteroskedasticity when the number of instruments cubed is negligible

compared to sample size, $d_z^3/n \to 0$. Meanwhile, recent tests proposed in Mikusheva and Sun (2021), Crudu et al. (2021), Matsushita and Otsu (2022), and Lim et al. (2022) allow the number of instruments to be proportional to sample size, $d_z/n \to \varrho \in [0, 1)$, but require that the number of instruments be large, $d_z \to \infty$. In practice, these conditions can be difficult to interpret and in settings with a moderate number of instruments it may be unclear which test, if any, is applicable. As examples, consider the analyses of Derenoncourt (2022), where $d_z = 9$ and $n = 239$, and Paravisini et al. (2014), where $d_z = 10$ and $n = 5,995$. The number of instruments cubed is non-negligible relative to the sample size, but asymptotic approximations based on $d_z \to \infty$ seem unlikely to resemble the finite sample distribution.

In contrast, the test considered in this paper can be applied in any of the settings described above. To test the null hypothesis I borrow an idea from Kleibergen (2002, 2005) and use first stage estimates that are uncorrelated with the structural error under the null hypothesis. These first-stage estimates are constructed using a jackknife ridge procedure and the structural errors are partialled out via an auxiliary conditional slope parameter. Combining the jackknife and partialing out approaches allows me to asymptotically remove dependence of the first-stage estimates not only from an observations own structural error but also from the structural error of other observations. So long as this auxiliary parameter can be consistently estimated, the proposed test statistic has a limiting chi-squared distribution with degrees of freedom equal to the number of structural parameters. The conditional slope parameter is simple to estimate with out-of-the-box methods, and consistency is achievable under approximate sparsity even when the number of instruments is much larger than the sample size. This approximate sparsity assumption is trivially satisfied when the first- and second-stage errors are homoskedastic.

Kleibergen's original analysis relies on applying central limit and continuous mapping theorems to show that variables in the model can be treated as if they were normally distributed. The limiting distribution of Kleibergen's K-statistic is then derived by conditioning on the first

stage estimates, exploiting the fact that uncorrelated jointly Gaussian random variables are independent. When the number of instruments is large, however, standard asymptotic theory cannot be applied. Instead, I develop new interpolation arguments to directly show that, in local neighborhoods of the null characterized by a local power index, the distribution of my proposed test statistic can be uniformly approximated by that of an analog statistic which replaces each observation in the expression of the test statistic with a Gaussian version. The interpolation arguments are based on Lindeberg's interpolation method (Lindeberg, 1922), but are modified to accommodate a "divide-by-zero" problem that arises under weak identification. These modifications are adaptable to other settings and may be of independent interest to a growing literature on direct Gaussian approximation techniques (Chatterjee, 2006; Chernozhukov et al., 2017; Celentano et al., 2020). Interestingly, the interpolation approach applied in this paper requires minimal conditions on the first-stage estimates. In particular, these estimates are not required to be consistent so the researcher has some flexibility in choosing how she constructs the first stage estimates. However, analysis of local power suggests a bias-variance trade-off which guides the recommendation of using ridge regression in the first stage.

When there is a single endogenous variable, a leading case in empirical applications, analysis of limiting behavior is considerably simplified by taking advantage of the particular form of the test statistic. In this case I show that, under an additional regularity condition, an infeasible version of the test that could be constructed if the auxiliary parameter was known to the researcher is consistent whenever the local power index diverges. When the local power index is bounded, I examine the limiting power of the test by examining the behavior of the analog statistic. Under the alternative hypothesis the analog statistic has a nearly non-central χ^2 distribution conditional on the first-stage estimates. The noncentrality parameter is proportional to the correlation between the true first-stage model and the first-stage estimates, but inversely proportional to the second moment of the first-stage estimates. Unfortunately, partialling out the structural error may introduce bias into the

first-stage estimates under the alternate hypothesis. Against certain alternatives this bias can be particularly pronounced and erase the first-stage signal from the instruments. This issue is pointed out by Moreira (2001), Andrews et al. (2006), and Andrews (2016) in the context of Kleibergen's original K-statistic.

To address this, I propose a simple combination of the jackknife K-statistic with the sup-score statistic of Belloni et al. (2012a) based on a thresholding rule. As with the Anderson-Rubin statistic, while the sup-score statistic does not incorporate first-stage information, it does not suffer from a loss of power against any particular alternative (Andrews et al., 2006; Andrews, 2016). The combination test decides whether the jackknife K-test or the sup-score test should be run by comparing the value of a conditioning statistic to a predetermined cutoff value. In the approximating Gaussian regime, this conditioning statistic is marginally independent of both the jackknife K-statistic and the sup-score statistic. This allows me to show that the combination test controls size under the null without having to require that the conditioning statistic converges in distribution to a stable limit. In a simulation study, I find that taking this cutoff value to be the 75^{th} quantile of the distribution of the conditioning statistic delivers a reasonable balance of power against local and distant alternatives. Using results in Chernozhukov et al. (2017) and Belloni et al. (2018) this quantile can be simulated via a multiplier bootstrap procedure.

When there are multiple endogenous variables, I cannot take advantage of the simplified form of the test statistic. Instead, I use a more involved interpolation argument that relies on strengthened moment conditions. This modified argument has a clean geometric interpretation explained in Section 1.6. Under these strengthened conditions I derive the limiting chi-squared distribution of the jackknife K-statistic in the larger context and propose a generalization of the thresholding test to improve power properties.

I apply the proposed testing procedures to the data of Gilchrist and Sands (2016) to generate weak instrument-robust confidence intervals for the effect of social spillovers in movie

consumption. Following Belloni et al. (2012a), the authors' initial analysis uses conventional heteroskedasticity-robust standard errors after estimating the first-stage via post-LASSO. The validity of this analysis depends on the structural parameter being strongly identified. Using a simple numerical demonstration, I argue that the first-stage F-statistics reported by the authors may not be reliable indicators of identification strength when LASSO is used to select instruments. The identification-robust confidence intervals generated by inverting the jackknife K-statistic are larger than those implied by the initial analysis but do not rule out the authors' point estimates. Moreover, for the author's main specification the confidence intervals obtained using my proposed testing procedures are considerably smaller than those obtained through inverting the sup-score test.

Finally, I examine the applicability of the theoretical results in this paper through a simulation study. While existing tests seem to face size distortions in alternate regimes, the test based on the jackknife K-statistic is has nearly exact size in a variety of settings. While the jackknife K-statistic may have diminished power against certain alternatives, this deficiency is ameliorated by combining the jackknife K-statistic with the sup-score statistic via the thresholding test. Compared to the many-instrument tests of Mikusheva and Sun (2021) and Matsushita and Otsu (2022) and the sup-score test of Belloni et al. (2012a), the tests proposed in this paper also appear to have favorable power properties, particularly when the instruments are highly correlated.

The outline of this paper is as follows. Section 1.3 formally defines the model considered and introduces the jackknife K-statistic. Section 1.4 provides an overview of the Gaussian approximation approach with a single endogenous variable and characterizes the limiting behavior of the test statistic in this setting. Section 1.5 uses this characterization to examine the power properties of the test and introduces the combination test to address power deficiencies against certain alternatives. Section 1.6 extends the analyses of Sections 1.4 and 1.5 to the case of multiple endogenous variables and outlines the Gaussian approximation argument

in this setting. Section 1.7 contains the empirical application while Section 1.8 provides evidence from simulation study. Proofs of the main results are deferred to Sections 1.10.1–1.10.4.

Notation. For any $n \in \mathbb{N}$ let $[n]$ denote the set $\{1, \ldots, n\}$. I work with a sequence of probability measures P_n on the data $\{(y_i, x_i, z_i) : i \in [n]\}$. To accommodate independent but not identically distributed observations, let $\mathbb{E}_n[f_i] = n^{-1} \sum_{i=1}^{n} f_i$ denote the empirical expectation and $\bar{\mathbb{E}}[f] = \mathbb{E}_n[\mathbb{E}[f_i]]$ denote the average expectation operator.

1.2. Prior Literature and Empirical Practice

When the first-stage F-statistic is small, standard asymptotic approximations may fail to accurately describe the finite-sample behavior of IV estimates. This was first pointed out by Nelson and Startz (1990) and Bound et al. (1995) who consider the finite-sample behavior of two-stage least squares (2SLS) in alternate settings where the IV is only weakly correlated with the endogenous variable. In a seminal paper, Staiger and Stock (1997) capture this phenomena in an asymptotic framework by considering a sequence of first-stage models that shrink to zero with the sample size. Under this framework, standard IV estimates are no longer consistent and inference procedures based on these statistics fail to control size. Because of these results, there has been a large interest in developing tests for the structural parameter that control size regardless of identification strength.

To test hypotheses about the structural parameter when instruments are suspected to be weak, Staiger and Stock (1997) propose the use of the Anderson-Rubin statistic, which does not require any assumptions about identification strength to control size. Noting that the Anderson-Rubin test is inefficient in overidentified models, Moreira (2001) and Kleibergen (2002, 2005) propose the use of the (non-jackknife) K-statistic, which has a limiting null distribution that does not depend on the number of instruments. Compared to the Anderson-Rubin statistic, these tests have improved power in local neighborhoods of the null but can

perform poorly against certain alternatives. To address this, Moreira (2003) and Kleibergen (2005) suggest combinations of the K-statistic and Anderson-Rubin statistic based on a conditioning statistic that is independent of them both under the null. Andrews et al. (2006) characterize the power envelope in a homoskedastic weakly identified IV model and show that the test based on the conditional likelihood ratio statistic of Moreira (2003) has nearly optimal power in this setting. When errors are heteroskedastic, Andrews (2016) proposes alternate combinations of the K-statistic and the Anderson-Rubin statistic based on a minimax regret criterion.

These initial tests are developed under asymptotic frameworks that treat the number of instruments as fixed or growing slowly relative to the sample size (Han and Phillips, 2006; Newey and Windmeijer, 2009; Andrews and Stock, 2007). However, with the emergence of large datasets and more sophisticated research designs, researchers may encounter scenarios where the number of instruments may not be negligible as a ratio of sample size. A prominent example of this is in judge-design settings where the number of instruments is equal to the number of judges to whom an individual can be assigned to (Maestas et al., 2013; Sampat and Williams, 2019; Dobbie et al., 2018). Since each judge can handle only a finite number of cases the number of instruments is proportional to the sample size. Moreover, to flexibly model the first-stage, researchers may generate a large number of instruments by enriching a "small" initial set of instruments via polynomial (or other) transformations. Angrist and Krueger (1991) famously interact quarter-of-birth, state-of-birth, and year-of-birth dummies to construct a total of 180 instruments. Belloni et al. (2012a) show that, when identification is strong, researchers can use a potentially high-dimensional, $d_z \gg n$, set of first-stage instrument basis terms in conjunction with a regularized LASSO or post-LASSO estimate of the first-stage. This strategy has been successfully employed in practice by Paravisini et al. (2014), Gilchrist and Sands (2016), Derenoncourt (2022), and Jou and Morgan (2023).

To address these settings, there has been recent interest in developing weak instrument-robust

tests under asymptotic frameworks that do not require that the ratio of instruments to sample size tends to zero. Crudu et al. (2021), Mikusheva and Sun (2021), and Matsushita and Otsu (2022) take advantage of a new central limit theorem for quadratic forms developed in Chao et al. (2012) and propose weak identification-robust tests that are valid even when the number of instruments is proportional to sample size; $d_z/n \to \varrho \in [0, 1)$. Following the many instruments asymptotic framework first introduced by Bekker (1994), the analyses in these papers rely on the number of instruments diverging. When the number of instruments is fixed or diverges slowly to infinity, these asymptotic approximations may provide poor characterizations of the proposed test statistics' finite sample distribution.

Limited identification-robust testing procedures exist for the high-dimensional case, $d_z \gg n$. To my knowledge, the only two options available are the sup-score test of Belloni et al. (2012a) and the split-sample optimal instrument AR test developed in Mikusheva (2023).[1] The sup-score test makes use of Gaussian approximations for maxima of high-dimensional vectors developed in Chernozhukov et al. (2013) but suffers from the same issue as the Anderson-Rubin test in that its critical value is increasing with the number of instruments. The spilt sample optimal instrument AR test splits the dataset into two parts and uses one split to estimate an optimal instrument and the other to test the null hypothesis. This may lead to a loss of power as only half of the sample is being effectively used to test the null hypothesis.

Weak instrument-robust tests may be particularly interesting in high-dimensional and heteroskedastic settings due to a lack of clarity on how to pretest for identification strength. When the number of instruments is modeled as fixed and errors are homoskedastic, Stock and Yogo (2005) propose pretesting for the strength of identification via the first-stage F-statistic. Based on their results, common practice in empirical settings has been to use standard Wald tests whenever the first-stage F-statistic exceeds 10. Lee et al. (2022) point out this recommendation is not applicable in heteroskedastic models and update the recommended

[1]The sup-score test is also considered by Gautier and Rose (2021, 2022).

F-statistic cutoff. To pretest for weak identification in the many-instruments asymptotic framework, $d_z \to \infty$, Mikusheva and Sun (2021) propose a new $\tilde{F}$-statistic and suggest using identification-robust procedures when $\tilde{F} < 4.14$. When the number of instruments is larger than sample size there is no accepted full-sample pretest for identification strength.[2] In particular, I demonstrate in Section 1.7 that first-stage F-statistics resulting from first-stage post-LASSO procedures can be misleading even if they are larger the standard cutoff of 10.

Asymptotic Regime	Main Tests
Low-Dimensional: $d_z^3/n \to 0$	Anderson-Rubin K/Lagrange Multiplier Conditional Linear Combination
Many-Instruments: $d_z/n \to \phi \in [0,1)$ $d_z \to \infty$	Jackknife-AR Jackknife-LM Conditional Linear Combination
High-Dimensional: $\log^M(d_z n)/n \to 0$	Sup-Score Test Split-Sample AR

Table 1.2.1: Existing Identification and Heteroskedasticity Robust Tests for Linear IV models.

I contribute to these literatures by proposing a new identification-robust test for the structural parameter that can work in potentially high-dimensional settings ($d_z \gg n$) without requiring that the number of instruments diverges. The testing procedures in this paper may be particularly applicable in intermediate cases where the number of instruments cubed may not be negligible relative to sample size but it is unclear whether asymptotic approximations based on $d_z \to \infty$ will accurately describe finite sample behavior. Examples of such intermediate cases include the post-LASSO analyses of Derenoncourt (2022), where $d_z = 9$ and $n = 239$, Paravisini et al. (2014), $d_z = 10$ and $n = 5{,}995$, and Gilchrist and Sands (2016), $d_z = 52$ and

[2] Mikusheva (2023) suggests a split-sample pretest for use with the split-sample optimal-instrument AR test.

$n = 1{,}671$.

In addition to the literature on weak-instrument robust testing, I contribute to a growing literature on direct Gaussian approximation and interpolation techniques (Chatterjee, 2006, 2010; Pouzo, 2015; Chernozhukov et al., 2013, 2017; Celentano et al., 2020). These techniques have proven useful to approximate the behaviors of statistics in a variety of nonstandard settings, such as high-dimensional random vectors or spectral analysis of random matrices. Prior analysis of statistics via interpolation techniques has relied on the boundedness of the derivatives of these statistics with respect to individual observations. This condition does not hold in my setting as the denominator of my test-statistic is not bounded away from zero under weak identification and, as such, derivatives of the jackknife K-statistic with respect to terms in the denominator may be unbounded. This poses a number of technical challenges for my interpolation argument that must be overcome in order to characterize the limiting behavior of the jackknife K-statistic, particularly when $d_x > 1$. I contribute to this literature by proposing modifications of the original Lindeberg (1922) interpolation technique that can accommodate statistics with unbounded derivatives.

1.3. Model and Setup

Though the analysis below allows for exogenous regressors, to simplify the exposition I follow Mikusheva and Sun (2021) and assume that they have already been partialed out of both the outcome, y_i, and the endogenous regressors, x_i. As the controls are assumed to be of fixed dimension, this is without loss of generality.[1] Along with the structural equation in (1.1.1), the IV model can then be written with the first stage as a system of simultaneous equations:

$$\begin{aligned} y_i &= x_i'\beta + \varepsilon_i \\ x_i &= \Pi_i + v_i \end{aligned} \tag{1.3.1}$$

[1] For discussion refer to Section 1.11.

The researcher observes the outcome $y_i \in \mathbb{R}$, the endogenous variable $x_i \in \mathbb{R}^{d_x}$, and the instruments $z_i \in \mathbb{R}^{d_z}$ but neither the structural error $\varepsilon_i \in \mathbb{R}$ nor the first-stage errors $v_i \in \mathbb{R}^{d_x}$. The structural error is assumed to be conditional-mean independent of the instruments, $\mathbb{E}[\varepsilon_i|z_i] = 0$. I denote $\mathbb{E}[x_i|z_i]$ as $\Pi_i \coloneqq \mathbb{E}[x_i|z_i]$ and make no assumptions about the functional form of the conditional expectation so the instruments are allowed to affect the endogenous variable in a nonlinear fashion.

The random variables $\{(z_i, \varepsilon_i, v_i)\}_{i=1}^n$ are assumed to be independent across observations. Observations need not be identically distributed but the errors are assumed to have a common covariance structure conditional on the instruments z_i:

$$\mathrm{Var}((\varepsilon_i, v_i)'|z_i) \coloneqq \Omega(z_i) = \begin{pmatrix} \sigma^2_{\epsilon\epsilon}(z_i) & \Sigma_{v\epsilon}(z_i) \\ \Sigma_{\epsilon v}(z_i) & \Sigma_{vv}(z_i) \end{pmatrix} \in \mathbb{R}^{(1+d_x)\times(1+d_x)}$$

As $\Omega(z_i)$ is otherwise left unrestricted, the errors are allowed to be heteroskedastic. All results in this paper hold conditionally on a realization of the instruments $\boldsymbol{z} \coloneqq (z_1', \ldots, z_n') \in \mathbb{R}^{n \times d_z}$ so from this point forth they are treated as fixed and all expectations can be understood as conditional on the instruments.

Under this setup, the researcher wishes to test a two-sided restriction on the structural parameter:

$$H_0 : \beta = \beta_0 \quad \text{vs.} \quad H_1 : \beta \neq \beta_0$$

I am interested in constructing powerful tests for this null-alternate pair that are asymptotically valid under arbitrarily weak identification and with minimal restrictions on the number of instruments d_z. To this end, define the null errors $\varepsilon_i(\beta_0) \coloneqq y_i - x_i'\beta_0$. Using these, I construct a variable, r_i, that is a "partialed-out" version of the endogenous variable satisfying $\mathrm{Cov}(r_i, \epsilon_i(\beta_0)) = 0$:

$$r_i \coloneqq x_i - \rho(z_i)\epsilon_i(\beta_0), \quad \rho(z_i) \coloneqq \frac{\mathrm{Cov}(\epsilon_i(\beta_0), x_i)}{\mathrm{Var}(\epsilon_i(\beta_0))} \in \mathbb{R}^{d_x}$$

$$= \frac{\Sigma_{v\epsilon}(z_i) + \Sigma_{vv}(z_i)(\beta - \beta_0)}{(1, \beta - \beta_0)'\Omega(z_i)'(1, \beta - \beta_0)}.$$

Each element of the nuisance parameter $\rho(z_i)$, $\rho_\ell(z_i)$ for $\ell = 1, \ldots, d_x$, can be interpreted as the (conditional) slope coefficient from a simple linear regression of $x_{\ell i}$ on $\epsilon_i(\beta_0)$. Thus, if $\rho_\ell(\cdot)$ falls in some function class Φ it can be estimated directly under H_0 by solving empirical analogs of:[2]

$$\rho_\ell(z_i) = \arg\min_{\varphi \in \Phi} \bar{\mathbb{E}}[(x_{\ell i} - \epsilon_i(\beta_0)\varphi(z_i))^2].$$

I will largely work under the assumption that $\rho(z_i)$ has an approximately sparse representation in some (growing) basis $b(z_i) \coloneqq (b_1(z_i), \ldots, b_{d_b}(z_i))' \in \mathbb{R}^{d_b}$, that is $\rho_\ell(z_i) = b(z_i)'\phi_\ell + \xi_{\ell i}$ where $\xi_{\ell i}$ represents an approximation error that tends to zero with the sample size and ϕ_ℓ is sparse in the sense that many of its coefficients are zero. This allows for nesting of the low-dimensional case, where the number of instruments is fixed, and the high dimensional case, where the number of instruments is potentially much larger than the sample size, under a unified estimation procedure. Under homoskedasticity, $\rho_\ell(z_i)$ is a constant function and thus has a spare representation in any basis that contains a constant term. In general, the aproximate sparsity assumption can either be interpreted as an assumption that there are only a few instruments that are important for explaining variation in the covariance matrix $\Omega(z_i)$ or as an assumption that the function $\rho(z_i)$ can be accurately approximated using only a smaller set of basis terms in $b(z_i)$.

As in Chernozhukov et al. (2022), the parameter ϕ_ℓ can be estimated via LASSO:

$$\hat{\phi}_\ell = \arg\min_{\phi \in \mathbb{R}^{d_b}} \mathbb{E}_n[(x_{\ell i} - \epsilon_i(\beta_0)b(z_i)'\phi)^2] + \lambda\|\phi\|_1, \tag{1.3.2}$$

or via post-LASSO, refitting an unpenalized version of (1.3.2) using only the basis terms associated with nonzero coeffecients in the inital LASSO regression. The estimating procedure

[2] Under H_1, $\rho_\ell(z_i)$ can be estimated directly by solving empirical analogs of $\rho_\ell(z_i) = \arg\min_{\phi \in \Phi} \mathbb{E}[(x_{\ell i} - \eta_i(\beta_0)\phi(z_i))^2]$ where $\eta_i(\beta_0) = \epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)|z_i]$. This requires an initial estimate of $\mathbb{E}[\epsilon_i(\beta_0)|z_i]$, however.

in (1.3.2) is a simple ℓ_1-penalized regression of $x_{\ell i}$ against $\epsilon_i(\beta_0)b(z_i)$. It can be easily implemented using out-of-the-box software available on most platforms. Under standard conditions, this leads to a consistent estimate of $\rho_\ell(z_i)$ as long as the sparsity condition $s^2 \log^M(d_b n)/n \to 0$ where s is the number of nonzero elements of ϕ_ℓ and M is a positive constant that depends on the moment bounds imposed. The estimation procedure is discussed in more detail in Section 1.4.2. With $\hat{\rho}(z_i) \coloneqq b(z_i)'\hat{\phi}_\ell$, I construct the estimated version of $r_{\ell i}$, $\hat{r}_{\ell i} \coloneqq x_i - \hat{\rho}(z_i)\epsilon_i(\beta_0)$ for each $\ell \in [d_x]$.

1.3.1. Test Statistic

The test statistic is based on an arbitrary jackknife-linear estimate of the first stage,

$$\widehat{\Pi}_{\ell i} = \sum_{j \neq i} h_{ij} \hat{r}_{\ell j}, \quad \ell \in [d_x]$$

for some "hat" matrix $H = [h_{ij}] \in \mathbb{R}^{n \times n}$. The phrase "hat matrix" is borrowed from ordinary least squares (OLS) where the projection matrix, $\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'$, is sometimes referred to as the hat matrix in the sense that $\hat{x} = \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'x$. In practice, the hat matrix, H, can be any matrix that depends only on $\boldsymbol{z}$. It is important to note that while $\widehat{\Pi}_{\ell i}$ does not depend on $\hat{r}_{\ell i}$, it may depend on z_i through the hat matrix H. This gives the test power against alternatives where $\mathbb{E}[\epsilon_i(\beta_0)z_i] \neq 0$. For technical reasons, I will assume that $h_{ii} = 0$ for each $i \in [n]$ so that $\widehat{\Pi}_{\ell i}$ can be written as $\widehat{\Pi}_{\ell i} = \sum_{j=1}^n h_{ij} r_{\ell j}$.

Formally, the only structure I require on the hat matrix H is a balanced-design condition described in Section 1.4. However, for reasons explained in Section 1.5 it may be optimal to introduce some regularization in estimating the first-stage models $\widehat{\Pi}_{\ell i}$ so I suggest using the deleted diagonal ridge-regression hat matrix $H(\lambda^\star)$:

$$[H(\lambda^\star)]_{ij} = \begin{cases} [\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z} + \lambda^\star I_{d_z})^{-1}\boldsymbol{z}']_{ij} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{1.3.3}$$

where, following recommendations in Harrell (2015) and van Wieringen (2023), the penalty parameter $\lambda^\star$ is set so that the effective degrees of freedom of the resulting hat matrix is no more than a fraction of the sample size:

$$\lambda^\star = \inf\{\lambda \geq 0 : \text{trace}(\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z} + \lambda I_{d_z})^{-1}\boldsymbol{z}') \leq n/5\}$$

The ridge hat matrix has the benefit of being well defined even when the number of instruments is larger than the sample size. I stress, though, that the $\widehat{\Pi}_{\ell i}$ estimators are not required to be consistent and the researcher may use any other hat matrix that she believes will lead to plausible first-stage estimates. Other possible choices of hat matrix include the jackknife OLS hat matrix of Angrist et al. (1999), the deleted diagonal projection matrix introduced in Chao et al. (2012) and successfully used in Kline et al. (2020), Crudu et al. (2021), Mikusheva and Sun (2021), and Matsushita and Otsu (2022), or hat matrices based on selecting instruments via some preliminary unsupervised technique such as principal component analysis (PCA). Remark 1.4.1 below discusses how the balanced-design condition may be verified for arbitrary choices of hat matrices.

For each $i = 1, \ldots, n$, define $\widehat{\Pi}_i = (\widehat{\Pi}_{1i}, \ldots, \widehat{\Pi}_{d_x i}) \in \mathbb{R}^{d_x}$ and $\widehat{\Pi}_{\epsilon i} = \epsilon_i(\beta_0)\widehat{\Pi}_i$. Collect these in the matrices

$$\begin{aligned} \varepsilon(\beta_0) &= \big(\varepsilon_1(\beta_0), \ldots, \varepsilon_n(\beta_0)\big)' \in \mathbb{R}^n \\ \widehat{\Pi} &= \big(\widehat{\Pi}_1', \ldots, \widehat{\Pi}_n'\big)' \in \mathbb{R}^{n \times d_x} \\ \widehat{\Pi}_\varepsilon &= \big(\widehat{\Pi}_{\epsilon 1}', \ldots, \widehat{\Pi}_{\epsilon n}'\big)' \in \mathbb{R}^{n \times d_x} \end{aligned} \tag{1.3.4}$$

The jackknife K-statistic can then be defined

$$JK(\beta_0) = \epsilon(\beta_0)'\widehat{\Pi}\big(\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon\big)^{-1}\widehat{\Pi}'\epsilon(\beta_0) \times \mathbf{1}\{\lambda_{\min}(\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon) > 0\} \tag{1.3.5}$$

I will show that, under appropriate moment bounds and conditions on the hat matrix,

H, the limiting distribution of $JK(\beta_0)$ under H_0 is $\chi^2_{d_x}$. For exposition, I will largely focus on the case where $d_x = 1$, in which case the form of the test statistic simplifies to $JK(\beta_0) = \left(\sum_{i=1}^n \epsilon_i(\beta_0)\widehat{\Pi}_i\right)^2 / \sum_{i=1}^n \epsilon_i^2(\beta_0)\widehat{\Pi}_i^2$. The extension to $d_x > 1$ is not immediate but is possible under strengthened moment conditions and is explored in Section 1.6.

Remark 1.3.1. While use of first-stage estimates that are uncorrelated with the structural error is inspired by Kleibergen (2002, 2005), the form of the jackknife K-statistic is distinct from that of the original K-statistics. One major difference is in how both test statistics account for heteroskedasticity. The K-statistic of Kleibergen (2005) accounts for heteroskedastic errors using a $d_z \times d_z$ matrix, which cannot be consistently estimated when d_z is large. In contrast, the jackknife K-statistic uses the heteroskedasticity robust variance estimate $(\widehat{\Pi}_\epsilon'\widehat{\Pi}_\epsilon)^{-1} \in \mathbb{R}^{d_x \times d_x}$. Showing that these variance estimates can be used to account for heteroskedasticity is a feature of the direct Gaussian approximation approach. Under weak identification the distribution of the variance estimate is relevant to the distribution of the test-statistic. However, even when $d_z \ll n$, the distribution of this variance estimate would be difficult to analyze using traditional central limit theorems as it is not a continuous function of a sample mean or even of a quadratic form.

1.4. Limiting Behavior with a Single Endogenous Variable

The limiting behavior of the test statistic is analyzed via a direct Gaussian approximation technique. When there is a single endogenous variable this approach can be considerably simplified. In this section, I detail the approach and take advantage of the simplified analysis to characterize the limiting behavior of the test statistic under local alternatives to H_0. This direct approach has the advantage of not relying on any particular central limit theorem, which allows a great deal of flexibility in the choice of hat matrix H.

For each $i \in [n]$, let $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ be jointly Gaussian random variables generated (i) independently of each other and the data and (ii) with the same mean and covariance matrix as

$(\epsilon_i(\beta_0), r_i)'$. In addition, define $\tilde{\Pi}_i \coloneqq \sum_{j \neq i} h_{ij} \tilde{r}_j$. The goal will be to show that the quantiles of $JK(\beta_0)$ can be approximated by corresponding quantiles of the Gaussian statistic,

$$JK_G(\beta_0) \coloneqq \frac{(\sum_{i=1}^n \tilde{\epsilon}_i(\beta_0)\tilde{\Pi}_i)^2}{\sum_{i=1}^n \mathbb{E}[\epsilon_i^2(\beta_0)]\tilde{\Pi}_i^2} \tag{1.4.1}$$

Since uncorrelated jointly Gaussian random variables are independent, under H_0 the vector $(\tilde{\epsilon}_1(\beta_0), \ldots, \tilde{\epsilon}_n(\beta_0))'$ is mean zero and independent of $(\tilde{r}_1, \ldots, \tilde{r}_n)'$. The null distribution of $JK_G(\beta_0)$ conditional on any realization of $(\tilde{r}_1, \ldots, \tilde{r}_n)'$ is then χ_1^2 and so its unconditional null distribution is also χ_1^2.

1.4.1. Interpolation Approach

Error arising from estimation of $\rho(z_i)$ prevents immediate comparison of the distribution of $JK(\beta_0)$ to the distribution of $JK_G(\beta_0)$. As such, I begin by considering the distribution of an infeasible statistic, $JK_I(\beta_0)$, which could be constructed if $\rho(z_i)$ were known to the researcher:

$$JK_I(\beta_0) \coloneqq \frac{\left(\sum_{i=1}^n \epsilon_i(\beta_0)\widehat{\Pi}_i^I\right)^2}{\sum_{i=1}^n \epsilon_i^2(\beta_0)\left(\widehat{\Pi}_i^I\right)^2} \times \mathbf{1}\Big\{\sum_{i=1}^n \epsilon_i^2(\beta_0)\left(\widehat{\Pi}_i^I\right)^2 > 0\Big\}$$

where $\widehat{\Pi}_i^I = \sum_{j \neq i} h_{ij} r_j$. To show that the distribution of $JK_I(\beta_0)$ can be approximated by the distribution of $JK_G(\beta_0)$, I adapt Lindeberg's interpolation method, first introduced by Lindeberg (1922) in an elegant proof of the central limit theorem. This method consists of one-by-one replacment of the terms $(\epsilon_i(\beta_0), r_i)$ in the expression of $JK_I(\beta_0)$ with their Gaussian analogs, $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$, and bounding the resulting one-step distributional changes.

Applying the interpolation method directly on the statistics $JK_I(\beta_0)$ and $JK_G(\beta_0)$, however, is not tractable as it requires bounding expectations of derivatives with respect to terms in the denominator. When identification is weak, the denominators of $JK_I(\beta_0)$ and $JK_G(\beta_0)$ may both be arbitrarily close to zero with positive probability. Derivatives with respect to terms in the denominators thus may not have finite expectations.

Instead, I consider a different approach. For a scaling factor s_n, introduced below, define the scaled numerators and denominators

$$N \coloneqq \Big(\frac{s_n}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(\beta_0)\widehat{\Pi}_i^I\Big)^2 \qquad \tilde{N} \coloneqq \Big(\frac{s_n}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\epsilon}_i(\beta_0)\tilde{\Pi}_i\Big)^2$$
$$D \coloneqq \frac{s_n^2}{n}\sum_{i=1}^{n}\epsilon_i^2(\beta_0)\big(\widehat{\Pi}_i^I\big)^2 \qquad \tilde{D} \coloneqq \frac{s_n^2}{n}\sum_{i=1}^{n}\mathbb{E}[\epsilon_i^2(\beta_0)]\big(\tilde{\Pi}_i\big)^2$$

and for any $a \geq 0$, define the decomposed statistics

$$JK_I^a(\beta_0) \coloneqq N - aD \qquad JK_G^a(\beta_0) \coloneqq \tilde{N} - a\tilde{D}$$

Since $D = 0$ implies $N = 0$ and since $\tilde{D} \neq 0$ almost surely, the events $(\{JK_I(\beta_0) \leq a\}, \{JK_G(\beta_0) \leq a\})$ are almost surely equivalent to the events $(\{JK_I^a(\beta_0) \leq 0\}, \{JK_G^a(\beta_0) \leq 0\})$. The decomposed statistics no longer have denominators to be dealt with and are tractable for the interpolation argument. I show for any $\varphi(\cdot) \in C_b^3(\mathbb{R})$, the space of all thrice continuously differentiable functions with bounded derivatives up to the third order, that there is a fixed constant $M > 0$ such that

$$|\mathbb{E}[\varphi(JK_I^a) - \varphi(JK_G^a)]| \leq \frac{M(a^3 \vee 1)}{\sqrt{n}}(L_2(\varphi) + L_3(\varphi)) \tag{1.4.2}$$

where $L_2(\varphi) \coloneqq \sup_x |\varphi''(x)|$ and $L_3(\varphi) \coloneqq \sup_x |\varphi'''(x)|$. By taking $\varphi(\cdot)$ to be a sequence of functions approximating the indicator function, $\mathbf{1}\{x \leq 0\}$, the result in (1.4.2) can be used to show that the cumulative distribution function (CDF) of the infeasible statistic $JK_I(\beta_0)$ can be approximated by the CDF of the Gaussian statistic $JK_G(\beta_0)$ at each point $a \in \mathbb{R}$. A Glivenko-Cantelli type argument is then be applied to show the approximation holds uniformly over all points on the real line. The Lindeberg interpolation argument on the decomposed test statistics makes use of the fact that the numerator and denominator of the Gaussian test statistic are functions of quadratic forms in the random vectors $\epsilon(\beta_0) \coloneqq (\epsilon_1(\beta_0), \ldots, \epsilon_n(\beta_0))'$

and $r \coloneqq (r_1, \ldots, r_n)'$.[1]

Moving from approximation of expectations of smooth functions to approximation of the CDF relies on a particular anticoncentration bound on $\tilde{D}$. I show that that this bound can be established under either weak or strong identification. This allows for the limiting null distribution of the test statistic under various identification regimes to be derived via a unifying argument. Additionally, even though $(N, D, \tilde{N}, \tilde{D})$ may all have nonnegligible distributions when identification is weak, the interpolation argument does not require any of these to individually converge in distribution or probability anywhere stable. This allows for a wide range of possible hat matrices H to be used in constructing the first stage estimates, $(\widehat{\Pi}_1, \ldots, \widehat{\Pi}_n)$. In particular, no assumption need be made on the number of instruments used to construct H nor any requirement imposed that the first-stage estimates $(\widehat{\Pi}_1, \ldots, \widehat{\Pi}_n)$ are consistent.

I now detail the assumptions needed for the argument. Define $\eta_i \coloneqq (\beta - \beta_0)v_i + \epsilon_i$ and $\zeta_i \coloneqq v_i - \rho(z_i)\eta_i$, noting $\eta_i = \epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)]$ and $\zeta_i = r_i - \mathbb{E}[r_i]$. In what comes below $c > 1$ can be considered an arbitrary constant that may be updated upon each use but that does not depend on sample size n.

Assumption 1.4.1 (Moment Conditions)**.** *There is a fixed constant $c > 1$ such that (i) $\{|\Pi_i| + |(\beta - \beta_0)| + |\rho(z_i)|\} \leq c$, and (ii) for any $l, k \in \mathbb{N} \cup \{0\}$ such that $l + k \leq 6$, $c^{-1} \leq \mathbb{E}[|\eta_i|^l |\zeta_i|^k] \leq c$.*

Assumption 1.4.2 (Balanced Design)**.** *(i) For $s_n^{-2} = \max_i \mathbb{E}[(\widehat{\Pi}_i^I)^2]$ the following is bounded away from zero, $c^{-1} \leq \mathbb{E}[\frac{s_n^2}{n} \sum_{i=1}^n (\widehat{\Pi}_i^I)^2]$; (ii) $\max_i s_n^2 \sum_{j \neq i} h_{ji}^2 \leq c$; and (iii) the following ratio is bounded away from zero: $\frac{\sum_{k=2}^n \lambda_k^2(HH')}{\sum_{k=1}^n \lambda_k^2(HH')} \geq c^{-1}$ where $\lambda_k(HH')$ represents the k^{th} largest eigenvalue of the matrix HH'.*

Assumptions 1.4.1 and 1.4.2 allow characterization of the null distribution of $JK(\beta_0)$. As-

[1] See Pouzo (2015) for another example of the Lindeberg interpolation method applied to approximate the distribution of quadratic forms.

sumption 1.4.1 imposes light moment conditions on the random variables η_i and ζ_i, which in turn imply restrictions on $\epsilon_i(\beta_0)$ and r_i. In particular, Assumption 1.4.1(i) imposes that $\epsilon_i(\beta_0)$ and r_i have finite means while Assumption 1.4.1(ii) bounds, both from above and away from zero, the first through sixth central moments of the random variables.

Assumption 1.4.2(i) requires that the average second moment of the infeasible first-stage estimators be on the same order as the maximum first-stage estimator second moment. This is imposed mainly to rule out hat matrices that are all zeroes or nearly all zeros so that the effective number of observations used to test the null is growing with the sample size. Remark 1.4.1 below discusses how this assumption and Assumption 1.4.2(ii) may be verified in practice. Remark 1.4.2 compares this balanced design assumption to that in the many-instruments literature (Crudu et al., 2021; Mikusheva and Sun, 2021; Matsushita and Otsu, 2022; Lim et al., 2022), noting that their balanced design neither implies nor is implied by the one in this paper.

Assumption 1.4.2(ii) requires that the maximum leverage of any observation be bounded. When H is symmetric, it is automatically satisfied under Assumption 1.4.1(i) and the definition of s_n.[2] The scaling factor s_n captures both the "size" of the elements in the hat matrix H and the strength of identification. If elements of the hat matrix are on the same order as a constant, one would expect $s_n = O(n^{-1})$ under strong identification ($\Pi_i \propto 1$) while $s_n = O(n^{-1/2})$ under weak identification ($\Pi_i \lesssim n^{-1/2}$). Assumption 1.4.2(iii) can be viewed as a technical requirement that there be more than one "effective" instrument in the hat matrix.[3] This condition can be easily verified in practice by examining the eigenvalues of HH'.

In addition to characterizing the limiting distribution of $JK(\beta_0)$ under H_0, I also examine

[2]To see this, notice that $s_n^{-2} = \max_i \mathbb{E}[(\widehat{\Pi}_i^I)^2] \geq \max_i \text{Var}(\widehat{\Pi}_i^I) = \max_i \sum_{j \neq i} h_{ij}^2 \text{Var}(r_j)$. By Assumption 1.4.1, $\text{Var}(r_j)$ is bounded from below by c^{-1}. Inverting this chain of inequalities yields that $s_n^2 \sum_{j \neq i} h_{ij}^2$ is bounded from above uniformly over all $i \in [n]$.

[3]In the case of a standard projection matrix (no deleted diagonal), Assumption 1.4.2(iii) would be satisfied whenever $\text{rank}(z(z'z)^{-1}z) > 1$.

the behavior of $JK(\beta_0)$ in local neighborhoods of the null. These local neighborhoods are characterized by the local power index P, defined below, as well as an additional regularity condition that restricts the size of $\mathbb{E}[\epsilon_i(\beta_0)]$ relative to $\mathbb{E}[r_i]$.

$$P := (\beta - \beta_0)^2 \mathbb{E}\left[\left(\frac{s_n}{\sqrt{n}} \sum_{i=1}^{n} \Pi_i \widehat{\Pi}_i^I\right)^2\right]$$

Assumption 1.4.3 (Local Identification)**.** *(i) The local power index P is bounded, $P \leq c$; and (ii)* $\max_i \mathbb{E}[(s_n \sum_{j \neq i} h_{ji} \epsilon_j(\beta_0))^2] \leq c$.

Under H_0, Assumption 1.4.3 is trivially satisfied since $(\beta - \beta_0) = 0$ and $\sum_{j \neq i} s_n^2 h_{ji}^2 \leq c$. The local power index is the second moment of the scaled numerator, N and is a measure of the association between the true first stage Π_i and the first-stage estimates $\widehat{\Pi}_i$. In Section 1.5, I discuss how the strength of this association is related to the power of the test under local alternatives. Proposition 1.4.1 below shows that when Assumption 1.4.3(ii) holds, $P \to \infty$ implies that the test based on the infeasible statistic $JK_I(\beta_0)$ is consistent.

Assumption 1.4.3(ii) is an additional technical condition that requires that the maximum value of $\mathbb{E}[(\sum_{j \neq i} h_{ji} \epsilon_j(\beta_0))^2]$ be on the same or lesser order than the maximum value of $\mathbb{E}[(\sum_{j \neq i} h_{ij} r_j)^2]$. Using the moment bounds in Assumption 1.4.1 and Assumption 1.4.2(ii) one can verify that Assumption 1.4.3(ii) is equivalent to the existence of constants $C_1, C_2 > 0$ such that

$$\begin{aligned}\max_i \Big(\sum_{j \neq i} h_{ji} \mathbb{E}[\epsilon_j(\beta_0)]\Big)^2 &\leq C_1 \max_i \mathbb{E}\big[(\sum_{j \neq i} h_{ij} r_j)^2\big] + C_2 \\ &= C_1 \max_i \left\{ \sum_{j \neq i} h_{ij}^2 \operatorname{Var}(r_j) + \Big(\sum_{j \neq i} h_{ij} \mathbb{E}[r_j]\Big)^2 \right\} + C_2\end{aligned}$$

for all $i \in [n]$. It is always satisfied whenever $\mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta - \beta_0)$ is in a $\sqrt{n}$-neighborhood of zero in the sense that $|\Pi_i(\beta - \beta_0)| \leq C/\sqrt{n}$ for all $i \in [n]$ and some constant C. In general, Assumption 1.4.3(ii) can be roughly interpreted as requiring the local neighborhoods of H_0

considered to be those in which the means of $(\epsilon_1(\beta_0), \ldots, \epsilon_n(\beta_0))$ are of the same or lesser order than the means of $(r_1, \ldots, r_n)$.

Under Assumptions 1.4.1–1.4.3, I establish a main technical lemma stating that the CDF of the infeasible statistic, $JK_I(\beta_0)$, can be uniformly approximated by the CDF of the Gaussian statistic, $JK_G(\beta_0)$. This result does not require $JK_G(\beta_0)$ to have a fixed limiting distribution.

Lemma 1.4.1 (Infeasible Uniform Approximation)**.** *Suppose that Assumptions 1.4.1–1.4.3 hold. Then,*

$$\sup_{a \in \mathbb{R}} \left| \Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a) \right| \to 0$$

I additionally show that the test based on the $JK_I(\beta_0)$ statistic is consistent whenever the power index diverges, $P \to \infty$, and Assumption 1.4.3(ii) holds.

Proposition 1.4.1 (Consistency)**.** *Suppose that Assumptions 1.4.1, 1.4.2, and 1.4.3(ii) hold. Then if $P \to \infty$ the test based on $JK_I(\beta_0)$ is consistent; i.e for any fixed $a \in \mathbb{R}$,* $\Pr(JK_I(\beta_0) \leq a) \to 0$.

The dependence of the consistency result on Assumption 1.4.3(ii) is a nontrivial restriction because of the bias taken on in constructing r_i. In particular, against certain alternatives it is possible that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$ even under strong identification. This is an extreme case, however. In general, bias in $\mathbb{E}[r_i]$ does not imply a violation of Assumption 1.4.3(ii), which requires only that the *size* of $\mathbb{E}[r_i]$ be of a weakly greater order than that of $\mathbb{E}[\epsilon_i(\beta_0)]$. Moreover, as discussed in Remark 1.4.5, Proposition 1.4.1 does not necessarily rule out consistency when $P \to \infty$ but Assumption 1.4.3(ii) fails.

Regardless, bias taken on in constructing r_i has consequences for the power of the test in finite samples. This is particularly true when the mean of r_i is of a lesser order than that of $\epsilon_i(\beta_0)$ as will be discussed in Section 1.5. To rectify this deficiency in tests based on the jackknife K-statistic, I suggest a thresholding test that decides whether to use the jackknife K-statistic or the sup-score Belloni et al. (2012a) statistic based on the value of the conditioning statistic.

This conditioning statistic, in turn, is based on a test statistic for the null hypothesis that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$.

1.4.2. Limiting Behavior of Test Statistic

The final step in characterizing the limiting behavior of the feasible test statistic is to show that the difference between the infeasible and feasible statistics is negligible. I begin with a technical lemma stating that the difference between $JK(\beta_0)$ and $JK_I(\beta_0)$ is asymptotically negligible whenever the differences between the scaled numerators and the scaled denominators are asymptotically negligible. Define these differences:

$$\Delta_N := \frac{s_n}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0)(\widehat{\Pi}_i - \widehat{\Pi}_i^I)$$
$$\Delta_D := \frac{s_n^2}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0)(\widehat{\Pi}_i^2 - (\widehat{\Pi}_i^I)^2)$$

Lemma 1.4.2. *Suppose Assumptions 1.4.1–1.4.3 hold and* $(\Delta_N, \Delta_D)' \to_p 0$. *Then* $|JK(\beta_0) - JK_I(\beta_0)| \to_p 0$.

While Lemma 1.4.2 is a simple statement, it is not obvious. In particular, showing that the difference between the infeasible and feasible statistics is negligible requires showing that $1/(D + \Delta_D)$ is bounded in probability, where D represents the scaled denominator of $JK_I(\beta_0)$. In a standard analysis, this would be done by arguing that D converges in distribution to a stable limit and then applying the continuous mapping theorem.[4] This approach is not applicable here as neither the scaled numerator nor the scaled denominator has a limiting distribution.

Instead, I directly show that $1/(D+\Delta_D)$ is bounded in probability by showing $\Pr(D \leq \delta_n) \to 0$ for any sequence $\delta_n \to 0$. This is done by first establishing that quantiles of D can be approximated by quantiles of $\tilde{D}$, the scaled denominator of $JK_G(\beta_0)$. If the variance of $\tilde{D}$

[4]This is the approach taken by Kleibergen (2002, 2005)

is bounded away from zero, its density can also be bounded with new bounds on Gaussian quadratic form densities from Götze et al. (2019), which yields the result. Otherwise, if $\text{Var}(\tilde{D}) \to 0$, the result holds by an application of Chebyshev's inequality and $\mathbb{E}[D] > c^{-1}$ from Assumption 1.4.2(i). This particular anticoncentration bound for $\tilde{D}$ is also important in the proof of Lemma 1.4.1 to establish anticoncentration for the decomposed Gaussian test statistic.

Lemma 1.4.2 allows the researcher to use alternate choices of estimators for $\rho(z_i)$, so long as they can verify that $(\Delta_N, \Delta_D)' \to_p 0$. Below, I verify that this condition can be satisfied for the ℓ_1-penalized estimation procedure proposed in (1.3.2). This requires a strengthened moment condition on η_i. Given a random variable X and $\upsilon > 0$ the Orlicz (quasi-)norm is defined

$$\|X\|_{\psi_\upsilon} \coloneqq \inf\{t > 0 : \mathbb{E}\exp(|X|^\upsilon / t^\upsilon) \leq 2\}$$

Random variables with a finite Orlicz norm for some $\upsilon \in (0, 1] \cup \{2\}$ are termed α-sub-exponential random variables (Gotze et al., 2021; Sambale, 2022). This class encompasses a wide range of potential distributions including all bounded and sub-Gaussian random random variables (with $\upsilon = 2$), all sub-exponential random variables such as Poisson or noncentral χ^2 random variables (with $\upsilon = 1$), as well as random variables with "fatter" tails such as Weibull distributed random variables with shape parameter $\upsilon \in (0, 1]$.

Assumption 1.4.4 (Estimation Error)**.** *(i) There is a fixed constant $\upsilon \in (0, 1] \cup \{2\}$ such that $\|\eta_i\|_{\psi_\upsilon} \leq c$; (ii) The basis terms $b(z_i)$ are bounded, $\|b(z_i)\|_\infty \leq C$ for all $i = 1, \ldots, n$; (iii) the approximation error satisfies $(\mathbb{E}_n[\xi_i^2])^{1/2} = o(n^{-1/2})$; (iv) the researcher has access to an estimator $\widehat{\phi}$ of ϕ that satisfies $\log(d_b n)^{2/(\upsilon \wedge 1)} \|\widehat{\phi} - \phi\|_1 \to_p 0$; (v) the following moment bounds hold*

(va) $\max_{1 \leq \ell \leq d_b} \left| \mathbb{E}\left[\frac{s_n}{\sqrt{n}} \sum_{i=1}^n \sum_{j \neq i} h_{ij} \epsilon_i(\beta_0) b_\ell(z_j) \epsilon_j(\beta_0) \right] \right| \leq c$

(vb) $\max_{\substack{1 \leq i \leq n \\ 1 \leq \ell \leq d_b}} |\mathbb{E}[s_n \sum_{j \neq i} h_{ij} b_\ell(z_j) \epsilon_j(\beta_0)]| \leq c.$

Assumption 1.4.4(i) strengthens the moment condition on η_i to require that η_i be in the class of α-sub-exponential random variables. While this condition is more restrictive than the moment condition in Assumption 1.4.1, as discussed above, it still allows for a wide range of potential distributions. Assumption 1.4.4(ii) is a standard condition in ℓ_1-penalized estimation. At the cost of extra notation, it can be relaxed and the sup-norm of the basis terms can be allowed to grow slowly with the sample size to accommodate bases such as normalized b-splines or wavelets. Assumption 1.4.4(iii) is a bound on the rate of decay of the approximation error, similar to the approximate sparsity condition of Belloni et al. (2012a).

Assumption 1.4.4(iv) is a high-level condition on the rate of consistency of the parameter estimate $\hat{\phi}$ in the ℓ_1 norm. This can be verified under approximate sparsity for both the LASSO estimator in (1.3.2) or post-LASSO procedures based on refitting an unpenalized version of (1.3.2) only using the basis terms selected in a LASSO first stage. See Belloni et al. (2012a), van der Greer (2016), Tan (2017), and Chetverikov and Sørensen (2021) for references under various choices of penalty parameter. This condition allows for the dimensionality of the basis terms, d_b, to grow near exponentially as a function of the sample size. Following the analysis of Tan (2017) one can see that, under appropriate choice of penalty parameter, this may be satisfied as long as $s^2 \log^{2(\upsilon+1)/\upsilon}(d_b n)/n \to 0$, where the sparsity index s denotes the number of nonzero elements of ϕ.

Assumption 1.4.4(v) is a strengthening of the definition of local neighborhoods and can be interpreted similarly to Assumption 1.4.3(ii). Since the moment conditions in Assumption 1.4.4(va,vb) hold with $b_\ell(z_j)\epsilon_j(\beta_0)$ replaced with r_j, Assumption 1.4.4(v) can be interpreted as requiring that $|\mathbb{E}[\sum_{j\neq i} h_{ij} b_\ell(z_j)\epsilon_j(\beta_0)]|$ is on the same order as $|\mathbb{E}[\sum_{j\neq i} h_{ij} r_j]|$ for all $i = 1, \ldots, n$ and $\ell = 1, \ldots, d_b$. As with Assumption 1.4.3(ii), it is trivially satisfied under H_0 or, using the fact that $\max_i \sum_{j\neq i} s_n^2 h_{ij}^2 \leq c$, whenever $\mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta - \beta_0)$ is in a $\sqrt{n}$-neighborhood of zero.

Under Assumptions 1.4.1–1.4.4, I establish that the difference between the infeasible and

feasible statistics can be treated as negligible when the estimation procedure proposed in (1.3.2) is used.

Lemma 1.4.3. *Suppose that Assumptions 1.4.1–1.4.4 hold. Then* $(\Delta_N, \Delta_D)' \to_p 0$.

Lemmas 1.4.1–1.4.3 are combined for the main result, local approximation of the distribution of the feasible test statistic, $JK(\beta_0)$, by the distribution of the Gaussian statistic, $JK_G(\beta_0)$. An immediate corollary is that the limiting null distribution of $JK(\beta_0)$ is χ^2_1.

Theorem 1.4.1 (Uniform Approximation)**.** *Suppose that Assumptions 1.4.1–1.4.4 hold. Then*

$$\sup_{a\in\mathbb{R}} \left|\Pr(JK(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a)\right| \to 0$$

Corollary 1.4.1 (Size Control)**.** *Suppose that Assumptions 1.4.1, 1.4.2 and 1.4.4 hold. Then, under* H_0, $JK(\beta_0) \rightsquigarrow \chi^2_1$.

If the limiting $JK_G(\beta_0)$ had a fixed distribution under H_1, Theorem 1.4.1 would follow immediately from Lemmas 1.4.1–1.4.3, and an application of Slutsky's lemma. However, under H_1, there is nothing preventing the distribution of $JK_G(\beta_0)$ changing with the sample size. Instead I establish Theorem 1.4.1 directly using the fact that both $JK(\beta_0)$ and $JK_G(\beta_0)$ are bounded in probability and that $JK_G(\beta_0)$ has a density that is bounded uniformly over n.

While $JK_G(\beta_0)$ does not have a fixed distribution, examining its behavior is still tractable and allows for insight into the power properties of the jackknife K-test. In the next section, I use this result to analyze the local power of the proposed test. To improve power against certain alternatives, I suggest a combination with the sup-score statistic of Belloni et al. (2012a).

Remark 1.4.1. A sufficient condition for Assumption 1.4.2(i) is that there is some fixed quantile $q \in (0, 100)$ such that $(cq)^{-1} \leq \frac{q^{\text{th}}\text{-quantile of } \mathbb{E}[(\widehat{\Pi}_i^I)^2]}{\max_i \mathbb{E}[(\widehat{\Pi}_i^I)^2]}$. In practice this can be verified

by checking that there is some quantile q such that both

$$\frac{q^{\text{th}}\text{-quantile of } \sum_{j\neq i} h_{ij}^2}{\max_i \sum_{j\neq i} h_{ij}^2} \quad \text{and} \quad \frac{q^{\text{th}}\text{-quantile of } (\sum_{j\neq i} h_{ij}\hat{r}_j)^2}{\max_i (\sum_{j\neq i} h_{ij}\hat{r}_j)^2} \tag{1.4.3}$$

are bounded away from zero. Similarly, Assumption 1.4.2(ii) can be verified by checking that $\max_i \sum_{j\neq i} h_{ji}^2 / \max_i \sum_{j\neq i} h_{ij}^2$ is bounded from above.

Remark 1.4.2. The balanced-design condition in Assumption 1.4.2(i) is neither weaker nor stronger than that in the many instruments literature (Crudu et al., 2021; Mikusheva and Sun, 2021; Matsushita and Otsu, 2022; Lim et al., 2022). These papers require that the projection matrix $P = \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'$ satisfies $[P]_{ii} \leq \delta \leq 1$ for some value δ and all $i \in [n]$. Since P is idempotent, $[P]_{ii} = 1$ for some $i \in [n]$ implies that $[P]_{ij} = 0$ for $j \neq i$.[5] This would not violate Assumption 1.4.2 if one were to take H such that $h_{ij} = [P]_{ij}\mathbf{1}\{i \neq j\}$; $\mathbb{E}[(\widehat{\Pi}_i^I)^2] = 0$ is allowed for a constant share of $i \in [n]$. Conversely, if the instruments are fixed or grow slowly, it is possible to construct a projection matrix P of rank d_z where $[P]_{ii}$ is bounded away from one for all $i \in [n]$, but "most" of the rows are zero. I view this as a theoretical edge case, however, that seems unlikely to result from real data.

Remark 1.4.3. The Lindeberg interpolation method allows me to give a nearly uniform explicit bound on the Gaussian approximation error. In particular, using the bound in (1.4.2), I show that for any fixed value $\Delta > 0$;

$$\sup_{a\leq\Delta} \left|\Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a)\right| \leq Cn^{-2/13}$$

where C is a constant that depends only on (c, Δ). Lemma 1.4.1 makes use of the fact that the limiting statistic $JK_G(\beta_0)$ is bounded in probability and extends this result to show that the approximation error tends to zero uniformly over the real line. While it does not account for estimation error in $\hat{\rho}(\cdot)$, obtaining an explicit bound reflects an improvement over the

[5] Since P is idempotent, $[P]_{ii} = \sum_{i=1}^n [P]_{ij}^2 = [P]_{ii}^2 + \sum_{j\neq i} [P]_{ij}^2$.

original analyses of K-statistics in Kleibergen (2002, 2005). These original studies rely on continuous mapping theorems to obtain the limiting chi-squared distributions, making the rate of decay of the approximation error difficult to analyze.

Remark 1.4.4. The interpolation argument relies on the fact that the first and second moments of $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$ are the same as the first and second moments of $(\epsilon_i(\beta_0), r_i)$ to match the first and moments of one-step deviations with Gaussian analogs. Without the jackknife form of $\widehat{\Pi}_i^I$, these one step deviations would additionally contain cross-terms such as $h_{ii} r_i \epsilon_i(\beta_0)$, for $i \in [n]$. While the first moment of this cross-term is matched by the first moment of the Gaussian analog, $h_{ii}\tilde{\epsilon}_i(\beta_0)\tilde{r}_i$, the second moment is not matched. This is manageable, however, so long as the terms h_{ii} are "small." An example of when the h_{ii} terms are small is when H is taken to be the OLS projection matrix, $H = \boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}$, and the number of instruments satisfies $d_z^3/n \to 0$. See Sections 1.10.1 and 1.11 for details.

Remark 1.4.5. Proposition 1.4.1 does not necessarily rule out that a test based on $JK_I(\beta_0)$ is consistent when $P \to \infty$ but Assumption 1.4.3(ii) fails to hold. The proof of Proposition 1.4.1 relies on showing that, when $P \to \infty$ and Assumption 1.4.3(ii) holds, $\mathbb{E}[|N|] \to \infty$ while $\text{Var}(|N|)$ and $\mathbb{E}[D]$ are bounded. These facts can be combined to show that $\Pr(N^2 - aD \leq 0) \to 0$ for any fixed $a \in \mathbb{R}$. When Assumption 1.4.3(ii) fails, $P \to \infty$ may imply that $\text{Var}(|N|) \to \infty$ as well, making the limiting behavior of the test difficult to analyze. There is reason to believe that this issue can be overcome, Andrews et al. (2004) show that the K-statistic of Kleibergen (2002) is consistent against fixed alternatives under strong identification. However, a full consistency result is not pursued here and left to future work.

Remark 1.4.6. Approximate sparsity of $\rho(z_i)$ may be a particularly palatable assumption in cases where the instrument set is generated by functions of a smaller initial set of instruments, as in Angrist and Krueger (1991), Paravisini et al. (2014), Gilchrist and Sands (2016), and Derenoncourt (2022). In these cases, the dimensionality of the basis, d_b, may not need to be much larger than the dimensionality of the instruments, d_z, to provide a good approximation

of $\rho(z_i)$. Interestingly, if taking $b(z_i) = z_i$ provides a good approximation of $\rho(z_i)$, the Tan (2017) result suggests that consistency of $\hat{\rho}(\cdot)$ is achievable under $d_z^2 \log^{2(\upsilon+1)/\upsilon}(d_z n)/n \to 0$ even if ϕ is fully dense. This requirement is weaker than the $d_z^3/n \to 0$ requirement of the standard K-statistic.

1.5. Improving Power against Certain Alternatives

Using the characterization of the limiting behavior of the test statistic derived in Section 1.4, I analyze the local power properties of the test. Unfortunately, against certain alternatives the test statistic may have trivial power, a deficiency shared with the K-statistics of Kleibergen (2002, 2005). To combat this, I propose a simple combination with the sup-score statistic of Belloni et al. (2012a) based on a thresholding rule.

1.5.1. Local Power Properties

In local neighborhoods of H_0, as defined in Assumptions 1.4.3 and 1.4.4, Theorem 1.4.1 implies that the limiting behavior of $JK(\beta_0)$ can be analyzed by examining the behavior of the Gaussian analog statistic, $JK_G(\beta_0)$. Conditional on the vector $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_n)$, the distribution of $JK_G(\beta_0)$ is nearly non-central χ_1^2 with noncentrality parameter $\mu(\tilde{r})$, $JK_G(\beta_0)|\tilde{r} \sim A^2(\tilde{r}) \cdot \chi_1^2(\mu(\tilde{r}))$:

$$A(\tilde{r}) = \frac{\sum_{i=1}^n \text{Var}(\eta_i)\tilde{\Pi}_i^2}{\sum_{i=1}^n \{\Pi_i^2(\beta-\beta_0)^2 + \text{Var}(\eta_i)\}\tilde{\Pi}_i^2}$$
$$\mu^2(\tilde{r}) = (\beta-\beta_0)^2 \frac{\left(\sum_{i=1}^n \Pi_i\tilde{\Pi}_i\right)^2}{\sum_{i=1}^n \{\Pi_i^2(\beta-\beta_0)^2 + \text{Var}(\eta_i)\}\tilde{\Pi}_i^2}.$$

Under local alternatives, the terms $\Pi_i^2(\beta-\beta_0)^2 \to 0$ so that $A(\tilde{r}) \to 1$ and $|\mu^2(\tilde{r}) - \mu_\infty^2(\tilde{r})| \to 0$, where

$$\mu_\infty^2(\tilde{r}) = (\beta-\beta_0)^2 \frac{\left(\sum_{i=1}^n \Pi_i\tilde{\Pi}_i\right)^2}{\sum_{i=1}^n \text{Var}(\eta_i)\tilde{\Pi}_i^2}. \tag{1.5.1}$$

The numerator of $\mu^2_\infty(\tilde{r})$ suggests that power is maximized when the first-stage estimate $\tilde{\Pi}_i$ is close to the true first stage value Π_i. Indeed, when errors are homoskedastic $\mu^2_\infty(\tilde{r})$ is maximized by setting $\tilde{\Pi}_i = \Pi_i$ reflecting the classical result of Chamberlain (1987). The denominator of $\mu^2_\infty(\tilde{r})$ suggests that having first-stage estimates $\tilde{\Pi}_i$ with low second moments may increase power. This guides the recommendation for the use of ℓ_2-regularization in constructing the hat matrix, H.

Unfortunately, estimators of Π_i based on $r_i = x_i - \rho(z_i)\epsilon_i(\beta_0)$ may not be close to Π_i under H_1. This is because the mean of r_i will in general differ from Π_i

$$\mathbb{E}[r_i] = \Pi_i - \rho(z_i)\Pi_i(\beta - \beta_0)$$

This deficiency is inherited from the similarity of the $JK(\beta_0)$ statistic to the K-statistic. As pointed out by Moreira (2001), this need not be an issue as long as there is a fixed constant $C \neq 0$ such that $\mathbb{E}[r_i] = C\Pi_i$ for all $i \in [n]$. However, in general, this will introduce bias into the first-stage estimates $\widehat{\Pi}_i$ under H_1. The power implications of this bias are particularly pronounced when $\rho(z_i)$ is a constant $(\beta - \beta_0) = 1/\rho(z_i)$. In this case, $\mathbb{E}[r_i]$, and thus $\mathbb{E}[\tilde{\Pi}_i]$, will equal zero for each $i \in [n]$, and the $JK(\beta_0)$ statistic will select a direction completely at random to direct power into.[1]

1.5.2. A Simple Combination Test

To combat this loss of power for tests based on the K-statistic, a common strategy is to combine the K-statistic with the Anderson-Rubin statistic based on a conditioning statistic. While the Anderson-Rubin statistic does not have optimal power on its own, it has the benefit of directing power equally in all directions avoiding the pitfalls of the K-statistic which lacks power in certain directions. Prominent examples of such tests are the conditional likelihood ratio test of Moreira (2003), the GMM-M test of Kleibergen (2005), and the minimax regret

[1] Andrews et al. (2006) and Andrews (2016) point out this deficiency in the context of the K-statistics of Kleibergen (2002, 2005).

tests of Andrews (2016). These combinations make use of the fact that the Anderson-Rubin statistic is asymptotically independent of both the K-statistic and the conditioning statistic.

Unfortunately, the asymptotic validity of these tests under heteroskedasticity is based on the assumption that $d_z^3/n \to 0$, which may not reasonably describe many settings discussed above. Instead, to improve the power of tests based on the jackknife K-statistic, I consider a simple combination with the sup-score statistic of Belloni et al. (2012a). The test based on the sup-score statistic (1.5.2) is similar in spirit to the Anderson-Rubin test but controls size even when d_z grows near exponentially as a function of the sample size.

$$S(\beta_0) := \sup_{1 \leq \ell \leq d_z} \left| \frac{\sum_{i=1}^n \epsilon_i(\beta_0) z_{\ell i}}{(\sum_{i=1}^n z_{\ell i}^2)^{1/2}} \right| \tag{1.5.2}$$

A size $\theta \in (0,1)$ test based on the sup-score statistic rejects whenever $S(\beta_0) > c_{1-\theta}^S$ where, for $e_1, \ldots, e_n$ iid standard normal and generated independently of the data, $c_{1-\theta}^S$ is the simulated multiplier bootstrap critical value:

$$c_{1-\theta}^S := (1-\theta) \text{ quantile of } \sup_{1 \leq \ell \leq d_z} \left| \frac{\sum_{i=1}^n e_i \epsilon_i(\beta_0) z_{\ell i}}{(\sum_{i=1}^n z_{\ell i}^2)^{1/2}} \right| \text{ conditional on } \{(y_i, x_i, z_i)\}_{i=1}^n.$$

As with the Anderson-Rubin test, tests based on the sup-score statistic may have suboptimal power properties in overidentified models as it does not incorporate first-stage information. However, the sup-score statistic does retain the benefit of directing power evenly in all directions, avoiding pitfalls of tests based on $JK(\beta_0)$ against certain alternatives.

The combination test will be based on an attempt to detect whether the alternative β is such that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i = 1, \ldots, n$. When this is the case, the test based on $JK(\beta_0)$ will choose a direction completely at random to direct power into. It would then be optimal for the researcher to test the null hypothesis using the sup-score statistic. Detection of whether

$\mathbb{E}[\widehat{\Pi}_i^I] = 0$ is based on the conditioning statistic:

$$C = \max_{1 \le i \le n} \left| \frac{\sum_{j \neq i} h_{ij} \hat{r}_j}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right|. \tag{1.5.3}$$

Under the assumption that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$, quantiles of the conditioning statistic can be simulated analogously to the sup-score critical value. For a new set of $e_1, \ldots, e_n$ iid standard normal and generated independently of the data, and for any $\theta \in (0, 1)$, define the conditional quantile

$$c_{1-\theta}^C \coloneqq (1-\theta) \text{ quantile of } \max_{1 \le i \le n} \left| \frac{\sum_{j \neq i} e_i h_{ij} \hat{r}_j}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right| \text{ conditional on } \{(y_i, x_i, z_i)\}_{i=1}^n \tag{1.5.4}$$

Depending on the value of the conditioning statistic, the thresholding test decides whether the test based on $JK(\beta_0)$ or one based on $S(\beta_0)$ should be run.

$$T(\beta_0; \tau) = \begin{cases} \mathbf{1}\{JK(\beta_0) > \chi^2_{1;1-\alpha}\} & \text{if } C \ge \tau \\ \mathbf{1}\{S(\beta_0) > \ c^S_{1-\alpha}\} & \text{if } C < \tau \end{cases} \tag{1.5.5}$$

for some cutoff τ, which I take in the simulation study and empirical exercise to be the 75^{th} quantile of the distribution of C under the assumption that $\mathbb{E}[\widehat{\Pi}_i^I] = 0, \forall i \in [n]$.

To show that the thresholding test controls size, I compare the rejection probability to that of a Gaussian analog. In addition to $JK_G(\beta_0)$, defined in (1.4.1), define the Gaussian analogs of $S(\beta_0)$ and the conditioning statistic C:

$$S_G(\beta_0) \coloneqq \sup_{1 \le \ell \le d_z} \left| \frac{\sum_{i=1}^n \tilde{\epsilon}_i(\beta_0) z_{\ell i}}{(\sum_{i=1}^n z_{\ell i}^2)^{1/2}} \right| \qquad C_G \coloneqq \sup_{1 \le i \le n} \left| \frac{\sum_{j \neq i} h_{ij} \tilde{r}_j}{(\sum_{j \neq i} h_{ij}^2)^{1/2}} \right|$$

where, as in Section 1.4, $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ are generated independently of each other and the data following a Gaussian distribution with the same mean and covariance matrix as $(\epsilon_i(\beta_0), r_i)$. Since $\text{Cov}(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i) = 0$ under H_0, the statistics C_G and $S_G(\beta_0)$ are independent under the

null. Similarly, the null distribution of $JK_G(\beta_0)$ is the same conditional on any realization of $(\tilde{r}_1, \ldots, r_n)$; it is also independent of C_G under the null. The Gaussian analog thresholding test decides whether the researcher should run a test based on $S_G(\beta_0)$ or $JK_G(\beta_0)$ depending on the value of C_G as in (1.5.5).

The test statistics $JK_G(\beta_0)$ and $S_G(\beta_0)$ are only marginally independent of the conditioning statistic C_G under the null. This limits the ways in which the test statistics can be combined using the conditioning statistic while still controlling size. This marginal independence in the Gaussian limit is enough, however, for the asymptotic validity of the thresholding test, $T(\beta_0; \tau)$. To establish that the behavior of the pairs $(C, JK(\beta_0))$ and $(C, S(\beta_0))$ can be approximated by the behavior of $(C_G, JK_G(\beta_0))$ and $(C_G, S_G(\beta_0))$, respectively, I rely on the following assumption:

Assumption 1.5.1 (Combination Conditions)**.** *Assume that (i) there is a* $\upsilon \in (0, 1] \cup \{2\}$ *such that* $\|\zeta_i\|_{\psi_\upsilon} \leq c$*; (ii)* $\max_{i,j} |\frac{h_{ij}}{(\mathbb{E}_n[h_{ij}^2])^{1/2}}| + \max_{l,i} |\frac{z_{li}}{(\mathbb{E}_n[z_{li}^2])^{1/2}}| \leq c$*; and (iii)* $\log^{7+4/\upsilon}(d_z n)/n \to 0$.

Assumption 1.5.1(i) is a strengthening of the moment bound on r_i similar to that of Assumption 1.4.4(i). As discussed, while more restrictive than the condition in Assumption 1.4.1, this still allows for a wide range of potential distributions for r_i. Assumption 1.5.1(ii) requires that the number of observations used to test $\mathbb{E}[\widehat{\Pi}_i] = 0$ via the conditioning statistic and the number of observations used to test the null hypothesis via the sup-score test are both growing with the sample size. It can be verified by looking at the hat matrix H and the instruments. Finally, Assumption 1.5.1(iii) is a light requirement on the number of instruments d_z needed for the validity of the sup-score test. It allows the number of instruments to grow near exponentially as a function of sample size.

Theorem 1.5.1. *Suppose Assumptions 1.4.1–1.4.4 and 1.5.1 hold. Then,*

1. *the test based on* $T(\beta_0; \tau)$ *has asymptotic size* α *for any choice of cutoff* τ*, and*
2. *if* $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ *for all* $i \in [n]$*, there exist sequences* $\delta_n \searrow 0$ *and* $\beta_n \searrow 0$ *such that with*

probability at least $1 - \delta_n$,

$$\sup_{\theta \in (0,1)} \left| \Pr_e(C \leq c^C_{1-\theta}) - (1-\theta) \right| \leq \beta_n,$$

where $\Pr_e(\cdot)$ *denotes the probability with respect to only the variables* $e_1, \ldots, e_n$.

The first part of Theorem 1.5.1 establishes the asymptotic validity of the thresholding test $T(\beta_0; \tau)$ for any choice of cutoff τ. The proof of this statement follows the logic outlined above. The second part of Theorem 1.5.1 establishes the validity of the multiplier bootstrap procedure to approximate quantiles of the conditioning statistic. It follows directly from results in Belloni et al. (2018) after verifying that the conditions needed for error taken on from estimation of $\rho(z_i)$ can treated as negligible under Assumption 1.4.4.

In Section 1.8, I investigate the power properties of the thresholding test via simulation study. I find that combining the $JK(\beta_0)$ statistic with the sup-score statistic based on C improves power against distant alternatives and helps alleviate a power decline suffered by the $JK(\beta_0)$ statistic against a particular set of alternatives.

Remark 1.5.1. As mentioned by Andrews (2016) in the context of the standard K-statistic, this attempt to rectify the power deficiency via this particular conditioning statistic is not perfect. In particular, under heteroskedasticity, the means of the partialed-out endogenous variables, $\mathbb{E}[r_i]$, may not be scaled versions of the true first stages. However, as long as $\mathbb{E}[r_i] \neq 0$, one can still expect $\mathbb{E}[\widehat{\Pi}^I_i] = \sum_{j \neq i} h_{ij} \Pi_i + (\beta - \beta_0) \sum_{j \neq i} h_{ij} \rho(z_i) \Pi_i$ to be related to the true fist stage Π_i and for the test to have nontrivial power. Moreover, in light of the dependence of the consistency result in Proposition 1.4.1 on Assumption 1.4.3(ii), in the case where $\mathbb{E}[\widehat{\Pi}_i] = 0$ for all $i \in [n]$ it may be particularly important to avoid using the jackknife K-statistic to test H_0.

1.6. Analysis with Multiple Endogenous Variables

To analyze the limiting behavior of the test statistic when $d_x > 1$, I follow the basic idea of Section 1.4, which is to show that quantiles of the jackknife K-statistic can be approximated by analogous quantiles of the Gaussian statistic:

$$JK_G(\beta_0) \coloneqq \tilde{\epsilon}(\beta_0)\tilde{\Pi}(\tilde{\Pi}_\epsilon'\tilde{\Pi}_\epsilon)^{-1}\tilde{\Pi}'\tilde{\epsilon}(\beta_0);$$

where $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$ are Gaussian with the same mean and covariance matrix as $(\epsilon_i(\beta_0), r_i)'$ and for $\tilde{\Pi}_{\ell i} = \sum_{j \neq i} h_{ij}\tilde{r}_{\ell j}$ define $\tilde{\Pi}_i \coloneqq (\tilde{\Pi}_{1i}, \ldots, \tilde{\Pi}_{d_x i})' \in \mathbb{R}^{d_x}$, $\tilde{\Pi}_{\epsilon i} \coloneqq (\mathbb{E}[\epsilon_i^2(\beta_0)])^{1/2}\tilde{\Pi}_i$, and

$$\begin{aligned}
\tilde{\epsilon}(\beta_0) &\coloneqq (\tilde{\epsilon}_1(\beta_0), \ldots, \tilde{\epsilon}_n(\beta_0))' \in \mathbb{R}^n \\
\tilde{\Pi} &\coloneqq (\tilde{\Pi}_1, \ldots, \tilde{\Pi}_n)' \in \mathbb{R}^{n \times d_x} \\
\tilde{\Pi}_\epsilon &\coloneqq (\tilde{\Pi}_{\epsilon 1}, \ldots, \tilde{\Pi}_{\epsilon n})' \in \mathbb{R}^{n \times d_x}
\end{aligned}$$

As in Section 1.4, notice that, since uncorrelated random variables are independent, under H_0 the vector $\tilde{\epsilon}(\beta_0)$ is mean zero and independent of $(\tilde{\Pi}, \tilde{\Pi}_\epsilon)$. Conditional on any realization of $(\tilde{\Pi}, \tilde{\Pi}_\epsilon)$ the $JK_G(\beta_0)$ statistic then follows a $\chi^2_{d_x}$ distribution, and thus, its unconditional distribution is also $\chi^2_{d_x}$.

In addition to characterizing the local behavior of $JK(\beta_0)$ with multiple endogenous variables, I show that the thresholding test of Section 1.5.2 can be applied with multiple endogenous variables with a generalized conditioning statistic.

1.6.1. Modified Interpolation Approach

As with a single endogenous variable, error taken on from the estimation of $\rho(z_i)$ prevents immediate comparison of $JK(\beta_0)$ to $JK_G(\beta_0)$. Instead as an intermediate step consider showing that the quantiles of $JK_I(\beta_0)$ can be approximated by corresponding quantiles of

$JK_G(\beta_0)$ where $JK_I(\beta_0)$ is an infeasible statistic:

$$JK_I(\beta_0) \coloneqq \epsilon(\beta_0)(\widehat{\Pi}^I)((\widehat{\Pi}^I_\epsilon)'(\widehat{\Pi}^I_\epsilon))^{-1}(\widehat{\Pi}^I)'\epsilon(\beta_0),$$

for $\widehat{\Pi}^I$ and $\widehat{\Pi}^I_\epsilon$ defined the same way as $\widehat{\Pi}$ and $\widehat{\Pi}_\epsilon$ in (1.3.4), respectively, but using the true values $(r_1, \ldots, r_n)'$ in place of their estimates $(\hat{r}_1, \ldots, \hat{r}_n)'$.

When there are multiple endogenous variables, $d_x > 1$, I cannot take advantage of the simplified form of the test statistic to establish this approximation as in Section 1.4. Instead I deal directly with the test statistics themselves. Consider functions $\varphi_\gamma(\cdot) \in C^3_b(\mathbb{R})$ that approximate the indicators $\mathbf{1}\{\cdot \leq a\}$, where $a \in \mathbb{R}$ is arbitrary and γ is a scaling factor inversely proportional to the quality of the approximation but positively proportional to the derivatives of φ_γ. The goal is to show, for a sequence γ_n tending to zero, that

$$\mathbb{E}[\varphi_{\gamma_n}(JK_I(\beta_0)) - \varphi_{\gamma_n}(JK_G(\beta_0))] \to 0 \tag{1.6.1}$$

The classical interpolation argument of Lindeberg (1922) would attempt to show (1.6.1) by one-by-one replacement of each pair, $(\epsilon_i(\beta_0), r_i)'$, in the expression of $\varphi_{\gamma_n}(JK_I(\beta_0))$ with its Gaussian analog, $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)'$, and bounding of the size of each of these deviations. As mentioned in Section 1.4, the problem arises as the derivative of the test statistic, $JK_I(\beta_0)$, with respect to terms in the denominator matrix, $\widehat{\Pi}'_\epsilon\widehat{\Pi}_\epsilon$, may be as large as the inverse of the minimum eigenvalue of the denominator matrix. When identification is sufficiently weak, the denominator matrix will have a nonnegligible distribution and the inverse of its minimum eigenvalue may not have finite moments.

To get around this, I modify the argument by considering a "data-dependent" choice of approximation parameter γ_n. This choice of approximation parameter inversely scales with the determinant of the denominator matrix and thus, since the determinant is the product of

the eigenvalues, inversely scales with the minimum eigenvalue.[1] Geometrically, this approach can be thought of as "stretching out" the function $\varphi_{\gamma_n}(\cdot)$ in directions where the minimum eigenvalue of the denominator matrix is close to zero. Since the overall derivatives of $\varphi_{\gamma_n}(JK_I(\beta_0))$ with respect to $(\epsilon_i(\beta_0), r_i)'$ depend on the product of derivatives with respect to the test statistic and derivatives of $\varphi_{\gamma_n}(\cdot)$, which scale inversely with the approximation parameter, this adjustment of the approximation parameter allows control of the overall derivative. Details of this approach can be found in Section 1.10.4.

This approach relies on stronger moment conditions, which I detail below. These strengthened moment conditions are needed mainly needed to bound moments of the determinant of the denominator matrix. For all $\ell = 1, \ldots, d_x$ let $\zeta_{\ell i} \coloneqq v_i - \rho_\ell(z_i)\eta_i$, noting that $\zeta_{\ell i} = r_{\ell i} - \mathbb{E}[r_{\ell i}]$. Recall also the definition of $\eta_i = \epsilon_i - v_i'(\beta - \beta_0)$, which is equal to $\epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)]$.

Assumption 1.6.1 (Moment Conditions)**.** *Assume (i) there are constants $c > 1$ and $\upsilon \in (0, 1] \cup \{2\}$ such that $\|\epsilon_i\|_{\psi_a} \leq c$ and $\|\zeta_{\ell i}\|_{\psi_\upsilon} \leq c$, and (ii) $c^{-1} \leq \lambda_{\min}(\mathbb{E}[\eta_i\eta_i']) \leq \lambda_{\max}(\mathbb{E}[\eta_i\eta_i']) \leq c$.*

Assumption 1.6.2 (Balanced Design)**.** *(i) For any $\ell = 1, \ldots, d_x$ let $s_{\ell,n}^{-2} = \max_{1\leq i\leq n} \mathbb{E}[(\widehat{\Pi}_{\ell i}^I)^2]$; then, the minimum eigenvalue of the following matrix is bounded away from zero:*

$$c^{-1} \leq \lambda_{\min}\mathbb{E}\left(\frac{s_{\ell,n}s_{k,n}}{n}\sum_{i=1}^n(\widehat{\Pi}_{\ell i}^I)(\widehat{\Pi}_{ki}^I)\right)_{\substack{1\leq\ell\leq d_x\\ 1\leq k\leq d_x}}$$

(ii) $\max_i s_n \sum_{j\neq i} h_{ji}^2 \leq c$; and (iii) the following is bounded away from zero: $\frac{\sum_{k=2}^n \lambda_k^2(HH')}{\sum_{k=1}^n \lambda_k^2(HH')} \geq c^{-1}$ where $\lambda_k(HH')$ represents the k^{th} largest eigenvalue of the matrix HH'.

Assumption 1.6.1(i) strengthens Assumption 1.4.1 to require that the random variables (η_i, ζ_i), and thus, by extension, $(\epsilon_i(\beta_0), r_i)$ are υ-sub-exponential. As discussed below Assumption 1.4.4 this is more restrictive than the finite sixth moments needed to establish Lemma 1.4.1 but

[1]The determinant has the benefit of being a smooth function of elements of the matrix. This makes it nicer to work with than the minimum eigenvalue itself, which loses differentiability when the dimension of its eigenspace is larger than one.

still allows for a wide range of possible distributions. Assumption 1.6.1(ii) is a light regularity condition requiring that the random variables $(\eta_{1i}, \ldots, \eta_{d_x i})$ be linearly independent.

Assumption 1.6.2(i) is a natural extension of Assumption 1.4.2(i) to the setting where $d_x > 1$. It requires that the average second moment of any linear combination of the first-stage estimates is proportional to the maximum second moment of the same linear combination. Assumption 1.6.2(ii,iii) are the same conditions as Assumption 1.4.2(ii,iii) and can again be implicitly thought of as requiring that the maximum leverage of any one observation be bounded and there be than two effective instruments in the hat matrix. Assumption 1.6.2 thus reduces to Assumption 1.4.2 when $d_x = 1$.

Assumption 1.6.3 (Local Identification)**.** *(i) The local power index is bounded $P \leq c$ for*

$$P = \sum_{\ell=1}^{d_x} \mathbb{E}\left[\left(\frac{s_{\ell,n}}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\Pi}_{\ell i}^{I} \Pi_i'(\beta - \beta_0)\right)^2\right]$$

(ii) $\mathbb{E}[(s_{n,\ell} \sum_{j \neq i} h_{ji} \epsilon_j(\beta_0))^2] \leq c$ *for all* $\ell = 1, \ldots, d_x$.

Lemma 1.6.1 (Infeasible Uniform Approximation)**.** *Suppose that Assumptions 1.6.1–1.6.3 hold. Then*

$$\sup_{a \in \mathbb{R}} |\Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0)) \leq a)| \to 0$$

1.6.2. Limiting Behavior of Test Statistic

Having derived the limiting behavior of the infeasible statistic, I next present a high-level condition under which estimation error taken on from estimation of $\rho(z_i)$ can be treated as negligible. I then verify this high-level condition for the ℓ_1-regularized estimators proposed in (1.3.2). For any $\ell = 1, \ldots, d_x$ define the scaled differences

$$\Delta_{N,\ell} := \frac{s_{\ell,n}}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0)(\widehat{\Pi}_{\ell,i} - \widehat{\Pi}_{\ell,i}^{I})$$

$$\Delta_{D,\ell} \coloneqq \frac{s_{\ell,n}^2}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0)(\widehat{\Pi}_{\ell,i}^2 - (\widehat{\Pi}_{\ell,i}^I)^2)$$

As long as these scaled differences tend to zero, Lemma 1.6.2 shows that the difference between the feasible and infeasible test statistics converges to zero:

Lemma 1.6.2. *Suppose that Assumptions 1.6.1–1.6.3 hold and that* $(\Delta_{N,\ell}, \Delta_{D,\ell}) \to_p 0$ *for all* $\ell = 1, \ldots, d_x$. *Then* $|JK(\beta_0) - JK_I(\beta_0)| \to_p 0$.

As with Lemma 1.4.2, while Lemma 1.6.2 is a simple statement, it is not immediate. In particular, establishing Lemma 1.6.2 requires showing that $\lambda_{\max}(D^{-1})$ is bounded in probability, where D represents a scaled version of the denominator matrix. This requires some work as the scaled denominator matrix is not required to converge in distribution to a stable limit. Instead I directly show that $\lambda_{\max}(D^{-1})$ is bounded in probability by showing that $\Pr(\lambda_{\min}(D) \leq \delta_n) \to 0$ for any sequence $\delta_n \to 0$.

To do this, I first demonstrate that it is sufficient to show that $\Pr(a'Da \leq \delta_n) \to 0$ for any $\delta_n \to 0$ and fixed $a \in \mathcal{S}^{d_x - 1} = \{v \in \mathbb{R}^{d_x} : \|v\| = 1\}$. I then establish the claim for an arbitrary choice of a. As in Lemma 1.4.2 I do this by comparing the scaled quadratic form of the denominator matrix to a Gaussian analog and then establishing the corresponding result for the Gaussian analog. This corresponding result is again also useful for establishing the validity of the interpolation approach with a dynamic choice of approximation parameter.

I state conditions under which $(\Delta_{N,\ell}, \Delta_{D,\ell}) \to_p 0$ holds for the ℓ_1-regularized estimation procedure proposed in (1.3.2). These conditions are equivalent to those in Assumption 1.4.4 but hold for each the d_x estimation procedures.

Assumption 1.6.4 (Estimation Error)**.** *(i) The basis terms* $b(z_i)$ *are bounded,* $\|b(z_i)\|_\infty \leq C$ *for all* $i = 1, \ldots, n$; *(ii) the approximation error satisfies* $(\mathbb{E}_n[\xi_{\ell i}^2])^{1/2} = o(n^{-1/2})$; *(iii) the researcher has access to estimators* $\widehat{\phi}_\ell$ *of* ϕ_ℓ *that satisfy* $\log(d_b n)^{2/(\upsilon \wedge 1)} \|\widehat{\phi}_\ell - \phi_\ell\|_1 \to_p 0$ *for each* $\ell \in [d_x]$; *and (iv) locally identified in the sense that*

(iva) $\max_{\substack{1\le\ell\le d_x\\ 1\le k\le d_b}} \left|\mathbb{E}\left[\frac{s_{n,\ell}}{\sqrt{n}}\sum_{i=1}^n\sum_{j\neq i} h_{ij}\epsilon_i(\beta_0)b_k(z_j)\epsilon_j(\beta_0)\right]\right| \le c$

(ivb) $\max_{\substack{1\le i\le n\\ 1\le\ell\le d_b}} |\mathbb{E}[s_{n,\ell}\sum_{j\neq i} h_{ij}b_\ell(z_j)\epsilon_j(\beta_0)]| \le c.$

Under Assumption 1.6.4 the conditions of Lemma 1.6.2 are satisfied. If these conditions are satisfied, Lemmas 1.6.1 and 1.6.2 can be combined to analyze the behavior of $JK(\beta_0)$ statistics in local neighborhoods of the null.

Theorem 1.6.1 (Uniform Approximation)**.** *Suppose that Assumptions 1.6.1–1.6.4 hold. Then,*

$$\sup_{a\in\mathbb{R}} |\Pr(JK(\beta_0)\le a) - \Pr(JK_G(\beta_0)\le a)| \to 0$$

In particular, under H_0, $JK(\beta_0) \rightsquigarrow \chi^2_{d_x}$.

As in Lemma 1.4.1, the result in Theorem 1.6.1 does not require $JK_G(\beta_0)$ to have a stable limiting distribution under H_1.

1.6.3. Improving Power against Certain Alternatives

As discussed in Section 1.5.1, tests based on the jackknife K-statistic may suffer from suboptimal power properties. These properties are particularly bad whenever $\mathbb{E}[\widehat{\Pi}_{\ell i}] = 0$ for some $\ell \in [d_x]$ and all $i \in [n]$. To improve power in this direction, I propose a generalization of the thresholding test in Section 1.5.2 based on the conditioning statistic C

$$C := \min_{1\le\ell\le d_x}\max_{1\le i\le n}\left|\frac{\sum_{j\neq i} h_{ij}\hat{r}_{\ell j}}{(\sum_{j\neq i} h_{ij}^2)^{1/2}}\right| \tag{1.6.2}$$

The conditioning statistic C attempts to detect whether, for *some* $\ell \in [d_x]$, $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0$ for all $i \in [n]$. Under the assumption that $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0, \forall i \in [n], \ell \in [d_x]$, quantiles of C can be simulated by multiplier bootstrap. Let $e_1, \ldots, e_n$ be generated iid standard normal independent of the data and for any $\theta \in (0,1)$, define the conditional bootstrap quantile:

$$c^C_{1-\theta} := (1-\theta) \text{ quantile of } \min_{1\le\ell\le d_x}\max_{1\le i\le n}\left|\frac{\sum_{j\neq i} e_j h_{ij}\hat{r}_j}{(\sum_{j\neq i} h_{ij}^2)^{1/2}}\right| \text{ conditional on } \{(y_i,x_i,z_i)\}_{i=1}^n$$

Based on the value of the conditioning statistic the researcher can decide whether to run a test based on $JK(\beta_0)$ or a test based on the sup-score statistic $S(\beta_0)$.

$$T(\beta_0;\tau) := \begin{cases} \mathbf{1}\{JK(\beta_0) > \chi^2_{d_x;1-\alpha}\} & \text{if } C > \tau \\ \mathbf{1}\{S(\beta_0) > c^S_{1-\alpha}\} & \text{if } C \leq \tau \end{cases} \tag{1.6.3}$$

As with Theorem 1.5.1, I show the asymptotic validity of the thresholding test by first establishing that quantiles of $(JK(\beta_0), C)$ and $(S(\beta_0), C)$ can jointly be approximated by Gaussian analogs and then using the marginal independence of the Gaussian analog testing and conditioning statistics under the null; $(JK(\beta_0) \perp C)$ and $(S\beta_0) \perp C)$ under H_0.

Theorem 1.6.2. *Suppose that Assumptions 1.5.1(ii,iii), 1.6.1, 1.6.2, and 1.6.4 hold. Then,*

1. *the test based on $T(\beta_0;\tau)$ has asymptotic size α for any choice of cutoff τ, and*
2. *if $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0$ for all $i \in [n]$ and $\ell \in [d_x]$, there exist sequences $\delta_n \searrow 0$ and $\beta_n \searrow 0$ such that with probability at least $1 - \delta_n$,*

$$\sup_{\theta \in (0,1)} |\Pr_e(C \leq c^C_{1-\theta}) - (1-\theta)| \leq \beta_n$$

where $\Pr_e(\cdot)$ denotes the probability with respect to only the variables $e_1, \ldots, e_n$.

The first part of Theorem 1.6.2 establishes the validity of the test based on the thresholding statistic for any choice of cutoff τ. In practice, I recommend taking the cutoff, τ, to be the 75^{th} quantile of the distribution of C under the assumption that $\mathbb{E}[\widehat{\Pi}^I_{\ell i}] = 0$ for all $\ell \in [d_x]$ and $i \in [n]$. The second part of Theorem 1.6.2 establishes that this quantile can be simulated via the multiplier bootstrap procedure described above.

1.7. Empirical Application

I apply the testing procedures proposed in this paper to the data of Gilchrist and Sands (2016), who seek to determine the effect of social spillovers in movie consumption. The sample consists of all 1,671 opening weekend days[1] between January 1, 2002 and January 1, 2012. For each opening weekend, the authors observe gross ticket sales for all movies wide released in theaters in the United States.[2] The data are obtained through Box Office Mojo, a subsidiary of the Internet Movie Database (IMDb). To focus on movies in theaters long enough for social spillovers to be a relevant factor, the authors consider only movies that remain in theaters for at least six weeks.

The outcome variables of interest are gross ticket sales of movies that opened in a given weekend in the second through sixth weeks of their run, while the endogenous variable is the gross ticket sales of a movie in its opening weekend. To control for seasonal periodicity in both the supply of and demand for movies, a vector of date controls are included. Formally, Gilchrist and Sands (2016) are interested in the parameters β_w, $w = 2, \ldots, 7$ from the linear IV model(s):

$$\text{Sales}_{wi}^{\perp} = \beta_w \text{Sales}_{1i}^{\perp} + \epsilon_{wi} \tag{1.7.1}$$

where, for $i = 1, \ldots, 6$, $\text{Sales}_{wi}^{\perp}$ represents gross national ticket sales, after the partialing out of date controls and a constant, $7w$ days after day i, of movies that opened on the opening weekend of i. The variable $\text{Sales}_{7i}^{\perp} = \sum_{w=1}^{6} \text{Sales}_{wi}^{\perp}$ denotes the cumulative national ticket sales from the second through sixth running weekends of movies who opened in weekend i, after the partialing out of date controls and a constant. The parameter β_w represents the social spillover effect of strong opening weekend sales on sales in later weeks; more people seeing a movie on its opening weekend will mean more people telling their friends about the movie potentially leading to larger sales later on.

[1] An opening weekend day is a Friday, Saturday, or Sunday of opening weekend.
[2] A wide released movie is any movie that ever shows on 600 or more screens.

Because movies with high first-week sales may have high sales in succeeding weeks for reasons other than word of mouth spillover effects (e.g the movie may receive positive critical reviews prerelease or be part of a previously successful franchise), the parameter β_w cannot be plausibly recovered from ordinary least squares regression of $\text{Sales}_{wi}^{\perp}$ on $\text{Sales}_{1i}^{\perp}$. To identify the structural parameter, Gilchrist and Sands (2016) employ a vector of nationally aggregated weather measures. These weather measures reflect the proportion of movie theaters experiencing a particular type of weather on a particular weekend. The measures include the proportion of movie theaters experiencing maximum temperatures in 5° Fahrenheit bins on the interval $[10°, 100°]$, the proportion of movie theaters experiencing precipitation levels in 0.25 inch per hour increments on the interval $[0, 1.5]$, and the proportions of theaters experiencing any type of snow and of theaters experiencing any type of rain.

The nationally aggregated weather conditions on opening weekend days serve as plausibly exogenous instrumental variables, affecting ticket sales in later weeks only through their effect on opening-weekend-day sales. Same-day weather conditions may also have an effect on movie ticket sales: when the weather is particularly nice, people may be more inclined to engage in outdoor activities while in poorer, weather people may choose to stay indoors and see a new movie. Putting together the nationally aggregated weather measures leaves Gilchrist and Sands (2016) with a vector of 52 instrumental variables. After the partialing out a constant and the date controls, four of these are linearly dependent. I discard these and work with the remaining 48 partialed-out instruments in my analysis.

To handle the large number of instruments, the authors follow Belloni et al. (2012a) and employ a post-LASSO estimate of the first stage. In their main specifications, they set the first-stage penalty parameter so that the number of instrument selected is one, two, or three. The resulting first-stage F-statistics using the selected instrument(s), 38.80, 25.86, and 20.95, respectively, seem to indicate strong identification.[3] However, the first-stage F-statistic on the full set of instrumental variables is only 3.80. Moreover, since the LASSO objective is an

[3]Typical empirical practice is to use the Wald test when the first stage F-statistic is larger than 10.

ℓ_1 penalized version of the OLS loss, using the variables selected by LASSO may mechanically lead to higher F-statistics even if the underlying relationship between the instruments and the endogenous variables is weak.

Figure 1.7.1 provides evidence from a simple simulation experiment to demonstrate this. For the simulation experiment I generate an iid sample of 10 IVs, $\{Z_{1i}, \ldots, Z_{10i}\}_{i=1}^{n}$ from a normal distribution with a Toeplitz covariance structure, $\text{Cov}(Z_{\ell i}, Z_{ki}) = (1.1)^{-|j-k|}$, $1 \leq j, k \leq 10$. The endogenous variable is generated to only have a weak relationship with the instruments $X_i = \frac{1}{\sqrt{n}} \sum_{\ell=1}^{10} 0.7 \cdot Z_{\ell i} + v_i$, where the first-stage errors v_i are independent standard normals. From this initial set of 10 instrumental variables I generate an additional 55 technical instruments by squaring and taking all interactions between variables in the initial set. These generated instruments are correlated with the initial instruments but do not directly enter the first stage.

I then set the LASSO penalty so that only a certain number of instruments are chosen and report the resulting average first stage F-statistics over one thousand simulations. As seen in Figure 1.7.1, these first-stage F-statistics increase significantly as the number of selected instruments decreases. While the "true" F-statistic, computed with only the 10 initial instruments directly relevant for the first stage, is only 5.234, the average F-statistic on the selected variables can be larger than 40. The persistence of this pattern between sample sizes $n = 500$ and $n = 1000$ suggests that this is not a small-sample issue and that pretesting for weak identifications based on post-LASSO F-statistics may be problematic generally. Figure 1.7.2 shows how the first stage F-statistic changes with the number of LASSO-selected variables in the Gilchrist and Sands (2016) data. The pattern is similar to that seen in the Figure 1.7.1 simulation experiment.

Given a lack of clarity on the strength of identification, I seek to validate the results of Gilchrist and Sands (2016) using the weak identification testing procedures proposed in this paper. The setting is particularly suitable for weak IV testing using the jackknife K-statistic.

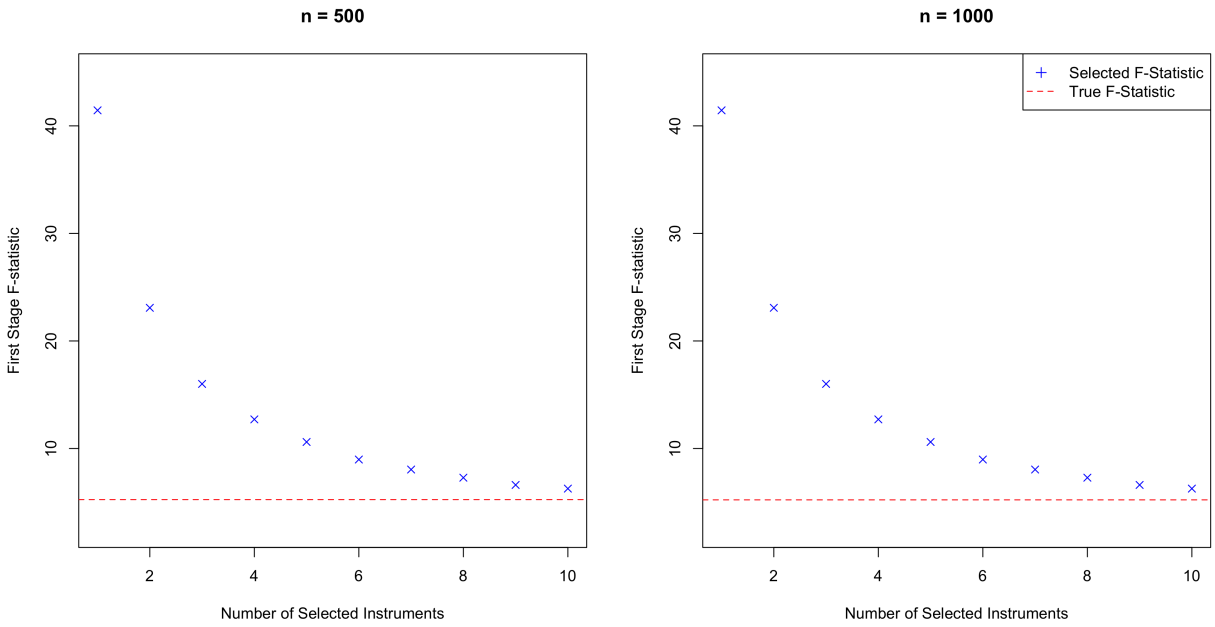

Figure 1.7.1: Results from Simulation Experiment. The endogenous variable is generated to be weakly related to a set of ten initial instruments. I take quadratic powers and interactions of these ten initial instruments to create an additional 55 technical instruments that do not directly enter the first stage. The LASSO penalty is then set to select a certain number of variables and I report the resulting average post-LASSO F-statistics over 1000 simulations. The average F-statistic using only the relevant ten initial instruments is 5.234 for both $n = 500$ and $n = 1000$.

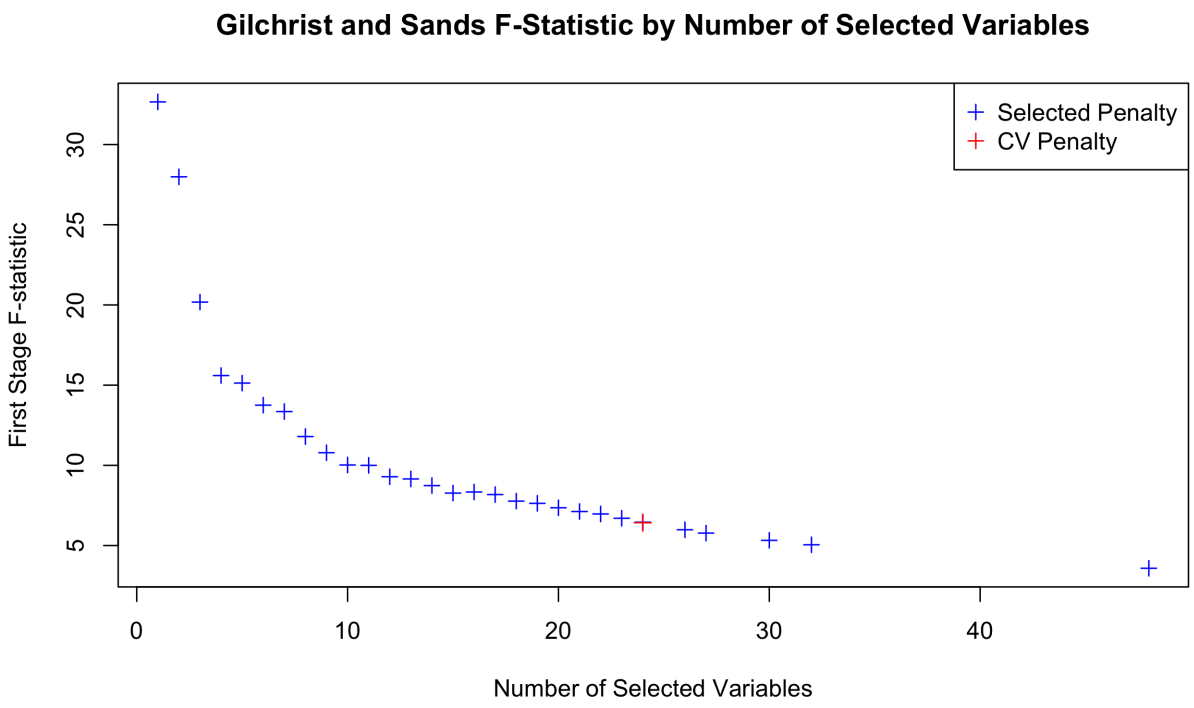

Figure 1.7.2: First-Stage F-statistic as Function of Number of LASSO-Selected Variables in the Data of Gilchrist and Sands (2016). When selecting variables using a cross-validated choice of LASSO penalty parameter, the first-stage F-statistic is 6.42.

With 48 instruments and a sample size of 1671, $d_z^3 = 110{,}592 \gg n$, making the tests of Moreira (2003, 2009), Kleibergen (2005), and Andrews (2016) inapplicable. On the other hand, it is unclear whether asymptotic approximations based on $d_z \to \infty$ will accurately describe the finite-sample distribution of test statistics with 48 instruments. Moreover, since fluctuations in movie theater attendance seem to be largely driven by either particularly cold or particularly hot weather (see Figure 4 in Gilchrist and Sands (2016)), the nuisance parameter $\rho(z_i)$ is plausibly approximately sparse.

Table 1.7.1 compares the 95% confidence intervals for $\beta_1, \ldots, \beta_7$ generated by the jackknife K test to the confidence intervals generated by the sup-score test of Belloni et al. (2012a) and the jackknife LM test (JLM) test of Matsushita and Otsu (2022). I form these confidence intervals by running the tests for each β_0 on a 300 point grid between zero and two and inverting the results; a point β_0 is included in the 95% confidence interval if the test fails to reject the null that $\beta_w = \beta_0$ at level $\alpha = 0.05$. For the $JK(\beta_0)$ statistic I use the choice of hat matrix in (1.3.3) and estimate the auxiliary parameter $\rho(z_i)$ as in (1.3.2). The penalty parameter λ is chosen with leave-one-out cross-validation using the `cv.glmnet` command from the `glmnet` package in R (R Core Team, 2021; Friedman et al., 2010). The critical value for the sup-score statistic $S(\beta_0)$ is simulated using 2,500 bootstrap draws. Confidence intervals based on the combination test, $T(\beta_0; \tau)$, are not directly reported as the pretesting procedure based on simulating the 75^{th} quantile of C as in (1.5.4) always suggests using the $JK(\beta_0)$ statistic.

For reference, I also provide point estimates and standard errors for $\beta_1, \ldots, \beta_7$ from Gilchrist and Sands (2016), Table 2. To facilitate comparison, these point estimates and standard errors come from a specification that uses all the instruments in the first stage of a 2SLS procedure. While the Gilchrist and Sands (2016) point estimates are always in the 95% confidence intervals generated by the $JK(\beta_0)$ and JLM tests, the confidence intervals from the identification-robust procedures are significantly wider than those generated with the 2SLS

standard errors. Interestingly, the confidence intervals from inverting the jackknife K-test tend to be quite similar to the confidence intervals from the JLM test. This is surprising given the distinct forms of the $JK(\beta_0)$ and the JLM test statistics.

For the parameters $\beta_2, \beta_4, \beta_5$, and β_6, the confidence intervals generated by the sup-score statistic are empty while the sup-score confidence interval for β_2 is nearly empty. This is also the case when using the jackknife AR-statistic of Crudu et al. (2021) and Mikusheva and Sun (2021), whose confidence intervals are not reported as they are always empty. With 48 instruments and a single parameter the linear IV model in (1.7.1) is overidentified and as such the empty confidence intervals could be interpreted as evidence of model misspecification. For the parameter β_7 the confidence interval generated by inverting the sup-score statistic is not empty and is instead 36% larger than the $JK(\beta_0)$ confidence interval and 41% larger than the JLM confidence interval. This result suggests that the jackknife K tests and JLM tests may have better power properties than the sup-score test in this setting.

Parameter	β_2	β_3	β_4	β_5	β_6	β_7
Estimate (s.e.)	0.475 (0.024)	0.269 (0.023)	0.164 (0.017)	0.121 (0.013)	0.093 (0.010)	1.222 (0.074)
$JK(\beta_0)$	$[0.436, 0.557]$	$[0.227, 0.334]$	$[0.134, 0.214]$	$[0.100, 0.167]$	$[0.080, 0.134]$	$[1.003, 1.391]$
$S(\beta_0)$	$\emptyset$	$[0.294, 0.334]$	$[0.087, 0.094]$	$\emptyset$	$\emptyset$	$[0.990, 1.518]$
JLM	$[0.436, 0.557]$	$[0.227, 0.334]$	$[0.134, 0.214]$	$[0.107, 0.167]$	$[0.087, 0.134]$	$[1.010, 1.384]$

Table 1.7.1: 95% Confidence Intervals based on inverting various test statistics. Instrument set used is the same as the Gilchrist and Sands (2016) instrument set less four collinear instruments; $d_z = 48$ with $n = 1,671$. Thresholding test confidence intervals are not reported as they coincide with confidence intervals for $JK(\beta_0)$.

Tables 1.7.2 and 1.7.3 repeat the analysis of Table 1.7.1 but with alternative instrument sets. The confidence intervals of Table 1.7.2 use only 5° Fahrenheit temperature bins ($d_z = 36$) while the confidence intervals of Table 1.7.3 include all the instruments used in Table 1.7.1 and all interactions between the 5° Fahrenheit temperature bins and the other weather measures for a total of 524 instruments.[4] For the most part, the confidence intervals generated by inverting the jackknife K-statistic are similar across Tables 1.7.1-1.7.3. The confidence intervals for

[4]The instrument set of Table 1.7.3 does not include interactions between temperature bins nor interactions between other weather measures.

the jackknife LM statistic however, become much narrower when using the largest set of instruments is used. This is interesting as the results from the $JK(\beta_0)$ test as well as the power analysis in Section 1.5 seem to suggest that use of the extra isntruments does not lead to better first-stage estimates. Interestingly, the JLM confidence intervals in for β_6, β_7 in Table 1.7.3 do not contain the point estimates for β_6 and β_7 from Gilchrist and Sands (2016). As with Table 1.7.1, Tables 1.7.2 and 1.7.3 do not report confidence intervals from $T(\beta_0; \tau)$ as these always agree with the $JK(\beta_0)$ confidence intervals and do not report jackknife AR confidence intervals as these are always empty.

Parameter	β_2	β_3	β_4	β_5	β_6	β_7
Estimate (s.e.)	0.475 (0.024)	0.269 (0.023)	0.164 (0.017)	0.121 (0.013)	0.093 (0.010)	1.222 (0.074)
$JK(\beta_0)$	$[0.449, 0.597]$	$[0.255, 0.389]$	$[0.148, 0.248]$	$[0.114, 0.194]$	$[0.094, 0.154]$	$[1.086, 1.555]$
$S(\beta_0)$	$\emptyset$	$[0.302, 0.329]$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
JLM	$[0.449, 0.597]$	$[0.255, 0.389]$	$[0.154, 0.248]$	$[0.114, 0.194]$	$[0.094, 0.154]$	$[1.092, 1.555]$

Table 1.7.2: 95% Confidence Intervals based on inverting various test statistics. Instrument set used includes only temperatures measures; $d_z = 36$, with $n = 1,671$. Thresholding test condidence intervals are not reported as they coincide with confidence intervals for $JK(\beta_0)$.

Parameter	β_2	β_3	β_4	β_5	β_6	β_7
Estimate (s.e.)	0.475 (0.024)	0.269 (0.023)	0.164 (0.017)	0.121 (0.013)	0.093 (0.010)	1.222 (0.074)
$JK(\beta_0)$	$[0.443, 0.604]$	$[0.215, 0.342]$	$[0.094, 0.228]$	$[0.087, 0.154]$	$[0.054, 0.121]$	$[0.916, 1.435]$
$S(\beta_0)$	$[0.416, 0.477]$	$\emptyset$	$\emptyset$	$[0.034, 0.121]$	$[0.121, 0.208]$	$[0.918, 1.562]$
JLM	$[0.463, 0.497]$	$[0.268, 0.282]$	$[0.161, 0.174]$	$[0.101, 0.107]$	$[0.063, 0.084]$	$[1.059, 1.137]$

Table 1.7.3: 95% Confidence Intervals based on inverting various test statistics. Instrument set used includes the original instrument set along with interactions of the temperature measures set with all other aggregated weather measures; $d_z = 524$, with $n = 1,671$. Thresholding test confidence intervals are not reported as they coincide with confidence intervals for $JK(\beta_0)$.

1.8. Simulation Study

In this simulation study, I examine the performance of tests based on the $JK(\beta_0)$ statistic and compare it with that of other tests that may be used in settings where the number of instruments is nonnegligible as a fraction of sample size. I consider a reduced-form data-generating process (DGP) similar to that of Matsushita and Otsu (2022). The outcome

variable, y_i, and endogenous variable, x_i, are generated according to

$$\begin{aligned} y_i &= x_i + \epsilon_i \\ x_i &= \Pi_i + v_i \end{aligned} \tag{1.8.1}$$

where $\Pi_i = \frac{1}{r_n}\sum_{k=1}^{5} \frac{3}{4}\bar{z}_{ki} + \frac{1}{4}\bar{z}_{ki}^2 + \frac{1}{4}\bar{z}_{ki}^3$ is a transformation of an initial set of instruments $\bar{z}_i \in \mathbb{R}^{10}$ generated as described below. The value of r_n varies depending on the strength of identification considered; for strong identification, $r_n = 1$, while under weak identification, $r_n = 1/\sqrt{n}$. To model heteroskedasticity, the errors (e_i, v_i) are generated $\epsilon_i = (1+\varrho_1(\bar{z}_{1i}^2+\bar{z}_{2i}^2+\bar{z}_{2i}\bar{z}_{3i}))e_{1i}$, and $v_i = \varrho_2(1+\bar{z}_{1i})\epsilon_i + (1-\varrho_2)^2 e_{2i}$ where e_{1i} and e_{2i} are generated independently of each other and other variables in the model according to a Laplace distribution with location parameter $\mu = 0$ and scale parameter $b = 1$.[1] Since the limiting χ^2 distribution of the jackknife K-statistic is exact when the errors are jointly Gaussian and $\rho(z_i)$ is known, I purposefully avoid normally distributed errors to investigate the quality of asymptotic approximations to the finite-sample behavior of the test. The parameters ϱ_1 and ϱ_2 control the degree of heteroskedasticity and endogeneity, respectively.

In addition to considering the behavior of tests under both weak and strong identification, I examine the size of the test under three different instrument regimes. In all three regimes, I begin with an initial set of instruments $\bar{z}_i = (\bar{z}_{1i}, \ldots, \bar{z}_{10i})'$ generated independently across indices according to a multivariate Gaussian distribution with Toeplitz covariance structure, $\text{Cov}(\bar{z}_{\ell i}, \bar{z}_{ki}) = 2^{-|\ell-k|}$. In the first regime, the full set instruments z_i is taken to be equal to $\bar{z}_i$ so that $d_z = 10$. In the second regime, the full set of instruments z_i additionally includes all quadratic and cubic terms, $(z_{\ell i}^2, z_{\ell i}^3)$, $\ell = 1, \ldots, 10$ so that in total $d_z = 30$. In the third regime, the full set of instrument includes the initial set of instruments, $\bar{z}_i$, and all quadratic terms (10 additional terms) and interactions of the initial set of instruments ($\binom{10}{2} = 45$

[1]The Laplace distribution is often referred to as a "double exponential" distribution. If X_1 and X_2 are independently distributed according Exponential(1), then $Y = X_1 - X_2$ has a Laplace distribution with parameters $\mu = 0$ and $b = 1$. If X has a Laplace distribution with parameters $\mu = 0$ and $b = 1$, then $|X| \sim$ Exponential(1).

DGP				Testing Procedure						
n	d_z	ϱ_1	ϱ_2	$JK(\beta_0)$	$S(\beta_0)$	$T(\beta_0;\tau_{0.3})$	$T(\beta_0;\tau_{0.75})$	A.Rbn.	JAR	JLM
200	10	0.2	0.3	0.0516	0.0352	0.0406	0.0406	0.0296	0.0766	0.0502
		0.2	0.6	0.0542	0.0306	0.0442	0.0384	0.0258	0.0748	0.0400
		0.5	0.3	0.0470	0.0338	0.0416	0.0418	0.0238	0.0784	0.0460
		0.5	0.6	0.0506	0.0350	0.0416	0.0390	0.0280	0.0676	0.0384
	30	0.2	0.3	0.0570	0.0124	0.0422	0.0200	0.0088	0.1000	0.0382
		0.2	0.6	0.0564	0.0126	0.0408	0.0208	0.0124	0.0962	0.0322
		0.5	0.3	0.0498	0.0100	0.0366	0.0190	0.0096	0.1090	0.0318
		0.5	0.6	0.0562	0.0118	0.0420	0.0216	0.0088	0.1104	0.0292
	65	0.2	0.3	0.0542	0.0316	0.0428	0.0370	0.0314	0.0764	0.0420
		0.2	0.6	0.0532	0.0366	0.0418	0.0398	0.0250	0.0780	0.0376
		0.5	0.3	0.0474	0.0308	0.0388	0.0362	0.0244	0.0748	0.0354
		0.5	0.6	0.0484	0.0324	0.0366	0.0388	0.0282	0.0708	0.0402
500	10	0.2	0.3	0.0590	0.0468	0.0478	0.0516	0.0376	0.0652	0.0452
		0.2	0.6	0.0530	0.0420	0.0460	0.0466	0.0366	0.0692	0.0434
		0.5	0.3	0.0496	0.0370	0.0408	0.0368	0.0338	0.0710	0.0464
		0.5	0.6	0.0512	0.0426	0.0456	0.0438	0.0334	0.0696	0.0404
	30	0.2	0.3	0.0522	0.0202	0.0386	0.0278	0.0238	0.0818	0.0322
		0.2	0.6	0.0558	0.0208	0.0408	0.0310	0.0266	0.0888	0.0342
		0.5	0.3	0.0554	0.0178	0.0392	0.0280	0.0174	0.0940	0.0272
		0.5	0.6	0.0570	0.0156	0.0426	0.0236	0.0206	0.0984	0.0280
	65	0.2	0.3	0.0542	0.0372	0.0434	0.0432	0.0384	0.0754	0.0464
		0.2	0.6	0.0584	0.0442	0.0482	0.0470	0.0334	0.0676	0.0438
		0.5	0.3	0.0614	0.0460	0.0504	0.0496	0.0316	0.0708	0.0434
		0.5	0.6	0.0526	0.0378	0.0434	0.0420	0.0298	0.0692	0.0358

Table 1.8.1: Simulated Size of Identification and Heteroskedasticity Robust Tests under Weak Identification. Each DGP is simulated 5000 times. Critical values of the sup-score statistic and quantiles of the conditioning statistic are calculated using 1000 multiplier bootstrap simulations.

additional terms), so that in total $d_z = 65$. Under each regime, the full set of instruments is passed to the test statistics with no indication about which instruments correspond to the initial set, and thus no indication about which instruments are relevant to the DGP.

I compare the simulated size of the jackknife K test and to the performance of the sup-score

DGP				Testing Procedure						
n	d_z	ϱ_1	ϱ_2	$JK(\beta_0)$	$S(\beta_0)$	$T(\beta_0;\tau_{0.3})$	$T(\beta_0;\tau_{0.75})$	A.Rbn.	JAR	JLM
200	10	0.2	0.3	0.0474	0.0420	0.0474	0.0468	0.0308	0.0728	0.0424
		0.2	0.6	0.0512	0.0386	0.0512	0.0506	0.0304	0.0764	0.0378
		0.5	0.3	0.0416	0.0318	0.0414	0.0414	0.0248	0.0794	0.0428
		0.5	0.6	0.0446	0.0342	0.0446	0.0442	0.0244	0.0806	0.0384
	30	0.2	0.3	0.0482	0.0122	0.0448	0.0264	0.0110	0.1048	0.0370
		0.2	0.6	0.0498	0.0120	0.0480	0.0312	0.0118	0.0980	0.0378
		0.5	0.3	0.0456	0.0126	0.0410	0.0262	0.0082	0.1146	0.0268
		0.5	0.6	0.0482	0.0110	0.0474	0.0308	0.0094	0.1090	0.0302
	65	0.2	0.3	0.0528	0.0380	0.0526	0.0510	0.0276	0.0696	0.0460
		0.2	0.6	0.0464	0.0360	0.0464	0.0468	0.0302	0.0728	0.0416
		0.5	0.3	0.0482	0.0298	0.0480	0.0466	0.0246	0.0738	0.0412
		0.5	0.6	0.0396	0.0320	0.0390	0.0386	0.0258	0.0748	0.0356
500	10	0.2	0.3	0.0524	0.0444	0.0524	0.0524	0.0394	0.0684	0.0472
		0.2	0.6	0.0476	0.0430	0.0476	0.0476	0.0400	0.0644	0.0490
		0.5	0.3	0.0434	0.0410	0.0434	0.0434	0.0340	0.0702	0.0404
		0.5	0.6	0.0448	0.0382	0.0448	0.0448	0.0350	0.0736	0.0432
	30	0.2	0.3	0.0502	0.0214	0.0502	0.0498	0.0240	0.0854	0.0368
		0.2	0.6	0.0522	0.0208	0.0522	0.0524	0.0224	0.0858	0.0392
		0.5	0.3	0.0456	0.0202	0.0456	0.0434	0.0220	0.0918	0.0264
		0.5	0.6	0.0500	0.0186	0.0500	0.0498	0.0204	0.0924	0.0268
	65	0.2	0.3	0.0490	0.0426	0.0490	0.0490	0.0350	0.0742	0.0472
		0.2	0.6	0.0522	0.0458	0.0522	0.0522	0.0436	0.0652	0.0442
		0.5	0.3	0.0542	0.0476	0.0542	0.0542	0.0294	0.0712	0.0446
		0.5	0.6	0.0438	0.0420	0.0438	0.0438	0.0306	0.0666	0.0500

Table 1.8.2: Simulated Size of Identification and Heteroskedasticity Robust Tests under Strong Identification. Each DGP is simulated 5000 times. Critical values of the sup-score statistic and quantiles of the conditioning statistic are calculated using 1000 multiplier bootstrap simulations.

test, $S(\beta_0)$, of Belloni et al. (2012a), the thresholding test introduced in Section 1.5.2, the standard Anderson-Rubin (A.Rbn.) test of Anderson and Rubin (1949) and Staiger and Stock (1997), the jackknife AR test (JAR) of Crudu et al. (2021) and Mikusheva and Sun (2021), and the jackknife LM test (JLM) of Matsushita and Otsu (2022). To estimate the parameter $\rho(z_i)$, I implement the ℓ_1-penalized procedure of (1.3.2) via the `glmnet` package in

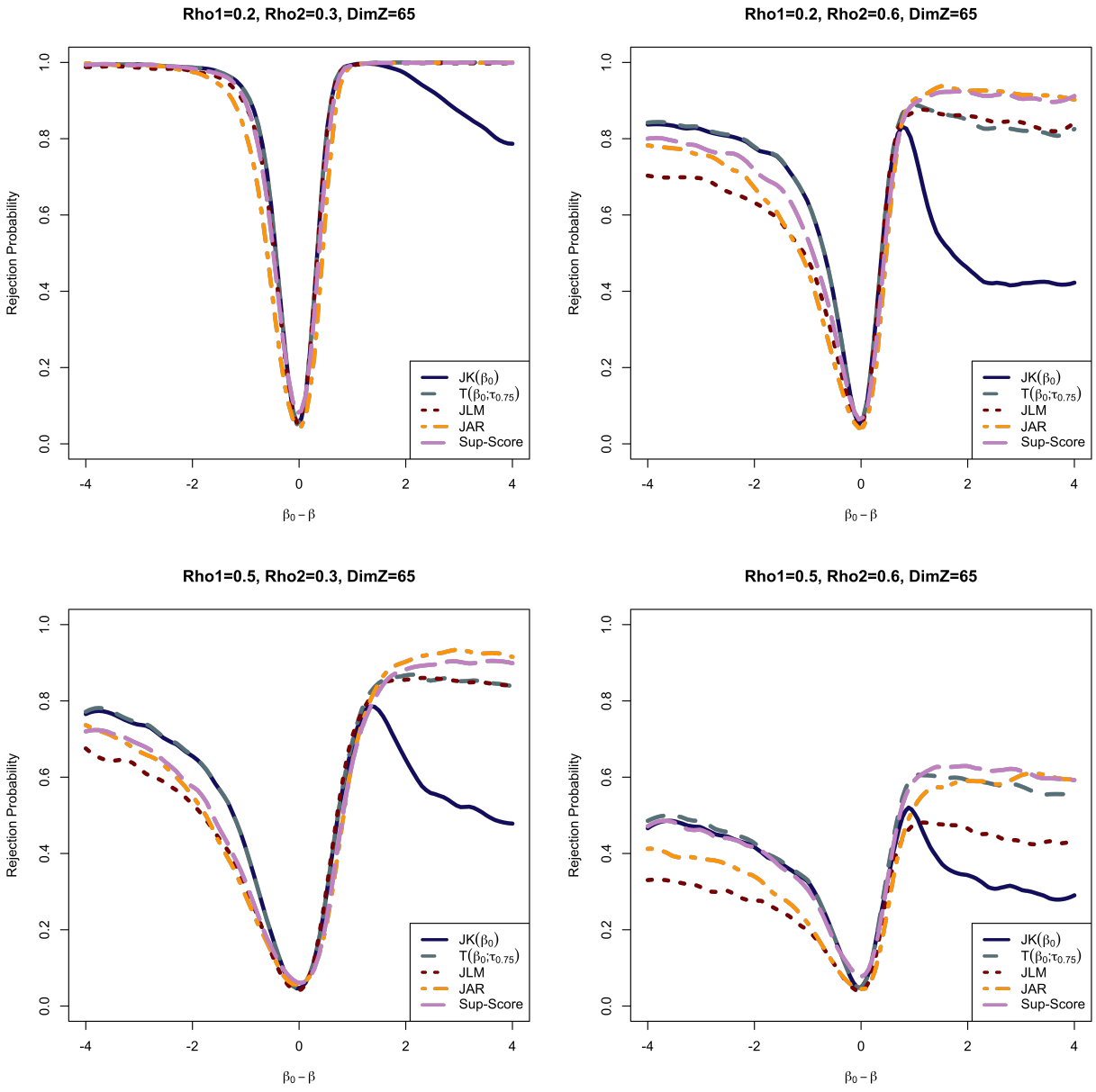

Figure 1.8.1: Calibrated Local Power Curves under Intermediate Identification Strength and 65 Instruments. Sample size is 500 and rejection probability is calculated on a grid of 100 $(\beta_0 - \beta)$ points between -4 and 4. At each point the DGP is simulated 2000 times.

`R` (Friedman et al., 2010). The penalty parameter λ is selected via tenfold cross-validation. I use the full vector of instruments as the basis to approximate $\rho(z_i)$. For the jackknife AR test I use cross-fit estimates of test statistic variances proposed and shown to improve power by Mikusheva and Sun (2021). Critical values of the sup-score and conditioning statistic are simulated with the procedures described in Section 1.5 with 1000 bootstrap replications. For

the combination test cutoff, I consider two different quantiles of the conditioning statistic under the assumption that $\mathbb{E}[\widehat{\Pi}_i^I] = 0$ for all $i \in [n]$; $\tau_{0.3}$ corresponding to the 30^{th} quantile and $\tau_{0.75}$ corresponding to the 75^{th} quantile.

Tables 1.8.1 and 1.8.2 report the simulated size for all tests under weak and strong identification, respectively. One can see that the $JK(\beta_0)$ statistic has nearly exact size in almost all the setups considered. In contrast, the jackknife AR test seems to overreject in nearly all the simulation setups considered. This is also the case in the simulation study of Matsushita and Otsu (2022) and so may be an artifact of the similarity of my simulation design to theirs.

The sup-score, jackknife AR, and jackknife LM test all seem to have particularly poor performance under both weak and strong identification when $d_z = 30$. This is the setup with the most correlation between the instruments. While tests based on the jackknife AR statistic can have a simulated size that is nearly double the nominal size in this setting, both the sup-score and and jackknife LM tests appear to be conservative. The size of the sup-score test is always under 0.025 while the size of the JLM test can be under half of the nominal size. Notably, the size properties of the sup-score test do seem to improve under both weak and strong identification when the sample size increases from $n = 200$ to $n = 500$. This is in line with theoretical results showing that the sup-score test has exact asymptotic size under standard conditions. In contrast, the size properties of the jackknife LM test do not seem to improve when the sample size increases and indeed worsen for three out of the four DGPs considered under both weak and strong identification. This suggests that the requirement of $d_z \to \infty$ may be important for the quality of finite-sample approximation by its limiting distribution.

The thresholding test seems to control size in all the setups considered. However, under weak identification the thresholding test appears to inherit the conservative nature of the sup-score test, even in the "large" sample size regime of $n = 500$. This is not the case under strong identification, suggesting that the thresholding-test is choosing to run tests

based on the $JK(\beta_0)$ with high probability in this regime. This behavior is similar to the conditional combination tests of Moreira (2003), Andrews (2016) which weigh the K-statistic more under strong identification. This behavior is optimal as the K-statistic yields effecient inference when the data is informative about the structural parameter (Andrews et al., 2004, 2006). When errors are homoskedastic and the number of instruments is fixed, the jackknife K-statistic can also be shown to yield effecient inference under strong identification.

Figure 1.8.1 plots calibrated local power curves under an intermediate-strength identification where the first stage is in a $n^{-1/3}$ neighborhood of zero, $d_z = 65$, $\varrho_1 \in \{0.2, 0.5\}$ and $\varrho_2 \in \{0.3, 0.6\}$. The critical value of each test is set to simulated 95^{th} quantile of the distribution of the corresponding test-statistic under H_0. I compare the calibrated local power curves of the $JK(\beta_0)$ test, the combination test with cutoff $\tau_{0.75}$, the jackknife AR test, the Jackknife LM test, and the sup-score test. The jackknife K-test appears to have stronger power than the jackknife AR, jackknife LM, and sup-score tests in local neighborhoods of the null as well as for negative values of $(\beta_0 - \beta)$. For values of $(\beta_0 - \beta)$ larger than 1.5, tests based on the jackknife K-statistic appear to suffer from a loss of power as described in Section 1.5. This power decline appears to be largerly ameliorated by combining the jackknife K-statistic with the sup-score statistic and the thresholding test appears to have good power properties over all alternatives considered. However, tests based on the jackknife AR or jackknife LM statistic may still provide better power than the threshholding test for very positive values of $(\beta_0 - \beta)$.

In order to consider the effect of correlated instruments on the power properties of the test in a setting with plausibly many instruments, I additionally examine local power under a fourth instrument regime. This setup adds the ten cubic terms $z_{\ell i}^3, \ell = 1, \ldots, 10$ to the interactions and quadratic terms of the third instrument regime for a total of 75 instruments, $d_z = 75$. Section 1.12 provides the simulated sizes of tests under this fourth regime. Figure 1.8.2 plots calibrated local power curves under this fourth instrument regime. While all tests have lower

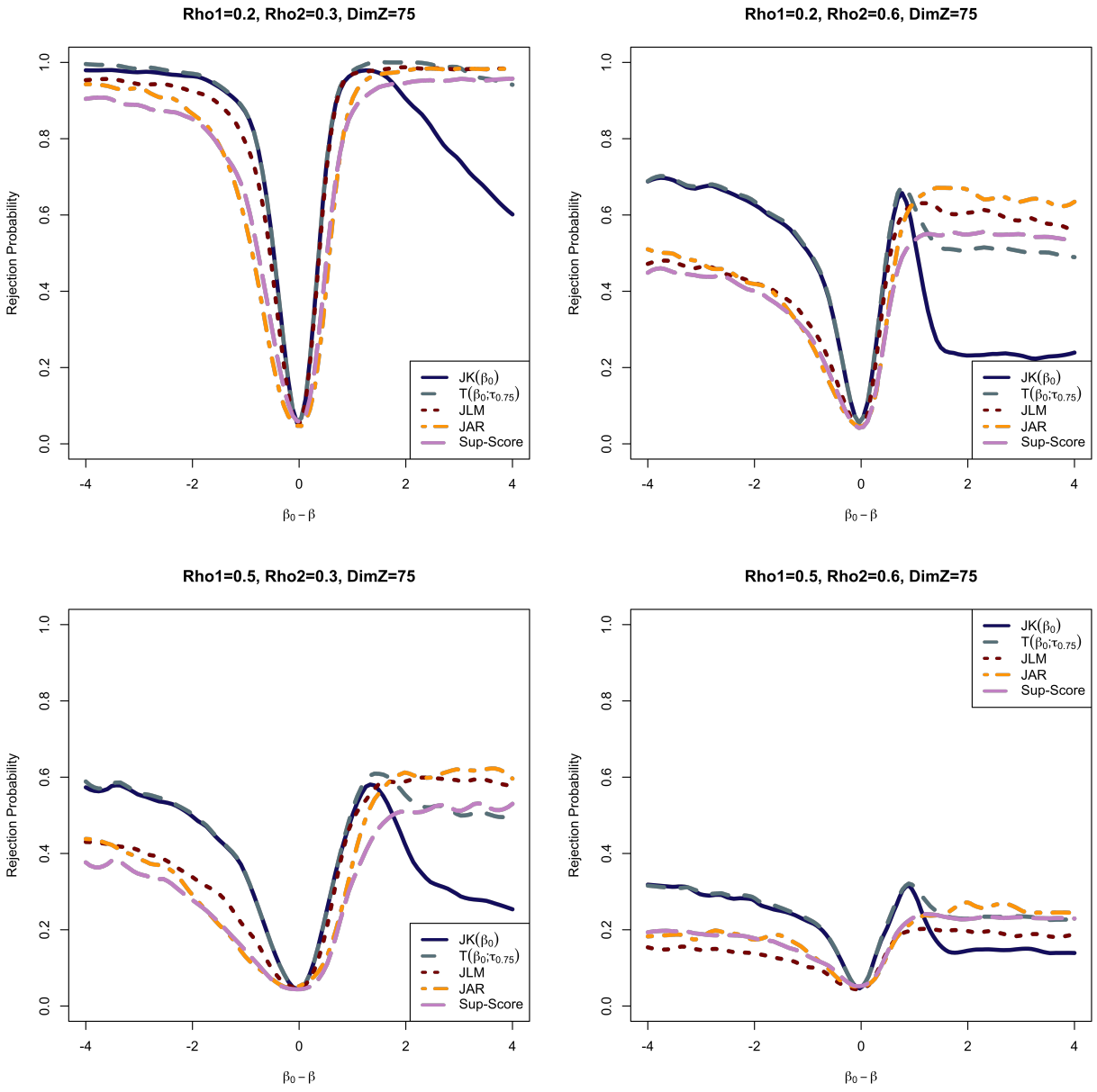

Figure 1.8.2: Calibrated Local Power Curves under Intermediate identification Strength and 75 Instruments. Sample size is 500 and rejection power is calculated on a grid of 100 $(\beta_0 - \beta)$ points between -4 and 4. At each point the DGP is simulated 2000 times.

power in this regime than in the regime considered in Figure 1.8.1, the many instrument jackknife AR and jackknife LM tests appear to face a steeper power decline than tests based on the jackknife K-statistic or the thresholding statistic.

These results should not be interpreted as critiques of the benchmark testing procedures of

Anderson and Rubin (1949), Staiger and Stock (1997), Belloni et al. (2012a), Crudu et al. (2021), Mikusheva and Sun (2021), and Matsushita and Otsu (2022), whose work I rely on and was inspired by.

1.9. Conclusion

I propose a new test for the structural parameter in a linear instrumental variables model. This test is based on a jackknife version of the K-statistic and the limiting behavior of the test is analyzed via a novel direct Gaussian approximation argument. I show that, as long as an auxiliary parameter can be consistently estimated, the test is robust to both the strength of identification and the number of instruments; the limiting distribution of the test statistic does not depend on either of these factors. Consistency of the auxiliary parameter can be achieved under approximate sparsity using simple-to-implement ℓ_1-penalized methods.

I characterize the behavior of the jackknife K-statistic in local neighborhoods of the null. To address a power deficiency that tests based on jackknife K-statistic inherit from their non-jackknife namesakes, I propose a testing procedure that decides whether the researcher should run a test via the jackknife K-statistic or one via the sup-score statistic based on the value of a conditioning statistic. While this combination may not fully address the power decline, I show that it works well in a simulation study and leave further refinements to future work.

1.10. Appendix: Proofs of Main Results

1.10.1. Proofs of Results in Section 1.4

Proof of Lemma 1.4.1

The statement $\sup_{a<0}\left|\Pr(JK_I(\beta_0)\leq a)-\Pr(JK_G(\beta_0)\leq a)\right|=0$ is immediate since both $JK_I(\beta_0)$ and $JK_G(\beta_0)$ are always weakly positive. It thus suffices to show

$$\sup_{a\geq 0}\left|\Pr(JK_I(\beta_0)\leq a)-\Pr(JK_G(\beta_0)\leq a)\right|\to 0$$

Before proceeding, we will introduce some notation. Let $\tilde{H}=s_nH$ and $\tilde{h}_{ij}=s_nh_{ij}$, where s_n is as in Assumption 1.4.2. Recall that $\tilde{h}_{ii}=0$ and define

$$N:=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(\beta_0)\sum_{j=1}^{n}\tilde{h}_{ij}r_j \qquad \tilde{N}:=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\epsilon}_i(\beta_0)\sum_{j=1}^{n}\tilde{h}_{ij}\tilde{r}_j$$
$$D:=\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(\beta_0)\Big(\sum_{j=1}^{n}\tilde{h}_{ij}r_j\Big)^2 \qquad \tilde{D}:=\frac{1}{n}\sum_{i=1}^{n}\kappa_i^2(\beta_0)\Big(\sum_{j=1}^{n}\tilde{h}_{ij}\tilde{r}_j\Big)^2$$

where $(\tilde{\epsilon}_i(\beta_0),\tilde{r}_i)$ are jointly Gaussian with the same mean and covariance matrix as $(\epsilon_i(\beta_0),r_i)$ and $\kappa_i^2(\beta_0)=\mathbb{E}[\epsilon_i^2(\beta_0)]$. Under this notation we can write $JK_I(\beta_0)=\frac{N^2}{D}\mathbf{1}_{\{D>0\}}$ and $JK_G(\beta_0)=\frac{\tilde{N}^2}{\tilde{D}}$. Dealing with these forms of the statistics is difficult for the interpolation argument, since the denominator is random. Instead, we will notice that since $D=0\implies N=0$ and $\Pr(\tilde{D}>0)=1$, for any $a\geq 0$ we can rewrite the events

$$\{JK_I(\beta_0)\leq a\}=\{N^2-aD\leq 0\}\quad\text{and}\quad\{JK_G(\beta_0)\leq a\}\overset{\text{a.s}}{=}\{\tilde{N}^2-a\tilde{D}\leq 0\}\tag{1.10.1}$$

With this in mind define

$$JK^a:=N^2-aD\quad\text{and}\quad\tilde{JK}^a:=\tilde{N}^2-a\tilde{D}$$

Showing Lemma 1.4.1 is then equivalent to showing that $\sup_a |\Pr(JK^a \leq 0) - \Pr(\tilde{JK}^a \leq 0)| \to 0$. We do so in a few lemmas, the final result being shown in Lemma 1.10.6 at the bottom of this subsection.

Lemma 1.10.1 (Lindeberg Interpolation)**.** *Suppose that Assumptions 1.4.1–1.4.3 hold. Let $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ be such that $\varphi(\cdot) \in C_b^3(\mathbb{R})$ with $L_2(\varphi) = \sup_x |\varphi''(x)|$ and $L_3(\varphi) = \sup_x |\varphi'''(x)|$. Then, there is a constant M that depends only on the constant c such that:*

$$|\mathbb{E}[\varphi(JK^a) - \varphi(\tilde{JK}^a)]| \leq \frac{M(a^3 \vee 1)}{\sqrt{n}}(L_2(\varphi) + L_3(\varphi))$$

Proof of Lemma 1.10.1. Begin by defining the leave-one-out numerator, denominator, and decomposed statistics

$$N_{-i} := \frac{1}{\sqrt{n}} \sum_{j \neq i} \dot{\epsilon}_j(\beta_0) \sum_{\ell \neq i} \tilde{h}_{j\ell} \dot{r}_\ell \qquad D_{-i} := \frac{1}{n} \sum_{j \neq i} \ddot{\epsilon}_j^2(\beta_0) \Big(\sum_{\ell \neq i} \tilde{h}_{j\ell} \dot{r}_\ell \Big)^2$$
$$JK_{-i} := N_{-i}^2 - aD_{-i}$$

where for each $\ell \in [n]$, $\dot{\epsilon}_\ell(\beta_0)$ is equal to $\epsilon_\ell(\beta_0)$ if $\ell > i$ and $\tilde{\epsilon}_\ell(\beta_0)$ if $\ell < i$, $\dot{r}_\ell$ is equal to r_ℓ if $\ell > i$ and $\tilde{r}_\ell$ if $\ell < i$, and $\ddot{\epsilon}_\ell^2(\beta_0)$ is equal to $\kappa_\ell^2(\beta_0)$ if $\ell < i$ and $\epsilon_\ell^2(\beta_0)$ if $\ell > i$. While the definitions of $\dot{\epsilon}_\ell, \dot{r}_\ell$, and $\ddot{\epsilon}_\ell$ depend on i because we will be considering only one deviation at a time, we will supress the dependence of these variables on i to simplify notation.

Next, define the one-step deviations

$$
\begin{aligned}
\Delta_{1i} &:= \epsilon_i(\beta_0)\sum_{j=1}^{n}\tilde{h}_{ij}\dot{r}_j + r_i\sum_{j=1}^{n}\tilde{h}_{ji}\dot{\epsilon}_j(\beta_0) \\
\tilde{\Delta}_{1i} &:= \tilde{\epsilon}_i(\beta_0)\sum_{j=1}^{n}\tilde{h}_{ij}\dot{r}_j + \tilde{r}_i\sum_{j=1}^{n}\tilde{h}_{ji}\dot{\epsilon}_j(\beta_0) \\
\Delta_{2i} &:= \underbrace{a\epsilon_i^2(\beta_0)(\sum_{j=1}^{n}\tilde{h}_{ij}\dot{r}_j)^2 + ar_i^2\sum_{j=1}^{n}\tilde{h}_{ji}^2\ddot{\epsilon}_j^2(\beta_0)}_{\Delta_{2i}^a} + \underbrace{2ar_i\sum_{j=1}^{n}\ddot{\epsilon}_j^2(\beta_0)\sum_{\ell\neq i}\tilde{h}_{j\ell}\tilde{h}_{ji}\dot{r}_\ell}_{\Delta_{2i}^b} \\
\tilde{\Delta}_{2i} &:= \underbrace{a\kappa_i^2(\beta_0)(\sum_{j=1}^{n}\tilde{h}_{ij}\dot{r}_j)^2 + a\tilde{r}_i^2\sum_{j=1}^{n}\tilde{h}_{ji}^2\ddot{\epsilon}_j^2(\beta_0)}_{\Delta_{2i}^a} + \underbrace{2a\tilde{r}_i\sum_{j=1}^{n}\ddot{\epsilon}_j^2(\beta_0)\sum_{\ell\neq i}\tilde{h}_{j\ell}\tilde{h}_{ji}\dot{r}_\ell}_{\tilde{\Delta}_{2i}^b}
\end{aligned}
\tag{1.10.2}
$$

These one-step deviations contain all the terms associated with observation i in the expression of the numerator and denominator of the test statistics. To demonstrate, note that these one-step deviations satisfy $N_{-1} + n^{-1/2}\Delta_{11} = N$ and $aD_{-1} + n^{-1}\Delta_{21} = aD$ as

$$
\begin{aligned}
N &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(\beta_0)\sum_{j=1}^{n}\tilde{h}_{ij}r_j \\
&= \frac{1}{\sqrt{n}}\sum_{j>1}\epsilon_j(\beta_0)\sum_{\ell=1}^{n}\tilde{h}_{j\ell}r_j + \epsilon_1(\beta_0)\frac{1}{\sqrt{n}}\sum_{j>1}\tilde{h}_{1j}r_j \\
&= \frac{1}{\sqrt{n}}\sum_{j>1}\epsilon_j(\beta_0)\Big\{\tilde{h}_{j1}r_1 + \sum_{\ell>1}h_{j\ell}r_\ell\Big\} + \epsilon_1(\beta_0)\frac{1}{\sqrt{n}}\sum_{j>1}\tilde{h}_{1j}r_j \\
&= \underbrace{\frac{1}{\sqrt{n}}\sum_{j>1}\epsilon_j(\beta_0)\sum_{\ell>1}h_{j\ell}r_\ell}_{N_{-1}} + \underbrace{\epsilon_1(\beta_0)\frac{1}{\sqrt{n}}\sum_{j>1}\tilde{h}_{1j}r_j + r_1\frac{1}{\sqrt{n}}\sum_{j>1}\tilde{h}_{j1}\epsilon_j(\beta_0)}_{n^{-1/2}\Delta_{11}}
\end{aligned}
$$

and

$$
\begin{aligned}
D &= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(\beta_0)\big(\sum_{j=1}^{n}\tilde{h}_{ij}r_j\big)^2 \\
&= \frac{1}{n}\sum_{j>1}\epsilon_j^2(\beta_0)\big(\sum_{\ell=1}^{n}\tilde{h}_{j\ell}r_\ell\big)^2 + \epsilon_1^2(\beta_0)\frac{1}{n}\big(\sum_{j>1}\tilde{h}_{1j}r_j\big)^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{j>1}\epsilon_j^2(\beta_0)\big(\tilde{h}_{j1}r_1 + \sum_{\ell\neq 1}\tilde{h}_{\ell j}r_\ell\big)^2 + \epsilon_1^2(\beta_0)\frac{1}{n}\big(\sum_{j>1}\tilde{h}_{1j}r_j\big)^2 \\
&= \underbrace{\frac{1}{n}\sum_{j>1}\epsilon_j^2(\beta_0)\big(\sum_{\ell>1}\tilde{h}_{\ell,j}r_\ell\big)^2}_{D_{-1}} \\
&\quad + \underbrace{\epsilon_1^2(\beta_0)\frac{1}{n}\big(\sum_{j>1}\tilde{h}_{1j}r_j\big)^2 + r_1^2\frac{1}{n}\sum_{j>1}\tilde{h}_{j1}^2\epsilon_j^2(\beta_0) + 2r_1\frac{1}{n}\sum_{j>1}\epsilon_j^2(\beta_0)\sum_{\ell>1}\tilde{h}_{\ell j}r_\ell}_{(an)^{-1}\Delta_{21}}
\end{aligned}
$$

Using the one-step deviations, write the difference $\mathbb{E}[\varphi(K^a) - \varphi(\tilde{K}^a)]$ as a telescoping sum, one by one replacing $(\Delta_{1i}, \Delta_{2i})$ with $(\tilde{\Delta}_{1i}, \tilde{\Delta}_{2i})$ in the expressions of $JK^a = N^2 - aD$ until we arrive at $\tilde{JK}^a = \tilde{N}^2 - a\tilde{D}$.

$$
\begin{aligned}
\mathbb{E}[\varphi(JK^a) - \varphi(\tilde{JK}^a)] = \sum_{i=1}^{n} &\mathbb{E}[\varphi(JK_{-i} + n^{-1/2}N_{-i}\Delta_{1i} + n^{-1}\Delta_{1i}^2 - n^{-1}\Delta_{2i})] \\
&- \mathbb{E}[\varphi(JK_{-i} + n^{-1/2}N_{-i}\tilde{\Delta}_{1i} + n^{-1}\tilde{\Delta}_{1i}^2 - n^{-1}\tilde{\Delta}_{2i})]
\end{aligned} \tag{1.10.3}
$$

Via a second-order Taylor expansion, we can write each term inside the summand

$$
\begin{aligned}
\mathbb{E}[\text{Term}_i] &= \mathbb{E}[\varphi'(JK_{-i})\{2n^{-1/2}N_{-i}(\Delta_{1i} - \tilde{\Delta}_{1i}) + n^{-1}(\Delta_{1i}^2 - \Delta_{1i}^2) - n^{-1}(\Delta_{2i} - \tilde{\Delta}_{2i})\}] \\
&\quad + \mathbb{E}[\varphi''(JK_{-i})\{4n^{-1}N_{-i}^2(\Delta_{1i}^2 - \tilde{\Delta}_{1i}^2) + n^{-2}(\Delta_{1i}^4 - \tilde{\Delta}_{1i}^4) - n^{-2}(\Delta_{2i}^2 - \Delta_{2i}^2)\}] \\
&\quad + \mathbb{E}[\varphi''(JK_{-i})\{4n^{-3/2}N_{-i}(\Delta_{1i}^3 - \tilde{\Delta}_{1i}^3) + 4n^{-3/2}N_{-i}(\Delta_{1i}\Delta_{2i} - \tilde{\Delta}_{1i}\tilde{\Delta}_{2i})\}] \\
&\quad + \mathbb{E}[\varphi''(JK_{-i})\{2n^{-2}(\Delta_{1i}^2\Delta_{2i} - \tilde{\Delta}_{1i}^2\tilde{\Delta}_{2i})\}] + R_i + \tilde{R}_i
\end{aligned}
$$

where R_i and $\tilde{R}_i$ are remainder terms to be examined later. Let $\mathcal{F}_{-i}$ denote the sigma algebra generated by all random variables whose index is not equal to i. Since (a) for each $i \in [n]$ the mean and covariance matrix of $(\epsilon_i(\beta_0), r_i)$ is the same as the mean and covariance matrix of $(\tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$, (b) $\mathbb{E}[\epsilon_i^2(\beta_0)] = \kappa_i^2(\beta_0)$, and (c) random variables are independent across indices,

we have that

$$\begin{aligned}\mathbb{E}[\Delta_{1i} - \tilde{\Delta}_{1i}|\mathcal{F}_{-i}] &= \mathbb{E}[\Delta_{1i}^2 - \tilde{\Delta}_{1i}^2|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{2i} - \tilde{\Delta}_{2i}|\mathcal{F}_{-i}] \\ &= \mathbb{E}[\Delta_{2i}^b - \tilde{\Delta}_{2i}^b|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{1i}\Delta_{2i}^b - \tilde{\Delta}_{1i}\tilde{\Delta}_{2i}^b|\mathcal{F}_{-i}] = 0\end{aligned} \tag{1.10.4}$$

Using this we can simplify the prior display

$$\begin{aligned}\mathbb{E}[\text{Term}_i] = &\underbrace{n^{-2}\mathbb{E}[\varphi''(JK_{-i})(\Delta_{1i}^4 - \Delta_{1i}^4)]}_{\mathbf{A}_i} - \underbrace{n^{-2}\mathbb{E}[\varphi''(JK_{-i})((\Delta_{2i}^a)^2 - (\tilde{\Delta}_{2i}^a)^2)]}_{\mathbf{B}_i} \\ &- \underbrace{2n^{-2}\mathbb{E}[\varphi''(JK_{-i})(\Delta_{2i}^a\Delta_{2i}^b - \tilde{\Delta}_{2i}^a\tilde{\Delta}_{2i}^b)]}_{\mathbf{C}_i} + \underbrace{4n^{-3/2}\mathbb{E}[\varphi''(JK_{-i})N_{-i}(\Delta_{1i}^3 - \tilde{\Delta}_{1i}^3)]}_{\mathbf{D}_i} \\ &+ \underbrace{4n^{-3/2}\mathbb{E}[\varphi''(JK_{-i})N_{-i}(\Delta_{1i}\Delta_{2i}^a - \tilde{\Delta}_{1i}\tilde{\Delta}_{2i}^a)]}_{\mathbf{E}_i} + \underbrace{2n^{-2}\mathbb{E}[\varphi''(JK_{-i})(\Delta_{1i}^2\Delta_{2i} - \tilde{\Delta}_{1i}^2\tilde{\Delta}_{2i})}_{\mathbf{F}_i} \\ &+ R_i + \tilde{R}_i\end{aligned}$$

where for some $\bar{JK}_{1i}$ and $\bar{JK}_{2i}$ we can write

$$\begin{aligned}R_i &= \mathbb{E}[\varphi'''(\bar{JK}_{1i})\{n^{-1/2}N_{-i}\Delta_{1i} + n^{-1}\Delta_{1i}^2 + n^{-1}\Delta_{2i}\}^3] \\ \tilde{R}_i &= \mathbb{E}[\varphi'''(\bar{JK}_{2i})\{n^{-1/2}N_{-i}\tilde{\Delta}_{1i} + n^{-1}\tilde{\Delta}_{1i}^2 + n^{-1}\tilde{\Delta}_{2i}\}^3]\end{aligned}$$

Applications of Lemmas 1.10.18 and 1.10.19, Cauchy-Schwarz, and the generalized Hölder inequality,[1] will allow us to bound for a fixed constant M that depends only on c,

$$\begin{aligned}|\mathbf{A}_i| &\le \frac{M}{n^2}L_2(\varphi) & |\mathbf{B}_i| &\le \frac{Ma^2}{n^2}L_2(\varphi) & |\mathbf{C}_i| &\le \frac{Ma^2}{n^{3/2}}L_2(\varphi) \\ |\mathbf{D}_i| &\le \frac{M}{n^{3/2}}L_2(\varphi) & |\mathbf{E}_i| &\le \frac{M(a \vee 1)}{n^{3/2}}L_2(\varphi) & |\mathbf{F}_i| &\le \frac{Ma^3}{n^{3/2}}L_2(\varphi)\end{aligned}$$

and

$$|R_i| + |\tilde{R}_i| \le \frac{M}{n^{3/2}}L_3(\varphi) + \frac{Ma^3}{n^3}L_3(\varphi)$$

[1] $\mathbb{E}[|fgk|]^3 \le \mathbb{E}[|f|^3]\mathbb{E}[|g|^3]\mathbb{E}[|k|^3]$

Combining these bounds and summing over n gives the result. □

Lemma 1.10.2 (Gaussian Denominator Anti-Concentration)**.** *Suppose that Assumptions 1.4.1 and 1.4.2 hold. Then for any sequence $\delta_n \searrow 0$,*

$$\Pr(\tilde{D} \leq \delta_n) \to 0$$

Proof of Lemma 1.10.2. By Assumption 1.4.1, we know that $\kappa_i^2(\beta_0) \in [c^{-1}, c]$ for all $i = 1, \ldots, n$ so that $\tilde{D} \geq \frac{c^{-1}}{n} \sum_{i=1}^n (\sum_{j=1}^n \tilde{h}_{ij} r_j)^2$. Then

$$\begin{aligned}
\Pr(\tilde{D} \leq \delta_n) &\leq \Pr\Big(\frac{1}{cn} \sum_{i=1}^n \big(\sum_{j=1}^n \tilde{h}_{ij} \tilde{r}_j\big)^2 \leq \tilde{\delta}_n\Big) \\
&= \Pr\big(\|\tilde{r}' \bar{H}^{1/2}\|^2 \leq \delta_n\big) \qquad (1.10.5)
\end{aligned}$$

where $\tilde{r} := (\tilde{r}_1, \ldots, \tilde{r}_n)' \in \mathbb{R}^n$ and $\bar{H} := \frac{1}{cn} \tilde{H} \tilde{H}' \in \mathbb{R}^{n \times n}$. $\bar{H}$ is symmetric and positive semidefinite so we can take $\bar{H}^{1/2}$ to be its symmetric square root, which will also be symmetric and positive semidefinite (and thus not necessarily equal to $\sqrt{\frac{c}{n}} \tilde{H}$). I provide two bounds on (1.10.5), the first of which corresponds to the strong identification setting while the second corresponds to weak identification.

First Bound. Since $\delta_n \searrow 0$ we will eventually have that $\delta_n < c^{-1}/2$. When this happens we can bound using Chebyshev's inequality and $c^{-1} < \mathbb{E}[r' \bar{H} r] < c$:

$$\begin{aligned}
\Pr(\tilde{r}' \bar{H} \tilde{r} \leq \delta_n) &= \Pr(\tilde{r}' \bar{H} \tilde{r} - \mathbb{E}[\tilde{r}' \bar{H} \tilde{r}] \leq \delta_n - \mathbb{E}[\tilde{r}' \bar{H} \tilde{r}]) \\
&\leq \Pr(\tilde{r}' \bar{H} \tilde{r} - \mathbb{E}[r' \bar{H} r] \geq \mathbb{E}[\tilde{r}' \bar{H} \tilde{r}] - \delta_n) \\
&\leq \Pr(|\tilde{r}' \bar{H} \tilde{r} - \mathbb{E}[r' \bar{H} r]| \geq \frac{1}{2c}) \\
&\leq 2c \operatorname{Var}(r' \bar{H} r) \qquad (1.10.6)
\end{aligned}$$

Under strong identification we will expect $\operatorname{Var}(r' \bar{H} r) \to 0$.

Second Bound. For the second bound, we will directly use bounds on the density of Gaussian quadratic forms from Götze et al. (2019). The vector $r'\bar{H}^{1/2}$ is Gaussian with covariance matrix $\Sigma_r = \bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2}$ where $\boldsymbol{R} = \text{diag}(\text{Var}(r_1), \ldots, \text{Var}(r_n))$. Let $\Lambda_1 = \sum_{k=1}^n \lambda_k^2(\Sigma_r)$ and $\Lambda_2 = \sum_{k=2}^n \lambda_k^2(\Sigma_r)$. By Assumption 1.4.2 and Lemma 1.10.37, Λ_2/Λ_1 is bounded away from zero. Using Theorem 1.10.4 we can then bound for some constant $C > 0$

$$\Pr(\|r'H\|^{1/2} \leq \delta_n) \leq C\delta_n\Lambda_1^{-1} \tag{1.10.7}$$

Combining Bounds. To combine the bounds in (1.10.6) and (1.10.7), first write

$$\text{Var}(\tilde{r}'\bar{H}\tilde{r}) = 2\text{trace}(\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H}) + 4\mu_r\bar{H}\boldsymbol{R}\bar{H}\mu_r$$

for $\mu_r = \mathbb{E}[r]$. Using the fact that $\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2}$ is symmetric positive definite we can bound:

$$\begin{aligned}
\mu_r'\bar{H}\boldsymbol{R}\bar{H}\mu_r &= (\mu_r'\bar{H}^{1/2})'(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})(\bar{H}^{1/2}\mu_r) \\
&\leq \lambda_1(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})\|\mu_r'\bar{H}^{1/2}\|^2 \\
&= \sqrt{\lambda_1^2(\bar{H}^{1/2}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2 \\
&= \sqrt{\lambda_1(\bar{H}^{1/2}\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2 \\
&\leq \sqrt{\text{trace}(\bar{H}^{1/2}\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H}^{1/2})}\|\mu_r'\bar{H}^{1/2}\|^2 \\
&= \sqrt{\text{trace}(\boldsymbol{R}\bar{H}\boldsymbol{R}\bar{H})}\|\mu_r'\bar{H}\|^2 \leq c^2\Lambda_1^{1/2}
\end{aligned} \tag{1.10.8}$$

where the first equality uses the symmetric square root of $\bar{H}$, the first inequality comes from Courant-Fischer minmax principle and the third equality uses the fact that the eigenvalues of A^2 are the squares of the eigenvalues of A, for any generic symmetric matrix A. The second inequality comes from the fact that a matrix times its transpose is always positive semidefinite and that for M psd, $\lambda_1(M) \leq \sqrt{\text{trace}(M^2)}$ since the trace is the sum of the (weakly positive) eigenvalues. The final inequality uses $\mu_r'\bar{H}\mu_r = \frac{c}{n}\sum_{i=1}^n(\mathbb{E}[\tilde{\Pi}_i])^2 \leq \frac{c}{n}\sum_{i=1}^n \mathbb{E}[(\tilde{\Pi}_i)^2] \leq c^2$.

Combining (1.10.6), (1.10.7), and (1.10.8) gives us

$$\Pr(\tilde{D} \leq \delta_n) \leq C \min\left\{\Lambda_1 + \Lambda_1^{1/2}, \delta_n \Lambda_1^{-1}\right\} \tag{1.10.9}$$

Regardless of the behavior of Λ_1, this tends to zero as $\delta_n \to 0$. □

Remark 1.10.1 (Final Anticoncentration Bound)**.** To give an explicit bound on (1.10.9) in terms of δ_n we note that, if $x^\star$ solves

$$x^\star + \sqrt{x^\star} = \frac{c}{x^\star}$$

then for any $x \geq 0$, $\min\{x + \sqrt{x}, c/x\} \leq x^\star + \sqrt{x^\star}$. Using this, notice that $(x^\star)^2 + (x^\star)^{3/2} = c$ so that $x^\star \leq \sqrt{c}$. This allows us to bound (1.10.9)

$$\Pr(\tilde{D} \leq \delta_n) \leq C \min\{\Lambda_1 + \Lambda_1^{1/2}, \delta_n \Lambda_1^{-1}\} \leq C(\delta_n^{1/2} + \delta_n^{1/4})$$

Lemma 1.10.3. *Let X_n and Y_n be two sequences of random variables and let $W_n = X_n/Y_n$. Then for any $c \in \mathbb{R}$ and any $\delta > 0$:*

$$\Pr(0 \leq X_n - cY_n \leq \delta) \leq \Pr(c \leq W_n \leq \delta^{1/2} + c) + \Pr(Y_n \leq \delta^{1/2})$$

and

$$\Pr(-\delta \leq X_n - cY_n \leq 0) \leq \Pr(c - \delta^{1/2} \leq W_n \leq c) + \Pr(Y_n \leq \delta^{1/2})$$

Proof. Define the event $\Omega = \{Y_n \geq \delta^{1/2}\}$. We can bound

$$\begin{aligned}\Pr(0 \leq X_n - cY_n \leq \delta) &= \Pr(cY_n \leq X_n \leq \delta + cY_n) \\ &\leq \Pr(\{cY_n \leq X_n \leq \delta + cY_n\} \cap \Omega) + \Pr(\Omega^c)\end{aligned}$$

$$= \Pr(\{c \leq W_n \leq \delta/Y_n + c\} \cap \Omega) + \Pr(\Omega^c)$$
$$\leq \Pr(c \leq W_n \leq \delta^{1/2} + c) + \Pr(\Omega^c)$$

The second statement of the lemma follows symmetrically. □

Lemma 1.10.4. *Suppose that X_n and Y_n are sequences of (real-valued) random variables such that $Y_n = O_p(1)$ and for any $x \in \mathbb{R}$*

$$|\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \to 0$$

Then $X_n = O_p(1)$.

Proof. Pick any $\epsilon > 0$, and let $M_{\epsilon/2}$ be such that $\Pr(Y_n > M_{\epsilon/2}) \leq \epsilon/2$ for all $n \geq N_\epsilon$. In addition, let $\tilde{N}_\epsilon$ be such that $|\Pr(X_n \leq M_{\epsilon/2}) - \Pr(Y_n \leq M_{\epsilon/2})| \leq \epsilon/2$ for all $n \geq \tilde{N}_\epsilon$. Then for all $n \geq N_\epsilon \vee \tilde{N}_{\epsilon/2}$,

$$\begin{aligned}\Pr(X_n > M_{\epsilon/2}) &\leq \Pr(Y_n > M_{\epsilon/2}) + |\Pr(X_n > M_{\epsilon/2}) - \Pr(Y_n > M_{\epsilon/2})| \\ &\leq \epsilon/2 + |\Pr(Y_n \leq M_{\epsilon/2}) - \Pr(X_n \leq M_{\epsilon/2})| \\ &\leq \epsilon/2 + \epsilon/2 = \epsilon\end{aligned}$$

□

Lemma 1.10.5. *Suppose that X_n and Y_n are sequences of (real-valued) random variables such that $Y_n = O_p(1)$ and for any $\Delta \in \mathbb{R}$*

$$\sup_{x \leq \Delta} |\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \to 0$$

Then $\sup_{x \in \mathbb{R}} |\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \to 0$.

Proof. Pick an $\epsilon > 0$. By Lemma 1.10.4, $X_n = O_p(1)$. Pick a constant $M_{\epsilon/3}$ such that

$\Pr(X_n > M_{\epsilon/3}) \leq \epsilon/3$ and $\Pr(Y_n > M_{\epsilon/3}) \leq \epsilon/3$. Then for any $x \in \mathbb{R}$ we can bound $|\Pr(X_n \leq x) - \Pr(Y_n \leq x)|$ by considering two cases:

Case 1. If $x \leq M_{\epsilon/3}$, then,

$$|\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \leq \sup_{x \leq M_{\epsilon/3}} |\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \tag{1.10.10}$$

by hypothesis, there is an N_ϵ such that for $n \geq N_\epsilon$ the RHS of (1.10.10) is less than ϵ.

Case 2. If $x > M_{\epsilon/3}$ we can bound

$$\begin{aligned} |\Pr(X_n \leq x) - \Pr(Y_n \leq x)| &\leq |\Pr(X_n \leq M_{\epsilon/3}) - \Pr(Y_n \leq M_{\epsilon/3})| \\ &\quad + |\Pr(M_{\epsilon/3} < X_n \leq x) - \Pr(M_{\epsilon/3} < Y_n \leq x)| \\ &\leq |\Pr(X_n \leq M_{\epsilon/3}) - \Pr(Y_n \leq M_{\epsilon/3})| + \epsilon/3 + \epsilon/3 \end{aligned} \tag{1.10.11}$$

By hypothesis, there is an $N_{\epsilon/3}$ such that $|\Pr(X_n \leq M_{\epsilon/3}) - \Pr(Y_n \leq N_{\epsilon/3})| \leq \epsilon/3$.

WLOG $N_{\epsilon/3} \geq N_\epsilon$. Combining the bounds in (1.10.10) and (1.10.11), for any $n \geq N_{\epsilon/3}$ and any $x \in \mathbb{R}$,

$$|\Pr(X_n \leq x) - \Pr(Y_n \leq x)| \leq \epsilon$$

Since this holds for all x, this gives the result. □

Lemma 1.10.6 (Approximate Distribution)**.** *Under Assumptions 1.4.1–1.4.3*

$$\sup_{a \in \mathbb{R}} |\Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a)| \to 0$$

Proof of Lemma 1.10.6. First, fix a $\Delta \geq 0$ and consider any $a \leq \Delta$. As in Lemma 1.10.2, let $\tilde{\varphi}(\cdot) : \mathbb{R} \to \mathbb{R}$ be three times continuously differentiable with bounded derivatives up to the third order such that $\tilde{\varphi}(x)$ is 1 if $x \leq 0$, $\tilde{\varphi}(x)$ is decreasing if $x \in (0, 1)$, and $\tilde{\varphi}(x)$ is zero if $x \geq 1$. Consider a sequence $\gamma_n \searrow 0$ slowly enough such that $(\gamma_n^{-2} + \gamma_n^{-3})/\sqrt{n} \to 0$ and define

$\varphi_n(x) = \tilde{\varphi}(\frac{x}{\gamma_n})$.

By Lemma 1.10.1 we can write for some constant M that depends only on Δ:

$$\begin{aligned}\Pr(JK_I(\beta_0) \le a) = \Pr(JK^a \le 0) &\le \mathbb{E}[\varphi_n(JK^a)] \\ &\le \mathbb{E}[\varphi_n(\tilde{JK}^a)] + \frac{M}{\sqrt{n}}(\gamma_n^2 + \gamma_n^{-3}) \\ &\le \Pr(\tilde{JK}^a \le 0) + \Pr(0 \le \tilde{N}^2 - a\tilde{D} \le \gamma_n) + \frac{M}{\sqrt{n}}(\gamma_n^2 + \gamma_n^{-3})\end{aligned}$$

Applying Lemma 1.10.3 and $\{\tilde{JK}^a \le 0\} = \{JK_G(\beta_0) \le a\}$ gives:

$$\begin{aligned}&\le \Pr(JK_G(\beta_0) \le a) + \underbrace{\Pr(a \le \tilde{N}^2/\tilde{D} \le a + \gamma_n^{1/2})}_{\mathbf{A}} \\ &+ \underbrace{\Pr(\tilde{D} \le \gamma_n^{1/2})}_{\mathbf{B}} + \frac{M}{\sqrt{n}}(\gamma_n^{-2} + \gamma_n^{-3})\end{aligned}$$

By Lemma 1.10.20, we can bound $\mathbf{A} \le M\gamma_n^{1/2}$ while by Lemma 1.10.2 and Remark 1.10.1, $\mathbf{B} \le M\gamma_n^{1/4}$. Since γ_n is chosen such that $\frac{M}{\sqrt{n}}(\gamma_n^{-2} + \gamma_n^{-3}) \to 0$ we can conclude that $\Pr(JK_I(\beta_0) \le a) \le \Pr(JK_G(\beta_0) \le a) + o(1)$. A symmetric argument with $\varphi_n(x) = \tilde{\varphi}(1 - \frac{x}{\gamma_n})$ gives a lower bound so that, in total

$$\Pr(JK_G(\beta_0) \le a) - \boldsymbol{e} \le \Pr(JK_I(\beta_0) \le a) \le \Pr(JK_G(\beta_0) \le a) + \boldsymbol{e}$$

where

$$\boldsymbol{e} = M\big(\frac{\gamma_n^{-2} + \gamma_n^{-3}}{\sqrt{n}} + \gamma_n^{1/2} + \gamma_n^{1/4}\big) = o(1)$$

Since the constant M depends only on Δ, this gives us that for any fixed $\Delta > 0$

$$\sup_{a \le \Delta} \big|\Pr(JK_I(\beta_0) \le a) - \Pr(JK_G(\beta_0) \le a)\big| \le C\big(\frac{\gamma_n^{-2} + \gamma_n^{-3}}{\sqrt{n}} + \gamma_n^{1/2} + \gamma_n^{1/4}\big) = o(1) \quad (1.10.12)$$

where C is a constant that depends only on Δ. Noting that the numerator $JK_G(\beta_0)$ is

$O_p(1)$ under Assumption 1.4.3 while the inverse of the denominator of $JK_G(\beta_0)$ is $O_p(1)$ by Lemma 1.10.2, we can apply Lemma 1.10.5. This step shows that the result in (1.10.12) implies that the approximation error tends to zero uniformly over the real line, which is the desired result. Optimizing over γ_n in the expression of (1.10.12) yields the rate of decay in Remark 1.4.3. □

Proof of Proposition 1.4.1

Proof of Proposition 1.4.1. As at the top of Section 1.10.1, recall that $\tilde{h}_{ii} = 0$, and define

$$N = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} r_j \qquad\qquad D = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0) (\sum_{j=1}^{n} \tilde{h}_{ij} r_j)^2$$

where $\tilde{h}_{ij} = s_n h_{ij}$. The goal is to show that $\Pr(JK_I(\beta_0) \leq a) \to 0$ for any fixed $a \in \mathbb{R}_+$. The event $\{JK_I(\beta_0) \leq a\}$ is equivalently expressed $\{N^2 - aD \leq 0\}$ so that $\Pr(JK(\beta_0) \leq a) = \Pr(N^2 - aD \leq 0)$. Under Assumptions 1.4.1 and 1.4.2, $aD = O_p(1)$ so by Lemma 1.10.8 it suffices to show that $\Pr(|N| \leq M) \to 0$ for any fixed $M \geq 0$. By assumption $P = \mathbb{E}[N^2] \to \infty$ so we move to show that $\text{Var}(N) = O(1)$ and then apply Lemma 1.10.7 to conclude. To this end, recall the definition of $\eta_i = \epsilon_i(\beta_0) - \mathbb{E}[\epsilon_i(\beta_0)]$, define $\mu_i = \mathbb{E}[\epsilon_i(\beta_0)] = \Pi_i(\beta - \beta_0)$, and let

$$N_1 \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \sum_{j=1}^{n} \tilde{h}_{ij} r_j \qquad\qquad N_2 \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mu_i \sum_{j=1}^{n} \tilde{h}_{ij} r_j$$

Notice that $N = N_1 + N_2$. To show that $\text{Var}(N_1) = O(1)$, define $\boldsymbol{a}_i = \eta_i \sum_{j=1}^{n} \tilde{h}_{ij} r_j$. Since $\mathbb{E}[\eta_i r_i] = 0$, we have that $\text{Cov}(\boldsymbol{a}_i, \boldsymbol{a}_j) = 0$ for $i \neq j$. Thus,

$$\text{Var}(N_1) = \text{Var}(\sum_{i=1}^{n} \boldsymbol{a}_i / \sqrt{n}) = n^{-1} \sum_{i=1}^{n} \text{Var}(\boldsymbol{a}_i) = n^{-1} \sum_{i=1}^{n} \text{Var}(\eta_i) \mathbb{E}[(\sum_{j=1}^{n} \tilde{h}_{ij} r_j)^2] \leq c^2$$

where the final inequality follows from an upper bound on $\text{Var}(\eta_i)$ from Assumption 1.4.1 and by definition of $\tilde{h}_{ij} = s_n h_{ij}$ from Assumption 1.4.2.

To show that $\text{Var}(N_2) = O(1)$ let $\boldsymbol{b}_i = \sum_{j=1}^n \tilde{h}_{ji}\tilde{\Pi}_j(\beta - \beta_0)$ and rewrite $N_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^n r_i \boldsymbol{b}_i$. Under Assumption 1.4.3(ii), $|\boldsymbol{b}_i| = |\mathbb{E}[\sum_{j=1}^n \tilde{h}_{ji}\epsilon_j(\beta_0)]| \leq c^{1/2}$, so we can bound

$$\text{Var}(N_2) = \text{Var}(\sum_{i=1}^n r_i \boldsymbol{b}_i/\sqrt{n}) = n^{-1}\sum_{i=1}^n \boldsymbol{b}_i^2 \,\text{Var}(r_i) \leq c^2$$

Since $\text{Var}(N) \leq 2\,\text{Var}(N_1) + 2\,\text{Var}(N_2)$, we can conclude. □

Lemma 1.10.7. *Suppose that X_n is a sequence of random variables such that $\mathbb{E}[X_n^2] \to \infty$ while $\text{Var}(X_n) = O(1)$. Then, for any $M \geq 0$, $\Pr(|X_n| \leq M) \to 0$.*

Proof. First, note that $\text{Var}(|X_n|) \leq \text{Var}(X_n)$ so $\text{Var}(|X_n|) = O(1)$. Moreover $\text{Var}(|X_n|) = \mathbb{E}[X_n^2] - (\mathbb{E}[|X_n|])^2$, so $\mathbb{E}[X_n^2] \to \infty$ and $\text{Var}(|X_n|) = O(1)$ implies that $\mathbb{E}[|X_n|] \to \infty$. Then,

$$\begin{aligned}
\Pr(|X_n| \leq M) &= \Pr(|X_n| - \mathbb{E}[|X_n|] \leq M - \mathbb{E}[|X_n|]) \\
&= \Pr(\mathbb{E}[|X_n|] - |X_n| \geq \mathbb{E}[|X_n| - M) \\
&\leq \Pr(|\mathbb{E}[|X_n|] - |X_n|| \geq \mathbb{E}[|X_n|] - M) \\
&\leq \frac{\text{Var}(|X_n|)}{\mathbb{E}[|X_n|] - M}
\end{aligned}$$

Since $\text{Var}(|X_n|) = O(1)$ but $\mathbb{E}[|X_n|] \to \infty$, this tends to zero. □

Lemma 1.10.8. *Suppose that X_n and Y_n are random variables such that $Y_n = O_p(1)$ and, for any $M \geq 0$, $\Pr(|X_n| \leq M) \to 0$. Then, for any $M_1 \geq 0$, $\Pr(X_n^2 - Y_n \leq M_1) \to 0$.*

Proof. Pick any $\epsilon > 0$. We want to show that, eventually, $\Pr(X_n^2 - Y_n > M_1) \geq 1 - \epsilon$. Since $Y_n = O_p(1)$, there is a fixed constant M_Y such that $\Pr(|Y_n| \leq M_Y) \geq 1 - \epsilon/2$. Since $\Pr(|X_n| \leq M) \to 0$ for any $M \geq 0$, there exists an N_X such that, for $n \geq N_X$, $\Pr(X_n^2 \leq M_1 + M_Y) \leq \epsilon/2$. A union bound completes the argument (on the eventuality

$n \geq N_X$):

$$\begin{aligned}\Pr(X_n^2 - Y_n > M) &\geq \Pr(X_n^2 > M_1 + M_Y, |Y_n| \leq M_Y) \\ &= 1 - \Pr(\{X_n^2 < M_1 + M_Y\} \cup \{|Y_n| > M_Y\}) \\ &\geq 1 - \epsilon/2 - \epsilon/2 = 1 - \epsilon\end{aligned}$$

□

Proof of Lemma 1.4.2

Proof of Lemma 1.4.2. For N and D defined at the top of Section 1.10.1 define $\widehat{N} = N + \Delta_N$ and $\widehat{D} = D + \Delta_D$. We can then write $JK(\beta_0) = \widehat{N}^2/\widehat{D}$ and rewrite

$$JK(\beta_0) - JK_I(\beta_0) = \frac{2ND\Delta_N + D\Delta_N - N^2\Delta_D}{D^2 + D\Delta_D}$$

Apply Lemma 1.10.19 to see that $N^2 = O_p(1)$ while under Assumption 1.4.2, $D = O_p(1)$. Thus, $2ND\Delta_n + D\Delta_n - N^2\Delta_D = o_p(1)$. Meanwhile, by Lemma 1.10.11, $\Pr(D^2 \leq \delta_n) \to 0$ for any sequence $\delta_n \to 0$. Apply Lemma 1.10.9 to conclude. □

Lemma 1.10.9. *Let A_n, B_n and Y_n be sequences of random variables such that $A_n = o_p(1)$ and $B_n = o_p(1)$. If Y_n is such that for any sequence $\delta_n \to 0$, $\Pr(|Y_n| \leq \delta_n) \to 0$, then,*

$$\left|\frac{A_n}{Y_n + B_n}\right| = o_p(1)$$

Proof. Fix any $\epsilon > 0$. We show that

$$\left|\frac{A_n}{Y_n + B_n}\right| \leq \epsilon$$

on an intersection of events whose probability tends to one. By Lemma 1.10.33 there is a

sequence $\epsilon_n \searrow 0$ such that

$$\Pr(|A_n| \leq \epsilon_n) \to 1 \quad \text{and} \quad \Pr(\epsilon|B_n| \leq \epsilon_n) \to 1$$

Consider the intersection of events $\Omega_1 \cap \Omega_2 \cap \Omega_3$ where

$$\Omega_1 := \{\epsilon|Y_n| \geq 2\epsilon_n\}, \quad \Omega_2 := \{\epsilon|B_n| \leq \epsilon_n\}, \quad \Omega_3 := \{|A_n| \leq \epsilon_n\}$$

By assumption, $\Pr(\Omega_1 \cap \Omega_2 \cap \Omega_3) \to 1$. On this event $|Y_n + B_n| \geq \epsilon_n/\epsilon > 0$ and $|A_n| \leq \epsilon_n$ so that $|A_n/(Y_n + B_n)| \leq |\epsilon_n/(\epsilon_n/\epsilon)| \leq \epsilon$. □

Lemma 1.10.10 (Denominator Interpolation)**.** *Suppose that Assumptions 1.4.1 and 1.4.2 hold. Let $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ be such that $\varphi(\cdot) \in C_b^3(\mathbb{R})$ with $L_2(\varphi) = \sup_x |\varphi''(x)|$ and $L_3(\varphi) = \sup_x |\varphi'''(x)|$. Then there is a constant M that depends only on the constant c such that:*

$$|\mathbb{E}[\varphi(D) - \varphi(\tilde{D})]| \leq \frac{M}{\sqrt{n}}(L_2(\varphi) + L_3(\varphi))$$

Proof of Lemma 1.10.10. We inherit the definitions of D_{-i}, Δ_{2i}^a, Δ_{2i}^b, $\tilde{\Delta}_{2i}^a$, and $\tilde{\Delta}_{2i}^b$ from the proof of Lemma 1.10.1 with $a = 1$. Then, as before we can write

$$\begin{aligned}\mathbb{E}[\varphi(D) - \varphi(\tilde{D})] = \sum_{i=1}^{n} &\mathbb{E}[\varphi(D_{-i} + n^{-1}\Delta_{2i}^a + n^{-1}\Delta_{2i}^b)] \\ &- \mathbb{E}[\varphi(D_{-i} + n^{-1}\tilde{\Delta}_{2i}^a + n^{-1}\tilde{\Delta}_{2i}^b)]\end{aligned}$$

We examine each term via a second-order Taylor expansion around D_{-i}

$$\begin{aligned}\mathbb{E}[\text{Term}_i] = &\frac{1}{n}\mathbb{E}[\varphi'(D_{-i})\{(\Delta_{2i}^a - \tilde{\Delta}_{2i}^a) + (\Delta_{2i}^b - \tilde{\Delta}_{2i}^b)\}] \\ &+ \frac{1}{2n^2}\mathbb{E}[\varphi''(D_{-i})\{((\Delta_{2i}^a)^2 - (\tilde{\Delta}_{2i}^a)^2) + 2(\Delta_{2i}^a\Delta_{2i}^b - \tilde{\Delta}_{2i}^a\tilde{\Delta}_{2i}^b) + ((\Delta_{2i}^b)^2 - (\Delta_{2i}^b)^2)\}] \\ &+ R_i + \tilde{R}_i\end{aligned}$$

where R_i and $\tilde{R}_i$ are remainder terms to be analyzed later. Using the restrictions in (1.10.4) we can simplify the above display:

$$\mathbb{E}[\text{Term}_i] = \underbrace{0.5n^{-2}\mathbb{E}[\varphi''(D_{-i})((\Delta_{2i}^a)^2 - (\tilde{\Delta}_{2i}^a)^2)]}_{\dot{\mathbf{A}}_i} + \underbrace{n^{-2}\mathbb{E}[\varphi''(K_{-i})(\Delta_{2i}^a\Delta_{2i}^b - \tilde{\Delta}_{2i}^a\tilde{\Delta}_{2i}^b)}_{\dot{\mathbf{B}}_i}$$
$$+ R_i + \tilde{R}_i$$

Using Lemma 1.10.18 we can bound

$$|\mathbf{A}_i| \leq \frac{M}{n^2}L_2(\varphi) \qquad\qquad |\mathbf{B}_i| \leq \frac{M}{n^{3/2}}L_2(\varphi)$$

For some $\bar{D}_{1i}$ and $\bar{D}_{2i}$ we can express

$$R_i = \mathbb{E}[\varphi'''(\bar{D}_{1i})\{n^{-1}\Delta_{2i}^a + \Delta_{2i}^b\}^3] \leq \frac{M}{n^{3/2}}L_3(\varphi) + \frac{M}{n^3}L_3(\varphi)$$
$$R_i = \mathbb{E}[\varphi'''(\bar{D}_{2i})\{n^{-1}\tilde{\Delta}_{2i}^a + \tilde{\Delta}_{2i}^b\}^3] \leq \frac{M}{n^{3/2}}L_3(\varphi) + \frac{M}{n^3}L_3(\varphi)$$

where the inequalities again come from applications of Lemma 1.10.18. Combining these bounds and summing over the n terms gives the result. □

Lemma 1.10.11 (Denominator anti-concentration)**.** *Suppose that Assumptions 1.4.1 and 1.4.2 hold. Then, for any sequence $\delta_n \searrow 0$,*

$$\Pr(D \leq \delta_n) \to 0$$

Proof of Lemma 1.10.11. Let $\tilde{\varphi}(\cdot) : \mathbb{R} \to \mathbb{R}$ be three times continuously differentiable with bounded derivatives up to the third order such that $\tilde{\varphi}(x)$ is 1 if $x \leq 0$, $\tilde{\varphi}(x)$ is decreasing if $x \in (0, 1)$, and $\tilde{\varphi}(x)$ is zero if $x \geq 1$. Consider a second sequence $\gamma_n \searrow 0$ slowly enough such that $(\gamma_n^{-2} + \gamma_n^{-3})/\sqrt{n} \to 0$. Take $\varphi_n(x) = \tilde{\varphi}(\frac{x-\delta_n}{\gamma_n})$. By Lemma 1.10.10 and since $\tilde{\varphi}(\cdot)$ has bounded derivatives up to the third order, there is a fixed constant $M_1 > 0$ that depends

only on c such that

$$\Pr(D \leq \delta_n) \leq \Pr(\tilde{D} \leq \delta_n + \gamma_n) + \frac{M_1}{\sqrt{n}}(\gamma_n^{-2} + \gamma_n^{-3})$$

Let γ_n be a sequence tending to zero such that $(\gamma_n^{-2} + \gamma_n^{-3})/\sqrt{n} \to 0$ and conclude by applying Lemma 1.10.2. □

Proof of Lemma 1.4.3

For any $j = 1, \ldots, d_b$ define the matrix $B_j = \text{diag}(b_j(z_1), \ldots, b_j(z_n))$ and collect observations $\epsilon(\beta_0) = (\epsilon_1(\beta_0), \ldots, \epsilon_n(\beta_0))' \in \mathbb{R}^n$, $r = (r_1, \ldots, r_n)' \in \mathbb{R}^n$, $\hat{r} = (\hat{r}_1, \ldots, \hat{r}_n)' \in \mathbb{R}^n$, and $\xi = (\xi_1, \ldots, \xi_n)' \in \mathbb{R}^n$. In addition, collect $b_\epsilon = (b_{\epsilon 1}, \ldots, b_{\epsilon n}) \in \mathbb{R}^{d_b \times n}$ where $b_{\epsilon i} = \epsilon_i(\beta_0) b(z_i) \in \mathbb{R}^{d_b}$. Finally, let $\boldsymbol{H} = \frac{s_n}{\sqrt{n}} H$, $\tilde{H} = s_n H$ and $\tilde{h}_{ij} = s_n h_{ij}$.

Step 1: $\Delta_N \to_p 0$. To show that $\Delta_N \to_p 0$ write

$$\begin{aligned}
\Delta_N &= |\epsilon(\beta_0)' \boldsymbol{H}(\hat{r} - r)| \\
&= |\epsilon(\beta_0)' \boldsymbol{H}(b_\epsilon' \hat{\gamma} - b_\epsilon' \gamma) - \epsilon(\beta_0)' \boldsymbol{H} \xi| \\
&\leq \underbrace{\max_{1 \leq j \leq d_b} |\epsilon(\beta_0)' \boldsymbol{H} B_j \epsilon(\beta_0)| \|\hat{\gamma} - \gamma\|_1}_{\mathbf{A}} + \underbrace{\|\epsilon(\beta_0)' \boldsymbol{H}\|_2 \|\xi_2\|_2}_{\mathbf{B}}
\end{aligned}$$

To bound **A** we move to apply Theorem 1.10.1 to the quadratic form $\epsilon(\beta_0)'(\boldsymbol{H} B_j)\epsilon(\beta_0)$. First notice that, under Assumption 1.4.4(v), we have

$$\|\mathbb{E}[\boldsymbol{H} b_j \epsilon(\beta_0)]\|_2 = \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E}[s_n \sum_{j \neq i} h_{ij} b(z_j) \epsilon_j(\beta_0)])^2 \leq c^2$$

In the notation of Theorem 1.10.1 this give us an upper bound on $\|\mathbb{E} f^{(1)}(X)\|_{\text{HS}}$. Next, Assumption 1.4.2 gives us that the Frobenius norm of $\boldsymbol{H} = \frac{s_n}{\sqrt{n}} \boldsymbol{H}$ is bounded, since the rows of $s_n H$ are square summable, $\sum_{j \neq i} (s_n h_{ij})^2 \leq c$ for all $i = 1, \ldots, n$. In the notation of Theorem 1.10.1 this gives us an upper bound on $\|\mathbb{E} f^{(2)}(X)\|_{\text{HS}}$. Applying Theorem 1.10.1

and a union bound then gives us that

$$\max_{1\leq j\leq d_b} |\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0) - \mathbb{E}[\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)]| = O_p(\log^{2/a}(d_b)) \tag{1.10.13}$$

Since $\max_{1\leq j\leq d_b} |\mathbb{E}[\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)]| \leq c$ under Assumption 1.4.4(v), (1.10.13) gives that

$$\max_{1\leq j\leq d_b} |\epsilon(\beta_0)'\boldsymbol{H}B_j\epsilon(\beta_0)| = O_p(\log^{2/a}(d_b))$$

Since $\log^{2/a}(d_b)\|\widehat{\gamma}-\gamma\|_1 \to_p 0$ by assumption, this yields that $\mathbf{A} \to_p 0$.

To bound $\mathbf{B}$ see that $\|\epsilon(\beta_0)'\boldsymbol{H}\|_2 = \frac{s_n^2}{n}\sum_{i=1}^n(\sum_{j\neq i} h_{ij}\epsilon_i(\beta_0))^2 = O_p(1)$ under Assumption 1.4.3(ii) while under Assumption 1.4.4 $\|\xi\|_2 = o(1)$.

Step 2: $\Delta_D \to_p 0$. Notice that $a^2-b^2 = 2b(a-b)+(a-b)^2$ and bound:

$$|\Delta_D| \leq \underbrace{\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\beta_0)\big|\sum_{j\neq i}\tilde{h}_{ij}r_j\big|}_{\mathbf{E}} \times \max_i |\sum_{j\neq i}\tilde{h}_{ij}(\hat{r}_j - r_j)|$$
$$+ \underbrace{\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\beta_0)}_{\mathbf{F}} \times \max_i |\sum_{j\neq i}\tilde{h}_{ij}(\hat{r}_j - r_j)|^2$$

Since both $\mathbf{E} = O_p(1)$ and $\mathbf{F} = O_p(1)$ under Assumptions 1.4.1 and 1.4.2, it suffices to show that

$$\max_i |\sum_{j\neq i}\tilde{h}_{ij}(\widehat{r}_j - r_j)| \to_p 0$$

To do so write

$$\max_i \big|\sum_{j\neq i}\tilde{h}_{ij}\{\hat{r}_j - r_j\}\big| \leq \underbrace{\max_{\substack{1\leq i\leq n\\ 1\leq j\leq d_b}} \big|\sum_{j\neq i}\tilde{h}_{ij}b(z_j)\epsilon_j(\beta_0)\big|\|\hat{\gamma}-\gamma\|_1}_{\mathbf{A}} + \underbrace{\max_{\substack{1\leq i\leq n\\ 1\leq j\leq d_b}} \big|\sum_{j\neq i}\tilde{h}_{ij}b(z_j)\xi_j\big|}_{\mathbf{B}}$$

To bound $\mathbf{A}$, note that by Assumption 1.4.4(v) $\max_{i,j}|\mathbb{E}[\sum_{j\neq i}\tilde{h}_{ij}b(z_j)\epsilon_j(\beta_0)| \leq c$. Under

Assumptions 1.4.2 and 1.4.4(ii), $\max_{i,j} \sum_{j \neq i} \tilde{h}_{ij}^2 b^2(z_j) \leq c^2$ so we can apply Theorem 1.10.1 and a union bound to obtain that

$$\max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq d_b}} \Big| \sum_{j \neq i} \tilde{h}_{ij} b(z_j) \epsilon_j(\beta_0) \Big| = O_p(\log^{1/a}(d_b n))$$

Along with the implied rate on $\|\hat{\gamma} - \gamma\|_1$ from Assumption 1.4.4(iv) this shows that $\mathbf{A} \to_p 0$.

To show that $\mathbf{B} \to 0$, use Cauchy-Schwarz, $\sum_{j \neq i} \tilde{h}_{ij}^2 b^2(z_j) \leq c$ for any i, j by Assumptions 1.4.2 and 1.4.4(ii), and $\sum_{i=1}^n \xi_i^2 = o(1)$ by Assumption 1.4.4(iii).

Proof of Theorem 1.4.1

Apply Lemma 1.10.12 with $X_n = JK(\beta_0)$, $Y_n = JK_I(\beta_0)$ and $Z_n = JK_G(\beta_0)$. The density of Z_n is uniformly bounded by Lemma 1.10.20.

Lemma 1.10.12. *Let X_n, Y_n, and Z_n be sequences of random variables such that $|X_n - Y_n| \to_p 0$, the distribution of Z_n is absolutely continuous with respect to Lebesgue measure and the density functions of Z_n are uniformly bounded and $\sup_{a \in \mathbb{R}} |\Pr(Y_n \leq a) - \Pr(Z_n \leq a)| \to 0$. Then $\sup_{a \in \mathbb{R}} |\Pr(X_n \leq a) - \Pr(Z_n \leq a)| \to 0$.*

Proof. For any $a \in \mathbb{R}$ and $\epsilon > 0$ we have that $\{X_n \leq a\} \subseteq \{Y_n \leq a + \epsilon\} \cup \{|X_n - Y_n| > \epsilon\}$; thus, by applying union bound and rearranging we obtain:

$$\begin{aligned} \Pr(X_n \leq a) &\leq \Pr(Y_n \leq a + \epsilon) + \Pr(|Y_n - X_n| > \epsilon) \\ &\leq \Pr(Z_n \leq a + \epsilon) + |\Pr(Y_n \leq a + \epsilon) - \Pr(Z_n \leq a + \epsilon)| \\ &\quad + \Pr(|Y_n - X_n| > \epsilon) \end{aligned}$$

so that

$$\Pr(X_n \leq a) - \Pr(Z_n \leq a) \leq \Pr(a < Z_n \leq a + \epsilon) + |\Pr(Y_n \leq a + \epsilon) - \Pr(Z_n \leq a + \epsilon)|$$

$$+ \Pr(|Y_n - X_n| > \epsilon)$$

Let $\epsilon_n \to 0$ be a sequence tending to zero such that $\Pr(|X_n - Y_n| > \epsilon_n) \to 0$ (Lemma 1.10.33). Applying a supremum to the above display yields

$$\begin{aligned}\sup_{a\in\mathbb{R}} \Pr(X_n \leq a) - \Pr(Z_n \leq a) \leq & \sup_{a\in\mathbb{R}} \Pr(a < Z_n \leq a + \epsilon_n) \\ & + \sup_{a\in\mathbb{R}} |\Pr(Y_n \leq a + \epsilon_n) - \Pr(Z_n \leq a + \epsilon_n)| \\ & + \Pr(|Y_n - X_n| > \epsilon_n)\end{aligned}$$

The first term goes to zero as $\epsilon_n \to 0$ since Z_n has a uniformly bounded density; the second term goes to zero by $\sup_{a\in\mathbb{R}} |\Pr(Y_n \leq a) - \Pr(Z_n \leq a)| \to 0$ and the third term goes to zero by definition of ϵ_n and $|Y_n - X_n| \to_p 0$.

We can apply a symmetric argument to show that $\sup_{a\in\mathbb{R}} \Pr(Z_n \leq a) - \Pr(X_n \leq a) \leq o(1)$ which completes the claim of the lemma. □

1.10.2. Proofs of Results in Section 1.5

The statement of Theorem 1.5.1 relies on showing

$$\begin{aligned}&\sup_{(a_1,a_2)\in\mathbb{R}^2} \left|\Pr(JK(\beta_0) \leq a_1, C \leq a_2) - \Pr(JK_G(\beta_0) \leq a_1, C_G \leq a_2)\right| \to 0 \\ \text{and} \quad &\sup_{(a_1,a_2)\in\mathbb{R}^2} \left|\Pr(S(\beta_0) \leq a_1, C \leq a_2) - \Pr(S_G(\beta_0) \leq a_1, C_G \leq a_2)\right| \to 0\end{aligned}$$

In particular, since $(JK_G(\beta_0) \perp C_G)$ and $(S_G(\beta_0) \perp C_G)$ under H_0, showing the above will imply the test based on $T(\beta_0; \tau)$ has asymptotic size α for any choice of cutoff τ. The second line in the above display follows imediately from Theorem 1.10.5 after verifying Assumption 1.10.2, below.

The first line in the top display relies on a joint interpolation of the infeasible $JK_I(\beta_0)$ test

statistic and the infeasible conditioning statistic C_I, which could be constructed if $\rho(z_i)$ was known to the researcher.

$$C_I := \max_{1\leq i\leq n} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_{ij}r_j/(n^{-1}\sum_{i=1}^{n} h_{ij}^2)^{1/2}\Big| \tag{1.10.14}$$

This joint interpolation argument is rather involved however, and deferred to Section 1.10.4. The interpolation argument for the conditioning statistic very closely follows the results in Chernozhukov et al. (2013). The results of Section 1.5 rely on showing that the difference between C and C_I can be treated as negligible. This in turn reduces to verifying Assumption 1.10.2, which is done in Lemma 1.10.13, below.

Lemma 1.10.13. *Suppose that Assumption 1.4.4 holds. Then there are sequences $\delta_n \searrow 0$, $\beta_n \searrow 0$ such that*

$$\Pr\big(\max_{i\in[n]} n^{-1}\sum_{j=1}^{n} \dot{h}_{ij}^2(\widehat{r}_j - r_j)^2 > \delta_n^2/\log^2(n)\big) \leq \beta_n$$

where $\dot{h}_{ij} = h_{ij}/(n^{-1}\sum_{j=1}^{n} h_{ij}^2)^{1/2}$.

Proof. In view of Lemma 1.10.33 it suffices to show

$$\max_{1\leq i\leq n} \frac{1}{n}\sum_{j=1}^{n} \dot{h}_{ij}^2(\hat{r}_i - r_i)^2 = o_p(1/\log^2(n)) \tag{1.10.15}$$

Notice that we can bound

$$\begin{aligned}\max_{1\leq i\leq n} \frac{1}{n}\sum_{j=1}^{n}(\hat{r}_i - r_i)^2 &= \max_{1\leq i\leq n} \big|(\widehat{\gamma}-\gamma)' n^{-1}\sum_{j=1}^{n} \epsilon_j^2(\beta_0)b(z_i)b(z_j)'(\widehat{\gamma}-\gamma)\big| \\ &\quad + \max_{1\leq i\leq n} |n^{-1}\sum_{j=1}^{n} \dot{h}_{ij}^2\xi_j^2| \\ &\leq \max_{\substack{1\leq i\leq n \\ 1\leq j,k\leq d_b}} \underbrace{\big|n^{-1}\sum_{j=1}^{n} \epsilon_j^2(\beta_0)b_j(z_j)b_k(z_j)\big|}_{\mathbf{A}_{ijk}} \|\widehat{\gamma}-\gamma\|_1^2\end{aligned}$$

$$+ n^{-1/2} \max_{1 \leq i \leq n} (n^{-1} \sum_{j=1}^{n} \dot{h}_{ij}^4)^{1/2} (\sum_{j=1}^{n} \xi_j^4)^{1/2}$$

Under Assumption 1.4.4(i,ii) each $\mathbf{A}_{ijk}$ is υ-sub-exponential by Theorem 1.10.1 (that is $\|\mathbf{A}_{ijk}\|_{\psi_\upsilon}$ is bounded). An application of Lemma 1.10.34 then yields that $\max_{i,j,k} |\mathbf{A}_{ijk}| = O_p(\log^{1/\nu}(d_b n))$. Along with Assumption 1.4.4(iv) this gives that $\max_{i,j,k} |\mathbf{A}_{ijk}| \|\widehat{\gamma} - \gamma\|_1 = O_p(\log^{-3/(\upsilon \wedge 1)}(d_b n)) = o_p(\log^{-2}(n))$. Meanwhile by definition of $\dot{h}_{ij}$, $\max_i (n^{-1} \sum_{j=1}^n \dot{h}_{ij}^4)^{1/2} = O(1)$ while by Assumption 1.4.4(iii) $(\sum_{j=1}^n \xi_j^4)^{1/2} = o(1)$. Since $\log^2(n)/\sqrt{n} \to 0$ this shows (1.10.15). □

Proof of Theorem 1.5.1

The first result in Theorem 1.5.1 with $JK(\beta_0)$ and C replaced with their infeasible analogs $JK_I(\beta_0)$ and C_I follows from the argument in Section 1.10.4. After verifying that $|JK(\beta_0) - JK(\beta_0)| \to_p 0$ via Lemma 1.4.3 and that Assumption 1.10.2 is satisfied via Lemma 1.10.13 follow the same steps as in the proof of Belloni et al. (2018), Theorem 2.1 to see that approximation result holds for the feasible $JK(\beta_0)$ and C.

For the second statement, I show that the conditions of Theorem 1.10.6 are satisfied. To see that Assumption 1.10.1(i,ii) is satisfied under Assumption 1.4.1 use (i) the definition of $\dot{h}_{ij} = h_{ij}/(n^{-1} \sum_{j=1}^n h_{ij}^2)^{1/2}$; (ii) that the variance of each r_j is bounded away from zero and (iii) that the fourth moments of r_j are bounded from above. Assumption 1.10.1(iii) is satisfied with $B_n = \log^{1/\upsilon}(n)$ by Assumption 1.5.1(i,iii) and Lemma 1.10.34. Finally Assumption 1.10.2 is satisfied by applying Lemma 1.10.13. Apply Theorem 1.10.6 to conclude.

1.10.3. Proofs of Results in Section 1.6

Throughout this section, define the scaled elements of the infeasible and gaussian numerators and denominators

$$N_\ell = \frac{s_{n,\ell}}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0) \sum_{j=1}^{n} h_{ij} r_j \qquad \tilde{N}_\ell = \frac{s_{n,\ell}}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\epsilon}_i(\beta_0) \sum_{j=1}^{n} h_{ij} \tilde{r}_j$$

$$D_{\ell k} = \frac{s_{\ell,n} s_{m,k}}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0)(\sum_{j=1}^{n} h_{ij} r_{\ell j})(\sum_{j=1}^{n} h_{ij} r_{kj}) \quad \tilde{D}_{\ell k} = \frac{s_{\ell,n} s_{m,k}}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0)(\sum_{j=1}^{n} h_{ij} \tilde{r}_{\ell j})(\sum_{j=1}^{n} h_{ij} \tilde{r}_{kj})$$

Collect these in $N = (N_1, \dots N_{d_x})' \in \mathbb{R}^{d_x}$, $\tilde{N} = (\tilde{N}_1, \dots, \tilde{N}_{d_x})' \in \mathbb{R}^{d_x}$, $D = [D_{\ell k}]_{\ell,k \in [d_x]} \in \mathbb{R}^{d_x \times d_x}$, and $\tilde{D} = [\tilde{D}_{\ell k}]_{\ell,k \in [d_x]} \in \mathbb{R}^{d_x \times d_x}$. After multiplying by scaling matrix $\text{diag}(s_{1,n}, \dots, s_{d_x,n})$ and the inverse of the scaling matrix we rewrite the infeasible and gaussian test statistics

$$JK_I(\beta_0) = N' D^{-1} N \mathbf{1}_{\{\lambda_{\min}(D) > 0\}} \qquad JK_G(\beta_0) = \tilde{N}' \tilde{D}^{-1} \tilde{N}$$

These are the representations of the test statsitics we will largely work through in this section.

Proof of Lemma 1.6.1

Lemma 1.6.1 follows immediately from the joint gaussian approximation argument established in Section 1.10.4.

Proof of Lemma 1.6.2

Define the matrix $\Delta_D = [(\Delta_D)_{\ell k}]_{\ell,k \in [d_x]}$ and the vector $\Delta_N = [(\Delta_N)_\ell]_{\ell \in [d_x]}$ where

$$(\Delta_D)_{\ell k} \coloneqq \frac{s_{\ell,n} s_{k,n}}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0) \big(\widehat{\Pi}_{\ell,i} \widehat{\Pi}_{k,i} - \widehat{\Pi}^I_{\ell,i} \widehat{\Pi}^I_{k,i} \big)$$

$$(\Delta_N)_\ell \coloneqq \frac{s_{\ell,n}}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0)(\widehat{\Pi}_{\ell,i} - \widehat{\Pi}^I_{\ell,i})$$

Under the conditions of Lemma 1.6.2 we have that $\|\Delta_D\| \to_p 0$ and $\|\Delta_N\| \to_p 0$. Using this notation, we can write the infeasible version of the test statistic as $JK^I(\beta_0) = N' D^{-1} N$ while the feasible version is written $JK(\beta_0) = (N + \Delta_N)'(D + \Delta_D)^{-1}(N + \Delta_N)$. Add and subtract D^{-1} to get

$$\begin{aligned} JK(\beta_0) &= \big(N + \Delta_N\big)' \big((D + \Delta_D)^{-1} \pm D^{-1} \big) \big(N + \Delta_N\big) \\ &= JK^I(\beta_0) + N' \big((D + \Delta_D)^{-1} - D^{-1} \big) N + \Delta_N \big((D + \Delta_D)^{-1} - D^{-1} \big) N \end{aligned}$$

$$+ \Delta_N'\big((D+\Delta_D)^{-1} - D^{-1}\big)\Delta_N + N'D^{-1}\Delta_N + \Delta_N D^{-1}N + \Delta_N D^{-1}\Delta_N$$

Via Lemma 1.10.15 we have that $\|D^{-1}\| = (\lambda_{\min}(D))^{-1} = O_p(1)$ and by assumption we have that $\Delta_N \to_p 0$. It therefore suffices to show that

$$\|(D+\Delta_D)^{-1} - D^{-1}\| \to_p 0 \tag{1.10.16}$$

To do so, we can use the following equality from Horn and Johnson (2012), p. 381.

$$\|(D+\Delta_D)^{-1} - D^{-1}\| \leq \frac{\|D^{-1}\|^2\|\Delta_D\|}{1 - \|D^{-1}\Delta_D\|}$$

Since $\|D^{-1}\| = O_p(1)$ and $\Delta_D \to_p 0$, this gives (1.10.16).

Proof of Theorem 1.6.1

Under Assumption 1.6.4, the conditions of Lemma 1.6.2 can be verified following the same steps as the proof of Lemma 1.4.3. Combine Lemma 1.6.2 and Lemma 1.6.1 as in the proof of Theorem 1.4.1 to conclude.

Proof of Theorem 1.6.2

Follows from the same argument as the proof of Theorem 1.6.1 using the joint interpolation of $JK(\beta_0)$ and C established in Section 1.10.4.

1.10.4. Joint Gaussian Approximation of $JK(\beta_0)$ and C

The main results of Sections 1.5 and 1.6 rely on a joint interpolation of the conditioning and testing statistics as well as a joint interpolation of the conditioning and testing statistics. The joint interpolation of $JK(\beta_0)$ and the conditioning statistic C is given in Section 1.10.4 after introducing some notation in Section 1.10.4. The joint gaussian approximation of $S(\beta_0)$ and C follows immediately from results in Belloni et al. (2018), Chernozhukov et al. (2017).

The result is presented below for the general form of the $JK(\beta_0)$ statistic under H_0 however the proof strategy is very similar when using the decomposed form of $JK(\beta_0)$ when $d_x = 1$. This proof is available on request.

Notation

Jackknife Statistic Definitions. Define $\tilde{h}_{\ell,ij} = s_{n,\ell} h_{ij}$ for each $\ell = 1, \ldots, d_x$ and the scaled leave-one-out quasi-numerator and denominators

$$U_{-i} = \left[\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{\epsilon}_j(\beta_0) \sum_{k \neq i} \tilde{h}_{\ell,jk} \dot{r}_{\ell k} \right]_{1 \leq \ell \leq d_x} \in \mathbb{R}^{d_x}$$

$$D_{-i} = \left[\frac{1}{n} \sum_{j=1}^{n} \ddot{\epsilon}_i^2(\beta_0) \big(\sum_{k \neq i} \tilde{h}_{\ell,ij} \dot{r}_{\ell j} \big) \big(\sum_{k \neq i} \tilde{h}_{\ell,ij} \dot{r}_{mj} \big) \right]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d_x}} \in \mathbb{R}^{d_x \times d_x}$$

where $\dot{\epsilon}_j(\beta_0)$ is equal to $\tilde{\epsilon}_j(\beta_0)$ if $j < i$ and equal to $\epsilon_j(\beta_0)$ if $j > i$, $\dot{r}_{\ell j}$ is equal to $\tilde{r}_{\ell j}$ if $j < i$ and equal to r_j if $j > i$, and $\ddot{\epsilon}_j(\beta_0)$ is equal to $\mathbb{E}[\epsilon_j^2(\beta_0)]$ if $j < i$ and equal to $\epsilon_j(\beta_0)$ if $j > i$. As in the proof of Lemma 1.4.1 while the definitions of $\dot{\epsilon}_j(\beta_0), \dot{r}_{\ell j}$, and $\ddot{\epsilon}_j(\beta_0)$ depend on i this dependence is suppressed to conslidate notation and since we only consider one step deviations at a time.

Also define the one step deviations

$$\Delta_{Ui} = \big[\epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{\ell,ij} \dot{r}_{\ell j} + r_{\ell i} \sum_{j=1}^{n} \tilde{h}_{\ell,ji} \dot{\epsilon}_j(\beta_0) \big]_{1 \leq \ell \leq d} \in \mathbb{R}^d$$

$$\tilde{\Delta}_{Ui} = \big[\tilde{\epsilon}_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{\ell,ij} \dot{r}_{\ell j} + \tilde{r}_{\ell i} \sum_{j=1}^{n} \tilde{h}_{\ell,ji} \dot{\epsilon}_j(\beta_0) \big]_{1 \leq \ell \leq d} \in \mathbb{R}^d$$

$$\Delta_{Di} = \underbrace{\big[(\Delta^a_{Di})_{\ell m} \big]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d}}}_{\Delta^a_{Di}} + \underbrace{\big[(\Delta^b_{Di})_{\ell m} \big]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d}}}_{\Delta^b_{Di}}$$

$$\tilde{\Delta}_{Di} = \underbrace{\big[(\tilde{\Delta}^a_{Di})_{\ell m} \big]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d}}}_{\tilde{\Delta}^a_{Di}} + \underbrace{\big[(\tilde{\Delta}^b_{Di})_{\ell m} \big]_{\substack{1 \leq \ell \leq d \\ 1 \leq m \leq d}}}_{\tilde{\Delta}^b_{Di}}$$

where

$$(\Delta^a_{Di})_{\ell m} = \epsilon_i^2(\beta_0)\big(\sum_{j=1}^n \tilde{h}_{\ell,ij} r_{\ell j}\big)\big(\sum_{j=1}^n \tilde{h}_{\ell,ij}\dot{r}_{\ell j}\big)\big(\sum_{j=1}^n h_{m,ij} r_{m,ij}\big)^2 + r_{\ell i} r_{ki} \sum_{j=1}^n \tilde{h}_{\ell,ij}\tilde{h}_{m,ij}\ddot{\epsilon}_j^2(\beta_0)$$

$$(\tilde{\Delta}^a_{Di})_{\ell m} = \tilde{\epsilon}_i^2(\beta_0)\big(\sum_{j=1}^n \tilde{h}_{\ell,ij} r_{\ell j}\big)\big(\sum_{j=1}^n \tilde{h}_{\ell,ij}\dot{r}_{\ell j}\big)\big(\sum_{j=1}^n h_{m,ij} r_{m,ij}\big)^2 + \tilde{r}_{\ell i} \tilde{r}_{ki} \sum_{j=1}^n \tilde{h}_{\ell,ij}\tilde{h}_{m,ij}\ddot{\epsilon}_j^2(\beta_0)$$

$$(\Delta^b_{Di})_{\ell m} = r_{\ell i} \sum_{j=1}^n \ddot{\epsilon}_j^2(\beta_0) \sum_{k\neq i} \tilde{h}_{\ell,ji}\tilde{h}_{m,jk}\dot{r}_{mk} + r_{ki} \sum_{j=1}^n \ddot{\epsilon}_j^2(\beta_0) \sum_{k\neq i} \tilde{h}_{\ell,ji}\tilde{h}_{m,jk}\dot{r}_{\ell k}$$

$$(\tilde{\Delta}^b_{Di})_{\ell m} = \tilde{r}_{\ell i} \sum_{j=1}^n \ddot{\epsilon}_j^2(\beta_0) \sum_{k\neq i} \tilde{h}_{\ell,ji}\tilde{h}_{m,jk}\dot{r}_{mk} + \tilde{r}_{ki} \sum_{j=1}^n \ddot{\epsilon}_j^2(\beta_0) \sum_{k\neq i} \tilde{h}_{\ell,ji}\tilde{h}_{m,jk}\dot{r}_{\ell k}$$

Notice that in this notation we can write the test statistic and gaussian test statistics, after scaling by $\mathrm{diag}(s_{n,1},\ldots,s_{n,d_x})$, as

$$C(\beta_0) = (U_{-1} + \Delta_{U1}/\sqrt{n})'(D_{-1} + \Delta_{D1}/n)^{-1}(U_{-1} + \Delta_{U1}/\sqrt{n})\mathbf{1}\{\lambda_{\min}(D_{-1} + \Delta_{D1})^{-1}) > 0\}$$

$$\tilde{C}(\beta_0) = (U_{-n} + \tilde{\Delta}_{Un}/\sqrt{n})'(\tilde{D}_{-1} + \tilde{\Delta}_{D1}/n)^{-1}(U_{-n} + \tilde{\Delta}_{U1}/\sqrt{n})$$

In this proof we will use these representations for the test statistics. Finally define

$$U = U_{-1} + \Delta_{U1}/\sqrt{n} \qquad \tilde{U} = U_{-n} + \tilde{\Delta}_{Un}/\sqrt{n}$$

$$D = D_{-1} + \Delta_{D1}/n \qquad \tilde{D} = D_{-n} + \Delta_{Dn}/n$$

Conditioning Statistic Definitions. Let $h_{\ell,ii} = 0$ for any $\ell = 1,\ldots,d_x$ and $i = 1,\ldots,n$. Define $\tilde{h}_{\ell,ij} = h_{\ell,ij}/\omega_{\ell i}$ for $\omega_{\ell i} = n^{-1}\sum_{j=1}^n |h_{\ell,ij}|$. Also define the one-step deviations:

$$\Delta_{Ci} \coloneqq (\tilde{h}_{1,ji} r_{1i}, -\tilde{h}_{1,ji} r_{1i}, \ldots, \tilde{h}_{d_x,ji} r_{d_x i}, -\tilde{h}_{d_x,ji} r_{d_x i})'_{1\leq j\leq n} \in \mathbb{R}^{2nd_x}$$

$$\Delta_{Ci} \coloneqq (\tilde{h}_{1,ji} \tilde{r}_{1i}, -\tilde{h}_{1,ji} \tilde{r}_{1i}, \ldots, \tilde{h}_{d_x,ji} \tilde{r}_{d_x i}, -\tilde{h}_{d_x,ji} \tilde{r}_{d_x i})'_{1\leq j\leq n} \in \mathbb{R}^{2nd_x}$$

And the leave-one-out vector

$$C_{-i} \coloneqq \frac{1}{\sqrt{n}} \sum_{j<i} \tilde{\Delta}_{Cj} + \frac{1}{\sqrt{n}} \sum_{j>i} \Delta_{Cj} \in \mathbb{R}^{2nd_x}$$

Notice that $C = \max_{1 \leq \iota \leq 2nd_x} (C_{-1} + \frac{1}{\sqrt{n}} \Delta_{C1})_\iota$ while $\tilde{C} = \max_{1 \leq \iota \leq 2nd_x} (C_{-n} + \Delta_{Cn})_\iota$.

Function Definitions. As in Chernozhukov et al. (2013) consider the "smooth max" function, $F_\beta : \mathbb{R}^p \to \mathbb{R}$ defined

$$F_\beta(z) = \beta^{-1} \log \left(\sum_{i=1}^{n} \exp(\beta z_i) \right)$$

which satisfies

$$0 \leq F_\beta(z) - \max_{1 \leq i \leq n} z_i \leq \beta^{-1} \log p.$$

Section 1.10.5 notes some useful properties of the smooth max function which we will use in the joint interpolation argument. In addition let $\varphi(\cdot) \in C_b^3(\mathbb{R})$ be such that $\varphi(x) = 1$ if $x \leq 0$, $\varphi'(x) < 0$ for $x \in (0, 1)$, and $\varphi(x) = 0$ for $x \geq 1$. For any $\gamma > 0$ and $a = (a_1, a_2)' \in \mathbb{R}^2$ define the function $\tilde{\varphi}(\cdot, \cdot, \cdot) : \mathbb{R}^{d_x} \times \text{vec}(\mathbb{R}^{d_x \times d_x}) \times \mathbb{R}^{2nd_x} \to \mathbb{R}$ via

$$\tilde{\varphi}_{\gamma,a}(u, \text{vec}(d), c) \coloneqq \phi_{\gamma,a_1}(u, \text{vec}(d)) \tau_{\gamma,a_2}(c) \tag{1.10.17}$$

where

$$\phi_{\gamma,a_1}(u, \text{vec}(d)) \coloneqq \varphi \left(\frac{u' d^{-1} u - a_1}{\gamma \lambda_{\min}^5(d)} \right)$$
$$\tau_{\gamma,a}(c) \coloneqq \varphi \left(\frac{F_{1/\gamma}(c) - a_2}{\gamma} \right)$$

The function $\tilde{\varphi}_{\gamma,a}(\cdot, \cdot, \cdot)$ is meant to approximate the indicator function $\mathbf{1}\{K(\beta_0) \leq a_1\}\mathbf{1}\{C \leq a_2\}$ with γ governing the quality of approximation. Where it is obvious, we will supress the

subscripts γ, a from our notation.

Main Argument

Lemma 1.10.14 (Joint Lindeberg Interpolation)**.** *Suppose that Assumptions 1.6.1–1.6.3 hold. Then there is a fixed constant M*

$$\left|\mathbb{E}[\tilde{\varphi}_{\gamma,a}(U, vec(D), C) - \tilde{\varphi}_{\gamma,a}(\tilde{U}, vec(\tilde{D}), \tilde{C})]\right| \leq \frac{M_1 \log^{M_2}(n)}{\sqrt{n}}(\gamma^{-1} + \gamma^{-2} + \gamma^{-3}) \tag{1.10.18}$$

Proof of Lemma 1.10.14. We can bound the difference on the left hand side of (1.10.18) using the telescoping sum

$$\begin{aligned}\sum_{i=1}^{n} \Big|\mathbb{E}[\tilde{\varphi}_{\gamma,a}(U_{-i} + \Delta_{Ui}/\sqrt{n}, \mathrm{vec}(D_{-i} + \Delta_{Di}/n), C_{-i} + \Delta_{Ci}/\sqrt{n})] \\ - \mathbb{E}[\tilde{\varphi}_{\gamma,a}(U_{-i} + \Delta_{Ui}/\sqrt{n}, \mathrm{vec}(D_{-i} + \Delta_{Di}/n), C_{-i} + \Delta_{Ci}/\sqrt{n})]\Big|\end{aligned} \tag{1.10.19}$$

By second degree Taylor expansion, we break each of the summands in (1.10.19) into first order, second order, and remainder terms; each of which are bounded below. We make use of the following moment conditions implied by (i) indpendence of observations across $i = 1, \ldots, n$ and (ii) the mean and covariance matrix of $(\epsilon_i(\beta_0), r_i)$ being equal to the mean and covariance matrix of $(\tilde{\epsilon}_i(\beta_0), r_i)$

$$\begin{aligned}0 &= \mathbb{E}[\Delta_{Ui} - \tilde{\Delta}_{Ui}|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{Ui}\Delta_{Ui}' - \tilde{\Delta}_{Ui}\tilde{\Delta}_{Ui}'|\mathcal{F}_{-i}] = \mathbb{E}[\mathrm{vec}(\Delta_{Di}) - \mathrm{vec}(\tilde{\Delta}_{Di})|\mathcal{F}_{-i}] \\ &= \mathbb{E}[\Delta_{Ci} - \tilde{\Delta}_{Ci}|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{Ui} \otimes \mathrm{vec}(\Delta_{Di}^b)' - \tilde{\Delta}_{Ui} \otimes \mathrm{vec}(\tilde{\Delta}_{Di}^b)'|\mathcal{F}_{-i}] \\ &= \mathbb{E}[\Delta_{Ci} \otimes \Delta_{Ui} - \tilde{\Delta}_{Ci} \otimes \tilde{\Delta}_{Ui}|\mathcal{F}_{-i}] = \mathbb{E}[\Delta_{Ci} \otimes \mathrm{vec}(\tilde{\Delta}_{Di}^b) - \tilde{\Delta}_{Ci} \otimes \mathrm{vec}(\tilde{\Delta}_{Di}^b)|\mathcal{F}_{-i}] \\ &= \mathbb{E}[\mathrm{vec}(\Delta_{Di}^b)\mathrm{vec}(\Delta_{Di}^b)' - \mathrm{vec}(\tilde{\Delta}_{Di}^b)\mathrm{vec}(\tilde{\Delta}_{Di}^b)'|\mathcal{F}_{-i}]\end{aligned} \tag{1.10.20}$$

where $\mathcal{F}_{-i}$ denotes the sub-sigma algebra generated by all observations not equal to i, $\otimes$ denotes the Kronecker product, and I apologize for the abuse of the equal sign in the above

display.

First Order Terms. First order terms can be expressed

$$\begin{aligned}\text{First Order}_i &= \sum_{\ell=1}^{d_x} \mathbb{E}\left[\frac{\partial}{\partial U_\ell}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Ui})_\ell - (\tilde{\Delta}_{Ui})_\ell)\right]/\sqrt{n} \\ &+ \sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x} \mathbb{E}\left[\frac{\partial}{\partial D_{\ell m}}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Di})_{\ell m} - (\tilde{\Delta}_{Di})_{\ell m})\right]/n \\ &+ \sum_{\ell=1}^{2nd_x} \mathbb{E}\left[\frac{\partial}{\partial C_\ell}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Ci})_\ell - (\tilde{\Delta}_{Ci})_\ell)\right]/\sqrt{n}\end{aligned}$$

These terms are all equal to zero after applying the matched moments in (1.10.20).

Second Order Terms. After canceling out terms using the matched moments in (1.10.20) the second order terms that remain can be expressed

$$\begin{aligned}\text{2nd Order}_i &= \frac{1}{n^{3/2}}\sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x}\sum_{n=1}^{d_x} \underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial U_\ell \partial D_{mn}}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Ui})_\ell(\Delta^a_{Di})_{mn} - (\tilde{\Delta}_{Ui})_\ell(\tilde{\Delta}^a_{Di})_{mn})\right]}_{\mathbf{A}_{\ell mn}} \\ &= \frac{1}{n^2}\sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x}\sum_{n=1}^{d_x}\sum_{o=1}^{d_x} \underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial U_\ell \partial D_{mn}}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta^a_{Di})_{\ell m}(\Delta^a_{Di})_{no} - (\tilde{\Delta}^a_{Di})_{\ell m}(\tilde{\Delta}^a_{Di})_{no})\right]}_{\mathbf{B}_{\ell mno}} \\ &= \frac{2}{n^2}\sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x}\sum_{n=1}^{d_x}\sum_{o=1}^{d_x} \underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial U_\ell \partial D_{mn}}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta^b_{Di})_{\ell m}(\Delta^a_{Di})_{no} - (\tilde{\Delta}^a_{Di})_{\ell m}(\tilde{\Delta}^b_{Di})_{no})\right]}_{\mathbf{C}_{\ell mno}} \\ &= \frac{1}{n^{3/2}}\sum_{\ell=1}^{2nd_x}\sum_{m=1}^{d_x}\sum_{n=1}^{d_x} \underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial C_\ell \partial D_{mn}}\tilde{\varphi}(U_{-i}, \text{vec}(D_{-i}), C_{-i})((\Delta_{Ci})_\ell(\Delta^a_{Di})_{mn} - (\tilde{\Delta}_{Ci})_\ell(\tilde{\Delta}^a_{Di})_{mn})\right]}_{\mathbf{D}_{\ell mn}}\end{aligned}$$

To bound each $\mathbf{A}_{\ell mn}$, $\mathbf{B}_{\ell mno}$, and $\mathbf{C}_{\ell mno}$ we use the fact that the second order derivatives of $\tilde{\varphi}$ are bounded up to a log power of n via repeated application of Lemmas 1.10.29 and 1.10.32. Under Assumption 1.6.1 the absolute value of terms $(\Delta_{Ui})_\ell, |\Delta^a_{Di}|_{mn}$, and $(\Delta^b_{Di}/\sqrt{n})_{no}$ can also be shown to have bounded third moments via the exact same steps as in the proof of Lemma 1.10.18. Putting these together with generalized Holder's inequality will yield a finite constants M_1 and M_2 such that $|\mathbf{A}_{lmn}| \leq M_1 \log^{M_2}(n)(\gamma^{-1} + \gamma^{-2})$, $\mathbf{B}_{\ell mno} \leq M_1 \log^{M_2}(n)(\gamma^{-1} + \gamma^{-2})$, and $|\mathbf{C}_{\ell mno}| \leq M_1 \log^{M_2}(n) n^{1/2}(\gamma^{-1} + \gamma^{-2})$. To bound $\mathbf{D}_{\ell mn}$ terms notice that

$$\sum_{\ell=1}^{2nd_x} \mathbf{D}_{\ell mn} = \sum_{\ell=1}^{2nd_x} \mathbb{E}\left[\frac{\partial}{\partial D_{mn}}\phi(U_{-i}, \text{vec}(D_{-i}))\frac{\partial}{\partial C_\ell}\tau(C_{-i})((\Delta_{C-i})_\ell(\Delta^a_{Di})_{mn} - (\tilde{\Delta}_{Ci})_\ell(\tilde{\Delta}^a_{Di})_{mn}))\right]$$

Apply Lemma 1.10.18 to bound Δ_{Di}^a, and Lemmas 1.10.29 and 1.10.32 to bound the derivative of $\phi(\cdot)$ and Cauchy-Schwarz to split up the Δ_{Ci} and Δ_{Di} terms

$$\leq \sqrt{M_1 \log^{M_2}(n)\gamma^{-2}}\mathbb{E}\left[\sum_{\ell=1}^{2nd_x}(\partial_\ell \tau(C_{-i}))^2((\Delta_{Ci})_\ell + (\tilde{\Delta}_{Ci})_\ell)^2\right]^{1/2}$$
$$\leq \sqrt{M_1 \log^{M_2}(n)\gamma^{-2}}\mathbb{E}\left[\max_{1\leq \ell \leq n}((\Delta_{Ci})_{2\ell} + (\tilde{\Delta}_{Ci})_{2\ell})^2 \sum_{\ell=1}^{2nd_x}(\partial_\ell \tau(C_{-i}))^2\right]^{1/2}$$

By Lemma 1.10.25 and chain rule we have that $\sum_{\ell=1}^{2nd_x}(\partial_\ell \tau(C_{-i}))^2 \leq \gamma^{-2}$. Moreover $(\Delta_{Ci})_\ell^{a/2}$ is sub-exponential so via Lemma 1.10.34 the second moment of the maximum is bounded by a power of $\log(n)$. After updating the constant M_1 and M_2 this yields

$$\leq M_1 \log^{M_2}(n)\gamma^{-2}$$

Putting these all together and summing over the remaining indices gives

$$|\text{Second Order}_i| \leq \frac{M_1 \log^{M_2}(n)}{n^{3/2}}(\gamma^{-1} + \gamma^{-2}) \tag{1.10.21}$$

Remainder Terms. The first remainder term can be expressed

$$\begin{aligned}
\text{Remainder}_i = &\frac{1}{n^{3/2}}\sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x}\sum_{n=1}^{d_x}\mathbb{E}\left[\frac{\partial^3}{\partial U_\ell \partial U_m \partial U_n}\tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Ui})_\ell(\Delta_{Ui})_m(\Delta_{Ui})_n\right] \\
&+ \frac{1}{n^3}\sum_{(\ell,m)}\sum_{(n,o)}\sum_{(q,p)}\mathbb{E}\left[\frac{\partial^3}{\partial D_{\ell m}\partial D_{no}\partial D_{pq}}\tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Di})_{\ell m}(\Delta_{Di})_{no}(\Delta_{Di})_{qp}\right] \\
&+ \frac{1}{n^{3/2}}\sum_{\ell=1}^{2nd_x}\sum_{m=1}^{2nd_x}\sum_{n=1}^{2nd_x}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial C_m \partial C_n}\tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Ci})_n\right] \\
&+ \frac{1}{n^2}\sum_{\ell=1}^{d_x}\sum_{m=1}^{d_x}\sum_{(n,o)}\mathbb{E}\left[\frac{\partial^3}{\partial U_\ell \partial U_m \partial D_{no}}\tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Ui})_\ell(\Delta_{Ui})_m(\Delta_{Di})_{no}\right] \\
&+ \frac{1}{n^{5/2}}\sum_{\ell=1}^{d_x}\sum_{(m,n)}\sum_{(o,p)}\mathbb{E}\left[\frac{\partial^3}{\partial U_\ell \partial D_{mn} \partial D_{op}}\tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Ui})_\ell(\Delta_{Di})_{mn}(\Delta_{Di})_{op}\right] \\
&+ \frac{1}{n^{5/2}}\sum_{\ell=1}^{2nd_x}\sum_{(m,n)}\sum_{(o,p)}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial D_{mn} \partial D_{op}}\tilde{\varphi}(\bar{U}, \text{vec}(\bar{D}), \bar{C})(\Delta_{Ci})_\ell(\Delta_{Di})_{mn}(\Delta_{Di})_{op}\right]
\end{aligned}$$

$$+\frac{1}{n^2}\sum_{\ell=1}^{2nd_x}\sum_{m=1}^{2nd_x}\sum_{(n,o)}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial C_m \partial D_{no}}\tilde{\varphi}(\bar{U},\text{vec}(\bar{D}),\bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Di})_{no}\right]$$
$$+\frac{1}{n^{3/2}}\sum_{\ell=1}^{2nd_x}\sum_{m=1}^{2nd_x}\sum_{n=1}^{d_x}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial C_m \partial U_n}\tilde{\varphi}(\bar{U},\text{vec}(\bar{D}),\bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Ui})_n\right]$$
$$+\frac{1}{n^2}\sum_{\ell=1}^{2nd_x}\sum_{m=1}^{2nd_x}\sum_{n=1}^{d_x}\mathbb{E}\left[\frac{\partial^3}{\partial C_\ell \partial C_m \partial U_n}\tilde{\varphi}(\bar{U},\text{vec}(\bar{D}),\bar{C})(\Delta_{Ci})_\ell(\Delta_{Ci})_m(\Delta_{Ui})_n\right]$$

where $\bar{U},\text{vec}(\bar{D})$, and $\bar{C}$ vary term by term but are always in the hyper-rectangles $[U_{-i},U+\Delta_{Ui}]$, $[\text{vec}(D_{-i}),\text{vec}(D_{-i}+\Delta_{Di})]$, and $[C_{-i},C_{-i}+\Delta_{Ci}]$, respectively. As such, any moment conditions that apply to U,D,C also apply to $(\bar{U},\bar{D},\bar{C})$. Repeated application of generalized Hölder inequality, Lemma 1.10.18 to bound moments of Δ_{Ui} and $(\Delta_{Di}/\sqrt{n})$, Lemma 1.10.32 to bound moments of the second and third derivatives of $\phi(\tilde{U},\text{vec}(\tilde{D}))$, Lemma 1.10.28 to bound the sums of derivatives of $\tau(\tilde{C})$, and Lemma 1.10.34 to bound moments of $\max_{1\le\ell\le n}(\Delta_{Ci})_\ell$ will yield that

$$|\text{Remainder}_i| \le \frac{M_1\log^{M_2}(n)}{n^{3/2}}(\gamma^{-1}+\gamma^{-2}+\gamma^{-3}) \tag{1.10.22}$$

Symmetric logic will bound the other remainder term. Summing (1.10.21) and (1.10.22) over indices gives the result. □

Lemma 1.10.15 (Denominator Anticoncentration)**.** *Suppose that Assumptions 1.6.1–1.6.3 hold. Then for any sequence $\delta_n\to 0$ we have that $\Pr(\lambda_{\min}(\tilde{D})\le\tilde{\delta}_n)\to 0$.*

Proof. By Lemma 1.10.17 it suffices to show that for any fixed $a\in\mathcal{S}^{d_x-1}$ and any $\delta_n\to 0$, $\Pr(a'Da\le\delta_n)\to 0$. For any such a write:

$$a'\tilde{D}a = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\epsilon_i^2(\beta_0)]\Big(\sum_{\ell=1}^{d_x}\sum_{j=1}^{n}a_\ell\tilde{h}_{\ell,ij}r_{\ell,j}\Big)^2$$
$$\ge \frac{1}{cn}\sum_{i=1}^{n}\Big(\sum_{\ell=1}^{d_x}\sum_{j=1}^{n}a_\ell\tilde{h}_{\ell,ij}r_{\ell,j}\Big)^2$$

Define $\dot{s}_{n,j} = \max_{\{\ell: a_\ell \neq 0\}} s_{n,\ell}$ and $\dot{h}_{ij} = s_n h_{ij}$

$$= \frac{1}{cn} \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} \dot{h}_{ij} \sum_{\ell=1}^{d_x} \frac{a_\ell s_{n,\ell}}{s_n} r_{\ell,j} \Big)^2$$

By Assumption 1.6.1 we have $\lambda_{\min}(\mathbb{E}[D]) \geq \underline{c}$ so $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \big(\sum_{\ell=1}^{d_x} \sum_{j=1}^{n} a_\ell \tilde{h}_{\ell,ij} r_{\ell,j} \big)^2] \geq c^{-1}$. Moreover, by Assumption 1.6.1, $\mathrm{Var}(\sum_{\ell=1}^{d_x} \frac{a_\ell s_{n,\ell}}{s_n})$ is bounded from above and below. Define the matrix $\tilde{H} = [\dot{h}_{ij}]_{ij}$ and follow the same steps as Lemma 1.10.15 to conclude. □

Lemma 1.10.16 (Gaussian Approximation)**.** *Suppose that Assumptions 1.6.1–1.6.3 hold. Then*

$$\sup_{a \in \mathbb{R}} \big| \Pr(JK_I(\beta_0) \leq a) - \Pr(JK_G(\beta_0) \leq a) \big| \to 0$$

Proof. Let $a = (a_1, a_2)$ and $\tilde{\phi}_{\gamma,a}$ be as in (1.10.17):

$$\begin{aligned}
\Pr(N' D^{-1} N \leq a_1, C \leq a_2) &\leq \mathbb{E}[\tilde{\phi}_{\gamma,a}(U, \mathrm{vec}(D), C)] \\
&\leq \mathbb{E}[\tilde{\phi}_{\gamma,a}(\tilde{U}, \mathrm{vec}(\tilde{D}), \tilde{C})] + \frac{M_1 \log_2^M(n)}{\sqrt{n}} (\gamma^{-1} + \gamma^{-2}) \\
&\leq \Pr(\tilde{N}' \tilde{D}^{-1} \tilde{N} \leq a_1, \tilde{C} \leq a_2) + \Pr(a_1 \leq \tilde{N}' \tilde{D}^{-1} N \leq a_1 + \gamma \lambda_{\min}^5(D)) \\
&\quad + \Pr(a_2 \leq C \leq a_2 + \gamma) + \frac{M_1 \log_2^{M_2}(n)}{\sqrt{n}} (\gamma^{-1} + \gamma^{-2} + \gamma^{-3}) \\
&\leq \Pr(\tilde{N}' \tilde{D}^{-1} \tilde{N} \leq a_1, \tilde{C} \leq a_2) + \Pr(a_1 \leq \tilde{N}' \tilde{D}^{-1} N \leq a_1 + \gamma \lambda_{\min}^5(D)) \\
&\quad + \Pr(a_2 \leq C \leq a_2 + \gamma) + \frac{M_1 \log_2^{M_2}(n)}{\sqrt{n}} (\gamma^{-1} + \gamma^{-2} + \gamma^{-3})
\end{aligned}$$

Let $\gamma \to 0$ at a rate such that $\frac{\log^{M_2}(n)}{\sqrt{n}} \gamma^{-3} \to 0$ and apply Lemmas 1.10.14 and 1.10.15 to conclude as in the proof of Lemma 1.10.6. A symmetric argument shows that the lower bound tends to zero.

□

Lemma 1.10.17. *Let $\Sigma_n \in \mathbb{R}^{d\times d}$ be a sequence of random positive-semidefinite matrices. Suppose that for any fixed $a \in \mathcal{S}^{d-1}$ and any $\delta_n \to 0$ we have that $\Pr(a'\Sigma_n a \leq \delta_n) \to 0$ and $\Pr(\lambda^2_{\max}(\Sigma_n) \geq \delta_n^{-1}) \to 0$. Then for any $\delta_n \to 0$, $\Pr(\lambda^2_{\min}(\Sigma_n) \leq \delta_n) \to 0$.*

Proof. Take any preliminary sequence $\delta_n \to 0$. It suffices to show that there is another sequence $\tilde{\delta}_n$ weakly larger than $\delta_n/2$ such that $\Pr(\lambda^2_{\min}(\Sigma_n) \leq \tilde{\delta}_n) \to 0$. For any $m \in \mathbb{N}$ let $\mathcal{A}_m$ be a set of points in $\mathcal{S}^{d-1}$ such that

$$\max_{a\in\mathcal{S}^{d-1}} \min_{\tilde{a}\in\mathcal{A}_m} \|a - \tilde{a}\| \leq \delta_m^2$$

From here let $\tilde{n}_j$ be defined

$$\tilde{n}_j = \inf\{n \geq j : \min_{\tilde{a}\in\mathcal{A}_{n,j}} \Pr(\tilde{a}'\Sigma_n a \leq 2\delta_{n_j}) < \delta_{n_j}\}$$

Define a new sequence $\tilde{\delta}_n \to 0$, weakly larger than δ_n, via

$$\tilde{\delta}_n = \begin{cases} 1 & \text{if } 0 \leq 0 \leq n < \tilde{n}_1 \\ \delta_i & \text{if } \tilde{n}_i \leq n < \tilde{n}_{i+1} \end{cases}$$

and notice that, by definition $\Pr(\min_{a\in\mathcal{A}_{\tilde{n}_j}} a'\Sigma_n a \leq 2\tilde{\delta}_n) < \delta_{\tilde{n}_j}$. We wish to show that $\lambda^2_{\min}(\Sigma_n) > \tilde{\delta}_n$ on an intersection of events whose probability tends to one. Since Σ_n is positive semi-definite, $\|x\|^2_{\Sigma_n} = x'\Sigma_n x$ defines a seminorm. By triangle inequality

$$\lambda^2_{\min}(\Sigma_{n_j}) \geq \min_{\mathcal{A}_{n_j}} a'\Sigma_{n_j} a - \lambda^2_{\max}(\Sigma_n)\tilde{\delta}^2_{n_j}$$

Define the events

$$\Omega_1 = \{\min_{\mathcal{A}_{\tilde{n}_j}} a'\Sigma_n a \geq 2\tilde{\delta}_n\} \quad \text{and} \quad \Omega_2 = \{\lambda_{\max}(\Sigma_n) \leq \tilde{\delta}_n^{-1/2}\}$$

On the intersection of these events, whose probabilities tend to one, we have $\lambda^2_{\min}(\Sigma_n) \geq \tilde{\delta}_n$. □

1.10.5. Relevant Moment Bounds

Moment Bounds for Section 1.4

Here I provide some lemmas that are useful in the proof of Lemmas 1.10.1–1.10.6

Lemma 1.10.18. *Let $\Delta_{1i}, \tilde{\Delta}_{1i}, \Delta^a_{2i}, \tilde{\Delta}^a_{2i}, \Delta^b_{2i}, \tilde{\Delta}^b_{2i}$ be as in (1.10.2). Then under Assumptions 1.4.1 and 1.4.2 there is a constant $M > 0$ such that for any $k = 1, \ldots, 6$:*

$$\mathbb{E}[|\Delta_{1i}|^k] \leq M \qquad \mathbb{E}[|\tilde{\Delta}_{1i}|^k] \leq M$$

and for any $k = 1, \ldots, 3$:

$$\mathbb{E}[|\Delta^a_{2i}|^k] \leq M\alpha^k \qquad \mathbb{E}[|\tilde{\Delta}^k_{2i}|] \leq M\alpha^k$$
$$\mathbb{E}[|\Delta^b_{2i}/\sqrt{n}|^k] \leq M\alpha^k \qquad \mathbb{E}[|\tilde{\Delta}^b_{2i}/\sqrt{n}|^k] \leq M\alpha^k$$

Proof. First, since

$$\sum_{j=1}^n h_{ij}^2 \mathbb{E}[(r_j - \mathbb{E}[r_j])^2 \leq \mathbb{E}[(\sum_{i=1}^n \tilde{h}_{ij} r_j)^2] \leq 1$$

the constants are bounded, $\sum_{i=1}^n \tilde{h}_{ij}^2 \leq c$. Applying Lemma 1.10.21 with $X_i = h_{ij} r_j$ and $X_i = h_{ij}\epsilon_j(\beta_0)$ we see that there is a constant A such that for any $k = 1, \ldots, 6$

$$\mathbb{E}\big[\big|\sum_{i=1}^n \tilde{h}_{ij} r_j\big|^k\big] \leq A \quad \text{and} \quad \mathbb{E}\big[\big|\sum_{i=1}^n \tilde{h}_{ij}\epsilon_j(\beta_0)\big|^k\big] \leq A \tag{1.10.23}$$

The bounds on $\mathbb{E}[|\Delta^k_{1i}|]$ and $\mathbb{E}[|\tilde{\Delta}^k_{1i}|]$ immediately follow from this result and the bounds on

moments of r_i and $\epsilon_i(\beta_0)$ in Assumption 1.4.1. The bounds on $\mathbb{E}[|\Delta_{2i}^a|^k]$ and $\mathbb{E}[|\tilde{\Delta}_{2i}^a|^k]$ also follow from (1.10.23) after noting that there is a finite constant B such that:

$$\mathbb{E}[(\sum_{i=1}^{n} \tilde{h}_{ij}^2 \epsilon_i^2(\beta_0))^k] \leq B$$

Finally to bound $\mathbb{E}[|\Delta_{2i}^b/\sqrt{n}|^k]$ and $\mathbb{E}[|\tilde{\Delta}_{2i}^b/\sqrt{n}|^k]$ apply Lemma 1.10.23 with

$$v_j = \epsilon_j^2(\beta_0) \sum_{k \neq i,j} \tilde{h}_{jk} r_k$$

, noting that $\mathbb{E}[|v_j|^3]$ is bounded by (1.10.23). □

Lemma 1.10.19. *Let N and N_{-i} be defined as in Section 1.10.1. Under Assumptions 1.4.1–1.4.3 there is a fixed constant M such that for all $i = 1, \ldots, n$ and any $k = 1, \ldots, 6$,*

$$\mathbb{E}[|N|^k] + \mathbb{E}[|N_{-i}|^k] \leq M$$

Proof. We show the bound for $\mathbb{E}[|N|^k]$ and note that the bound for N_{-i} follows from symmetric logic. Write $\epsilon_i(\beta_0) = \eta_i + \gamma_i$ where $\gamma_i = \Pi_i(\beta - \beta_0)$ and η_i is mean zero. Decompose $N = N_1 + N_2 + N_3$:

$$N_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \sum_{j=1}^{n} \tilde{h}_{ij} \dot{r}_j, \; N_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} r_i \sum_{j=1}^{n} \tilde{h}_{ji} \gamma_j, \quad \text{and} \quad N_3 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \sum_{j=1}^{n} \tilde{h}_{ij} \mathbb{E}[r_j]$$

where $\dot{r}_j = r_j - \mathbb{E}[r_j]$.

Since via Assumption 1.4.2, $\sum_{i=1}^{n} h_{ji}^2 \leq c$ and via Assumption 1.4.1, $|\gamma_j| \leq c$, we can bound,

$$(\sum_{j=1}^{n} h_{ji} \gamma_j / \sqrt{n})^4 \leq (\frac{c}{\sqrt{n}} \sum_{i=1}^{n} |h_{ji}|)^4 \leq c^8 \implies (\sum_{j=1}^{n} h_{ji} \gamma_j / \sqrt{n})^6 \leq c^8 (\sum_{j=1}^{n} h_{ji} \gamma_j / \sqrt{n})^2$$

Under Assumption 1.4.3, $\mathbb{E}[N_2^2] \leq c$ while Assumption 1.4.2 implies that $(\sum_{i=1}^{n} h_{ij} \mathbb{E}[r_j])^2 \leq c$

so that $\mathbb{E}[N_3^2] \leq c^2$.

An absolute bound on the higher moments of N_2 then follows from an application of Lemma 1.10.21 with $X_i = r_i \sum_{j=1}^n h_{ji}\gamma_j/\sqrt{n}$. An absolute bound on the higher moments of N_3 follows from symmetric logic.

To bound higher moments of N_1 define $v_i = \sum_{j<i}\{\eta_i h_{ij} r_j + \dot{r}_i h_{ji}\eta_j\}$ and write $N_1 = \frac{1}{\sqrt{n}}\sum_{i=2}^n v_i$. The sequence $v_2, \ldots, v_n$ is a martingale difference array. Via the same procedure as the bounds on $\mathbb{E}[|\Delta_{1i}|^k]$ as in Lemma 1.10.18 one can verify that there is a fixed constant M such that $\mathbb{E}[|v_i|^k] \leq M$ for all $k = 1, \ldots, 6$. The bound on the higher moments of N then follows from Lemma 1.10.24.

The bounds for moments of N_{-i} follow symmetric logic. □

Lemma 1.10.20. *Let $\tilde{N}$ and $\tilde{D}$ be defined as in Section 1.10.1. Let $f(\cdot, \tilde{r})$ be the density function of $\frac{\tilde{N}}{\tilde{D}^{1/2}}|\tilde{r}$. Under Assumptions 1.4.1 and 1.4.3 there is a constant $M > 0$ such that $\sup_x |f(x, \tilde{r})| \leq M$ for almost all $\tilde{r}$.*

Proof. Recall that

$$\tilde{N} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\epsilon}_i(\beta_0)\sum_{j=1}^n \tilde{h}_{ij}\tilde{r}_j \quad \text{and} \quad \tilde{D}^{1/2} = \sqrt{\frac{1}{n}\sum_{i=1}^n \kappa_i^2(\beta_0)(\sum_{j=1}^n \tilde{h}_{ij}\tilde{r}_j)^2}$$

The distribution of $\tilde{\epsilon}_i(\beta_0)|\tilde{r}_i$ is

$$\tilde{\epsilon}_i(\beta_0)|\tilde{r} \sim N\left(\mu_i(r_i), (1-\rho_i^2)\operatorname{Var}(\epsilon_i(\beta_0))\right)$$

where $\mu_i(r_i) = \Pi_i(\beta - \beta_0) + \frac{\operatorname{Cov}(\epsilon_i(\beta_0), r_i)}{\operatorname{Var}(r_i)}(r_i - \mathbb{E}[r_i])$ and $\rho_i = \operatorname{corr}(\epsilon_i(\beta_0), r_i)$. Define $\bar{\Pi}_i := \sum_{j=1}^n \tilde{h}_{ij}\tilde{r}_j$. Then, conditional on $\tilde{r}$,

$$\frac{\tilde{N}}{\tilde{D}^{1/2}} \sim N\left(\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \mu_i(r_i)\bar{\Pi}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n \kappa_i^2(\beta_0)\bar{\Pi}_i^2}}, \frac{\frac{1}{n}\sum_{i=1}^n (1-\rho_i^2)\operatorname{Var}(\epsilon_i(\beta_0))\bar{\Pi}_i^2}{\frac{1}{n}\sum_{i=1}^n \kappa_i^2(\beta_0)\bar{\Pi}_i^2}\right) \tag{1.10.24}$$

The maximum of the normal density is proportional to the inverse of the standard deviation so it suffices to show that the variance in (1.10.24) is bounded away from zero. To this end, notice that under Assumptions 1.4.1 and 1.4.3

$$(1-\delta^2)c^{-2} \leq (1-\rho_i^2)\frac{\mathrm{Var}(\epsilon_i(\beta_0))}{\kappa_i^2(\beta_0)} \leq c^2$$

By Lemma 1.10.40 to this gives that the conditional variance is also larger than $(1-\delta^2)c^{-2} > 0$.

□

Lemma 1.10.21. *Let $X_1, \ldots, X_n$ be such that $\mathbb{E}[X_i] = \mu_i$ and $\mathbb{E}[(\sum_{i=1}^n X_i)^2] \leq C$. Suppose that for any $i = 1, \ldots, n$ there is a constant U such that*

$$\mathbb{E}[(X_i-\mu_i)^3] \leq U\mathbb{E}[(X_i-\mu_i)^2] \quad \textit{and} \quad \mathbb{E}[(X_i-\mu_i)^6]^{1/3} \leq U\mathbb{E}[(X_i-\mu_i)^2]$$

Then $\mathbb{E}[(\sum_{i=1}^n X_i)^6] \leq 64U^3C^3 + 32C^3$.

Proof. First write

$$\mathbb{E}[(\sum_{i=1}^n X_i)^2] = \sum_{i=1}^n \mathbb{E}(X_i-\mu_i)^2 + (\sum_{i=1}^n \mu_i)^2 \leq C$$

To bound $\mathbb{E}[(\sum_{i=1}^n X_i)^6]$ expand out

$$\begin{aligned}
\mathbb{E}[(\sum_{i=1}^n X_i)^6] &= \mathbb{E}[(\sum_{i=1}^n (X_i-\mu_i) + \sum_{i=1}^n \mu_i)^6] \\
&\lesssim \mathbb{E}[(\sum_{i=1}^n (X_i-\mu_i))^6] + (\sum_{i=1}^n \mu_i)^6 \\
&= \sum_{i=1}^n \mathbb{E}[(X_i-\mu_i)^6] + \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[(X_i-\mu_i)^3(X_j-\mu_j)^3] \\
&\quad + \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[(X_i-\mu_i)^4(X_j-\mu_j)^2]
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq i,j}\mathbb{E}[(X_i-\mu_i)^2(X_j-\mu_i)^2(X_k-\mu_k)^2] + (\sum_{i=1}^{n}\mu_i)^6 \\
&\leq \sum_{i=1}^{n}\mathbb{E}[(X_i-\mu_i)^6] + \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[(X_i-\mu_i)^3]\mathbb{E}[(X_j-\mu_j)^3] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[(X_i-\mu_i)^6]^{4/6}\mathbb{E}[(X_j-\mu_j)^6]^{2/6} \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq i,j}\mathbb{E}[(X_i-\mu_i)^6]^{1/3}\mathbb{E}[(X_j-\mu_i)^6]^{1/3}\mathbb{E}[(X_k-\mu_k)^6]^{1/3} \\
&+ C^3 \\
&= \left(\sum_{i=1}^{n}(\mathbb{E}[(X_i-\mu_i)^6])^{1/3}\right)^3 + \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[(X_i-\mu_i)^3]\mathbb{E}[(X_j-\mu_j)^3] + C^3 \\
&\leq \left(\sum_{i=1}^{n}(\mathbb{E}[(X_i-\mu_i)^6])^{1/3}\right)^3 + \left(\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu_i)^3]\right)^2 + C^3 \\
&\leq 2U^3\left(\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu_i)^2]\right)^3 + C^3 \\
&\leq 2U^3C^3 + C^3
\end{aligned}
$$

where the implied constant in the second line is 32 by an application of Lemma 1.10.40, the third line comes from expanding out the power, the first inequality by application of Hölder's inequality, and the penultimate inequality comes from applying bounds on the third and sixth central moments in terms of the second moments. □

Lemma 1.10.22. *Let $h = (h_1, \ldots, h_n) \in \mathbb{R}^n$ be such that $\sum_{i=1}^n h_i^2 \leq b$. Suppose that $X_1, \ldots, X_n$ are such that $\mathbb{E}[|X_i|^k] \leq M$ for all $k = 1, 2, 3$. Then*

$$\mathbb{E}\big[\big|\sum_{i=1}^{n} h_i^2 X_i\big|^3\big] \leq b^3 M^3$$

Proof. We can expand out

$$\begin{aligned}\mathbb{E}\big[\big|\sum_{i=1}^n h_i^2 X_i\big|^3\big] &\leq \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n h_i^2 h_j^2 h_k^2 \mathbb{E}[|X_i||X_j||X_k|] \\ &\leq M^3 \sum_{i=1}^n h_i^2 \sum_{j=1}^n h_j^2 \sum_{k=1}^n h_k^2 \\ &\leq M^3 \big(\sum_{i=1}^n h_i^2\big)^3 \leq c^3 M^3\end{aligned}$$

□

Lemma 1.10.23. *Let $v_1, \ldots, v_n$ be random variables such that $\mathbb{E}[|v_i|^3] \leq M$ for all $i = 1, \ldots, n$. Let $h = (h_1, \ldots, h_n) \in \mathbb{R}^n$ be a vector of weights such that $\|h\|_2 \leq c$. Then*

$$\mathbb{E}\big[\big|\frac{1}{\sqrt{n}}\sum_{i=1}^n h_i v_i\big|^3\big] \leq c^3 M$$

Proof. We can expand out

$$\begin{aligned}\mathbb{E}\big[\big|\frac{1}{\sqrt{n}}\sum_{i=1}^n h_i v_i\big|^3\big] &\leq \frac{1}{n^{3/2}}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n |h_i||h_j||h_k|\mathbb{E}[|v_i||v_j||v_k|] \\ &\leq \frac{M}{n^{3/2}}\sum_{i=1}^n |h_i| \sum_{j=1}^n |h_j| \sum_{k=1}^n |h_k| \leq \frac{M}{n^{3/2}}\|h\|_1^3 \leq Mc^3\end{aligned}$$

where the second inequality follows from generalized Hölder's inequality,

$$|\mathbb{E}[fgh]| \leq (\mathbb{E}[|f|^3]\mathbb{E}[|g|^3]\mathbb{E}[|h|^3])^{1/3}$$

and the fourth inequality from $\|h\|_1 \leq \sqrt{n}\|h\|_2$. □

Lemma 1.10.24. *Let $v_1, \ldots, v_n$ be a martingale difference array such that $\mathbb{E}[|v_i|^l] \leq M$ for*

all $l = 1, \dots, k$. Then there is a fixed constant C_k that only depends on k such that

$$\mathbb{E}[(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i)^k] \leq C_k M$$

Proof. We move to apply Theorem 1.10.3 with $X_t = \sum_{i=1}^{t} v_i/\sqrt{n}$.

$$\begin{aligned}\mathbb{E}[(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i)^k] &\leq \mathbb{E}[(\max_{s \leq n} \sum_{t=1}^{s} X_s)^k] \\ &\leq C_k \mathbb{E}\big[\big(\sum_{i=1}^{n} v_i^2/n\big)^{k/2}\big] \leq C_k \mathbb{E}\big[\frac{1}{n} \sum_{i=1}^{n} v_i^k\big] \leq C_k M\end{aligned}$$

where the second inequality comes from Theorem 1.10.3 and the third comes from an application of Jensen's inequality to the sample mean. □

Useful Properties of Smooth Max

Lemma 1.10.25 (Chernozhukov et al. (2013), Lemma A.2)**.** *For every $1 \leq j, k, l \leq p$,*

$$\partial_j F_\beta(z) = \pi_j(z), \qquad \partial_j \partial_k F_\beta(z) = \beta w_{jk}(z), \qquad \partial_j \partial_k \partial_l F_\beta(z) = \beta^2 q_{jkl}(z)$$

where for $\delta_{jk} := \mathbf{1}\{j = k\}$,

$$\pi_j(z) := e^{\beta z_j} \Big/ \sum_{i=1}^{n} e^{\beta z_i}, \qquad w_{jk} := (\pi_j \delta_{jk} - \pi_j \pi_k)(z)$$

$$q_{jkl}(z) := (\pi_j \delta_{jl} \delta_{jk} - \pi_j \pi_l \delta_{jk} - \pi_j \pi_k (\delta_{jl} + \delta_{kl}) + 2\pi_j \pi_k \pi_l)(z)$$

Moreover,

$$\pi_j(z) \geq 0, \qquad \sum_{j=1}^{p} \pi_i(z) = 1, \qquad \sum_{j,k=1}^{p} |w_{jk}(z)| \leq 2, \qquad \sum_{j,k,l=1}^{p} |q_{jkl}| \leq 6$$

Lemma 1.10.26 (Chernozhukov et al. (2013), Lemma A.3)**.** *For every $x, z \in \mathbb{R}^p$,*

$$|F_\beta(x) - F_\beta(z)| \leq \max_{1 \leq j \leq p} |x_j - z_j|.$$

Lemma 1.10.27 (Chernozhukov et al. (2013), Lemma A.4)**.** *Let $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ be such that $\varphi \in C_b^3(\mathbb{R})$ and define $m : \mathbb{R}^p \to \mathbb{R}$, $z \mapsto \varphi(F_\beta(z))$. The derivatives (up to the third order) of m are given*

$$\begin{aligned}
\partial_j m(z) &= (\partial g(F(\beta))\pi_j)(z) \\
\partial_j \partial_k m(z) &= (\partial^2 g(F_\beta)\pi_j \pi_k + \partial g(F_\beta)\beta w_{jk})(z) \\
\partial_j \partial_k \partial_l m(z) &= (\partial^3 g(F_\beta)\pi_j \pi_k \pi_l + \partial^2 g(F_\beta)\beta(w_{jk}\pi_l + w_{jl}\pi_k + w_{kl}\pi_j) + \partial g(F_\beta)\beta^2 q_{jkl})(z)
\end{aligned}$$

where π_j, w_{jk}, q_{jkl} are as described in Lemma 1.10.25.

Lemma 1.10.28 (Chernozhukov et al. (2013), Lemma A.5)**.** *Define*

$$L_1(\varphi) = \sup_x |\varphi'(x)|, L_2(\varphi) = \sup_x |\varphi''(x)|, \quad \text{and } L_3(\varphi) = \sup_x |\varphi'''(x)|$$

For every $1 \leq j, k, l \leq p$,

$$|\partial_j \partial_k m(z)| \leq U_{jk}(z) \quad \text{and} \quad |\partial_j \partial_k \partial_l m(z)| \leq U_{jkl}(z)$$

where for $W_{jk}(z) := (\pi_j \delta_{jk} + \pi_j \pi_k)(z)$,

$$\begin{aligned}
U_{jk}(z) &:= (L_2 \pi_j \pi_k + L_1 \beta W_{jk}(z) \\
U_{jkl}(z) &:= (L_3 \pi_j \pi_k \pi_l + L_2 \beta(W_{jk}\pi_l + W_{jl}\pi_k + W_{kl}\pi_j) + L_1 \beta^2 Q_{jkl})(z) \\
Q_{jkl}(z) &:= (\pi_j \delta_{jl} \delta_{jk} + \pi_j \pi_k \delta_{jk} + \pi_j \pi_k (\delta_{jl} + \delta_{kl}) + 2\pi_j \pi_k \pi_l)(z).
\end{aligned}$$

Moreover,

$$\sum_{j,k=1}^{p} U_{jk}(z) \leq (L_2 + 2L_1\beta) \quad \textit{and} \quad \sum_{j,k,l=1}^{p} U_{jkl}(z) \leq (L_3 + 6L_2\beta + 6L_1\beta^2).$$

Moment Bounds for Sections 1.5 and 1.6

Lemma 1.10.29. *Suppose that Assumption 1.6.1 holds and let N and D be as defined at the top of Section 1.10.4 Then under H_0, for any k there is a fixed constant C_k such that for any $\ell = 1, \ldots, d_x$*

$$\mathbb{E}[|N_\ell|^k] \leq C_k \quad \textit{and} \quad \mathbb{E}[|D_{\ell\ell}|^k] \leq C_k \log^{2k/a}(n)$$

Proof. Let $\eta_{\ell i} = r_i - \mathbb{E}[r_i]$ and write

$$N_\ell = \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} \eta_{\ell j}}_{N_\ell^1} + \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i(\beta_0) \sum_{j=1}^{n} \tilde{h}_{ij} \mathbb{E}[r_{\ell j}]}_{N_\ell^2}$$

To bound moments of N_ℓ^1 use the fact that N_ℓ^1 is a quadratic form in mean-zero a-sub-exponential variables. By Theorem 1.10.1, N_ℓ^1 is therefore also a-sub-exponential with parameter $a/2$; thus $(N_\ell^1)^{a/2}$ is sub-exponential and Lemma 1.10.34 provides the moment bound for arbitrary moments. To bound moments of N_ℓ^2 we use the fact that $\max_i \left| \sum_{j=1}^n \tilde{h}_{ij} \mathbb{E}[r_{\ell j}] \right|$ is bounded by assumption and apply Burkholder-Davis-Gundy (Theorem 1.10.3) after adding and subtracting $\mathbb{E}[\epsilon_i(\beta_0)]$.

To bound moments of $D_{\ell\ell}$ we decompose

$$|D| \leq \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2(\beta_0) \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} h_{ij} r_j \right|^2$$

Apply Theorem 1.10.1 to see that $\sum_{j=1}^n h_{ij} r_j$ is α-sub-exponential and Lemma 1.10.34 to bound the RHS by a log-power of n. □

Matrix Derivative Lemmas

The purpose of this section is largely to establish some matrix derivative expressions that will be useful for the Lindeberg interpolation in

Lemma 1.10.30. *Let $D \in \mathbb{R}^{d\times d}$ be a symmetric, real matrix such that $\det(D) \neq 0$. Let $N \in \mathbb{R}^d$ be a vector. The derivatives up to the derivatives of quadratic form $N'D^{-1}N$ are given.*
First Order:

$$\frac{\partial}{\partial N_l} = 2\sum_{j=1}^{d}(D^{-1})_{jl}N_j, \qquad \frac{\partial}{\partial D_{lm}} = -2\sum_{j=1}^{d}\sum_{k=1}^{d}(D^{-1})_{jl}(D^{-1})_{km}N_jN_k,$$

Second Order:

$$\frac{\partial^2}{\partial N_l N_m} = 2(D^{-1})_{lm}, \qquad \frac{\partial^2}{\partial N_l \partial D_{pq}} = -2\sum_{j=1}^{d}(D^{-1})_{jp}(D^{-1})_{ql}N_j,$$
$$\frac{\partial^2}{\partial D_{lm}\partial D_{qj}} = \sum_{j=1}^{d}\sum_{k=1}^{d}\left\{(D^{-1})_{lp}(D^{-1})_{qj})(D^{-1})_{km} + (D^{-1})_{kp}(D^{-1})_{mq}(D^{-1})_{lj}\right\}N_jN_k$$

Third Order:

$$\frac{\partial^3}{\partial N_l \partial N_m \partial N_p} = 0, \qquad \frac{\partial^3}{\partial N_l \partial N_m \partial D_{pq}} = -2(D^{-1})_{lp}(D^{-1})_{qm}$$
$$\frac{\partial^3}{\partial D_{lm}\partial D_{pq}\partial N_r} = 2\sum_{j=1}^{d}\left\{(D^{-1})_{lp}(D^{-1})_{qj}(D^{-1})_{rm} + (D^{-1})_{rp}(D^{-1})_{mq}(D^{-1})_{lj}\right\}N_j$$
$$\begin{aligned}\frac{\partial^3}{\partial D_{lm}D_{pq}D_{rs}} = 2\sum_{j=1}^{d}\sum_{j=1}^{d}\Big\{&(D^{-1})_{lr}(D^{-1})_{ps}(D^{-1})_{qj}(D^{-1})_{km} + (D^{-1})_{lp}(D^{-1})_{qr}(D^{-1})_{js}(D^{-1})_{km}\\ &+ (D^{-1})_{lp}(D^{-1})_{qj}(D^{-1})_{kr}(D^{-1})_{ms} + (D^{-1})_{kr}(D^{-1})_{ps}(D^{-1})_{mq}(D^{-1})_{lj}\\ &+ (D^{-1})_{kp}(D^{-1})_{mr}(D^{-1})_{qs}(D^{-1})_{lj} + (D^{-1})_{rp}(D^{-1})_{mq}(D^{-1})_{lr}(D^{-1})_{js}\Big\}N_jN_k\end{aligned}$$

Proof. The derivative of an element of the the inverse of a matrix $\boldsymbol{X}$ can be expressed (Petersen and Pedersen, 2012)

$$\frac{\partial(\boldsymbol{X}^{-1})_{kl}}{\partial \boldsymbol{X}_{ij}} = -(\boldsymbol{X}^{-1})_{ki}(\boldsymbol{X}^{-1})_{jl} \tag{1.10.25}$$

repeated application of this identity as well as the expression of the quadratic form

$$N'D^{-1}N = \sum_{j=1}^{d}\sum_{k=1}^{d}(D^{-1})_{jk}N_jN_k$$

leads to the result, bearing in mind that the inverse of a symmetric matrix is symmetric. □

Lemma 1.10.31. *Let D be a symmetric positive definite matrix. Then, for any $p > 3$, the derivatives of $(\det(D))^p$ are given up to the third order by*

$$\begin{aligned}
\frac{\partial\,(\det(D))^p}{\partial D_{lm}} &= p(\det(D))^{p-1}(D^{-1})_{lm} \\
\frac{\partial^2\,(\det(D))^p}{\partial D_{lm}\partial D_{pq}} &= \frac{p!}{(p-2)!}(\det(D))^{p-2}(D^{-1})_{pq}(D^{-1})_{lm} \\
&\quad + p(\det(D))^{p-1}(D^{-1})_{lp}(D^{-1})_{mq} \\
\frac{\partial^3\,(\det(D))^p}{\partial D_{lm}\partial D_{pq}\partial D_{rs}} &= \frac{p!}{(p-3)!}(\det(D))^{p-3}(D^{-1})_{rs}(D^{-1})_{pq}(D^{-1})_{lm} \\
&\quad + \frac{p!}{(p-2)!}(\det(D))^{p-2}\Big\{(D^{-1})_{pq}(D^{-1})_{lr}(D^{-1})_{ps} + (D^{-1})_{pr}(D^{-1})_{qs}(D^{-1})_{lm} \\
&\qquad\qquad + (D^{-1})_{rs}(D^{-1})_{lp}(D^{-1})_{mq}\Big\} \\
&\quad + p(\det(D))^{p-1}\Big\{(D^{-1})_{lr}(D^{-1})_{qs}(D^{-1})_{mq} + (D^{-1})_{lp}(D^{-1})_{mr}(D^{-1})_{qs}\Big\}
\end{aligned}$$

Proof. We can express the derivative of the detrminant (Petersen and Pedersen, 2012),

$$\frac{\partial,\det(\boldsymbol{X})}{\partial\,\boldsymbol{X}_{ij}} = \det(\boldsymbol{X})(\boldsymbol{X}^{-1})_{ij} \tag{1.10.26}$$

Repeated application of this and (1.10.25) yields the result. □

Lemma 1.10.32. *For any $p > 4$ define the function $\gamma(N, vec(D)) : \mathbb{R}^d \times \mathbb{R}^{d^2}$ by*

$$\gamma(N, vec(D)) := \begin{cases} (\det(D))^p(N'D^{-1}N - c) & \text{if } \det(D) \neq 0 \\ 0 & \text{if } \det(D) = 0 \end{cases}$$

This function is thrice continously differentiable. Futher the k^{th} moments of all partial derivatives of this function up to the third order are bounded

$$\mathbb{E}[(\partial^{\alpha}\gamma(N, vec(D))^{k}] \leq C_k(\max_{\iota \leq d}\mathbb{E}[|D_{\iota\iota}|^{2pdk}] \vee \max_{\iota \leq d}\mathbb{E}[|N_{\iota\iota}|^{6k})$$

where C_k is a positive constant that only depends on k and d.

Proof. The first statement is clear by examination of the derivatives in Lemmas 1.10.30 and 1.10.31 as well as the inequality (1.10.27) below. For the moment bounds, we may extensive use of following bounds on elements of D^{-1} for a positive-definite D^{-1}:

$$\begin{aligned} |\det(D)(D^{-1})_{jk}| \leq \det(D)\text{trace}(D^{-1}) &\leq d\lambda_{\max}(D^{-1})\big(\prod_{m=1}^{d}\lambda_m(D)\big) \\ &= d\prod_{m=2}^{d}\lambda_m(D) \\ &\leq d\big(\sum_{m=2}^{d}\lambda_m(D)\big)^{d-1} \\ &\leq d(\text{trace}(D))^{d-1} \end{aligned} \tag{1.10.27}$$

where the first inequality uses the fact that the largest element of a positive semidefinite matrix is on the diagonal and the fact that the diagonal elements of a positive semidefinite matrix are weakly positive, the second inequality uses the fact that the trace is the sum of the eigenvalues and the determinant is the product of the eigenvalues, the equality comes from $\frac{1}{\lambda_{\min}(D)} = \lambda_{\max}(D^{-1})$, the third inequality uses the AM-GM inequality and the fourth again uses that the trace is the sum of the (weakly positive) eigenvalues.

The moment bounds follow from (1.10.27) and the expressions in Lemmas 1.10.30 and 1.10.31. We give an example of how this is done for the first order derivatives, higher order derivatives follow from similar logic. For the following let A be an arbitrary random variable. *First*

Order.

$$\begin{aligned}
\mathbb{E}\left|A\frac{\partial\gamma}{\partial N_l}\right|^k &\lesssim \sum_{j=1}^{d}\mathbb{E}|(\operatorname{trace}(D))^{kdp}N_j^kA^k| \\
&\lesssim \sum_{j=1}^{d}\sum_{\iota=1}^{d}\mathbb{E}[D_{\iota\iota}^{kdp}N_j^kA^k] \\
&\leq \sum_{j=1}^{d}\sum_{\iota=1}^{d}\gamma^{2kdp}\mathbb{E}[N_j^{2k}A^{2k}] \\
\mathbb{E}\left|A\frac{\partial\gamma}{\partial D_{lm}}\right|^k &= p\mathbb{E}\left|A\det(D)^{p-1}\sum_{j=1}^{d}\sum_{j'=1}^{d}(D^{-1})_{lm}(D^{-1})_{jj'}N_jN_{j'}\right|^k \\
&\lesssim p\sum_{j=1}^{d}\sum_{j'=1}^{d}\mathbb{E}[|(\operatorname{trace}(D))^{2k(d-1)+(p-3)kd}A^kN_j^kN_{j'}^k| \\
&\leq \sum_{j=1}^{d}\sum_{j'=1}^{d}\gamma^{2kd(p-1)}\mathbb{E}[A^{2k}N_j^{2k}N_{j'}^{2k}]
\end{aligned}$$

□

1.10.6. Technical Lemmas

Probability Lemmas

Lemma 1.10.33. *Let X_n be a sequence of random variables such that $X_n = o_p(1)$, that is for any $\delta > 0$, $\Pr(|X_n| \geq \delta) \to 0$. Then, there is a sequence $\delta_n \to 0$ such that $\Pr(|X_n| \geq \delta_n) \to 0$.*

Proof. Take a preliminary sequence $\tilde{\delta}_n \to 0$ and define

$$\tilde{n}_j = \inf\{n : \Pr(|X_n| > \tilde{\delta}_j) < \tilde{\delta}_j\}$$

Because $\Pr(|X_n| > \delta) \to 0$ for any fixed δ, we know that n_j is finite. Define a new sequence

$\delta_n \to 0$ as below:

$$\delta_n = \begin{cases} 1 & \text{if } 0 \le n < \tilde{n}_1 \\ \tilde{\delta}_i & \text{if } \tilde{n}_i \le n < \tilde{n}_{i+1} \end{cases} \tag{1.10.28}$$

By construction, this sequence satisfies $\Pr(X_n \ge \delta_n) \le \delta_n$ whenever $n \ge n_1$. □

Lemma 1.10.34. *Suppose that $X_1, \dots, X_n$ are α-subexponential such that $\Pr(|X_i| \ge t) \le 2\exp(-t^\alpha/K)$ for all $t \ge 0$ and fixed constants K. For any $p \ge 1$ there is a constant C that depends only on p, K such that:*

$$\mathbb{E}\left[\max_{i \le n} \frac{|X_j|^p}{(1+\log i)^{p/\alpha}} \right] \le C$$

As a consequence

$$\mathbb{E}\big[\max_{i \le n} |X_i|^p \big] \le C(\log n)^{p/\alpha}$$

Proof. Argument below is provided for $\alpha = 1$. This can be extended to $\alpha \ne 1$ by noting that if $\Pr(|X_i| \ge t) \le 2\exp(-t^\alpha/K)$ for some $\alpha > 0$ then $\Pr(|X_i|^\alpha \ge t) \le 2\exp(-t/K)$.

$$\begin{aligned}
\mathbb{E} \max_{i \le n} \frac{|X_i|^p}{(1+\log i)^p} &= \int_0^\infty \Pr\left(\max_i \frac{|X_i|^p}{(1+\log i)^p} > t \right) dt \\
&= \int_0^{2^{p/\alpha}} \Pr\left(\max_i \frac{|X_i|^p}{(1+\log i)^p} > t \right) dt + \int_{2^{p/\alpha}}^\infty \Pr\left(\max_i \frac{|X_i|^p}{(1+\log i)^p} > t \right) dt \\
&\le 2^p + \int_{2^{p/\alpha}}^\infty \sum_{i=1}^n \Pr\left(\frac{|X_i|}{1+\log i} > t^{1/p} \right) dt \\
&\le 2^p + \int_{2^p}^\infty \sum_{i=1}^n 2\exp\left(-\frac{t^{1/p}(1+\log i)}{K} \right) dt \\
&= 2^p + 2\sum_{i=1}^n \int_{2^p}^\infty \exp\left(-\frac{t^{1/p}}{K} \right) i^{-t^{1/p}}\, dt \\
&\le 2^p + 2\sum_{i=1}^n \int_{2^p}^\infty \exp(-t^{-1/p}/K) i^{-2}\, dt
\end{aligned}$$

$$\leq 2^p + 2\left(\sum_{i=1}^{n} i^{-2}\right)\left(\int_{2^p}^{\infty} \exp(-t^{-1/p}/K)\, dt\right)$$

Both the integral and the summation are bounded, which gives the result. □

Matrix Lemmas

Lemma 1.10.35. *Given a matrix M and a matrix P of full rank, the matrix M and the matrix $P^{-1}MP$ have the same eigenvalues.*

Proof. Suppose λ is a eigenvalue of $P^{-1}MP$ with eigenvector p. Then

$$P^{-1}MPv = \lambda v \implies M(Pv) = \lambda Pv$$

Hence Pv is an eigenvector of M with eigenvalue λ. Similarly, given an eigenvector v of M, it can be shown that $P^{-1}v$ is an eigenvector of $P^{-1}MP$;

$$P^{-1}MP(P^{-1}v) = P^{-1}Mv = \lambda P^{-1}v$$

□

Lemma 1.10.36. *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be real symmetric positive semidefinite matrices. For an arbitary square matrix M let $\lambda_k(M)$ denote the k^{th} largest eigenvalue of M. Then for any $k = 1, \ldots, n$:*

$$\lambda_k(A)\lambda_n(B) \leq \lambda_k(AB) \leq \lambda_k(A)\lambda_1(B)$$

Lemma 1.10.37. *Let $D \in \mathbb{R}^{n\times n}$ be a diagonal real matrix such that $d_{ii} \in [u, U]$ for all $i = 1, \ldots, n$. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric real matrix. For an arbitrary square matrix M,*

let $\lambda_k(M)$ denote the k^{th} largest eigenvalue of M. Then for any $k = 1, \ldots, n$:

$$u\lambda_k(A^2) \leq \lambda_k(ADA) \leq U\lambda_k(A^2)$$

Proof. Consider any vector $a \in \mathbb{R}^n$ and define $\boldsymbol{a} = a'H$. Then

$$\begin{aligned}\alpha' HDH\alpha = \boldsymbol{a}' D\boldsymbol{a} = \sum_{i=1}^{n} d_{ii}(\boldsymbol{a}_i)^2 &\in \left[u\sum_{i=1}^{n}(\boldsymbol{a}_i)^2,\, U\sum_{i=1}^{n}(\boldsymbol{a}_i)^2\right] \\ &= \left[u \times a'H^2a,\, U \times a'H^2a\right]\end{aligned}$$

The result then follows from an application of Courant-Fischer-Weyl min-max principle. □

Lemma 1.10.38. *Let $X_1, \ldots, X_n$ denote i.i.d standard normal random variables and $a_1, \ldots, a_n$ denote weakly positive constants. Then*

$$\Pr\left(\sum_{i=1}^{n} a_i X_i^2 \leq \epsilon \sum_{i=1}^{n} a_i\right) \leq \sqrt{e\epsilon}$$

Miscellaneous Lemmas

Lemma 1.10.39. *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be two sequences of real numbers. If $a_i \leq Ub_i$ for some $U > 0$, then $\sum_i a_i / \sum_i b_i \leq U$. Conversely if $a_i \geq Lb_i$ for some $L > 0$ then $\sum_i a_i / \sum_i b_i \geq L$.*

Proof. Replace $a_i \leq Ub_i$ for the upper bound and $a_i \geq Lb_i$ for the lower bound. □

The following is a standard bound, but it is used a lot so it is restated here.

Lemma 1.10.40. *Let $a_1, \ldots, a_m$ be constants and $p > 1$. Then*

$$|a_1 + \ldots a_m|^p \leq m^{p-1} \sum_{i=1}^{m} |a_i|^p$$

Proof. Apply Hölder's inequality with $\frac{1}{p} + \frac{p-1}{p} = 1$ to the vectors $(a_1, \ldots, a_m) \in \mathbb{R}^m$ and $(1, \ldots, 1) \in \mathbb{R}^m$ □

1.10.7. Assorted Results from Literature

Concentration Inequalities and Tail Bounds

Theorem 1.10.1 (Gotze et al. (2021)*Theorem 1.2)**.** *Let $X_1, \ldots, X_n$ be independent random variables satisfying $\|X_i\|_{\Psi_a} \leq M$ for some $a \in (0, 1] \cup \{2\}$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be a polynomial of total degree $D \in \mathbb{N}$. Then for all $t > 0$;*

$$\Pr(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2 \exp\left(- \frac{1}{C_{D,a}} \min_{1 \leq d \leq D} \left(\frac{t}{M^d \|\mathbb{E} f^{(d)}(X)\|_{HS}} \right)^{a/d} \right)$$

In particular, if $\|\mathbb{E} f^{(d)}(X)\|_{HS} \leq 1$ for $d = 1, \ldots D$, then

$$\mathbb{E} \exp\left(\frac{C_{D,a}}{M^a} |f(X)|^{\frac{a}{D}} \right) \leq 2,$$

or equivalently

$$\|f(X)\|_{\Psi_{\frac{a}{D}}} \leq C_{d,a} M^D$$

Theorem 1.10.2 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent, mean-zero sub-gaussian random variables, and let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Then, for every $t \geq 0$, we have*

$$\Pr\left\{ \left| \sum_{i=1}^n a_i X_i \right| \geq t \right\} \leq 2 \exp\left(- \frac{ct^2}{K^2 \|a\|_2^2} \right)$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Theorem 1.10.3 (Burkholder-Davis-Gurdy for Discrete Time Martingales)**.** *For any $1 \leq k < \infty$ there exist positive constants c_k and C_k such that for all local martingales with $X_0 = 0$*

and stopping times τ

$$c_k \mathbb{E}\big[\big(\sum_{t=1}^{\tau}(X_t - X_{t-1})^2\big)^{k/2}\big] \leq \mathbb{E}\big[(\sup_{t\leq\tau} X_t)^k\big] \leq C_k \mathbb{E}\big[\big(\sum_{t=1}^{\tau}(X_t - X_{t-1})^2\big)^{k/2}\big]$$

Anticoncentration Bounds

Let $\xi \in \mathbb{R}^n$ follow a normal distribution on $\mathbb{R}^n$ with mean zero and covariance matrix Σ_ξ. Order the eigenvalues of Σ_ξ in non-increasing order $\lambda_{1\xi} \geq \lambda_{2\xi} \geq ... \geq \lambda_{n\xi}$. Define the quantities

$$\Lambda_{k\xi}^2 = \sum_{j=k}^{\infty} \lambda_{j\xi}^2, \quad k = 1, 2$$

Theorem 1.10.4 (Götze et al. (2019), Theorem 2.6)**.** *Let* ξ *be a gaussian element with zero mean and covariance* Σ_ξ*. Then it holds for any* $\boldsymbol{a} \in \mathbb{R}^n$ *that*

$$\sup_{x\geq 0} p_\xi(x, \boldsymbol{a}) \lesssim (\Lambda_{1\xi}\Lambda_{2\xi})^{-1/2}$$

where $p_\xi(x, a)$ *denotes the p.df of* $\|\xi - \boldsymbol{a}\|^2$*.*

We use the following anticoncentration lemma from Nazarov (2003) noted in Chernozhukov et al. (2017).

Lemma 1.10.41. *Let* $Y = (Y_1, \ldots, Y_p)'$ *be a centered Gaussian random vector in* $\mathbb{R}^p$ *such that* $\mathbb{E}[Y_j^2] \geq b$ *for all* $j = 1, \ldots, p$ *and some constant* $b > 0$*. Then for every* $y \in \mathbb{R}^p$ *and* $a > 0$*,*

$$\Pr(Y \leq y + a) - \Pr(Y \leq y) \leq Ca\sqrt{\log(p)}$$

where C *is a constant only depending on* b*.*

Gaussian Comparasions and Approximations

We also use the following gaussian approximation results from Belloni et al. (2018), Chernozhukov et al. (2017). Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be independent, mean zero, random vectors and let $Y_1, \ldots, Y_n \in \mathbb{R}^p$ be independent random vectors such that $Y_i \sim N(0, \mathbb{E}[X_i X_i'])$. Suppose that the researcher does not directly observe $X_1, \ldots, X_n$ but instead observes noisy estimates $\widehat{X}_1, \ldots, \widehat{X}_n \in \mathbb{R}^p$.

Define the sums

$$S_n^X = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{X}_i \qquad\qquad S_n^Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$$

Let $\mathcal{A}^{\text{re}}$ be the class of all hyperrectangles in $\mathbb{R}^p$; that is, $\mathcal{A}^{\text{re}}$ consists of all sets A of the form

$$A = \{w \in \mathbb{R}^p : a_j \leq w_j \leq b_j \text{ for all } j = 1, \ldots, p\}$$

for some $-\infty \leq a_j \leq b_j \leq \infty$, $j = 1, \ldots, p$. Define

$$\rho_n(\mathcal{A}^{\text{re}}) \coloneqq \sup_{A \in \mathcal{A}^{\text{re}}} \left| \Pr(S_n^X \in A) - \Pr(S_n^Y \in A) \right|$$

Bounding $\rho_n(\mathcal{A}^{\text{re}})$ relies on the following moment conditions:

Assumption 1.10.1. *Suppose there are constants $B_n \geq 1$, $b > 0, q > 0$ such that*

(i) $n^{-1} \sum_{i=1}^n \mathbb{E}[X_{ij}^2] \geq b$ *for all* $j = 1, \ldots, p$

(ii) $n^{-1} \sum_{i=1}^n \mathbb{E}[|X_{ij}|^{2+k}] \leq B_n^k$ *for all* $j = 1, \ldots, p$ *and* $k = 1, 2$.

(iii) $\mathbb{E}[(\max_{1 \leq j \leq p} |X_{ij}|/B_n)^4] \leq 1$ *for all* $i = 1, \ldots, n$ *and* $\left(\frac{B_n^4 \ln^7(pn)}{n}\right)^{1/6} \leq \delta_n$.

as well as the following bounds on the estiamtion error

Assumption 1.10.2. *The estimates $\hat{X}_1, \ldots, \hat{X}_n$ satisfy*

$$\Pr\left(\max_{1\leq j\leq p} \mathbb{E}_n[(\widehat{X}_{ij} - X_{ij})^2] > \delta_n^2/\log^2(pn)\right) \leq \beta_n$$

Theorem 1.10.5 (Belloni et al. (2018), Theorem 2.1)**.** *Suppose that Assumptions 1.10.1 and 1.10.2 hold. Then there is a constant C which depends only on b such that*

$$\rho_n(\mathcal{A}^{re}) \leq C\{\delta_n + \beta_n\}$$

Let $e_1, \ldots, e_n \overset{\text{iid}}{\sim} N(0,1)$ be generated independently of the data. A gaussian bootstrap draw is defined

$$S_n^{X,\star} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i \widehat{X}_i$$

Theorem 1.10.6 (Belloni et al. (2018), Theorem 2.2)**.** *Suppose that Assumptions 1.10.1 and 1.10.2 hold. Then there is a constant C which depends only on b such that*

$$\sup_{A\in\mathcal{A}^{re}} \left|\Pr_e(S_n^{X,\star} \in A) - \Pr(S_n^Y \in A)\right| \leq C\delta_n$$

with probability at least $1 - \beta_n - (\log n)^{-2}$ where $\Pr_e(\cdot)$ denotes the probability measure only taken with respect to the variables $e_1, \ldots, e_n$ conditional on the data used to estimate $\widehat{X}$.

1.11. Appendix: Incorporating Exogenous Controls

In this section, I analyze the model with exogeneous controls. To this end, define the vector $z_2 = (z_{21}', \ldots, z_{2n}')' \in \mathbb{R}^{n\times d_c}$. Let $P_2 = z_2(z_2'z_2)^{-1}z_2' \in \mathbb{R}^{n\times n}$ denote the projection onto the column space of z_2 and $M_2 = I_n - P_2$ denote the projection onto to orthocomplement of the column space. Focus will be on the case where $d_x = 1$ to simplify notation, but the basic concepts apply generally to $d_x > 1$.

For $y \coloneqq (y_1, \ldots, y_n)' \in \mathbb{R}^n$ and $x \coloneqq (x_1', \ldots, x_n')' \in \mathbb{R}^{n\times}$ define $y^\perp \coloneqq M_2 y$ and $x^\perp \coloneqq M_2 x$ as the "partialled out" versions of y and x, respectively. Let $y_i^\perp$ be the i^{th} element of $y^\perp$ and $x_i^\perp$ be the i^{th} element of $x^\perp$. From here we can define $\epsilon(\beta_0) \coloneqq y - x\beta_0$, $\epsilon^\perp(\beta_0) = M_2\epsilon(\beta_0)$ and $r^\perp \coloneqq M_2 r$ where as in the main text $r = (r_1, \ldots, r_n)'$ is constructed $r_i = x_i - \rho(z_i)\epsilon_i(\beta_0)$. The definition of $\rho(z_i)$ does not change after partialling out z_2 since all expectations are understood to be conditional on the instruments z. Notice that $\epsilon^\perp(\beta_0)$ is mean zero. Finally I assume that the controls have been partialled out of hat matrix so that the effective hat matrix is $M_2 H$ and the vector $\widehat{\Pi} \in \mathbb{R}^n$ is defined $\widehat{\Pi} = (M_2 H)(M_2 r)$. This does not make a difference for the numerator of the $JK(\beta_0)$ statistic but does affect the denominator slightly. When this is not done, inference may be conservative.

Using matrix notation in the numerator to make things clear, we can write the version of the $JK(\beta_0)$ statistic with the partialled out vectors, $\epsilon^\perp(\beta_0)$ and $r^\perp$, in terms of the original vectors, $\epsilon(\beta_0)$ and r,

$$\begin{aligned} JK_I(\beta_0) &= \frac{\left(\frac{1}{\sqrt{n}}\epsilon(\beta_0)' M_2 \tilde{H} M_2 r\right)^2}{\frac{1}{n}\sum_{i=1}^n (\epsilon_i^\perp(\beta_0))^2 \left(\sum_{j=1}^n \boldsymbol{h}_{ij} r_j\right)^2} \\ &= \frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(\beta_0)\sum_{j=1}^n \boldsymbol{h}_{ij} r_j\right)^2}{\frac{1}{n}\sum_{i=1}^n (\epsilon_i^\perp(\beta_0))^2 \left(\sum_{j=1}^n \boldsymbol{h}_{ij} r_j\right)^2} \end{aligned}$$

where $\boldsymbol{h}_{ij} = [M_2 \tilde{H} M_2]_{ij}$, $\tilde{H} = s_n H$, and $m_{ij} = [M_2]_{ij}$. I seek to characterize the limiting distribution of $JK(\beta_0)$ under H_0. To do so, we show that quantiles $JK(\beta_0)$ can be approximated by quantiles of the gaussian analog statistic

$$JK_G(\beta_0) = \frac{\left(\frac{1}{\sqrt{n}}\tilde{\epsilon}(\beta_0)' M_2 \tilde{H} M_2 \tilde{r}\right)^2}{\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\epsilon_i) \left(\sum_{j=1}^n \boldsymbol{h}_{ij} \tilde{r}_j\right)^2}$$

where $(\tilde{\epsilon}_i, \tilde{\epsilon}_i(\beta_0), \tilde{r}_i)$ are generated gaussian independent of the data and with the same mean and covariance as $(\epsilon_i, \epsilon_i(\beta_0), r_i)$. Since $\mathrm{Var}(\tilde{\epsilon}(\beta_0)) = \mathrm{Var}(\epsilon_i)$ under H_0, $\mathbb{E}[\tilde{\epsilon}(\beta_0)' M_2] = 0$, and $\tilde{r} \perp \tilde{\epsilon}(\beta_0)$, this gaussian analog statistic has a χ^2_1 distribution conditional on any realization

of $\tilde{r}$ and thus its unconditional distribution is also χ^2_1.

Showing that quantiles of $JK(\beta_0)$ can be approximated by quantiles of $\tilde{JK}(\beta_0)$ proceeds in two steps. In the first step, we show that $JK(\beta_0)$ converges in probability to an intermediate statistic.

$$JK^{\text{int}}(\beta_0) = \frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(\beta_0)\sum_{j=1}^n \boldsymbol{h}_{ij} r_j\right)^2}{\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(\sum_{j\neq i}\boldsymbol{h}_{ij} r_j)^2}$$

We will then show that quantiles of this intermediate statistic can be approximated by quantiles of $\tilde{JK}(\beta_0)$. In view of Lemma 1.4.2, it suffices to show for the first step that $\Delta_D \to_p 0$, where

$$\Delta_D = \frac{1}{n}\sum_{i=1}^n ((\epsilon_i^{\perp}(\beta_0))^2 - \epsilon_i^2)\widehat{\Pi}_i^2$$

To do this, notice that under H_0 we can write $\epsilon_i^{\perp}(\beta_0) = \epsilon_i + z_{2i}'(\widehat{\Gamma} - \Gamma)$ where $\widehat{\Gamma} = (z_2'z_2)^{-1}z_2\epsilon(\beta_0)$ is a $\sqrt{n}$-consistent estimate of Γ. Exploiting this fact we get

$$\Delta_D = (\widehat{\Gamma} - \Gamma)'\frac{1}{n}\sum_{i=1}^n (\widehat{\Pi}_i)^2 z_{2i}z_{2i}'(\widehat{\Gamma} - \Gamma) + 2(\widehat{\Gamma} - \Gamma)'\frac{1}{n}\sum_{i=1}^n \epsilon_i z_{2i}\widehat{\Pi}_i$$

Both of these terms will tend to zero by the consistency $\widehat{\Gamma}$ to Γ, giving that $\Delta_D \to_p 0$.

In our second step, we argue that quantiles of $JK^{\text{int}}(\beta_0)$ can be approximated by quantiles of $JK_G(\beta_0)$. To make this comparasion, we can follow almost exactly the same steps as in Section 1.10.1. The only difference between analysis in this case and analysis in the original case is that the partialling out of controls leads the test statistic to not strictly have a jackknife form; the effective hat matrix M_2HM_2 no longer has a deleted diagonal. However, as I will argue below, this will not make a difference in the interpolation argument since the diagonal terms of $[P_2]_{ii}$ are small in the sense that they sum to d_c.

The (1.10.2) analog one step deviations for the numerator are given

$$\Delta_{1i} = \epsilon_i(\beta_0)\sum_{j\neq i}\boldsymbol{h}_{ij}\dot{r}_j + r_i\sum_{j\neq i}\boldsymbol{h}_{ji}\dot{\epsilon}_j(\beta_0) + \boldsymbol{h}_{ii}\epsilon_i(\beta_0)r_i$$
$$\tilde{\Delta}_{1i} = \tilde{\epsilon}_i(\beta_0)\sum_{j\neq i}\boldsymbol{h}_{ij}\dot{r}_j + \tilde{r}_j\sum_{j\neq i}\boldsymbol{h}_{ji}\dot{\epsilon}_j(\beta_0) + \boldsymbol{h}_{ii}\tilde{\epsilon}_i(\beta_0)\tilde{r}_i$$

where as Section 1.10.1, a dotted variable is equal to the gaussian analog if $j > i$ but equal to the standard version otherwise. The first and second moments of the first two terms in Δ_{1i} can be matched with their gaussian analog terms as in the proof of Lemma 1.10.1. While we cannot match seconds moments of the third term in the one step deviation, this sum of all these third terms can be treated as negligible after scaling by $1/\sqrt{n}$ as $\sum_{i=1}^n |\boldsymbol{h}_{ii}| \lesssim d_c$. This is because $M_2\tilde{H}M_2 = \tilde{H} - P_2\tilde{H} - \tilde{H}P_2 - P_2\tilde{H}P_2$. The matrix $\tilde{H}$ has zeros on it's diagonal. Meanwhile

$$|[P_2\tilde{H}]_{ii}|^2 = \Big|\sum_{j=1}^n [P_2]_{ij}\tilde{H}_{ji}\Big|^2 \leq \Big(\sum_{j=1}^n [P_2]_{ij}^2\Big)\Big(\sum_{j\neq i} H_{ji}^2\Big) \lesssim [P_2]_{ii}$$

where the final inequality comes because the matrix P_2 is symmetric and idempotent and since $\Big(\sum_{j\neq i} H_{ji}^2\Big) \lesssim 1$ by Assumption 1.4.2(ii). A similar argument can be used to show that $[P_2\tilde{H}P_2]_{ii}^2 \lesssim [P_2]_{ii}$. Since P_2 is a projection matrix we must have that $\|P_2He_j\| \leq \|He_j\|$ for any basis vector $e_j \in \mathbb{R}^n$. Thus $\sum_{j=1}^n [P_2H]_{ji}^2 \leq \sum_{j=1}^n [H]_{ji}^2$. Finally, we can use the fact that the trace of P_2 is equal to its rank to show that $\sum_{i=1}^n |\boldsymbol{h}_{ii}| \lesssim d_c$

The one step deviations in the denominator can be bounded using the same logic. These one step deviations are given

$$\Delta_{2i} = \epsilon_i^2(\sum_{j\neq i}\boldsymbol{h}_{ij}\dot{r}_j)^2 + r_i^2\sum_{j\neq i}\boldsymbol{h}_{ji}^2\ddot{\epsilon}_j^2 + r_i\sum_{j\neq i}\ddot{\epsilon}_j(\sum_{k\neq j,i}\boldsymbol{h}_{ji}\boldsymbol{h}_{jk}r_k)$$
$$+ \epsilon_i^2\big(\boldsymbol{h}_{ii}^2r_i^2 + 2\boldsymbol{h}_{ii}r_j\sum_{j\neq i}\boldsymbol{h}_{ij}r_j\big)^2$$

$$\tilde{\Delta}_{2i} = \tilde{\epsilon}_i^2 (\sum_{j \neq i} \boldsymbol{h}_{ij} \dot{r}_j)^2 + \tilde{r}_i^2 \sum_{j \neq i} \boldsymbol{h}_{ji}^2 \ddot{\epsilon}_j^2 + \tilde{r}_i \sum_{j \neq i} \ddot{\epsilon}_j \big(\sum_{k \neq j,i} \boldsymbol{h}_{ji} \boldsymbol{h}_{jk} r_k \big)$$
$$+ \epsilon_i^2 \big(\boldsymbol{h}_{ii}^2 r_i^2 + 2 \boldsymbol{h}_{ii} r_j \sum_{j \neq i} \boldsymbol{h}_{ij} r_j)^2$$

where $\ddot{\epsilon}_j$ is equal to $\mathrm{Var}(\epsilon_j)$ if $j < i$ and equal to ϵ_j if $j > i$. The first three terms in this expansion are can be dealt with exactly as in the proof of Lemma 1.10.1. The fourth term is new, however summing over the fourth terms and scaling by $1/n$ will be negligible as $\sum_{i=1}^n |\boldsymbol{h}_{ii}| \lesssim d_c$. After showing the lindeberg interpolation step, the rest of the proof follows exactly as in Section 1.10.1.

1.12. Appendix: Additional Tables from Simulation Study

DGP				Testing Procedure						
n	d_z	ϱ_1	ϱ_2	$JK(\beta_0)$	$S(\beta_0)$	$T(\beta_0;\tau_{0.3})$	$T(\beta_0;\tau_{0.75})$	A.Rbn.	JAR	JLM
200	10	0.2	0.3	0.0516	0.0352	0.0406	0.0406	0.0296	0.0766	0.0502
		0.2	0.6	0.0542	0.0306	0.0442	0.0384	0.0258	0.0748	0.0400
		0.5	0.3	0.0470	0.0338	0.0416	0.0418	0.0238	0.0784	0.0460
		0.5	0.6	0.0506	0.0350	0.0416	0.0390	0.0280	0.0676	0.0384
	30	0.2	0.3	0.0570	0.0124	0.0422	0.0200	0.0088	0.1000	0.0382
		0.2	0.6	0.0564	0.0126	0.0408	0.0208	0.0124	0.0962	0.0322
		0.5	0.3	0.0498	0.0100	0.0366	0.0190	0.0096	0.1090	0.0318
		0.5	0.6	0.0562	0.0118	0.0420	0.0216	0.0088	0.1104	0.0292
	65	0.2	0.3	0.0542	0.0316	0.0428	0.0370	0.0314	0.0764	0.0420
		0.2	0.6	0.0532	0.0366	0.0418	0.0398	0.0250	0.0780	0.0376
		0.5	0.3	0.0474	0.0308	0.0388	0.0362	0.0244	0.0748	0.0354
		0.5	0.6	0.0484	0.0324	0.0366	0.0388	0.0282	0.0708	0.0402
	75	0.2	0.3	0.0512	0.0122	0.0364	0.0210	0.0150	0.0972	0.0422
		0.2	0.6	0.0564	0.0162	0.0416	0.0272	0.0152	0.0974	0.0414
		0.5	0.3	0.0488	0.0136	0.0368	0.0208	0.0168	0.1144	0.0380
		0.5	0.6	0.0516	0.0128	0.0390	0.0224	0.0122	0.1166	0.0390
500	10	0.2	0.3	0.0590	0.0468	0.0478	0.0516	0.0376	0.0652	0.0452
		0.2	0.6	0.0530	0.0420	0.0460	0.0466	0.0366	0.0692	0.0434
		0.5	0.3	0.0496	0.0370	0.0408	0.0368	0.0338	0.0710	0.0464
		0.5	0.6	0.0512	0.0426	0.0456	0.0438	0.0334	0.0696	0.0404
	30	0.2	0.3	0.0522	0.0202	0.0386	0.0278	0.0238	0.0818	0.0322
		0.2	0.6	0.0558	0.0208	0.0408	0.0310	0.0266	0.0888	0.0342
		0.5	0.3	0.0554	0.0178	0.0392	0.0280	0.0174	0.0940	0.0272
		0.5	0.6	0.0570	0.0156	0.0426	0.0236	0.0206	0.0984	0.0280
	65	0.2	0.3	0.0542	0.0372	0.0434	0.0432	0.0384	0.0754	0.0464
		0.2	0.6	0.0584	0.0442	0.0482	0.0470	0.0334	0.0676	0.0438
		0.5	0.3	0.0614	0.0460	0.0504	0.0496	0.0316	0.0708	0.0434
		0.5	0.6	0.0526	0.0378	0.0434	0.0420	0.0298	0.0692	0.0358
	75	0.2	0.3	0.0522	0.0234	0.0428	0.0316	0.0280	0.0818	0.0430
		0.2	0.6	0.0518	0.0252	0.0412	0.0318	0.0274	0.0916	0.0422
		0.5	0.3	0.0500	0.0240	0.0400	0.0316	0.0274	0.1028	0.0470
		0.5	0.6	0.0522	0.0220	0.0434	0.0328	0.0230	0.1002	0.0434

Table 1.12.1: Simulated Size of Identification and Heteroskedasticity Robust Tests under Weak Identification. Each DGP is simulated 5000 times. Critical values of the sup-score statistic and quantiles of the conditioning statistic are calculated using 1000 multiplier bootstrap simulations.

DGP				Testing Procedure						
n	d_z	ϱ_1	ϱ_2	$JK(\beta_0)$	$S(\beta_0)$	$T(\beta_0;\tau_{0.3})$	$T(\beta_0;\tau_{0.75})$	A.Rbn.	JAR	JLM
200	10	0.2	0.2	0.0474	0.0420	0.0474	0.0468	0.0308	0.0728	0.0424
		0.2	0.6	0.0512	0.0386	0.0512	0.0506	0.0304	0.0764	0.0378
		0.5	0.2	0.0416	0.0318	0.0414	0.0414	0.0248	0.0794	0.0428
		0.5	0.6	0.0446	0.0342	0.0446	0.0442	0.0244	0.0806	0.0384
	30	0.2	0.2	0.0482	0.0122	0.0448	0.0264	0.0110	0.1048	0.0370
		0.2	0.6	0.0498	0.0120	0.0480	0.0312	0.0118	0.0980	0.0378
		0.5	0.2	0.0456	0.0126	0.0410	0.0262	0.0082	0.1146	0.0268
		0.5	0.6	0.0482	0.0110	0.0474	0.0308	0.0094	0.1090	0.0302
	65	0.2	0.2	0.0528	0.0380	0.0526	0.0510	0.0276	0.0696	0.0460
		0.2	0.6	0.0464	0.0360	0.0464	0.0468	0.0302	0.0728	0.0416
		0.5	0.2	0.0482	0.0298	0.0480	0.0466	0.0246	0.0738	0.0412
		0.5	0.6	0.0396	0.0320	0.0390	0.0386	0.0258	0.0748	0.0356
	75	0.2	0.2	0.0516	0.0120	0.0498	0.0406	0.0188	0.1070	0.0414
		0.2	0.6	0.0444	0.0130	0.0436	0.0392	0.0198	0.1052	0.0408
		0.5	0.2	0.0416	0.0100	0.0408	0.0328	0.0128	0.1094	0.0412
		0.5	0.6	0.0480	0.0128	0.0474	0.0432	0.0122	0.1096	0.0430
500	10	0.2	0.2	0.0524	0.0444	0.0524	0.0524	0.0394	0.0684	0.0472
		0.2	0.6	0.0476	0.0430	0.0476	0.0476	0.0400	0.0644	0.0490
		0.5	0.2	0.0434	0.0410	0.0434	0.0434	0.0340	0.0702	0.0404
		0.5	0.6	0.0448	0.0382	0.0448	0.0448	0.0350	0.0736	0.0432
	30	0.2	0.2	0.0502	0.0214	0.0502	0.0498	0.0240	0.0854	0.0368
		0.2	0.6	0.0522	0.0208	0.0522	0.0524	0.0224	0.0858	0.0392
		0.5	0.2	0.0456	0.0202	0.0456	0.0434	0.0220	0.0918	0.0264
		0.5	0.6	0.0500	0.0186	0.0500	0.0498	0.0204	0.0924	0.0268
	65	0.2	0.2	0.0490	0.0426	0.0490	0.0490	0.0350	0.0742	0.0472
		0.2	0.6	0.0522	0.0458	0.0522	0.0522	0.0436	0.0652	0.0442
		0.5	0.2	0.0542	0.0476	0.0542	0.0542	0.0294	0.0712	0.0446
		0.5	0.6	0.0438	0.0420	0.0438	0.0438	0.0306	0.0666	0.0500
	75	0.2	0.2	0.0480	0.0220	0.0480	0.0480	0.0314	0.0880	0.0394
		0.2	0.6	0.0492	0.0284	0.0492	0.0492	0.0278	0.0874	0.0470
		0.5	0.2	0.0404	0.0190	0.0404	0.0404	0.0254	0.0992	0.0426
		0.5	0.6	0.0470	0.0226	0.0470	0.0468	0.0182	0.0960	0.0418

Table 1.12.2: Simulated Size of Identification and Heteroskedasticity Robust Tests under Strong Identification. Each DGP is simulated 5000 times. Critical values of the sup-score statistic and quantiles of the conditioning statistic are calculated using 1000 multiplier bootstrap simulations.

Chapter 2

Doubly-Robust Inference for Conditional Average Treatment Effects with High-Dimensional Controls

2.1. Introduction

Consider a potential outcomes framework (Rubin, 1974a, 1978a) where an observed outcome $Y \in \mathbb{R}$ and treatment $D \in \{0, 1\}$ are related to two latent potential outcomes $Y_1, Y_0 \in \mathbb{R}$ via $Y = DY_1 + (1 - D)Y_0$. To account for unobserved confounding factors a common strategy is to assume the researcher has access to a vector of covariates, $Z = (Z_1', X')' \in \mathcal{Z}_1 \times \mathcal{X} \subseteq \mathbb{R}^{d_z - d_x} \times \mathbb{R}^{d_x}$, such that the potential outcomes are independent of the treatment decision after conditioning on the observed covariates, $(Y_1, Y_0) \perp D|Z$. In this setting, we are interested in estimation of and inference on the conditional average treatment effect (CATE):

$$\mathbb{E}[Y_1 - Y_0 \mid X = x]. \tag{2.1.1}$$

Estimation of the CATE generally requires first fitting propensity score and/or outcome regression models. When the number of control variables Z is large ($d_z \gg n$), these first-stage models must be estimated using regularized methods which converge slower than the nonparametric rate and typically rely on the correctness of parametric specifications for

consistency.[1]

Fortunately, if both models are correctly specified, one can obtain a nonparametric-rate consistent estimator and valid inference procedure for the CATE by using the popular augmented inverse propensity weighted (aIPW) signal (Semenova and Chernozhukov, 2021; Fan et al., 2022). This is because the aIPW signal obeys an orthogonality condition at, crucially, the true nuisance model values that limits the first-stage estimation error passed on to the second-stage estimator. Moreover, estimators based on the aIPW signal are doubly-robust; consistency of the resulting second-stage estimators requires correct specification of only one of the first-stage propensity score or outcome regression models. However inference based on these estimators is not doubly-robust. The orthogonality of the aIPW signal fails under misspecification and the resulting testing procedures and confidence intervals are rendered invalid.

This paper proposes a doubly-robust estimator and inference procedure for the conditional average treatment effect when the number of control variables, d_z, is potentially much larger than the sample size, n. The dimensionality of the conditioning variable, d_x, remains fixed in our analysis. Our approach is based on Tan (2020) wherein doubly-robust inference is developed for the average treatment effect. We take a series approach to estimating the CATE, using a quasi-projection of the aIPW signal onto a growing set of basis functions. By assuming a logistic form for the propensity score model and a linear form for the outcome regression model, we construct novel ℓ_1-regularized first-stage estimating equations to recover a partial orthogonality of the aIPW signal at the limiting values of the first-stage estimators. So long as the limiting values of the first stage estimators have sparse representations this restricted orthogonality is enough to achieve doubly-robust pointwise and uniform inference; pointwise and uniform confidence intervals centered at the second-stage estimator are valid even if one of the logistic or linear functional forms is misspecified.

[1]Recent works by Bauer and Kohler (2019), Schmidt-Hieber (2020) provide some limited nonparametric results in high-dimensional settings using deep neural networks.

To achieve this restricted orthogonality at all points in the support of the conditioning variable, we employ distinct first-stage estimating equations for each basis term used in the second-stage series approximation. This results in the number of first-stage estimators growing with the number of basis terms. These estimators converge uniformly to limiting values under standard conditions in high-dimensional analysis. Improving on prior work in doubly-robust inference, our ℓ_1 regularized first-stage estimation incorporates a data-dependent penalty parameter based on the work of Chetverikov and Sørensen (2021). This allows practical implementation of our proposed estimation procedure with minimal knowledge of the underlying data generating process.

The use of multiple pairs of nuisance parameter estimates leaves us with multiple limiting values for the aIPW signals. So long as one of the nuisance models is correctly specified these limiting values share a conditional mean function. However, the various limiting values may all have different error terms describing their deviations from the conditional mean. This limits our ability to straightforwardly apply existing nonparametric results for series estimators (Newey, 1997; Belloni et al., 2015). Under modified conditions, we analyze the asymptotic properties of our second-stage series estimator to re-derive pointwise and uniform inference results. These modified conditions are in general slightly stronger than those of Belloni et al. (2015), though in certain special cases collapse exactly to the conditions of Belloni et al. (2015).

Prior Literature. Chernozhukov et al. (2018) analyze the general problem of estimating finite dimensional target parameters in the presence of potentially high-dimensional nuisance functions. Using score functions that are Neyman-orthogonal with respect to nuisance parameters they show that it is possible to obtain target parameter estimates that are $\sqrt{n}$-consistent and asymptotically normal so long as the nuisance parameters are consistent at rate $n^{-1/4}$, a condition satisfied by many machine learning-based estimators. Semenova and Chernozhukov (2021) take advantage of new results for series estimation in Belloni et al.

(2015) and consider series estimation of functional target parameters after high-dimensional nuisance estimation.[2] The inference results of these papers are highly dependent on the orthogonality of their second stage estimators to first stage estimation error, making it difficult to directly extend these analyses when the first stage estimators are not consistent and the orthogonality cannot be applied.

In the same setting as this paper, Tan (2020), Bradic et al. (2019) consider estimation of the average treatment effect. After assuming a logistic form for the propensity score and a linear form for the outcome regression, both papers propose ℓ_1-regularized first-stage estimators that allow for partial control of the derivative of the aIPW signal away from true nuisance values and thus allow for doubly-robust inference. Bradic et al. (2019) differs from Tan (2020) in their use of sample splitting, which allows them to achieve a "sparsity double robust" estimate of the ATE; so long as one nuisance model is sufficiently sparse the other may be more dense. Smucler et al. (2019) extends the analysis of Tan (2020) to consider doubly-robust inference for a larger class of finite dimensional target parameters with bilinear influence functions. Wu et al. (2021) provide doubly-robust inference procedures for covariate-specific treatment effects with discrete conditioning variables; their results depend on exact representation assumptions that are unlikely to hold with continuous covariates. Moreover, no uniform inference procedures are described.

These papers pioneered the approach that we will employ below, which is to directly use the first order conditions of the first stage estimators to control second stage estimation error. However, it is not a priori clear how to extend this approach to control the estimation error passed onto an infinite dimensional target parameter like the CATE. As discussed above, our analysis requires re-deriving pointwise and uniform inference results for nonparametric series estimators under modified conditions. We do not consider the sample splitting approach of Bradic et al. (2019), which may allow for relaxed sparsity conditions on our nuisance parameter estimates, but consider this an interesting future extension.

[2] Fan et al. (2022) provides a similar analysis using a second-stage kernel estimator.

Chetverikov and Sørensen (2021) propose a data-driven "bootstrap after cross-validation" approach to penalty parameter selection that is modified for and implemented in our setting. This work is related to other work on the lasso (Tibshirani, 1996; Bickel et al., 2009; Belloni and Chernozhukov, 2013; Chetverikov et al., 2021) and ℓ_1-regularized M-estimation in high-dimensional settings (van der Greer, 2016; Tan, 2017).

Paper Structure. This paper proceeds as follows. Section 2.2 defines the problem and introduces our methods for estimation and inference. Section 2.3 provides intuition for how the first-stage estimation procedure allows for doubly-robust estimation and inference on the CATE as well as formally establishes the necessary first-stage convergence. Section 2.4 presents the main results: valid pointwise and uniform inference for the second-stage series estimator if either the first-stage logistic propensity score model or linear outcome regression model is correctly specified. Section 2.5 ties up a technical detail. Section 2.6 applies our proposed estimator to examine the effect of maternal smoking on infant birth weight while Section 2.7 provides evidence from simulation study. Section 2.8 concludes. Proofs of main results are deferred to Section 2.9.

Notation. For any measure F and any function f, define the L^2 norm, $\|f\|_{F,2} = (\mathbb{E}_F[f^2])^{1/2}$ and the L^∞ norm $\|f\|_{F,\infty} = \operatorname{ess\,sup}_F |f|$. For any vector in $\mathbb{R}^p$ let $\|\cdot\|_p$ for $p \in [1,\infty]$ denote the ℓ_p norm, $\|a\|_p = (\sum_{l=1}^p a_l^p)^{1/p}$ and $\|a\|_\infty = \max_{1\le l\le\infty} |a_l|$. If the subscript is unspecified, we are using the ℓ_2 norm. For two vectors $a, b \in \mathbb{R}^p$, let $a \circ b = (a_i b_i)_{i=1}^p$ denote the Hadamard (element-wise) product. We adopt the convention that for $a \in \mathbb{R}^p$ and $c \in \mathbb{R}$, $a+c = (a_i+c)_{i=1}^p$. For a matrix $A \in \mathbb{R}^{m\times n}$ let $\|A\| = \max_{\|v\|_{\ell_2}\le 1} \|Av\|_{\ell_2}$ denote the operator norm and $\|A\|_\infty = \sup_{1\le r\le m, 1\le s\le n} |A_{rs}|$. For any real valued function f let $\mathbb{E}_n[f(X)] = \frac{1}{n}\sum_{i=1}^n f(X_i)$ denote the empirical expectation and $\mathbb{G}_n[f(X)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}[X_i])$ denote the empirical process. For two sequences of random variables $\{a_n\}_\mathbb{N}$ and $\{b_n\}_\mathbb{N}$, we say $a_n \lesssim_P b_n$ or $a_n = O_p(b_n)$ if a_n/b_n is bounded in probability and say $a_n = o_p(b_n)$ if $a_n/b_n \to_p 0$.

2.2. Setup

In this section, we formally define the setting and identification strategy that we consider. We then introduce our doubly-robust estimator and inference procedure. The parameter of interest is the conditional average treatment effect: $\mathbb{E}[Y_1 - Y_0 \mid X = x]$. However, for this paper we largely focus on estimation and inference for the conditional average counterfactual outcome:

$$g_0(x) \coloneqq \mathbb{E}[Y_1 \mid X = x]. \tag{2.2.1}$$

Doubly-robust estimation and inference on $\mathbb{E}[Y_0 \mid X = x]$ follows a similar procedure and is described in Section 2.5. The procedures can be combined for doubly-robust estimation and inference for the CATE.

2.2.1. Setting

We assume the researcher observes i.i.d data and conditioning on Z is sufficient to control for all confounding factors affecting both the treatment decision D and the potential outcomes, Y_1 and Y_0. Our analysis allows the dimensionality of the controls, $Z = (Z_1, X)$, to grow much faster than the sample size ($d_z \gg n$), while assuming the dimensionality of the conditioning variables, X, remains fixed ($d_x \ll n$).

Assumption 2.2.1 (Identification)**.**

(i) $\{Y_i, D_i, Z_i\}_{i=1}^n$ *are independent and identically distributed.*

(ii) $(Y_1, Y_0) \perp D \mid Z$.

(iii) There exists a value $\eta \in (0, 1)$ *such that* $\eta < \mathbb{E}[D \mid Z = z] < 1 - \eta$ *almost surely in* Z.

To obtain doubly-robust estimation and inference we use the augmented inverse propensity

weighted (aIPW) signal,

$$Y(\pi, m) = \frac{DY}{\pi(Z)} - \left(\frac{D}{\pi(Z)} - 1\right) m(Z), \tag{2.2.2}$$

which is a function of a fitted propensity score model, $\pi(Z)$, and a fitted outcome regression model, $m(Z)$, whose true values are given $\pi^\star(Z) \coloneqq \mathbb{E}[D \mid Z]$ and $m^\star(Z) \coloneqq \mathbb{E}[Y \mid D = 1, Z]$. Under Assumption 2.2.1, the aIPW signal $Y(\cdot, \cdot)$ provides doubly-robust identification of $g_0(x)$. That is, for integrable $\pi \neq \pi^\star$ and $m \neq m^\star$,

$$\begin{aligned} \mathbb{E}[Y_1 \mid X = x] &= \mathbb{E}[Y(\pi^\star, m^\star) \mid X = x] \\ &= \mathbb{E}[Y(\pi\,, m^\star) \mid X = x] \\ &= \mathbb{E}[Y(\pi^\star, m\,) \mid X = x]. \end{aligned} \tag{2.2.3}$$

We use a series approach to estimate $g_0(x)$, taking a quasi-projection of the aIPW signal onto a growing set of k weakly positive basis terms:

$$p^k(x) \coloneqq \big(p_1(x), \ldots, p_k(x)\big)' \in \mathbb{R}^k_+. \tag{2.2.4}$$

The basis terms are required to be weakly positive as they are used as weights within the convex first-stage estimators estimating equations.[1] Examples of weakly positive basis functions are B-splines or shifted polynomial series terms. To ensure that the basis terms are well behaved, we assume regularity conditions on $\xi_{k,\infty} \coloneqq \sup_{x \in \mathcal{X}} \|p^k(x)\|_\infty$, $\xi_{k,2} \coloneqq \sup_{x \in \mathcal{X}} \|p^k(x)\|_2$, and the eigenvalues of the design matrix $Q \coloneqq \mathbb{E}[p^k(x)p^k(x)']$.

For each basis term $p_j(x), j = 1, \ldots, k$, we estimate a separate propensity score model, $\widehat{\pi}_j(Z)$, and outcome regression model, $\widehat{m}_j(Z)$. Under standard moment and sparsity conditions, these converge uniformly over $j = 1, \ldots, k$ to limiting values $\bar{\pi}_j(Z)$ and $\bar{m}_j(Z)$. If the propensity score model and outcome regression models are correctly specified these limiting values coincide with the true values $\pi^\star(Z)$ and $m^\star(Z)$. However, in general the limiting and true

values may differ. The double robustness of the aIPW signal allows for identification of the CATE even if only one of the nuisance models is correctly specified. If either $\bar{\pi}_j = \pi^\star$ or $\bar{m}_j = m^\star$, we can write for all $j = 1, \ldots, k$:

$$\begin{aligned} Y(\bar{\pi}_j, \bar{m}_j) &= g_0(x) + \epsilon_j, \qquad\qquad \mathbb{E}[\epsilon_j \mid X] = 0 \\ &= g_k(x) + r_k(x) + \epsilon_j \end{aligned} \tag{2.2.5}$$

where $g_0(x)$ is the conditional counterfactual outcome (2.2.1), $g_k(x) \coloneqq p^k(x)'\beta^k$ is the projection of $g_0(x)$ onto the first k basis terms, and $r_k(x) \coloneqq g_0(x) - g_k(x)$ denotes the approximation error from this projection. Note the separate error terms for each $j = 1, \ldots, k$ in (2.2.5), which are collected together in the vector $\epsilon^k \coloneqq (\epsilon_1, \ldots, \epsilon_k)$. As long as one of the first-stage models is correctly specified, the least squares parameter β^k governing the projection in $g_k(x)$ can be identified by the projection of the aIPW signal onto the basis terms $p^k(x)$:

$$\begin{aligned} \beta^k &\coloneqq Q^{-1}\mathbb{E}[p^k(X)Y_1] \\ &= Q^{-1}\mathbb{E}[p^k(X)Y(\pi^\star, m^\star)] \\ &= Q^{-1}\mathbb{E}[p^k(X)Y(\bar{\pi}_j, \bar{m}_j)], \quad \forall j = 1, \ldots, k. \end{aligned} \tag{2.2.6}$$

[1] In case the researcher wants to use a second-stage basis that cannot be transformed to be weakly positive, we have shown a slightly modified method of constructing our doubly-robust estimator and inference procedure that does not require the first-stage weights to directly be the second-stage basis terms. This is available on request.

2.2.2. Estimator and Inference Procedure

We assume a logistic regression form for the propensity score model and a linear form for the outcome regression model:

$$\begin{aligned} \pi(Z;\gamma) &= \left(1+\exp(-\gamma' Z)\right)^{-1}, \\ m(Z;\alpha) &= \alpha' Z. \end{aligned} \tag{2.2.7}$$

For each $j = 1, \ldots, k$, the parameters of (2.2.7), $\gamma, \alpha \in \mathbb{R}^{d_z}$, are estimated, respectively, by

$$\widehat{\gamma}_j \coloneqq \arg\min_{\gamma} \mathbb{E}_n[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}] + \lambda_{\gamma,j}\|\gamma\|_1, \tag{2.2.8}$$

$$\widehat{\alpha}_j \coloneqq \arg\min_{\alpha} \mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j' Z}(Y-\alpha' Z)^2]/2 + \lambda_{\alpha,j}\|\alpha\|_1. \tag{2.2.9}$$

The penalty parameters $\lambda_{\gamma,j}$ and $\alpha_{\gamma,j}$ are chosen via a data dependent technique described below. Under standard assumptions the parameter estimators $\widehat{\gamma}_j, \widehat{\alpha}_j$ will converge uniformly over $j = 1, \ldots, k$ to population minimizers

$$\bar{\gamma}_j \coloneqq \arg\min_{\gamma} \mathbb{E}[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}], \tag{2.2.10}$$

$$\bar{\alpha}_j \coloneqq \arg\min_{\alpha} \mathbb{E}[p_j(Z)De^{-\bar{\gamma}_j' Z}(Y-\alpha' Z)^2]. \tag{2.2.11}$$

which we assume are sufficiently sparse. Our first-stage estimators are then $\widehat{\pi}_j(Z) \coloneqq \pi(Z;\widehat{\gamma}_j)$ and $\widehat{m}_j(Z) \coloneqq m(Z;\widehat{\alpha}_j)$ with limiting values $\bar{\pi}_j(Z) \coloneqq \pi(Z;\bar{\gamma}_j)$ and $\bar{m}_j(Z) \coloneqq m(Z;\bar{\alpha}_j)$, respectively.

After plugging in the functional forms of $\bar{\pi}_j(Z)$ and $\bar{m}_j(Z)$ into the aIPW signal one can verify that the derivatives of the aIPW signal with respect to the parameters γ_j and α_j are almost identical to the first order conditions of the minimization problems in (2.2.10)-(2.2.11) . Optimality of $\bar{\gamma}_j$ and $\bar{\alpha}_j$ will thus imply that the gradient of the limiting aIPW signal, weighted by $p_j(X)$, is mean zero even when the limiting values $\bar{\pi}_j(Z)$ and $\bar{m}_j(Z)$ differ from

the true values $\pi^\star(Z)$ and $m^\star(Z)$. This allows us to control how sensitive the second stage CATE estimator is to first stage nuisance model estimation error even under misspecification and achieve doubly robust inference. The importance of this fact and why it is useful is discussed at greater depth in Section 2.3.

Our second-stage estimator is defined $\widehat{g}(x) := p^k(x)'\widehat{\beta}^k$ where $\widehat{\beta}^k$ is an estimate of the population projection parameter, β^k, obtained by combining all k pairs of first-stage estimators

$$\widehat{\beta}^k := \widehat{Q}^{-1}\mathbb{E}_n \begin{bmatrix} p_1(X)Y(\widehat{\pi}_1, \widehat{m}_1) \\ \vdots \\ p_k(X)Y(\widehat{\pi}_k, \widehat{m}_k) \end{bmatrix}, \tag{2.2.12}$$

and $\widehat{Q} := \mathbb{E}_n[p^k(X)p^k(X)']$. We estimate the variance of $\widehat{g}(x)$ using $\widehat{\sigma}(x) := \|\widehat{\Omega}^{1/2}p^k(x)\|/\sqrt{n}$ where

$$\widehat{\Omega} := \widehat{Q}^{-1}\mathbb{E}_n[\{p^k(X) \circ \widehat{\epsilon}^k\}\{p^k(X) \circ \widehat{\epsilon}^k\}']\widehat{Q}^{-1}, \tag{2.2.13}$$

and $\circ$ represents the Hadamard element-wise product. The vector $\widehat{\epsilon}^k$ collects the various estimated error terms; $\widehat{\epsilon}^k := (\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_k)$ for $\widehat{\epsilon}_j := Y(\widehat{\pi}_j, \widehat{m}_j) - \widehat{g}(x)$, $j = 1, ..., k$. Inference is based on the $100(1-\eta)\%$ confidence bands

$$\left[\underline{i}(x), \bar{i}(x)\right] := \left[\widehat{g}(x) - c^\star\left(1-\eta/2\right)\widehat{\sigma}(x),\ \widehat{g}(x) + c^\star\left(1-\eta/2\right)\widehat{\sigma}(x)\right]. \tag{2.2.14}$$

For pointwise inference, the critical value $c^\star(1-\eta/2)$ is taken as the $(1-\eta/2)$ quantile of a standard normal distribution. For uniform inference $c^\star(1-\eta/2)$ is taken

$$c_u^\star(1-\eta/2) := (1-\eta/2)\text{-quantile of } \sup_{x\in\mathcal{X}}\left|\frac{p^k(x)\widehat{\Omega}^{1/2}}{\widehat{\sigma}(x)}N_k^b\right|$$

where N_k^b is a bootstrap draw from $N(0, I_k)$. Sections 2.3 and 2.4 show that, under standard sparsity and moment conditions, these pointwise and uniform inference procedures remain valid even under misspecification of either first-stage model.

2.2.3. Penalty Parameter Selection

To select the penalty parameters $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$ in (2.2.8)-(2.2.9) we propose a data driven two-step procedure based on the work of Chetverikov and Sørensen (2021). For each $j = 0, 1 \ldots, k$, we start with pilot penalty parameters given by

$$\lambda_{\gamma,j}^{\text{pilot}} = c_{\gamma,j} \times \sqrt{\frac{\ln^3(d_z)}{n}} \quad \text{and} \quad \lambda_{\alpha,j}^{\text{pilot}} = c_{\alpha,j} \times \sqrt{\frac{\ln^3(d_z)}{n}} \tag{2.2.15}$$

for some constants $c_{\gamma,j}, c_{\alpha,j}$ selected from the interval $[\underline{c}_n, \bar{c}_n]$ with $\underline{c}_n > 0$. In practice, the researcher has a fair bit of flexibility in choosing these constants. The optimal choice of these constants may depend on the underlying data generating process. We recommend using cross validation to pick these constants from a fixed-cardinality set of possible values. In line with Assumption 2.3.1(vi), the values in the set should be chosen to be on the order of the maximum value of $\|p^k(X_i)\|_\infty$ observed in the data.

Using $\lambda_{\gamma,j}^{\text{pilot}}$ and $\lambda_{\alpha,j}^{\text{pilot}}$ in lieu of $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$ in (2.2.8)-(2.2.9) we generate pilot estimators $\widehat{\gamma}_j^{\text{pilot}}$ and $\widehat{\alpha}_j^{\text{pilot}}$. These pilot estimators are used to generate plug in estimators $\widehat{U}_{\gamma,j}$ and $\widehat{U}_{\alpha,j}$ of the residuals

$$\begin{aligned} \widehat{U}_{\gamma,j} &:= -p_j(X)\{D(1 + e^{-\widehat{\gamma}_j^{\text{pilot}\prime} Z}) - 1\} \\ \widehat{U}_{\alpha,j} &:= -p_j(X) D e^{-\widehat{\gamma}_j^{\text{pilot}\prime} Z}(Y - \widehat{\alpha}_j^{\text{pilot}\prime} Z). \end{aligned} \tag{2.2.16}$$

whose true values are given

$$\begin{aligned} U_{\gamma,j} &:= -p_j(X)\{D(1 + e^{-\bar{\gamma}_j' Z}) - 1\} \\ U_{\alpha,j} &:= -p_j(X) D e^{-\bar{\gamma}' Z}(Y - \bar{\alpha}' Z) \end{aligned} \tag{2.2.17}$$

These true residuals are the derivatives of the minimization problems in (2.2.10)-(2.2.11) evaluated at minimizing values $\bar{\gamma}_j$ and $\bar{\alpha}_j$. After generating the residual estimates, we use a

multiplier bootstrap procedure to select final penalty parameters $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$.

$$\begin{aligned}\lambda_{\gamma,j} &= c_0 \times (1-\epsilon)\text{-quantile of } \max_{1\leq l\leq d_z} |\mathbb{E}_n[e_i \widehat{U}_{\gamma,j} Z_l]| \text{ given } \{Y_i, D_i, Z_i\}_{i=1}^n, \\ \lambda_{\alpha,j} &= c_0 \times (1-\epsilon)\text{-quantile of } \max_{1\leq l\leq d_z} |\mathbb{E}_n[e_i \widehat{U}_{\alpha,j} Z_l]| \text{ given } \{Y_i, D_i, Z_i\}_{i=1}^n\end{aligned} \tag{2.2.18}$$

where $e_1, \ldots, e_n$ are independent standard normal random variables generated independently of the data $\{Y_i, D_i, X_i\}_{i=1}^n$ and $c_0 > 1$ is a fixed constant.[2] In line with other work we find $c_0 = 1.1$ works well in simulations. So long as our residual estimates converge in empirical mean square to limiting values and $k\epsilon \to 0$, the choice of penalty parameters in (2.2.18) will ensure that the penalty parameters dominate the noise with probability approaching one uniformly over the k first stage estimation procedures. This allows for consistent variable selection and coefficient estimation.

2.3. Theory Overview

We begin with a main technical lemma which provides a bound on rate at which first-stage estimation error is passed on to the second-stage CATE and variance estimators. This bound is comparable to others seen in the inference after model-selection literature (Belloni et al., 2013; Tan, 2020) and is achieved under standard conditions in the ℓ_1-regularized estimation literature (Bickel et al., 2009; Bühlmann and van de Geer, 2011; Belloni and Chernozhukov, 2013; Chetverikov and Sørensen, 2021). However, this bound is achieved at the limiting values of the propensity score and outcome regression models which may differ from the true values $\pi^\star$ and $m^\star$ under misspecification.

The potential misspecification of the first-stage models means we cannot directly apply orthogonality of the aIPW signal, discussed below, to show that the effect of first-stage estimation error on the second-stage is negligible. Instead, we use the first order conditions for

[2]The constant c_0 can be different for the propensity score and outcome regression models and can also vary for each $j = 1, \ldots, k$. All that matters is that each constant satisfies the requirements of Lemma 2.3.1. This complicates notation, however.

$\widehat{\gamma}_j$ and $\widehat{\alpha}_j$ to directly control this quantity. After presenting the lemma Section 2.3.2 provides some intuition for how this is done. Controlling the rate at which first-stage estimation error is passed on to the second-stage estimator even at points away from the true values $\pi^\star$ and $m^\star$ is key for obtaining doubly-robust inference for the CATE.

2.3.1. Uniform First-Stage Convergence

To show uniform convergence of the first-stage estimators and thus uniform control of the bias passed on from the first-stage estimation to the second-stage estimator we rely on Assumption 2.3.1, below. The conditions in Assumption 2.3.1(v,vi) depend on the sup-norm of the basis functions, $\xi_{k,\infty} = \sup_{x\in\mathcal{X}} \|p^k(x)\|_\infty$.

Assumption 2.3.1 (First-Stage Convergence)**.**

(i) The regressors Z are bounded, $\max_{1\leq l\leq d_z} |Z_l| \leq C_0$ almost surely.

(ii) The errors $Y_1 - \bar{m}_j(Z)$ are uniformly subgaussian conditional on Z in the following sense. There exists fixed positive constants G_0 and G_1 such that for any j:

$$G_0\mathbb{E}\left[\exp\left(\{Y_1 - \bar{m}_j(Z)\}^2/G_0^2\right) - 1 \mid Z\right] \leq G_1^2$$

almost surely.

(iii) There is a constant B_0 such that $\bar{\gamma}_j' Z \geq B_0$ almost surely for all j.

(iv) There exists fixed constants $\xi_0 > 1$ and $1 > \nu_0 > 0$ such that for each $j = 1, \ldots, k$ the following empirical compatibility condition holds for the empirical hessian matrix $\tilde{\Sigma}_{\gamma,j} := \mathbb{E}_n[De^{-\bar{\gamma}_j' Z} ZZ']$. For any $b \in \mathbb{R}^{d_z}$ and $\mathcal{S}_j = \{l : |\bar{\gamma}_{j,l}| \vee |\bar{\alpha}_{j,l}| \neq 0\}$:

$$\sum_{l\notin\mathcal{S}_j} |b_l| \leq \xi_0 \sum_{l\in\mathcal{S}_j} |b_l| \implies \nu_0^2 \Big(\sum_{l\in\mathcal{S}_j} |b_l|\Big)^2 \leq |\mathcal{S}_j| \left(b'\tilde{\Sigma}_{\gamma,j} b\right).$$

(v) There exists fixed constants c_u and $C_U > 0$ such that for all $j = 1, \ldots, k$, $\mathbb{E}[U_{\gamma,j}^4] \leq$

$(\xi_{k,\infty} C_U)^4$ *and* $\min_{1 \leq l \leq d_z} \mathbb{E}[U_{\gamma,j}^2 Z_l^2] \geq c_u$.

(vi) The constant $\underline{c}_n$ *is chosen such that* $\xi_{k.\infty} \lesssim \underline{c}_n$ *and the following sparsity bounds hold for* $s_k = \max_{1 \leq j \leq k} |\mathcal{S}_j|$

$$\frac{\xi_{k,\infty} s_k^2 \bar{c}_n^2 \ln^5(d_z n)}{n} \to 0, \quad \textit{and} \quad \frac{\xi_{k,\infty}^4 \ln^7(d_z k n)}{n} \to 0.$$

The first part of Assumption 2.3.1 assumes that the regressors are bounded while the second assumes that tail behavior of the outcome regression errors are uniformly thin. Both of these can be relaxed somewhat with sufficient moment conditions on the tail behavior of the controls and errors. We should note that compactness of $\mathcal{X}$ is generally required by nonparametric estimators. The third part of the assumption bounds all limiting propensity scores $\bar{\pi}_j(Z)$ away from zero uniformly. The fourth assumption is an empirical compatibility condition on the weighted first-stage design matrix. It is slightly weaker than the restricted eigenvalue conditions often assumed in the literature (Bickel et al., 2009; Belloni et al., 2012b). The penultimate condition is an identifiability constraint that limits the moments of the noise and bounds it away from zero uniformly over all estimation procedures. Many of the constants in Assumption 2.3.1 are assumed to be fixed across all j. This is mainly to simplify the exposition of the results below and in practice all constants can be allowed to grow slowly with k. However, the growth rate of these terms affects the required first-stage sparsity.

The last condition is required for the validity of the bootstrap penalty parameter selection procedure and is comparable to the requirements needed for the bootstrap after cross validation technique described by Chetverikov and Sørensen (2021). The main difference is the additional assumption on the growth rate of the basis functions, $\xi_{k,\infty}$ which is to ensure uniform stability of the estimation procedures (2.2.8)-(2.2.9) as well as some assumptions on the order of the constants $c_{\gamma,j}$ and $c_{\alpha,j}$ in (2.2.15).

Assumptions 2.3.1(v,vi) depend on the sup-norm of the basis functions, $\xi_{k,\infty}$. This growth

rate of this quantity will depend on the form of basis used for the second stage nonparametric estimator. In both our simulation study as well as our empirical exercise we use B-splines for which $\xi_{k,\infty} \lesssim \sqrt{k}$. Other common bases used in nonparametric estimation are polynomial series for which $\xi_k \lesssim k$, or wavelets for which $\xi_{k,\infty} \lesssim \sqrt{k}$. Belloni et al. (2015) provide a discussion for other choices of basis terms.

Lemma 2.3.1 (First-Stage Convergence)**.** *Suppose that Assumption 2.3.1 holds. In addition assume that* $c_0 > (\xi_0 + 1)/(\xi_0 - 1)$, $k/n \to 0$, $k\epsilon \to 0$, *and there is a fixed constant* $c > 0$ *such that for all* j, $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$.[1] *Then the following weighted means converge uniformly in absolute value at least at rate:*

$$\max_{1 \leq j \leq k} \left| \mathbb{E}_n[p_j(X) Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X) Y(\bar{\pi}_j, \bar{m}_j)] \right| \lesssim_P \frac{s_k \, \xi_{k,\infty}^2 \ln(d_z)}{n} \tag{2.3.1}$$

and in empirical mean square at least at rate:

$$\max_{1 \leq j \leq k} \mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] \lesssim_P \frac{s_k^2 \, \xi_{k,\infty}^4 \ln(d_z)}{n} \tag{2.3.2}$$

Lemma 2.3.1 provides a tight bound on the first-stage estimation error passed on to the second-stage estimator even when the first-stage estimators converge to values that are not the true propensity score or outcome regression. In particular under the sparsity bound $s_k \xi_{k,\infty}^2 k^{1/2} \ln^2(d_z)/\sqrt{n} \to 0$, any linear combination of the means in both (2.3.1) and (2.3.2) is $o_p(\sqrt{n})$. This allows us to obtain doubly-robust inference for the CATE. This sparsity bound is similar in form to others in the literature (Belloni et al., 2012b; van der Greer, 2016; Chetverikov and Sørensen, 2021) however is somewhat stronger due to the additional dependence on $\xi_{k,\infty}^2 k^{1/2}$.

[1] The requirement $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ may seem a bit unnatural, but it can be enforced in practice without upsetting any assumptions by using the alternative linear penalty $\lambda_{\alpha,j}^{\text{ratio}} := \max\{\lambda_{\gamma,j} \geq c, \lambda_{\alpha,j}\}$. In simulations, we find this constraint is rarely binding. The constant c here is arbitrary, it is only important that the ratio $\lambda_{\gamma,j}/\lambda_{\alpha,j}$ is bounded from above.

2.3.2. Managing First-Stage Bias

We now provide some intuition for how this result is obtained and the role our particular estimating equations play in establishing this fact. We focus on control of the vector $\mathbf{B}^k$, defined in (2.3.3), which measures the bias passed on from first-stage estimation to the second-stage estimate $\widehat{\beta}^k$. Limiting the size of $\mathbf{B}^k$ is crucial in showing convergence of $\widehat{\beta}^k$ to the true parameter β^k and thus consistency of the nonparametric estimator $\widehat{g}(x)$.

$$\mathbf{B}^k := \mathbb{E}_n \begin{bmatrix} p_1(X) \left\{ Y(\widehat{\pi}_1, \widehat{m}_1) - Y(\bar{\pi}_1, \bar{m}_1) \right\} \\ \vdots \\ p_k(X) \left\{ Y(\widehat{\pi}_k, \widehat{m}_k) - Y(\bar{\pi}_k, \bar{m}_k) \right\} \end{bmatrix}. \tag{2.3.3}$$

For exposition, we consider a single term of (2.3.3), $\mathbf{B}_j^k$, which roughly measures the first-stage estimation bias taken on from adding the j^{th} basis term to our series approximation of $g_0(x)$. The discussion that follows is a bit informal, instead of considering the derivatives with respect to the true parameters below our proof strategy will directly use the Kuhn-Tucker conditions of the optimization routines in (2.2.8)-(2.2.9). However, the general intuition is the same as is used in the proofs.

In addition to the doubly-robust identification property (2.2.3), the aIPW signal is typically useful in the high-dimensional setting because it obeys an orthogonality condition at the true values $(\pi^\star, m^\star)$:[2]

$$\mathbb{E}[\nabla_{\pi,m} Y(\pi^\star, m^\star) \mid Z] = 0. \tag{2.3.4}$$

When both the propensity score model and outcome regression model are correctly specified we can (loosely speaking) examine the bias $\mathbf{B}_j^k$ by replacing $\bar{\pi}_j = \pi^\star$ and $\bar{m}_j = m^*$ and

[2]Robustness and orthogonality are indeed closely related, see Theorem 6.2 in Newey and McFadden (1994) for a discussion.

considering the following first order expansion:

$$\begin{aligned}\mathbf{B}_j^k &= \mathbb{E}_n[p_j(X)Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\pi^\star, m^\star)] \\ &= \underbrace{\mathbb{E}_n[p_j(X)\nabla_{\pi,m} Y(\pi^\star, m^\star)]}_{O_p(n^{-1/2}) \text{ by (2.3.4)}} \begin{bmatrix} \widehat{\pi}_j - \pi^\star \\ \widehat{m}_j - m^\star \end{bmatrix} + o_p(n^{-1/2}).\end{aligned} \tag{2.3.5}$$

By orthogonality of the aIPW signal the gradient term is close to zero, which guarantees that the bias is asymptotically negligible even if the nuisance parameters converge slowly to the true values, $\pi^\star$ and $m^\star$.[3] This allows the researcher to ignore first-stage nuisance parameter estimation error and treat $\pi^\star$ and $m^\star$ as known when analyzing the asymptotic properties of the second-stage series estimator. Indeed, since the aIPW signal orthogonality holds conditional on $Z = (Z_1, X)$, if both models are correctly specified only a single pair of first-stage estimators would be needed to provide control over all the elements in $\mathbf{B}^k$. This is the approach followed by Semenova and Chernozhukov (2021).

So long as either one of $\bar{\pi}_j = \pi^\star$ or $\bar{m}_j = m^\star$ we still have that $\mathbb{E}[p_j(X)Y_1] \approx \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)]$ by double-robustness of the aIPW signal (2.2.3). However, the aIPW orthogonality tells us nothing about the expectation of the gradient away from the true parameters, $\pi^\star, m^\star$; if either $\bar{\pi}_j \neq \pi^\star$ or $\bar{m}_j \neq m^\star$ there is no reason to believe that the gradient on the right hand side of (2.3.5) is mean zero when evaluated instead at $Y(\bar{\pi}_j, \bar{m}_j)$. In general, the bias $\mathbf{B}_j^k$ will then diminish at the rate of convergence of our nuisance parameters. Because we have high dimensional controls, this convergence rate will generally be much slower than the standard nonparametric rate (Newey, 1997; Belloni et al., 2015).

To get around this, we design the first-stage objective functions (2.2.8)-(2.2.9) such that the resulting first-order conditions control the bias passed on to the second-stage. Consider the

[3] Typically all that is required is that $\|\hat{\pi}_j - \pi^\star\| = o_p(n^{-1/4})$ and $\|\hat{m}_j - m^\star\| = o_p(n^{-1/4})$ in order to make the second order remainder term $\sqrt{n}$-negligible

following expansion instead around the limiting parameters $\bar{\gamma}_j$ and $\bar{\alpha}_k$.

$$\begin{aligned}\mathbf{B}_j^k &= \mathbb{E}_n[p_j(X)Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)] \\ &= \mathbb{E}_n[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] \begin{bmatrix} \widehat{\gamma}_j - \bar{\gamma}_j \\ \widehat{\alpha}_j - \bar{\alpha}_j \end{bmatrix} + o_p(n^{-1/2})\end{aligned} \tag{2.3.6}$$

After substituting the forms of $\bar{\pi}_j(z) = \pi(z; \bar{\gamma}_j)$ and $\bar{m}_j(z) = m(z; \bar{\alpha}_j)$ described in (2.2.7) and differentiating with respect to γ_j and α_j we obtain

$$\mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] = \mathbb{E}\begin{bmatrix} -p_j(X)De^{-\bar{\gamma}_j'Z}(Y - \bar{\alpha}_j'Z)Z \\ -p_j(x)\{D(1 + e^{-\bar{\gamma}_j'Z})Z - Z\} \end{bmatrix} \tag{2.3.7}$$

However, by definition $\bar{\gamma}_j$ and $\bar{\alpha}_j$ solve the minimization problems defined in (2.2.10)-(2.2.11), the population analogs of our finite sample estimating equations. The first order conditions of these minimization problems yield

$$\mathbb{E}\overbrace{\underbrace{\begin{bmatrix} -p_j(X)\{D(1 + e^{\bar{\gamma}'Z})Z - Z\} \\ -p_j(X)De^{-\bar{\gamma}'Z}(DY - \bar{\alpha}'Z)Z \end{bmatrix}}_{\text{First order condition of } \bar{\alpha}_j}}^{\text{First order condition of } \bar{\gamma}_j} = 0 \implies \mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] = 0 \tag{2.3.8}$$

Examining the first order conditions in (2.3.8), we see that they exactly give us control over the gradient (2.3.7). Under suitable convergence of the first-stage parameter estimates, this guarantees the bias examined in expansion (2.3.6) is negligible even under misspecification of the propensity score or outcome regression models.

Control of this gradient under misspecification is not provided using other estimating equations, such as maximum likelihood for the logistic propensity score model or ordinary least squares for the linear outcome regression model. Moreover, control over the gradient of $\mathbf{B}_j^k$ from

(2.3.3) is not provided by the first-order conditions for $\bar{\gamma}_l$ and $\bar{\alpha}_l$ for $l \neq j$:

$$\begin{aligned}\mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] &= \mathbb{E}\overbrace{\begin{bmatrix} -p_j(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)Z \\ -p_j(X)\{D(1+e^{\bar{\gamma}'Z})Z - Z\}\end{bmatrix}}^{\text{First order condition of } \bar{\gamma}_l} \\ &\neq \mathbb{E}\underbrace{\begin{bmatrix} -p_l(X)\{D(1+e^{\bar{\gamma}'Z})Z - Z\} \\ -p_l(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)Z\end{bmatrix}}_{\text{First order condition of } \bar{\alpha}_l}.\end{aligned} \tag{2.3.9}$$

Showing that the inference procedure of Section 2.2 remains valid at all points $x \in \mathcal{X}$ under misspecification requires showing negligible first-stage estimation bias for any linear transformation of the vector (2.3.3). As outlined above, this requires using k separate pairs of nuisance parameter estimator to obtain k separate pairs of first order conditions, one for each term of the vector.

2.4. Main Results

In this section, we present the main consistency and distributional results for our second-stage estimator $\widehat{g}(x)$ described in Section 2.2. A full set of second-stage results, including pointwise and uniform linearization lemmas and uniform convergence rates, can be found in the Online Appendix. The first set of results is established under the following condition, which limits the bias passed from first-stage estimation onto the second-stage estimator. In particular, Condition 1 implies that the bias vector $\mathbf{B}^k$ from (2.3.3) satisfies $\|\mathbf{B}^k\| = o_p(n^{-1/2})$.

Condition 1 (No Effect of First-Stage Bias)**.**

$$\max_{1\leq j\leq k} \left|\mathbb{E}_n[p_j(X)Y(\widehat{\pi}_j, \widehat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)]\right| = o_p(n^{-1/2}k^{-1/2}). \tag{2.4.1}$$

Via Lemma 2.3.1 we can see that is a logistic propensity score model and a linear outcome

regression model and estimating the first-stage models using the estimating equations (2.2.8)-(2.2.9), Condition 1 can be achieved under Assumption 2.3.1 and the sparsity bound

$$\frac{s_k \, \xi_{k,\infty}^2 k^{1/2} \ln(d_z)}{\sqrt{n}} \to 0. \tag{2.4.2}$$

If the researcher were to assume different parametric forms for the first-stage model, different first estimating equations would have to be used to obtain doubly-robust estimation and inference. However, so long as the Condition 1 can be established at the limiting values of the first-stage models, the results of this section hold.

Having dealt with the first-stage estimation error, the main complication remaining is that under misspecification the aIPW signals $Y(\hat{\pi}_j, \hat{m}_j)$ for $j = 1, \ldots, k$ do not all converge to the same limiting values. However, so long as at least one of the first-stage models is correctly specified, all of the limiting aIPW signals have the same conditional mean, $g_0(x)$. In the standard setting, consistency of nonparametric estimator relies on certain conditions on the error terms. In our setting, we require that these assumptions hold uniformly over k the error terms. We note though that there is a non-trivial dependence structure between that limiting aIPW signals. This strong dependence gives plausibility to our uniform conditions. For example, if the logistic propensity score model is correctly specified and the difference between the limiting outcome regression models is bounded, $|\max_{1 \leq j \leq k} \bar{m}_j(Z) - \min_{1 \leq j \leq k} m_j(Z)| \leq C$ almost surely, our conditions reduce exactly to the conditions of Belloni et al. (2015). In general, however, the uniform conditions suggest that a degree of undersmoothing is optimal when implementing our estimation procedure; the optimal choice of k may be smaller than in standard nonparametric regression.

2.4.1. Pointwise Inference

Pointwise inference relies on the following assumption in tandem with Condition 1.

Assumption 2.4.1 (Second-Stage Pointwise Assumption)**.** *Let* $\bar{\epsilon}_k \coloneqq \max_{1 \leq j \leq k} |\epsilon_j|$. *Assume*

that

(i) Uniformly over all n, the eigenvalues of $Q = \mathbb{E}[p^k(x)p^k(x)']$ are bounded from above and away from zero.

(ii) The conditional variance of the error terms is uniformly bounded in the following sense. There exists constants $\underline{\sigma}^2$ and $\bar{\sigma}^2$ such that for any $j = 1, 2 \ldots$ we have that $\underline{\sigma}^2 \leq \text{Var}(\epsilon_j \mid X) \leq \bar{\sigma}^2 < \infty$;

(iii) For each n and k there are finite constants c_k and ℓ_k such that for each $f \in \mathcal{G}$

$$\|r_k\|_{L,2} = (\mathbb{E}[r_k(x)^2])^{1/2} \leq c_k \quad \textit{and} \quad \|r_k\|_{L,\infty} = \sup_{x \in \mathcal{X}} |r_k(x)| \leq \ell_k c_k.$$

(iv) $\sup_{x \in \mathcal{X}} \mathbb{E}[\bar{\epsilon}_k^2 \mathbf{1}\{\bar{\epsilon}_k + \ell_k c_k > \delta\sqrt{n}/\xi_k\} \mid X = x] \to 0$ as $n \to \infty$ and $\sup_{x \in \mathcal{X}} \mathbb{E}[\ell_k^2 c_k^2 \mathbf{1}\{\bar{\epsilon}_k + \ell_k c_k > \delta\sqrt{n}/\xi_k\} \mid X = x] \to 0$ as $n \to \infty$ for any $\delta > 0$.

As mentioned, these are exactly the conditions required by Belloni et al. (2015), with the modification that the bounds on conditional variance and other moment conditions on the error term hold uniformly over $j = 1, \ldots, k$. The assumptions on the series terms being used in the approximation can be shown to be satisfied by a number of commonly used functional bases, such as polynomial bases or splines, under adequate normalizations and smoothness of the underlying regression function. Readers should refer to Newey (1997), Chen (2007), or Belloni et al. (2015) for a more in depth discussion of these assumptions.[1]

Under these assumptions, the variance of our second-stage estimator is governed by one of the following variance matrices:

$$\begin{aligned} \tilde{\Omega} &:= Q^{-1}\mathbb{E}[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']Q^{-1} \\ \Omega_0 &:= Q^{-1}\mathbb{E}[\{p^k(x) \circ \epsilon^k\}\{p^k(x) \circ \epsilon^k\}']Q^{-1} \end{aligned} \tag{2.4.3}$$

[1] In practice, we recommend the use of B-splines in order to to satisfy the first requirement that the basis functions are weakly positive and to reduce instability of the convex optimization programs described in (2.2.8)-(2.2.9).

where $\circ$ represents the Hadamard (element-wise) product and, abusing notation, for a vector $a \in \mathbb{R}^k$ and scalar $c \in \mathbb{R}$ we let $a + c = (a_i + c)_{i=1}^k$. Later on, we establish the validity of the plug-in analog $\widehat{\Omega}$ (2.2.13), as an estimator of these matrices.

Theorem 2.4.1 (Pointwise Normality)**.** *Suppose that Condition 1 and Assumption 2.4.1 hold. In addition suppose that $\xi_k^2 \log k/n \to 0$. Then so long as either the logistic propensity score model or linear outcome regression model is correctly specified, for any $\alpha \in S^{k-1}$:*

$$\sqrt{n}\frac{\alpha'(\widehat{\beta}^k - \beta^k)}{\|\alpha'\Omega^{1/2}\|} \to_d N(0,1) \tag{2.4.4}$$

where generally $\Omega = \tilde{\Omega}$ but if $\ell_k c_k \to 0$ then we can set $\Omega = \Omega_0$. Moreover, for any $x \in \mathcal{X}$ and $s(x) \coloneqq \Omega^{1/2} p^k(x)$,

$$\sqrt{n}\frac{p^k(x)'(\widehat{\beta}^k - \beta^k)}{\|s(x)\|} \to_d N(0,1) \tag{2.4.5}$$

and if the approximation error is negligible relative to the estimation error, namely $\sqrt{n} r_k(x) = o(\|s(x)\|)$, then

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|s(x)\|} \to_d N(0,1) \tag{2.4.6}$$

Theorem 2.4.1 shows that the estimator proposed in Section 2.2 has a limiting gaussian distribution even under misspecification of either first-stage model. This allows for doubly-robust pointwise inference after establishing a consistent variance estimator.

2.4.2. Uniform Convergence

Next, we turn to strengthening the pointwise results to hold uniformly over all points $x \in \mathcal{X}$. This requires stronger conditions. We make the following assumptions on the tail behavior of the error terms which strengthens Assumption 2.4.1.

Assumption 2.4.2 (Uniform Limit Theory)**.** *Let $\bar{\epsilon}_k = \sup_{1 \le j \le k} |\epsilon_j|$, $\alpha(x) \coloneqq p^k(x)/\|p^k(x)\|$,*

and let

$$\xi_k^L := \sup_{\substack{x,x' \in \mathcal{X} \\ x \neq x'}} \frac{\|\alpha(x) - \alpha(x')\|}{\|x - x'\|}.$$

Further for any integer s let $\bar{\sigma}_k^s = \sup_{x \in \mathcal{X}} \mathbb{E}[|\bar{\epsilon}_k|^s | X = x]$. For some $m > 2$ assume

(i) The regression errors satisfy $\sup_{x \in \mathcal{X}} \mathbb{E}[\max_{1 \leq i \leq n} |\bar{\epsilon}_{k,i}|^m \mid X = x] \lesssim_P n^{1/m}$

(ii) The basis functions are such that (a) $\xi_k^{2m/(m-2)} \log k / n \lesssim 1$, (b) $(\bar{\sigma}_k^2 \vee \bar{\sigma}_k^m) \log \xi_k^L \lesssim \log k$, and (c) $\log \bar{\sigma}_k^m \xi_k \lesssim \log k$.

As before, Assumption 2.4.2 is very similar to its analogue in Belloni et al. (2015), with the modification that the conditions are required to hold for $\bar{\epsilon}_k$ as opposed to ϵ_k. Under this assumption, we derive doubly-robust uniform rates of convergence uniform inference procedures for the conditional counterfactual outcome $g_0(x)$.

Theorem 2.4.2 (Strong Approximation by a Gaussian Process)**.** *Assume that Condition 1 holds and that Assumptions 2.4.1-2.4.2 hold with $m \geq 3$. In addition assume that (i) $\bar{R}_{1n} = o_p(a_n^{-1})$ and (ii) $a_n^6 k^4 \xi_k^2 (\bar{\sigma}_k^3 + \ell_k^3 c_k^2)^2 \log^2 n / n \to 0$ where*

$$\bar{R}_{1n} := \sqrt{\frac{\xi_k^2 \log k}{n}} (n^{1/m} \sqrt{\log k} + \sqrt{k} \ell_k c_k) \quad \text{and} \quad \bar{R}_{2n} := \sqrt{\log k} \cdot \ell_k c_k$$

Then so long as either the propensity score model or outcome regression model is correctly specified, for some $\mathcal{N}_k \sim N(0, I_k)$:

$$\sqrt{n} \frac{\alpha(x)'(\widehat{\beta} - \beta)}{\|\alpha(x)'\Omega^{1/2}\|} =_d \frac{\alpha(x)'\Omega^{1/2}}{\|\alpha(x)'\Omega^{1/2}\|} N_k + o_p(a_n^{-1}) \ \text{ in } \ell^\infty(\mathcal{X}) \tag{2.4.7}$$

so that for $s(x) := \Omega^{1/2} p^k(x)$

$$\sqrt{n} \frac{p^k(x)'(\widehat{\beta} - \beta)}{\|s(x)\|} =_d \frac{s(x)}{\|s(x)\|} N_k + o_p(a_n^{-1}) \ \text{ in } \ell^\infty(\mathcal{X}) \tag{2.4.8}$$

and if $\sup_{x\in\mathcal{X}} \sqrt{n}|r_k(x)|/\|s(x)\| = o(a_n^{-1})$*, then*

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|}N_k + o_p(a_n^{-1}) \quad \textit{in } \ell^\infty(\mathcal{X}) \tag{2.4.9}$$

where in general we take $\Omega = \tilde{\Omega}$ *but if* $\bar{R}_{2n} = o_p(a_n^{-1})$ *then we can set* $\Omega = \Omega_0$ *where* $\tilde{\Omega}$ *and* Ω_0 *are as in* (2.4.3)*.*

Theorem 2.4.2 establishes conditions under which we obtain a doubly-robust strong approximation of the empirical process $x \mapsto \sqrt{n}(\widehat{g}(x) - g_0(x))$ by a Gaussian process. After establishing consistent estimation of the matrix Ω, this strong approximation result allows us to show validity of the uniform confidence bands described in Section 2.2. As noted by Belloni et al. (2015), this is distinctly different from a Donsker type weak convergence result for the estimator $\widehat{g}(x)$ as viewed as a random element of $\ell^\infty(X)$. In particular, the covariance kernel is left completely unspecified and in general need not be well behaved.

2.4.3. Matrix Estimation and Uniform Inference

We establish that the estimator $\widehat{\Omega}$ proposed in (2.2.13) is a consistent estimator of the true limiting variance Ω, where $\Omega = \tilde{\Omega}$ in general but if $\bar{R}_{2n} = o_p(a_n^{-1})$ then $\Omega = \Omega_0$. To do so, we rely on the second-stage assumptions Assumptions 2.4.1 and 2.4.2 as well as the following condition limiting the first-stage estimation error passed on to the variance estimator $\widehat{\Omega}$.

Condition 2 (Variance Estimation)**.** Let $m > 2$ be as in Assumption 2.4.2. Then,

$$\xi_{k,\infty} \max_{1\le j\le k} \mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] = o_p(k^{-2}n^{-1/m}) \tag{2.4.10}$$

Via Lemma 2.3.1 we can establish Condition 2 under Assumption 2.3.1 as well as the additional

sparsity bound[2]

$$\frac{\xi_{k,\infty}^5 s_k^2 k^2 \ln(d_z)}{n^{(m-1)/m}}. \tag{2.4.11}$$

Theorem 2.4.3 (Matrix Estimation)**.** *Suppose that Conditions 1 and 2 and Assumptions 2.4.1-2.4.2 hold. In addition, assume that* $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$*. Then, so long as either the propensity score model or outcome regression model is correctly specified then for* $\widehat{\Omega} = \widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1}$*:*

$$\|\widehat{\Omega} - \Omega\| \lesssim_P (v_n \vee \ell_k c_k)\sqrt{\frac{\xi_k^2 \log k}{n}} = o(1)$$

Theorem 2.4.3 establishes that pointwise inference based on the test statistic described in Section 2.2, obtained by replacing Ω in Theorem 2.4.1 with the consistent estimator $\widehat{\Omega}$, is doubly-robust. Hypothesis tests based on the test statistic as well as pointwise confidence intervals for $g_0(x)$ remain valid even if one of the first-stage parameters is misspecified.

We now establish the validity of uniform inference based on the gaussian bootstrap critical values $c_u^\star(1-\alpha)$ defined in Section 2.2.

Theorem 2.4.4 (Validity of Uniform Confidence Bands)**.** *Suppose Conditions 1 and 2 are satisfied and Assumptions 2.4.1–2.4.2 hold with* $m \geq 4$*. In addition suppose (i)* $R_{1n} + R_{2n} \lesssim \log^{1/2} n$*, (ii)* $\xi_k \log^2 n / n^{1/2-1/m} = o(1)$*, (iii)* $\sup_{x\in\mathcal{X}} |r_k(x)|/\|p^k(x)\| = o(\log^{-1/2} n)$*, and (iv)* $k^4\xi_k^2(1+l_k^3 r_k^3)^2 \log^5 n/n = o(1)$*. Then, so long as either the propensity score model or outcome regression model is satisfied*

$$\Pr\left(\sup_{x\in\mathcal{X}} \left|\frac{\widehat{g}(x) - g(x)}{\widehat{\sigma}(x)}\right| \leq c^\star(1-\alpha)\right) = 1 - \alpha + o(1).$$

[2]The sparsity bound (2.4.11) required for consistent variance estimation can be significantly sharpened if the researcher is willing to use a cross fitting procedure, using one sample to estimate the nuisance parameters and another to evaluate the aIPW signal. This is because one could more directly follow Semenova and Chernozhukov (2021) and control alternate quantities with bounds that converge more quickly to zero.

As a result, uniform confidence intervals formed in (2.2.14) *satisfy*

$$\Pr(g(x) \in [\underline{i}(x), \bar{i}(x)],\ \forall x \in \mathcal{X}) = 1 - \alpha + o(1).$$

In conjunction with Lemma 2.3.1, Theorem 2.4.1 and Theorem 2.4.3, Theorem 2.4.4 shows the validity of the uniform inference procedure described in Section 2.2.

2.5. Estimation of the Conditional Average Treatment Effect

Up to now, we have mainly focused on doubly-robust estimation and model-assisted inference for the function

$$g_0(x) = \mathbb{E}[Y_1 \mid X = x].$$

We conclude by noting that we can use a symmetric procedure to obtain model-assisted inference for the additional conditional counterfactual outcome

$$\tilde{g}_0(x) = \mathbb{E}[Y_0 \mid X = x].$$

To do so, we use the alternate aIPW signal

$$Y_0(\pi_0, m_0) = \frac{(1-D)Y}{1-\pi_0(Z)} + \left(\frac{1-D}{1-\pi_0(Z)} - 1\right) m_0(Z)$$

where as before the true value for $\pi_0^\star(z) = \Pr(D = 1 \mid Z = z)$ but now $m_0^\star(z) = \mathbb{E}[Y \mid D = 0, Z = z]$. To estimate these nuisance models we again assume a logistic form for the propensity score model $\pi_0(z) = \pi(z; \gamma^0)$ and a linear form for the outcome regression model $m_0(z) = m(z, \alpha^0)$ as in (2.2.7) and use a separate estimation procedure for each basis term in our series approximation of $\tilde{g}_0(x)$. The estimating equations we use to estimate each γ_j^0

and α_j^0 differ from those in (2.2.8)-(2.2.9) however, and are instead given

$$\widehat{\gamma}_j^0 \coloneqq \arg\min_{\gamma} \mathbb{E}_n[p_j(X)\{(1-D)e^{\gamma' Z} - D\gamma' Z\}] + \lambda_{\gamma,j}\|\gamma\|_1$$
$$\widehat{\alpha}_j^0 \coloneqq \arg\min_{\alpha} \mathbb{E}_n[p_j(Z)(1-D)e^{\widehat{\gamma}_j^{0\prime} Z}(Y - \alpha' Z)^2]/2 + \lambda_{\alpha,j}\|\alpha\|_1$$

which under the natural analog of Assumption 2.3.1 converge uniformly to population minimizers:

$$\bar{\gamma}_j^0 \coloneqq \arg\min_{\gamma} \mathbb{E}[p_j(X)\{(1-D)e^{\gamma' Z} - D\gamma' Z\}]$$
$$\bar{\alpha}_j^0 \coloneqq \arg\min_{\alpha} \mathbb{E}[p_j(Z)(1-D)e^{\bar{\gamma}_j^{0\prime} Z}(Y - \alpha' Z)^2]$$

Letting $\bar{\pi}_{0,j}(z) = \pi(z, \bar{\gamma}_j^0)$, and $\bar{m}_{0,j}(z) = m(z, \bar{\alpha}_j^0)$ we can repeat the decomposition of Section 2.3, expressing $\tilde{Y}(\bar{\pi}_{0,j}, \bar{m}_{0,j})$ as functions of the parameters $\bar{\gamma}_j^0$ and $\bar{\alpha}_j^0$ and show that the first order conditions for $\bar{\gamma}_j^0$ and $\bar{\alpha}_j^0$ directly control the bias passed on to the second stage nonparametric estimator for $\tilde{g}_0(x)$. Convergence rates and validity of inference then follow from symmetric analysis of the results in Sections 2.3 and 2.4. Combining estimation and inference of the two conditional counterfactual outcomes then gives a doubly-robust estimator and inference procedure for the CATE. To perform inference on the CATE we can use the variance matrix

$$\bar{\Omega} = \Omega_0 + \Omega_1 - 2\Omega_2$$

where Ω_0 is as in (2.4.3) but Ω_1 and Ω_2 are given

$$\begin{aligned} \Omega_1 &= Q^{-1}\mathbb{E}[\{p^k(x) \circ \epsilon_0^k\}\{p^k(x) \circ \epsilon_0^k\}']Q^{-1} \\ \Omega_2 &= Q^{-1}\mathbb{E}[\{p^k(x) \circ \epsilon^k\}\{p^k(x) \circ \epsilon_0^k\}']Q^{-1} \end{aligned} \tag{2.5.1}$$

where $\epsilon_{0,j}^k = Y_0(\bar{\pi}_{0,j}, \bar{m}_{0,j}) - \tilde{g}_0(x)$ and $\epsilon_0^k = (\epsilon_{0,1}^k, \ldots, \epsilon_{0,k}^k)'$. These matrices can be consistently estimated using their natural empirical analogs as in (2.2.13).

2.6. Empirical Application

We apply the model assisted estimator to estimate the effect of maternal smoking on infant birthweight conditional on the age of the mother. We use the Cattaneo (2010) dataset which can be found online on the Stata website.[1] The dataset describes each infant's birthweight in grams, Y, whether or not the mother smoked during pregnancy, $D = 1$ indicating smoking, and a number of covariates containing information on the mother's health and socioeconomic background, $Z = (X, Z_1)$, where X represents the conditioning variable, maternal age. The dataset includes a base of 21 control variables. We additionally construct quadratic powers and interactions of continuous control variables to generate an additional 29 control variables so that in total $d_z = 50$. A full summary of the data used as well as additional details/analysis from our empirical analysis can be found in Section 2.11.

We compare the model assisted estimator of the CATE against one where standard MLE and OLS loss functions are used to estimate the first stage propensity score and outcome regression models. We also qualitatively compare our results to Zimmert and Lechner (2019), who use a kernel based approach to estimate the CATE in this setting. While this sort of comparison is not perfect since we do not know the true DGP, this setting is advantageous for analysis since we strongly expect that (i) the effect of smoking on birthweight will be negative and (ii) this effect should grow stronger in magnitude as the age of the mother increases. These hypotheses have been corroborated by other work that examines the conditional average treatment effect in this setting (Zimmert and Lechner, 2019; Abrevaya, 2006; Lee et al., 2017).

2.6.1. Empirical Results

Figure 2.6.1 displays our main results from implementing both the model assisted and standard MLE/OLS estimation procedures. After removing the top 3% and bottom 3% of smoker and non-smoker birthweights by maternal age, we select the penalty parameters

[1] The dataset can be downloaded here.

for the first stage models via the bootstrap procedure described in Section 2.4. The pilot penalty parameters are uniformly taken to be equal to zero, so that the residuals used in the bootstrap procedure are generated from non-regularized estimations. We take $c_0 = 2$ in (2.2.18) and select the first stage penalty parameters using the 90^{th}, 85^{th}, and 80^{th} quantiles of the bootstrap distribution. For the second stage basis functions we implement second degree b-splines with 3 knots via the splines2 package in `R` (Wang and Yan, 2021).

Consistent with prior work, both estimators of the CATE suggest that the effect of smoking on birthweight becomes more negative with age. Both estimation procedures also generally produces negative estimates for the CATE, but it should be noted that for the lowest levels of penalization the model assisted CATE estimate suggests a slightly positive effect of smoking for particularly young mothers, though this difference is not significantly different from zero. The shapes of the estimated functions remain relatively stable under various sizes of the penalty parameter, though the model assisted procedure is more sensitive to the level of regularization introduced.[2] Overall, the magnitude of the CATE estimates produced by the model assisted estimator seem to be more reasonable those produced by the standard estimator.

For the most part, the effects found here are similar to those found in Zimmert and Lechner (2019), though the effects estimated using standard first stage loss functions have somewhat larger magnitudes and in general both series estimation procedures seem to give less reasonable results on the boundaries. An advantage of using a series second stage however, in contrast to the kernel second stage of Zimmert and Lechner (2019), is the existence of the uniform confidence bands displayed. Reassuringly, the estimates of Zimmert and Lechner (2019) seem to be within the 95% uniform confidence bands generated by the model assisted estimator.

As a robustness check, we also try estimating the treatment effect via second degree splines with five knots and first degree splines with seven knots. These results are displayed in

[2]Numerically solving the minimization problems in (2.2.8)-(2.2.9) also typically requires more iterations to converge than solving the standard MLE/OLS minimization problems.

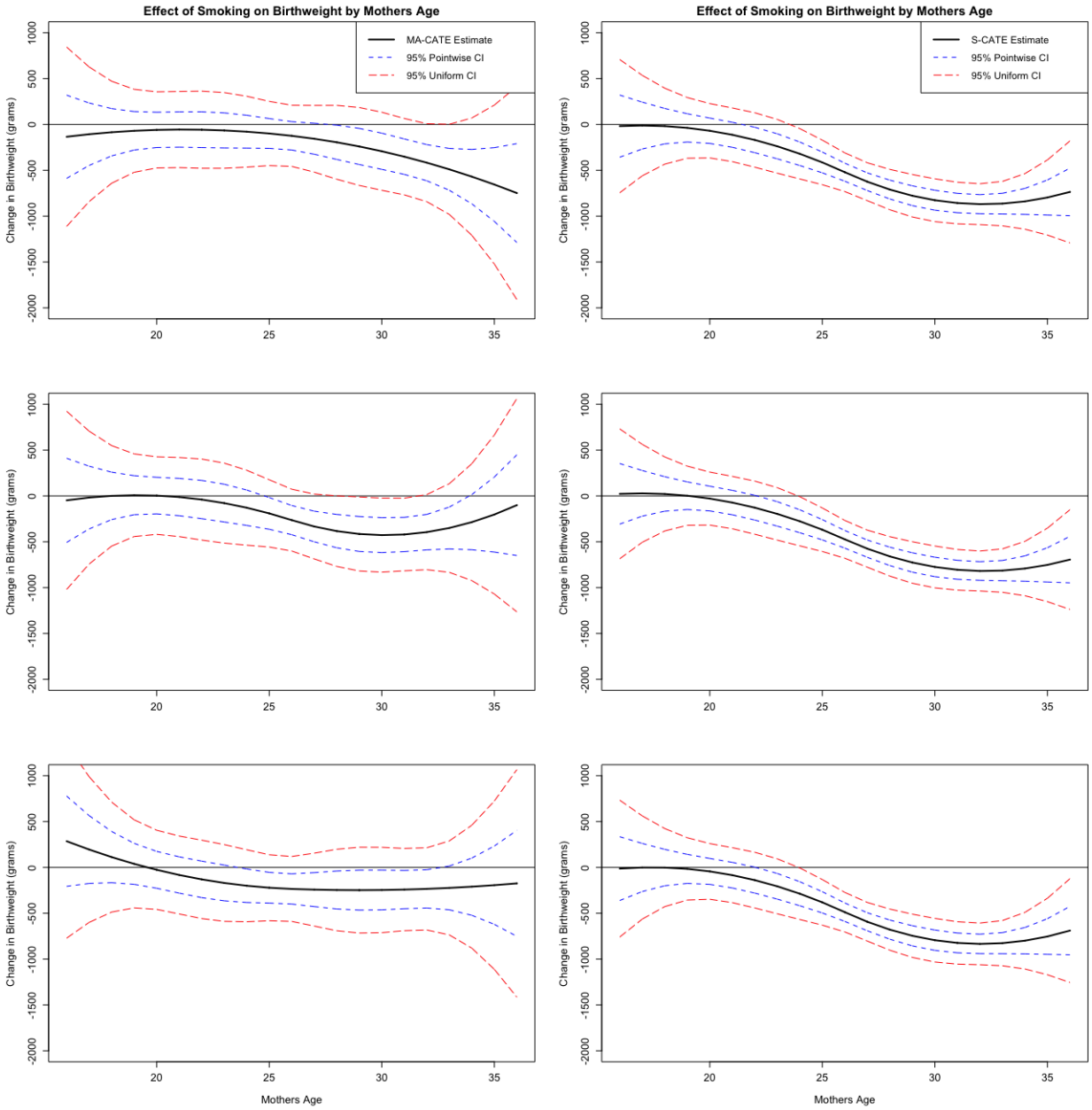

Figure 2.6.1: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 90^{th} quantile of the bootstrap distribution to select the penalty parameters, second row uses 85^{th} quantile, and final row uses the 80^{th} quantile. Second stage is computed using b-splines of the second degree with 3 knots. 95% pointwise confidence intervals are displayed in blue short dashes and 95% uniform confidence bands are displayed in long red dashes.

Figures 2.6.2 and 2.6.3, respectively. Again, we find that the effect of smoking on child birthweight is almost uniformly negative regardless of estimation procedure used or choice of penalty parameter. The shape of the estimated CATE function remains fairly stable under both alternative specifications. Again, the confidence bands from the model assisted procedure remain larger than the confidence bands from the standard procedure. However, the in the first degree spline specification the uniform confidence bands for the standard procedure suggest a significantly positive CATE for some values of maternal age; an implausible result.

Finally, Table 2.6.1 reports the smoothed average treatment effect estimates taken from averaging the model assisted CATE estimates from Figure 2.6.1 across observations. Again, these estimates are in line with prior work

Table 2.6.1: Smoothed Model Assisted ATE Estimates

Bootstrap Penalty Qt.	90^{th}	85^{th}	80^{th}
Implied ATE	-163.257	-222.431	-207.827

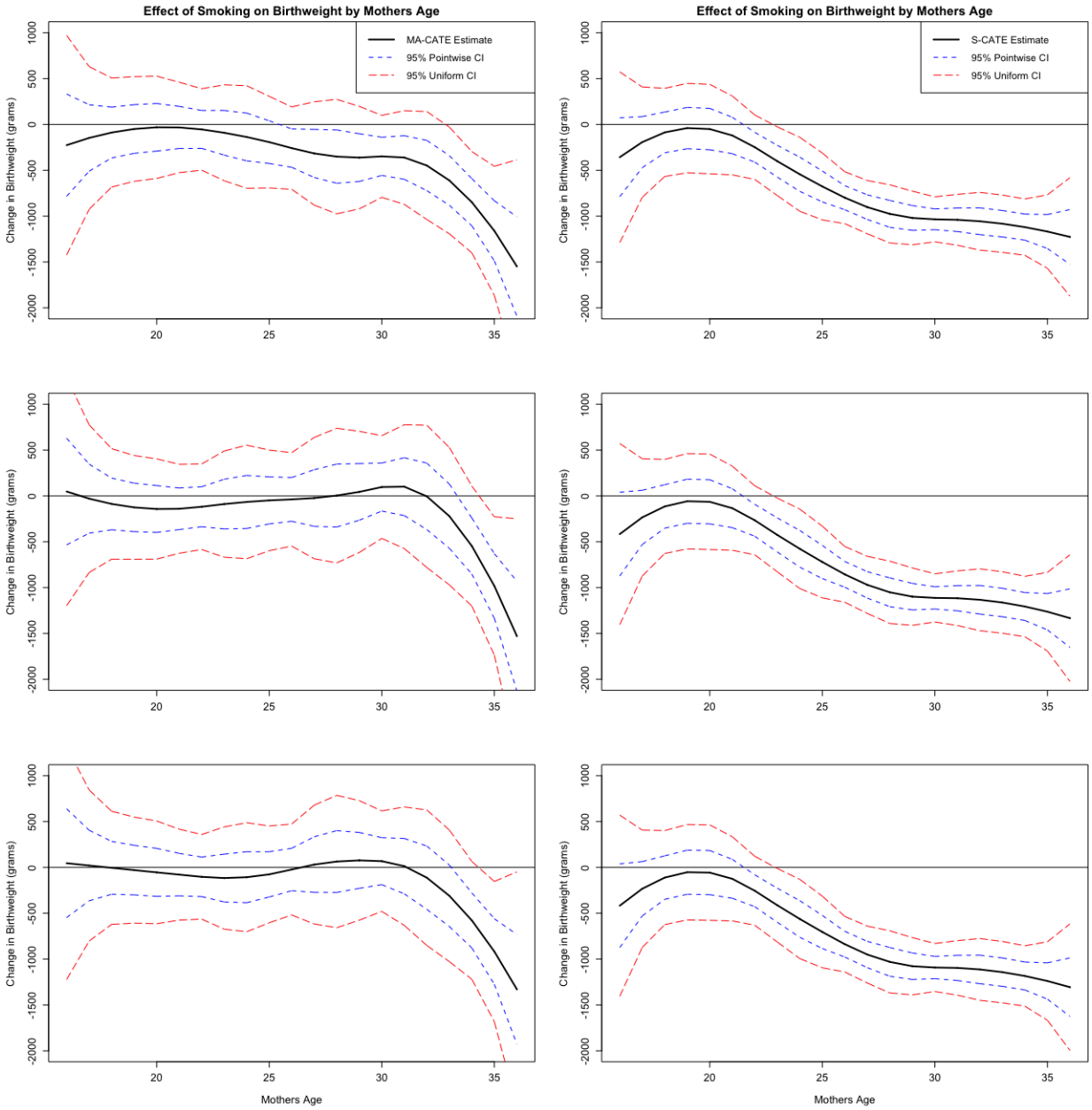

Figure 2.6.2: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 95^{th} quantile of the bootstrap distribution to select the penalty parameters, second row uses 90^{th} quantile, and final row uses the 85^{th} quantile. Second stage is computed using b-splines of the second degree with 5 knots. 95% pointwise confidence intervals are displayed in blue short dashes and 95% uniform confidence bands are displayed in long red dashes.

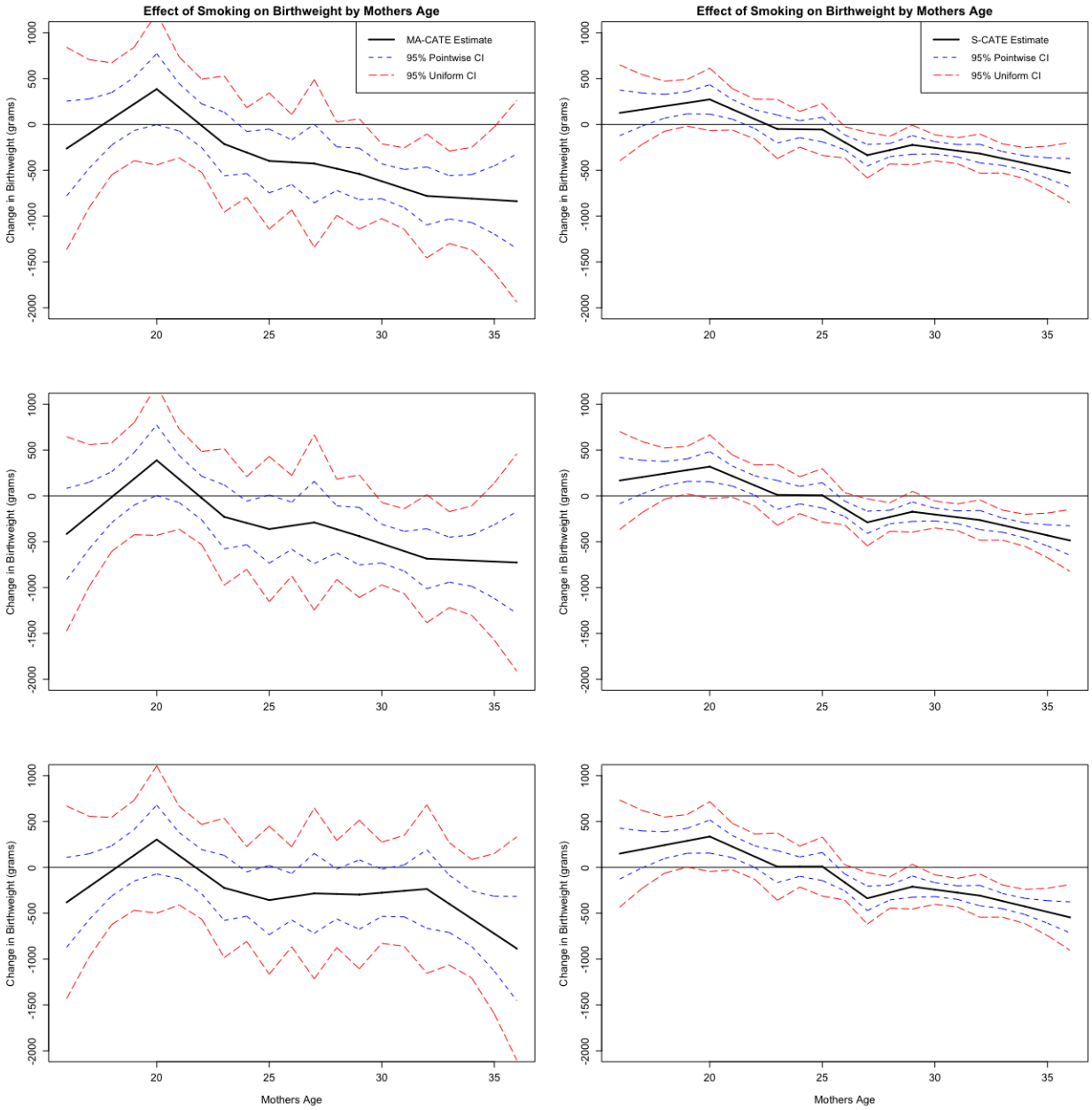

Figure 2.6.3: CATE of maternal smoking estimated using model assisted estimating equations (left) and standard MLE/OLS estimating equations (right). Top row uses the 95^{th} quantile of the bootstrap distribution to select the penalty parameters, second row uses 90^{th} quantile, and final row uses the 85^{th} quantile. Second stage is computed using b-splines of the first degree with 5 knots. 95% pointwise confidence intervals are displayed in blue short dashes and 95% uniform confidence bands are displayed in long red dashes.

2.7. Simulation Study

We investigate the finite-sample performance of the doubly-robust estimator and inference procedure via simulation study. We find that our proposed estimation procedure retains good coverage properties even under misspecification.

2.7.1. Simulation Design

Observations are generated i.i.d. according to the following distributions The error term is generated following $\epsilon \sim N(0,1)$. The controls are set $Z_i = (Z_{1i}, X_i) \in \mathbb{R}^{d_z}$ where $d_z = 100$, $X \sim U(1,2)$, and the independent regressors Z_1 are jointly centered Gaussian with a covariance matrix of the Toeplitz form

$$\text{Cov}(Z_{1,j}, Z_{1,k}) = \mathbb{E}[Z_{1,j} Z_{1,k}] = 2^{-|j-k|}, \quad 3 \leq j, k \leq d_z.$$

To capture misspecification, we let $Z^\dagger$ be a transformation of the regressors in Z_1 where $Z_j^\dagger = Z_j + \max(0, 1 + Z_j)^2, \ \forall \ j = 3, \ldots, d_z$. Let `sparsity` control the number of regressors in $Z = (Z_1, X)$ entering the DGP.

(S1) *Correct specification*: Generate D given Z from a Bernoulli distribution with $\Pr(D = 1|Z) = \{1 + \exp(p_1 - X - 0.5X^2 - \gamma' Z_1)\}^{-1}$ and $Y = D(1 + X + 0.5X^2 + \gamma' Z_1) + \epsilon$.

(S2) *Propensity score model correctly specified, but outcome regression model misspecified*: Generate D given Z as in (S1), but $Y = D(1 + X + 0.5X^2 + \gamma' Z_1^\dagger) + \epsilon$.

(S3) *Propensity score model misspecified, but outcome regression model correctly specified*: Generate Y according to (S1), but generate D given Z from a Bernoulli distribution with $\Pr(D = 1|Z) = \{1 + \exp(p_2 - X - 0.5X^2 + \gamma' Z_1^\dagger)\}^{-1}$.

where the constants p_1 and p_2 differ in various simulation setups but are always set so that the average probability of treatment is about one half. To consider various degrees of

high-dimensionality, we implement $N \in \{500, 1000\}$ with $d_z = 100$. For (S1), `sparsity`$= 6$; for (S2), `sparsity`$= 4$; and, for (S3), `sparsity`$= 5$. Results are reported for $S = 1,000$ repeated simulations.

2.7.2. Estimators and Implementation

To select the first stage penalty parameters, we implement the multiplier bootstrap procedure described in Section 2.2.3. The constants $c_{\gamma,j}$ and $c_{\alpha,j}$ in the pilot penalty parameters (2.2.15) are selected via cross validation from a set of size 5. To select the final bootstrap penalty parameter we set $c_0 = 1.1$ and select the 95^{th} quantile of $B = 10000$ bootstrap replications. In our second-stage estimation, we use a b-spline basis of size $k = 3$. B-splines are implemented from the R package `splines2` (Wang and Yan, 2021), which uses the specification detailed in Perperoglou et al. (2019). In the tables below, we refer to our method as *MA-DML* (model assisted double machine learning).

We compare our proposed estimator and inference procedure to that of Semenova and Chernozhukov (2021), which projects a single aIPW signal onto a growing series of basis terms. In implementing this *DML* method, we use the standard ℓ_1-penalized maximum likelihood (MLE) and ordinary least squares (OLS) loss functions to estimate the first stage propensity score and outcome regression models, respectively.[1]

Estimation error is studied for the target parameter $g_0(x) = \mathbb{E}[Y|D = 1, X = x]$ over a grid of 100 points spaced across $x \in [1, 2]$, i.e. the support of X. We study average coverage across simulations of each method's pointwise (at $x = 1.5$) and uniform confidence intervals. To compare the estimation error for the target parameter $g(x)$ across the two different estimators $\widehat{g}_s(x)$ for each simulation $s = 1, \ldots, S$, we utilize integrated bias, variance, and mean-squared

[1]Vira Semenova provides several example `R` scripts implementing *DML*: `https://sites.google.com/view/semenovavira/research`.

error where $\bar{g}(x) = S^{-1} \sum_{s=1}^{S} \widehat{g}_s(x)$,

$$\begin{aligned}
\text{IBias}^2 &= \int_0^1 (\bar{g}(x) - g_0(x))^2 dx, \\
\text{IVar} &= S^{-1} \sum_{s=1}^{S} \int_0^1 (\widehat{g}_s(x) - \bar{g}(x))^2 dx, \\
\text{IMSE} &= S^{-1} \sum_{s=1}^{S} \int_0^1 (\widehat{g}_s(x) - g_0(x))^2 dx.
\end{aligned}$$

2.7.3. Simulation Results

Table 2.7.1 presents the simulation results for all three specifications (S1)-(S3) for $n = 500$ and $n = 1000$. Integrated squared bias, variance, and mean squared error are presented in columns (1)-(3), respectively. Pointwise and uniform coverage results are presented in columns (4)-(7).

For pointwise and uniform coverage under correct specification regime (S1), *MA-DML* has some slight improvements. Under misspecification DGPs (S2) and (S3), the pointwise coverage of *MA-DML* is closer to the targets except in the $N = 1000$ and (S2) case where it slightly underperforms. However, *MA-DML* has a notable improvement over *DML* in the (S3) case when $N = 1000$. Similarly, *MA-DML* outperforms *DML* in three of the four misspecified regimes, i.e. all but (S3) when $N = 500$ where *MA-DML* has over-coverage. Under (S2) when $N = 1000$, both methods are markedly deterioated uniform coverage, although *MA-DML* is noticably closer to target.

In regards to estimation error, in four of the six settings, *MA-DML* has a lower MSE than *DML* where regardless of sample size *MA-DML* underperforms in (S3). Notably, it does appear *MA-DML* has substantially smaller IBias^2 across the DGPs.

Finally, we were surprised to find for both estimators that coverage properties, in general,

Table 2.7.1: Simulation study.

DGP	Estimator	IBias^2 (1)	IVar (2)	IMSE (3)	Cov90 (4)	Cov95 (5)	UCov90 (6)	UCov95 (7)
		K=3, n=500, $d_z = 100$						
(S1)	DML	0.04	0.31	0.35	0.92	0.96	1.00	1.00
	MA-DML	$\sim$0.0	0.34	0.34	0.93	0.97	1.00	1.00
(S2)	DML	0.16	2.17	2.33	0.92	0.97	0.83	0.86
	MA-DML	0.03	2.12	2.15	0.90	0.94	0.88	0.91
(S3)	DML	0.03	0.55	0.59	0.87	0.93	0.95	0.97
	MA-DML	0.01	0.79	0.80	0.91	0.95	0.99	0.99
		K=3, n=1000, $d_z = 100$						
(S1)	DML	0.12	0.20	0.32	0.83	0.90	0.96	0.96
	MA-DML	0.01	0.22	0.23	0.83	0.90	0.99	0.99
(S2)	DML	0.40	2.1	2.5	0.84	0.91	0.33	0.39
	MA-DML	0.19	2.07	2.26	0.83	0.89	0.50	0.55
(S3)	DML	0.11	0.34	0.46	0.74	0.82	0.80	0.84
	MA-DML	0.01	0.53	0.54	0.84	0.89	0.89	0.91

Note: DGP refers to the three various data generating processes introduced above. IBias^2, IVar, and IMSE refer to integrated squared bias, variance, and mean squared error, respectively. Cov90, Cov95, UCov90, and UCov95 refer to the coverage proportion of the 90% and 95% pointwise and uniform confidence intervals across simulations. K refers to the number of series terms, N to the sample size, and d_z to the dimensionality of the random variable Z_1.

improve under the higher-dimensional regime of $N = 500$ with $d_z = 100$ compared to $N = 1,000$ and $d_z = 100$. In particular, with a higher ratio of covariates to observations, the uniform coverage properties under regime (S2) were substantially better. The estimation error results were in line with our priors as the higher-dimensional regime sees in general higher estimation errors for both methods.

For coverage under correct specification, we did anticipate the underperformance of *MA-DML* given it is designed to handle misspecification with the cost of other estimators outperforming under correct specification. Additionally, we attribute the poor uniform coverage in DGP (S2) for both estimators under $N = 1,000$ to a lack of a rich enough cross-validation given the performance was improved under a more difficult regime when the number of observations drops to $N = 500$. The integrated bias of *MA-DML* is lower across the various DGPs compared to *DML*. Following the discussion in Section 2.3 this is expected since the first stage estimating equations for the model assisted procedure are specifically designed to minimize the bias passed on to the second stage estimator. However, the model assisted procedure has higher values of integrated variance compared to the standard procedure, which could be attributable to the use of k distinct first-stage estimations.

Our findings should not be interpreted as a critique of the Semenova and Chernozhukov (2021) benchmark method, whose work we rely on and were inspired by.

2.8. Conclusion

Estimation of conditional average treatment effects with high dimensional controls typically relies on first estimating two nuisance parameters: a propensity score model and an outcome regression model. In a high-dimensional setting, consistency of the nuisance parameter estimators typically relies on correctly specifying their functional forms. While the resulting second-stage estimator for the conditional average treatment effect typically remains consistent even if one of the nuisance parameters is inconsistent, the confidence intervals may no longer

be valid.

In this paper, we consider estimation and valid inference on the conditional average treatment effect in the presence of high dimensional controls and nuisance parameter misspecification. We present a nonparametric estimator for the CATE that remains consistent at the non-parametric rate, under slightly modified conditions, even under misspecification of either the logistic propensity score model or linear outcome regression model. The resulting Wald-type confidence intervals based on this estimator also provide valid asymptotic coverage under nuisance parameter misspecification.

2.9. Appendix: Proofs for Results in Main Text

Here we provide proofs of the main results in Sections 2.3-2.4. The proofs for Section 2.4 rely on an assortment of supporting lemmas proved in Section 2.9.3.

2.9.1. Proofs for Main First Stage Results

Proof of Lemma 2.3.1

The proof of Lemma 2.3.1 relies on a series of non-asymptotic bounds that are established in Online Appendix Lemmas 2.9.5 and 2.9.6 that hold on $\bigcap_{m=1}^{6} \Omega_{k,m}$ and depend on the quantity

$$\bar{\lambda}_k = M\xi_{k,\infty}\sqrt{\frac{\log(d_z/\epsilon)}{n}}$$

where M is a fixed constant. In addition let $\tilde{\Sigma}^1_{\alpha,j} := \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{\alpha}_j'Z|ZZ']$ and $\Sigma^1_{\alpha,j} := \mathbb{E}\tilde{\Sigma}^1_{\alpha,j}$ and define the event

$$\Omega_{k,7} := \{\|\tilde{\Sigma}^1_{\alpha,j} - \Sigma^1_{\alpha,j}\|_\infty \leq \bar{\lambda}_k, \forall j \leq k\} \tag{2.9.1}$$

In Online Section 2.9.3 we show that $\Pr(\bigcap_{m=1}^{7}) \geq 1 - o(1)$. Under these events, Lemma 2.9.1, below provides the bound needed for first statement of Lemma 2.3.1 while Lemma 2.9.2

provides the bound needed for the second statement.

Lemma 2.9.1 (Nonasymptotic Bounds for Weighted Means)**.** *Suppose that Assumption 2.3.1 holds, $\xi_0 > (c_0+1)/(c_0-1)$, and $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$. In addition, assume there is a constant $c > 0$ such that $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ for all $j \leq k$. Then, under the event $\bigcap_{m=1}^{7}\Omega_{k,m}$, there is a constant M_2 that does not depend on k such that*

$$\max_{1\leq j\leq k}|\mathbb{E}_n[p_j(X)Y(\widehat{\pi}_j,\widehat{m}_j)] - \mathbb{E}_n[p_j(X)Y(\bar{\pi}_j,\bar{m}_j)]| \leq M_2 s_k\bar{\lambda}_k^2 \tag{2.9.2}$$

Proof. We show that the bound of (2.9.2) holds for any $j = 1,\ldots,k$ in a couple steps. To save notation, define

$$\begin{aligned}\mu_j(\pi,m) &:= \mathbb{E}_n\left[p_j(X)Y(\pi,m)\right] \\ &= \mathbb{E}_n\left[p_j(X)\left\{\frac{DY}{\pi(Z)} + \left(\frac{D}{\pi(Z)} - 1\right)m(Z)\right\}\right]\end{aligned}$$

Step 1: Decompose Difference and Use Logistic FOCs. Consider the following decomposition

$$\begin{aligned}\mu_j(\widehat{\pi}_j,\widehat{m}_j) - \mu_j(\bar{\pi}_j,\bar{m}_j) &= \mathbb{E}\left[p_j(X)\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}\left(1 - \frac{D}{\bar{\pi}_j(X)}\right)\right] \\ &\quad + \mathbb{E}_n\left[p_j(X)D\{Y - \bar{m}_j(Z)\}\left(\frac{1}{\widehat{\pi}_j(Z)} - \frac{1}{\bar{\pi}_j(Z)}\right)\right] \\ &\quad + \mathbb{E}_n\left[p_j(X)\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}\left(\frac{D}{\bar{\pi}_j(Z)} - \frac{D}{\widehat{\pi}_j(Z)}\right)\right] \\ &:= \delta_{1,j} + \delta_{2,j} + \delta_{3,j}\end{aligned}$$

Notice that $\delta_{1,j} + \delta_{3,j} = (\widehat{\alpha}_j - \bar{\alpha}_j)'\mathbb{E}_n[p_j(X)(1 - D/\widehat{\pi}_j(Z))Z]$. By the first order conditions for $\widehat{\gamma}_j$ we have that

$$|\mathbb{E}_n[p_j(X)\{Z_l - DZ_l/\widehat{\pi}_j(Z)\}]| \leq \lambda_{\gamma,j}\ \forall l = 1,\ldots,d_z \implies \|\mathbb{E}_n[p_j(X)\{Z_l - DZ_l/\widehat{\pi}_j(Z)\}]\|_\infty \leq \lambda_{\gamma,j}.$$

Applying Hölder's inequality to $\delta_{1,j} + \delta_{3,j}$ then gives us that on the event $\Omega_{k,2}$

$$|\delta_{1,j} + \delta_{3,j}| \leq \|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \lambda_{\gamma,j} \leq \|\widehat{\alpha}_j - \bar{\alpha}_j\| \bar{\lambda}_k.$$

By Lemma 2.9.6 on the event $\bigcap_{m=1}^{6} \Omega_{k,m}$ and under the conditions of Lemma 2.9.1, $\|\widehat{\alpha}_j - \bar{\alpha}_j\| \leq M_1 s_k \bar{\lambda}_k$ where M_1 is a constant that does not depend on k. So

$$|\delta_{1,j} + \delta_{3,j}| \leq M_1 s_k \bar{\lambda}_k^2 \tag{M.1}$$

Step 2: Use Outcome Regression Score Domination to Bound $\delta_{2,j}$. Now deal with the term $\delta_{2,j}$. By first order Taylor expansion, for some $u \in (0,1)$

$$\begin{aligned}
\delta_{2,j} &= -(\widehat{\gamma}_j - \bar{\gamma}_j)' \mathbb{E}_n[p_j(X) D\{Y - \bar{m}_j(Z)\} e^{-\bar{\gamma}_j' Z} Z] \\
&\quad + (\widehat{\gamma}_j - \bar{\gamma}_j)' \mathbb{E}_n[p_j(X) D\{Y - \bar{m}_j(Z)\} e^{-u\widehat{\gamma}_j' Z - (1-u)\bar{\gamma}_j' Z} ZZ'](\widehat{\gamma}_j - \bar{\gamma}_j)/2 \\
&:= \delta_{21,j} + \delta_{22,j}
\end{aligned}$$

In the event $\Omega_{k,1} \cap \Omega_{k,2} \cap \Omega_{k,3} \cap \Omega_{k,4}$ we have by score domination of the linear outcome regression model and Lemma 2.9.5 that $\delta_{21} \leq M_0 s_k \bar{\lambda}_k^2$.

The term $\delta_{22,j}$ is second order. On the event $\Omega_{k,0} \cap \Omega_{k,1}$ where $\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{\lambda}_k \leq M_0 \eta / C_0$ it can be bounded with

$$\begin{aligned}
\delta_{22,j} &\leq e^{C_0 \|\widehat{\gamma}_j - \bar{\gamma}_j\|_1} \mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z} |Y - \bar{m}_j(Z)| \{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}^2] \\
&\leq e^{M_0 \eta} \mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z} |Y - \bar{m}_j(Z)| \{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}^2].
\end{aligned}$$

This in turn is bounded in a few steps. First note on the event $\Omega_{k,7}$

$$(\mathbb{E}_n - \mathbb{E})[p_j(X) D e^{-\bar{\gamma}_j' Z} |Y - \bar{m}_j(Z)| \{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}^2] \leq \bar{\lambda}_k \|\widehat{\gamma}_j - \bar{\lambda}_j\|_1^2.$$

By Assumption 2.3.1 we have that $G_0^2 E[D|Y - \bar{m}_j(Z)| \mid Z] \leq G_1^2/G_0 + G_0$ so that,

$$\mathbb{E}[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{m}_j(Z)|\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq (G_1^2/G_0 + G_0)\mathbb{E}[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2].$$

On the event $\Omega_{k,6}$ we have that

$$(\mathbb{E}_n - \mathbb{E})[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1.$$

Putting these all together gives

$$\begin{aligned}
&\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}|Y - \bar{m}_j(Z)|\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \\
&\qquad \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2 + (G_1^2/G_0 + G_0)\bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\
&\qquad\quad + (G_1^2/G_0 + G_0)\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}]
\end{aligned} \tag{M.2}$$

To bound (M.2) note again that in the event $\Omega_{k,1} \cap \Omega_{k,2}$, $\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{\lambda}_k$ and that using by (O.4) in Online Appendix Lemma 2.9.6:

$$\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq e^{-M_0\eta}M_0 s_k \bar{\lambda}_k^2.$$

Plugging these into (M.2) gives

$$\delta_{22,j} \leq e^{M_0\eta}M_0^2 s_k^2 \bar{\lambda}_k^3 + e^{M_0\eta}(G_1^2/G_0 + G_0)M_0^2 s_k^2 \bar{\lambda}_k^3 + (G_1^2/G_0 + G_0)M_0 s_k \bar{\lambda}_k^2 \tag{M.3}$$

so that in total $\delta_{2,j} = \delta_{21,j} + \delta_{22,j}$ is bouned

$$\delta_{2,j} \leq M_0 s_k (G_1^2/G_0 + G_0 + 1)\bar{\lambda}_k^2 + e^{M_0\eta}M_0^2 s_k^2 (G_1^2/G_0 + G_0 + 1)\bar{\lambda}_k^3 \tag{M.4}$$

Step 3: Combine Terms. Putting this together yields

$$\begin{aligned}|\delta_{1,j} + \delta_{2,j} + \delta_{3,j}| \leq \{M_1 + M_0(G_1^2/G_0 + G_0 + 1)\} s_k \bar{\lambda}_k^2 \\ + e^{M_0 \eta}(G_1^2/G_0 + G_0) M_0^2 s_k^2 \bar{\lambda}_k^3\end{aligned} \tag{M.5}$$

Use the fact that $s_k \bar{\lambda}_k \leq \eta < 1$ to simplify the last term of this expression

$$\begin{aligned}|\delta_{1,j} + \delta_{2,j} + \delta_{3,j}| \leq \{M_1 + M_0(G_1^2/G_0 + G_0 + 1)\} s_k \bar{\lambda}_k^2 \\ + e^{M_0 \eta}(G_1^2/G_0 + G_0) M_0^2 s_k \bar{\lambda}_k\end{aligned} \tag{M.6}$$

This gives the result (2.9.2) after taking $M_2 = M_1 + M_0(G_1^{/} G_0 + G_0 + 1) + e^{M_0 \eta}(G_1^2/G_0 + G_0) M_0^2$.

□

Lemma 2.9.2 (Nonasymptotic Bounds for Variance Estimation)**.** *Suppose that Assumption 2.3.1 hold, $\xi_0 > (c_0 + 1)/(c_0 - 1)$, and $2C_0 \nu_0^{-2} s_k \bar{\lambda}_k \leq \eta < 1$. In addition, assume there is a constant $c > 0$ such that $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ for all $j \leq k$. Then, under the event $\bigcap_{m=1}^{7} \Omega_{k,m}$, there is a constant M_3 that does not depend on k such that*

$$\max_{1 \leq j \leq k} \mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] \leq M_3 \xi_{k,\infty}^2 s_k^2 \bar{\lambda}_k^2 \tag{2.9.3}$$

Proof. We show the bound holds for each $j = 1, \ldots, k$. We start by decomposing

$$\begin{aligned}p_j(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j)) &= p_j(X)\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}\left(1 - \frac{D}{\bar{\pi}_j(X)}\right) \\ &\quad + p_j(X) D\{Y - \bar{m}_j(Z)\}\left(\frac{1}{\widehat{\pi}_j(Z)} - \frac{1}{\bar{\pi}_j(Z)}\right) \\ &\quad + p_j(X)\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}\left(\frac{D}{\bar{\pi}_j(Z)} - \frac{D}{\widehat{\pi}_j(Z)}\right) \\ &:= \tilde{\delta}_{1,j} + \tilde{\delta}_{2,j} + \tilde{\delta}_{3,j}\end{aligned}$$

We will use the fact that $(a+b+c)^2 \leq 4a^2 + 4b^2 + 4c^2$ to bound

$$\mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] \leq 4\mathbb{E}_n[\tilde{\delta}_{1,j}^2] + 4\mathbb{E}_n[\tilde{\delta}_{2,j}^2] + 4\mathbb{E}_n[\tilde{\delta}_{3,j}^2]. \tag{V.1}$$

To bound $\mathbb{E}_n[\tilde{\delta}_{2,j}]$ use the mean value equation (O.2) in Online Appendix Lemma 2.9.6 and the lower bound on $\bar{g}_j(z)$ from Assumption 2.3.1

$$\begin{aligned}\mathbb{E}_n[\tilde{\delta}_{2,j}^2] &= \mathbb{E}_n[p_j^2(X)D\{Y - \bar{m}_j(Z)\}^2\{\widehat{\pi}_j^{-1}(Z) - \bar{\pi}_j^{-1}(Z)\}^2] \\ &\leq \xi_{k,\infty} e^{-B_0} \left(1 + e^{C_0\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1}\right)^2 \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{m}_j(Z)\}^2\{\widehat{g}_j(Z) - \bar{g}_j(Z)\}^2]\end{aligned}$$

Applying (O.8) in Online Appendix Lemma 2.9.6, Online Appendix Lemma 2.9.5, and $s_k\bar{\lambda}_k \leq \eta < 1$ there is a constant $\tilde{M}_1$ that does not depend on k such that in the event $\bigcap_{m=1}^7 \Omega_{k,m}$ this is bounded

$$\leq \tilde{M}_1 \xi_{k,\infty} s_k \bar{\lambda}_k^2 \tag{V.2}$$

To bound $\mathbb{E}_n[\tilde{\delta}_{3,j}]$ write $\widehat{\pi}_j^{-1}(Z) - \bar{\pi}_j^{-1}(Z) = e^{-\bar{\gamma}_j'Z}\{e^{-\widehat{\gamma}_j'Z + \bar{\gamma}_j'Z} - 1\}$ and use the lower bound on $\bar{g}_j(z)$ from Assumption 2.3.1:

$$\begin{aligned}\mathbb{E}_n[\tilde{\delta}_{3,j}^2] &= \mathbb{E}_n[p_j^2(X)D\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}^2\{\widehat{\pi}_j^{-1}(Z) - \bar{\pi}_j^{-1}(Z)\}^2] \\ &\leq \xi_{k,\infty} e^{-B_0} \left(1 + e^{C_0\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1}\right)^2 \mathbb{E}_n[p_j(X)e^{-\bar{\gamma}_j'Z}\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}^2]\end{aligned}$$

Applying Online Appendix Lemma 2.9.6, there is a constant $\tilde{M}_2$ that does not depend on k such that on the event $\bigcap_{m=1}^6 \Omega_{k,m}$ this is bounded

$$\leq \tilde{M}_2 \xi_{k,\infty} s_k \bar{\lambda}_k^2 \tag{V.3}$$

Finally, to bound $\mathbb{E}_n[\tilde{\delta}_{1,j}^2]$ again use the lower bound on $\bar{g}_j(z)$ and decompose

$$\begin{aligned}\mathbb{E}_n[\tilde{\delta}_{1,j}^2] &= \mathbb{E}_n[p_j^2(X)\{\widehat{m}(z) - \bar{m}(z)\}^2\{1 - D/\bar{\pi}_j(Z)\}^2] \\ &\leq \xi_{k,\infty}^2(1 + e^{-B_0})^2\mathbb{E}_n[\{\widehat{m}_j(Z) - \bar{m}_j(Z)\}^2] \\ &\leq \xi_{k,\infty}^2(1 + e^{-B_0})^2 C_0^2\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1^2\end{aligned}$$

Again on the event $\bigcap_{m=1}^6 \Omega_{k,m}$ apply Online Appendix Lemma 2.9.6 this is bounded, for some constant $\tilde{M}_3$ that does not depend on k by

$$\leq \tilde{M}_3\xi_{k,\infty}^2 s_k^2\bar{\lambda}_k^2 \tag{V.4}$$

The result (2.9.3) follows by collecting (V.1)-(V.4). □

2.9.2. Proofs of Main Second Stage Results

The proofs for Section 2.4 closely follow those of Belloni et al. (2015) with some modifications to deal with the various error terms. They also rely on some additional second stage results proved in Online Section 2.10 .

Proof of Theorem 2.4.1

Equation (2.4.5) follows from applying (2.4.4) with $\alpha = p(x)/\|p(x)\|$ and (2.4.6) follows from (2.4.5). So it suffices to prove (2.4.4).

For any $\alpha \in S^{k-1}$, $1 \lesssim \|\alpha'\Omega^{1/2}\|$ because of the conditional variance of $\bar{\epsilon}_j^2$ is bounded from below and from above and under the positive semidefinite ranking

$$\Omega \geq \Omega_0 \geq \underline{\sigma}^2 Q^{-1}.$$

Moreover, by condition (ii) of the theorem and Lemma 2.10.2, $R_{1n}(\alpha) = o_p(1)$. So we can

write

$$\begin{aligned}\sqrt{n}\alpha'(\widehat{\beta}-\beta) &= \frac{\sqrt{n}\alpha'}{\|\alpha'\Omega^{1/2}\|}\mathbb{G}_n[p^k(x)\circ(\epsilon^k+r_k)]+o_p(1)\\ &= \sum_{i=1}^{n}\frac{\alpha'}{\sqrt{n}\|\alpha'\Omega^{1/2}\|}\{p^k(x)\circ(\epsilon^k+r_k)\}.\end{aligned}$$

Goal will be to verify Lindberg's condition for the CLT. Throughout the rest of the proof, it will be helpful to make the following notations. First, for any vector $a=(a_1,\ldots,a_k)'\in S^{k-1}$, let $|a|=(|a_1|,\ldots,|a_k|)'$ and note that $|a|\in S^{k-1}$ as well:

$$\tilde{\alpha}_n'=\frac{\alpha'}{\sqrt{n}\|\alpha_n'\Omega^{1/2}\|},\quad \omega_n:=|\tilde{\alpha}|'p^k(x),\quad \text{and}\quad \bar{\epsilon}_k:=\sup_{1\le j\le k}|\epsilon_j|$$

Now, by the definition of Ω we have that

$$\text{Var}\left(\sum_{i=1}^{n}\frac{\alpha'}{\sqrt{n}\|\alpha'\Omega^{1/2}\|}\{p^k(x)\circ(\epsilon^k+r_k)\}\right)=1.$$

Second for each $\delta>0$

$$\begin{aligned}&\sum_{i=1}^{n}\mathbb{E}\left[(\tilde{\alpha}_n'\{p^k(x)\circ(\epsilon^k+r_k)\})^2\mathbf{1}\left\{|\tilde{\alpha}_n'\{p^k(x)\circ(\epsilon^k+r_k)\}|>\delta\right\}\right]\\ &\quad\le\sum_{i=1}^{n}\mathbb{E}\left[\omega_n^2\mathbb{E}\left[\bar{\epsilon}_k^2\,\mathbf{1}\{|\omega_n||\bar{\epsilon}_k+\ell_kc_k|>\delta\}\mid X=x\right]\right]\end{aligned}\tag{2.9.4}$$

What we are using here is the following. Suppose α is a nonrandom vector in $\mathbb{R}^k$, a is a (positive) random vector in $\mathbb{R}^k$ and b is a random vector in $\mathbb{R}^k$. Then,

$$\{\alpha'(a\circ b)\}=\sum_{j=1}^{k}\alpha_ja_jb_j\le\|b\|_\infty\sum_{j=1}^{k}|\alpha_j|a_j=(|\alpha|'a)\|b\|_\infty.\tag{2.9.5}$$

To bound the right hand side of (2.9.4) use the fact that $1 \lesssim \|\alpha'\Omega^{1/2}\|$ because $1 \lesssim \underline{\sigma}^2$ and

$$\Omega \geq \Omega_0 \geq \underline{\sigma}^2 Q^{-1}$$

in the positive semidefinite sense. Using these two we have

$$n\mathbb{E}|\omega_n|^2 \leq \mathbb{E}[(|\alpha|'p^k(x))^2]/(\alpha'\Omega\alpha) \lesssim 1.$$

By the bounded eigenvalue condition and using the trace operator:

$$\mathbb{E}[(|\alpha|p^k(x))^2] = \text{trace}(\mathbb{E}[|\alpha|'p^k(x)'p^k(x)|\alpha|]) = |\alpha|'Q|\alpha| \lesssim \|\alpha\| = 1$$

Further note, $|\omega_{ni}| \lesssim \frac{\xi_k}{\sqrt{n}}$. Using $(a+b)^2 \leq 2a^2+2b^2$, the right hand side of (2.9.4) is bounded by

$$2n\mathbb{E}[|\omega_n|^2\bar{\epsilon}_k^2\,\mathbf{1}\{|\bar{\epsilon}_k| + \ell_k c_k > \delta/|\omega_n|\}] + 2n\mathbb{E}[|\omega_n|^2\ell_k^2 c_k^2 \mathbf{1}\{|\bar{\epsilon}_k| + \ell_k c_k > \delta/|\omega_n|\}]$$

and both terms converge to zero. Indeed, to bound the first term note that, for some $c > 0$:

$$\begin{aligned} 2n\mathbb{E}[|\omega_n|^2\bar{\epsilon}_k^2\,\mathbf{1}\{|\bar{\epsilon}_k| + \ell_k c_k > \delta/|\omega_n|\}] &\lesssim n\mathbb{E}[|\omega_n|^2] \sup_{x\in\mathcal{X}} \mathbb{E}[\bar{\epsilon}_k^2 \mathbf{1}\{\bar{\epsilon}_k^2 + \ell_k c_k > c\delta\sqrt{n}/\xi_k\} \mid X = x] \\ &= o(1) \end{aligned}$$

where here we use the first part of Assumption 2.4.1(iv). To show the second term converges to zero, follow the same steps as for the first term, but apply the second part of Assumption 2.4.1(iv).

Proof of Theorem 2.4.2

We apply Yurinskii's coupling lemma (Pollard, 2001)

Yurinskii's Coupling Lemma

Let $\xi_1, \ldots, \xi_n$ be independent random k-vectors with $\mathbb{E}[\xi_i] = 0$ and $\beta := \sum_{i=1}^n \mathbb{E}[\|\xi_i\|^3]$ finite. Let $S := \xi_1 + \cdots + \xi_n$. For each $\delta > 0$ there exists a random vector T with a $N(0, \mathrm{var}(S))$ distribution such that

$$\mathbb{P}(|S - T| > 3\delta) \leq C_0 B \left(1 + \frac{|\log(1/B)|}{k}\right) \quad \text{where } B := \beta k \delta^{-3} \tag{YC}$$

for some universal constant C_0.

In order to apply the coupling, we want to consider a first order approximation to the estimator

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i, \quad \zeta_i = \Omega^{-1/2} p^k(x) \circ (\epsilon^k + r_k).$$

When $\bar{R}_{2n} = o_p(a_n^{-1})$ a similar argument can be used with $\zeta_i = \Omega^{-1/2} p^k(x) \circ (\epsilon^k + r_k)$ replaced with $\Omega^{-1/2} p^k(x) \circ \epsilon^k$. As before, the eigenvalues of Ω are bounded away from zero, therefore

$$\begin{aligned}
\mathbb{E}\|\zeta_i\|^3 &\lesssim \mathbb{E}[\|p^k(x) \circ (\epsilon^k(x) + r_k)\|^3] \\
&\lesssim \mathbb{E}[\|p^k(x)\|^3 (|\bar{\epsilon}_k|^3 + |r_k|^3)] \\
&\lesssim \mathbb{E}[\|p^k(x)\|^3](\bar{\sigma}_k^3 + \ell_k^3 c_k^3) \\
&\lesssim \mathbb{E}[\|p^k(x)\|^3]\xi_k(\bar{\sigma}_k^3 + \ell_k^3 c_k^3) \\
&\lesssim k\xi_k(\bar{\sigma}_k^3 + \ell_k^3 c_k^3)
\end{aligned}$$

Therefore, by Yurinskii's coupling lemma (YC), for each $\delta > 0$,

$$\begin{aligned}
\Pr\left\{\|\sum_{i=1}^n \zeta_i/\sqrt{n} - \mathcal{N}_k\| > 3\delta a_n^{-1}\right\} &\lesssim \frac{nk^2\xi_k(\bar{\sigma}_m^3 + \ell_k^3 c_k^3)}{(\delta a_n^{-1}\sqrt{n})^3}\left(1 + \frac{\log(k^3\xi_k(\bar{\sigma}_k^3 + \ell_k^3 c_k^3))}{k}\right) \\
&\lesssim \frac{a_n^3 k^2 \xi_k(\bar{\sigma}_k^3 + \ell_k^3 c_k^3)}{\delta^3 n^{1/2}}\left(1 + \frac{\log n}{k}\right) \to 0.
\end{aligned}$$

because $a_n^6 k^2 \xi_k (\bar{\sigma}_m^3 + \ell_k^3 c_k^3) \log^2 n/n \to 0$. Using the first two results from Lemma 2.10.3, (2.10.6)-(2.10.7), we obtain that

$$\|\sqrt{n}\alpha(x)'(\widehat{\beta}^k - \beta^k) - \alpha(x)'\Omega^{1/2}\mathcal{N}_k\| \leq \|1/\sqrt{n}\sum_{i=1}^{n}\alpha(x)'\Omega^{1/2}\zeta_i - \alpha(x)'\Omega^{1/2}\mathcal{N}_k\| + \bar{R}_{1n} = o_p(a_n^{-1}).$$

uniformly over $x \in \mathcal{X}$. Since $\|\alpha(x)'\Omega^{1/2}\|$ is bounded from below uniformly over $x \in \mathcal{X}$ we obtain the first statetment of Theorem 2.10.2 from which the second statement directly follows.

Finally, under the assumption that $\sup_{x\in\mathcal{X}} n^{1/2}|r(x)|/\|s(x)\| = o_p(a_n^{-1})$,

$$\frac{\sqrt{np(x)'(\widehat{\beta}^k - \beta^k)}}{\|s(x)\|} - \frac{\sqrt{n}(\widehat{g}(x) - g_0(x))}{\|s(x)\|} = o_p(a_n^{-1})$$

so that the third statement, (2.4.9) holds.

Proof of Theorem 2.4.3

Preliminaries for Proof of Theorem 2.4.3

Lemma (Symmetrization). *Let $Z_1, \ldots, Z_n$ be independent stochastic processes with mean zero and let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher random variables generated independetly of the data. Then*

$$\mathbb{E}^*\Phi\left(\frac{1}{2}\|\sum_{i=1}^n \epsilon_i Z_i\|_{\mathcal{F}}\right) \leq \mathbb{E}^*\Phi\left(\|\sum_{i=1}^n Z_i\|_{\mathcal{F}}\right) \leq \mathbb{E}^*\Phi\left(2\|\epsilon_i(Z_i - \mu_i)\|_{\mathcal{F}}\right), \tag{SI}$$

for every nondecreasing, convex $\Phi : \mathbb{R} \to \mathbb{R}$ and arbitrary functions $\mu_i : \mathcal{F} \to \mathbb{R}$.

For $p \geq 1$ consider the Shatten norm S_p on symmetrix $k \times k$ matrices Q defined by $\|Q\|_{S_p} = (\sum_{j=1}^k |\lambda_j(Q)|^p)^{1/p}$ where $\lambda_1(Q), \ldots, \lambda_k(Q)$ are the eigenvalues of Q. The case $p = \infty$ recovers the operator norm and $p = 2$ recovers the Frobenius norm.

Lemma (Khinchin's Inequality for Matrices). *For symmetric $k \times k$ matrices Q_i, $i = 1, \ldots, n$, $2 \leq p \leq \infty$, and an i.i.d sequence of Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$ we have*

$$\left\| \left(\mathbb{E}_n[Q_i^2]\right)^{1/2} \right\|_{S_p} \leq \left(\mathbb{E}_\epsilon \|\mathbb{G}_n[\epsilon_i Q_i]\|_{S_p}^p\right)^{1/p} \leq C\sqrt{p} \left\| \left(\mathbb{E}_n[Q_i^2]\right)^{1/2} \right\|_{S_p} \tag{KI-1}$$

where C is an absolute constant. So, for $k \geq 2$,

$$\mathbb{E}_\epsilon[\|\mathbb{G}_n[\epsilon_i Q_i]\|] \leq C\sqrt{\log k} \|(\mathbb{E}_n[Q_i^2])^{1/2}\| \tag{KI-2}$$

for some (possibly different) absolute constant C.

We will establish consistent estimation of

$$\Sigma = \mathbb{E}[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']$$

using

$$\widehat{\Sigma} = \mathbb{E}_n[\{p^k(x) \circ \widehat{\epsilon}^k\}\{p^k(x) \circ \widehat{\epsilon}^k\}']$$

Consistency of $\widehat{\Omega}$ will then follow from the consistency of $\widehat{Q}$ established by Lemma 2.10.1. To save notation, define the vectors

$$\widehat{Y} := \begin{bmatrix} Y(\widehat{\pi}_1, \widehat{m}_1) \\ \vdots \\ Y(\widehat{\pi}_k, \widehat{m}_k) \end{bmatrix} \quad \text{and} \quad \widehat{Y} := \begin{bmatrix} Y(\widehat{\pi}_1, \widehat{m}_1) \\ \vdots \\ Y(\widehat{\pi}_k, \widehat{m}_k) \end{bmatrix} \tag{2.9.6}$$

Also define $\dot{\epsilon}^k := (\dot{\epsilon}^k_1, \ldots, \dot{\epsilon}^k_k)$ so that $\dot{\epsilon}^k_j := Y(\bar{\pi}_j, \bar{m}_j) - \widehat{g}(x)$. Ideally, we would like to use $\dot{\epsilon}^k$ to estimate $\widehat{\Sigma}$, but we don't observe $\dot{\epsilon}^k$. Define $\Delta := \widehat{\epsilon}^k - \dot{\epsilon}^k = \widehat{Y}^k - \bar{Y}^k \in \mathbb{R}^k$.

Using this, we can decompose

$$\begin{aligned} \widehat{\Sigma} &= \mathbb{E}_n[\{p^k(x) \circ (\Delta + \dot{\epsilon}^k)\}\{p^k(x) \circ (\Delta + \dot{\epsilon}^k)\}] \\ &= \underbrace{\mathbb{E}_n[\{p^k(x) \circ \Delta\}\{p^k(x) \circ \Delta\}']}_{\Sigma_1} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \dot{\epsilon}^k\}\{p^k(x) \circ \Delta\}']}_{\Sigma_2} \\ &\quad + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \Delta\}\{p^k(x) \circ \dot{\epsilon}^k\}']}_{\Sigma_3} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ \dot{\epsilon}^k\}\{p^k(x) \circ \dot{\epsilon}^k\}]}_{\Sigma_4} \end{aligned} \tag{2.9.7}$$

We first show that $\|\Sigma_4 - \Sigma\| \to_p 0$. This is nonstandard because of the Hadamard product.

Lemma 2.9.3 (Psuedo-Variance Estimator Consistency)**.** *Suppose Assumption 2.4.1 and Assumption 2.4.2 hold. Further, define $v_n = \mathbb{E}[\max_{1 \le i \le n} |\bar{\epsilon}_k|^2]^{1/2}$. In addition, assume that $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$. Then,*

$$\|\widehat{Q} - Q\| \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}} = o(1)$$
$$\text{and} \quad \|\Sigma_4 - \Sigma\| \lesssim_P (v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2 \log k}{n}}$$

Proof. The first result is established by Lemma 2.10.1 (Matrix LLN). Rest of proof will follow proof of Theorem 4.6 in Belloni et al. (2015). Like in (2.9.7) we can define $\dot{\Delta} \equiv \dot{\epsilon}^k - \epsilon^k = g_0(x) - \widehat{g}(x)$[1] and decompose

$$\Sigma_4 = \underbrace{\mathbb{E}_n[p^k(x)p^k(x)'\dot{\Delta}^2]}_{\Sigma_{41}} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \cdot \dot{\Delta}\}']}_{\Sigma_{42}}$$
$$+ \underbrace{\mathbb{E}_n[\{p^k(x) \cdot \dot{\Delta}\}\{p^k(x) \circ (\epsilon^k + r_k)\}']}_{\Sigma_{43}} + \underbrace{\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}]}_{\Sigma_{44}}$$

The terms Σ_{41}, Σ_{42} and Σ_{43} are simple to show are negligible.

$$\begin{aligned}
&\|\Sigma_{41} + \Sigma_{42} + \Sigma_{43}\| \\
&\quad \leq \|\mathbb{E}_n[\{p^k(x)'(\widehat{\beta}^k - \beta^k)\}p^k(x)p^k(x)']\| + \|\mathbb{E}_n[\{p^k(x) \circ (\epsilon^k + r_k)\}p^k(x)'\{p^k(x)'(\widehat{\beta}^k - \beta^k)\}]\| \\
&\quad\quad + \|\mathbb{E}_n[p^k(x)\{p^k(x)'(\widehat{\beta}^k - \beta^k)\}\{p^k(x) \circ (\epsilon^k + r_k)\}']\| \\
&\quad \leq \max_{1\leq i\leq n} |p^k(x)(\widehat{\beta}^k - \beta^k)|^2 \|\mathbb{E}_n[p^k(x)p^k(x)']\| \\
&\quad\quad + 2 \max_{1\leq i\leq n} |\bar{\epsilon}_{k,i}| + |r_{k,i}| \max_{1\leq i\leq n} |p^k(x)'(\widehat{\beta} - \beta)| \|\mathbb{E}_n[p^k(x)p^k(x)']\|
\end{aligned}$$

By Theorem 2.10.2 $|\max_{1\leq i\leq n} |p^k(x)'(\widehat{\beta}^k - \beta^k)| \lesssim_P \xi_k^2(\sqrt{\log k} + \bar{R}_{1n} + \bar{R}_{2n})^2/n$, by Assumption 2.4.1 the approximation error is bounded $\max_{1\leq i\leq n} |r_{k,i}| \leq \ell_k c_k$, by Assumption 2.4.2 and Markov's inequality the errors are bounded $\max_{1\leq i\leq n} |\bar{\epsilon}_{k,i}| \lesssim_p v_n^2$. Finally, by the first part of Lemma 2.9.3 $\|\widehat{Q}\| \lesssim_P \|Q\| \lesssim 1$. Putting this all together with $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$ and $\xi_k^2 \log k/n \to 0$ gives

$$\|\Sigma_{41} + \Sigma_{42} + \Sigma_{43}\| \lesssim_P (v_n \vee 1 + \ell_k c_k)\sqrt{\frac{\xi_k^2 \log k}{n}}.$$

Next, we want to control $\Sigma_{44} - \Sigma$. To do this, let $\eta_1, \ldots, \eta_n$ be independent Rademacher random variables generated independently from the data. Then for $\eta = (\eta_1, \ldots, \eta_n)$

[1] It is useful to recall that $\dot{\epsilon}^k = \bar{Y}^k - \widehat{g}(x)$ and $\epsilon^k = \bar{Y}^k - g_0(x)$

$$
\begin{aligned}
&\mathbb{E}[\|\mathbb{E}_n[\{p^k(x)\circ(\epsilon^k+r_k)\}\{p^k(x)\circ(\epsilon^k+r_k)\}'] - \Sigma\|] \\
&\quad\lesssim \mathbb{E}[\mathbb{E}_\eta[\mathbb{E}_n\|\eta\{p^k(x)\circ(\epsilon^k+r_k)\}\{p^k(x)\circ(\epsilon^k+r_k)\}'\|]] \\
&\quad\lesssim \sqrt{\frac{\log k}{n}}\mathbb{E}[(\|\mathbb{E}_n[\|p^k(x)\|^2(\bar{\epsilon}_k+r_k)^2\{p^k(x)\circ(\epsilon^k+r_k)\}\{p^k(x)\circ(\epsilon^k+r_k)\}'\|])^{1/2}] \\
&\quad\lesssim \sqrt{\frac{\xi_k^2\log k}{n}}\mathbb{E}[\max_{1\le i\le n}|\bar{\epsilon}_{k,i}+r_k|(\|\mathbb{E}_n[\{p^k(x)\circ(\epsilon^k+r_k)\}\{p^k(x)\circ(\epsilon^k+r_k)\}'\|])^{1/2}] \\
&\quad\le \sqrt{\frac{\xi_k^2\log k}{n}}(\mathbb{E}[\max_{1\le i\le n}|\bar{\epsilon}_{k,i}+r_k|^2])^{1/2}\times(\mathbb{E}[\|\mathbb{E}_n[\{p^k(x)\circ(\epsilon^k+r_k)\}\{p^k(x)\circ(\epsilon^k+r_k)\}'\|]])^{1/2}
\end{aligned}
$$

where the first inequality holds from Symmetrization (SI), the second from Khinchin's inequality (KI-1), the third by $\max_{1\le i\le n}\|p^k(x)\|\le\xi_k$ and the fourth by Cauchy-Schwarz inequality.

Since for any positive numbers a, b and R, $a\le R(a+b)^{1/2}$ implies $a\le R^2+R\sqrt{b}$, the expression above and the triangle inequality yields

$$
\begin{aligned}
&\mathbb{E}[\|\mathbb{E}_n[\{p^k(x)\circ(\epsilon^k+r_k)\}\{p^k(x)\circ(\epsilon^k+r_k)\}'] - \Sigma\|] \\
&\qquad\lesssim \frac{\xi_k^2\log k}{n}(v_n^2+\ell_k^2c_k^2) + \left(\frac{\xi_k^2\log k}{n}\{v_n^2+\ell_k^2c_k^2\}\right)^{1/2}\|\Sigma\|^{1/2}
\end{aligned}
$$

and so, because $\|\Sigma\|\lesssim 1$ and $(v_n^2+\ell_k^2c_k^2)\xi_k^2\log k/n\to 0$ we have

$$
\mathbb{E}[\|\mathbb{E}_n[\{p^k(x)\circ(\epsilon^k+r_k)\}\{p^k(x)\circ(\epsilon^k+r_k)\}'] - \Sigma\|]\lesssim (v_n\vee 1+\ell_kc_k)\sqrt{\frac{\xi_k^2\log k}{n}}.
$$

The second result of Lemma 2.9.3 follows from Markov's inequality. □

Now, we need to take care of the terms

$$
\begin{aligned}
\Sigma_1 &= \mathbb{E}_n[\{p^k(x)\circ\Delta\}\{p^k(x)\circ\Delta\}'] \\
\Sigma_2 &= \mathbb{E}_n[\{p^k(x)\circ\dot{\epsilon}^k\}\{p^k(x)\circ\Delta\}'] \\
\Sigma_3 &= \mathbb{E}_n[\{p^k(x)\circ\Delta\}\{p^k(x)\circ\dot{\epsilon}^k\}']
\end{aligned}
$$

where $\Delta = \widehat{Y}^k - \bar{Y}^k$ and $\dot{\epsilon}^k = \bar{Y}^k - \widehat{g}(x) = \widehat{g}(x) - g^k(x) + \epsilon^k$. To do so we will use Condition 2.

Lemma 2.9.4 (Negligible Variance Bias)**.** *Suppose that Condition 2, Assumption 2.4.1 and Assumption 2.4.2 hold. Then*

$$\|\Sigma_1 + \Sigma_2 + \Sigma_3\| = o_p(1).$$

Proof. From Condition 2, the term Σ_1 being negligible immediately follows from Cauchy-Schwarz. Notice that

$$\begin{aligned}
\|\Sigma_1\| &\leq k \sup_{\substack{1\leq l\leq k\\ 1\leq j\leq k}} |\mathbb{E}_n[p_l(X)(Y(\hat{\pi}_l, \hat{m}_l) - Y(\bar{\pi}_j, \bar{m}_j))p_l(X)(Y(\widehat{\pi}_l, \widehat{m}_l) - Y(\bar{\pi}_l, \bar{m}_l))]| \\
&\leq k \sup_{1\leq l\leq k} (\mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2])^{1/2} \sup_{1\leq j\leq k} (\mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2])^{1/2} \\
&= o_p(1).
\end{aligned}$$

To see that Σ_2 is negligible notice that

$$\begin{aligned}
\|\Sigma_2\| &\leq k \sup_{\substack{1\leq l\leq k\\ 1\leq j\leq k}} \mathbb{E}_n[p_l(X)(\epsilon_l + p^k(x)'(\widehat{\beta}^k - \beta^k))p_j(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))] \\
&\leq k \sup_{1\leq l\leq k} \mathbb{E}_n[p_l(X)^2(\epsilon_l + p^k(x)'(\widehat{\beta} - \beta))^2]^{1/2} \mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2]^{1/2} \\
&\leq \xi_{k,\infty}(\max_{1\leq i\leq n} |\bar{\epsilon}_k| + \max_{1\leq i\leq n} p^k(x)'(\widehat{\beta} - \beta)) \mathbb{E}_n[p_j(X)^2(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2]^{1/2}
\end{aligned}$$

Applying Assumption 2.4.2 and Theorem 2.10.2 gives

$$\lesssim_P k\xi_{k,\infty} n^{1/m} \mathbb{E}[p_j(X)^2 Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2]^{1/2} = o_p(1)$$

where the final line is via Condition 2. Showing negligibility of Σ_3 follows the same steps. □

Proof of Theorem 2.4.4

Follows from the exact same steps as Theorem 3.5 in Semenova and Chernozhukov (2021) after establishing strong approximation by a gaussian process as in Theorem 2.4.2 and consistent variance estimation as in Theorem 2.4.3.

2.9.3. Supporting Lemmas for First Stage

Here we provide supporting lemmas and their proofs. We start off with non-asymptotic bounds for first stage parameters and means.

Nonasymptotic Bounds for the First Stage

The nonasymptotic bounds for the first stage will depend on certain events. In Section 2.9.3 we will show that under Assumption 2.3.1 these events happen with probability approaching one. To control sparsity, define $\mathcal{S}_{\gamma,j} := \{j : \bar{\alpha}_j \neq 0\}$, $\mathcal{S}_{\alpha,j} := \{j : \bar{\alpha}_j \neq 0\}$. Recall $s_k := \max_{1\leq j\leq k}\{|\mathcal{S}_{\gamma,j}| \vee |\mathcal{S}_{\alpha,j}|\}$. Define the scores

$$\begin{aligned} S_{\gamma,j} &:= \mathbb{E}_n[U_{\gamma,j}Z] \\ S_{\alpha,j} &:= \mathbb{E}_n[U_{\alpha,j}Z] \end{aligned} \tag{2.9.8}$$

With these in mind, we will consider nonasymptotic bounds under the events:

$$\begin{aligned} \Omega_{k,1} &:= \{\lambda_{\gamma,j} \geq c_0 \cdot \|S_{\gamma,j}\|_\infty, \forall j \leq k\} \\ \Omega_{k,2} &:= \{\lambda_{\gamma,j} \leq \bar{\lambda}_k, \forall j \leq k\} \end{aligned} \tag{2.9.9}$$

Following Chetverikov and Sørensen (2021), the first event is referred to as "score domination" while the second event is referred to as "penalty majorization".

Bounds will be established on the ℓ_1 convergence rate of the estimated coefficient vector as

well as on the symmetrized Bregman divergences, $D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j)$ and $D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \gamma_j)$, defined by

$$\begin{aligned} D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) &:= \mathbb{E}_n \left[p_j(X) D\{e^{-\widehat{\gamma}_j' Z} - e^{-\bar{\gamma}_j' Z}\}\{\bar{\gamma}_j' Z - \widehat{\gamma}_j' Z\} \right], \\ D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \widehat{\gamma}) &:= \mathbb{E}_n \left[p_j(X) D e^{-\widehat{\gamma}_j' Z} (\bar{\alpha}_j' Z - \widehat{\alpha}_j' Z)^2 \right]. \end{aligned} \tag{2.9.10}$$

Lemma 2.9.5 (Nonasymptotic Bounds for Logistic Model)**.** *Suppose that Assumption 2.3.1 holds with $\xi_0 > (c_0+1)/(c_0-1)$ and $2C_0\nu_0^{-2} s_k \bar{\lambda}_k \leq \eta < 1$. Then, under the events $\Omega_{k,1} \cap \Omega_{k,2}$ defined in (2.9.9), there exists a finite constant M_0 that does not depend on k such that*

$$\max_{1 \leq j \leq k} D^{\ddagger}(\bar{g}, \widehat{g}) \leq M_0 s_k \bar{\lambda}_k^2 \quad \textit{and} \quad \max_{1 \leq j \leq k} \|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{\lambda}_k \tag{2.9.11}$$

Proof. We show that the bound of (2.9.11) holds for each $j = 1, \ldots, k$. For any $\gamma \in \mathbb{R}^d$ define $\tilde{\ell}_j(\gamma) := \mathbb{E}_n[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}]$. By optimality of $\widehat{\gamma}_j$ we must have, for any $u \in (0, 1]$:

$$\tilde{\ell}_j\left(\widehat{\gamma}_j\right) + \lambda_{\gamma,j}\|\widehat{\gamma}_j\|_1 \leq \tilde{\ell}\left((1-u)\widehat{\gamma}_j + u\bar{\gamma}_j\right) + \lambda_{\gamma,j}\|(1-u)\widehat{\gamma}_j + u\bar{\gamma}_j\|_1.$$

Using convexity of the ℓ_1 norm $\|\cdot\|_1$, this gives after rearrangment

$$\tilde{\ell}_j\left(\widehat{\gamma}_j\right) - \tilde{\ell}\left((1-u)\widehat{\gamma}_j + u\bar{\gamma}_j\right) + \lambda_{\gamma,j} u\|\widehat{\gamma}_j\|_1 \leq \lambda_{\gamma,j} u\|\bar{\gamma}_j\|_1.$$

Divide both sides by u and let $u \to^+ 0$

$$\mathbb{E}_n[p_j(X) D\{e^{-\widehat{\gamma}' Z} + (1-D)\}\{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}] + \lambda_{\gamma,j}\|\widehat{\gamma}_j\|_1 \leq \lambda_{\gamma,j}\|\bar{\gamma}_j\|_1.$$

By direct calculation, we have that $D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j)$ from (2.9.10) can be expressed

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) = \mathbb{E}_n[p_j(X) D\{e^{-\widehat{\gamma}' Z} + (1-D)\}\{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}] - \mathbb{E}_n[p_j(X) D\{e^{-\bar{\gamma}' Z} + (1-D)\}\{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}].$$

Combining the last two displays yields

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + \mathbb{E}_n[p_j(X)D\{e^{-\bar{\gamma}'Z} + (1-D)\}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}] + \lambda_{\gamma,j}\|\widehat{\gamma}_j\|_1 \leq \lambda_{\gamma,j}\|\bar{\gamma}_j\|_1 \quad \text{(L.1)}$$

In the event $\Omega_{k,1}$ we have that

$$|\mathbb{E}_n[p_j(X)D\{e^{-\bar{\gamma}'Z} + (1-D)\{\widehat{\gamma}'Z - \bar{\gamma}'Z\}\}] \leq c_0^{-1}\lambda_{\gamma,j}\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \quad \text{(L.2)}$$

Combining (L.1) and (L.2) yields

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + \lambda_{\gamma,j}\|\widehat{\gamma}_j\|_1 \leq \lambda_{\gamma,j}\|\bar{\gamma}_j\| + c_0^{-1}\lambda_{\gamma,j}\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1.$$

Expanding $\|\gamma_j\|_1 = \sum_{l\in\mathcal{S}_{\gamma,j}}|\gamma_l| + \sum_{l\notin\mathcal{S}_{\gamma,j}}|\gamma_l|$ for $\gamma = \widehat{\gamma}_j, \bar{\gamma}_j$ and applying the triangle inequalities $|\widehat{\gamma}_{j,l}| \geq |\bar{\gamma}_{j,l}| - |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|$ for $l \in \mathcal{S}_{\gamma,j}$ and the equality $\widehat{\gamma}_{j,l} = \widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}$ gives

$$\begin{aligned} D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + \lambda_{\gamma,j}\Big\{ \sum_{l\in\mathcal{S}_{\gamma,j}} |\bar{\gamma}_{j,l}| - \sum_{l\in\mathcal{S}_{\gamma,j}} |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| + \sum_{j\notin\mathcal{S}_{\gamma,j}} |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \Big\} \\ \leq \lambda_{\gamma,j}\Big\{ \sum_{l\in\mathcal{S}_{\gamma,j}} |\bar{\gamma}_{j,l}| + c_0^{-1}\sum_{l\in\mathcal{S}_{\gamma,j}} |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| + c_0^{-1}\sum_{j\notin\mathcal{S}_{\gamma,j}} |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \Big\} \end{aligned}$$

Rearrange to get

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\sum_{l\notin\mathcal{S}_\beta} |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \leq (1 + c_0)^{-1}\lambda_{\gamma,j}\sum_{l\in\mathcal{S}_{\gamma,j}} |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|.$$

Adding $(1 - c_0^{-1})\lambda_{\gamma,j}\sum_{l\in\mathcal{S}_{\gamma,j}}|\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}|$ gives

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq 2\lambda_{\gamma,j}\sum_{l\in\mathcal{S}_{\gamma,j}} |\widehat{\gamma}_{j,l} - \bar{\gamma}_{j,l}| \quad \text{(L.3)}$$

By Lemma 4 in Appendix V.3 of Tan (2017) we have that for $\delta_j := \widehat{\gamma}_j - \bar{\gamma}_j$

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) \geq \frac{1 - e^{-C_0\|\delta_j\|_1}}{C_0\|\widehat{\delta}_j\|} \left(\delta_j' \tilde{\Sigma}_{\gamma,j} \delta_j\right) \tag{L.4}$$

By (L.3) and $\xi_0 > (c_0+1)/(c_0-1)$ we have that $\sum_{l \notin \mathcal{S}_{\gamma,j}} |\delta_{j,l}| \leq \xi_0 \sum_{l \in \mathcal{S}_{\gamma,j}} |\delta_{j,l}|$. Applying the empirical compatability condition from Assumption 2.3.1 to (L.3) then yields

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\|\delta_j\|_1 \leq 2\lambda_{\gamma,j}\nu_0^{-1}|\mathcal{S}_{\gamma,j}|^{1/2}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2} \tag{L.5}$$

Combining (L.4) and (L.5) to get an upper bound on $(\delta_j'\tilde{\Sigma}\delta_j)^{1/2}$ gives

$$\nu_0\|\delta_j\|_2 \leq (\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2} \leq 2\lambda_{\gamma,j}\nu_0^{-1}|\mathcal{S}_{\gamma,j}|^{1/2}\frac{C_0\|\delta_j\|_1}{1 - e^{-C_0\|\delta_j\|_1}}.$$

Plugging the second bound into (L.5) gives

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\|\delta_j\|_1 \leq 2\lambda \sum_{l \in \mathcal{S}_{\gamma,j}} |\delta_{j,l}| \leq 4\lambda_{\gamma,j}^2\nu_0^{-2}|\mathcal{S}_{\gamma,j}|\frac{C_0\|\delta_j\|_1}{1 - e^{-C_0\|\delta_j\|_1}}.$$

The second inequality and $\sum_{l \notin \mathcal{S}_{\gamma,j}} |\delta_{j,l}| \leq \xi_0 \sum_{l \in \mathcal{S}_{\gamma,j}} |\delta_{j,l}|$ imply $1 - e^{-C_0\|\delta_j\|_1} \leq 2C_0\lambda_{\gamma,j}\nu_0^{-2}|\mathcal{S}_{\gamma,j}| \leq \eta$ so,

$$\frac{1 - e^{-C_0\|\delta_j\|_1}}{C_0\|\delta_j\|_1} = \int_0^1 e^{-C_0\|\delta_j\|_1 u}\, du \geq e^{-C_0\|\delta_j\|_1} \geq 1 - \eta.$$

Combining the last two displays gives

$$D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) + (1 - c_0^{-1})\lambda_{\gamma,j}\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq 4\lambda_{\gamma,j}^2\nu_0^{-2}(1 - \eta)|\mathcal{S}_{\gamma,j}| \tag{L.6}$$

Applying $\Omega_{k,2}$ to bound $\lambda_{\gamma,j} \leq \bar{\lambda}_k$ and noting that $|\mathcal{S}_{\gamma,j}| \leq s_k$ by definition gives (2.9.11) with $M_0 = \frac{4\nu_0^{-1}(1-\eta)}{1-c_0^{-1}}$. □

For each j, consider the matrices,

$$\begin{aligned}\tilde{\Sigma}_{\alpha,j} &:= \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}(Y-\bar{\alpha}_j'Z)^2ZZ'] \\ \tilde{\Sigma}_{\gamma,j} &:= \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}ZZ']\end{aligned} \tag{2.9.12}$$

In addition define $\Sigma_{\alpha,j} := \mathbb{E}\tilde{\Sigma}_{\alpha,j}$ and $\Sigma_{\gamma,j} := \mathbb{E}\tilde{\Sigma}_{\gamma,j}$. For the outcome regression model, we will consider nonasymptotic bounds under the following additional events:

$$\begin{aligned}\Omega_{k,3} &:= \{\lambda_{\alpha,j} \geq c_0\|S_{\alpha,j}\|_\infty, \forall j \leq k\} \\ \Omega_{k,4} &:= \{\lambda_{\alpha,j} \leq \bar{\lambda}_k, \forall j \leq k\} \\ \Omega_{k,5} &:= \{\|\tilde{\Sigma}_{\alpha,j} - \Sigma_{\alpha,j}\|_\infty \leq \bar{\lambda}_k, \forall j \leq k\} \\ \Omega_{k,6} &:= \{\|\tilde{\Sigma}_{\gamma,j} - \Sigma_{\gamma,j}\|_\infty \leq \bar{\lambda}_k, \forall j \leq k\}\end{aligned} \tag{2.9.13}$$

Lemma 2.9.6 (Nonasymptotic Bounds for Linear Model)**.** *Suppose that Assumption 2.3.1 holds, $\xi_0 > (c_0+1)/(c_0-1)$, and $2C_0\nu_0^{-2}s_k\bar{\lambda}_k \leq \eta < 1$. In addition, assume there is a constant $c > 0$ such that $\lambda_{\alpha,j}/\lambda_{\gamma,j} \geq c$ for all $j \leq k$. Then, under the event $\bigcap_{m=1}^6 \Omega_{k,m}$ there is a constant M_1 that does not depend on k such that*

$$\max_{1\leq j\leq k} D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \leq M_1 s_k \bar{\lambda}_k^2 \quad \textit{and} \quad \max_{1\leq j\leq k} \|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \leq M_1 s_k \bar{\lambda}_k \tag{2.9.14}$$

Proof. We show that the bound of (2.9.14) holds for each $j = 1, \ldots, k$. We proceed in a few steps.

Step 1: Optimization Step. Let $\tilde{\ell}_j(\alpha; \widehat{\gamma}_j) := \mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j'Z}\{Y-\alpha'Z\}^2]/2$. Optimality of $\widehat{\alpha}_j$ implies that for any $u \in (0,1]$:

$$\tilde{\ell}_j\left(\widehat{\alpha}_j; \widehat{\gamma}_j\right) - \tilde{\ell}_j\left((1-u)\widehat{\alpha}_j + u\bar{\alpha}_j; \widehat{\gamma}_j\right) + \lambda_{\alpha,j}\|\widehat{\alpha}_j\|_1 \leq \lambda_{\alpha,j}\|(1-u)\widehat{\alpha}_j + u\bar{\alpha}_j\|_1.$$

Convexity of the ℓ_1 norm $\|\cdot\|_1$ gives

$$\tilde{\ell}_j\left(\widehat{\alpha}_j;\widehat{\gamma}_j\right)-\tilde{\ell}_j\left((1-u)\widehat{\alpha}_j+u\bar{\alpha}_j;\widehat{\gamma}_j\right)+\lambda_{\alpha,j}u\|\widehat{\alpha}_j\|_1\leq\lambda_{\alpha,j}u\|\bar{\alpha}_j\|_1.$$

Dividing both sides by u and letting $u\to 0^+$ gives:

$$-\mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j'Z}\{Y-\widehat{\alpha}_j'Z\}\{\widehat{\alpha}_j'Z-\bar{\alpha}_j'Z\}]+\lambda_{\alpha,j}\|\widehat{\alpha}_j\|_1\leq\lambda_{\alpha,j}\|\bar{\alpha}_j\|_1.$$

Rearranging using the form of $D^{\ddagger}_{\alpha,j}$ in (2.9.10) yields:

$$D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j,\bar{\alpha}_j;\widehat{\gamma}_j)+\lambda_{\alpha,j}\|\widehat{\alpha}_j\|_1\leq(\widehat{\alpha}_j-\bar{\alpha}_j')\mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}'Z}\{Y-\bar{\alpha}_j'Z\}Z]+\lambda_{\alpha,j}\|\bar{\alpha}_j\|_1 \quad \text{(O.1)}$$

Step 2: Quasi-Score Domination and relating $\bar{\gamma}_j$ to $\widehat{\gamma}_j$. For this step, we will use the fact that we are in the event $\Omega_{k,1}\cap\Omega_{k,2}\cap\Omega_{k,3}\cap\Omega_{k,5}\cap\Omega_{k,6}$. Using the expression for $D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j,\bar{\gamma}_j)$ from (2.9.10) we find that for some $u\in(0,1)$:

$$\begin{aligned}D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j,\bar{\gamma}_j)&=-\mathbb{E}_n[p_j(X)D\{e^{-\widehat{\gamma}_j'Z}-e^{-\bar{\gamma}_j'Z}\}\{\widehat{\gamma}_j'Z-\bar{\gamma}_j'Z\}]\\&=\mathbb{E}_n[p_j(X)De^{-u(\widehat{\gamma}_j-\bar{\gamma}_j)'Z}e^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z-\bar{\gamma}_j'Z\}^2]\end{aligned}$$

where the second step uses the mean value theorem:

$$e^{-\widehat{\gamma}_j'Z}-e^{-\bar{\gamma}_j'Z}=e^{-u\widehat{\gamma}_j'Z-(1-u)\bar{\gamma}_j'Z}(\widehat{\gamma}_j-\bar{\gamma}_j)'Z \quad \text{(O.2)}$$

In the event $\Omega_{k,1}\cap\Omega_{k,2}$ using the bound in Online Appendix Lemma 2.9.5 and the fact that $C_0\nu_0^{-2}s_k\bar{\lambda}_k\leq\eta<1$ gives us that

$$C_0\|\widehat{\gamma}_j-\bar{\gamma}_j\|_1\leq C_0M_0s_k\bar{\lambda}_k\leq M_0\eta. \quad \text{(O.3)}$$

In the event $\Omega_{k,1}\cap\Omega_{k,2}$ the bound in (L.6) also gives us that $D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j,\bar{\gamma}_j)\leq M_0s_k\lambda^2_{\gamma,j}$.

Combining the above displays then yields

$$M_0 s_k \lambda_{\gamma,j}^2 \geq D_{\gamma,j}^{\ddagger}(\widehat{\gamma}_j, \bar{\gamma}_j) \geq e^{M_0 \eta} \mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z} \{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}^2]. \tag{O.4}$$

Again applying the bound on $C_0 \|\widehat{\gamma}_j - \bar{\gamma}_j\|_1$ (O.3) gives

$$\begin{aligned} D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \widehat{\gamma}_j) &= \mathbb{E}_n[p_j(X) D e^{-\widehat{\gamma}_j' Z} (\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z)^2] \\ &= \mathbb{E}_n[p_j(X) D e^{-(\widehat{\gamma}_j - \bar{\gamma}_j)' Z} e^{-\bar{\gamma}_j' Z} (\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z)^2] \\ &\geq e^{-M_0 \eta} D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \end{aligned} \tag{O.5}$$

Decomposing the empirical expectation on the RHS of (O.1) gives

$$\begin{aligned} (\widehat{\alpha}_j - \bar{\alpha}_j)' \mathbb{E}_n[p_j(X) D e^{-\widehat{\gamma}_j' Z} \{Y - \bar{\alpha}_j' Z\} Z] = &\underbrace{(\widehat{\alpha}_j - \bar{\alpha}_j)' \mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z} \{Y - \bar{\alpha}_j' Z\} Z]}_{\delta_{1,j}} \\ &+ \underbrace{\mathbb{E}_n[p_j(X) D \{e^{-\widehat{\gamma}_j' Z} - e^{-\bar{\gamma}_j' Z}\} \{Y - \bar{\alpha}_j' Z\} \{\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z\}]}_{\delta_{2,j}} \end{aligned}$$

By Hölder's inequality, in the event $\Omega_{k,3}$, $\delta_{1,j}$ is bounded

$$\delta_{1,j} \leq c_0^{-1} \|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \lambda_{\alpha,j} \tag{O.6}$$

By the mean value equation (O.2) and the Cauchy-Schwarz inequality, $\delta_{2,j}$ can be bounded from above by

$$\begin{aligned} \delta_{2,j} \leq e^{C_0 \|\widehat{\gamma}_j - \bar{\gamma}_j\|_1} &\times \mathbb{E}_n^{1/2}[p_j(X) D e^{-\bar{\gamma}_j' Z} \{\widehat{\alpha}' Z - \bar{\alpha}' Z\}^2] \\ &\times \mathbb{E}_n^{1/2}[p_j(X) D e^{-\bar{\gamma}_j' Z} \{Y - \bar{\alpha}_j' Z\}^2 \{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}^2] \end{aligned} \tag{O.7}$$

Using (O.3) the first term in (O.7) can be bounded by $e^{M_0 \eta}$. The second term is exactly the square root of $D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)$. The third term is bounded in a few steps. First, in the event

$\Omega_{k,5}$ we have that

$$(\mathbb{E}_n - \mathbb{E})[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}^2\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2.$$

By Assumption 2.3.1 and Lemma 2.10.7 we have that $\mathbb{E}[D\{Y - \bar{\alpha}_j'Z\}^2] \leq G_0^2 + G_1^2$ so that:

$$\mathbb{E}[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}^2\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq (G_0^2 + G_1^2)\mathbb{E}[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2].$$

In the event $\Omega_{k,6}$ we have that

$$(\mathbb{E}_n - \mathbb{E})[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1.$$

and we can bound $\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2]$ using (O.4). Putting this together gives

$$\begin{aligned}\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}^2\{\widehat{\gamma}_j'Z - \bar{\gamma}_j'Z\}^2] \leq \bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\ +(G_0^2 + G_1^2)\bar{\lambda}_k\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2 \\ + (G_0^2 + G_1^2)e^{-M_0\eta}M_0 s_k\lambda_{\gamma,j}^2\end{aligned} \tag{O.8}$$

Applying convexity of $\sqrt{\cdot}$ and the bounds on $\|\widehat{\gamma}_j - \bar{\gamma}_j\|_1^2$ in the event $\Omega_{k,1} \cap \Omega_{k,2}$ from (L.6) gives

$$\begin{aligned}\delta_{2,j} &\leq \{e^{M_0\eta}(1 + (G_0^2 + G_1^2)^{1/2})(M_0\bar{\lambda}_k\lambda_{\gamma,j}s_k)^{1/2} + (G_0^2 + G_1)^2(M_0 s_k\lambda_{\gamma,j}^2)^{1/2}\}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)^{1/2} \\ &\leq \tilde{C}\{(\bar{\lambda}_k\lambda_{\gamma,j}s_k)^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)^{1/2}\end{aligned} \tag{O.9}$$

where $\tilde{C} = \max\{e^{M_0\eta}M_0^{1/2}(1 + G_0 + G_1), (G_0^2 + G_1^2)M_0^{1/2}\}$. Combining (O.6) and (O.9) gives

a bound on the empirical expectation on the RHS of (O.1).

$$(\widehat{\alpha}_j - \bar{\alpha}_j)'\mathbb{E}_n[p_j(X)De^{-\widehat{\gamma}_j'Z}\{Y - \bar{\alpha}_j'Z\}Z] \leq \underbrace{c_0^{-1}\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1\lambda_{\alpha,j}}_{\text{Bound on } \delta_{1,j} \text{ from (O.6)}} + \underbrace{\tilde{C}\{(\bar{\lambda}_k\lambda_{\gamma,j}s_k)^{1/2} + (s_k\lambda_{\gamma,j}^2)^{1/2}\}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)^{1/2}}_{\text{Bound on } \delta_{2,j} \text{ from (O.9)}} \tag{O.10}$$

For convenience, we will sometimes continue to refer to the bound on $\delta_{2,j}$ from (O.9) as simply $\delta_{2,j}$.

Step 3: Express Minimization Constraint in Terms of $\bar{\gamma}_j$ and Simplify. We use the results from *Step 2* to rewrite the minimization bound (O.1) from *Step 1*. Using (O.5) and (O.10) together with the minimization bound (O.1) yields

$$e^{-M_0\eta}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) + \lambda_{\alpha,j}\|\widehat{\alpha}_j\|_1 \leq c_0^{-1}\lambda_{\alpha,j}\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 + \lambda_{\alpha,j}\|\bar{\alpha}_j\|_1 + \delta_{2,j} \tag{O.11}$$

Apply the triangle inequality $|\widehat{\alpha}_{j,l}| \geq |\bar{\alpha}_{j,l}| - |\widehat{\alpha}_{j,l} - \bar{\alpha}_{j,l}|$ for $l \in \mathcal{S}_{\alpha,j}$ and $|\widehat{\alpha}_{j,l}| = |\widehat{\alpha}_{j,l} - \bar{\alpha}_{j,l}|$ for $l \notin \mathcal{S}_{\alpha,j}$ to the above to obtain

$$e^{-M_0\eta}D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) + (1 - c_0^{-1})\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \leq 2\lambda_{\alpha,j}\sum_{l\in\mathcal{S}_{\alpha,j}}|\widehat{\alpha}_{j,l} - \bar{\alpha}_{j,l}| + \delta_{2,j}.$$

Let $\delta_j = \widehat{\alpha}_j - \bar{\alpha}_j$. We use the form $D_{\alpha,j}^{\ddagger}(\widehat{\alpha}_j, \bar{\alpha}_j) = \mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}\{\widehat{\alpha}_j'Z - \bar{\alpha}_j'Z\}^2] = \delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j$ to expand out

$$e^{-M_0\eta}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j) + (1 - c_0^{-1})\lambda_{\alpha,j}\|\delta\|_1 \leq 2\lambda_{\alpha,j}\sum_{l\in\mathcal{S}_{\alpha,j}}|\delta_{j,l}| + \tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2} \tag{O.12}$$

Step 4: Apply Empirical Compatability Condition. Let $\delta_{3,j} := \tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}$

and $D^{\star}_{\alpha,j} := e^{-M_0\eta}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\delta_j\|_1$. In the even $\Omega_{k,1}\cap\Omega_{k,2}\cap\Omega_{k,3}\cap\Omega_{k,5}\cap\Omega_{k,6}$ that (O.12) holds, there are two possibilities. For $\xi_2 = 1 - 2c_0/\{(\xi_1+1)(c_0-1)\} \in (0,1]$ either

$$\xi_2 D^{\star}_{\alpha,j} \leq \delta_{3,j}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2} \tag{O.13}$$

or $(1-\xi_2)D^{\star}_{\alpha,j} \leq 2\lambda_{\alpha,j}\sum_{l\in\mathcal{S}_{\alpha,j}}|\delta_{j,l}|$, that is

$$D^{\star}_{\alpha_j} \leq (\xi_1+1)(c_0-1)c_0^{-1}\lambda_{\alpha,j}\sum_{l\in\mathcal{S}_{\alpha,j}}|\delta_{j,l}| \tag{O.14}$$

We deal with these two cases separately. First, if (O.14) holds, then $\sum_{l\notin\mathcal{S}_{\alpha,j}}|\delta_{j,l}| \leq \xi_1\sum_{l\in\mathcal{S}_{j,l}}|\delta_{j,l}|$. We can apply the empirical compatability of Assumption 2.3.1 to (O.14) to obtain.

$$e^{-M_0\eta}(\delta_j'\tilde{\Sigma}_{\gamma,j}\delta_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\delta_{j,l}\| \leq \nu_1(\xi_1+1)(\xi_1-1)\lambda_{\alpha,j}(s_j\delta_j\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2}.$$

Inverting for $(\delta_j\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2}$ and plugging in gives

$$e^{-M_0\eta}D^{\ddagger}_{\alpha,j}(\widehat{\alpha},\bar{\alpha}_j;\bar{\gamma}_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\widehat{\alpha}_j-\bar{\alpha}_j\|_1 \leq \tilde{M}s_k\lambda^2_{\alpha,j} \tag{O.15}$$

where $\tilde{M} = e^{M_0\eta}(\xi_1+1)(c_0-1)c_0^{-1}$. Next, assume that (O.13) holds. In this case, we can directly invert for $(\delta_j\tilde{\Sigma}_{\gamma,j}\delta_j)^{1/2}$ to get that

$$e^{-M_0\eta}D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j,\bar{\alpha}_j;\bar{\gamma}_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\widehat{\alpha}_j-\bar{\alpha}_j\|_1 \leq \xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda^2_{\gamma,j})^{1/2}\}^2 \tag{O.16}$$

Combining (O.15) and (O.16) gives

$$\begin{aligned} e^{-M_0\eta}D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j,\bar{\alpha}_j;\bar{\gamma}_j) + (1-c_0^{-1})\lambda_{\alpha,j}\|\widehat{\alpha}_j-\bar{\alpha}_j\|_1 &\leq \tilde{M}s_k\lambda^2_{\alpha,j} \\ &\quad + \xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda^2_{\gamma,j})^{1/2}\}^2 \end{aligned} \tag{O.17}$$

Step 5: Apply Penalty Majorization and Bounded Penalty Ratio. Use the fact that $\lambda_{\gamma,j}/\lambda_{\alpha,j} \leq c^{-1}$ to express (O.17) as

$$D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \leq e^{M_0\eta}\tilde{M}s_k\lambda^2_{\alpha,j} + e^{M_0\eta}\xi_2^{-1}\tilde{C}\{(s_k\bar{\lambda}_k\lambda_{\gamma,j})^{1/2} + (s_k\lambda^2_{\gamma,j})^{1/2}\}^2$$
$$\|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \leq (1-c_0^{-1})^{-1}\tilde{M}s_k\lambda_{\alpha,j} + (1-c_0^{-1})^{-1}c^{-1}\tilde{C}\{(s_k\bar{\lambda}_k)^{1/2} + (s_k\lambda_{\gamma,j})^{1/2}\}^2$$

In the event $\Omega_{k,2} \cap \Omega_{k,3}$ we have that $\lambda_{\gamma,j} \vee \lambda_{\alpha,j} \leq \bar{\lambda}_k$, so that the above simplifies to

$$\begin{aligned} D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) &\leq M_1 s_k \bar{\lambda}^2_k \\ \|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 &\leq M_1 s_k \bar{\lambda}_k \end{aligned} \tag{O.18}$$

for $M_1 = \max\{e^{M_0\eta}, c^{-1}(1-c_0^{-1})^{-1}\}(\tilde{M} + 2e^{M_0\eta}\xi_2^{-1}\tilde{C})$. This completes the result (2.9.14). □

Nonasymptotic Bounds for Residual Estimation

We now provide nonasymptotic bounds on the empirical mean square error between the estimated residuals $\widehat{U}_{\gamma,j}$ and $\widehat{U}_{\alpha,j}$ and the true residuals

$$\begin{aligned} U_{\gamma,j} &:= -p_j(X)\{De^{-\bar{\gamma}_j'Z} + (1-D)\} \\ U_{\alpha,j} &:= p_j(X)De^{-\bar{\gamma}_j'Z}(Y - \bar{\alpha}_j^{\text{pilot}'}Z), \end{aligned} \tag{2.9.15}$$

These bounds will be shown under the events in (2.9.9), (2.9.13), and (2.9.1) using the results in Lemmas 2.9.5 and 2.9.6.

Lemma 2.9.7 (Nonasymptotic Logistic Residual Bound)**.** *Suppose that Assumption 2.3.1 and the conditions of Lemma 2.9.5 hold. Then, in the event $\Omega_{k,1} \cap \Omega_{k,2}$ described on (2.9.9) there is a constant $M_{\gamma,r}$ that does not depend on k such that:*

$$\max_{1\leq j\leq k} \mathbb{E}_n[(\widehat{U}_{\gamma,j} - U_{\gamma,j})^2] \leq M_{\gamma,r}\xi_{k,\infty}s_k\bar{\lambda}^2_k. \tag{2.9.16}$$

Proof. Consider each j separately. By applying the mean value theorem (O.2) and Lemma 2.9.5,

we can write

$$
\begin{aligned}
(\widehat{U}_{\gamma,j} - U_{\gamma,j})^2 &= p_j(X)^2 D\{e^{-\widehat{\gamma}_j' Z} - e^{-\bar{\gamma}_j' Z}\}\{e^{-\widehat{\gamma}_j' Z} - e^{-\bar{\gamma}_j' Z}\} \\
&\leq \xi_{k,\infty} p_j(X) D\{e^{-\widehat{\gamma}_j' Z} - e^{-\bar{\gamma}_j' Z}\} e^{-\bar{\gamma}_j' Z - u(\widehat{\gamma}_j - \bar{\gamma}_j)' Z}\{\bar{\gamma}_j' Z - \widehat{\gamma}_j' Z\} \\
&\leq \xi_{k,\infty} e^{-B_0 + M_0 \eta} D\{e^{-\widehat{\gamma}_j' Z} - e^{-\bar{\gamma}_j' Z}\}\{\bar{\gamma}_j' Z - \widehat{\gamma}_j' Z\}
\end{aligned}
$$

So that

$$
\begin{aligned}
\mathbb{E}_n[(\widehat{U}_{\gamma,j} - U_{\gamma,j})^2] &\leq e^{-B_0 + M_0 \eta} \xi_{k,\infty} \underbrace{\mathbb{E}_n[p_j(X) D\{e^{-\widehat{\gamma}_j' Z}\}\{\widehat{\gamma}_j' Z - \bar{\gamma}_j' Z\}]}_{= D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j)} \\
&\leq e^{-B_0 + M_0 \eta} \xi_{k,\infty} s_k \bar{\lambda}_k^2
\end{aligned}
$$

□

Lemma 2.9.8 (Nonasymptotic Linear Residual Bound)**.** *Suppose that Assumption 2.3.1 and the conditions of Lemma 2.9.6 hold. Then, in the event* $\bigcap_{m=1}^{6} \Omega_{k,m}$*, there is a constant* $M_{\alpha,r}$ *that does not depend on* k *such that*

$$
\max_{1 \leq j \leq k} \mathbb{E}_n[(\widehat{U}_{\alpha,j} - U_{\alpha,j})^2] \leq M_{\alpha,r} \xi_{k,\infty}^2 s_k^2 \bar{\lambda}_k^2 \tag{2.9.17}
$$

Proof. Recall that $\widehat{U}_{\alpha,j} = p_j(X) D e^{-\widehat{\gamma}_j' Z}(Y - \widehat{\alpha}_j' Z)$ and $U_{\alpha,j} = p_j(X) D e^{-\bar{\gamma}_j' Z}(Y - \bar{\alpha}_j' Z)$. As an intermediary, define $\dot{U}_{\gamma,j} = p_j(X) D e^{-\widehat{\gamma}_j' Z}(Y - \bar{\alpha}_j' Z)$. We will show a bound on the empirical mean square error between $\widehat{U}_{\alpha,j}$ and $\dot{U}_{\alpha,j}$ as well as on the empirical mean square error between $\dot{U}_{\alpha,j}$ and $U_{\alpha,j}$. The bound in (2.9.17) will then follow from $(a+b)^2 \leq 2a^2 + 2b^2$.

First consider $(\widehat{U}_{\alpha,j} - \dot{U}_{\alpha,j})^2$:

$$
\begin{aligned}
\mathbb{E}_n[(\widehat{U}_{\alpha,j} - \bar{U}_{\alpha,j})^2] &= \mathbb{E}_n p_j^2(X) D e^{-2\widehat{\gamma}_j' Z}(\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z)^2] \\
&= \mathbb{E}_n[p_j^2(X) D e^{-2(\bar{\gamma}_j' Z - (\widehat{\gamma}_j - \bar{\gamma}_j)' Z)}(\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z)^2]
\end{aligned}
$$

$$\leq \xi_{k\infty} e^{-B_0} e^{2M_0\eta} \underbrace{\mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z}(\widehat{\alpha}_j' Z - \bar{\alpha}_j' Z)]}_{= D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j)}$$

$$\leq e^{2M_0\eta - B_0} M_1 \xi_{k,\infty} s_k \bar{\lambda}_k^2$$

Where the last empirical expectation is bounded by Lemma 2.9.6. Next, consider $(\dot{U}_{\alpha,j} - U_{\alpha,j})^2$:

$$\begin{aligned}
\mathbb{E}_n[(\dot{U}_{\alpha,j} - U_{\alpha,j})^2] &= \mathbb{E}_n[p_j^2(X) D \{e^{-\widehat{\gamma}' Z} - e^{-\bar{\gamma}' Z}\}^2 \{Y - \bar{\alpha}_j' Z\}^2] \\
&= \mathbb{E}_n[p_j^2(X) D \{e^{-\bar{\gamma}' Z - u(\widehat{\gamma} - \bar{\gamma})' Z} (\bar{\gamma}_j' Z - \widehat{\gamma}_j' Z)\}^2 (Y - \bar{\alpha}_j' Z)^2] \\
&\leq 2 e^{M_0\eta - B_0} C_0^2 \xi_{k,\infty} (M_1 s_k \bar{\lambda}_k)^2 \mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z} (Y - \bar{\alpha}_j' Z)^2]
\end{aligned}$$

To proceed we assume that Z contains a constant. That is $Z = (1, Z_2, \ldots, Z_{d_z})$. However, this is not necessary it just simplifies the proof a bit. We bound the final empirical expectation in the event $\Omega_{k,5}$. In this event we can bound

$$\begin{aligned}
&\mathbb{E}_n[p_j(X) D e^{-\bar{\gamma}_j' Z} (Y - \bar{\alpha}_j' Z)^2] \\
&\qquad = (\mathbb{E}_n - \mathbb{E})[p_j(X) D e^{-\bar{\gamma}_j' Z} (Y - \bar{\alpha}_j' Z)^2] + \mathbb{E}[p_j(X) D e^{-\bar{\gamma}_j' Z} (Y - \bar{m}_j(X))^2] \\
&\qquad \leq \bar{\lambda}_k + \xi_{k,\infty} e^{-B_0} (D_0 + D_1)^2.
\end{aligned}$$

Combining the above, and using the fact that $s_k \bar{\lambda}_k \leq \eta < 1$ completes the reult.

□

Probability Bounds for the First Stage

In this section we establish that each of the events in (2.9.9), (2.9.13), and (2.9.1) occurs under Assumption 2.3.1 with probability approaching one.

Lemma 2.9.9 (Logistic Score Domination and Penalty Majorization)**.** *Suppose Assumption 2.3.1 holds and that the penalty parameter $\lambda_{\gamma,j}$ is chosen as described in Section 2.2.*

Then, for n sufficiently large, the event $\Omega_{k,1}$ holds with probability $1-\epsilon-\rho_{\gamma,n}$ where

$$\rho_{\gamma,n} = C\max\left\{\frac{4kn+4k}{n^2}, \left(\frac{\tilde{M}\xi_{k,\infty}s_{k,\gamma}\bar{c}_n^2\ln^5(d_z n)}{n}\right)^{1/2}, \left(\frac{\tilde{M}\xi_{k,\infty}^4\ln^7(d_z kn)}{n}\right)^{1/6}, \frac{1}{\ln^2(d_z kn)}\right\}. \tag{2.9.18}$$

where $C, \tilde{M}$ are absolute constants that do not depend on k. In particular so long as $\epsilon \to 0$ as $n \to \infty$, this shows that $\Pr(\Omega_{k,1}) = 1 - o(1)$ under the rate conditions of Assumption 2.3.1.

Moreover, with probability at least $1 - \frac{5k}{n} - \frac{4k}{n^2}$ there is a constant M_2 that does not depend on k such that $\Omega_{k,2}$ holds with

$$\bar{\lambda}_k = \max\{M_2, M_4, M_5, M_6, M_7\}\xi_{k,\infty}\sqrt{\frac{\ln(d_z n)}{n}} \tag{2.9.19}$$

where M_4, M_5, M_6 and M_7 are all constants that also do not depend on k described in Lemma 2.9.10 and Lemmas 2.9.11-2.9.13. In particular, so long as $k/n \to 0$, $\Pr(\Omega_{k,2}) = 1 - o(1)$.

Proof. Collecting the logistic nonasymptotic residual bound from Lemma 2.9.7 and the probability bounds from Lemmas 2.9.11–2.9.14 we find that, (eventually) with probability at least $1 - \frac{4k}{n} - \frac{4k}{n^2}$:

$$\max_{\substack{1\le j\le k\\ 1\le l\le d_z}} \mathbb{E}_n[(\widehat{U}_{\gamma,j}Z_l - U_{\gamma,j}Z_l)^2] \le M_{\gamma,r}C_0^2\frac{\xi_{k,\infty}s_{k,\gamma}\bar{c}_n^2\ln^3(d_z n)}{n}. \tag{P.1}$$

where $M_{\gamma,r}$ is a constant that does not depend on k. Define the vectors

$$\begin{aligned} W_k &:= (U_{\gamma,1}Z', \ldots, U_{\gamma,k}Z')' \in \mathbb{R}^{kd_z} \\ &:= (W_{k,1}', \ldots, W_{k,k}')' \\ \widehat{W}_k &:= (\widehat{U}_{\gamma,1}Z', \ldots, \widehat{U}_{\gamma,k}Z')' \in \mathbb{R}^{kd_z} \\ &:= (\widehat{W}_{k,1}', \ldots, \widehat{W}_{k,k}')'. \end{aligned}$$

Notice by optimality of $\bar{\gamma}_1, \ldots, \bar{\gamma}_k$ that W_k is a mean zero vector. Under our assumptions the covariance matrix $\Sigma_k = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[W_k W_k']$ exists and is finite. Define the sequences of constants

$$\begin{aligned}\delta_{\gamma,n}^2 &:= M_{\gamma,r} C_0^2 \xi_{k,\infty} s_{k,\gamma} \bar{c}_n^2 \ln^5(d_z n)/n \\ \beta_{\gamma,n} &:= \frac{4k}{n} + \frac{4k}{n^2}\end{aligned}$$

Then, by (P.1) we have that with probability at least $1 - \beta_{\gamma,n}$

$$\Pr\left(\|\mathbb{E}_n[(\widehat{W}_k - W_k)^2]\|_\infty > \delta_n^2 / \ln^2(d_z n)\right) \leq \beta_n. \tag{P.2}$$

Let $e_1, \ldots, e_n$ be i.i.d normal random variables generated independently of the data. Define the scaled random variables and the multiplier bootstrap process

$$\begin{aligned}\widehat{S}_{\gamma,n}^e &:= n^{-1/2} \sum_{i=1}^n e_i \widehat{W}_{k,i} \\ &:= (\widehat{S}_{\gamma,1}^{e\prime}, \ldots, \widehat{S}_{\gamma,k}^{e\prime})'\end{aligned}$$

and let $\Pr_e$ denote the probability measure with respect to the $e_i's$ conditional on the observed data. Assumption 2.3.1 implies that the conditions of (2.9.22) hold for $Z = W_k$ with b replaced by c_u and B_n replaced by $B_k = (\xi_{k,\infty} C_0 C_U)^3 \vee 1$. Further, via (P.2) the residual estimation requirement of with δ_n and β_n replaced by $\delta_{\gamma,n}$ and $\beta_{\gamma,n}$.

Let $\widehat{q}_{\gamma,j}(\alpha)$ be the α quantile of $\|\widehat{S}_{\gamma,j}^{e\prime}\|$ *conditional* on the data Z_i and the estimates $\widehat{Z}_i$. Theorem 2.9.4 then shows that there is a finite constant depending only on c_u such that

$$\max_{1\leq j\leq k} \sup_{\alpha\in(0,1)} \left|\Pr(\|S_{\gamma,j}\| \geq \widehat{q}_{\gamma,j}(\alpha)) - \alpha\right| \leq C \max\left\{\beta_{\gamma,n}, \delta_{\gamma,n}, \left(\frac{B_k^4 \ln^7(k d_z n)}{n}\right)^{1/6}, \frac{1}{\ln^2(k d_z n)}\right\}.$$

This gives the first claim of Lemma 2.9.9 by construction of $\lambda_{\gamma,j}$. The second claim follows Lemma 2.9.16. For this second claim we will consider the marginal convergence of each $U_{\gamma,j} Z$

as opposed to their joint convergence (the convergence of W_k). First, notice that condiitonal on the data, the random vector $\mathbb{E}_n[e\widehat{U}_{\gamma,j}Z]$ is centered gaussian in $\mathbb{R}^{d_z}$. Lemma 2.9.16 then shows that

$$\widehat{q}_{\gamma,j}(\epsilon) \leq (2+\sqrt{2})\sqrt{\frac{\ln(d_z/\epsilon)}{n}\max_{1\leq l\leq d_z}\mathbb{E}_n[\widehat{U}^2_{\gamma,j}Z_l^2]}.$$

Furthermore, with probability at least $1-\beta_{\gamma,n}-\frac{1}{n}$ we have that, for all $j=1,\ldots,k$:

$$\begin{aligned}\max_{1\leq l\leq d_z}\mathbb{E}_n[\widehat{U}^2_{\gamma,j}Z_l^2] &\leq C_0^2\mathbb{E}_n[\widehat{U}^2_{\gamma,j}] \\ &\leq 2C_0^2(\mathbb{E}_n[U^2_{\gamma,j}]+\mathbb{E}_n[(\widehat{U}^2_{\gamma,j}-U_{\gamma,j})^2]) \leq 4C_0^2\xi^2_{k,\infty}C_U^2+\delta^2_{\gamma,n}/\ln^2(d_zn))\end{aligned}$$

Under the rate conditions of Assumption 2.3.1, $\delta^2_{\gamma,n}/\ln^2(d_zn)$ will eventually be smaller than 1 and so the claim in (2.9.19) holds with $M_2 = 8C_0^2C_U^2 \vee 1$. □

Lemma 2.9.10 (Linear Score Domination and Penalty Majorization)**.** *Suppose Assumption 2.3.1 holds and that the penalty parameters $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$ are chosen as described in Section 2.2. Then, for n sufficiently large, the event $\Omega_{k,3}$ holds with probability $1-\epsilon-\rho_{\alpha,n}$ where:*

$$\rho_{\alpha,n} = C\max\left\{\frac{4kn+4k}{n^2}, \left(\frac{\tilde{M}\xi^2_{k,\infty}s^2_{k,\alpha}\bar{c}_n^2\ln^5(d_zn)}{n}\right)^{1/2}, \left(\frac{\tilde{M}\xi^4_{k,\infty}\ln^7(d_zkn)}{n}\right)^{1/6}, \frac{1}{\ln^2(d_zkn)}\right\}. \tag{2.9.20}$$

where $C, \tilde{M}$ are absolute constants that do not depend on k. In particular so long as $\epsilon \to 0$ as $n\to\infty$, this shows that $\Pr(\Omega_{k,3}) = 1-o(1)$ under Assumption 2.3.1.

Moreover, with probability at least $1-\frac{5k}{n}-\frac{4k}{n^2}$ there is a constant M_4 that does not depend on k such that $\Omega_{k,4}$ holds with

$$\bar{\lambda}_k = \max\{M_2, M_4, M_5, M_6, M_7\}\xi_{k,\infty}\sqrt{\frac{\ln(d_zn)}{n}} \tag{2.9.21}$$

where M_2, M_5, M_6 and M_7 are all constants that also do not depend on k described in

Lemma 2.9.9 and Lemmas 2.9.11-2.9.13. In particular, so long as $k/n \to 0$, $\Pr(\Omega_{k,4}) = 1 - o(1)$.

Proof. Apply the same steps as the proof of Lemma 2.9.9 with

$$\delta_{\alpha,n}^2 = M_{\alpha,r} C_0^2 \xi_{k,\infty}^2 s_k^2 \bar{c}_n^2 \ln^5(d_z n)/n$$
$$\beta_{\alpha,n} = \frac{4}{n} + \frac{4}{n^2}$$

□

Lemma 2.9.11 (Probabilistic Bound on $\Omega_{k,5}$)**.** *Let* $\tilde{\Sigma}_{\alpha,j}$ *and* $\Sigma_{\alpha,j} = \mathbb{E}\tilde{\Sigma}_{\alpha,j}$ *be as in* (2.9.12). *Under Assumption 2.3.1 if*

$$\bar{\lambda}_k \geq 4\xi_{k,\infty}(G_0^2 + G_0 G_1) C_0^2 \left[G_0^2 \log(d_z/\epsilon)/n + G_0 G_1 \sqrt{\log(d_z/\epsilon)/n} \right]$$

Then $\Pr(\Omega_{k,5}) \geq 1 - 2k\epsilon^2$. *In particular, there is a constant* M_5 *that does not depend on* k, *such that if* $\bar{\lambda}_k \geq \xi_{k,\infty} M_5 \sqrt{\log(d_z/\epsilon)/n}$ *and* $k\epsilon^2 \to 0$ *as* $n \to \infty$ *then under the conditions of Assumption 2.3.1,* $\Pr(\Omega_{k,5}) = 1 - o(1)$.

Proof. We show that this happens with probability $1 - 2\epsilon^2$ for each $j = 1, \ldots, k$. For any $l, h = 1, \ldots, d_z$, the variable

$$p_j(X) e^{-\bar{\gamma}' Z} D\{Y - \bar{m}_j(Z)\}^2 Z_l Z_h$$

is the product of $p_k(X) e^{-\bar{\gamma}_j' Z} Z_l Z_h$, which is bounded in absolute value by $\xi_{k,\infty} C_0^2 e^{-B_0}$, and $D\{Y - \bar{m}_j(Z)\}$, which is uniformly sub-gaussian conditional on Z. By Lemma 2.10.8 we have:

$$\mathbb{E}\left[|(\tilde{\Sigma}_{\alpha,j})_{lh} - (\tilde{\Sigma}_{\alpha,j})_{lh}|^k \right] \leq \frac{k!}{2} (2\xi_{k,\infty} C_0^{-2} e^{-B_0} G_0^2)^{k-2} (2\xi_{k,\infty} C_0^2 e^{-B_0} G_0 G_1)^2, \quad k = 2, 3, \ldots.$$

Apply the above and Lemma 2.10.6 with $t = \log(d_z^2/\epsilon^2)/n$ to obtain

$$\Pr\left(|(\tilde{\Sigma}_{\alpha,j})_{lh} - (\tilde{\Sigma}_{\alpha,j})_{lh}| > 2e^{-B_0}\xi_{k,\infty}C_0^2G_0^2t + 2e^{-B_0}\xi_{k,\infty}C_0^2G_0G_1\sqrt{2t}\right) \leq 2\epsilon^2/d_z^2.$$

A union bound completes the argument. □

Lemma 2.9.12 (Probabilistic Bound on $\Omega_{k,6}$)**.** *Let $\tilde{\Sigma}_{\gamma,j}$ and $\Sigma_{\gamma,j} = \mathbb{E}\tilde{\Sigma}_{\gamma,j}$ be as in (2.9.12). Under Assumption 2.3.1 if*

$$\bar{\lambda}_k \geq \xi_{k,\infty}\sqrt{2}(e^{-B_0}+1)C_0\sqrt{\log(d_z/\epsilon)/n},$$

then $\Pr(\Omega_{k,6}) \leq 1 - 2k\epsilon^2$. In particular, there is a constant M_6 that does not depend on k, such that if $\bar{\lambda}_k \geq \xi_{k,\infty}M_6\sqrt{\log(d_z/\epsilon)/n}$ and $k\epsilon^2 \to 0$ as $n \to \infty$ then under the conditions of Assumption 2.3.1, $\Pr(\Omega_{k,6}) = 1 - o(1)$.

Proof. Consider each j separately. For any $l, h = 1, \ldots, d_z$, note $|(\tilde{\Sigma}_{\gamma,j})_{lh}| = |p_j(X)De^{-\bar{\gamma}_j'Z}Z_lZ_h| \leq \xi_{k,\infty}C_0^2e^{-B_0}$ so that $(\tilde{\Sigma}_{\gamma,j})_{lh} - (\Sigma_{\gamma,j})_{lh}$ is mean zero and bounded in abosulte values by $2\xi_{k,\infty}C_0^2e^{-B_0}$. Applying Lemma 2.10.4 with $\bar{\lambda}_k \geq 4\xi_{k,\infty}C_0^2e^{-B_0}\sqrt{\log(d_z/\epsilon)/n}$ yields:

$$\Pr\left(|(\tilde{\Sigma}_{\gamma,j})_{lh} - (\Sigma_{\gamma,j})_{lh}| \geq \bar{\lambda}_k\right) \leq 2\epsilon^2/d_z^2.$$

A union bound completes the argument. □

Lemma 2.9.13 (Probabilitstic Bound on $\Omega_{k,7}$)**.** *Let $\tilde{\Sigma}^1_{\alpha,j}$ and $\Sigma^1_{\alpha,j} = \mathbb{E}\tilde{\Sigma}^1_{\alpha,j}$ be as in (2.9.1). Under Assumption 2.3.1 if*

$$\bar{\lambda}_k \geq \xi_{k\infty}4(G_0^2+G_1^2)^{1/2}e^{-B_0}C_0^2\sqrt{\log(d_z/\epsilon)/n},$$

then $\Pr(\Omega_{k,7}) \geq 1 - 2k\epsilon^2$. In particular, there is a constant M_7 that does not depend on k such that if $\bar{\lambda}_k \geq \xi_{k,\infty}M_7\sqrt{\log(d_z/\epsilon)/n}$ and $k\epsilon^2 \to 0$ as $n \to \infty$ then, under the conditions of

Assumption 2.3.1, $\Pr(\Omega_{k,7}) \geq 1 - o(1)$.

Proof. We deal with each j term separately. The variables $p_j(X)e^{-\bar{\gamma}_j' Z}|Y - \bar{m}_j(Z)|Z_l Z_h$ are uniformly sub-gaussian conditional on Z because $|p_j(X)e^{-\bar{\gamma}_j' Z} Z_l Z_h| \leq \xi_{k,\infty} e^{-B_0} C_0^2$ and $D|Y - \bar{m}_j(Z)|$ is uniformly sub-gaussian. Applying Lemma 2.10.5 for

$$\bar{\lambda}_k \geq e^{-B_0} \xi_{k,\infty} C_0^2 \sqrt{8(G_0^2 + G_1)^2} \sqrt{\log(d_z/\epsilon)/n}$$

yields

$$\Pr\left(|(\tilde{\Sigma}_{\gamma,j})_{lh} - (\Sigma_{\gamma,j})_{lh}| \geq \bar{\lambda}_k\right) \leq 2\epsilon^2/d_z^2.$$

A union bound completes the argument. □

Probability Bounds for Residual Estimation

For showing consistent residual estimation, we employ the following two lemmas.

Lemma 2.9.14 (Deterministic Logistic Score Domination)**.** *Under Assumption 2.3.1 let*

$$\bar{\lambda}_k \geq \xi_{k,\infty} \sqrt{2}(e^{-B_0} + 1) C_0 \sqrt{\ln(d_z/\epsilon)/n}.$$

Then if for all $j = 1, \ldots, k$ we let $\lambda_{\gamma,j} \equiv \bar{\lambda}_k$, $\Pr(\Omega_{k,1} \cap \Omega_{k,2}) \geq 1 - 2k\epsilon$. In particular, there is a constant M_8^p that does not depend on k such that if $\bar{\lambda}_k \geq M_8^p \xi_{k,\infty} \sqrt{\ln(d_z n)/n}$ $\Pr(\Omega_{k,1} \cap \Omega_{k,2}) \geq 1 - 2k/n^p$.

Proof. Let us recall that

$$\|S_j\|_\infty = \max_{1 \leq l \leq d_z} |\mathbb{E}_n[p_j(X)\{-De^{-\bar{\gamma}_j' Z} + (1 - D)\} Z_j]|.$$

Notice for each $1 \leq l \leq d_z$, $S_{j,l} = p_j(X)\{-De^{-\bar{\gamma}_j' Z} + (1 - D)\} Z_l$ is bounded in absolute value by $C_0 \xi_{k,\infty}(e^{-B_0} + 1)$ and is mean zero by optimality of $\bar{\gamma}_j$. For $\bar{\lambda}_k \geq 2(e^{-B_0} + 1) C_0 \sqrt{\ln(d_z/\epsilon)/n}$

apply Lemma 2.10.4 to see the result. □

Lemma 2.9.15 (Deterministic Linear Score Domination)**.** *Under Assumption 2.3.1 let*

$$\bar{\lambda}_k \geq \xi_{k,\infty}(e^{-B_0}C_0)\sqrt{8(G_0^2+G_1^2)}\sqrt{\ln(d_z/\epsilon)/n}.$$

Then if for all $j = 1,\ldots,k$ *we let* $\lambda_{\gamma,j} \equiv \bar{\lambda}_k$, $\Pr(\Omega_{k,3}\cap\Omega_{k,4}) \geq 1-2k\epsilon$. *In particular, there is a constant* M_9^p *that does not depend on* k *such that if* $\bar{\lambda}_k \geq M_9^p\xi_{k,\infty}\sqrt{\ln(d_z n)/n}$, $\Pr(\Omega_{k,3}\cap\Omega_{k,4}) \geq 1-2k/n^p$.

Proof. Notice $S_{j,l} = p_j(X)De^{-\bar{\gamma}_j'Z}\{Y-\bar{m}_j(Z)\}Z_l$ for $l=1,\ldots,p$. By optimality of $\bar{\alpha}_j$, $S_{j,l}$ is mean zero. Under Assumption 2.3.1, $|S_{j,l}| \leq e^{-B_0}C_0|D\{Y-\bar{m}_j(Z)\}|$ so by Assumption 2.3.1 the variables $S_{j,l}$ are uniformly sub-gaussian conditional on Z in the following sense:

$$\max_{l=1,\ldots,p}\tilde{G}_0^2\mathbb{E}[\exp(S_{j,l}^2/\tilde{G}_0^2)-1] \leq \tilde{G}_1^2$$

for $\tilde{G}_0 = \xi_{k,\infty}C_0G_0e^{-B_0}$ and $\tilde{G}_1 = \xi_{k,\infty}C_0G_1e^{-B_0}$. Apply Lemma 2.10.5 for $\bar{\lambda}_k$ defined above in the statement of Lemma 2.9.15 and union bound to obtain the result. □

2.9.4. Supporting High Dimensional Probability Results

High Dimensional Central Limit and Bootstrap Theorems

Lemma 2.9.16 (Gaussian Quantile Bound)**.** *Let* $Y = (Y_1,\ldots,Y_p)$ *be centered Gaussian in* $\mathbb{R}^p$ *with* $\sigma^2 \leq \max_{1\leq j\leq p}\mathbb{E}[Y_j^2]$ *and* $\rho \geq 2$. *Let* $q^Y(1-\epsilon)$ *denote the* $(1-\epsilon)$*-quantile of* $\|Y\|_\infty$ *for* $\epsilon\in(0,1)$. *Then* $q^Y(1-\epsilon) \leq (2+\sqrt{2})\sigma\sqrt{\ln(p/\epsilon)}$.

Proof. See Chetverikov and Sørensen (2021), Lemma D.2. □

Now let $Z_1,\ldots,Z_n$ be independent, mean zero random variables in $\mathbb{R}^p$, and denote their

scaled average and variance by

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \quad \text{and} \quad \Sigma := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Z_i Z_i'].$$

For $\mathbb{R}^p$ values random variables U and V, define the distributional measure of distance

$$\rho(U, V) := \sup_{A \in \mathcal{A}_p} \left| \Pr(U \in A) - \Pr(V \in A) \right|$$

where $\mathcal{A}_p$ denotes the collection of all hyperrectangles in $\mathbb{R}^p$. For any symmetric positive matrix $M \in \mathbb{R}^{p \times p}$, write $N_M := N(\mathbf{0}, M)$.

Theorem 2.9.1 (High-Dimensional CLT)**.** *If, for some finite constants $b > 0$ and $B_n \geq 1$,*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Z_{ij}^2] \geq b, \ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[|Z_{ij}|^{2+k}] \leq B_n^k \ \text{ and } \ \mathbb{E}\left[\max_{1 \leq j \leq p} Z_{ij}^4 \right] \leq B_n^4. \tag{2.9.22}$$

for all $i \in \{1, \ldots, n\}, j \in \{1, \ldots, p\}$ and $k \in \{1, 2\}$, then there exists a finite constant C_b, depending only on b, such that:

$$\rho(S_n, N_\Sigma) \leq C_b \left(\frac{B_n^4 \ln^7(pn)}{n} \right)^{1/6}.$$

Proof. See Chernozhukov et al. (2017), Proposition 2.1. □

Let $\widehat{Z}_i$ be an estimator of Z_i and let $e_1, \ldots, e_n$ be i.i.d $N(0, 1)$ and independent of both the Z_i's and $\widehat{Z}_i$'s. Define $\widehat{S}_n^e := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i \widehat{Z}_i$ and let $\Pr_e$ denote the conditional probability measure computed with respect to the $e_i's$ for fixed Z_i's and $\widehat{Z}_i$'s. Also abbreviate

$$\tilde{\rho}(\widehat{S}_n^e, N_\Sigma) := \sup_{A \in \mathcal{A}_p} \left| \Pr_e\left(\widehat{S}_n^e \in A \right) - \Pr\left(N_\Sigma \in A \right) \right|.$$

Theorem 2.9.2 (Multiplier Bootstrap for Many Approximate Means)**.** *Let* (2.9.22) *hold*

for some finite constants $b > 0$ and $B_n \geq 1$, and let $\{\beta_n\}_{\mathbb{N}}$ and $\{\delta_n\}_{\mathbb{N}}$ be sequences in $\mathbb{R}_{++}$ converging to zero such that

$$\Pr\left(\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n}(\widehat{Z}_{ij}-Z_{ij})^2 > \frac{\delta_n^2}{\ln^2(pn)}\right)\leq \beta_n \tag{2.9.23}$$

Then, there exists a finite constant C_b depending only on b such that with probability at least $1-\beta_n - 1/\ln^2(pn)$,

$$\tilde{\rho}(\widehat{S}_n^e, N_\Sigma) \leq C_b \max\left\{\delta_n, \left(\frac{B_n \ln^6(pn)}{n}\right)^{1/6}\right\}.$$

Proof. See Belloni et al. (2018), Theorem 2.2 or Chetverikov and Sørensen (2021) Theorem D.2. □

We now consider a partition of Z and $\widehat{Z}$ into k subvectors.

$$Z := (Z_1', \ldots, Z_k')' \in \mathbb{R}^{d_1,\ldots,d_k} \quad \text{and} \quad \widehat{Z} := (\widehat{Z}_1', \ldots, Z_k')' \in \mathbb{R}^{d_1,\ldots,d_k}$$

where $\sum_{j=1}^k d_j = p$. Given such a partition, for any symmetric, positive definite $M \in \mathbb{R}^{p\times p}$ let $N_{M,j}$ denote the marginal distribution of the subvector of N_M corresponding the the indices of partition j. In other words, N_{M_1} would denote the marginal distribution of the first d_1 elements of an $\mathbb{R}^p$ vector with distribution N_M, N_2 would denote the marginal distribution of the next d_2 elements and so on. For each $j = 1, \ldots, k$ define $q_{M,j}^N : \mathbb{R} \to \bar{\mathbb{R}}$ as the (extended) quantile function of $\|N_{M,j}\|_\infty$,

$$q_{M,j}^N(\epsilon) := \inf\left\{t \in \mathbb{R} : \Pr(\|N_{M,j}\|_\infty \leq t) \geq \epsilon\right\}.$$

Define $q_{M,j}^N(\epsilon) = +\infty$ if $\epsilon \geq 1$ and $-\infty$ if $\epsilon \leq 0$ so that $q_{M,j}^N$ is always montone (strictly) increasing.

Lemma 2.9.17. *Let $M \in \mathbb{R}^{p\times p}$ be symmetric positive definite, let U be a random variable in $\mathbb{R}^p$. Partition U into k subvectors, $U = (U_1', \ldots, U_k')' \in \mathbb{R}^{d_1,\ldots,d_k}$ where $d_1 + \cdots + d_k = p$. For each $j = 1, .., k$ let q_j denote the quantile function of $\|U_j\|_\infty$. Then for any $j = 1, \ldots, k$,*

$$q^N_{M,j}(\epsilon - 2\rho(U, N_M)) \leq q_j(\epsilon) \leq q^N_{M,j}(\epsilon + \rho(U, N_M)) \quad \textit{for all } \epsilon \in (0, 1).$$

Proof. Proof is a slight modification of that of Lemma D.3 in Chetverikov and Sørensen (2021). Main idea is to add and substract a $\|N_M\|_\infty$ term and use the fact that the approximation is achieved over all hyperrectangles. We show the bound holds for each $j = 1, \ldots, k$. Without loss of generality, consider U_1. Let $N_{M,1}$ denote the maginal distribution of the first d_1 elements of a $\mathbb{R}^p$ vector with distribution N_M.

$$\begin{aligned}
\Pr(\|U_1\|_\infty \leq t) &= \Pr(\|N_{M,1}\|_\infty \leq t) + \Pr(\|U_1\|_\infty \leq t) - \Pr(\|N_{M,1}\|_\infty \leq t) \\
&= \Pr(\|N_{M,1}\|_\infty \leq t) + \Big(\Pr(U \in [-t, t]^p \times \mathbb{R}^{p-d_1}) - \Pr(N_M \in [-t, t]^p \times \mathbb{R}^{p-d_1})\Big) \\
&\leq \Pr(\|N_{M,1}\|_\infty \leq t) + \rho(U, N_M)
\end{aligned}$$

for any $t \in \mathbb{R}$. A similar construction will give that

$$\Pr(\|U_1\|_\infty \leq t) \geq \Pr(\|N_{M,1}\|_\infty \leq t) - \rho(U, N_M).$$

Substituting $t = q^N_{M,1}(\epsilon - 2\rho(U, N_M))$ into the upper bound on $\Pr(\|U_1\|_\infty \leq t)$ gives the lower bound statement, while $t = q^N_{M,1}(\epsilon + \rho(U, N_M))$ and using the lower bound on $\Pr(\|U_1\|_\infty \leq t)$ gives the upper bound statement. □

As with Z partition S_n and $\widehat{S}^e_n$ into

$$S_n = (S'_{n,1}, \ldots, S'_{n,k})' \in \mathbb{R}^{d_1,\ldots,d_k} \quad \text{and} \quad \widehat{S}^e_n = (\widehat{S}^{e\prime}_{n,1}, \ldots, \widehat{S}^{e\prime}_{n,k})' \in \mathbb{R}^{d_1,\ldots,d_k}.$$

For each $j = 1, \dots, k$ define $q_{n,j}(\epsilon)$ as the ϵ-quantile of $\|S_{n,j}\|_\infty$

$$q_{n,j}(\epsilon) := \inf\{t \in \mathbb{R} : \Pr(\|S_{n,j}\|_\infty \leq t) \geq \epsilon\} \quad \text{for } \epsilon \in (0,1).$$

Let $\widehat{q}_{n,j}(\epsilon)$ be the ϵ-quantile of $\|\widehat{S}^e_{n,j}\|_\infty$, computed conditionally on X_i and $\widehat{X}_i$'s,

$$\widehat{q}_{n,j}(\epsilon) := \inf\{t \in \mathbb{R} : \Pr_e(\|\widehat{S}^e_{n,j}\|_\infty \leq t) \geq \epsilon\} \text{ for } \epsilon \in (0,1).$$

Theorem 2.9.3 (Quantile Comparasion)**.** *If* (2.9.22) *holds for some finite constants* $b > 0$ *and* $B_n \geq 1$*, and*

$$\rho_n := 2C_b \left(\frac{B_n^4 \ln^7(pn)}{n} \right)^{1/6}$$

denotes the upper bound in Theorem 2.9.1 multiplied by two, then for all $j = 1, \dots, k$

$$q^N_{\Sigma,j}(1 - \epsilon - \rho_n) \leq q_{n,j}(1-\epsilon) \leq q^N_{\Sigma,j}(1 - \epsilon + \rho_n) \quad \textit{for all } \epsilon \in (0,1).$$

If, in addition, (2.9.23) *holds for some sequences* $\{\delta_n\}_\mathbb{N}$ *and* $\{\beta_n\}_\mathbb{N}$ *converging to zero, and*

$$\rho'_n \leq 2C'_b \max\left\{ \delta, \left(\frac{B_n^4 \ln^6(pn)}{n} \right)^{1/6} \right\}$$

denotes the upper bound in Theorem 2.9.2 multiplied by two, then with probability at least $1 - \beta_n - 1/\ln^2(pn)$ *we have for all* $j = 1, \dots, k$*,*

$$q^N_{\Sigma,j}(1 - \epsilon - \rho'_n) \leq \widehat{q}_{n,j}(1-\epsilon) \leq q^N_{\Sigma,j}(1 - \epsilon + \rho'_n) \quad \textit{for all } \epsilon \in (0,1).$$

Proof. From Lemma 2.9.17 with $U = S_n$ we obtain

$$q^N_{\Sigma,j}(1 - \epsilon - 2\rho(S_n, N_\Sigma)) \leq q_{n,j}(1-\epsilon) \leq q^N_{\Sigma,j}(1 - \epsilon + \rho(S_n, N_\Sigma)).$$

The first chain of inequalities then follows from $2\rho(S_n, N_\Sigma) \leq \rho_n$ by Theorem 2.9.1.

For the second claim, apply Lemma 2.9.17 with $U = \widehat{S}_n^e$ and condition on the Z_i's and $\widehat{Z}_i$'s obtain

$$q_{\Sigma,j}^N(1-\epsilon-2\tilde{\rho}(\widehat{S}_n^e, N_\Sigma)) \leq \widehat{q}_n(1-\epsilon) \leq q_{\Sigma,j}^N(1-\epsilon+\tilde{\rho}(\widehat{S}_n^e, N_\Sigma)).$$

The second chain of inequalities then follows on the event $2\tilde{\rho}(\widehat{S}_n^e, N_\Sigma) \leq \rho_n'$, which by Theorem 2.9.2 happens with probability at least $1 - \beta_n - 1/\ln^2(pn)$. □

Theorem 2.9.4 (Multiplier Bootstrap Consistency)**.** *Let* (2.9.22) *and* (2.9.23) *hold for some constants* $b > 0$ *and* $B_n \geq 1$ *and some sequences* $\{\delta_n\}_\mathbb{N}$ *and* $\{\beta_n\}_\mathbb{N}$ *in* $\mathbb{R}_{++}$ *converging to zero. Then, there exists a finite constant* C_b*, depending only on* b*, such that*

$$\max_{1\leq j\leq k} \sup_{\epsilon\in(0,1)} \left|\Pr(\|S_{n,j}\|_\infty \geq \widehat{q}_{n,j}(1-\alpha)) - \alpha\right| \leq C_b \max\left\{\beta_n, \delta_n, \left(\frac{B_n^4 \ln^7(pn)}{n}\right)^{1/6}, \frac{1}{\ln^2(pn)}\right\}.$$

Proof. By Theorem 2.9.1 and Theorem 2.9.3,

$$\begin{aligned}
\Pr(\|S_{n,j}\|_\infty \leq \widehat{q}_{n,j}(1-\epsilon)) &\leq \Pr(\|S_{n,j}\|_\infty \leq q_{\Sigma,j}^N(1-\epsilon+\rho_n')) + \beta_n + \frac{1}{\ln^2(pn)} \\
&\leq \Pr(\|N_{\Sigma,j}\|_\infty \leq q_{\Sigma,j}^N(1-\epsilon+\rho_n')) + \rho_n + \beta_n + \frac{1}{\ln^2(pn)} \\
&\leq 1 - \epsilon + \rho_n' + \rho_n + \beta_n + \frac{1}{\ln^2(pn)}
\end{aligned}$$

Where the second inequality is making use of the same rectangle argument as before. A parallel argument shows that

$$\Pr(\|S_{n,j}\|_\infty \leq \widehat{q}_{n,j}(1-\epsilon)) \geq 1 - \epsilon - \left(\rho_n' + \rho_n + \beta_n + \frac{1}{\ln^2(pn)}\right).$$

Combining these two inequalities gives the result.

□

2.10. Appendix: Additional Second Stage Results

Theorem 2.10.1 (Integrated Rate of Convergence)**.** *Assume that Condition 1 and Assumption 2.4.1 hold. In addition suppose that* $\xi_k^2 \log k/n \to 0$ *and* $c_k \to 0$*. Then if either the propensity score our outcome regression model are correctly specified:*

$$\|\widehat{g}_k - g_0\|_{L,2} = (\mathbb{E}[(\widehat{g}(x) - g_0(x))^2])^{1/2} \lesssim_p \sqrt{k/n} + c_k \tag{2.10.1}$$

Proof. We begin with a matrix law of large numbers from Rudelson (1999), which is used to show $\widehat{Q} \to_p Q$.

Lemma 2.10.1 (Rudelson's LLN for Matrices)**.** *Let* $Q_1, \ldots, Q_n$ *be a sequence of independent, symmetric, non-negative* $k \times k$ *matrix valued random variables with* $k \geq 2$ *such that* $Q = \mathbb{E}[\mathbb{E}_n Q_i]$ *and* $\|Q_i\| \leq M$ *a.s. Then for* $\widehat{Q} = \mathbb{E}_n[Q_i]$*,*

$$\Delta := \mathbb{E}\|\widehat{Q} - Q\| \lesssim \frac{M \log k}{n} + \sqrt{\frac{M\|Q\| \log k}{n}}.$$

In particular if $Q_i = p_i p_i'$ *with* $\|p_i\| \leq \xi_k$ *almost surely, then*

$$\Delta := \mathbb{E}\|\widehat{Q} - Q\| \lesssim \frac{\xi_k^2 \log k}{n} + \sqrt{\frac{\xi_k^2\|Q\| \log k}{n}}.$$

Now, to prove Theorem 2.10.1 we have that:

$$\begin{aligned}\|\widehat{g}_k - g_0\|_{L,2} &\leq \|p^k(x)'\widehat{\beta}^k - p^k(x)'\beta^k\|_{L,2} + \|p^k(x)'\beta^k - g\|_{L,2} \\ &\leq \|p^k(x)'\widehat{\beta}^k - p^k(x)'\beta^k\|_{L,2} + c_k\end{aligned}$$

where under the normalization $Q = I_k$ we have that

$$\|p'\widehat{\beta} - p'\beta\|_{L,2} = \|\widehat{\beta} - \beta\|$$

Further,

$$\begin{aligned}\|\widehat{\beta}^k - \beta^k\| &= \|\widehat{Q}^{-1}\mathbb{E}[p^k(x) \circ (\widehat{Y} - \bar{Y})]\| + \|\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ (\epsilon^k + r_k)]\| \\ &\leq \|\widehat{Q}^{-1}\mathbb{E}[p^k(x) \circ (\widehat{Y} - \bar{Y})]\| + \|\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ \epsilon^k]\| + \|\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) r_k]\|\end{aligned}$$

By the matrix LLN (Lemma 2.10.1) we have that since $\xi_k^2 \log k/n \to 0$, $\|\widehat{Q} - Q\| \to_p 0$. This means that with probability approaching one all eigenvalues of $\widehat{Q}$ are boundedaway from zero, in particular they are larger than $1/2$. So w.p.a 1

$$\lesssim \|\mathbb{E}[p^k(x) \circ (\widehat{Y} - \bar{Y})]\| + \|\mathbb{E}_n[p^k(x) \circ \epsilon^k]\| + \|\mathbb{E}_n[p^k(x) r_k]\|$$

Under Condition 1 the first term is $o_p(\sqrt{k/n})$. By equation (A.48) in Belloni et al. (2015) the third term is bounded in probability by c_k. For the second term apply the third condition in Assumption 2.4.1 to see

$$\mathbb{E}\|\mathbb{E}_n[p^k(x) \circ \epsilon^k]\|^2 = \mathbb{E}\sum_{j=1}^{k} \epsilon_j^2 p_j(x)^2/n \leq \bar{\sigma}^2 \mathbb{E}_n[p^k(x)p^k(x)'/n] \lesssim \mathbb{E}[p^k(x)p^k(x)'/n] = k/n.$$

This gives $\|\mathbb{E}_n[p^k(x) \circ \epsilon^k]\| \lesssim_p \sqrt{k/n}$ and thus shows (2.10.1). □

Lemma 2.10.2 (Pointwise Linearization)**.** *Suppose that Condition 1 and Assumption 2.4.1, hold. In addition assume that* $\xi_k^2 \log k/n \to 0$. *Then for any* $\alpha \in S^{k-1}$,

$$\sqrt{n}\alpha'(\widehat{\beta}^k - \beta^k) = \alpha'\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] + R_{1n}(\alpha) \tag{2.10.2}$$

where the term $R_{1n}(\alpha)$, *summarizing the impact of unknown design, obeys*

$$R_{1n}(\alpha) \lesssim_p \sqrt{\frac{\xi_k^2 \log k}{n}}(1 + \sqrt{k}\ell_k c_k) \tag{2.10.3}$$

Moreover,

$$\sqrt{n}\alpha'(\widehat{\beta}^k - \beta^k) = \alpha'\mathbb{G}_n[p^k(x) \circ \epsilon^k] + R_{1n}(\alpha) + R_{2n}(\alpha) \tag{2.10.4}$$

where the term R_{2n}, summarizing the impact of approximation error on the sampling error of the estimator, obeys

$$R_{2n}(\alpha) \lesssim_p \ell_k c_k \tag{2.10.5}$$

Proof. Decompose as before,

$$\begin{aligned}\sqrt{n}\alpha'(\widehat{\beta}^k - \beta^k) = \sqrt{n}\alpha'\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ (\widehat{Y} - \bar{Y})] \\ + \alpha'\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] \\ + \alpha'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)].\end{aligned}$$

The first term is $o_p(1)$ under Condition 1, we can just include this term in $R_{1n}(\alpha)$. Now bound $R_{1n}(\alpha)$ and $R_{2n}(\alpha)$.

Step 1. Conditional $X = [x_1, \ldots, x_n]$, the term

$$\alpha'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ \epsilon^k].$$

has mean zero and variance bounded by $\bar{\sigma}^2\alpha'[\widehat{Q}^{-1} - I]\widehat{Q}^{-1}[\widehat{Q}^{-1} - I]\alpha$. Next, by Lemma 2.10.1, with probability approaching one, all eigenvalues of $\widehat{Q}^{-1}$ are bounded from above and away zero. So,

$$\bar{\sigma}^2\alpha'[\widehat{Q}^{-1} - I_k]\widehat{Q}^{-1}[\widehat{Q}^{-1} - I_k]\alpha \lesssim \bar{\sigma}^2\|\widehat{Q}\|\|\widehat{Q}^{-1}\|^2\|\widehat{Q}^{-1} - I_k\|^2 \lesssim_p \frac{\xi_k^2 \log k}{n}.$$

so by Chebyshev's inequality,

$$\alpha'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ \epsilon^k] \lesssim_p \sqrt{\frac{\xi_k^2 \log k}{n}}.$$

Step 2. From the proof of Lemma 4.1 in Belloni et al. (2015), we get that

$$\alpha'(\widehat{Q}^{-1} - I_k)\mathbb{G}_n[p^k(x)r_k] \lesssim_p \sqrt{\frac{\xi_k^2 \log k}{n}}\ell_k c_k \sqrt{k}$$

This completes the bound on $R_{1n}(\alpha)$ and gives (2.10.2)-(2.10.3). Next, also from the proof of Lemma 4.1 from Belloni et al. (2015),

$$R_{2n}(\alpha) = \alpha'\mathbb{G}_n[p^k(x)r_k] \lesssim_p \ell_k c_k,$$

which gives (2.10.4)-(2.10.5). □

The following lemma shows that, after adding Assumption 2.4.2 the linearization of our coefficient estimator $\widehat{\beta}^k$ established in Lemma 2.10.2 holds uniformly over all points $x \in \mathcal{X}$. That is to say the error from linearization is bounded in probability uniformly over all $x \in \mathcal{X}$. It will form an important building block in uniform consistency and strong approximation results presented in Theorems 2.10.2 and 2.4.2.

Lemma 2.10.3 (Uniform Linearization)**.** *Suppose that Condition 1 and Assumption 2.4.1-2.4.2 hold. Then if either the propensity score model our outcome regression model is correctly specified:*

$$\sqrt{n}\alpha(x)'(\widehat{\beta}^k - \beta^k) = \alpha(x)'\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] + R_{1n}(\alpha(x)) \tag{2.10.6}$$

where $R_{1n}(\alpha(x))$ describes the design error and satisfies

$$R_{1n}(\alpha(x)) \lesssim_p \sqrt{\frac{\xi_k^2 \log k}{n}}(n^{1/m}\sqrt{\log k} + \sqrt{k}\ell_k c_k) := \bar{R}_{1n} \tag{2.10.7}$$

uniformly over $x \in \mathcal{X}$. Moreover,

$$\sqrt{n}\alpha(x)'(\widehat{\beta}^k - \beta^k) = \alpha(x)'\mathbb{G}_n[p^k(x) \circ \epsilon^k] + R_{1n}(\alpha(x)) + R_{2n}(\alpha(x)) \tag{2.10.8}$$

where $R_{2n}(\alpha(x))$ describes the sampling error and satisfies, uniformly over $x \in \mathcal{X}$:

$$R_{2n}(\alpha(x)) \lesssim_P \sqrt{\log k} \cdot \ell_k c_k := \bar{R}_{2n} \tag{2.10.9}$$

Proof. As in the proof of Lemma 2.10.2, we decompose

$$\begin{aligned}\sqrt{n}\alpha(x)'(\widehat{\beta}^k - \beta^k) = \sqrt{n}\alpha(x)'\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ (\widehat{Y} - \bar{Y})] \\ + \alpha(x)'\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)] \\ + \alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ (\epsilon^k + r_k)].\end{aligned} \tag{2.10.10}$$

Using Condition 1, the matrix LLN (Lemma 2.10.1), and bounded eigenvalues of the design matix, we have that:

$$\sup_{x \in \mathcal{X}} \sqrt{n}\alpha(x)'\widehat{Q}^{-1}\mathbb{E}_n[p^k(x) \circ (\widehat{Y} - \bar{Y})] = o_p(1).$$

Since this is $o_p(1)$, we can simply include this term in $R_{1n}(\alpha(x))$. Now derive bounds on $R_{1n}(\alpha(x))$ and $R_{2n}(\alpha(x))$.

Step 1: Conditional on the data let

$$T := \left\{ t = (t_1, \ldots, t_n) \in \mathbb{R}^n : t_i = \alpha(x)'(\widehat{Q}^{-1} - I)p^k(x) \circ \epsilon^k, x \in \mathcal{X} \right\}.$$

Define the norm $\|\cdot\|_{n,2}$ on $\mathbb{R}^n$ by $\|t\|_{n,2}^2 = n^{-1}\sum_{i=1}^n t_i^2$. For an $\varepsilon > 0$ an ε-net of the normed space $(T, \|\cdot\|_{n,2})$ is a subset T_ε of T such that for every $t \in T$ there is a point $t_\varepsilon \in T_\varepsilon$ such that $\|t - t_\varepsilon\|_{n,2} < \varepsilon$. The covering number $N(T, \|\cdot\|_{n,2}, \varepsilon)$ of T is the infimum of the cardinality of ε-nets of T.

Let $\eta_1, \ldots, \eta_n$ be independent Rademacher random variables that are independent of the data. Let $\eta = (\eta_1, \ldots, \eta_n)$. Let $\mathbb{E}_\eta[\cdot]$ denote the expectation with respect to the distribution

of η. By Dudley's inequality (Dudley, 1967),

$$\mathbb{E}_\eta\left[\sup_{x\in\mathcal{X}}\left|\alpha(x)'[\widehat{Q}^{-1}-I]\mathbb{G}_n[\eta_i p^k(x)\circ\epsilon^k]\right|\right]\lesssim\int_0^\theta\sqrt{\log N(T,\|\cdot\|_{n,2},\varepsilon)}\,d\varepsilon.$$

where

$$\begin{aligned}\theta&:=2\sup_{t\in T}\|t\|_{n,2}\\&=2\sup_{x\in\mathcal{X}}\left(\mathbb{E}_n[(\alpha(x)'(\widehat{Q}^{-1}-I)p^k(x)\circ\epsilon^k)^2]\right)^{1/2}\\&\leq 2\max_{1\leq i\leq n}|\bar{\epsilon}_{k,i}|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2},\end{aligned}$$

by (2.9.5). Now, for any $x\in\mathcal{X}$,

$$\begin{aligned}&\left(\mathbb{E}_n[(\alpha(x)'(\widehat{Q}^{-1}-I)p^k(x)\circ\epsilon^k-\alpha(\tilde{x})'(\widehat{Q}^{-1}-I)p^k(x)\circ\epsilon^k)^2]\right)^{1/2}\\&\quad\leq\max_{1\leq i\leq n}|\bar{\epsilon}_{k,i}|\|\alpha(x)-\alpha(\tilde{x})\|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2}\\&\quad\leq\xi_k^L\max_{1\leq i\leq n}|\bar{\epsilon}_{k,i}|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2}\|x-\tilde{x}\|\end{aligned}$$

So, for some $C>0$,

$$N(T,\|\cdot\|_{n,2},\varepsilon)\leq\left(\frac{C\xi_k^L\max_{1\leq i\leq n}|\bar{\epsilon}_{k,i}|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2}}{\varepsilon}\right)^{d_x}.$$

This gives us that

$$\int_0^\theta\sqrt{\log(N(T,\|\cdot\|_{2,n},\varepsilon))}\,d\varepsilon\leq\max_{1\leq i\leq n}|\bar{\epsilon}_{k,i}|\|\widehat{Q}^{-1}-I\|\|\widehat{Q}\|^{1/2}\int_0^2\sqrt{d_x\log(C\xi_k^L/\varepsilon)}\,d\varepsilon.$$

By Assumption 2.4.2 we have that $\mathbb{E}[\max_{1\leq i\leq n}|\bar{\epsilon}_{k,i}|\mid X]\lesssim_P n^{1/m}$ where $X=(x_1,\ldots,x_n)$. In addition $\xi_k^{2m/(m-2)}\log k/n\lesssim 1$ for $m>2$ gives that $\xi_k^2/\log k/n\to 0$. So, $\|\widehat{Q}^{-1}-I\|\lesssim_P$

$(\xi_k^2 \log k/n)^{1/2}$ and $\|\widehat{Q}^{-1}\| \lesssim_P 1$. Combining this all with $\log \xi_k^L \lesssim \log k$ implies

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} \left|\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ \epsilon^k]\right| \mid X\right] \leq 2\mathbb{E}\left[\mathbb{E}_\eta \sup_{x\in\mathcal{X}} \left|\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[\eta_i p^k(x) \circ \epsilon^k]\right| \mid X\right]$$
$$\lesssim_P n^{1/m}\sqrt{\frac{\xi_k^2 \log^2 k}{n}}$$

where the first line is due to symmetrization inequality. This gives us

$$\sup_{x\in\mathcal{X}} \left|\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) \circ \epsilon^k]\right| \lesssim_p n^{1/m}\sqrt{\frac{\xi_k^2 \log^2 k}{n}} \tag{2.10.11}$$

Step 2: Now simply report the results on approximation error from Belloni et al. (2015) . Since the approximation error is the same for all signals $Y(\bar{\pi}_k, \bar{m}_k)$, there is no Hadamard product to deal with.

$$\sup_{x\in\mathcal{X}} \left|\alpha(x)'[\widehat{Q}^{-1} - I]\mathbb{G}_n[p^k(x) r_k]\right| \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}} \ell_k c_k \sqrt{k} \tag{2.10.12}$$

$$\sup_{x\in\mathcal{X}} \left|\alpha(x)'\mathbb{G}_n[p^k(x) r_k]\right| \lesssim_P \ell_k c_k \sqrt{\log k} \tag{2.10.13}$$

Looking at (2.10.10) and combining (2.10.11)-(2.10.12) gives the bound on $R_{1n}(\alpha(x))$ while (2.10.13) gives the bound on $R_{2n}(\alpha(x))$. □

Theorem 2.10.2 gives conditions under which our estimator converges in probability to the true conditional counterfactual outcome $g_0(x)$. In particular, this convergence happens uniformly at the rates defined in (2.10.15)-(2.10.16). If these two terms go to zero, the entire estimator will converge uniformly to the true conditional expectation of interest.

Theorem 2.10.2 (Uniform Rate of Convergence)**.** *Suppose that Condition 1 and Assumptions 2.4.1-2.4.2 hold. Then so long as either the propensity score model or outcome regression*

model is correctly specified:

$$\sup_{x\in\mathcal{X}}\left|\alpha(x)'\mathbb{G}_n[p^k(x)\circ\epsilon^k]\right| \lesssim_P \sqrt{\log k} \tag{2.10.14}$$

Moreover, for

$$\bar{R}_{1n} := \sqrt{\frac{\xi_k^2\log k}{n}}(n^{1/m}\sqrt{\log k}+\sqrt{k}\ell_k c_k)$$
$$\bar{R}_{2n} := \sqrt{\log k}\cdot\ell_k c_k$$

we have that

$$\sup_{x\in\mathcal{X}}\left|p^k(x)'(\widehat{\beta}^k-\beta^k)\right| \lesssim_P \frac{\xi_k}{\sqrt{n}}\left(\sqrt{\log k}+\bar{R}_{1n}+\bar{R}_{2n}\right) \tag{2.10.15}$$

and

$$\sup_{x\in\mathcal{X}}\left|\widehat{g}(x)-g_0(x)\right| \lesssim_P \frac{\xi_k}{\sqrt{n}}\left(\sqrt{\log k}+\bar{R}_{1n}+\bar{R}_{2n}\right)+\ell_k c_k \tag{2.10.16}$$

Proof. The goal will be to apply the following two theorems from Giné and Koltchinskii (2006) and der Vaart and Wellner (1996).

Preliminaries for Proof of Theorem 2.10.2

Theorem (Gine and Koltchinskii, 2006). *Let $\xi_1, \dots, \xi_n$ be i.i.d random variables taking values in a measurable space $(S, \mathscr{S})$ with a common distribution P defined on the underlying n-fold product space. Let $\mathcal{F}$ be a measurable class of functions mapping $S \to \mathbb{R}$ with a measurable envelope F. Let σ^2 be a constant such that $\sup_{f\in\mathcal{F}} \operatorname{Var}(f) \le \sigma^2 \le \|F\|^2_{L^2(P)}$. Suppose there exist constats $A > e^2$ and $V \ge 2$ such that $\sup_Q N(\mathcal{F}, L^2(Q), \varepsilon\|F\|_{L^2(Q)}) \le (A/\varepsilon)^V$ for all $0 < \varepsilon \le 1$. Then*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \{f(\xi_i) - \mathbb{E}[f(\xi_1)]\}\right\|_{\mathcal{F}}\right] \le C\left[\sqrt{n\sigma^2 V \log\frac{A\|F\|_{L^2(P)}}{\sigma}} + V\|F\|_\infty \log\frac{A\|F\|_{L^2(P)}}{\sigma}\right]. \tag{GK}$$

where C is a universal constant.

Theorem (VdV&W 2.14.1). *Let $\mathcal{F}$ be a P-measurable class of measurable functions with a measurable envelope function F. Then for any $p \ge 1$,*

$$\left\|\|\mathbb{G}_n\|^*_{\mathcal{F}}\right\|_{P,p} \lesssim \|J(\theta_n, \mathcal{F})\|F\|_n\|_{P,p} \lesssim J(1, \mathcal{F})\|F\|_{P,2\vee p} \tag{VW}$$

where $\theta_n = \left\|\|f\|_n\right\|^*_{\mathcal{F}} / \|F\|_n$, where $\|\cdot\|_n$ is the $L_2(\mathbb{P}_n)$ seminorm and the inequalities are valid up to constants depending only on the p in the statement. The term $J(\cdot,\cdot)$ is given

$$J(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\mathcal{F}, \|\cdot\|_{L^2(Q)}, \varepsilon\|F\|_{L^2(Q)})}\, d\varepsilon.$$

We would like to apply these theorems to bound $\sup_{x\in\mathcal{X}} |\alpha(x)'\mathbb{G}_n[p^k(x)\circ\epsilon^k]|$ and thus show (2.10.14). The other two statements of Theorem 2.10.2 follow from this. To this end, let's consider the class of functions

$$\mathcal{G} := \{(\epsilon^k, x) \mapsto \alpha(v)'(p^k(x)\circ\epsilon^k), v \in \mathcal{X}\}.$$

Let's note that $|\alpha(v)'p^k(x)| \le \xi_k$, $\operatorname{Var}(\alpha(v)'p^k(x)) = 1$, and for any $v, \tilde{v} \in \mathcal{X}$

$$|\alpha(v)'(p^k(x)\circ\epsilon^k) - \alpha(\tilde{v})'(p^k(x)\circ\epsilon^k)| \le |\bar{\epsilon}_k|\xi_k^L\xi_k\|v - \tilde{v}\|,$$

where $\bar{\epsilon}_k = \|\epsilon^k\|_\infty$. Then, taking $G(\epsilon^k, x) \le \bar{\epsilon}_k\xi_k$ we have that

$$\sup_Q N(\mathcal{G}, L^2(Q), \varepsilon\|G\|_{L^2(Q)}) \le \left(\frac{C\xi_k^L}{\varepsilon}\right)^d. \tag{2.10.17}$$

Now, for a $\tau \ge 0$ specified later define $\epsilon_k^- = \epsilon^k\mathbf{1}\{|\bar{\epsilon}_k| \le \tau\} - \mathbb{E}[\epsilon^k\mathbf{1}\{|\bar{\epsilon}_k| \le \tau\} \mid X]$ and

$\epsilon_k^+ = \epsilon^k \mathbf{1}\{|\bar{\epsilon}_k| > \tau\} - \mathbb{E}[\epsilon^k \mathbf{1}\{|\bar{\epsilon}_k| > \tau\} \mid X]$. Since $\mathbb{E}[\epsilon^k \mid X] = 0$ we have that $\epsilon^k = \epsilon_k^- + \epsilon_k^+$. Using this decompose:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha(v)'(p^k(x) \circ \epsilon^k) = \sum_{i=1}^n \alpha(v)'(p^k(x) \circ \epsilon_k^-)/\sqrt{n} + \sum_{i=1}^n \alpha(v)'(p^k(x) \circ \epsilon_k^+)/\sqrt{n}.$$

We deal with each of these terms individually, in two steps.

Step 1: For the first term, we set up for an application of (GK). Equation (2.10.17) gives us the constants $A = C\xi_k^L$ and $V = d_x \vee 2$. To get σ^2 note that for any $v \in \mathcal{X}$,

$$\begin{aligned}
\mathrm{Var}(\alpha(v)'(p^k(x) \circ \epsilon_k^-)/\sqrt{n}) &\leq \mathbb{E}[(\alpha(v)'(p^k(x) \circ \epsilon_k^-)/\sqrt{n})^2] \\
&\leq \frac{1}{n} \mathbb{E}[(\alpha(v)' p^k(x))^2] \sup_{x \in \mathcal{X}} \mathbb{E}[\|\epsilon_k^-\|_\infty^2 \mid X = x] \\
&\leq \frac{\bar{\sigma}_k^2 \wedge \tau^2}{n}
\end{aligned}$$

Finally note that we can take the envelope $G = \|\epsilon_k^-\|_\infty \xi_k/\sqrt{n}$ where $\|G\|_{L^2(P)} \leq \frac{\bar{\sigma}_k \wedge \tau}{\sqrt{n}}$ and $\|G\|_\infty \leq \tau \xi_k/\sqrt{n}$.

We can now apply (GK) to get that

$$\mathbb{E}[\sup_{x \in \mathcal{X}} |\alpha(x)' \mathbb{G}_n[p^k(x) \circ \epsilon_k^-]|] \lesssim \sqrt{\bar{\sigma}_k^2 \wedge \tau^2 \log(\xi_k^L)} + \frac{\tau \xi_k \log(\xi_k^L)}{\sqrt{n}}.$$

Step 2: For the second term, we set up for an application of (VW) with the envelope function $G = \|\epsilon_k^+\|_\infty \xi_k/\sqrt{n}$ and note that

$$\mathbb{E}[\|\epsilon_k^+\|_\infty^2] \leq \mathbb{E}[\bar{\epsilon}_k^2 \mathbf{1}\{|\bar{\epsilon}_k| > \tau\}] \leq \tau^{-m+2} \mathbb{E}[|\bar{\epsilon}_k|^m]$$

We can now use (VW) to bound

$$\begin{aligned}
\mathbb{E}\left[\left\|\sup_{x \in \mathcal{X}} |\alpha(x)' \mathbb{G}_n[p^k(x) \circ \epsilon_k^+]|\right\|\right] &\lesssim \sqrt{\mathbb{E}[|\bar{\epsilon}_k|^m]} \tau^{-m/2+1} \xi_k \int_0^1 \sqrt{\log(\xi_k^L/\varepsilon)}\, d\varepsilon \\
&\lesssim \sqrt{\sigma_k^m} \tau^{-m/2+1} \xi_k \sqrt{\log(\xi_k^L)}.
\end{aligned}$$

Step 3: Let $\tau = \xi_k^{2/(m-2)}$ and apply Markov's inequality. The bounds from step one and two

become

$$\sup_{x\in\mathcal{X}}|\alpha(x)'\mathbb{G}_n[p^k(x)\circ\epsilon_k^-]| \lesssim_P \sqrt{\bar{\sigma}_k^2\log(\xi_k^L)} + \frac{\xi_k^{2m/(m-2)}\log(\xi_k^L)}{\sqrt{n}}$$
$$\sup_{x\in\mathcal{X}}|\alpha(x)'\mathbb{G}_n[p^k(x)\circ\epsilon_k^+]| \lesssim_P \sqrt{\bar{\sigma}_k^m\log(\xi_k^L)}$$

Applying Assumption 2.4.2 along with the inequality

$$\frac{\xi_k^{m/(m-2)}\log k}{\sqrt{n}} = \sqrt{\log k}\sqrt{\frac{\xi_k^{2m/(m-2)}\log k}{n}} \lesssim \log k$$

completes the proof. □

Theorem 2.10.3 (Validity of Gaussian Bootstrap)**.** *Suppose that the assumptions of Theorem 2.4.2 hold with $a_n = \log n$ and the assumptions of Theorem 2.4.3 hold with $a_n = O(n^{-b})$ for some $b > 0$. In addition, suppose that there exists a sequence ξ_n' obeying $1 \lesssim \xi_n' \lesssim \|p^k(x)\|$ uniformly for all $x \in \mathcal{X}$ such that $\|p^k(x) - p^k(x')\|/\xi_n' \le L_n\|x - x'\|$, where $\log L_n \lesssim \log n$. Let N_k^b be a bootstrap draw from $N(0, I_k)$ and $P^\star$ be the distribution conditional on the observed data $\{Y_i, D_i, Z_i\}_{i=1}^n$. Then the following approximation holds uniformly in $\ell^\infty(\mathcal{X})$:*

$$\frac{p^k(x)'\widehat{\Omega}^{1/2}}{\widehat{\Omega}^{1/2}p^k(x)}N_k^b =^d \frac{p^k(x)'\Omega^{1/2}}{\|\Omega^{1/2}p^k(x)\|} + o_{P^\star}(\log^{-1} N) \tag{2.10.18}$$

Proof. See Theorem 3.4 in Semenova and Chernozhukov (2021). □

2.10.1. Concentration and Tail Bounds

We make use of the following concentration and tail bounds. Lemmas 2.10.4–2.10.8 can be found in Bühlmann and van de Geer (2011). The proof of Lemma 2.10.9 is trivial but provided here.

Lemma 2.10.4. *Let $(Y_1, \ldots, Y_n)$ be independent random variables such that $\mathbb{E}[Y_i] = 0$ for*

$i = 1, \ldots, n$ *and* $\max_{i=1,\ldots,m} |Y_i| \leq c_0$ *for some constant* c_0*. Then, for any* $t > 0$*,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| > t\right) \leq 2\exp\left(-\frac{nt^2}{2c_0^2}\right).$$

Lemma 2.10.5. *Let* $(Y_1, \ldots, Y_n)$ *be independent random variables such that* $\mathbb{E}[Y_i] = 0$ *for* $i = 1, \ldots,$ *and* $(Y_1, \ldots, Y_n)$ *are uniformly sub-gaussian:* $\max_{1 \leq i \leq n} c_1^2 \mathbb{E}[\exp(Y_i^2/c_1^2) - 1] \leq c_2^2$ *for some constants* (c_1, c_2)*. Then for any* $t > 0$*,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| > t\right) \leq 2\exp\left(-\frac{nt^2}{8(c_1^2 + c_2^2)}\right).$$

Lemma 2.10.6. *Let* $(Y_1, \ldots, Y_n)$ *be independent variables such that* $\mathbb{E}[Y_i] = 0$ *for* $i = 1, \ldots, n$ *and*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[|Y_i|^k] \leq \frac{k!}{2} c_3^{k-2} c_4^2, \quad k = 2, 3, \ldots,$$

for some constants (c_3, c_4)*. Then, for any* $t > 0$*,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| > c_3 t + c_4\sqrt{2t}\right) \leq 2\exp(-nt).$$

Lemma 2.10.7. *Suppose that* Y *is sub-gaussian:* $c_1^2 \mathbb{E}[\exp(Y^2/c_1^2) - 1] \leq c_2^2$ *for some constants* (c_1, c_2)*. Then*

$$\mathbb{E}[|Y|^k] \leq \Gamma\left(\frac{k}{2} + 1\right)(c_1^2 + c_2^2) c_1^{k-2}, \quad k = 2, 3, \ldots.$$

Lemma 2.10.8. *Suppose that* X *is bounded,* $|X| \leq c_0$*, and* Y *is sub-gaussian,* $c_2^2 \mathbb{E}[\exp(Y^2/c_1^2) - 1] \leq c_2^2$ *for some constants* (c_1, c_2)*. Then* $Z = XY^2$ *satisfies*

$$\mathbb{E}\left[|Z - \mathbb{E}[Z]|^k\right] \leq \frac{k!}{2} c_3^{k-2} c_4^2, \quad k = 2, 3, \ldots,$$

for $c_3 = 2c_0 c_1^2$ *and* $c_4 = 2c_0 c_1 c_2$*.*

Lemma 2.10.9. *Suppose that Y is sub-gaussian in the following sense, there exist positive constants $c_0, c_1 > 0$ such that $c_0^2 \mathbb{E}[\exp(Y^2/c_0^2) - 1] \leq c_1^2$. Then*

$$\mathbb{E}[|Y|] \leq c_1^2/c_0 + c_0.$$

Proof. Use the fact that $e^{x^2} > |x|$ and the characterization of sub-gaussian. □

2.11. Appendix: Additional Details on Empirical Application

As mentioned in the setup, to avoid outlier contamination we drop the top 3% and bottom 3% of birthweights by maternal age. We also drop ages for which there are fewer than 10 smoker or non smoker observations. The result is a dataset with 4107 (of an initial 4602) observations on the outcome variable, birthweight. In addition to the 21 control variables (Z) available in the dataset, we further generate an additional 29 interaction/higher order variables that we believe may be useful in controlling for confounding as well as a constant. Table 2.11.1 provides a summary of the initial 21 control variables.[1]

In addition to these 21 control variables, we include the folowing interactions: mbsmoke × alcohol, medu × fedu, mage × fage, msmoke^2, msmoke × alcohol, mage^2, mage × mmarried, mage × medu, mage × fedu, monthslb^2, msmoke ×monthslb^2, monthslb^2 × msmoke ^{2}, msmoke^2 × prenatal^2, msmoke^2 × mage^2, mage^2 × monthslb^2, mage^2 × fage, fage^2 × mage^2, fage^2 × mage, mage^2 × mrace, fage^2 × frace, msmoke^2 × alcohol, mage^2 × alcohol, fage^2 × alcohol, monthslb^2 × alcohol, mage^2 × mhisp, fage^2 × fhisp, medu × mage^2. We also include indicators for the month of birth.

In conducting analysis, we found it quite helpful to the stability of the final model assisted estimator to do some light trimming of the estimated propensity score and outcome regression models. In particular we trim the estimated propensity score(s) to be between 0.01 and

[1] This table is generated using the wonderful stargazer package in `R` (Hlavac, 2022).

Table 2.11.1: Summary of Data used in Emprical Exercise

Statistic	N	Mean	St. Dev.	Min	Max
bweight	4,107	3,384.354	447.616	1,544	4,668
mmarried	4,107	0.708	0.455	0	1
mhisp	4,107	0.034	0.181	0	1
fhisp	4,107	0.038	0.192	0	1
foreign	4,107	0.054	0.226	0	1
alcohol	4,107	0.031	0.174	0	1
deadkids	4,107	0.252	0.434	0	1
mage	4,107	26.125	5.025	16	36
medu	4,107	12.703	2.470	0	17
fage	4,107	27.000	9.022	0	60
fedu	4,107	12.324	3.624	0	17
nprenatal	4,107	10.822	3.613	0	40
monthslb	4,107	21.938	30.255	0	207
order	4,107	1.858	1.056	0	12
msmoke	4,107	0.390	0.890	0	3
mbsmoke	4,107	0.183	0.386	0	1
mrace	4,107	0.847	0.360	0	1
frace	4,107	0.822	0.382	0	1
prenatal	4,107	1.204	0.507	0	3
birthmonth	4,107	6.556	3.352	1	12
lbweight	4,107	0.025	0.155	0	1
fbaby	4,107	0.443	0.497	0	1
prenatal1	4,107	0.803	0.398	0	1

0.99 and trim the estimated mean regression models so that they take a value no more than roughly 12.5% higher or lower than the maximum or minimum value of Y observed in the data.

Because the control variables are all of different magnitudes, it is common to do some normalization before estimating the ℓ_1-regularized propensity score and outcome regression models so that all variables are "punished" equally by the penalty. We normalize our data by scaling each variable to take on values between zero and one.

2.12. Appendix: Consistency between First Stage and Second Stage Assumptions

In this section, we examine the consistency between the first stage and second stage assumptions on the basis terms $p^k(x)$. In particular, we are interested in finding a positive basis that also satisfies the bounded eigenvalue condition on the design matrix in Assumption 2.4.1. We also discuss how to construct the model assisted estimator with weights in (2.2.8)-(2.2.9) that are not directly the second stage basis terms in case the researcher is worried about their choice of basis terms satisfying the first stage and second stage stage assumptions simultaneously.

Suppose that $\mathcal{X} = [0, 1]$. First, note that the first stage non-negativity and second stage design assumptions can be trivially satisfied by using a locally constant basis; that is by taking

$$p_j(x) = \mathbf{1}_{[\ell_{j-1}, \ell_j)}(x) \tag{2.12.1}$$

for some $0 = \ell_0 < \ell_1 < \cdots < \ell_t = 1$. While this basis may have poor approximation qualities, the general principle can be extended to any basis whose elements have disjoint (or limitedly overlapping) supports. Higher order piecewise polynomial approximations can often be implemented using *B-splines* which are orthonormalized regression splines. See De Boor (2001) for an in-depth discussion or Newey (1997) for an application of B-splines to nonparametric series regression.

These higher order splines can be defined recursively. For a given (weakly increasing) knot sequence $\ell := (\ell_j)_{j=1}^t$ we define the "first-order" B-splines denoted $B_{1,1}(x), \ldots, B_{t,1}(x)$ using (2.12.1), that is $B_{j,1}(x) = p_j(x)$. On top of these functions, we can define higher order B-splines via the recursive relation (De Boor (2001), p.90)

$$B_{j,d+1} := \omega_{j,d}(x) B_{j,d}(x) + [1 - \omega_{j+1,d}(x)] B_{j+1,d}(x). \tag{2.12.2}$$

where

$$\omega_{j,d}(x) := \begin{cases} \frac{x-\ell_j}{\ell_{j+d}-\ell_j} & \text{if } \ell_{j+d} \neq \ell_j \\ 0 & \text{otherwise} \end{cases}.$$

If X is continuously distributed on an open set containing the knots (ℓ_j), De Boor (2001) shows that the B-spline basis is almost surely positive. Moreover, B-splines is locally supported in the sense each $B_{j,d}$ is positive on (ℓ_j, ℓ_{j+d}), zero off this support and for each d:

$$\sum_{j=1}^{t} B_{j,d} = 1 \qquad \text{on } [0,1].$$

where the summation is taken pointwise (see De Boor (2001), p.36). From the final property we can see the B-spline basis using $k = td$ basis terms, $p^k(x) = (B_{j,l}(x))_{\substack{j=1,..,t \\ l=1,...,d}}$ are totally bounded so that.

B-splines used directly in this manner, however, do not lead to a design matrix $Q = \mathbb{E}[p^k(x)p^k(x)']$ with eigenvalues which are bounded away from zero. To achieve this, the basis fucntions must be divided by their ℓ_2 norm. In practice, this leads to b-spline terms who are grown at rate $\xi_{k,\infty} \lesssim \sqrt{k}$. The pilot penalty constants can be chosen from a set whose bounds are on the order of $\sqrt{k}$ and the sparsity bounds of Assumption 2.3.1 reduce to

$$\frac{s_k\, k^{3/2} \ln^5(d_z n)}{n} \to 0 \quad \text{and} \quad \frac{k^2 \ln^7(d_z k n)}{n} \to 0$$

while the bounds in (2.4.2) and (2.4.11) reduce respectively to

$$\frac{s_k\, k^{3/2} \ln(d_z)}{\sqrt{n}} \to 0 \quad \text{and} \quad \frac{s_k^2\, k^{7/2} \ln(d_z)}{n^{(m-1)/m}} \to 0.$$

2.12.1. Alternate Weighting

So long as the second stage basis $p^k(x)$ contains a constant term, it is possible to weight the estimating equations (2.2.8)-(2.2.9) by some $p^k(x) = p^k(x) + c_k$ with minimal modification to

the model assisted estimator. The constants $c_k \in \mathbb{R}$ can be allowed to grow with k so long as we replace $\xi_{k,\infty}$ with the maximum of $\tilde{\xi}_{k,\infty} := \sup_{x \in \mathcal{X}} \|\tilde{p}^k(x)\|_\infty$ and $\xi_{k,\infty}$ in the sparsity bounds of Section 2.4. Without loss of generality we will assume that the first basis term is a constant so that $p_1(x) \equiv 1$

After estimating the models $(\hat{\pi}_1, \hat{m}_1), \ldots, (\hat{\pi}_k, \hat{m}_k)$ using $(\tilde{p}_1(x), \ldots, \tilde{p}_k(x))$ in (2.2.8)-(2.2.9) we would construct the second stage estimate $\hat{\beta}^k$

$$\tilde{\beta}^k = \widehat{Q}^{-1} \mathbb{E}_n \begin{bmatrix} \tilde{p}_1(x) Y(\hat{\pi}_1, \hat{m}_1) - c_k Y(\hat{\pi}_1, \hat{m}_1) \\ \tilde{p}_2(x) Y(\hat{\pi}_2, \hat{m}_2) - c_k Y(\hat{\pi}_1, \hat{m}_1) \\ \vdots \\ \tilde{p}_k(x) Y(\hat{\pi}_k, \hat{m}_k) - c_k Y(\hat{\pi}_1, \hat{m}_1) \end{bmatrix}.$$

Via the same analysis of Sections 2.3 and 2.4 we will still be able to show that the bias passed on from first stage estimation to the second stage parameter $\tilde{\beta}^k$ remains negligible even under misspecification of either first stage model. This is because Lemma 2.3.1 will establish that

$$\begin{aligned} &\max_{1 \le j \le k} |\mathbb{E}_n[\tilde{p}_j(x) Y(\hat{\pi}_j, \hat{m}_j)] - \mathbb{E}_n[\tilde{p}_j(x) Y(\bar{\pi}_j, \bar{m}_j)]| = o_p(n^{-1/2} k^{-1/2}) \quad \text{and} \\ &\max_{1 \le j \le k} \tilde{\xi}_{k,\infty} \max_{1 \le j \le k} \mathbb{E}_n[\tilde{p}_j(x)^2 (Y(\hat{\pi}_j, \hat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] = o_p(k^{-2} n^{-1/m}) \end{aligned}.$$

Using the first statement, we can immediately establish via the triangle inequality that

$$\max_{1 \le j \le k} |\mathbb{E}_n[\tilde{p}_j(x) Y(\hat{\pi}_j, \hat{m}_j) - c_k Y(\hat{\pi}_1, \hat{m}_1)] - \mathbb{E}_n[\tilde{p}_j(x) Y(\bar{\pi}_j, \bar{m}_j) - c_k Y(\bar{\pi}_1, \bar{m}_1)]| = o_p(n^{-1/2} k^{-1/2})$$

which is the exact analog of Condition 1 needed to establish consistency at the nonparameteric rate of the modified model assisted estimator. Similarly, using the second statement and $(a+b)^2 \le 2a^2 + 2b^2$ we can immediately establish that

$$\max_{1 \le j \le k} \mathbb{E}_n[(\tilde{p}_j(x) Y(\hat{\pi}_j, \hat{m}_j) - c Y(\hat{\pi}_1, \hat{m}_1) - \tilde{p}_j(x) Y(\bar{\pi}_j, \bar{m}_j) + c Y(\bar{\pi}_j, \bar{m}_j))^2] = o_p(k^{-2} n^{-1/m})$$

which is the exact analog of Condition 2 needed to establish a consistent variance estimator when $\tilde{\beta}^k$ is used instead of the $\hat{\beta}^k$ from (2.2.12).

This logic can be extended slightly if the researcher would like to weight the estimating equations (2.2.8)-(2.2.9) by some $\tilde{p}^k(x) = G^k p^k(x)$ for an invertible and bounded sequence of linear operators $G^k : \mathbb{R}^k \to \mathbb{R}^k$. In this case, one would again use $\tilde{p}^k(x)$ in place of $p^k(x)$ in (2.2.8)-(2.2.9) and construct the second stage coeffecients via

$$\tilde{\beta}^k := \widehat{Q}^{-1} G^{k,-1} \mathbb{E}_n \begin{bmatrix} \tilde{p}_1(x) Y(\hat{\pi}_1, \hat{m}_1) \\ \vdots \\ \tilde{p}_k(x) Y(\hat{\pi}_k, \hat{m}_k) \end{bmatrix}$$

After constructing the second stage estimator using $\tilde{\beta}^k$, inference procedures would proceed normally as described in Section 2.2.

2.13. Appendix: Alternative CV-Type Method for Penalty Parameter Selection

In this section we consider a procedure for penalty parameter selection where we use the pilot penalty parameters described in (2.2.15) directly, after choosing constants $c_{\gamma,j}$ and $c_{\alpha,j}$ from a (finite) set via cross validation. For each j we will assume that

$$c_{\gamma,j}, c_{\alpha,j} \in \Lambda_n \subseteq [\underline{c}_n, \bar{c}_n] \tag{2.13.1}$$

where $|\Lambda_n|$ can be fairly large (on the order of n^2/k).

2.13.1. Theory Overview

Let $M_5, M_6, M_7, M_8^2, M_9^2$ be constants that do not depend on k as in Lemmas 2.9.11–2.9.15. Whenever

$$\underline{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}} \geq \xi_{k,\infty} \max\left\{M_5, M_6, M_7, M_8^2, M_9^2\right\} \sqrt{\frac{\ln(d_z n)}{n}}. \tag{2.13.2}$$

we will have that, under Assumption 2.3.1(i)-(iv) the event $\bigcap_{k=1}^{7} \Omega_{k,7}$ occurs with probability at least $1 - 10k/n^2$ for the $2k$ pilot penalty parameters chosen with any values $c_{\gamma,j}, c_{\alpha,j} \in \Lambda_n$ and

$$\bar{\lambda}_k := \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}.$$

In this event, apply Lemmas 2.9.5 and 2.9.6 to obtain the following finte sample bounds for the parameter estimates

$$\max_{1\leq j\leq k} D^{\ddagger}_{\gamma,j}(\widehat{\gamma}_j, \bar{\gamma}_j) \leq M_0 \frac{s_k \bar{c}_n^2 \ln^3(d_z n)}{n} \quad \text{and} \quad \max_{1\leq j\leq k} \|\widehat{\gamma}_j - \bar{\gamma}_j\|_1 \leq M_0 s_k \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}$$

$$\max_{1\leq j\leq k} D^{\ddagger}_{\alpha,j}(\widehat{\alpha}_j, \bar{\alpha}_j; \bar{\gamma}_j) \leq M_1 \frac{s_k \bar{c}_n^2 \ln^3(d_z n)}{n} \quad \text{and} \quad \max_{1\leq j\leq l} \|\widehat{\alpha}_j - \bar{\alpha}_j\|_1 \leq M_1 s_k \bar{c}_n \sqrt{\frac{\ln^3(d_z n)}{n}}$$

and Lemma 2.9.1 to obtain the following finite sample bound for the weighted means:

$$\max_{1\leq j\leq k} |\mathbb{E}_n[p_j(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))]| \leq M_2 \frac{\bar{c}_n^2 s_k \ln^3(d_z n)}{n} \tag{2.13.3}$$

$$\max_{1\leq j\leq k} |\mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] \leq M_3 \frac{\xi_{k,\infty}^2 \bar{c}_n^2 s_k^2 \ln^3(d_z n)}{n} \tag{2.13.4}$$

Combining (2.13.2) and (2.13.3) we can see that Condition 1 can be obtained under Assumption 2.3.1(i)-(iv) and the following modified sparsity bounds

$$\frac{k|\Lambda_n|}{n^2} \to 0, \quad \frac{\underline{c}_n^{-1} \xi_{k,\infty}}{\ln(d_z n)} \to 0 \quad \text{and} \quad \frac{\bar{c}_n^2 s_k k^{1/2} \ln^3(d_z n)}{\sqrt{n}} \to 0. \tag{2.13.5}$$

Simlarly combining (2.13.2) and (2.13.4), Condition 2 can additionally be obtained by strengthening the rates in (2.13.5) to include

$$\frac{\xi_{k,\infty}^2 \bar{c}_n^2 s_k k^2 \ln^3(d_z n)}{n^{(m-1)/m}} \to 0 \tag{2.13.6}$$

for $m > 2$ as in Assumption 2.4.2. These rates are comparable and in certain cases may be more palatable than those presented in the main text, Assumption 2.3.1(vi). They come at the cost of slower rates of convergence for the weighted means as seen by comparing eqs. (2.13.3)–(2.13.4) to eqs. (2.3.1) and (2.3.2).

2.13.2. Practical Implementation

In practice, the constants $M_5, M_6, M_7, M_8^2, M_9^2$ from Lemmas 2.9.11–2.9.15 are roughly on the order of $\|Z\|_\infty$. We therefore reccomend setting

$$\begin{aligned}
\underline{c}_n &= \frac{1}{2\log^{1/2}(d_z n)} \max_{1\le i\le n} \|p^k(X_i)\|_\infty \max_{1\le i\le n} \|Z_i\|_\infty \\
\bar{c}_n &= \frac{3\log^{1/2}(d_z n)}{2} \max_{1\le i\le n} \|p^k(X_i)\|_\infty \max_{1\le i\le n} \|Z_i\|_\infty
\end{aligned}$$

and letting Λ_n be a set of points evenly spaced between $\underline{c}_n$ and $\bar{c}_n$. The cross validation procedure then can be implemented in the following steps.

1. *Split the sample into K_1 folds.*
2. *Consider a single pair of values for c_α, c_γ and designate a fold to hold out.*
3. *Estimate nuisance model parameters using $\lambda_{\gamma,j}^{pilot}$ and $\lambda_{\alpha,j}^{pilot}$ on the remaining folds.*
4. *Evaluate the resulting models on held out fold using non-penalized loss functions.*
5. *Repeat K times and record average loss over all folds.*
6. *Choose values of $c_{\gamma,j}$ and $c_{\alpha,j}$ with the lowest average loss.*

In practice we find this procedure works well with small K_1, around $K = 5$ and with $|\Lambda_n|$ on the order of about 10-20.

Chapter 3

Ordered, Unordered, and Minimal Monotonicity

3.1. Introduction

Economists have long used instrumental variables (IV) to identify the causal effect of an endogenous treatment choice on outcomes of interest. An IV is characterized by two core properties: it is an exogenous variable, and it affects the outcome only through its impact on the treatment choice. These properties, however, are insufficient to identify treatment effects. Early IV literature secures identification by invoking strong function form assumptions (Theil, 1953), which impose homogeneous treatment effects (Heckman and Robb, 1985).

In an influential paper, Imbens and Angrist (1994) investigate weaker assumptions that secure the nonparametric identification of a causal effect in a binary choice model with heterogenous agents. They introduce the monotonicity condition, which states that a change in the instrument must induce all agents to shift their choices towards the same treatment status. The monotonicity condition has several desirable properties: it is simple, intuitive, and renders the identification of the Local Average Treatment Effect (LATE) via the Two-Stage Least Squares (2SLS) estimand. The condition gives rise to notions such as the marginal treatment effect and response functions Heckman and Vytlacil (1999, 2005). Finally, the monotonicity condition of Imbens and Angrist (1994) is frequently regarded as the minimal

criteria necessary to ascribe causal interpretation to the 2SLS regression.[1]

Contrary to what one might expect, identification in IV models with multiple treatment choices is not a straightforward extension of the binary choice case. Multiple choices allow for a variety of distinct monotonicity conditions that collapse into the same condition when the treatment is limited to only two choices. Notably, Angrist and Imbens (1995) directly apply the monotonicity of Imbens and Angrist (1994) to the case of multiple treatments. Vytlacil (2006) shows that the condition is equivalent to assuming an ordered choice model. In contrast, the unordered monotonicity of Heckman and Pinto (2018) applies to settings with unordered treatment choices; it neither implies nor is implied by the monotonicity of Angrist and Imbens (1995). Despite their specific motivations, both conditions are equivalent to the Imbens and Angrist (1994) monotonicity in the case of binary choices and both allow for a causal interpretation for 2SLS estimands.

The properties of the ordered monotonicity of Angrist and Imbens (1995) and the unordered monotonicity of Heckman and Pinto (2018) raise several questions: Do these conditions share some of the choice restrictions they imply? Is there a weaker monotonicity condition underlying both conditions? If this weaker condition exists, is it equivalent to Imbens and Angrist (1994) in the binary choice model? Are there some criteria that enable us to characterize the similarities and differences among monotonicity conditions of IV models with multiple choices? What is the economic rationale that justifies the differences among these monotonicity conditions?

While the two ordered and unordered monotonicity criteria discussed above share similarities, little is known about the relationship between them. The IV literature seldom considers a meta-analysis across monotonicity conditions in settings with multiple treatments. This paper fills this gap by considering two independent but linked inquiries.

The first inquiry is on the relationship between ordered and unordered monotonicity. The

[1] Huber and Mellace (2012) is an example of a paper that considers identification of the LATE via 2SLS under alternate conditions.

shared key properties that suggest a deeper connection between the two criteria. We update the equivalence results of Vytlacil (2006) and Heckman and Pinto (2018) to provide symmetric characterizations of ordered and unordered monotonicity. These characterizations enable us to note some useful common properties of ordered and unordered monotonicity and set the stage for joint analysis of the two conditions.

The second inquiry considers whether these two monotonicity conditions can be subsumed by a broader criterion that would still enable useful causal analysis. By leveraging their symmetric characterizations, we show that both conditions share a common property which we term the *minimal monotonicity condition*. This minimal monotonicity condition is precisely what is required for the two stage least squares between any two instrument values to identify an interpretable causal parameter. Moreover, in general in no weaker condition would allow for such causal interpretability of two stage least squares estimands.[2] We provide a characterization for minimal monotonicity that allows the researcher to easily verify whether the condition holds.

In addition, we show that minimal monotonicity can be justified by a notion of choice rationality significantly weaker than those displayed by agents in ordered and unordered choice models. Settings where ordered and unordered monotonicity hold can thus be seen as particular instances of a broad class of choice models described by the minimal monotonicity condition. By analyzing the properties of minimal monotonicity, we thus hope to facilitate development of monotonicity conditions that may be suitable in a range of economic settings that are not neatly described by ordered or unordered choice models. We provide some natural economic examples where ordered and unordered monotonicity fail, but where minimal monotonicity may still allow researchers to conduct meaningful causal analysis.

This paper contributes to the theoretical literature on ordered and unordered choice models. It adds to the literature that extends the understanding and usage of monotonicity conditions

[2]Indeed, in cases where treatment is binary, we show that minimal monotonicity reduces exactly to the monotonicity of Imbens and Angrist (1994).

(Kamat, 2021; Mogstad et al., 2018; Mogstad and Torgovitsky, 2018; Hull, 2018). Our analyses are informative to a growing literature on empirical economics that examines non-standard monotonicity conditions to aid the identification and evaluation of treatment effects (Pinto, 2021; Kline and Walters, 2016; Mountjoy, 2021; Feller et al., 2016; Brinch et al., 2017; Kirkeboen et al., 2016). We additionally contribute to the literature tying monotonicity criterion to particular structural models (Vytlacil, 2002, 2006; Heckman and Pinto, 2018) by showing the minimal monotonicity is implied by a basic model of rationality.

This paper proceeds as follows. Section 3.2 reviews the prior literature on monotonicity conditions. Section 3.3 describes the IV model and introduces our notation. Section 3.4 discusses the content of ordered and unordered monotonicity conditions and revisits the equivalence results for ordered and unordered choice models. It explores the symmetry of equivalence results between these two models to motivate a novel monotonicity condition. Section 3.5 discusses the properties of the Minimal Monotonicity Condition. Section 3.6 discusses the economic content of the minimal monotonicity condition. Section 3.7 discuss some applications of monotonicity criteria that are economically justified. Section 3.8 concludes.

3.2. Literature Review

Historically, the traditional approach to evaluating IV models has been to use structural equations to describe the agent's choice (Goldberger, 1972; Heckman, 1976, 1979). Imbens and Angrist (1994) departed from the traditional IV literature based on structural equations. Using the language of potential outcomes (Rubin, 1974b, 1978a), they introduce the notion of *monotonicity*, which formalizes an intuitive assumption stating that an IV change induces all agents toward choosing the same treatment choice.[1]

[1] See also Angrist et al. (1996).

Angrist and Imbens (1995) extend this monotonicity condition to the case of multiple choices. They show that their monotonicity provides a causal interpretation of the conventional Two-Stage Square Least Squares (2SLS) estimand in models with endogenous choices and heterogeneous treatment responses. Their work spiked a substantial literature on both empirical and theoretical aspects of monotonicity conditions (Angrist et al., 2000; Barua and Lang, 2016; Dahl et al., 2017; Huber and Mellace, 2012, 2015; Imbens and Rubin, 1997; Klein, 2010; Small and Tan, 2007; Aliprantis, 2012; de Chaisemartin, 2017).[2]

Vytlacil (2002, 2006) bridge the gap between IV models that rely on monotonicity conditions and the previous literature that invokes structural equations. Vytlacil (2002) shows that the monotonicity condition of Imbens and Angrist (1994) is equivalent to the random threshold crossing model of Heckman and Vytlacil (1999, 2005, 2007a). Vytlacil (2006) shows the monotonicity criterion of Angrist and Imbens (1995) is equivalent to an ordered choice model with random thresholds. This model is examined by Cameron and Heckman (1998) and further studied by Carneiro et al. (2003), Cunha et al. (2007).

Unordered choice models have been studied mostly by literature on structural equations. A common approach is to assume that the equations that govern the treatment are generated by additively separable threshold-crossing models. Examples of this literature are Heckman and Vytlacil (2007b), Heckman et al. (2006, 2008). A substantial contribution to this literature is due to Lee and Salanié (2018), who studied the identification of causal effects for choice models defined by an arbitrary set of threshold-crossing rules. Heckman and Pinto (2018) connect the structural and monotonicity approaches. They present an economically motivated condition termed *unordered monotonicity* which applies to treatment values that do not have a natural order. Building upon Vytlacil (2002), they further show that unordered monotonicity can be equivalently expressed as a multivariate choice model with latent crossing thresholds.

Little is known about the shared features of ordered and unordered choice models. The

[2] Huber et al. (2017) consider weaker assumptions at the principal strata level, which are also employed by Frölich (2007).

rationale that generates an ordered choice model is considerably different from the motivation that justifies unordered choices. Not surprisingly, each model often carries distinct mathematical formalizations. A rare example of a comparative discussion between ordered and unordered choice models is Heckman et al. (2006). Their ordered choice model employs a partition of the real line by non-stochastic thresholds. The treatment choice indicates the interval that the latent stochastic index lies in this partition. In contrast, their unordered choice model employs a set of latent indexes that are additive in the observed and unobserved characteristics of the agent.

As mentioned, we perform a comparative analysis between ordered and unordered monotonicities. To do so, we revisit the monotonicity condition of Angrist and Imbens (1995) using new tools of analysis developed in Heckman and Pinto (2018).

3.3. Setup

Our IV model consists of three observed variables: a categorical instrument Z that takes N_Z values in $\mathcal{Z} = \{z_1, ..., z_{N_Z}\}$; a multiple treatment choice T that takes N_T values in $\mathcal{T} = \{t_1, ..., t_{N_T}\}$; and a real-valued outcome $Y \in \mathbb{R}$.[1] These variables belong to the probability space $(\mathcal{I}, \mathcal{F}, P)$ where $i \in \mathcal{I}$ denotes an individual. Most of the IV literature describes model assumptions via the potential outcome framework (Rubin, 1978b; Holland, 1986), where $Y_i(t, z)$ denotes the potential outcome for individual i when Z_i, T_i take values z, t and $T_i(z)$ denotes the potential choice for i when the instrument Z_i is set to the value z. The assumptions that characterize the core properties of the instrument are:

$$\text{Exogeneity Condition:} \quad Z_i \perp\!\!\!\perp (T_i(z), Y_i(t)) \text{ for all } (z, t, i) \in \mathcal{Z} \times \mathcal{T} \times \mathcal{I} \tag{3.3.1}$$

$$\text{Exclusion Restriction:} \quad Y_i(t, z) = Y_i(t, z') \text{ for all } z, z' \in \mathcal{Z}, \text{ and } (t, i) \in \mathcal{T} \times \mathcal{I} \tag{3.3.2}$$

$$\text{IV Relevance:} \quad T \not\perp\!\!\!\perp Z \text{ (Not statistically independent)} \tag{3.3.3}$$

[1] We suppress pre-treatment variables X from the model for the sake of notational simplicity. The analysis can be understood as conditioned on these variables.

The exogeneity assumption means that the instrument Z is as good as randomly assigned. The exclusion restriction implies that Z dos not directly cause Y and the IV relevance states that T and Z correlate. The IV model can be equivalently described by structural equations. The structural approach enable us to represent the individual's unobserved characteristics that generate selection bias by an unobserved random vector $\boldsymbol{V}$. The IV model is described by the following equations:

$$\text{Choice Equation}: \quad T = f_T(Z, \boldsymbol{V}), \tag{3.3.4}$$

$$\text{Outcome Equation}: \quad Y = f_Y(T, \boldsymbol{V}, \epsilon), \tag{3.3.5}$$

$$\text{Independence Condition}: \quad Z, \boldsymbol{V}, \epsilon \text{ are statistically independent} \tag{3.3.6}$$

The choice equation (3.3.4) means that T is cased by the instrument Z and unobserved characteristics $\boldsymbol{V}$, while the outcome equation (3.3.5) states that Y is caused by choice T, unobserved characteristics $\boldsymbol{V}$ and an unobserved error term ϵ that is exogenous.[2] Functions $f_T(\cdot)$ and $f_Y(\cdot)$ are not observed and can take arbitrary functional forms. The counterfactual (or potential) choice when the instrument were fixed to a value $z \in \mathcal{Z}$ is given by $T(z) \equiv f_T(z, \boldsymbol{V})$ and the counterfactual outcome generated by fixing T to a value t, that is $Y(t) \equiv f_Y(t, \boldsymbol{V}, \epsilon)$.[3] The independence condition (3.3.6) states that Z is statistically independent of the individual's unobserved characteristics $\boldsymbol{V}$ and the error term ϵ. The condition implies the exogeneity condition (3.3.1) of the potential outcome framework.[4] It also implies the

[2] Error term ϵ is used so that Y conditioned on $\boldsymbol{V}$ and Z is not deterministic.

[3] See Heckman and Pinto (2014) and Pinto and Heckman (2021) for a discussion on causal models and the fixing operator.

[4] Indeed, the counterfactuals $Y(t), T(z)$ are a function of $\boldsymbol{V}$ and ϵ, which, according to (3.3.6), are statistically independent of Z. The exclusion restriction (3.3.2) arises because Z is not an argument in the outcome equation (3.3.6). IV relevance (3.3.3) corresponds to the assumption that Z causes T in the choice equation (3.3.4).

following matching (or unconfoundness) condition:[5]

$$\text{Matching Condition:} \quad Y(t) \perp\!\!\!\perp T|\boldsymbol{V} \text{ for all } t \in \mathcal{T}. \tag{3.3.7}$$

Condition (3.3.7) states that $Y(t)$ becomes statistically independent of T when conditioning for $\boldsymbol{V}$. If $\boldsymbol{V}$ were observable, we would be able to evaluate the counterfactual outcome $Y(t)$ by conditioning the outcome Y on $T = t$ and $\boldsymbol{V}$.

The primary lesson of the matching condition (3.3.7) is that the identification of causal effects hinges on controlling for the unobserved characteristics $\boldsymbol{V}$. Identification strategies of IV methods can be understood as econometric procedures that seek to exploit the exogenous variation of Z to control for $\boldsymbol{V}$. Controlling for unobserved characteristics is a daunting task since $\boldsymbol{V}$ is unobserved and can have an arbitrary dimension. The *response vector* $\boldsymbol{S}$ in (3.3.8) facilitates this task:

$$\boldsymbol{S} = [T(z_1), ..., T(z_{N_Z})]^\intercal, \quad \text{supp}(\boldsymbol{S}) \equiv \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_S}\} \tag{3.3.8}$$

The response vector $\boldsymbol{S}$ is defined as the unobserved random vector of dimension $N_Z \times 1$ that stacks the counterfactual choices $T(z)$ across the IV-values z in $\mathcal{Z}$. Elements of the support of the response vector, $\boldsymbol{s} \in \text{supp}(\boldsymbol{S})$, are called *response-types*. Each element $T(z)$ in $\boldsymbol{S}$ may take any of the N_T values in $\mathcal{T}$. Thus, the number of possible response-types totals $N_T^{N_Z}$. The LATE model of Imbens and Angrist (1994) investigates the case of a binary instrument $\mathcal{Z} = \{z_0, z_1\}$ and a binary treatment $\mathcal{T} = \{t_0, t_1\}$. The response vector $\boldsymbol{S} = [T(z_0), T(z_1)]^\intercal$, admits four possible response-types: never-takers $\boldsymbol{s}_{\text{nt}} = [t_0, t_0]^\intercal$, compliers $\boldsymbol{s}_{\text{c}} = [t_0, t_1]^\intercal$, always-takers $\boldsymbol{s}_{\text{at}} = [t_1, t_1]^\intercal$, and defiers $\boldsymbol{s}_{\text{d}} = [t_1, t_0]^\intercal$. Note that the choice T is fully determined by the instrument Z for any given response-type $\boldsymbol{s}$. For instance, compliers $\boldsymbol{s}_{\text{c}}$ choose $T = t_1$ for $Z = z_1$ and $T = t_0$ for $Z = z_0$. We can express the treatment choice as a function of $\boldsymbol{S}$ and Z

[5] The independence condition (3.3.6) implies that $\epsilon \perp\!\!\!\perp Z|\boldsymbol{V}$. Thereby $f_Y(t, \boldsymbol{V}, \epsilon) \perp\!\!\!\perp f_T(Z, \boldsymbol{V})|\boldsymbol{V} \Rightarrow Y(t) \perp\!\!\!\perp T|\boldsymbol{V}$.

as $T = \left[\mathbf{1}[Z = z_1], ..., \mathbf{1}[Z = z_{N_Z}]\right] \cdot \boldsymbol{S}$, where $\mathbf{1}[\cdot]$ denotes then indicator function.

The response vector $\boldsymbol{S}$ simplifies the identification problem by playing the role of a balancing score for $\boldsymbol{V}$. This means that $\boldsymbol{S}$ is a function of $\boldsymbol{V}$ since each counterfactual $T(z)$ is a function of $\boldsymbol{V}$, and that $\boldsymbol{S}$ preserves the matching property (3.3.7). Indeed, $Y(t) \perp\!\!\!\perp T|\boldsymbol{S}$ holds because, given $\boldsymbol{S}$, T depends only on Z which is independent of $Y(t)$. This matching property enables us to connect observed expectations from data with the unobserved counterfactuals we seek to evaluate via the following equation:[6]

$$\underbrace{E(Y|T=t,Z=z)P(T=t|Z=z)}_{\text{Observed}} = \sum_{\boldsymbol{s}\in\text{supp}(\boldsymbol{S})} \underbrace{\mathbf{1}[T=t|\boldsymbol{S}=\boldsymbol{s},Z=z]}_{\text{Known}} \cdot \underbrace{E(Y(t)|\boldsymbol{S}=\boldsymbol{s})P(\boldsymbol{S}=\boldsymbol{s})}_{\text{Unobserved}}. \tag{3.3.9}$$

The left-hand side of equation (3.3.9) comprises of the observed quantities, namely, the conditional expectation $E(Y|T=t,Z=z)$ and propensity score $P(T=t|Z=z)$.[7] The first term of the right-hand side of the equation is nonrandom since T is a deterministic function of the instrument Z and the response type $\boldsymbol{S}$. The second term on the right-hand side is unobserved. It comprises expected value of counterfactual outcomes conditioned on response-types $E(Y(t)|\boldsymbol{S}=\boldsymbol{s})$ and response-type probabilities $P(\boldsymbol{S}=\boldsymbol{s})$.

Equation (3.3.9) characterizes the identification problem of IV models as the solution of a system of linear equations. We seek to identify the unobserved quantities on the right-hand side of equation (3.3.9) (outcome counterfactuals and response-type probabilities) using the observed quantities on the left-hand side of equation (3.3.9) (outcome expectations and propensity scores). The general solution to this problem requires some matrix notation.

The response matrix $\boldsymbol{R}$ is central to our analysis. It organizes the response-types in supp($\boldsymbol{S}$)

[6] See Heckman and Pinto (2018) for a proof.

[7] Equation (3.3.9) holds for any real-valued function $g : \mathbb{R} \to \mathbb{R}$ and for $(z,t) \in \mathcal{Z} \times \mathcal{T}$, that is:

$$E(g(Y)|T=t,Z=z)P(T=t|Z=z) = \sum_{\boldsymbol{s}\in\text{supp}(\boldsymbol{S})} \mathbf{1}[T=t|\boldsymbol{S}=\boldsymbol{s},Z=z] \cdot E(g(Y(t))|\boldsymbol{S}=\boldsymbol{s})P(\boldsymbol{S}=\boldsymbol{s}).$$

Setting $g(Y) = \mathbf{1}[Y = y]; y \in \mathbb{R}$ generates an equation for the probabilities of counterfactual outcomes. Setting $g(Y) = 1$ generates an equation that relates propensity scores and response-types probabilities.

into a $N_Z \times N_S$ array where each column displays a response-type and each row corresponds to an instrument value:

$$\boldsymbol{R} \equiv [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_S}] \in \mathcal{T}^{N_Z \times N_S}. \tag{3.3.10}$$

We use $\boldsymbol{B}_t = \mathbf{1}[\boldsymbol{R} = t]$ for the $N_Z \times N_S$ binary matrix that takes value one if the respective entry in $\boldsymbol{R}$ is t and zero otherwise. The value in the z-th row and $\boldsymbol{s}$-th column of $\boldsymbol{B}_t$ is given by $\boldsymbol{B}_t[z, \boldsymbol{s}] = \mathbf{1}[T = t | \boldsymbol{S} = \boldsymbol{s}, Z = z]$. The response matrix in the LATE model is given by:

$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_{\text{nt}} & \boldsymbol{s}_{\text{c}} & \boldsymbol{s}_{\text{at}} & \boldsymbol{s}_{\text{d}} \end{matrix} \\ \begin{bmatrix} t_0 & t_0 & t_1 & t_1 \\ t_0 & t_1 & t_1 & t_0 \end{bmatrix} \end{array} \begin{matrix} z_0 \\ z_1 \end{matrix} \quad \therefore \boldsymbol{B}_{t_0} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_{\text{nt}} & \boldsymbol{s}_{\text{c}} & \boldsymbol{s}_{\text{at}} & \boldsymbol{s}_{\text{d}} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array}, \boldsymbol{B}_{t_1} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_{\text{nt}} & \boldsymbol{s}_{\text{c}} & \boldsymbol{s}_{\text{at}} & \boldsymbol{s}_{\text{d}} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \end{array}. \tag{3.3.11}$$

In this notation, we can rewrite equation (3.3.9) as:

$$\boldsymbol{Q}_Z(t) \odot \boldsymbol{P}_Z(t) = \boldsymbol{B}_t \cdot \big(\boldsymbol{Q}_S(t) \odot \boldsymbol{P}_S\big) \quad \text{for all } t \in \mathcal{T}, \tag{3.3.12}$$

where $\boldsymbol{Q}_Z(t) = [E(Y|T = t, Z = z_1), ..., E(Y|T = t, Z = z_{N_Z})]$, $\boldsymbol{P}_Z(t) = [P(T = t|Z = z_1), ..., P(T = t|Z = z_{N_Z})]$ are the observed vectors of outcome expectations and propensity scores. $\boldsymbol{Q}_S(t) = [E(Y(t)|\boldsymbol{S} = \boldsymbol{s}_1), ..., E(Y(t)|\boldsymbol{S} = \boldsymbol{s}_{N_S})]$, $\boldsymbol{P}_S = [P(\boldsymbol{S} = \boldsymbol{s}_1), ..., P(\boldsymbol{S} = \boldsymbol{s}_{N_S})]$ are the unobserved vectors of counterfactual outcomes and response-type probabilities. The symbol $\odot$ denotes element-wise (Hadamard) multiplication.

Equation (3.3.12) serves two purposes in our analysis. Firstly, it establishes that the identification of causal parameters stems from the rank properties of the binary matrices $\boldsymbol{B}_t$. If $\boldsymbol{B}_t$ were a square full-rank matrix, we would be able to identify counterfactual outcomes by inverting $\boldsymbol{B}_t$, that is, $\boldsymbol{Q}_S(t) = (\boldsymbol{B}_t^{-1}\boldsymbol{Q}_Z(t) \odot \boldsymbol{P}_Z(t)) \div (\boldsymbol{B}_t^{-1}\boldsymbol{P}_Z(t))$, where $\div$ denotes element-wise division. More broadly, Heckman and Pinto (2018) show that, for any given

subset of response-types $\mathcal{S} \subset \text{supp}(\boldsymbol{S})$,

$$E(Y(t)|\boldsymbol{S} \in \mathcal{S}) \text{ is identified if and only if } \boldsymbol{b}(\mathcal{S})'(\boldsymbol{I} - \boldsymbol{B}_t^+ \boldsymbol{B}_t)\boldsymbol{b}(\mathcal{S}) = 0, \tag{3.3.13}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{B}_t^+$ is the Moore-Penrose pseudo-invesrse of $\boldsymbol{B}_t$ and $\boldsymbol{b}(\mathcal{S}) = [\mathbf{1}[\boldsymbol{s}_1 \in \mathcal{S}], ..., \mathbf{1}[\boldsymbol{s}_{N_S} \in \mathcal{S}]]'$ is a binary vector that indicates which response-type belongs to $\mathcal{S}$.[8]

Equation (3.3.12) also helps characterize how monotonicity conditions secure the identification of causal parameters. The equation entails a fundamental identification problem. The number of known parameters in its left-hand side totals $N_Z \cdot N_T$. The number of unknown parameters in the right-hand side is proportional to the number response-types in $\mathcal{S}$, which totals $N_T^{N_Z}$. An identification problem arises since the number of known parameters grows linearly in N_Z while the number of unknowns grows exponentially in N_Z. Monotonicity conditions solve this problem by assuming choice restrictions that systematically eliminate potential response-types in $\mathcal{S}$. They effectively equate the growth rate of known and unknown parameters. For instance, the monotonicity condition of Imbens and Angrist (1994) reduces the number of response-types of the binary choice model from 2^{N_Z} to $N_Z + 1$.

3.4. Ordered and Unordered Monotonicity

Monotonicity conditions are choice restrictions that eliminate response-types systematically. As discussed previously, Angrist and Imbens (1995) and Heckman and Pinto (2018) provide monotonicity criterions for ordered and unordered choice models, respectively. We will refer to these conditions as ordered monotonicity (3.4.1) and unordered monotonicity (3.4.2) for the sake of clarity:

Ordered Monotonicity (OM): For any $z, z' \in \mathcal{Z}$ either,

[8] If $E(Y(t)|\boldsymbol{S} \in \mathcal{S})$ is identified, then it can be evaluated by the expression $E(Y(t)|\boldsymbol{S} \in \mathcal{S}) = \frac{\boldsymbol{b}(\mathcal{S})'\boldsymbol{B}_t^+\left(\boldsymbol{Q}_Z(t) \odot \boldsymbol{P}_Z(t)\right)}{\boldsymbol{b}(\mathcal{S})'\boldsymbol{B}_t^+\boldsymbol{P}_Z(t)}$.

$$\begin{aligned} T_i(z) \geq T_i(z') \text{ for all } i \in \mathcal{I} \\ \text{or} \quad T_i(z) \leq T_i(z') \text{ for all } i \in \mathcal{I}. \end{aligned} \tag{3.4.1}$$

Unordered Monotonicity (UM): For any $z, z' \in \mathcal{Z}$ and any $t \in \mathcal{T}$ either,

$$\begin{aligned} \mathbf{1}[T_i(z) = t] \geq \mathbf{1}[T_i(z') = t] \text{ for all } i \in \mathcal{I} \\ \text{or} \quad \mathbf{1}[T_i(z) = t] \leq \mathbf{1}[T_i(z) = t] \text{ for all } i \in \mathcal{I} \end{aligned} \tag{3.4.2}$$

OM (3.4.1) captures the notion that a change in instrumental values produces incentives that either move all agents towards weakly "higher" treament values or move all agents towards weakly "lower" treatment values. The condition can be understood as stating that an instrumental change that induces one agent to increase their treatment choice cannot cause another agent to decrease their treatment choice. The condition requires an ordinal treatment, such as years of schooling.

UM (3.4.2) states that for each treatment, each instrumental change must either move all agents weakly towards that treatment or weakly away from the treatment. This differs from OM (3.4.1) as it compares *the indicator function* of the treatment instead of the treatment value itself. Because of this, UM (3.4.2) does not require ordered treatments, making it relevant for settings where the treatment has no natural ordering such as analysis of college major choice or neighborhood effects.[1]

Importantly, both OM (3.4.1) and UM (3.4.2) enable the researcher to identify a mixture of Local Average Treatment Effects (LATEs) with identifiable weights and both conditions ascribe causal interpretations to the estimands of Two-Stage Least Squares (2SLS) regressions.

[1]Additionally, Heckman and Pinto (2018) show that UM occurs naturally in economic settings where choice incentives weakly increase among all treatment choices as the instrument varies.Buchinsky and Pinto (2021) use revealed preference analysis to show how choice incentives induced by the instrumental variable generate a range of monotonicity conditions.

3.4.1. Expressing Monotonicities as Sequences of Counterfactual Choices

Because the definition of OM (3.4.1) compares treatment values, it requires that $\mathcal{T}$ be an ordered set. We propose a slightly more inclusive definition of ordered monotonicity that does not require an ordered treatment. The central property of ordered monotonicity is a mapping between a sequence of IV values and some sequence of treatment values in which higher rankings of Z correspond to higher rankings of Z. The following formula expresses this criterion:

OM Sequence: There exist a sequencing of $\mathcal{Z}$, $(z_1, \ldots, z_{N_Z})$, and a strict ordering on $\mathcal{T}$ such that:[2]

$$(T_i(z_1), \ldots, T_i(z_{N_Z})) \text{ is an increasing sequence in } \mathcal{T} \text{ for any } i \in \mathcal{I}. \tag{3.4.3}$$

The OM sequential criteria (3.4.3) generates the OM condition (3.4.1) whenever the ordering $\mathcal{T}$ is assumed, however it does not require a specific ordering on $\mathcal{T}$ a priori. In Section 3.7 we will demonstrate the usefulness of this more inclusive definition with a plausible research design that generates OM-Sequence (3.4.3) on a treatment space that has no natural ordering.

We can also characterize the UM condition in (3.4.2) in terms of a sequence of counterfactual choices:

UM Sequence: For each $t \in \mathcal{T}$ there exists a sequencing of $\mathcal{Z}$, $(z_1^{(t)}, \ldots, z_{N_Z}^{(t)})$ such that:

$$(\mathbf{1}[T_i(z_1^{(t)}) = t], \ldots, \mathbf{1}[T_i(z_{N_Z}^{(t)}) = t]) \text{ is weakly increasing for any } i \in \mathcal{I}. \tag{3.4.4}$$

UM Sequence (3.4.4) differs from OM Sequence (3.4.3) in two significant ways. First, the sequence of IV values in the unordered case can differs across treatment values while the IV sequence of ordered case remains the same for all $t \in \mathcal{T}$. Second, UM Sequence (3.4.4)

[2]A strict ordering is one such that for any $t, t' \in \mathcal{T}$ with $t \neq t'$ exactly one of $(t' \geq t)$ or $(t \geq t')$ is true.

utilizes treatment indicators, while the OM Sequence (3.4.3) employs the treatment values themselves.

It is easy to see that the OM and UM sequence characterizations in (3.4.3) and (3.4.4) are equivalent for a binary treatment. However, this equivalence between ordered and unordered monotonicity breaks down for choice models with three or more treatment choices. In general, ordered monotonicity does not imply unordered monotonicity nor vice versa. To partially demonstrate why this is the case, consider an ordering on the treatments where $t_1 \leq t_2 \leq t_3$ and the following two pairs of treatment response patterns:

$$\begin{matrix} \boldsymbol{s}_a & \boldsymbol{s}_b \\ \begin{pmatrix} t_1 & t_2 \\ t_2 & t_3 \end{pmatrix} & \begin{matrix} z \\ z' \end{matrix} \end{matrix} \qquad \begin{matrix} \boldsymbol{s}_c & \boldsymbol{s}_d \\ \begin{pmatrix} t_1 & t_3 \\ t_2 & t_2 \end{pmatrix} & \begin{matrix} z \\ z' \end{matrix} \end{matrix} \tag{3.4.5}$$

The treatment response patterns displayed by agents $\boldsymbol{s}_a$ and $\boldsymbol{s}_b$ are natural under ordered monotonicity, they could can be rationalized by a research design where z' provides uniformly greater treatment incentives than z. However, they cannot both exist in a response matrix that satisfies unordered monotonicity. By examining UM Sequence (3.4.4) we can see that there would be no possible sequencing of the instruments that would allow the sequence of t_2 indicators to be weakly increasing for all agents.

Similarly, the treatment response patterns displayed by agents $\boldsymbol{s}_c$ and $\boldsymbol{s}_d$ in (3.4.5) are not prohibited by unordered monotonicity; they can both be rationalized by a research design where instrument z' explicitly incentivizes treatment t_2. Despite this, they cannot both be present in a response matrix that satisfies ordered monotonicity under the ordering on the treatments given above. Since the switch from instrument z to instrument z' induces agent $\boldsymbol{s}_c$ to switch to a "lower" treatment while inducing agent $\boldsymbol{s}_d$ to switch to a "higher" treatment, there can be no sequencing on the instruments satisfying OM Sequence (3.4.3).

The two conditions are also not mutually exclusive, it is possible for a response matrix to

satsify both ordered and unordered monotonicity, even when the treatment is multi-valued. Moreover, the lack of nesting between ordered and unordered monotonicity does not change if we consider all possible orderings on the treatment space. For a complete example of this and more in depth discussion, refer to Section 3.11.

3.4.2. Characterizations of Unordered and Ordered Monotonicity

We first present an updated version of the unordered equivalence result in Heckman and Pinto (2018). This updated version is presented symmetrically to a later equivalence result for ordered monotonicity and will facilitate comparasion of the two conditions.

Theorem 3.4.1 (Unordered Equivalence)**.** *The following statements are equivalent:*

(i). For each $t \in \mathcal{T}$ there is a sequence of instruments $(z_1^{(t)}, ..., z_{N_T}^{(t)})$ such that UM Sequence (3.4.4) holds.

(ii). Given any $t \in \mathcal{T}$ and any $k \in \{1, ..., N_Z - 1\}$, we have that

$$\mathbf{1}[T_i(z_{k+1}^{(t)}) = t] \geq \mathbf{1}[T_i(z_k^{(t)}) = t] \text{ for all } i \in \mathcal{I}..$$

(iii). For any $t \in \mathcal{T}$ and $t', t'' \neq t$ there are no 2×2 submatrices in $\boldsymbol{R}$ of the form:

$$\begin{pmatrix} t & t'' \\ t' & t \end{pmatrix} \text{ or } \begin{pmatrix} t' & t \\ t & t'' \end{pmatrix}. \tag{3.4.6}$$

(iv). For the unordered verification matrix $\Psi_{\boldsymbol{U}}$ defined below, $\|\Psi_{\boldsymbol{U}}\| = 0$;

$$\Psi_{\boldsymbol{U}} \equiv \left((\mathbb{1} - \boldsymbol{U})^\intercal \boldsymbol{U}\right) \odot \left((\mathbb{1} - \boldsymbol{U})^\intercal \boldsymbol{U}\right)^\intercal. \tag{3.4.7}$$

where $\mathbb{1}$ denotes a matrix of all ones, $\odot$ denotes the Hadamard (element-wise) product

and:

$$\boldsymbol{U} \equiv \begin{bmatrix} \boldsymbol{B}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \boldsymbol{B}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \cdots & \boldsymbol{B}_{N_T} \end{bmatrix}. \tag{3.4.8}$$

(v). For each $t \in \mathcal{T}$, there are real-valued functions $\varphi(\cdot, t)$ and $\zeta(\cdot, t)$ such that the treatment choice T can be rationalized by:

$$\mathbf{1}[T = t] = \mathbf{1}\left[\zeta(Z, t) \geq \varphi(\boldsymbol{V}, t)\right],$$

where $\zeta(z_{k+1}^{(t)}, t) > \zeta(z_k^{(t)}, t)$ for $k = 1, ..., N_Z - 1$ and any t.

Proof. See Appendix A. □

The first two items of Theorem 3.4.1 reflect the discussion in Section 3.4 relating UM-Sequence (3.4.4) to the classical definition of unordered monotonicity introduced in Heckman and Pinto (2018) and restated in (3.4.1).

Item *(iii)* resembles the discussion above and states that unordered monotonicity can be verified by individually checking each 2×2 submatrix of $\boldsymbol{R}$. We can see that the appearance of one of the restricted submatrices in (3.4.6) prevents the existence of a sequence of instruments that would make the sequence of treatment t indicators increasing for all agents. Unfortunately, this requirement may be difficult to verify in practice, since the number of 2×2 submatices is growing exponentially with the dimensions of the response matrix. Item *(iv)* of the theorem provides a practical method of verifying the condition using matrix algebra. Item *(v)* is faimilar to the literature and provides an equivalence between unordered monotonicity and separability conditions such as in Vytlacil (2002). For further discussion, Heckman and Pinto

(2018) describe other useful aspects of each of the equivalent statements, *(iii)* and *(v)* above.

We next present an equivalence result for Ordered Monotonicity.

Theorem 3.4.2 (Ordered Equivalence)**.** *The following statements are equivalent:*

(i). There is is a sequence on $\mathcal{Z}$, $(z_1, \ldots, z_{N_Z})$ and a strict ordering on $\mathcal{T}$ that satisfies the requirement of OM-Sequence (3.4.3).

(ii). There is a strict ordering on $\mathcal{T}$ such that for any $k \in \{1, \ldots, N_Z - 1\}$ and any t:

$$\mathbf{1}[T_i(z_{k+1}) \geq t] \geq \mathbf{1}[T_i(z_k) \geq t] \text{ for all } i \in \mathcal{I}.$$

(iii). There is a strict ordering on $\mathcal{T}$ such that for any $t < t''$ and $t' > t'''$ there are no 2×2 submatrices of $\boldsymbol{R}$ of the form either

$$\begin{pmatrix} t & t' \\ t'' & t''' \end{pmatrix} \quad or \quad \begin{pmatrix} t' & t \\ t''' & t'' \end{pmatrix}; \tag{3.4.9}$$

(iv). There is a strict ordering on $\mathcal{T}$ such that for the ordered verification matrix $\Psi_{\boldsymbol{O}}$ defined below, $\|\Psi_{\boldsymbol{O}}\| = 0$;

$$\Psi_{\boldsymbol{O}} \equiv \left((\mathbb{1} - \boldsymbol{O})^\intercal \boldsymbol{O}\right) \odot \left((\mathbb{1} - \boldsymbol{O})^\intercal \boldsymbol{O}\right)^\intercal, \tag{3.4.10}$$

where $\mathbb{1}$ indicates a matrix of all ones, $\odot$ represents the Hadamard (element-wise) product, and:

$$\boldsymbol{O} \equiv \left[\boldsymbol{B}^*_{t_1}, \ldots, \boldsymbol{B}^*_{t_{N_T}}\right];$$

(v). There is a strict ordering on $\mathcal{T}$ such that for some real-valued functions $\varphi(\cdot, t)$ and $\zeta(\cdot, t)$

the treatment choice can be rationalized by

$$\mathbf{1}[T \geq t] = \mathbf{1}[\zeta(Z,t) \geq \varphi(\boldsymbol{V},t)],$$

where $\zeta(z_{k+1},t) > \zeta(z_k,t)$ *for* $k = 1,\ldots,N_Z - 1$ *and any* t*.*

Proof. See Appendix A □

Theorem 3.4.2 extends Vytlacil (2006) in a fashion that enables us to compare ordered and unordered monotonicity conditions. The first and second items of the ordered equivalence result reconcile the two notions of ordered monotonicity presented above. It shows that if OM-Sequence (3.4.3) holds, we can find an ordering on $\mathcal{T}$ that satisfies the typical definition of ordered monotonicity and vice versa; if there is an ordering on $\mathcal{T}$ that satisfies ordered monotonicity we can find a sequence on $\mathcal{Z}$ to satisfy OM-Sequence (3.4.3).[3] Item *(iii)* of Theorem 3.4.2 provides a similar insight to Item *(iii)* of Theorem 3.4.1, namely that ordered monotonicity can be verified simply by looking at the 2×2 submatrices of the response matrix $\boldsymbol{R}$. Item *(iv)* provides a tractable method for verifying this property.

The final item of the theorem restates the equivalence result of Vytlacil (2006) and shows that assuming ordered monotonicity is equivalent to taking an ordered choice behavioral model. While this result is familar to the literature, we provide an alternative proof in Section 3.9 using properties of *lonesum binary matrices*; a concept we borrow from the information theory literature (Ryser, 1957).

Symmetries between Ordered and Unordered Monotonicity

The characterizations of ordered monotonicity in Theorem 3.4.2 are symmetric to those of unordered monotonicity in Theorem 3.4.1. We have already discussed the usefulness of some of these specific symmetries above. For example, the symmetry between the sequential characterizations of ordered and unordered monotonicities provides an easy way of seeing that

[3] In particular we can take the sequence that orders z' after z if $T_i(z') \geq T_i(z)$ for all $i \in \mathcal{I}$.

ordered and unordered monotonicity are equivalent in the case of a binary treatment. Other symmetries are new to our discussion and are worth briefly mentioning. The symmetric matrix verification characterizations provides an easy way to verify if a response matrix satisfies both the ordered and unordered monotonicity conditions by checking if $\|\Psi_{\boldsymbol{U}}\| + \|\Psi_{\boldsymbol{O}}\| = 0$. Verifying this allows researchers to take advantage of both sets of identification results.

The symmetry between the restricted 2×2 submatrices in ordered and unordered monotonicity provides insight on how a response matrix could satisfy ordered monotonicity but not unordered monotonicity and vice versa. More importantly, however, the similarity between the two restricted submatrices in Theorems 3.4.1 and 3.4.2 suggests a common condition shared by both criteria. In particular, note that both restricted patterns in (3.4.6) and (3.4.9) prevent any two agents from having exactly opposite treatment responses for the same instrumental variable shift. We term this common restriction the *Minimal Monotonicity Condition* and analyze its properties in Section 3.5.

3.5. The Minimal Monotonicity Condition

The minimal monotonicity (MM) condition (3.5.1) is a weak criteria shared by both ordered and unordered conditions. It is determined by a symmetric restriction that is common to Theorems 3.4.1 and 3.4.2. Indeed, it turns out that minimal monotonicity is the core common property of ordered and unordered monotonicity that enables the 2SLS estimand to achieve causal interpretability.

Minimal Monotonicity (MM): For any pair of instruments $z, z' \in \mathcal{Z}$ and any pair of treatments $t, t' \in \mathcal{T}$ either

$$\begin{aligned} &\mathbf{1}[T_i(z) = t]\mathbf{1}[T_i(z') = t'] \geq \mathbf{1}[T_i(z) = t']\mathbf{1}[T_i(z') = t] \,\forall i \ \in \mathcal{I} \\ \text{or} \quad &\mathbf{1}[T_i(z) = t]\mathbf{1}[T_i(z') = t'] \leq \mathbf{1}[T_i(z) = t']\mathbf{1}[T_i(z') = t] \,\forall\, i \in \mathcal{I}. \end{aligned} \tag{3.5.1}$$

The first row in (3.5.1) states that an instrumental change from z to z' incentives *all* agents to shift their choice away from t and towards t'. The second row in (3.5.1) describes the opposite behavior. In summary, the MM condition states that an intrumental change that induces an agent to switch their choice from t to t' cannot induce another another agent to switch their choice from t' to t. Lemma 3.5.1 provides an equivalent characterization of the MM condition in terms of response-types.

Lemma 3.5.1. *Minimal monotonicity MM holds if and only if for all distinct instruments $z, z' \in \mathcal{Z}$ and all distinct treatments $t, t' \in \mathcal{T}$, there are* no *response-types $\boldsymbol{s}, \boldsymbol{s}' \in \text{supp}(\boldsymbol{S})$ such that*

$$\begin{pmatrix} \boldsymbol{s}[z] & \boldsymbol{s}'[z] \\ \boldsymbol{s}[z'] & \boldsymbol{s}'[z'] \end{pmatrix} = \begin{matrix} \boldsymbol{s} & \boldsymbol{s}' \\ \begin{pmatrix} t & t' \\ t' & t \end{pmatrix} & \begin{matrix} z \\ z' \end{matrix} \end{matrix} \tag{3.5.2}$$

Proof. See Appendix A □

Lemma 3.5.1 presents the prohibited pattern of 2×2 submatrices of the response matrix $\boldsymbol{R}$ induced by MM. The pattern is the common intersection between the submatrix characterizations in item *(iii)* of Theorems 3.4.1 and 3.4.2. Lemma 3.5.2 establishes that MM is *strictly* weaker than MM and OM.

Lemma 3.5.2. *The following relationships are true of ordered, unordered, and minimal monotonicity:*

1. *UM* $\Rightarrow$ *MM, but MM* $\not\Rightarrow$ *UM*
2. *OM* $\Rightarrow$ *MM, but MM* $\not\Rightarrow$ *OM*

Proof. See Appendix A □

To see why minimal monotonicity is crucial for the interpretability of 2SLS, it is useful to quickly define and discuss interpretable causal parameters.

3.5.1. Interpretable Causal Parameters

We follow an established literature that defines a meaningful causal parameter τ as a weighted average of local average treatment effects with positive weights:[1]

$$\tau = \sum_{\{t,t'\}, t \neq t'} \omega_{t,t'} \mathbb{E}[Y(t) - Y(t') \mid \boldsymbol{S} \in \mathcal{S}_{t,t'}] \quad \text{with } \omega_{t,t'} = 0 \text{ or } \omega_{t',t} = 0. \tag{3.5.3}$$

Here $\mathcal{S}_{t,t'}$ denotes a set of response types may vary according to the treatments being compared and $\omega_{t,t'} \geq 0$ are positive weights. The defining idea is that each treatment pair is only represented once, so we cannot have a positive weight on both $\mathbb{E}[Y(t) - Y(t') \mid \boldsymbol{S} \in \mathcal{S}_{t,t'}]$ and $\mathbb{E}[Y(t') - Y(t) \mid \boldsymbol{S} \in \mathcal{S}_{t',t}]$. The absence of negative weights allows this causal parameter to give us meaningful insight into the direction of the treatment effects.

Angrist and Imbens (1995) demonstrate that, under ordered monotonicity, the 2SLS estimand identifies such a meaningful causal parameter using a binary instrument with multiple treatments. Heckman and Pinto (2018) show a similar result for unordered monotonicity using comparisons of the outcome Y for any two instruments $z, z' \in \mathcal{Z}$. The equivalence result for minimal monotonicity in Theorem 3.5.1 establishes that it is indeed MM that is the driving force behind both of these identification results.

Lemma 3.5.2 provides some intuition for why this is the case. Consider the difference in average outcome between two values of the instrument, as in the numerator of a 2SLS estimand. This difference always has a unique decomposition into a weighted average of treatment effects among all the (ordered) pairs of possible treatment values.[2] The restriction

[1] For examples of works that adopt this criteria, see Angrist and Imbens (1995), Heckman and Urzúa (2010), Kirkeboen et al. (2016), Mogstad et al. (2021).

[2] See Section 3.10 for a discussion of the exact forms of this decomposition as well as the 2SLS estimands

on the response matrix imposed by minimal monotonicity (3.5.2) means that if one pair of treatment values, (t, t'), is represented in this weighted sum, the opposite pair, (t', t), cannot also be represented. So, minimal monotonicity is sufficient for this difference to satisfy the condition for an interpretable causal parameter (3.5.3). Moreover, because this decomposition is unique, we can show that minimal monotonicity is necessary for interpretability as well.

3.5.2. Equivalence Results

We now provide a set of equivalent characterizations of the minimal monotonicity condition in the spirit of the results for unordered and ordered monotonicity in Theorems 3.4.1 and 3.4.2.

Theorem 3.5.1 (Minimal Monotonicity Equivalence)**.** *The following statements are equivalent:*

(i). For any distinct pair of instruments $z, z' \in \mathcal{Z}$ and any pair of treatments, $t, t' \in \mathcal{T}$, we have either:

$$\begin{aligned} &\mathbf{1}[T_i(z) = t]\mathbf{1}[T_i(z') = t'] \geq \mathbf{1}[T_i(z) = t']\mathbf{1}[T_i(z') = t] \; \forall i \; \in \mathcal{I} \\ \text{or} \quad &\mathbf{1}[T_i(z) = t]\mathbf{1}[T_i(z') = t'] \leq \mathbf{1}[T_i(z) = t']\mathbf{1}[T_i(z') = t] \; \forall \, i \in \mathcal{I}. \end{aligned} \tag{3.5.4}$$

(ii). There are no 2×2 submatrices of $\boldsymbol{R}$ of the form in (3.5.2).

(iii). For the matrix $\Psi_{\boldsymbol{M}}$ defined below, $\|\Psi_{\boldsymbol{M}}\| = 0$

$$\Psi_{\boldsymbol{M}} \equiv \sum_{t \neq t'} \left(\boldsymbol{B}_t^{\intercal} \boldsymbol{B}_{t'}\right) \odot \left(\boldsymbol{B}_t^{\intercal} \boldsymbol{B}_{t'}\right)^{\intercal}. \tag{3.5.5}$$

where $\odot$ represents the Hadamard (element-wise) multiplication.[3]

for ordered and unordered monotonicity mentioned above.

[3]We use the short-hand notation $\sum_{t \neq t'} \xi(t, t') \equiv \sum_{t \in \mathcal{T}} \sum_{t' \in \mathcal{T} \setminus \{t\}} \xi(t, t')$.

(iv). For any *pair of instruments* z, z' *the 2SLS type estimand*

$$\beta_{z,z'} = \mathbb{E}[Y \mid Z = z] - \mathbb{E}[Y \mid Z = z']$$

identifies an interpretable causal parameter as described in (3.5.3).

Proof. See Appendix A □

Many features of this equivalence result are symmetric to the unordered and ordered equivalence results of Theorems 3.4.1 and 3.4.2. Item *(i)* defines the complete version of the MM condition. Items *(ii)* and *(iii)* of Theorem 3.5.1 provide ways of verify the MM condition symmetric to counterparts for unordered and ordered monotonicity in Theorems 3.4.1 and 3.4.2. Item *(ii)* presents a general response matrix condition. It states that no 2×2 submatrix of the response-matrix $\boldsymbol{R}$ presents the prohibited pattern in (3.5.2). Item *(iii)* provides a tractable method of verifying the MM condition. The verification requires an order of $\mathcal{T}^2$ matrix operations.

The last item of Theorem 3.5.1 is the empirically relevant feature of the MM condition. It provides a solution to our initial inquiry on a weak mononocity criteria that ensures interpretable causal parameters for the widely used method of 2SLS. Indeed, there can be no weaker monotonicity criterion that guarantees such causal interpretability.

3.5.3. Relationship Between Monotonicity Criterion

The three monotonicity conditions are equivalent in the case of a binary treatment. In this special case the definition of MM (3.5.1) reduces to:[4]

$$\begin{aligned} \mathbf{1}[T_i(z) = t] \geq \mathbf{1}[T_i(z') = t] \quad \text{for all } i \in \mathcal{I} \\ \text{or } \mathbf{1}[T_i(z) = t] \leq \mathbf{1}[T_i(z') = t] \quad \text{for all } i \in \mathcal{I}, \end{aligned}$$

[4]This is done by replacing $\mathbf{1}[T_i(z') = t']$ with $(1 - \mathbf{1}[T_i(z) = t'])$ on the left hand side and $\mathbf{1}[T_i(z) = t']$ with $(1 - \mathbf{1}[T_i(z) = t])$ on the right hand side. Afterwards, distribute and simplify.

which is exactly the requirement imposed by both ordered and unordered monotonicity. However, as demonstrated by Lemma 3.5.2, when there are multiple treatments the three monotonicity criterion are distinct.

We can gain further interpretation of the monotonicity restrictions by examining the relation between the verification matrices $\Psi_{\boldsymbol{U}}, \Psi_{\boldsymbol{O}}$, and $\Psi_{\boldsymbol{M}}$ of Theorems 3.4.1, 3.4.2 and 3.5.1. We express the verification matrices in terms of a primitive component defined by:

$$\Psi(t, t', t'', t''') \equiv \left(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'}\right) \odot \left(\boldsymbol{B}_{t''}^\intercal \boldsymbol{B}_{t'''}\right). \tag{3.5.6}$$

$\Psi(t, t', t'', t''')$ is a function of four binary matrices $(\boldsymbol{B}_t, \boldsymbol{B}_{t'}, \boldsymbol{B}_{t''}, \boldsymbol{B}_{t'''})$ that returns a primitive verification matrix of dimension $N_Z \times N_S$ whose elements are either zeros or natural numbers.[5] Under this notation, the verification matrix $\Psi_{\boldsymbol{M}}$ can be expressed as:

$$\Psi_{\boldsymbol{M}} = \sum_{t \neq t'} \Psi(t, t', t', t). \tag{3.5.7}$$

Equation (3.5.7) explains the content of the verification matrix $\Psi_{\boldsymbol{M}}$. Theorem 3.5.1 states that MM (3.5.1) holds if and only if $\|\Psi_{\boldsymbol{M}}\| = 0$. By definition the matrix $\Psi_{\boldsymbol{M}}$ is the sum of the primitive verification matrices $\Psi(t, t', t', t)$ across all $N_T \cdot (N_T - 1)$ binary combinations of two distinct treatment choices $t, t' \in \mathcal{T}$. The elements of the primitive verification matrices are weakly positive and so $\|\Psi_{\boldsymbol{M}}\| = 0$ if and only if $\|\Psi(t, t', t', t)\| = 0$ for all distinct treatment values t and t'. Thus a necessary and sufficient condition for MM to hold is that each primitive verification matrix $\Psi(t, t', t', t)$ contains only zero elements for all $t, t' \in \mathcal{T}$ such that $t \neq t'$. Indeed, $\|\Psi(t, t', t', t)\| = 0$ is a equivalent to the nonexistence of any 2×2 submatrix of the response matrix $\boldsymbol{R}$ is of the form $\left(\begin{smallmatrix} t & t' \\ t' & t \end{smallmatrix}\right)$.

Theorem 3.5.2 relates the UM verification matrix to the primitive verification matrices and

[5]The output of function Ψ remains the same if first two inputs can be switched by the last two inputs $\Psi(t, t', t'', t''') = \Psi(t'', t''', t, t')$. We also have that $\Psi(t, t', t'', t''')^\intercal = \Psi(t', t, t''', t'')$, namely, the transpose of $\Psi(t, t', t'', t''')$ is equal to the matrix $\Psi(t', t, t''', t'')$ in which we switch t by t' and t'' by t'''.

the MM verification matrix.

Theorem 3.5.2 (Decomposing Unordered Verification)**.** *The following relation holds for any response matrix* $\boldsymbol{R} \in \mathcal{T}^{N_Z \times N_S}$:

$$\|\Psi_{\boldsymbol{U}}\| = 0 \quad \Leftrightarrow \quad \|\Psi_{\boldsymbol{M}}\| + \|\Psi_{\boldsymbol{U}\setminus\boldsymbol{M}}\| = 0, \tag{3.5.8}$$

$$\textit{where} \quad \Psi_{\boldsymbol{U}\setminus\boldsymbol{M}} \equiv \sum_{t\neq t'\neq t''} \Psi(t,t',t'',t). \tag{3.5.9}$$

Proof. See Appendix A □

Lemma 3.5.2 explains that UM imposes extra constraints in addition to those required for MM to hold. Theorem 3.5.2 clarifies these additional constraints. Equation (3.5.8) decomposes the UM verification ($\|\Psi_{\boldsymbol{U}}\| = 0$) into two verification requirements. The first verification criterion, $\|\Psi_{\boldsymbol{M}}\| = 0$, means that MM must hold. The additional constraint, $\|\Psi_{\boldsymbol{U}\setminus\boldsymbol{M}}\| = 0$, means that the elements of the matrix $\Psi(t,t',t'',t)$ must be zero for any selection of three distinct treatment choices $t,t',t'' \in \mathcal{T}$. This additional criterion rules out violations of the prohibited pattern in item *(iii)* of Theorem 3.4.1 that involve three distinct treatment values.

Theorem 3.5.2 offers a combinatorial interpretation of monotonicity conditions MM and UM. MM imposes $\binom{N_T}{2}$ constraints on the primitive verification matrix $\Psi(t,t',t',t)$ across the combination of treatment choices taken *two* at a time. UM imposes an additional $\binom{N_T}{3}$ constraints on $\Psi(t,t',t'',t)$ across the combination of treatment choices taken *three* at a time.

Theorem 3.5.3 decomposes the verification criterion of ordered monotonicity ($\|\Psi_{\boldsymbol{O}}\| = 0$) into the verification criterion of the MM condition, $\|\Psi_{\boldsymbol{M}}\| = 0$, and two other criteria corresponding to matrices $\Psi^{(1)}_{\boldsymbol{O}\setminus\boldsymbol{M}}$ and $\Psi^{(2)}_{\boldsymbol{O}\setminus\boldsymbol{M}}$. The requirement that $\|\Psi_{\boldsymbol{M}}\| = 0$ means that OM implies MM, as in Lemma 3.5.2. Matrix $\Psi^{(1)}_{\boldsymbol{O}\setminus\boldsymbol{M}}$ contains a subset of the unordered monotonicity constrains in matrix $\Psi_{\boldsymbol{U}\setminus\boldsymbol{M}}$ of Theorem 3.5.2. Matrix $\Psi^{(2)}_{\boldsymbol{O}\setminus\boldsymbol{M}}$ contains constrains that are not in $\Psi_{\boldsymbol{U}\setminus\boldsymbol{M}}$. This is as expected since OM does not imply nor is implied by UM.

Theorem 3.5.3 (Decomposing Ordered Verification)**.** *The following relation holds for any*

response matrix $\boldsymbol{R} \in \mathcal{T}^{N_Z \times N_S}$:

$$\|\Psi_{\boldsymbol{O}}\| = 0 \quad \Leftrightarrow \quad \|\Psi_{\boldsymbol{M}}\| + \|\Psi^{(1)}_{\boldsymbol{O}\setminus\boldsymbol{M}}\| + \|\Psi^{(2)}_{\boldsymbol{O}\setminus\boldsymbol{M}}\| = 0, \tag{3.5.10}$$

where

$$\Psi^{(1)}_{\boldsymbol{O}\setminus\boldsymbol{M}} \equiv \sum_{t_1 < min(t_2,t_3)} \Psi(t_1,t_2,t_3,t_1),$$

$$\Psi^{(2)}_{\boldsymbol{O}\setminus\boldsymbol{M}} \equiv \sum_{t_1 < t_2 \leq t_3} \Psi(t_1,t_3,t_2,t_2) + \sum_{t_1 < t_2 < t_3} \Psi(t_1,t_3,t_3,t_2) + \sum_{t_4 < t_2, t_1 < t_3} \Psi(t_1,t_2,t_3,t_4).$$

Proof. See Appendix A □

3.6. An Economic Interpretation for Monotonicity Conditions

This section explores the economic content of the monotonicity criteria. We show that the minimal monotonicity condition (3.5.1) can be linked to a broad notion of rationality regarding treatment choices. This contrasts to ordered and unordered monotonicity, which as seen in Theorems 3.4.1 and 3.4.2 are equivalent to assuming particular ordered and unordered choice models.

Our analysis is based on the method of Pinto (2021), Buchinsky and Pinto (2021) who use revealed preference analysis to ascribe economic interpretation to response matrices.[1] The method uses the concept of an incentive matrix $\boldsymbol{L}$ that characterizes the choice incentives (columns) generated by the IV-values (rows). Each column $\boldsymbol{L}[\cdot, t]$ displays the relative ranking of incentives towards choice $t \in \text{supp}(t)$ across the IV-values $z \in \mathcal{Z}$. $\boldsymbol{L}[z', t] < \boldsymbol{L}[z, t]$ means that the IV-value z yields strictly greater incentives towards t than IV-value z'. The matrix is ordinal, monotonic transformations characterize equivalent choice incentives.

To fix ideas, consider the binary LATE model of Sections 3.4 where T denotes college enrollment; $T = t_1$ for college enrollment and $T_i = t_0$ otherwise. The instrument Z denotes a

[1] This analysis is also similar to the encouragment designs proposed by Zelen (1979, 1990).

randomly assigned tuition discount such that $Z = z_1$ if the discount is granted and $Z = z_0$ otherwise. Incentive matrix (3.6.1) characterizes the choice incentives of the LATE model. $\boldsymbol{L}[z_0, t_0] = \boldsymbol{L}[z_1, t_0] = 0$ means that the voucher offers no incentives for choice t_0 (no college). $\boldsymbol{L}[z_0, t_1] < \boldsymbol{L}[z_1, t_1]$ indicates that the tuition discount z_1 incentives college enrollment t_1.

$$\text{LATE Incentive Matrix} \quad \boldsymbol{L} = \begin{array}{c} \begin{matrix} t_0 & t_1 \end{matrix} \\ \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \end{array} \begin{matrix} z_0 \\ z_1 \end{matrix} \tag{3.6.1}$$

Pinto (2021), Buchinsky and Pinto (2021) use revealed preference analysis to translate the incentive matrix into choice restrictions. They invoke the Weak Axiom of Revealed Preferences (WARP) and Choice Normality to generate the following choice rule:[2]

$$\text{Choice Rule: } T_i(z) = t \text{ and } \underbrace{\boldsymbol{L}[z', t'] - \boldsymbol{L}[z, t'] \leq \boldsymbol{L}[z', t] - \boldsymbol{L}[z, t]}_{\substack{\text{Switch from } z \text{ to } z' \text{ provides greater} \\ \text{incentives for } t \text{ than } t'}} \implies T_i(z') \neq t' \tag{3.6.2}$$

Choice Rule (3.6.2) formalizes an intuitive behavioral restriction. If an agent i chooses t when exposed to instrument z, and the IV-shift from z to z' yields greater incentives towards t than t', then agent i does not choose t' under z'. Otherwise stated, each instrumental value z is associated with an incentive gap between t and t'. If an agent decides for choice t given z, then t is revealed preferred to t'. The agent should t' only if the incentive gap between t and t' increases.

Applying choice rule (3.6.2) to the LATE incentive matrix (3.6.1) generates the following *choice restriction*:

$$T_i(z_0) = t_1 \text{ and } \underbrace{\boldsymbol{L}[z_1, t_0] - \boldsymbol{L}[z_0, t_0]}_{=0} \leq \underbrace{\boldsymbol{L}[z_1, t_1] - \boldsymbol{L}[z_0, t_1]}_{=1} \implies T_i(z_1) \neq t_0. \tag{3.6.3}$$

[2] [Define both] The choice rule would have a strict inequality if we were to assume WARP only.

Choice restriction (3.6.3) is summarized by $T_i(z_0) = t_1 \Rightarrow T_i(z_1) \neq t_0$. It states that if an agent chooses to attend college when not offered any incentives ($T_i(z_0) = t_1$) they must also chooses choose to attend college when offered the tuition discount ($T_i(z_1) = t_1$). The restriction is equivalent to the monotonicity condition of Imbens and Angrist (1994), discussed in Section 3.3, which eliminates the defiers and enables the identification of LATE.

Buchinsky and Pinto (2021) characterize the class of incentive matrices that produce OM and UM conditions. For instance, they demonstrate that choice incentives characterized by lonesum matrices produce unordered choice models while increasing incentives such as those described by a Vandermonde matrix produce ordered choice models.

Incentives and Minimal Monotonicity

It is natural to inquire about which types of incentive designs ensure the MM condition. It turns out that the Choice Rule (3.6.2) itself ensures the MM condition. Theorem 3.6.1 asserts that the MM condition always arises whenever we apply the revealed preference analysis encoded by the choice rule to any choice incentives.

Theorem 3.6.1. *The minimal monotonicity condition* (3.5.1) *holds for all choice models generated by applying Choice Rule* (3.6.2) *to an arbitrary Incentive Matrix* $\boldsymbol{L}$.

Theorem 3.6.1 draws a sharp distinction in the interpretation of the monotonicity conditions. In essence, OM and UM can be interpreted as monotonicity conditions that arise when agents that display a rational behavior face a particular a class of choice incentives. This paradigm does not apply to MM, since MM is not a property ascribed to any particular pattern of incentives. Instead, MM is a supra-condition that can be justified by a weak notion of rationality itself.

With this in mind, the MM condition can be seen not as a final goal, but rather a starting point for generating and interpreting monotonicity criteria. To be more precise, minimal monotonicity ensures that a range of monotonicity conditions obtained by combining a broad

notion of choice rationality with specific choice incentives satisfy particular basic properties, including interpretability of 2SLS-type estimands. Section 3.7 uses this insight to illustrate the flexibility of the MM condition in empirical analysis.

3.7. Economic Examples of Monotonicity Conditions

We present several examples of response matrices generated by combining specific incentive structures with the choice rule (3.6.2) defined in Section 3.6. The first two examples demonstrate specific incentive designs that generate response matrices satisfying unordered and ordered monotonicity, respectively. Afterwards, we present some natrual research designs that generate response matrices that do not comply with either ordered or unordered monotonicity. However, by Theorem 3.6.1, these response matrices still satisfy the minimal monotonicity condition. We discuss how minimal monotonicity may still enable the researcher to gain insight into causal effects under these research designs. Our minimal monotonicity examples include the popular Extensive Margin Compliers (EMCO) of Angrist and Imbens (1995) and a double RCT design. In all examples we consider incentive designs for a three-valued treatment choice $\mathcal{T} = \{t_1, t_2, t_3\}$ and four instrumental values $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$.

3.7.1. A Case of Choice Incentives that Justify Unordered Monotonicity

In this example, let T denote the student's decision among college majors: t_1 for humanities, t_2 for social sciences, and t_3 for the STEM fields of science, technology, engineering, and math. The instrumental variable Z represents a randomly assigned vouchers that offers a tuition discount that may apply to one, several or none of the majors. For example, consider the social experiment that randomly assigns one of the four vouchers z_1, z_2, z_3, z_4 to college students:

1. Voucher z_1 offers no tuition discount.
2. Voucher z_2 applies only to STEM (t_3).
3. Voucher z_3 applies to either STEM (t_3) or social sciences (t_2).

4. Voucher z_4 applies to all majors.

Assuming that the voucher amount is always the same, this design is characterized by incentive matrix $\boldsymbol{L}$ in (3.7.1).[1]

$$\boldsymbol{L} = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.7.1}$$

We use this first example to describe the machinery that translates choice incentives into monotonicity conditions and identification results. We adopt a more parsimonious approach in the subsequent examples. We place detailed derivations in Appendix 3.12.

Choice rule (3.6.2) converts the Incentive Matrix (3.7.1) into choice restrictions that determine the model response matrix $\boldsymbol{R}$. The choice rule applies to any two instrument-treatment pairs; $((z,t),(z',t')) \in (\mathcal{Z} \times \mathcal{T})^2$. To exemplify how this is done, Table 3.7.1 displays all the restrictions generated by applying Choice Rule (3.6.2) to an agent who chooses a humanities major (t_1) when offered no tuition discount (z_1). The first row of Table 3.7.1 applies the choice rule to the two instrument-treatment pairs, (z_1, t_1) and (z_2, t_2). By applying the choice rule, we see that an agent i who chooses a humanties major (t_1) when offered no tuition discount (z_1) would not chose treatment a social sciences major (t_2) when offered instrument a tuition discount only for STEM majors (z_2). The incentives for choosing either t_1 or t_2 remain the same when the IV switches from z_1 to z_2. The incentive inequality in (3.6.2) is satisfied and the choice restriction $T_i(z_1) \neq t_2$ holds.

The other lines of Table 3.7.1 apply this same logic to all other instrument-treatment pairs. We can see that, in total, under the incentive structure summarized in (3.7.1) the Choice

[1] Elements one indicate the presence of incentive (the tuition discount) while elements zero indicate the lack of it.

Rule (3.6.2) places the following restrictions on an agent who chooses a humanities major when offered no tuition discount

$$T_i(z_1) = t_1 \implies T_i(z_2) \neq t_2 \text{ and } T_i(z_4) \notin \{t_2, t_3\}$$

This analysis can be repeated for all types of agents, the sum total of all choice restrictions generated by applying the choice rule in this research design are presented in Table 3.7.2. All choice restrictions of Table 3.7.2 eliminate a total of 74 out of 81 possible response-types. The seven response-types that survive this elimination procedure are presented as columns of the following response matrix, $\boldsymbol{R}$, in (3.7.2):

Table 3.7.1: Applying Choice Rule (3.6.2) to $T_i(z_1) = t_1$ and Incentive Matrix (3.7.1)

Counterfactual Choice	Incentive Condition			Choice Restriction
$T(z_1) = t_1$	$\boldsymbol{L}[z_2, t_2] - \boldsymbol{L}[z_1, t_2]$	$= 0 \leq 0 =$	$\boldsymbol{L}[z_2, t_1] - \boldsymbol{L}[z_1, t_1]$	$\Rightarrow$ $T(z_2) \neq t_2$
$T(z_1) = t_1$	$\boldsymbol{L}[z_2, t_3] - \boldsymbol{L}[z_1, t_3]$	$= 1 \not\leq 0 =$	$\boldsymbol{L}[z_2, t_1] - \boldsymbol{L}[z_1, t_1]$	$\Rightarrow$ No Restriction
$T(z_1) = t_1$	$\boldsymbol{L}[z_3, t_2] - \boldsymbol{L}[z_1, t_2]$	$= 1 \not\leq 0 =$	$\boldsymbol{L}[z_3, t_1] - \boldsymbol{L}[z_1, t_1]$	$\Rightarrow$ No Restriction
$T(z_1) = t_1$	$\boldsymbol{L}[z_3, t_3] - \boldsymbol{L}[z_1, t_3]$	$= 1 \not\leq 0 =$	$\boldsymbol{L}[z_3, t_1] - \boldsymbol{L}[z_1, t_1]$	$\Rightarrow$ No Restriction
$T(z_1) = t_1$	$\boldsymbol{L}[z_4, t_2] - \boldsymbol{L}[z_1, t_2]$	$= 1 \leq 1 =$	$\boldsymbol{L}[z_4, t_1] - \boldsymbol{L}[z_1, t_1]$	$\Rightarrow$ $T(z_4) \neq t_2$
$T(z_1) = t_1$	$\boldsymbol{L}[z_4, t_3] - \boldsymbol{L}[z_1, t_3]$	$= 1 \leq 1 =$	$\boldsymbol{L}[z_4, t_1] - \boldsymbol{L}[z_1, t_1]$	$\Rightarrow$ $T(z_4) \neq t_3$

This table presents all the choice restrictions generated by applying the choice rule (3.6.2) to each of the tuples $((z_1, t_1), (z', t'))$ where $z' in \{z_2, z_3, z_4\}$ and $t' in \{t_2, t_3, t_4\}$ according to the choice incentives displayed in the incentive matrix (3.7.1).

$$\boldsymbol{R} = \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \end{matrix} \\ \begin{bmatrix} t_1 & t_1 & t_1 & t_1 & t_2 & t_2 & t_3 \\ t_1 & t_1 & t_3 & t_3 & t_2 & t_3 & t_3 \\ t_1 & t_2 & t_2 & t_3 & t_2 & t_2 & t_3 \\ t_1 & t_1 & t_1 & t_1 & t_2 & t_2 & t_3 \end{bmatrix} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.7.2}$$

Using the characterizations in Theorem 3.4.1, we can verify that unordered monotonicity holds for the response matrix $\boldsymbol{R}$ presented in (3.7.2). Notice that there is no 2×2 matrix of the matrix is of the type of the form $\left(\begin{smallmatrix} t & t' \\ t'' & t \end{smallmatrix}\right)$ where $t' \neq t$, or $t'' \neq t$. Appendix 3.12.1 corroborates the UM property using the verification matrix from item (iv) of Theorem 3.4.1.

Table 3.7.2: Choice Restrictions generated by Incentive Matrix (3.7.1)

1	$T_i(z_1) = t_1$	$\Rightarrow$	$T_i(z_2) \neq t_2$ and $T_i(z_4) \notin \{t_2, t_3\}$
2	$T_i(z_2) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_3) \neq t_3$ and $T_i(z_4) \notin \{t_2, t_3\}$
3	$T_i(z_3) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_2) \notin \{t_2, t_3\}$ and $T_i(z_4) \notin \{t_2, t_3\}$
4	$T_i(z_4) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_2) \neq t_2$
5	$T_i(z_1) = t_2$	$\Rightarrow$	$T_i(z_2) \neq t_1$ and $T_i(z_3) \notin \{t_1, t_3\}$ and $T_i(z_4) \notin \{t_1, t_3\}$
6	$T_i(z_2) = t_2$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_3\}$ and $T_i(z_3) \notin \{t_1, t_3\}$ and $T_i(z_4) \notin \{t_1, t_3\}$
7	$T_i(z_3) = t_2$	$\Rightarrow$	$T_i(z_1) \neq t_3$ and $T_i(z_4) \neq t_3$
8	$T_i(z_4) = t_2$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_3\}$ and $T_i(z_2) \neq t_1$ and $T_i(z_3) \notin \{t_1, t_3\}$
9	$T_i(z_1) = t_3$	$\Rightarrow$	$T_i(z_2) \notin \{t_1, t_2\}$ and $T_i(z_3) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
10	$T_i(z_2) = t_3$	$\Rightarrow$	$T_i(z_3) \neq t_1$
11	$T_i(z_3) = t_3$	$\Rightarrow$	$T_i(z_1) \neq t_2$ and $T_i(z_2) \notin \{t_1, t_2\}$ and $T_i(z_4) \neq t_2$
12	$T_i(z_4) = t_3$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_2\}$ and $T_i(z_2) \notin \{t_1, t_2\}$ and $T_i(z_3) \notin \{t_1, t_2\}$

This table presents all the choice restrictions generated by applying the choice rule (3.6.2) to each of the tuples $((z,t),(z',t')) \in (\{t_1,t_2,t_3\} \times \{z_1,z_2,z_3,z_4\})^2$. according to the incentive matrix (3.7.1).

While the incentive design described in (3.7.1) is plausible and the application of the choice rule a minimal behavioral requirement on the agents, it is helpful to remark on how this specific incentive structure generates unordered monotonicity. Note that the incentive structure increases in the sense that each successive instrument provides weakly more incentives for all treatment choices. No change in the instrument from z to z' would strictly decrease incentives for one treatment while strictly increasing incentives for another.

This property is crucial for generating unordered monotonicity. Under WARP, if an agent would choose treatment t under instrument value z, they must also choose treatment t under instrument value z' whenever the switch from z to z' weakly increases incentives for t relative to all other incentives. Because of the increasing nature of the incentive structure, each switch in the instrument value either increases the incentives for a choice t relative to *all* other treatments or decreases the incentives for the choice t relative to all other treatments. This, along with the choice rule, prevents an instrumental switch from moving one agent strictly towards choosing t while moving another agent strictly away from choice t and towards another treatment t'.

3.7.2. A Case of Choice Incentives that Justify Ordered Monotonicity

In this example, suppose the CEO of a company is interested in whether higher health insurance premiums lead to moral hazard in employees' safety behavior. Employees decide among three health insurance policies t_1, t_2, t_3 that have increasing premiums. The co-pay of each policy off-sets the increasing premium such that all policies cost the same.

To study this, the CEO randomly assigns agents to one of four groups, z_1, z_2, z_3, z_4, that incentivize (say by offering an additional week of vacation) various insurance plan options. We consider the following scheme of choice incentives:

1. Group z_1 is incentivized to choose treatment t_1.
2. Group z_2 is offered no incentives.
3. Group z_3 is offered incentives for all choices.
4. Group z_4 is incentivized to choose treatment t_3.

Equation (3.7.3) presents the incentive matrix $\boldsymbol{L}$ that characterizes the design of choice incentives.[2] This incentive design is rather peculiar because it is tailored to generate the OM criteria. Equation (3.7.3) also presents the corresponding response matrix $\boldsymbol{R}$ generated by the method of revealed preference analysis described in Sections 3.6 and 3.7.1. Detailed derivations are presented in Appendix 3.12.2.

$$\boldsymbol{L} = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \quad \Rightarrow \quad \boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 & \boldsymbol{s}_8 \end{matrix} \\ \begin{bmatrix} t_1 & t_1 & t_1 & t_1 & t_1 & t_2 & t_2 & t_3 \\ t_1 & t_1 & t_2 & t_2 & t_3 & t_2 & t_2 & t_3 \\ t_1 & t_1 & t_2 & t_2 & t_3 & t_2 & t_2 & t_3 \\ t_1 & t_3 & t_2 & t_3 & t_3 & t_2 & t_3 & t_3 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.7.3}$$

Using Theorem 3.4.2 it is easy to check that OM holds for response matrix (3.7.3). The indices of the treatment choices weakly increase as the instrument ranges along $z_1 \rightarrow$

[2] Elements one indicate the presence of incentive (an additional one week vacation) while elements zero indicate the lack of it.

$z_2 \to z_3 \to z_4$. Ordered monotonicity is satisfied by assigning treatment values that satisfy $t_1 \leq t_2 \leq t_3$. By applying the identification results of Angrist and Imbens (1995), one can verify that the 2SLS has the causal interpretation of a weighted average of LATEs of the type $E(Y(t_{k+1}) - Y(t_k)|\boldsymbol{S})$; $k \in \{1, 2\}$. Similarly via Theorem 3.4.1 it is also easy to verify that UM does not hold for response matrix (3.7.3). The 2×2 submatrix of rows (z_1, z_4) and columns $(\boldsymbol{s}_3, \boldsymbol{s}_7)$ displays the values $\left(\begin{smallmatrix} t_1 & t_2 \\ t_2 & t_3 \end{smallmatrix}\right)$ which violates item (iii) of Theorem 3.4.1; the shift of IV-values from z_1 to z_4 induces some agents towards choice t_2 while inducing others away from t_2.

Interestingly, the natural ranking on the treatment space is not important for generating ordered monotonicity in this way. We could just as easily have considered a treatment space with no natural ranking, such as the choice of neighborhood to live in. The incentive design summarized by $\boldsymbol{L}$ in (3.7.3) would have still generated the response matrix $\boldsymbol{R}$ that satisfies ordered monotonicity. This demonstrates the usefulness of considering the slightly broader characterization of ordered monotonicity presented in OM Sequence (3.4.3). Had ordered monotonicity been ruled out a priori, the researcher may not have been able to take advantage of the Angrist and Imbens (1995) identification results.

3.7.3. Beyond Ordered or Unordered Monotonicity

The MM condition provides a theoretical foundation for a wide range choice behaviors that do not exactly conform to the paradigm imposed by ordered or unordered choices. It offers the necessary flexibility to examine economic settings that are not neatly described by ordered or unordered monotonicity. We illustrate this fact in the following examples.

The Double Randomization Design

A basic inquiry in social science is to evaluate the causal effect of a treatment t_1 versus its absence. The standard IV experiment that would allow us to assess this effect is to randomly offer a voucher that incentives a set of agents to choose a treatment choice t_1. This experiment

can be described by the binary LATE model discussed in Section 3.3.

A straightforward extension of this setup is to insert a second treatment t_2 and randomize a second voucher that incentives t_2 for the same set of agents. The combination of the two randomization runs generate four groups according to the voucher assignments. Notationally, our experiment consists of three choices $T \in \{t_0, t_1, t_2\}$, where t_0 represents not choosing either treatment t_1 or t_2, and four instrumental values $\{z_1, z_2, z_3, z_4\}$ that classify the voucher recipients into:

1. Group z_1 comprise agents that do not receive any voucher.
2. Group z_2 comprise agents that receive a voucher that incentivizes choice (t_2).
3. Group z_3 comprise agents that receive a voucher that incentivizes choice (t_1).
4. Group z_4 are those that receive two vouchers, one for t_1 and another for t_2.

Equation (3.7.4) presents an incentive matrix corresponding to this research design and the corresponding response matrix generated by the revealed preference analysis described in Section 3.7.1. See Appendix 3.12.3 for detailed derivations.

$$\boldsymbol{L} = \begin{array}{c} \begin{matrix} t_0 & t_1 & t_2 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \Rightarrow \boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 & \boldsymbol{s}_8 & \boldsymbol{s}_9 \end{matrix} \\ \begin{bmatrix} t_0 & t_0 & t_0 & t_0 & t_0 & t_1 & t_1 & t_2 & t_2 \\ t_0 & t_0 & t_2 & t_2 & t_2 & t_1 & t_2 & t_2 & t_2 \\ t_0 & t_1 & t_0 & t_1 & t_1 & t_1 & t_1 & t_1 & t_2 \\ t_0 & t_1 & t_2 & t_1 & t_2 & t_1 & t_1 & t_2 & t_2 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.7.4}$$

Response matrix (3.7.4) satisfies neither UM nor OM. UM does not hold because the 2×2 submatrix generated by rows (z_2, z_3) and columns $(\boldsymbol{s}_2, \boldsymbol{s}_3)$ displays matrix $\left(\begin{smallmatrix} t_0 & t_2 \\ t_1 & t_0 \end{smallmatrix}\right)$, which violates item (iii) of Theorem 3.4.1. For the ordering $t_0 \leq t_1 \leq t_2$ on $\mathcal{T}$, columns $(\boldsymbol{s}_2, \boldsymbol{s}_3)$ also preclude OM. Under this ordering, no matter how we order the IV values, we cannot generate that the sequence of treatments is increasing for both $\boldsymbol{s}_2$ and $\boldsymbol{s}_3$. In fact, no matter which ordering we take on $\mathcal{T}$, we can always find a pair of response types whose treatment

uptake patterns violate OM Sequence.[3]

However, as guaranteed by Theorem 3.6.1, the response matrix $\boldsymbol{R}$ in (3.7.4) satisfies minimal monotonicity (MM). This can be further verified via the minimal monotonicity characterization in Theorem 3.5.1 by checking that there is no 2×2 submatrix in (3.7.4) of the form $\left(\begin{smallmatrix} t & t' \\ t' & t \end{smallmatrix}\right)$. Appendix 3.12.3 additionally corroborates this fact by evaluating the verification matrix $\Psi_{\boldsymbol{M}}$ from Theorem 3.5.1(*iii*).

Despite the fact that neither UM nor OM holds, we can still explore causal relationships in this choice model. In particular, the response matrix still enables us to identify causal parameters using 2SLS type estimands. For example, if the researcher was interested in the effect of treatment t_2 against alternate treatments, the 2SLS-type estimand

$$\beta_{z_4,z_3} \equiv \mathbb{E}[Y \mid Z = z_4] - \mathbb{E}[Y \mid Z = z_3].$$

recovers a weighted average of $\mathbb{E}[Y(t_2) - Y(t_1) \mid \boldsymbol{S} \in \mathcal{S}_{2,1}]$ and $\mathbb{E}[Y(t_2) - Y(t_0) \mid \boldsymbol{S} \in \mathcal{S}_{2,0}]$, with positive weights, for some sets of response types $\mathcal{S}_{2,1}$ and $\mathcal{S}_{2,0}$ that can be found using the decomposition in Section 3.10. If one was alternatively interested in the effect of t_1 against alternate treatments, the 2SLS-type estimand

$$\beta_{z_3,z_1} \equiv \mathbb{E}[Y \mid Z = z_3] - \mathbb{E}[Y \mid Z = z_2]$$

recovers a weighted average of $\mathbb{E}[Y(t_1) - Y(t_0) \mid \boldsymbol{S} \in \mathcal{S}_{1,0}]$ and $\mathbb{E}[Y(t_1) - Y(t_2) \mid \boldsymbol{S} \in \mathcal{S}_{1,2}]$ for two alternate sets of response types $\mathcal{S}_{1,0}$ and $\mathcal{S}_{1,2}$.

Incentives that Justify Extensive Margin Compliers Only

We next exemplify how the choice rationale can be used to justify monotonicity criteria that

[3]Using the violation of unordered monotonicity with t_0 we have that t_0 cannot be ranked highest or lowest in any ordering that satisfies OM; this would mean that some agents increase their treatment as the instrument ranges from z_2 to z_3 while other agents decrease their treatment. We thus only have to consider the orderings $t_1 \leq t_0 \leq t_2$ and $t_2 \leq t_0 \leq t_1$, which can both be eliminated by considering the alternating patterns displayed by response types $\boldsymbol{s}_4$ and $\boldsymbol{s}_7$.

are more restrictive than UM and OM. Consider a group of students of a technical college that decide among three possible majors: computer science (t_1), electrical engineering (t_2), or mechanical engineering (t_3).

College administration perform a double randomization of two types of tuition vouchers. The first voucher offers a tuition discount for computer science (t_1) while the other for the engineering courses (t_2 or t_3). Students can be divided into four groups according to the voucher assignment:

1. Group z_1 receives no voucher;
2. Group z_2 receives the voucher for computer science (t_1) only;
3. Group z_3 receives the voucher that incentivizes electrical (t_2) or mechanical (t_3) engineering.
4. Group z_4 receives both vouchers which offers incentives to all three choices.

Equation (3.7.5) presents the incentive matrix associated with this experimental design and its corresponding response matrix. See Appendix 3.12.4 for derivation details.

$$\boldsymbol{L} = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \quad \Rightarrow \quad \boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \end{matrix} \\ \begin{bmatrix} t_1 & t_1 & t_1 & t_2 & t_2 & t_3 & t_3 \\ t_1 & t_1 & t_1 & t_1 & t_2 & t_1 & t_3 \\ t_1 & t_2 & t_3 & t_2 & t_2 & t_3 & t_3 \\ t_1 & t_1 & t_1 & t_2 & t_2 & t_3 & t_3 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.7.5}$$

Response matrix (3.7.5) is an example where both OM and UM are satisfied. We can check that OM holds by assigning values $(1, 2, 3)$ to (t_1, t_2, t_3) and reordering the IV-values from z_1, z_2, z_3, z_4 to z_2, z_1, z_4, z_3. The resulting response matrix is presented in (3.7.6) which shows that treatment values weakly increase as Z ranges along its values. It is easy to check that each of the binary matrices $\boldsymbol{B}_t = \mathbf{1}[\boldsymbol{R} = t]; t \in \{1, 2, 3\}$ that indicate the treatment choices of response matrix (3.7.6) is triangular (i.e. lonesum). This implies that UM holds by item (iv) of Theorem 3.4.1.

$$\text{Reordered } \boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 1 & 3 \\ 1 & 1 & 1 & 2 & 2 & 3 & 3 \\ 1 & 1 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 3 & 2 & 2 & 3 & 3 \end{bmatrix} \end{array} \begin{matrix} z_2 \\ z_1 \\ z_4 \\ z_3 \end{matrix} \tag{3.7.6}$$

Response matrix (3.7.5) has a special property beyond UM and OM: each of its compliers takes only two treatment values, one of them being t_1. Specifically, the matrix has four response-types that display a variation of treatment choice, these are the compliers $(\boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4, \boldsymbol{s}_6)$. The choice values of response-types $\boldsymbol{s}_2, \boldsymbol{s}_4$ are t_1 or t_2 and the choice values of $\boldsymbol{s}_3, \boldsymbol{s}_6$ are t_1 or t_3 This special property is called Extensive Margin Compliers Only (EMCO) which is formalized in (3.7.7).[4]

EMCO: There exists a treatment choice $t_1 \in \mathcal{T}$ such that for any $z, z' \in \text{supp}(Z)$ we have that

$$T_i(z) \neq T_i(z') \Rightarrow T_i(z) = t_1 \text{ or } T_i(z') = t_1 \text{ for all } i \in \mathcal{I} \tag{3.7.7}$$

EMCO (3.7.7) simplifies the multiple-choice decision of compliers into a binary decision that debates between choosing t_1 or not. In our example, compliers $\boldsymbol{s}_2, \boldsymbol{s}_4$ debate between choosing computer science t_1 or electrical engineering t_2 while compliers $\boldsymbol{s}_3, \boldsymbol{s}_6$ debate between computer science t_1 or mechanical engineering t_3. None of the compliers debate between electrical or mechanical engendering. Instead, they decide between choosing computer science or not.

EMCO (3.7.7) enables us to recode the multiple choice $T_i \in \{t_1, t_2, t_3\}$ into a binary choice $D_i = \mathbf{1}[T_i = t_1]$ that indicates if the agent i chooses t_1. The 2SLS regression that uses the binary indicator as the endogenous treatment evaluates a weighted average of LATE-type effects between choosing t_1 and not across compliers.

In particular, the comparison between two IV-values identifies the causal effect for of choosing

[4] See Rose and Shem-Tov (2021), Angrist and Imbens (1995).

t_1 versus not choosing t_1 for a sub-set of compliers. For instance, consider the IV-values z_1 and z_2. We can use equation (3.3.9) to obtain the following identification result:

$$\frac{E(Y|Z=z_2)-E(Y|Z=z_1)}{P(T=t_2|Z=z_2)-P(T=t_2|Z=z_1)} = \tag{3.7.8}$$

$$\frac{E(Y(t_1)-Y(t_2)|\boldsymbol{S}=\boldsymbol{s}_4)P(\boldsymbol{S}=\boldsymbol{s}_4)+E(Y(t_1)-Y(t_3)|\boldsymbol{S}=\boldsymbol{s}_6)P(\boldsymbol{S}=\boldsymbol{s}_6)}{P(\boldsymbol{S}=\boldsymbol{s}_4)+P(\boldsymbol{S}=\boldsymbol{s}_6)}. \tag{3.7.9}$$

Equations (3.7.8)–(3.7.9) show that the comparison between IV-values z_1 and z_2 identifies the causal effect of choosing t_1 versus not choosing t_1 conditional on response-types $\boldsymbol{s}_4, \boldsymbol{s}_6$. The equations are similar to the LATE identification equation of Imbens and Angrist (1994). They imply that we can evaluate the causal effect via the 2SLS regression that uses the sub-sample of agents assigned to z_1 and z_2.

Orthogonal Array Design

We additionally examine an IV choice model based on the orthogonal array experimental design. Orthogonal arrays are a widely popular experimental design developed by C.D. Rao (Rao, 1946a,b, 1947, 1949). Orthogonal arrays are widely used in Agricultural and Industrial sciences to determine the optimum mix of treatments that maximize production yield. The method is based on the random assignment of a combinatorial arrangements of treatments for each randomization arm. We adapt this setup to an instrumental variable setting by exogenously providing incentives for one or more treatments instead of directly assigning agents to treatment arms. Below, we will see that this incentive structure allows for a broad range of identification results.

Formally, a binary orthogonal array is a matrix of zeros and ones such that any two-column submatrix displays all possible combinations of zeros and ones. In other words, the tuples

$$\{(0,0),(0,1),(1,0),(1,1)\}$$

are all rows in any two-column submatrix. An orthogonal array incentive design if its

associated incentive matrix is a binary orthogonal array. The incentive matrix in (3.7.10) displays an example of an orthogonal array incentive design and the corresponding response matrix generated by applying Choice Rule (3.6.2).

$$\boldsymbol{L} = \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \Rightarrow \boldsymbol{R} = \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 & \boldsymbol{s}_8 & \boldsymbol{s}_9 \end{matrix} \begin{bmatrix} t_1 & t_2 & t_2 & t_2 & t_2 & t_3 & t_3 & t_3 & t_3 \\ t_1 & t_1 & t_2 & t_2 & t_2 & t_1 & t_3 & t_3 & t_3 \\ t_1 & t_1 & t_2 & t_2 & t_2 & t_1 & t_1 & t_2 & t_3 \\ t_1 & t_1 & t_1 & t_2 & t_3 & t_1 & t_3 & t_3 & t_3 \end{bmatrix} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.7.10}$$

This response matrix satisfies neither unordered nor ordered monotonicity. When the instrument switches from z_1 to z_4, agents in response type $\boldsymbol{s}_5$ move from treatment t_2 to treatment t_3 while agents in response type $\boldsymbol{s}_6$ move away from t_3 and towards t_1. This represents a violation of ordered monotonicity and also prevents t_3 from being ordered the highest or lowest in any ordering on $\mathcal{T}$ that would satisfy ordered monotonicity.[5] Similarly we can see a switch from z_3 to z_4 induces agents in response type $\boldsymbol{s}_3$ to move from treatment t_2 to treatment t_1 while inducing agents in response type $\boldsymbol{s}_7$ to move away from treatment t_1 and towards treatment t_3. This again represents a violation of unordered monotonicity and prevents t_1 from being ordered either the highest or the lowest in any ordering $\mathcal{T}$ that would satisfy ordered monotonicity. Since all orderings on $\mathcal{T} = \{t_1, t_2, t_3\}$ must have either t_1 or t_3 as the largest or smallest element, this means there is no ordering on $\mathcal{T}$ that satisfies ordered monotonicity.

Again, however, Theorem 3.6.1 guarantees that the response matrix satisfies minimal monotonicity (MM). Any simple 2SLS estimands exploiting the variation between any two instruments (as in Theorem 3.5.1) identifies an interpretable causal parameter. For example, if the researcher was interested in the effect of treatment t_1 against alternative treatments,

[5] If t_3 is ranked highest a movement away from t_3 represents moving towards a lower treatment while a towards t_3 represents moving towards a higher treatment. Vice versa, if t_3 is ranked lowest a movement towards t_3 represents moving towards a lower treatment while a movement away from t_3 represents moving towards a higher treatment.

the 2SLS estimand $\beta_{z_1,z_2} = \mathbb{E}[Y \mid Z = z_1] - \mathbb{E}[Y \mid Z = z_2]$ recovers a weighted average of $\mathbb{E}[Y(t_1) - Y(t_2) \mid \boldsymbol{S} = \boldsymbol{s}_2]$ and $\mathbb{E}[Y(t_1) - Y(t_3) \mid \boldsymbol{S} = \boldsymbol{s}_6]$.

Interestingly, the response matrix $\boldsymbol{R}$ in (3.7.10) also satisfies the extensive margin compliers only (EMCO) condition of (3.7.5) when considering any three instruments at a time. As in the last example, this allows for 2SLS estimands to additionally be interpreted as recovering the causal effect of one treatment against another.

3.8. Conclusion

Analysis of ordered and unordered IV choice models has largely been conducted in parallel, with little overlap between the two strands of the literature. Ordered choice models are commonly analyzed using ordered monotonicity (OM), introduced by Angrist and Imbens (1995), while unordered choice models are commonly analyzed using the unordered monotonicity (UM) of Heckman and Pinto (2018).

This paper bridges the gap between analysis of ordered and unordered IV choice models. We note symmetric features of ordered and unordered monotonicity and use them to derive symmetric characterizations of the two. The symmetric characterizations offer deep insights into the relationship between the two monotonicity criterion. Moreover they provide computationally tractable ways to verify the two monotonicity criterion, which may be useful to researchers who wish to utilize both sets of identification results.

The symmetric characterizations illuminate a shared monotonicity property, which we term the minimal monotonicity (MM) condition. We characterize this novel criterion and show it is the minimal requirement needed to identify interpretable causal parameters using 2SLS type estimands. Moreover, minimal monotonicity can be associated with a notion of rationality that enables the investigation of a range of economic choice models that do not comply with ordered or unordered monotonicity.

3.9. Appendix: Proofs of Main Results

3.9.1. Lonesum Matrix Characterizations

Using the characterization of unordered monotonicity in UM-Sequence (3.4.4), unordered monotonicity is equivalent to there being a permutation of the rows of $\boldsymbol{B}_t$ such that each column of $\boldsymbol{B}_t$ is weakly increasing.[1] Existence of such a reordering characterizes a class of binary matrices known as lonesum matrices, which are a generalization of lower triangular binary matrices. The lonesum property, and various characterizations of it, end up forming the basis of much of our proof strategy. We discuss the property breifly below and note a useful characterization of the property.

Lonesum Matrices

Following Ryser (1957), a binary matrix $\boldsymbol{A}$ is lonesum if each of its entries is uniquely determined by its column and row sums. Matrix $\boldsymbol{A}$ below is an example of such a lonesum matrix:

$$\boldsymbol{A} = \begin{array}{c} \\ \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \\ \text{column} \quad c_1\ c_2\ c_3\ c_4\ c_5 \\ \text{column-sum} \quad 0\ 3\ 1\ 2\ 1 \end{array} \begin{array}{l} \text{row} \\ r_1 \\ r_2 \\ r_3 \end{array} \begin{array}{l} \text{row-sum} \\ 2 \\ 4 \\ 1 \end{array} \quad \Rightarrow \quad \begin{array}{c} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix} \\ c_1\ c_3\ c_5\ c_4\ c_2 \\ 0\ 1\ 1\ 2\ 3 \end{array} \begin{array}{l} \text{row} \\ r_3 \\ r_1 \\ r_2 \end{array} \begin{array}{l} \text{row-sum} \\ 1 \\ 2 \\ 4 \end{array}$$

Original Matrix — Reordered Matrix

We can reorder the rows of the matrix $\boldsymbol{A}$ such that the elements of each column are weaklly increasing. We can also reorder the columns so that the matrix $\boldsymbol{A}$ is lower triangular, which is why the lonesum property is considered a generalization of binary lower triangular matrices. The lonesum matrix property can also be productively characterized in the following ways:

Lemma 3.9.1 (Lonesum Matrices)**.** *A binary matrix $\boldsymbol{A} \in \{0,1\}^{m \times n}$ is lonesum if and only*

[1] This permutation can differ for each $\boldsymbol{B}_t$, but there must be such a permutation for each $t \in \mathcal{T}$.

if:

(i). Matrix $\boldsymbol{A}$ *is lower-triangular under column and row permutations.*

(ii). There are no 2×2 *submatrix in* $\boldsymbol{A}$ *of the form either:*

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{3.9.1}$$

(iii). $\boldsymbol{\iota}^\intercal\big((\mathbb{1} - \boldsymbol{A})^\intercal \boldsymbol{A}\big) \odot \big((\mathbb{1} - \boldsymbol{A})^\intercal \boldsymbol{A}\big)^\intercal \boldsymbol{\iota} = 0$*, where* $\boldsymbol{\iota}$ *is a* n*-dimensional vector of elements ones and* $\mathbb{1}$ *is a* $m \times n$ *matrix of element ones.*

(iv). Let $r_i(\boldsymbol{A})$ *and* $c_j(\boldsymbol{A})$ *represent the row sum of row* i *and the column sum of column* j*, respectively. Each entry* $\boldsymbol{A}[i, j]$*, for* $1 \leq i \leq m$ *and* $1 \leq i \leq n$*, can be expressed as:*

$$\boldsymbol{A}[i, j] = \mathbf{1}\left[r_i(\boldsymbol{A}) \ \geq \ \sum_{j'=1}^{n} \mathbf{1}\left[c_{j'}(\boldsymbol{A}) \geq c_j(\boldsymbol{A})\right]\right] \tag{3.9.2}$$

The sum $\sum_{j'=1}^{n} \mathbf{1}\left[c_{j'}(\boldsymbol{A}) \geq c_j(\boldsymbol{A})\right]$ *represents the number of columns of* $\boldsymbol{A}$ *with a weakly larger column sum than column* j

Proof. The second item comes from Ryser (1957). We next show a series of equivalences.

(iii) $\iff$ *(ii).* Notice that *(iii)* is equivalent to the matrix

$$(\mathbb{1} - \boldsymbol{A})^\intercal \boldsymbol{A} \odot \boldsymbol{A}^\intercal (\mathbb{1} - \boldsymbol{A})$$

being a matrix of all zeros. By direct calculation, the ij^{th} element of $(\mathbb{1} - \boldsymbol{A})^\intercal \boldsymbol{A}$ is given

$$\sum_{k=1}^{m} \boldsymbol{A}[k, i] - \boldsymbol{A}[k, j]\boldsymbol{A}[k, i] = \sum_{k=1}^{m} \boldsymbol{A}[k, i](1 - \boldsymbol{A}[k, j]).$$

This is non-zero if and only if $\boldsymbol{A}[k, i] = 1$ and $\boldsymbol{A}[k, j] = 0$ for some k. Conversely, the ij^{th}

element of $\boldsymbol{A}^\intercal(\mathbb{1} - \boldsymbol{A})$ can be expressed

$$\sum_{k=1}^{m} \boldsymbol{A}[k,j] - \boldsymbol{A}[k,i]\boldsymbol{A}[k,j] = \sum_{k=1}^{m} \boldsymbol{A}[k,j](1 - \boldsymbol{A}[k,i]).$$

This is non-zero if and only if $\boldsymbol{A}[k,i] = 0$ and $\boldsymbol{A}[k,j] = 1$ for some k. The ij^{th} element of $(\mathbb{1} - \boldsymbol{A})^\intercal \boldsymbol{A} \odot \boldsymbol{A}^\intercal(\mathbb{1} - \boldsymbol{A})$ is non-zero if and only if each of the terms above are nonzero, which in turn is equivalent to their existing two rows k, k' such that

$$\begin{pmatrix} \boldsymbol{A}[k,i] & \boldsymbol{A}[k,j] \\ \boldsymbol{A}[k',i] & \boldsymbol{A}[k',j] \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

which is equivalent to one of the restricted submatrices in item (ii) up to a relabeling of k and k'.

(iv) $\iff$ *(ii).* For the forward direction, notice that if *(iv)* holds, then the matrix $\boldsymbol{A}$ can clearly be reproduced uniquely from its row and column sums, so it is lonesum and thus *(ii)* holds. We then want to show that *(ii)* $\implies$ *(iv).* Consider the contrapositive. First, suppose that there is some element $\boldsymbol{A}[i,j]$ such that $\boldsymbol{A}[i,j] = 1$ but

$$r_i(\boldsymbol{A}) < \sum_{j'=1}^{n} \mathbf{1}[c_{j'}(\boldsymbol{A}) \geq c_j(\boldsymbol{A})]. \tag{3.9.3}$$

That is, the row sum of row i is less than the number of columns j' with column sum larger than column j. For this to be the case, there must be some column j' such that $\boldsymbol{A}[i,j'] = 0$ but $c_{j'}(\boldsymbol{A}) \geq c_j(\boldsymbol{A})$.[2] Because $\boldsymbol{A}[i,j'] = 0$ we know that $j \neq j'$.

Because $\boldsymbol{A}[i,j] = 1$ but $\boldsymbol{A}[i,j'] = 0$, for the column sum of j' to be weakly larger than the column sum of j, there must be some other row $i' \neq i$ such that $\boldsymbol{A}[i',j] = 0$ but $\boldsymbol{A}[i',j'] = 1$. This generates the restricted pattern which violates *(ii)* (up to a relabeling of columns j and

[2]There must be some column j' that contributes to the right hand side of (3.9.3) but not to the row sum on the left hand side.

j').

$$\begin{pmatrix} \boldsymbol{A}[i,j] & \boldsymbol{A}[i,j'] \\ \boldsymbol{A}[i',j] & \boldsymbol{A}[i',j'] \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Conversely, suppose that $\boldsymbol{A}[i,j] = 0$ but

$$r_i(\boldsymbol{A}) \geq \sum_{j'=1}^{n} \mathbf{1}[c_{j'}(\boldsymbol{A}) \geq c_j(\boldsymbol{A})] \tag{3.9.4}$$

By similar logic as the previous case, this means that there must be some column j' such that $\boldsymbol{A}[i,j'] = 1$ but $c_{j'}(\boldsymbol{A}) < c_j(\boldsymbol{A})$. Since column j does not contribute to the row sum on the right hand side of (3.9.4) but does contribute to the column sum on the right hand side, there must be some other column that does the opposite in order for (3.9.4) to hold.

Because $\boldsymbol{A}[i,j'] = 1$ and $\boldsymbol{A}[i,j] = 0$ but $c_{j'}(\boldsymbol{A}) < c_j(\boldsymbol{A})$, there must be some other row $i' \neq i$ such that $\boldsymbol{A}[i,j'] = 1$ but $\boldsymbol{A}[i',j'] = 0$. As before, we can show that this generates the restricted pattern which violates *(ii)* (up to a relabeling of columns j and j').

(i) $\iff$ *(ii)*. That *(i)* $\implies$ *(ii)* is clear, since the existence of the restricted pattern is stable under row and column permutations and the existence of such a restricted pattern prevents a matrix from being lower triangular. To see that *(ii)* $\implies$ *(i)*, suppose that $\boldsymbol{A}$ is lonesum and let $\tilde{\boldsymbol{A}}$ be the matrix generated by ordering the rows of $\boldsymbol{A}$ is increasing row-sum and the colums of $\boldsymbol{A}$ in decreasing column sum. Now, notice that $\tilde{\boldsymbol{A}}$ must also be lonesum, since if the restricted pattern appears in $\tilde{\boldsymbol{A}}$ it must also appear in $\boldsymbol{A}$; we cannot generate the restricted pattern using row and column permutations of a lonesum matrix. If $\tilde{\boldsymbol{A}}$ is lonesum, then *(iv)* must hold, that is

$$\tilde{\boldsymbol{A}}[i,j] = \mathbf{1}\left[r_i(\tilde{\boldsymbol{A}}) \;\geq\; \sum_{j'=1}^{n} \mathbf{1}\left[c_{j'}(\tilde{\boldsymbol{A}}) \geq c_j(\tilde{\boldsymbol{A}})\right]\right]. \tag{3.9.5}$$

Since the columns of $\tilde{\boldsymbol{A}}$ are ordered in increasing column sum, we must have by (3.9.5) that $\tilde{\boldsymbol{A}}[i,j] \leq \tilde{\boldsymbol{A}}[i,j']$ for $j' > j$. Similarly, since the rows are ordered in decreasing row

sum, (3.9.5) implies that $\tilde{\boldsymbol{A}}[i,j] \leq \tilde{\boldsymbol{A}}[i',j]$ for $i' < i$. Together, these imply that $\tilde{\boldsymbol{A}}$ is lower triangular.

□

3.9.2. Proofs of Results in Section 3.4

Proof of Theorem 3.4.1

We prove Theorem 3.4.1 via a series of implications:

(i) $\implies$ *(ii).* If there is a response type $\boldsymbol{s}$ and a treatment t for which *(ii)* is not true, then the sequence

$$\left(\mathbf{1}[\boldsymbol{s}[z_1^{(t)}] = t], \ldots, \mathbf{1}[\boldsymbol{s}[z_{N_Z}^{(t)}] = t]\right)$$

is not increasing.

(ii) $\implies$ *(iii).* If there is a 2×2 matrix of $\boldsymbol{R}$ of the form

$$\begin{array}{cc} \boldsymbol{s} & \boldsymbol{s}' \\ \begin{pmatrix} t & t' \\ t'' & t \end{pmatrix} & \begin{array}{c} z \\ z' \end{array} \end{array}$$

then we have $\mathbf{1}[\boldsymbol{s}[z] = t] > \mathbf{1}[\boldsymbol{s}[z'] = t]$ while $\mathbf{1}[\boldsymbol{s}'[z] = t] < \mathbf{1}[\boldsymbol{s}'[z'] = t]$, a violation of *(ii)*. A symmetric argument holds for the other restricted submatrix of $\boldsymbol{R}$.

(iii) $\implies$ *(iv).* First notice that *(iv)* is equivalent to the matrix $\boldsymbol{U}$ being lonesum by part three of Lemma 3.9.1. Further notice that the matrix $\boldsymbol{U}$ is lonesum if and only if each $\boldsymbol{B}_t$ is lonesum. Because there are no restricted submatrices of $\boldsymbol{R}$ of the form in *(iii)* there are no submatrices of any $\boldsymbol{B}_t$ of the form either

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

By part *(ii)* of Lemma 3.9.1, this is equivalent to each $\boldsymbol{B}_t$ being lonesum.

(iv) $\implies$ *(v)*. Item *(iv)* is equivalent to each $\boldsymbol{B}_t$ being lonesum. We seek to use the fourth item of Lemma 3.9.1. With this in mind, define the functions

$$\zeta(z,t) \equiv \text{row sum of the } z^{\text{th}} \text{ row of the matrix } \boldsymbol{B}_t$$

$$\varphi(\boldsymbol{s},t) \equiv \#\text{ of columns of } \boldsymbol{B}_t \text{ with a larger column sum than that of its } \boldsymbol{s}^{\text{th}} \text{ column.}$$

Because $\boldsymbol{S}$ is implicitly a function of unobserved confounders $\boldsymbol{V}$, we can also write $\varphi(\cdot,t)$ as a function of $\boldsymbol{V}$. By the definition of $\boldsymbol{B}_t = \mathbf{1}[\boldsymbol{R}=t]$ and the fourth item of Lemma 3.9.1, we can then write

$$\mathbf{1}[T=t] = \mathbf{1}[\zeta(Z,t) \geq \varphi(\boldsymbol{V},t)].$$

By the first item of Lemma 3.9.1, there is a reordering of the rows and columns of $\boldsymbol{B}_t$ such that $\boldsymbol{B}_t$ is lower triangular. Let $z_1^{(t)}$ denote the instrument associated with the "top" row of this matrix, $z_2^{(t)}$ denote the second row of this matrix, and so on till $z_{N_Z}^{(t)}$.

Then, the function ζ satisfies $\zeta(z_{k+1}^{(t)},t) \geq \zeta(z_k^{(t)},t)$ for $k=1,...,N_Z-1$, by definition of the sequence $(z_1^{(t)},\ldots,z_{N_Z}^{(t)})$; weakly more response types must be taking up treatment t for each successive value of this sequence.

(v) $\implies$ *(i)*. Since $\zeta(z_{k+1}^{(t)},t) > \zeta(z_k^{(t)},t)$ for $k=1,...,N_Z-1$, if *(v)* holds we must have that

$$(\mathbf{1}[T_i(z_1^{(t)})=t],\ldots,\mathbf{1}[T_i(z_{N_Z}^{(t)})=t])$$

is a weakly increasing sequence in $\{0,1\}$ for all $i \in \mathcal{I}$.

Proof of Theorem 3.4.2

We prove Theorem 3.4.2 via a series of implications:

(i) $\implies$ *(ii)*. Take any strict ordering on $\mathcal{T}$. Suppose there is a violation of *(ii)* for some

$i \in \mathcal{I}$. Then for that particular $i \in \mathcal{I}$ the sequence

$$(T_i(z_1), \ldots, T_i(z_{N_Z}))$$

is not increasing with respect to the ordering on $\mathcal{T}$. If there is no such ordering satisfying *(ii)*, then *(i)* cannot be satisfied.

(ii) $\implies$ *(iii)*. Suppose there is a 2×2 submatrix of $\boldsymbol{R}$ of the form:

$$\begin{array}{c} \begin{matrix} \boldsymbol{s} & \boldsymbol{s}' \end{matrix} \\ \begin{pmatrix} t & t' \\ t'' & t''' \end{pmatrix} \begin{matrix} z \\ z' \end{matrix} \end{array}$$

for some $t'' > t$ and $t''' < t'$. This means that $\boldsymbol{s}[z'] > \boldsymbol{s}[z]$ while $\boldsymbol{s}'[z] > \boldsymbol{s}'[z']$. These two statememnts cannot both be true under *(ii)*. A symmetric argument applies for the other 2×2 restricted submatrix.

(iii) $\implies$ *(iv)*. Notice that by part 3 of Lemma 3.9.1 that *(iv)* is equivalent to $\boldsymbol{O}$ being lonesum. Suppose that $\boldsymbol{O} = [\boldsymbol{B}^*_{t_1}, \ldots, \boldsymbol{B}^*_{t_{N_T}}]$ is not lonesum. That is, by the lonesum characterization (3.9.1) there is a 2×2 submatrix of $\boldsymbol{O}$ of the form

$$\begin{array}{c} \begin{matrix} \boldsymbol{s} & \tilde{\boldsymbol{s}} \end{matrix} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{matrix} z \\ z' \end{matrix} \end{array} \quad \text{or} \quad \begin{array}{c} \begin{matrix} \boldsymbol{s} & \tilde{\boldsymbol{s}} \end{matrix} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{matrix} z \\ z' \end{matrix} \end{array}.$$

WLOG suppose there is a 2×2 submatrix of $\boldsymbol{O}$ of the first form. This indicates that for some $t \in \mathcal{T}$, the instrumental switch from z to z' induces agents of response type $\boldsymbol{s}$ to switch from a treatment weakly below t to a treatment strictly greater than t. That is $\boldsymbol{s}[z] < \boldsymbol{s}[z']$. Conversely, for some treatment $t' \in \mathcal{T}$ the instrumental switch from z to z' induces agents of

response $\tilde{\boldsymbol{s}}$ to switch from a treatment strictly greater than t' to a treatment weakly lower than t'. That is $\tilde{\boldsymbol{s}}[z] < \tilde{\boldsymbol{s}}[z']$.

These two statements are incompatible with each other, so we cannot have that $\boldsymbol{s} = \tilde{\boldsymbol{s}}$. Leting $t = \boldsymbol{s}[z], t' = \tilde{\boldsymbol{s}}[z], t'' = \boldsymbol{s}[z']$, and $t''' = \tilde{\boldsymbol{s}}[z']$, this implies a 2×2 submatrix of $\boldsymbol{R}$ of the form

$$\begin{matrix} \boldsymbol{s} & \tilde{\boldsymbol{s}} \\ \begin{pmatrix} t & t' \\ t'' & t''' \end{pmatrix} & \begin{matrix} z \\ z' \end{matrix} \end{matrix}$$

with $t < t''$ and $t' > t'''$. This is a violation of *(iii)*. Similarly, considering 2×2 submatrices of $\boldsymbol{O}$ of the second form, we can find a violation of the other pattern restricted by *(iii)*.

(iv) $\implies$ *(v)*. Item *(iv)* is equivalent to $\boldsymbol{O}$ being lonesum. This in turn implies that $\boldsymbol{B}_t^* = \mathbf{1}[\boldsymbol{R} \geq t]$ is lonesum for each $t \in \mathcal{T}$. With this in mind, define the functions

$$\zeta(z, t) \equiv \text{row sum of the } z^{\text{th}} \text{ row of the matrix } \boldsymbol{B}_t^*$$
$$\varphi(\boldsymbol{s}, t) \equiv \#\text{ of columns of } \boldsymbol{B}_t^* \text{ with a larger column sum than that of its } \boldsymbol{s}^{\text{th}} \text{ column.}$$

By the first item of Lemma 3.9.1, there is a reordering of the rows and columns of $\boldsymbol{O}$ such that $\boldsymbol{O}$ is lower triangular. Let $z_1^{(t)}$ denote the instrument associated with the "top" row of this matrix, $z_2^{(t)}$ denote the second row of this matrix, and so on till $z_{N_Z}^{(t)}$.

Because $\boldsymbol{S}$ is implicity a function of unobserved confounders $\boldsymbol{V}$, we can also write $\varphi(\cdot, t)$ as a function of $\boldsymbol{V}$. By the definition of $\boldsymbol{B}_t^* = \mathbf{1}[\boldsymbol{R} \geq t]$ and the fourth item of Lemma 3.9.1, we can then write

$$\mathbf{1}[T \geq t] = \mathbf{1}[\zeta(Z, t) \geq \varphi(\boldsymbol{V}, t)]$$

for each t. Moreover, by definition of the sequence $(z_1, \ldots, z_{N_Z})$ we know that, for each treatment $t \in \mathcal{T}$, weakly more response types take up treatments larger than t as the

instrument cycles through the sequence $(z_1, \ldots, z_{N_Z})$. By definition of the $\zeta(\cdot, \cdot)$ function then, $\zeta(z_{k+1}, t) > \zeta(z_k, t)$ for $k = 1, \ldots, N_Z - 1$ and all t.

(v) $\implies$ *(i)*. Since $\zeta(z_{k+1}, t) > \zeta(z_k, t)$ for $k = 1, \ldots, N_Z - 1$ and all t, if holds *(v)* we must have that, for all t

$$(\mathbf{1}[T_i(z_1) \geq t], \ldots, \mathbf{1}[T_i(z_{N_Z}) \geq t])$$

is a weakly increasing sequence in $\{0, 1\}$ for all $i \in \mathcal{I}$. This implies that

$$(T_i(z_1), \ldots, T_i(z_{N_Z}))$$

must be a weakly increasing sequence with respect to the ordering on $\mathcal{T}$ for all $i \in \mathcal{I}$.

3.9.3. Proofs of Results in Section 3.5

Proof of Theorem 3.5.1

We show a system of implications.

(i) $\iff$ *(ii)*. This is provided by Lemma 3.5.1.

(ii) $\iff$ *(iv)*. This follows from the discussion in Section 3.10, namely the decomposition of $\beta_{z,z'}$ in equation (3.10.6), and the definition of an interpretable causal parameter in (3.5.3). The decomposition of $\beta_{z,z'}$ gives the forward direction. The definition of an interpretable causal parameter gives the backwards direction: if there is a negative weight there must be a pair of treatments t, t' such that some agents that are switching from t to t' as the instrument ranges from z to z' whereas that same instrumental switch moves other agents from t' to t.

(iii) $\implies$ *(ii)*. We consider the contrapositive. Suppose there is no binary matrix $\boldsymbol{B}$ element-wise less than or equal to $\sum_{t'' \neq t,t'} \boldsymbol{B}_{t''}$ such that $\boldsymbol{B}_t + \boldsymbol{B}$ is lonesum. This means

that $\boldsymbol{B}_t$ is lonesum, so that by Theorem 3.4.1 there is a 2×2 submatrix $\boldsymbol{R}$ of either the form

$$\begin{pmatrix} t & t''' \\ t'' & t \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} t'' & t \\ t & t''' \end{pmatrix},$$

for some $t'', t''' \neq t$. If either $t'' \neq t'$ or $t''' \neq t'$, then we can find a binary matrix that is element wise less than $\sum_{t'' \neq t, t'} \boldsymbol{B}_{t'''}$ to "fill in the gap" and get rid of the restricted pattern. In particular we can take $\tilde{\boldsymbol{B}}$ to be the matrix that is equal to one at the position of either t'' or t''' and zero everywhere else. If there is no such matrix that eliminates the restricted pattern then both $t'' = t'$ and $t''' = t'$. So we have the restricted pattern (3.5.2) in $\boldsymbol{R}$.

(ii) $\iff$ *(iii)*. The ij^{th} element of $\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'}$ is given

$$\sum_{z=z_1}^{N_Z} \mathbf{1}[\boldsymbol{s}_i[z] = t]\mathbf{1}[\boldsymbol{s}_j[z] = t'],$$

this is nonzero if and only if we have $\boldsymbol{s}_i[z] = t$ and $\boldsymbol{s}_j[z] = t'$ for some instrument value z. Similarly, the ij^{th} element of $(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'})^\intercal = \boldsymbol{B}_{t'}^\intercal \boldsymbol{B}_t$ is given

$$\sum_{z=z_1}^{N_Z} \mathbf{1}[\boldsymbol{s}_i[z] = t']\mathbf{1}[\boldsymbol{s}_j[z] = t].$$

This is non-zero if and only if we have $\boldsymbol{s}_i[z'] = t'$ and $\boldsymbol{s}_j[z'] = t$ for some insrument value z'.

Then, the ij^{th} element of the Hadamard product $(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'}) \odot (\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'})^\intercal$ is non-zero if and only if the ij^{th} elements of both $(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'})$ and $(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'})^\intercal$ are non-zero. By the characterizations above and because each response type is a well defined function, this is equivalent to $\boldsymbol{s}_i[z] = t, \boldsymbol{s}_i[z'] = t'$ but $\boldsymbol{s}_j[z] = t', \boldsymbol{s}_j[z'] = t$ for some $z' \neq z$. This is equivalent to the restricted pattern (3.5.2) existing between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ for instrument values z, z' and the specific treatment values t and t'.

All elements of $(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'}) \odot (\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'})^\intercal$ being equal to zero is then equivalent to their being

no restricted patterns (3.5.2) between the specific treatment values t and t' in the response matrix $\boldsymbol{R}$.

Because each $(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'}) \odot (\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'})^\intercal$ has weakly positive entries, checking whether

$$\iota^\intercal \left(\sum_{(t,t') \in \mathcal{C}_2(\mathcal{T})} (\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'}) \odot (\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'})^\intercal \right) \iota = 0$$

is equivalent to checking whether each $(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'}) \odot (\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'})^\intercal$ is equal to the zero matrix. By the discussion above, this is equivalent to checking whether there are no matrices of the form (3.5.2) for any $t, t' \in \mathcal{T}$.

Proof of Theorem 3.5.2

Let $\Psi_{\boldsymbol{U}}(t) = \big((\tilde{\mathbb{1}} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big) \odot \big((\mathbb{1} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big)^\intercal$, where $\tilde{\mathbb{1}}$ is a $N_Z \times N_S$ matrix of element ones. Using this notation, we can rewrite the UM verification in Item (iv) of Theorem 3.4.1 as the following sum:

$$\begin{aligned}
\|\Psi_{\boldsymbol{U}}\| &= \|\big((\mathbb{1} - \boldsymbol{U})^\intercal \boldsymbol{U}\big) \odot \big((\mathbb{1} - \boldsymbol{U})^\intercal \boldsymbol{U}\big)^\intercal\| \\
&= \|\sum_{t \in \mathcal{T}} \big((\tilde{\mathbb{1}} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big) \odot \big((\tilde{\mathbb{1}} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big)^\intercal\| \\
&= \sum_{t \in \mathcal{T}} \|\big((\tilde{\mathbb{1}} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big) \odot \big((\tilde{\mathbb{1}} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big)^\intercal\| \\
&= \sum_{t \in \mathcal{T}} \|\Psi_{\boldsymbol{U}}(t)\|
\end{aligned}$$

The first equality is from the definition of the verification in Theorem 3.4.1. The second equality arises from the construction of matrix $\boldsymbol{U}$. The third equality is due to the fact that all elements of $\Psi_{\boldsymbol{U}}(t)$ are either zero or a natural number.

We can use the fact that $\sum_{t \in \mathcal{T}} \boldsymbol{B}_t = \tilde{\mathbb{1}}$ to express matrix $\Psi_{\boldsymbol{U}}(t)$ as:

$$\Psi_{\boldsymbol{U}}(t) = \Big((\tilde{\mathbb{1}} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\Big) \odot \Big((\tilde{\mathbb{1}} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\Big)^\intercal$$

$$
\begin{aligned}
&= \left(\left(\left(\sum_{t' \in \mathcal{T}} \boldsymbol{B}_{t'} \right) - \boldsymbol{B}_t \right)^\intercal \boldsymbol{B}_t \right) \odot \left(\left(\left(\sum_{t' \in \mathcal{T}} \boldsymbol{B}_{t'} \right) - \boldsymbol{B}_t \right)^\intercal \boldsymbol{B}_t \right)^\intercal \\
&= \left(\left(\sum_{t' \in \mathcal{T} \setminus \{t\}} \boldsymbol{B}_{t'} \right)^\intercal \boldsymbol{B}_t \right) \odot \left(\left(\sum_{t' \in \mathcal{T} \setminus \{t\}} \boldsymbol{B}_{t'} \right)^\intercal \boldsymbol{B}_t \right)^\intercal \\
&= \left(\sum_{t' \in \mathcal{T} \setminus \{t\}} \boldsymbol{B}_{t'}^\intercal \boldsymbol{B}_t \right) \odot \left(\sum_{t' \in \mathcal{T} \setminus \{t\}} \boldsymbol{B}_{t'}^\intercal \boldsymbol{B}_t \right)^\intercal \\
&= \left(\sum_{t' \in \mathcal{T} \setminus \{t\}} \boldsymbol{B}_{t'}^\intercal \boldsymbol{B}_t \right) \odot \left(\sum_{t' \in \mathcal{T} \setminus \{t\}} \boldsymbol{B}_t^\intercal \boldsymbol{B}_{t'} \right) \\
&= \sum_{t' \in \mathcal{T} \setminus \{t\}} \left(\boldsymbol{B}_{t'}^\intercal \boldsymbol{B}_t \right) \odot \left(\boldsymbol{B}_{t'}^\intercal \boldsymbol{B}_t \right)^\intercal + 2 \sum_{t', t'' \in \mathcal{T} \setminus \{t\}} \left(\boldsymbol{B}_{t'}^\intercal \boldsymbol{B}_t \right) \odot \left(\boldsymbol{B}_t^\intercal \boldsymbol{B}_{t''} \right)^\intercal \\
&= \sum_{t' \in \mathcal{T} \setminus \{t\}} \Psi(t', t, t, t') + 2 \sum_{t', t'' \in \mathcal{T} \setminus \{t\}} \Psi(t', t, t, t'')
\end{aligned}
$$

The derivation above use simple rules of matrix algebra and the formula for the product of sums. We use the fact that $\Psi(t', t, t, t'')^\intercal = \Psi(t, t', t'', t)$ and express the transpose of $\Psi_{\boldsymbol{U}}(t)$ as:

$$
\Psi_{\boldsymbol{U}}(t)^\intercal = \sum_{t' \in \mathcal{T} \setminus \{t\}} \Psi(t, t', t', t) + 2 \sum_{(t', t'') \in \mathcal{C}_2(\mathcal{T} \setminus \{t\})} \Psi(t, t', t'', t) \tag{3.9.6}
$$

Recall that the elements of matrix $\Psi(t, t', t'', t''')$ are either zero or natural numbers. Thus, equation (3.9.6) implies that $\|\Psi_{\boldsymbol{U}}(t)\| = 0$ (or equivalently $\|\Psi_{\boldsymbol{U}}(t)^\intercal\| = 0$) if and only if:

$$
\|\Psi(t, t', t', t)\| = 0 \text{ for all } t' \in \mathcal{T} \setminus \{t\} \tag{3.9.7}
$$

$$
\text{and } \|\Psi(t, t', t'', t)\| = 0 \text{ for all combinations of } t', t'' \in \mathcal{T} \setminus \{t\}. \tag{3.9.8}
$$

Now $\|\Psi_{\boldsymbol{U}}\| = 0$ only and only if $\|\Psi_{\boldsymbol{U}}(t)\| = 0$ for all $t \in \mathcal{T}$, which completes the proof.

3.9.4. Proof of Theorem 3.5.3

Lemmas

Proof of Lemma 3.5.1. Suppose there is a violation of MM (3.5.1). This is equivalent to there being pair of response types $\boldsymbol{s}, \boldsymbol{s}'$, a pair of treatments z, z', and a pair of treatments t, t' such that

$$\begin{aligned} &\mathbf{1}[\boldsymbol{s}[z] = t]\mathbf{1}[\boldsymbol{s}'[z'] = t'] > \mathbf{1}[\boldsymbol{s}[z] = t']\mathbf{1}[\boldsymbol{s}'[z'] = t] \\ \text{and} \quad &\mathbf{1}[\boldsymbol{s}'[z] = t]\mathbf{1}[\boldsymbol{s}[z'] = t'] < \mathbf{1}[\boldsymbol{s}'[z] = t']\mathbf{1}[\boldsymbol{s}[z'] = t]. \end{aligned}$$

This is in turn equivalent to $\boldsymbol{s}[z] = t$, $\boldsymbol{s}[z'] = t'$ and $\boldsymbol{s}'[z] = t', \boldsymbol{s}'[z'] = t$, which is equivalent (up to a relabeling of $\boldsymbol{s}$ and $\boldsymbol{s}'$) to the restricted pattern (3.5.2) appearing in the response matrix $\boldsymbol{R}$.

Proof of Lemma 3.5.2. Follows from Theorems 3.5.2 and 3.5.3. That MM holds when OM and UM fail can be seen via examples in Section 3.7.

3.9.5. Proof of Results in Section 3.6

Proof of Theorem 3.6.1

Consider any pair of instrument values $z, z' \in \mathcal{Z}$ and any pair of treatments $t, t' \in \mathcal{T}$. Without loss of generality, it is enough show there are no 2×2 submatrices of the form

$$\begin{array}{cc} \boldsymbol{s} & \boldsymbol{s}' \\ \begin{pmatrix} t & t' \\ t' & t \end{pmatrix} & \begin{array}{c} z \\ z' \end{array} \end{array}.$$

Define $\Delta_t \equiv \boldsymbol{L}[z', t] - \boldsymbol{L}[z, t]$ and $\Delta_{t'} \equiv \boldsymbol{L}[z', t'] - \boldsymbol{L}[z, t']$. There are two scenarios. Either $\Delta_t \leq \Delta_{t'}$ or $\Delta_t \geq \Delta_{t'}$. In each case, we have the following behavioral restrictions from the

Choice Rule (3.6.2).

$$\text{If} \quad \Delta_t \leq \Delta_{t'} \quad \text{then} \quad T_i(z) = t' \implies T_i(z') \neq t$$
$$\text{If} \quad \Delta_{t'} \leq \Delta_t \quad \text{then} \quad T_i(z) = t \implies T_i(z') \neq t'$$

The first restriction would eliminate the response type $\boldsymbol{s}'$ from the matrix $\boldsymbol{R}$ while the second restriction would eliminate the response type $\boldsymbol{s}$ from the response matrix $\boldsymbol{R}$. In either case, we cannot have the restricted 2×2 submatrix displayed at the top of the proof.

3.10. Appendix: 2SLS Analysis

3.10.1. Interpretation of 2SLS under Ordered and Unordered Monotonicity

Under ordered monotonicity and a binary instruments, Angrist and Imbens (1995) show that the 2SLS estimand identifies the following:

$$\begin{aligned} \beta_{\text{2SLS}} &= \frac{\mathbb{E}[Y \mid Z = z_1] - \mathbb{E}[Y \mid Z = z_0]}{\mathbb{E}[T \mid Z = z_1] - \mathbb{E}[T \mid Z = z_0]} \\ &= \sum_{j=1}^{N_T} \omega_{t_j, t_{j-1}} \mathbb{E}[Y(t_j) - Y(t_{j-1}) \mid \boldsymbol{S} \in \mathcal{S}_{t_j, t_{j-1}}] \end{aligned} \tag{3.10.1}$$

where $\mathcal{S}_{t_j, t_{j-1}} \equiv \{\boldsymbol{s} \in \mathcal{S}; \boldsymbol{s}[z_1] \geq t_j > \boldsymbol{s}[z_0]\}$, that is the sets of response types for whom a change in instrument receipt from z_0 to z_1 induces a change in treatment from strictly "below" t_j to weakly "above" t_j. The weights $\omega_{t_j, t_{j-1}}$ are positive and given:

$$\omega_{t_j, t_{j-1}} = \frac{\Pr(\boldsymbol{S} \in \mathcal{S}_{t_j, t_{j-1}})}{\sum_{j=1}^{N_T} \Pr(\boldsymbol{S} \in \mathcal{S}_{t_j, t_{j-1}})}. \tag{3.10.2}$$

Unordered monotonicity also allows 2SLS type estimands to be expressed in terms of a weighted average of LATE parameters with positive weights.[1] In this setting the 2SLS

[1] In addition, Buchinsky and Pinto (2021) show that any variation in the instrumental variable can be used to identify a meaningful counterfactual outcome mean. For instance, the 2SLS estimate that uses a choice indicator for t and any IV-values $z, z' \in \mathcal{Z}$, such that $P(T = t|Z = z) > (T = t'|Z = z)$, identifies the

numerator can be decomposed

$$\begin{aligned}\beta^{u}_{\text{2SLS}}(z,z') &= \mathbb{E}[Y \mid Z=z] - \mathbb{E}[Y \mid Z=z'] \\ &= \sum_{\{t,t'\},t\neq t'} \omega^{u}_{t,t'} \mathbb{E}[Y(t) - Y(t') \mid \boldsymbol{S} \in \mathcal{S}_{t,t'}(z,z')]\end{aligned} \tag{3.10.3}$$

where $\mathcal{S}_{t,t'}(z,z') \equiv \{\boldsymbol{s} : \boldsymbol{s}[z]=t, \boldsymbol{s}[z']=t'\}$ is the set of response types that switch treatments from t to t' as the instrument varies from z to z'. Under unordered monotonicity, the same instrument switch cannot induce some agents to switch towards choice t while inducing others to switch away from choice t. So, we must have either $\mathcal{S}_{t,t'}(z,z') = \emptyset$ or $\mathcal{S}_{t',t}(z,z') = \emptyset$ (or both). The weights $\omega^{u}_{t,t'}$ are weakly positive and given $\omega^{u}_{t,t'} = \Pr(\boldsymbol{S} \in \mathcal{S}_{t,t'})$.

3.10.2. General Unique Decomposition

Using the identification equality in (3.3.9) we can rewrite

$$\mathbb{E}[Y \mid Z=z] = \sum_{\boldsymbol{s}\in\text{supp}(\boldsymbol{S})} \sum_{t\in\mathcal{T}} \mathbf{1}[\boldsymbol{s}[z]=t] \mathbb{E}[Y(t) \mid \boldsymbol{S}=\boldsymbol{s}] \Pr(\boldsymbol{S}=\boldsymbol{s})$$

$$\mathbb{E}[Y \mid Z=z'] = \sum_{\boldsymbol{s}\in\text{supp}(\boldsymbol{S})} \sum_{t\in\mathcal{T}} \mathbf{1}[\boldsymbol{s}[z']=t] \mathbb{E}[Y(t) \mid \boldsymbol{S}=\boldsymbol{s}] \Pr(\boldsymbol{S}=\boldsymbol{s})$$

Using these, we can express the quasi-2SLS estimand as the following

$$\beta_{z,z'} \equiv \mathbb{E}[Y \mid Z=z] - \mathbb{E}[Y \mid Z=z'] \tag{3.10.4}$$

$$= \sum_{\boldsymbol{s}\in\text{supp}(\boldsymbol{S})} \sum_{t\in\mathcal{T}} \big(\mathbf{1}[\boldsymbol{s}[z]=t] - \mathbf{1}[\boldsymbol{s}[z']=t]\big) \mathbb{E}[Y(t) \mid \boldsymbol{S}=\boldsymbol{s}] \Pr(\boldsymbol{S}=\boldsymbol{s}) \tag{3.10.5}$$

$$= \sum_{\{t,t'\},t\neq t'} \mathbb{E}[Y(t) - Y(t') \mid \boldsymbol{S} \in \mathcal{S}_{t,t'}(z,z')] \Pr(\boldsymbol{S} \in \mathcal{S}_{t,t'}(z,z')), \tag{3.10.6}$$

---

following parameter:

$$\beta^{t}_{\text{2SLS}}(z,z') = \frac{\mathbb{E}[Y\mathbf{1}[T=t] \mid Z=z] - \mathbb{E}[Y\mathbf{1}[T=t] \mid Z=z']}{\Pr(T=t \mid Z=z) - \Pr(T=t \mid Z=z')} = E(Y(t) | \boldsymbol{S} \in \mathcal{S}^{t}_{z,z'}),$$

where $\mathcal{S}^{t}_{z,z'} = \{\boldsymbol{s} : \boldsymbol{s}[z]=t, \boldsymbol{s}[z'] \neq t\}$ is the set of response types that switch from treatment choice t to any other treatment choice as the instrument varies from z to z'.

where the last equality is due to the fact that sets $\mathcal{S}_{t,t'}(z,z')$, defined below (3.10.3), form a partition of $\text{supp}(\boldsymbol{S})$ as t,t' ranges in $\mathcal{T}$. Equation (3.10.6) holds regardless of any monotonicity assumption. That is, no matter what the restriction is on the support of $\boldsymbol{S}$, we will always be able to rewrite the 2SLS numerator as in (3.10.6).

In view of Lemma 3.5.1, a violation of MM is equivalent to there being a pair of treatments t,t' such that the sets $\mathcal{S}_{t,t'}(z,z')$ and $\mathcal{S}_{t',t}(z,z')$ are *both* nonempty. This induces negative weights in the 2SLS estimand; both $\mathbb{E}[Y(t)-Y(t') \mid \mathcal{S}_{t,t'}(z,z')]$ and $\mathbb{E}[Y(t')-Y(t) \mid \mathcal{S}_{t',t}(z,z')]$ are represented in the decomposition (3.10.6). This in turn limits our ability to use the 2SLS estimand to gain useful insight into the direction of causal effects. The partial minimal monotonicity criterion is then crucial for interpreting $\beta_{z,z'}$ as a type of interpretable causal parameter defined in (3.5.3).

3.11. Appendix: Ordered vs. Unordered Example

We consider a setup with three treatments, $\mathcal{T}=\{t_1,t_2,t_3\}$, and three instruments, $\mathcal{Z}=\{z_1,z_2,z_3\}$. Response matrices (3.11.1)–(3.11.2) below are useful to understand the difference between ordered and unordered monotonicity conditions:

$$\boldsymbol{R}_1 = \overbrace{\underbrace{\begin{bmatrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \\ t_1 & t_2 & t_3 & t_1 & t_1 & t_2 & t_1 \\ t_1 & t_2 & t_3 & t_2 & t_3 & t_2 & t_1 \\ t_1 & t_2 & t_3 & t_3 & t_3 & t_3 & t_2 \end{bmatrix}}_{\text{Ordered AND Unordered}} \begin{matrix} z_1 \\ z_2 \\ z_3 \end{matrix}}^{\text{Ordered but NOT Unordered}} \tag{3.11.1}$$

$$\boldsymbol{R}_2 = \overbrace{\underbrace{\begin{bmatrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7^* \\ t_1 & t_2 & t_3 & t_1 & t_1 & t_2 & t_1 \\ t_1 & t_2 & t_3 & t_2 & t_3 & t_2 & t_2 \\ t_1 & t_2 & t_3 & t_3 & t_3 & t_3 & t_1 \end{bmatrix}}_{\text{Unordered but NOT Ordered}} \begin{matrix} z_1 \\ z_2 \\ z_3 \end{matrix}}^{\text{Ordered AND Unordered}} \tag{3.11.2}$$

Columns $\boldsymbol{s}_1, ..., \boldsymbol{s}_7$ of response matrix $\boldsymbol{R}_1$ (3.11.1) denote response-types. Each column describes the sequence of counterfactual choices, $(T_i(z_1), T_i(z_2), T_i(z_3))$, for an agent in that column's response type. The counterfactual treatment in each of these sequences is weakly increasing with respect to the ordering $t_1 \leq t_2 \leq t_3$; for any agent $i \in \mathcal{I}$, $T_i(z_1) \leq T_i(z_2) \leq T_i(z_3)$. Thus OM-Sequence (3.4.3) holds.

However, response types $\boldsymbol{s}_6, \boldsymbol{s}_7$ in $\boldsymbol{R}_1$ violate the sequential representation of unordered monotonicity in (3.4.4) for choice t_2. Consider two agents $i, i' \in \mathcal{I}$ such that $\boldsymbol{S}_i = \boldsymbol{s}_6$ and $\boldsymbol{S}_{i'} = \boldsymbol{s}_7$. The sequence of t_2-indicators for agent i is weakly decreasing while the sequence for agent i' is weakly increasing

$$\big(\mathbf{1}[T_i(z_1) = t_2], \mathbf{1}[T_i(z_2) = t_2], \mathbf{1}[T_i(z_3) = t_2]\big) = (1, 1, 0)$$
$$\big(\mathbf{1}[T_{i'}(z_1) = t_2], \mathbf{1}[T_{i'}(z_2) = t_2], \mathbf{1}[T_{i'}(z_3) = t]\big) = (0, 0, 1).$$

This represents a violation of UM-Sequence (3.4.4) for the sequencing of $\mathcal{Z}$, (z_1, z_2, z_3). Moreover, because the switch from z_2 to z_3 induces agent i to move strictly away from treatment choice t_2 while moving agent i' strictly towards treatment choice t_2, there is no other sequencing of $\mathcal{Z}$ that would satisfy the requirement of UM Sequence (3.4.4). We can conclude that the response matrix $\boldsymbol{R}_1$ does not satisfy unordered monotonicity.

Response matrix $\boldsymbol{R}_2$ (3.11.2) replaces $\boldsymbol{s}_7$ in $\boldsymbol{R}_1$ with $\boldsymbol{s}_7^*$. The treatment indexes in $\boldsymbol{s}_7^*$ are not weakly increasing with respect to the ordering $t_1 \leq t_2 \leq t_3$. Thus the ordered monotonicity that held for $\boldsymbol{R}_1$ does not hold for $\boldsymbol{R}_2$. Indeed, the reader can confirm that there is no ordering $\mathcal{T}$ that satisfies OM-Sequence (3.4.3). In this case, though, the response matrix $\boldsymbol{R}_2$ satisfies unordered monotonicity. Equations (3.11.3)–(3.11.5) are instructive in establishing this fact.

$$\text{Reordered rows of } \boldsymbol{R}_2 \text{ for } t_1 \;=\; \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7^* \end{matrix} \\ \begin{bmatrix} t_1 & t_2 & t_3 & t_2 & t_3 & t_2 & t_2 \\ t_1 & t_2 & t_3 & t_3 & t_3 & t_3 & t_1 \\ t_1 & t_2 & t_3 & t_1 & t_1 & t_2 & t_1 \end{bmatrix} \end{array} \begin{matrix} z_2 \\ z_3 \\ z_1 \end{matrix}, \tag{3.11.3}$$

$$\text{Reordered rows of } \boldsymbol{R}_2 \text{ for } t_2 \;=\; \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7^* \end{matrix} \\ \begin{bmatrix} t_1 & t_2 & t_3 & t_3 & t_3 & t_3 & t_1 \\ t_1 & t_2 & t_3 & t_1 & t_1 & t_2 & t_1 \\ t_1 & t_2 & t_3 & t_2 & t_3 & t_2 & t_2 \end{bmatrix} \end{array} \begin{matrix} z_3 \\ z_1 \\ z_2 \end{matrix}, \tag{3.11.4}$$

$$\text{Reordered rows of } \boldsymbol{R}_2 \text{ for } t_3 \;=\; \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7^* \end{matrix} \\ \begin{bmatrix} t_1 & t_2 & t_3 & t_1 & t_1 & t_2 & t_1 \\ t_1 & t_2 & t_3 & t_2 & t_3 & t_2 & t_2 \\ t_1 & t_2 & t_3 & t_3 & t_3 & t_3 & t_1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \end{matrix}. \tag{3.11.5}$$

Equation (3.11.3) reorders the rows of $\boldsymbol{R}_2$ in (3.11.2) from (z_1, z_2, z_3) to (z_2, z_3, z_1). At each move along the sequence (z_2, z_3, z_1), additional response types switch to treatment t_1 and no response types switch away from t_1. This means that

$$\mathbf{1}[T_i(z_2) = t_1] \leq \mathbf{1}[T_i(z_3) = t_1] \leq \mathbf{1}[T_i(z_1) = t_1]$$

holds for all agents $i \in \mathcal{I}$, regardless of response type. Thus UM Sequence (3.4.4) holds for t_1. By symmetric logic, equation (3.11.4) demonstrates that UM Sequence (3.4.4) holds for t_2 using the IV sequence (z_3, z_1, z_2) and equation (3.11.5) shows that UM Sequence holds for t_3 using sequence (z_1, z_2, z_3). We conclude that unordered monotonicity holds.

Response matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ in (3.11.1)–(3.11.2) show that ordered monotonicity does not imply unordered monotonicity nor vice-versa. Ordered monotoncity holds for $\boldsymbol{R}_1$ but not for $\boldsymbol{R}_2$ while unordered monotonicity holds for $\boldsymbol{R}_2$ but not for $\boldsymbol{R}_1$. The two monotonicity conditions can intersect, both ordered and unordered monotonicity hold for the submatrix generated by response-types $\boldsymbol{s}_1$ to $\boldsymbol{s}_6$.

3.12. Appendix: Additional Information Regarding the Examples of Section 3.7

3.12.1. Verifying Unordered Monotonicity

We seek to show that the response matrix 3.7.2 is a case of UM (3.4.4) using the verification matrix of item (*iv*) of Theorem 3.4.1. The matrix is presented below for convenience.

$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \end{matrix} \\ \begin{bmatrix} t_1 & t_1 & t_1 & t_1 & t_2 & t_2 & t_3 \\ t_1 & t_1 & t_3 & t_3 & t_2 & t_3 & t_3 \\ t_1 & t_2 & t_2 & t_3 & t_2 & t_2 & t_3 \\ t_1 & t_1 & t_1 & t_1 & t_2 & t_2 & t_3 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix}$$

Let $\boldsymbol{B}_t = \mathbf{1}[\boldsymbol{R} = t]; t \in \{t_1, t_2, t_3\}$ denote the binary matrices corresponding to response matrix (3.7.2). Those are displayed below:

$$\boldsymbol{B}_{t_1} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix}$$

$$\boldsymbol{B}_{t_2} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix}$$

$$\boldsymbol{B}_{t_3} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix}$$

Unordered monotonicity holds if and only if the binary matrices $\boldsymbol{B}_{t_1}, \boldsymbol{B}_{t_2}, \boldsymbol{B}_{t_3}$ are lonesum. For item (*iv*) of Theorem 3.4.1 to hold, it suffices to show that $\|\Psi_{\boldsymbol{U}}(t)\| = 0$ for all $t \in \{t_1, t_2, t_3\}$ where $\Psi_{\boldsymbol{U}}(t)$ is given by $\Psi_{\boldsymbol{U}}(t) \equiv \big((\mathbb{1} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big) \odot \big((\mathbb{1} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big)^\intercal$. It is useful to express $\Psi_{\boldsymbol{U}}(t_1)$ as $\Psi_{\boldsymbol{U}}(t) = \tilde{\Psi}_{\boldsymbol{U}}(t) \odot \tilde{\Psi}_{\boldsymbol{U}}(t)^\intercal$ where $\tilde{\Psi}_{\boldsymbol{U}}(t) = \big((\mathbb{1} - \boldsymbol{B}_t)^\intercal \boldsymbol{B}_t\big)$.

The matrices $\tilde{\Psi}_{\boldsymbol{U}}(t_1), \tilde{\Psi}_{\boldsymbol{U}}(t_2), \tilde{\Psi}_{\boldsymbol{U}}(t_3)$ are computed below:

Note that $\|\Psi_{\boldsymbol{U}}(t)\| = 0$ if $\tilde{\Psi}_{\boldsymbol{U}}(t)$ is a triangular matrix with a zero diagonal. Thus it suffices to evaluate matrix $\tilde{\Psi}_{\boldsymbol{U}}(t)$ for $t \in \{t_1, t_2, t_3\}$.

$$\tilde{\Psi}_{\boldsymbol{U}}(t_1) = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^{\intercal}}_{(\mathbb{1} - \boldsymbol{B}_{t_1})^{\intercal}} \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}}_{\boldsymbol{B}_{t_1}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 4 & 3 & 2 & 2 & 0 & 0 & 0 \\ 4 & 3 & 2 & 2 & 0 & 0 & 0 \\ 4 & 3 & 2 & 2 & 0 & 0 & 0 \end{bmatrix}$$

$$\tilde{\Psi}_{\boldsymbol{U}}(t_2) = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}^{\intercal}}_{(\mathbb{1} - \boldsymbol{B}_{t_2})^{\intercal}} \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}_{\boldsymbol{B}_{t_2}} = \begin{bmatrix} 0 & 1 & 1 & 0 & 4 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 & 2 & 0 \\ 0 & 1 & 1 & 0 & 4 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 4 & 3 & 0 \end{bmatrix}$$

$$\tilde{\Psi}_{\boldsymbol{U}}(t_3) = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}^{\intercal}}_{(\mathbb{1} - \boldsymbol{B}_{t_3})^{\intercal}} \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\boldsymbol{B}_{t_3}} = \begin{bmatrix} 0 & 0 & 1 & 2 & 0 & 1 & 4 \\ 0 & 0 & 1 & 2 & 0 & 1 & 4 \\ 0 & 0 & 0 & 1 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 1 & 2 & 0 & 1 & 4 \\ 0 & 0 & 0 & 1 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

It is easy to see that in $\tilde{\Psi}_{\boldsymbol{U}}(t_1) \odot \tilde{\Psi}_{\boldsymbol{U}}(t_1)^{\intercal}$ is equal to a matrix of zeros. Indeed, the matrix $\tilde{\Psi}_{\boldsymbol{U}}(t_1)$ is triangular with a zero diagonal. Thus, when we perform the element wise multiplication of $\tilde{\Psi}_{\boldsymbol{U}}(t_1)$ and its transpose, at least one of the elements of the multiplication will be zero. The same occurs for matrices $\tilde{\Psi}_{\boldsymbol{U}}(t_2)$ and $\tilde{\Psi}_{\boldsymbol{U}}(t_3)$.

3.12.2. A Case of Choice Incentives for Ordered Monotonicity

The can summarize the above incentive structure the binary incentive matrix given below:

$$\boldsymbol{L} = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.12.1}$$

We use this first example to describe the machinery that translates choice incentives into monotonicity conditions and identification results. We adopt a more parsimonious approach in the subsequent examples.

Choice rule (3.6.2) converts the Incentive Matrix (3.12.1) into choice restrictions that determine the model response matrix $\boldsymbol{R}$. These choice restrictions are displayed in Table 3.12.1. Choice restrictions in Table 3.12.1 are in turn used to eliminate the response-types that are not economically justifiable.

Each counterfactual choice $T(z)$ of the response vector $\boldsymbol{S} = [T(z_1), T(z_2), T(z_3), T(z_4)]'$ takes up to three values in $\{t_1, t_2, t_3\}$. Thus, there are $3^4 = 81$ potential response types. The combination of all choice restrictions of Table 3.12.1 eliminate a total of 74 out of the 81 potential response-types. Response matrix $\boldsymbol{R}$ in (3.12.2) displays the resulting seven response-types that survive the elimination process.

$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 & \boldsymbol{s}_8 \end{matrix} \\ \begin{bmatrix} t_1 & t_1 & t_1 & t_1 & t_1 & t_2 & t_2 & t_3 \\ t_1 & t_1 & t_2 & t_2 & t_3 & t_2 & t_2 & t_3 \\ t_1 & t_1 & t_2 & t_2 & t_3 & t_2 & t_2 & t_3 \\ t_1 & t_3 & t_2 & t_3 & t_3 & t_2 & t_1 & t_3 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.12.2}$$

We use equations (3.3.12)–(3.3.13) to evaluate the causal parameters identified by response matrix (3.7.3). The response-matrix (3.7.3) enables the identification of eight response-type probabilities:

Table 3.12.1: Choice Restrictions generated by Incentive Matrix (3.12.1)

1	$T_i(z_1) = t_1$	$\Rightarrow$	$\varnothing$
2	$T_i(z_2) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_3) \notin \{t_2, t_3\}$ and $T_i(z_4) \neq t_2$
3	$T_i(z_3) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_2) \notin \{t_2, t_3\}$ and $T_i(z_4) \neq t_2$
4	$T_i(z_4) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_2) \notin \{t_2, t_3\}$ and $T_i(z_3) \notin \{t_2, t_3\}$
5	$T_i(z_1) = t_2$	$\Rightarrow$	$T_i(z_2) \notin \{t_1, t_3\}$ and $T_i(z_3) \notin \{t_1, t_3\}$ and $T_i(z_4) \neq t_1$
6	$T_i(z_2) = t_2$	$\Rightarrow$	$T_i(z_1) \neq t_3$ and $T_i(z_3) \notin \{t_1, t_3\}$ and $T_i(z_4) \neq t_1$
7	$T_i(z_3) = t_2$	$\Rightarrow$	$T_i(z_1) \neq t_3$ and $T_i(z_2) \notin \{t_1, t_3\}$ and $T_i(z_4) \neq t_1$
8	$T_i(z_4) = t_2$	$\Rightarrow$	$T_i(z_1) \neq t_3$ and $T_i(z_2) \notin \{t_1, t_3\}$ and $T_i(z_3) \notin \{t_1, t_3\}$
9	$T_i(z_1) = t_3$	$\Rightarrow$	$T_i(z_2) \notin \{t_1, t_2\}$ and $T_i(z_3) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
10	$T_i(z_2) = t_3$	$\Rightarrow$	$T_i(z_1) \neq t_2$ and $T_i(z_3) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
11	$T_i(z_3) = t_3$	$\Rightarrow$	$T_i(z_1) \neq t_2$ and $T_i(z_2) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
12	$T_i(z_4) = t_3$	$\Rightarrow$	$\varnothing$

This table presents all the choice restrictions generated by applying the choice rule (3.6.2) to each of the combination of choices $(t, t') \in \{t_1, t_2, t_3\}$ and instrumental values $(z, z') \in \{z_1, z_2, z_3, z_4\}$ of the incentive matrix (3.12.1).

Point Identified	$P(\boldsymbol{S} = \boldsymbol{s}_1), P(\boldsymbol{S} = \boldsymbol{s}_2), P(\boldsymbol{S} = \boldsymbol{s}_5), P(\boldsymbol{S} = \boldsymbol{s}_8)$.
Partially Identified	$P(\boldsymbol{S} \in \{\boldsymbol{s}_3, \boldsymbol{s}_4\}), P(\boldsymbol{S} \in \{\boldsymbol{s}_3, \boldsymbol{s}_6\}), P(\boldsymbol{S} \in \{\boldsymbol{s}_4, \boldsymbol{s}_7\}), P(\boldsymbol{S} \in \{\boldsymbol{s}_6, \boldsymbol{s}_7\})$.

as well as the following counterfactual outcomes.

Always-takers	$\mathbb{E}(Y(t_1)\|\boldsymbol{S} = \boldsymbol{s}_1)$	-	$\mathbb{E}(Y(t_3)\|\boldsymbol{S} = \boldsymbol{s}_8)$
Switchers	$\mathbb{E}(Y(t_1)\|\boldsymbol{S} = \boldsymbol{s}_2)$	-	$\mathbb{E}(Y(t_3)\|\boldsymbol{S} = \boldsymbol{s}_5)$
Partially Identified	$\mathbb{E}(Y(t_1)\|\boldsymbol{S} \in \{\boldsymbol{s}_3, \boldsymbol{s}_4, \boldsymbol{s}_5\})$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S} \in \{\boldsymbol{s}_4, \boldsymbol{s}_7\})$	$\mathbb{E}(Y(t_3)\|\boldsymbol{S} \in \{\boldsymbol{s}_2, \boldsymbol{s}_4, \boldsymbol{s}_7\})$
		$\mathbb{E}(Y(t_2)\|\boldsymbol{S} \in \{\boldsymbol{s}_6, \boldsymbol{s}_7\})$	
		$\mathbb{E}(Y(t_2)\|\boldsymbol{S} \in \{\boldsymbol{s}_3, \boldsymbol{s}_6\})$	
		$\mathbb{E}(Y(t_2)\|\boldsymbol{S} \in \{\boldsymbol{s}_3, \boldsymbol{s}_4\})$	

The identification results above state that only four out of nine response-type probabilities are point-identified. Most of the counterfactual outcomes are partially identified. Only four counterfactual outcome means are point-identified, none of these for choice t_2. In contrast, the unordered response matrix (3.7.2) secures the point-identification of all response-type probabilities and most of the counterfactual outcome means.

3.12.3. MM under the Double Randomization Design

We consider the emergence of MM in a "Double Randomization" design in which two vouchers are randomly assigned to the same sample of prospective students. The first voucher offers a tuition discount that applies to a natural science major. The second one applies to social science majors. We can divide the students into four groups:

1. Group z_1 does not receive any voucher.
2. Group z_2 receives only the social sciences voucher (t_3).
3. Group z_3 receives only the natural sciences voucher (t_2).
4. Group z_4 receives both the social sciences and natural sciences voucher.

Assuming the social sciences and natural sciences vouchers are of the same amount and that students cannot double major (so that they can only apply one voucher at a time), the IV design described above can be summarized by the incentive matrix in (3.12.3).

$$\boldsymbol{L} = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.12.3}$$

Applying the Choice Rule (3.6.2) from above generates the choice restrictions of Table 3.12.2. These in turn generate the response matrix $\boldsymbol{R}$ in (3.12.4).

$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 & \boldsymbol{s}_8 & \boldsymbol{s}_9 \end{matrix} \\ \begin{bmatrix} t_1 & t_1 & t_1 & t_1 & t_1 & t_2 & t_2 & t_3 & t_3 \\ t_1 & t_1 & t_3 & t_3 & t_3 & t_2 & t_3 & t_3 & t_3 \\ t_1 & t_2 & t_1 & t_2 & t_2 & t_2 & t_2 & t_2 & t_3 \\ t_1 & t_2 & t_3 & t_2 & t_3 & t_2 & t_2 & t_3 & t_3 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.12.4}$$

Table 3.12.2: Choice Restrictions generated by Incentive Matrix (3.7.4)

1	$T_i(z_1) = t_1$	$\Rightarrow$	$T_i(z_2) \neq t_2$ and $T_i(z_3) \neq t_3$
2	$T_i(z_2) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_3) \neq t_3$ and $T_i(z_4) \neq t_3$
3	$T_i(z_3) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_2) \neq t_2$ and $T_i(z_4) \neq t_2$
4	$T_i(z_4) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_2) \notin \{t_2, t_3\}$ and $T_i(z_3) \notin \{t_2, t_3\}$
5	$T_i(z_1) = t_2$	$\Rightarrow$	$T_i(z_2) \neq t_1$ and $T_i(z_3) \notin \{t_1, t_3\}$ and $T_i(z_4) \notin \{t_1, t_3\}$
6	$T_i(z_2) = t_2$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_3\}$ and $T_i(z_3) \notin \{t_1, t_3\}$ and $T_i(z_4) \notin \{t_1, t_3\}$
7	$T_i(z_3) = t_2$	$\Rightarrow$	$T_i(z_4) \neq t_1$
8	$T_i(z_4) = t_2$	$\Rightarrow$	$T_i(z_1) \neq t_3$ and $T_i(z_3) \notin \{t_1, t_3\}$
9	$T_i(z_1) = t_3$	$\Rightarrow$	$T_i(z_2) \notin \{t_1, t_2\}$ and $T_i(z_3) \neq t_1$ and $T_i(z_4) \notin \{t_1, t_2\}$
10	$T_i(z_2) = t_3$	$\Rightarrow$	$T_i(z_4) \neq t_1$
11	$T_i(z_3) = t_3$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_2\}$ and $T_i(z_2) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
12	$T_i(z_4) = t_3$	$\Rightarrow$	$T_i(z_1) \neq t_2$ and $T_i(z_2) \notin \{t_1, t_2\}$

This table presents all the choice restrictions generated by applying the choice rule (3.6.2) to each of the combination of choices $(t, t') \in \{t_1, t_2, t_3\}$ and instrumental values $(z, z') \in \{z_1, z_2, z_3, z_4\}$ of the incentive matrix (3.7.4).

Applying equations (3.3.12)–(3.3.13) to response-matrix (3.12.4) gives that all response-type probabilities are identified, $P(\boldsymbol{S} = \boldsymbol{s}_j)$; $j = 1, ..., 9$, as well as the following counterfactual outcomes:

Always-takers	$\mathbb{E}(Y(t_0)\|\boldsymbol{S} = \boldsymbol{s}_1)$	$\mathbb{E}(Y(t_1)\|\boldsymbol{S} = \boldsymbol{s}_6)$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S} = \boldsymbol{s}_9)$
Switchers	$\mathbb{E}(Y(t_0)\|\boldsymbol{S} = \boldsymbol{s}_2)$	$\mathbb{E}(Y(t_1)\|\boldsymbol{S} = \boldsymbol{s}_7)$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S} = \boldsymbol{s}_8)$
	$\mathbb{E}(Y(t_0)\|\boldsymbol{S} = \boldsymbol{s}_3)$		
Partially Identified	$\mathbb{E}(Y(t_0)\|\boldsymbol{S} \in \{\boldsymbol{s}_4, \boldsymbol{s}_5\})$	$\mathbb{E}(Y(t_1)\|\boldsymbol{S} \in \{\boldsymbol{s}_2, \boldsymbol{s}_4\})$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S} \in \{\boldsymbol{s}_3, \boldsymbol{s}_5\})$
		$\mathbb{E}(Y(t_1)\|\boldsymbol{S} \in \{\boldsymbol{s}_5, \boldsymbol{s}_8\})$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S} \in \{\boldsymbol{s}_4, \boldsymbol{s}_7\})$

3.12.4. MM under the Extensive Margin Compliers Only (EMCO) Design

We revisit the incentive design described in (3.7.5), presented again in $\boldsymbol{L}$ (3.12.5) below

$$\boldsymbol{L} = \begin{array}{c} \begin{matrix} t_0 & t_1 & t_2 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.12.5}$$

Applying the Choice Rule (3.6.2) to the incentive design summarized in (3.12.5) we generate the following choice restrictions. These choice restrictions will in turn be used to eliminate response types, i.e restrict supp($\boldsymbol{S}$).

Table 3.12.3: Choice Restrictions generated by Incentive Matrix (3.12.5)

1	$T_i(z_1) = t_1$	$\Rightarrow$	$T_i(z_2) \notin \{t_2, t_3\}$ and $T_i(z_4) \notin \{t_2, t_3\}$
2	$T_i(z_2) = t_1$	$\Rightarrow$	$\varnothing$
3	$T_i(z_3) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_2) \notin \{t_2, t_3\}$ and $T_i(z_4) \notin \{t_2, t_3\}$
4	$T_i(z_4) = t_1$	$\Rightarrow$	$T_i(z_1) \notin \{t_2, t_3\}$ and $T_i(z_2) \notin \{t_2, t_3\}$
5	$T_i(z_1) = t_2$	$\Rightarrow$	$T_i(z_2) \neq t_3$ and $T_i(z_3) \notin \{t_1, t_3\}$ and $T_i(z_4) \notin \{t_1, t_3\}$
6	$T_i(z_2) = t_2$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_3\}$ and $T_i(z_3) \notin \{t_1, t_3\}$ and $T_i(z_4) \notin \{t_1, t_3\}$
7	$T_i(z_3) = t_2$	$\Rightarrow$	$T_i(z_1) \neq t_3$ and $T_i(z_2) \neq t_3$ and $T_i(z_4) \neq t_3$
8	$T_i(z_4) = t_2$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_3\}$ and $T_i(z_2) \neq t_3$ and $T_i(z_3) \notin \{t_1, t_3\}$
9	$T_i(z_1) = t_3$	$\Rightarrow$	$T_i(z_2) \neq t_2$ and $T_i(z_3) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
10	$T_i(z_2) = t_3$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_2\}$ and $T_i(z_3) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
11	$T_i(z_3) = t_3$	$\Rightarrow$	$T_i(z_1) \neq t_2$ and $T_i(z_2) \neq t_2$ and $T_i(z_4) \neq t_2$
12	$T_i(z_4) = t_3$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_2\}$ and $T_i(z_2) \neq t_2$ and $T_i(z_3) \notin \{t_1, t_2\}$

This table presents all the choice restrictions generated by applying the choice rule (3.6.2) to each of the combination of choices $(t, t') \in \{t_1, t_2, t_3\}$ and instrumental values $(z, z') \in \{z_1, z_2, z_3, z_4\}$ of the incentive matrix (3.12.5).

After exhausting the choice restrictions in Table 3.12.3 we are left with 7 out of a possible 81 response types. These response types are consolidated and displayed in the response matrix $\boldsymbol{R}$ (3.12.6) below.

$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 \end{matrix} \\ \begin{bmatrix} t_1 & t_1 & t_1 & t_2 & t_2 & t_3 & t_3 \\ t_1 & t_1 & t_1 & t_1 & t_2 & t_1 & t_3 \\ t_1 & t_2 & t_3 & t_2 & t_2 & t_3 & t_3 \\ t_1 & t_1 & t_1 & t_2 & t_2 & t_3 & t_3 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.12.6}$$

We can apply the identification results of Heckman and Pinto (2018) to response-matrix (3.12.6) in order to identify all the response-type probabilities $P(\boldsymbol{S} = \boldsymbol{s}_j)$; $j = 1, ..., 7$ as well as the following counterfactual outcomes:

Always-takers	$\mathbb{E}(Y(t_1)\|S = s_1)$	$\mathbb{E}(Y(t_2)\|S = s_5)$	$\mathbb{E}(Y(t_3)\|S = s_7)$
Switchers		$\mathbb{E}(Y(t_2)\|S = s_2)$	$\mathbb{E}(Y(t_3)\|S = s_3)$
		$\mathbb{E}(Y(t_2)\|S = s_4)$	$\mathbb{E}(Y(t_3)\|S = s_6)$
Partially Identified	$\mathbb{E}(Y(t_1)\|S \in \{s_2, s_3\})$		
	$\mathbb{E}(Y(t_1)\|S \in \{s_4, s_6\})$		

3.12.5. MM under Orthogonal Array Design

We additionally examine an IV choice model based on the popular orthogonal array experimental design. Orthogonal arrays are a widely popular experimental design developed by CD Rao (Rao, 1946a,b, 1947, 1949). Orthogonal arrays are widely used in Agricultural and Industrial sciences to determine the optimum mix of treatments that maximize production yield. The method is based on the random assignment of a combinatorial arrangements of treatments for each randomization arm. We adapt this setup to an instrumental variable setting by exogenously providing incentives for one or more treatments instead of directly assigning agents to treatment arms. Below, we will see that this incentive structure allows for a broad range of identification results.

Formally, a binary orthogonal array is a matrix of zeros and ones such that any two-column submatrix displays all possible combinations of zeros and ones. In other words, the tuples

$$\{(0,0),(0,1),(1,0),(1,1)\}$$

are all rows in any two-column submatrix. An orthogonal array incentive design if its associated incentive matrix is a binary orthogonal array. The incentive matrix in (3.7.10)

displays an example of an orthogonal array incentive design. In context of the college choice example, we can rationalize the orthogonal array incentive design (3.7.10) with the following research design:

1. Group z_1 receives a cash voucher if they choose to major in the natural sciences (t_2) or the social sciences (t_3).
2. Group z_2 receives no cash voucher.
3. Group z_3 receives a cash voucher if they do not go to college (t_1) or if they major in the natural sciences (t_2).
4. Group z_3 receives a cash voucher if they do not go to college (t_1) or if they major in the social sciences (t_3).

Table 3.12.4 displays the choice restrictions generated by applying the Choice Rule (3.6.2) to the orthogonal array incentive design (3.7.10). After using these choice restrictions to eliminate response types, we are left with nine total response types summarized in the response matrix (3.12.7).

Table 3.12.4: Choice Restrictions generated by Incentive Matrix (3.7.10)

1	$T_i(z_1) = t_1$	$\Rightarrow$	$T_i(z_2) \notin \{t_2, t_3\}$ and $T_i(z_3) \notin \{t_2, t_3\}$ and $T_i(z_4) \notin \{t_2, t_3\}$
2	$T_i(z_2) = t_1$	$\Rightarrow$	$T_i(z_3) \notin \{t_2, t_3\}$ and $T_i(z_4) \notin \{t_2, t_3\}$
3	$T_i(z_3) = t_1$	$\Rightarrow$	$T_i(z_2) \neq t_2$ and $T_i(z_4) \neq t_2$
4	$T_i(z_4) = t_1$	$\Rightarrow$	$T_i(z_2) \neq t_3$ and $T_i(z_3) \neq t_3$
5	$T_i(z_1) = t_2$	$\Rightarrow$	$T_i(z_2) \neq t_3$ and $T_i(z_3) \neq t_3$
6	$T_i(z_2) = t_2$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_3\}$ and $T_i(z_3) \notin \{t_1, t_3\}$
7	$T_i(z_3) = t_2$	$\Rightarrow$	$T_i(z_1) \neq t_1$ and $T_i(z_2) \neq t_1$
8	$T_i(z_4) = t_2$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_3\}$ and $T_i(z_2) \notin \{t_1, t_3\}$ and $T_i(z_3) \notin \{t_1, t_3\}$
9	$T_i(z_1) = t_3$	$\Rightarrow$	$T_i(z_2) \neq t_2$ and $T_i(z_4) \neq t_2$
10	$T_i(z_2) = t_3$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
11	$T_i(z_3) = t_3$	$\Rightarrow$	$T_i(z_1) \notin \{t_1, t_2\}$ and $T_i(z_2) \notin \{t_1, t_2\}$ and $T_i(z_4) \notin \{t_1, t_2\}$
12	$T_i(z_4) = t_3$	$\Rightarrow$	$T_i(z_1) \neq t_1$ and $T_i(z_2) \neq t_1$

This table presents all the choice restrictions generated by applying the choice rule (3.6.2) to each of the combination of choices $(t, t') \in \{t_1, t_2, t_3\}$ and instrumental values $(z, z') \in \{z_1, z_2, z_3, z_4\}$ of the incentive matrix (3.7.10).

$$\boldsymbol{R} = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & \boldsymbol{s}_4 & \boldsymbol{s}_5 & \boldsymbol{s}_6 & \boldsymbol{s}_7 & \boldsymbol{s}_8 & \boldsymbol{s}_9 \end{matrix} \\ \begin{bmatrix} t_1 & t_2 & t_2 & t_2 & t_2 & t_3 & t_3 & t_3 & t_3 \\ t_1 & t_1 & t_2 & t_2 & t_2 & t_1 & t_3 & t_3 & t_3 \\ t_1 & t_1 & t_2 & t_2 & t_2 & t_1 & t_1 & t_2 & t_3 \\ t_1 & t_1 & t_1 & t_2 & t_3 & t_1 & t_3 & t_3 & t_3 \end{bmatrix} \end{array} \begin{matrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{matrix} \tag{3.12.7}$$

This response matrix satisfies neither unordered nor ordered monotonicity. When the instrument switches from z_1 to z_4, agents in response type $\boldsymbol{s}_3$ move from treatment t_2 to treatment t_3 while agents in response type $\boldsymbol{s}_6$ move away from t_3 and towards t_1. This represents a violation of ordered monotonicity and also prevents t_3 from being ordered the highest or lowest in any ordering on $\mathcal{T}$ that would satisfy ordered monotonicity.[1] Similarly we can see a switch from z_3 to z_4 induces agents in response type $\boldsymbol{s}_3$ to move from treatment t_2 to treatment t_1 while inducing agents in response type $\boldsymbol{s}_7$ to move away from treatment t_1 and towards treatment t_3. This again represents a violation of unordered monotonicity and precents t_1 from being ordered either the highest or the lowest in any ordering $\mathcal{T}$ that would satisfy ordered monotonicity. Since all orderings on $\mathcal{T} = \{t_1, t_2, t_3\}$ must have either t_1 or t_3 as the largest or smallest element, this means there is no ordering on $\mathcal{T}$ that satisfies ordered monotonicity.

Despite this, we can once again use Theorem 3.5.1 to verify that this matrix does indeed satisfy MM. Thus we can still use 2SLS type estimands to recover interpretable causal parameters as defined in (3.5.3). Moreover, by applying (3.3.12)–(3.3.13) we can see that all response types probabilities $\mathbb{P}(\boldsymbol{S} = \boldsymbol{s}_j)$, $j = 1, \ldots, 9$ are identified. Additionally, using (3.3.12)–(3.3.13) we obtain that the following counterfactual outcomes are identified

[1] If t_3 is ranked highest a movement away from t_3 represents moving towards a lower treatment while a towards t_3 represents moving towards a higher treatment. Vice versa, if t_3 is ranked lowest a movement towards t_3 represents moving towards a lower treatment while a movement away from t_3 represents moving towards a higher treatment.

Always-takers	$\mathbb{E}(Y(t_1)\|\boldsymbol{S}=\boldsymbol{s}_1)$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S}=\boldsymbol{s}_4)$	$\mathbb{E}(Y(t_3)\|\boldsymbol{S}=\boldsymbol{s}_9)$
Switchers	$\mathbb{E}(Y(t_1)\|\boldsymbol{S}=\boldsymbol{s}_3)$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S}=\boldsymbol{s}_2)$	$\mathbb{E}(Y(t_3)\|\boldsymbol{S}=\boldsymbol{s}_5)$
	$\mathbb{E}(Y(t_1)\|\boldsymbol{S}=\boldsymbol{s}_7)$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S}=\boldsymbol{s}_8)$	$\mathbb{E}(Y(t_3)\|\boldsymbol{S}=\boldsymbol{s}_6)$
Partially Identified	$\mathbb{E}(Y(t_1)\|\boldsymbol{S}\in\{\boldsymbol{s}_2,\boldsymbol{s}_6\})$	$\mathbb{E}(Y(t_2)\|\boldsymbol{S}\in\{\boldsymbol{s}_3,\boldsymbol{s}_5\})$	$\mathbb{E}(Y(t_3)\|\boldsymbol{S}\in\{\boldsymbol{s}_7,\boldsymbol{s}_8\})$

Bibliography

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics 21*(4), 489–519.

Aliprantis, D. (2012). Compulsory schooling laws and educational attainment. *Journal of Educational and Behavioral Statistics 37*, 316–338.

Anderson, T. W. and H. Rubin (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics 20*(1), 46 – 63.

Andrews, D. W., M. Moreira, and J. H. Stock (2004, August). Optimal invariant similar tests for instrumental variables regression. (299).

Andrews, D. W. and J. H. Stock (2007). Testing with many weak instruments. *Journal of Econometrics 138*(1), 24–46. 50th Anniversary Econometric Institute.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica 74*(3), 715–752.

Andrews, I. (2016). Conditional linear combination tests for weakly identified models. *Econometrica 84*(6), 2155–2182.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000, July). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies 67*(3), 499–527.

Angrist, J. D. and G. W. Imbens (1995, June). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association 90*(430), 431–442.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics 14*(1), 57–67.

Angrist, J. D., G. W. Imbens, and D. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Barua, R. and K. Lang (2016). School entry, educational attainment, and quarter of birth: A cautionary tale of a local average treatment effect. *Journal of Human Capital 10*(3), 347–376.

Bauer, B. and M. Kohler (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics 47*(4), 2261 – 2285.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica 62*(3), 657–681.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012a). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012b). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521 – 547.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018). High-dimensional econometrics and regularized gmm.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–366. High Dimensional Problems in Econometrics.

Belloni, A., V. Chernozhukov, and C. Hansen (2013, 11). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705 – 1732.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association 90*(430), 443–450.

Bradic, J., S. Wager, and Y. Zhu (2019, May). Sparsity Double Robust Inference of Average Treatment Effects. Papers 1905.00744, arXiv.org.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017). Beyond late with a discrete instrument. *Journal of Political Economy 125*(4), 985–1039.

Buchinsky, M. and R. Pinto (2021). Economics of monotonicity conditions. *NBER Working Paper*.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Cameron, S. V. and J. J. Heckman (1998, April). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political Economy 106*(2), 262–333.

Carneiro, P., K. Hansen, and J. J. Heckman (2003, May). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review 44*(2), 361–422.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics 155*(2), 138–154.

Celentano, M., A. Montanari, and Y. Wu (2020, 09–12 Jul). The estimation error of general first order methods. In J. Abernethy and S. Agarwal (Eds.), *Proceedings of Thirty Third Conference on Learning Theory*, Volume 125 of *Proceedings of Machine Learning Research*, pp. 1078–1141. PMLR.

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

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012). Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory 28*(1), 42–86.

Chatterjee, S. (2006). A generalization of the Lindeberg principle. *The Annals of Probability 34*(6), 2061 – 2076.

Chatterjee, S. (2010). A new approach to strong embeddings.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics* (1 ed.), Volume 6B, Chapter 76, pp. 5549–5632. Elsevier.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics 41*(6), 2786 – 2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability 45*(4), 2309–2352.

Chernozhukov, V., W. K. Newey, and R. Singh (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica 90*(3), 967–1027.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021). On cross-validated Lasso in high dimensions. *The Annals of Statistics 49*(3), 1300 – 1317.

Chetverikov, D. and J. R.-V. Sørensen (2021). Analytic and bootstrap-after-cross-validation methods for selecting penalty parameters of high-dimensional m-estimators. *ArXiv NA*, 1–50.

Crudu, F., G. Mellace, and Z. Sándor (2021). Inference in instrumental variable models with heteroskedasticity and many instruments. *Econometric Theory 37*(2), 281–310.

Cunha, F., J. J. Heckman, and S. Navarro (2007, November). The identification and economic content of ordered choice models with stochastic cutoffs. *International Economic Review 48*(4), 1273–1309.

Dahl, C. M., M. Huber, and G. Mellace (2017). It's never too late: A new look at local average treatment effects with or without defiers. *Discussion Papers on Business and Economics*.

De Boor, C. (2001). *A practical guide to splines; rev. ed.* Applied mathematical sciences. Berlin: Springer.

de Chaisemartin, C. (2017). Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics 8*(2), 367–396.

der Vaart, A. V. and J. Wellner (1996). *Weak Convergence and Empirical Processes* (1 ed.). Springer Series in Statistics. Springer, New York, NY.

Derenoncourt, E. (2022, February). Can you move to opportunity? evidence from the great migration. *American Economic Review 112*(2), 369–408.

Dobbie, W., J. Goldin, and C. S. Yang (2018, February). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review 108*(2), 201–40.

Dudley, R. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis 1*(3), 290–330.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics 40*(1), 313–327.

Feller, A., T. Grindal, L. Miratrix, and L. C. Page (2016). Compared to what? variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics 10*(3), 1245–1285.

Friedman, J., R. Tibshirani, and T. Hastie (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics 139*(1), 35–75. Endogeneity, instruments and identification.

Gautier, E. and C. Rose (2021). High-dimensional instrumental variables regression and confidence sets.

Gautier, E. and C. Rose (2022). Fast, robust inference for linear instrumental variables models using self-normalized moments.

Gilchrist, D. S. and E. G. Sands (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy 124*(5), 1339–1382.

Giné, E. and V. Koltchinskii (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability 34*(3), 1143 – 1216.

Goldberger, A. S. (1972, November). Structural equation methods in the social sciences. *Econometrica 40*(6), 979–1001.

Götze, F., A. Naumov, V. Spokoiny, and V. Ulyanov (2019). Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli 25*(4A), 2538 – 2563.

Gotze, F., H. Sambale, and A. Sinulis (2021). Concentration inequalities for polynomials in alpha-sub-exponential random variables. *Electronic Journal of Probability 26*(none), 1 – 22.

Han, C. and P. C. B. Phillips (2006). Gmm with many moment conditions. *Econometrica 74*(1), 147–192.

Harrell, F. E. (2015). *Regression Modeling Strategies*. Spring Series in Statistics. Springer Cham.

Heckman, J. and R. Pinto (2018). Unordered monotonicity. *Econometrica 86*, 1–35.

Heckman, J. J. (1976, December). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement 5*(4), 475–492.

Heckman, J. J. (1979, January). Sample selection bias as a specification error. *Econometrica 47*(1), 153–162.

Heckman, J. J. and R. Pinto (2014). Causal analysis after haavelmo. *Econometric Theory*, 1–37.

Heckman, J. J. and R. Robb (1985, October-November). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics 30*(1–2), 239–267.

Heckman, J. J. and S. Urzúa (2010, May). Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics 156*(1), 27–37.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J., S. Urzúa, and E. J. Vytlacil (2008). Instrumental variables in models with multiple outcomes: The general unordered case. *Annales d'Economie et de Statistique 91–92*, 151–174.

Heckman, J. J. and E. J. Vytlacil (1999, April). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*(8), 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007a). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, Chapter 70, pp. 4779–4874. Amsterdam: Elsevier B. V.

Heckman, J. J. and E. J. Vytlacil (2007b). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs, and to forecast their effects in new environments. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, Chapter 71, pp. 4875–5143. Amsterdam: Elsevier B. V.

Hlavac, M. (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Social Policy Institute. R package version 5.2.3.

Holland, P. W. (1986, December). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Horn, R. and C. Johnson (2012). *Matrix Analysis*. Cambridge University Press.

Huber, M., L. Laffers, and G. Mellace (2017). Sharp iv bounds on average treatment effects on the treated and other populations under endogeneity and noncompliance. *Journal of Applied Econometrics 32*(1), 56–79.

Huber, M. and G. Mellace (2012). Relaxing monotonicity in the identification of local average treatment effects. Unpublished manuscript, Department of Economics, University of St. Gallen.

Huber, M. and G. Mellace (2015). Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics 97*, 398–411.

Hull, P. (2018). Isolateing: Identifying counterfactual-specific treatment effects with cross-stratum comparisons. *Umpublished Manuscript*.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and D. B. Rubin (1997, October). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies 64*(4), 555–574.

Jou, A. and T. Morgan (2023). Do relief programs compensate affected populations? evidence from the great depression and the new deal. *Working Paper*.

Kamat, V. (2021). Identifying the effects of a program offer with an application to head start. *Umpublished Manuscrip*.

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics 131*(3), 1057–1111.

Kleibergen, F. (2002, 02). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica 70*, 1781–1803.

Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica 73*(4), 1103–1123.

Klein, T. J. (2010). Heterogeneous treatment effects: Instrumental variables without monotonicity? *Journal of Econometrics 155*, 99–116.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.

Kline, P. and C. R. Walters (2016, 07). Evaluating public programs with close substitutes: The case of head start. *The Quarterly Journal of Economics 131*(4), 1795–1848.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022, October). Valid t-ratio inference for iv. *American Economic Review 112*(10), 3260–90.

Lee, S., R. Okui, and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics 32*(7), 1207–1225.

Lee, S. and B. Salanié (2018). Identifying effects of multivalued treatments. *Econometrica 86*, 1939–1963.

Lim, D., W. Wang, and Y. Zhang (2022). A conditional linear combination test with many weak instruments.

Lindeberg, J. W. (1922). Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift 15*, 211–225.

Maestas, N., K. J. Mullen, and A. Strand (2013, August). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *American Economic Review 103*(5), 1797–1829.

Matsushita, Y. and T. Otsu (2022). A jackknife lagrange multiplier test with many weak instruments. *Econometric Theory*, 1–24.

Mikusheva, A. (2023). Many weak instruments in time series econometrics. *Working Paper*.

Mikusheva, A. and L. Sun (2021, 12). Inference with many weak instruments. *The Review of Economic Studies 89*(5), 2663–2686.

Mogstad, M., A. Santos, and A. Torgovitsky (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica 86*(5), 1589–1619.

Mogstad, M. and A. Torgovitsky (2018). Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics 10*(1), 577–613.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review 111*(11), 3663–3698.

Moreira, M. (2009, 10). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics 152*, 131–140.

Moreira, M. J. (2001). *Tests with correct size when instruments can be arbitrarily weak*. Citeseer.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica 71*(4), 1027–1048.

Mountjoy, J. (2021). Community colleges and upward mobility. *Unpublished Manuscript*.

Nazarov, F. (2003). *On the Maximal Perimeter of a Convex Set in R^n with Respect to a Gaussian Measure*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Nelson, C. R. and R. Startz (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica 58*(4), 967–976.

Newey, W. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Newey, W. K. and F. Windmeijer (2009). Generalized method of moments with many weak moment conditions. *Econometrica 77*(3), 687–719.

Paravisini, D., V. Rappoport, P. Schnabl, and D. Wolfenzon (2014, 09). Dissecting the Effect of Credit Supply on Trade: Evidence from Matched Credit-Export Data. *The Review of Economic Studies 82*(1), 333–359.

Perperoglou, A., W. Sauerbrei, M. Abrahamowicz, and M. Schmid (2019). A review of spline function procedures in r. *BMC medical research methodology 19*(1), 1–16.

Petersen, K. B. and M. S. Pedersen (2012, nov). The matrix cookbook. Version 20121115.

Pinto, R. (2021). Beyond intention to treat: Using the incentives in moving to opportunity to identify neighborhood effects. *NBER Working Paper*.

Pinto, R. and J. Heckman (2021). The econometric model for causal policy analysis. *Manuscript Prepared for Annual Review of Economics (2021) edition*.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Pouzo, D. (2015). Bootstrap consistency for quadratic forms of sample averages with increasing dimension. *Electronic Journal of Statistics 9*(2), 3046 – 3097.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. (1946a). Difference sets and combinatorial arrangements derivable from finite geometries. *Proceedings of the Indian National Science Academy 12*(3), 123–135.

Rao, C. R. (1946b). Hypercubes of strength 'd' leading to confounded designs in factorial experiments. *Bulletin of the Calcutta Mathematical Society 38*, 67–78.

Rao, C. R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *Supplement to the Journal of the Royal Statistical Society 9*(1), 128–139.

Rao, C. R. (1949). On a class of arrangements. *Proceedings of the Edinburgh Mathematical Society 8*(3), 119–125.

Rose, E. K. and Y. Shem-Tov (2021). On recoding ordered treatments as binary indicators. *Umpublished Manuscript*.

Rubin, D. B. (1974a). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. B. (1974b, October). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Rubin, D. B. (1978a). Bayesian inference for causal effects. *The Annals of Statistics 6*(1), 34–58.

Rubin, D. B. (1978b, January). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics 6*(1), 34–58.

Rudelson, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal 164*, 60–72.

Ryser, H. (1957). Combinatorial properties of matrices of zeros and ones. *Canadian Journal of Mathematics 9*, 371–377.

Sambale, H. (2022). Some notes on concentration for α-subexponential random variables.

Sampat, B. and H. L. Williams (2019, January). How do patents affect follow-on innovation? evidence from the human genome. *American Economic Review 109*(1), 203–36.

Schmidt-Hieber, J. (2020, 08). Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics 48*, 1875–1897.

Semenova, V. and V. Chernozhukov (2021, 08). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal 24*, 264–289. utaa027.

Small, D. S. and Z. Tan (2007). A stochastic monotonicity assumption for the instrumental variables method. *Technical report, Department of Statistics, Wharton School, University of Pennsylvania.*.

Smucler, E., A. Rotnitzky, and J. M. Robins (2019). A unifying apptoach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *ArXiv NA*, 1–125.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Stock, J. and M. Yogo (2005). *Testing for Weak Instruments in Linear IV Regression*. New York: Cambridge University Press.

Tan, Z. (2017). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *ArXiv NA*, 1–60.

Tan, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics 48*(2), 811 – 837.

Theil, H. (1953). *Esimation and Simultaneous Correlation in Complete Equation Systems.* The Hague: Central Planning Bureau. Mimeographed memorandum.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

van der Greer, S. (2016). *Estimation and Testing under Sparsity.* Lecture Notes in Mathematics. Springer, New York, NY.

van Wieringen, W. N. (2023). Lecture notes on ridge regression.

Vytlacil, E. J. (2002, January). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Vytlacil, E. J. (2006, August). Ordered discrete-choice selection models and local average treatment effect assumptions: Equivalence, nonequivalence, and representation results. *Review of Economics and Statistics 88*(3), 578–581.

Wang, W. and J. Yan (2021). Shape-restricted regression splines with R package splines2. *Journal of Data Science 19*(3), 498–517.

Wu, P., Z. Tan, W. Hu, and X.-H. Zhou (2021). Model-assisted inference for covariate-specific treatment effects with high-dimensional data.

Zelen, M. (1979). A new design for randomized clinical trials. *New England Journal of Medicine 300*(22), 1242–1245. PMID: 431682.

Zelen, M. (1990). Randomized consent designs for clinical trials: An update. *Statistics in Medicine 9*(6), 645–656.

Zimmert, M. and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding.