CERGE - EI
Center for Economic Research and Graduate Education –
Economics Institute

Essays on Instrumental Variables

Maksim Smirnov

Dissertation

Prague 2026

Dissertation Committee:
Stanislav Anatolyev (CERGE-EI, Chair)
Vasily Korovkin (Universitat Pompeu Fabra)
Nikolas Mittag (CERGE-EI)

Referees:
Kirill Borusyak (University of California, Berkeley)
Laurent Davezies (CREST-ENSAE)

Contents

Acknowledgements **iv**

Abstract **vi**

Introduction **1**

1 Treatment Effects Identification and Testing via Reduced-Form Projections **2**

1.1 Introduction 2
1.2 Model 4
1.2.1 Setup 4
1.2.2 Marginal treatment effects 6
1.2.3 MTE and causal parameters 7
1.2.4 MTE and welfare maximization 8
1.3 Identification by projection 11
1.3.1 Identification of regression derivatives by projections 11
1.3.2 MTE identification by projections 15
1.4 Adaptive nonparametric tests 16
1.4.1 Null hypothesis 16
1.4.1.1 Heterogeneity test 16
1.4.1.2 Monotone decreasing MTE test 17
1.4.2 Test statistics 17
1.4.2.1 Test function by quadratic projection 17
1.4.2.2 Test statistics and adaptivity 18
1.4.3 Critical values and asymptotic validity 20
1.4.4 Implications and extensions 23
1.4.4.1 Implications for optimal welfare 23
1.4.4.2 Redirecting power 24
1.4.5 Detection of subpopulations with nontrivial heterogeneity 24
1.5 Feasible test 26
1.6 Simulations 27
1.7 Application 29
1.8 Conclusion 29
1.A Appendix: auxiliary results 29

1.B Appendix: proofs of main results 34
1.C Appendix: details of the simulation design 44
1.D Appendix: redirection of power 44

2 Many Instruments Estimation and Inference under Clustered Dependence 46
2.1 Introduction 46
2.2 Model 48
2.2.1 Setup 48
2.2.2 Three cases against JIVE 49
2.3 Weighted leave-cluster-out IV 51
2.3.1 Leaving cluster out 51
2.3.2 Weighting by cluster sizes 53
2.3.3 Asymptotic variance estimation 54
2.4 Asymptotic theory 55
2.5 Simulations 59
2.6 Application 59
2.7 Conclusion 60
2.Q Appendix: derivations for Example 2 60
2.A Appendix: useful lemmas and main results 62
2.B Appendix: auxiliary results 87
2.C Appendix: Asymptotic Properties of Large Projection Matrices 91
2.C.1 Introduction 91
2.C.2 Setup and assumptions 93
2.C.3 Relation of $\vec{p}$ to leave-k-out analog 94
2.C.4 Asymptotic distribution of $\vec{p}$ 96
2.C.5 Some implications 96
2.C.6 Proofs 98

References 110

Acknowledgements

This dissertation would be impossible without the help, advice, and support from multiple great people I have been lucky to meet on this journey. I am grateful to my main advisor Stanislav Anatolyev, who introduced me to the world of modern econometrics. Stanislav has been very generous with his time and advice, supported me throughout all the years of my studies, and has become a multiple-times coauthor. I am indebted to my external advisors, Dmitry Arkhangelsky and Anna Mikusheva, for their guidance, support, and kindness. Anna and Dmitry have been a constant inspiration for me and I have learnt a lot from our many chats, which most definitely shaped both my research and my personality. I also thank my dissertation committee members, Vasily Korovkin and Nikolas Mittag, who have been very helpful to me, have given some thought-provoking feedback and great ideas, which helped me greatly improve this dissertation. I am thankful to two referees, Kirill Borusyak and Laurent Davezies, for their valuable time and invaluable feedback. Further, I am also grateful to Sergey Kokovin, under whose guidance I made my first steps in research and who is primarily culpable for my interest in Economics. I am exceptionally grateful to my family and my wife Vlada who have been supporting me throughout my studies – in my highs and in my lows – both emotionally and intellectually. I thank the Study Affairs Office and the Grant Office for their help with administrative matters and Andrea Downing for her great effort in editing the text of this dissertation. I gratefully acknowledge the support of the Grant Agency of Charles University Project No. 68724.

Prague, Czech Republic Maksim Smirnov

July, 2026

Abstract

In this thesis, I study instrumental variable models and propose new methods for their analysis. I derive novel theoretical results, based on which I propose new statistical procedures that are relevant for applied research. In addition to analytical derivations, I run simulations and apply the proposed methods to relevant empirical contexts. In Chapter 1, I consider a nonparametric instrumental variable model with binary treatment and propose a simple test of unobserved heterogeneity of treatment effects. The test is built on a novel statistical result that allows for the identification of the average second regression derivative by quadratic least-squares projections. Based on this statistical result, I propose a test of heterogeneity of average treatment effects across margins of participation, using the marginal treatment effects (MTE) methodology. In particular, I exploit the link between the slope of the MTE curve and the second derivative of an unknown regression function. The test is easy to implement and can be run using off-the-shelf software. Further, I illustrate the performance of my test with simulations based on an influential empirical setup. Finally, I propose a version of my test that checks for a monotone MTE curve and, more generally, monotone regression derivatives. Both versions of my test have implications for welfare maximization and policy analysis. In Chapter 2, we study the setting with many weak instrumental variables and clustered dependence. We propose a new estimator of structural parameters, which is simple in implementation. The estimator combines leave-cluster-out methodology with inverse-root-cluster-size reweighting, which together provide robustness and desirable asymptotic properties. We also propose an asymptotic variance estimator, which allows for valid inferences based on pivotized test statistics. Finally, we show the properties of our procedure using simulations and real data from an influential study.

Introduction

Instrumental variables are one of the main tools in empirical practice and are important for causal analysis. When the variable of interest is endogeneous, conventional regression-based methods for estimation and testing are invalid and a popular identification strategy is to use some instrumental variables, which are exogeneous and correlated with the treatment of interest. In these chapters, I propose new econometric methods for estimation and testing in instrumental variable models. In addition to deriving novel theoretical results, I illustrate the properties of my proposed methods with simulations, based on important empirical settings, and directly with real data.

In Chapter 1, I study unobserved heterogeneity of average treatment effects. In particular, I study a nonparametric instrumental variable model with a binary treatment and develop new methods for testing treatment effect heterogeneity. I propose tests for (i) constant marginal treatment effects (MTE) and (ii) monotone decreasing MTE curves. The analysis builds on a novel identification result showing that the average second derivative of an unknown regression function can be recovered via a quadratic least-squares projection. This result enables identification of the average slope of the MTE curve through a reduced-form projection. Building on these insights, I construct simple, projection-based tests for constant and monotone MTEs that are easy to implement and have direct implications for policy evaluation and empirical welfare maximization. Monte Carlo simulations and an empirical application demonstrate the tests' finite-sample performance and practical relevance. The procedures I propose can be directly applied more generally to test shape restrictions on the underlying regression function outside of the MTE framework. In particular, one can test for concavity or convexity of the unknown regression function.

In Chapter 2, co-authored with Stanislav Anatolyev, we study the high-dimensional setting with many possibly weak instruments under data clustering. Clustered dependence is relevant for applications but it complicates inferences about structural parameters and might threaten the consistency of conventional estimates. We propose an alternative approach to both estimation of and making inferences about structural parameters, which is computationally attractive and allows general structures of intra-cluster correlations, presence of many instruments and possibly weak identification. We use the natural extension of jackknifing, the leave-cluster-out methodology, applied to the instrument projection matrix. On top of that, we reweigh the elements of the projection matrix by an inverse geometric average of cluster sizes to flexibly adjust for cluster size heterogeneity, which helps avoid typically strong requirements on their growth rate. We set out a formal asymptotic framework to analyze the proposed weighted leave-cluster-out instrumental variables (WLCOIV) estimator, with an increasing number of clusters, possibly increasing cluster sizes, and presence of many possibly weak instruments. We prove a central limit theorem for the influence function embedded in the WLCOIV estimator under both strong and weak identification, construct an associated WLCOIV variance estimator, and show its consistency. Finally, we run a simulation experiment to verify the ability to control the actual size by the t-test based on WLCOIV estimation, and apply our methodology to an influential applied paper by Angrist and Krueger (1991).

1 Treatment Effects Identification and Testing via Reduced-Form Projections

1.1 Introduction

A crucial question in policy analysis is that of external validity. Namely, can we extrapolate the average effects of specific subpopulations to other, possibly more general, subpopulations? The answer crucially depends on the process of selection into treatment. In particular, if individuals select themselves into treatment based on their perceived effects of this treatment, the average effect for participants might be markedly different from the average effect for non-participants. In these cases, it is useful to have an instrumental variable (IV) that is exogenous, i.e. affects outcome only through the treatment selection, because it identifies average effects for certain subpopulations of compliers, i.e. individuals whose treatment status changes with the instrument. At this stage, a straightforward question that applied researchers face is the one of external validity. Specifically, is this average effect for compliers representative for other subpopulations or for the general population? If the selection into treatment decision is based on individual effects, then there is meaningful heterogeneity of average effects across individuals with different willingness to participate in treatment, and hence the average effect for compliers is not scalable to other subpopulations. However, if the data hints at the absence of selection on individual effects, such extrapolations of average effects might be justified. In this paper, I propose a novel test of treatment effect heterogeneity, that might be used to assess if the extrapolation of average effects is appropriate. The test is simple in implementation and is asymptotically valid.

Application of IV models to policy evaluation is important under non-compliance and selection into treatment on unobservables. While regression-based methods like OLS or difference-in-differences identify the average treatment effect (ATE) or the average treatment effect on the treated (ATT), a conventional IV-based two-stage least squares estimator in general identifies the *local* average treatment effect (LATE, Imbens and Angrist, 1994). ATE and ATT have straightforward interpretations: the average effect over the whole population and the average effect in the subpopulation of treated individuals (actual effect of the policy), respectively. At the same time, LATE has the interpretation of the average treatment effect for compliers, i.e. the unobserved subpopulation whose treatment status responds to a change in the instrument: e.g. for a binary instrument Z, compliers select into treatment if and only if $Z = 1$.[1] While LATE is identified by conventional methods, its policy relevance might be questioned. There is no way to identify compliers from the data, and the policy relevance of the effects for compliers has been a focus of some discussion (see Imbens, 2010; Deaton, 2009; Heckman and Urzua, 2009). When a policy maker has a specific policy in mind, she might be naturally interested in its effect on those who will be affected by this policy, i.e. PRTE (policy relevant treatment effect, Heckman and Vytlacil, 2005, Heckman et al., 2006). Unfortunately, PRTE is usually not point identified, and there is a clear distinction between what we can identify (LATE) and what we care about (usually PRTE, ATE or ATT) – see

[1] If the instrumental variable Z has k distinct values, e.g. $z_1, .., z_k$, there exist $k(k-1)/2$ groups of compliers, each corresponds to a pair of values of Z

Mogstad and Torgovitsky (2024) for a recent review.

All causal parameters discussed above (LATE, PRTE, ATE, ATT) have the same underlying concept – MTE (marginal treatment effect, Björklund and Moffitt, 1987). MTE shows how average effects vary with the margin of participation: from most willing to participate in treatment to least willing (see Section 2 for formal discussion). In fact, LATE, PRTE, ATE and ATT are different averages of MTE (see Heckman et al., 1999 and Heckman and Vytlacil, 2005). I use this insight to develop a novel test of heterogeneity in IV models: when MTE is constant, all averages of it, including LATE and PRTE, are the same, and hence LATE is externally valid.

I contribute to several strands of literature. First, I contribute to the literature on MTE identification. So far, this literature has largely focused on the identification of the whole MTE curve. Specifically, MTE is nonparametrically identified by reduced-form regression derivative (Heckman et al., 1999) when the propensity score has full support (Heckman et al., 2006). When the propensity score has limited support (e.g. with a binary or a multivalued instrumental variable), MTE is identified under linearity or additive separability (Brinch et al., 2017). Instead of focusing on the identification of the MTE curve, I focus on the identification of its *average slope*: I show that a quadratic projection of outcome on propensity score identifies the average MTE slope.

Second, I contribute to the literature on the identification of causal effects in IV models. Usually, in such models only LATE (or an average of LATEs) is point identified (Imbens and Angrist, 1994, Angrist et al., 1996, Mogstad et al., 2024). Mogstad et al. (2018) developed a procedure that partially identifies causal parameters different from LATE (e.g. PRTE or ATE), and Voronin (2025) discusses inference on these parameters. In this paper, instead of identifying the bounds on policy relevant causal parameters like Mogstad et al. (2018), I test if these policy relevant causal parameters are equal to LATE. Thus, my test provides guidance when simple extrapolation of LATE is justified.

Third, I contribute to the literature on testing for treatment effect heterogeneity in the MTE framework. Two workhorse tests were developed by Heckman et al. (2010) and are based on the linearity of reduced-form regression in propensity score under no heterogeneity. One test requires the regression estimation of polynomial approximations of different orders and testing of every one of them. Another test is a conditional moment test based on Bierens (1990). Compared to these two tests in Heckman et al. (2010), my test is easier in implementation, and I also formally show the asymptotic validity of my test. Another test that can be used was developed by Li and Nie (2008), and uses local linear regressions to nonparametrically estimate the reduced-form regression and test for its linearity. In contrast, I use local *quadratic* regressions to nonparametrically estimate the *average slope* of the reduced-form regression and test that it is zero. Further, I use a different test statistic: instead of the residual sum of squares of Li and Nie (2008), I use the maximum, following the literature on many moment inequalities (Chetverikov, 2018, Chernozhukov et al., 2019, Chetverikov, 2019). Another distinct feature of my procedure is that I outline a simple extension of it that aims at the detection of observable subpopulations with nontrivial heterogeneity. Namely, by utilizing concepts from multiple hypotheses testing literature I suggest applying my test to multiple subpopulations defined by values of covariates and aggregate p-values by the sequential procedure of Benjamini and Hochberg (1995) to control the false discovery rate.

Fourth, I contribute to the literature on welfare maximization. In the MTE framework, Liu (2022) and Sasaki and Ura (2024) derived welfare maximizing rules for general shapes of MTE in terms of incentive assignment and treatment assignment, respectively. In contrast, I focus on two shapes of MTE, namely constant MTE and monotonically decreasing MTE. In these two cases, the optimal policy is easy to derive because it does not require the unconstrained estimation of the MTE curve. My procedure directly tests for constant MTE and its straightforward extension tests for the monotonically decreasing MTE, providing guidance on the optimal incentive assignment.

Fifth, I contribute to the literature on regression derivatives identification. In particular, I derive a novel statistical result that connects quadratic least squares projections with the average second regression derivative. While it has been known that the linear projection identifies the average regression derivative (see Yitzhaki, 1996, Angrist and Krueger, 1999, Angrist and Pischke, 2009), I extend this result to quadratic projection and a second regression derivative: I show that the quadratic projection coefficient is the average second regression derivative if the explanatory variable belongs to $[0, 1]$ interval.

The rest of the paper is organized as follows. Section 2 formally introduces the model. Section 3 presents the novel statistical result on the connection of projections and unknown regression. Section 4 develops testing procedures and shows their validity. Section 5 proposes a feasible version of the tests. Sections 6 and 7 illustrate the procedures using both simulations and data. Section 8 concludes.

1.2 Model

In this Section, I formally lay out the model and introduce marginal treatment effects (MTE) and some of their applications. In particular, in Sections 2.3-2.4, I discuss the implications of MTE for heterogeneity analysis and optimal policy and motivate the tests of constant MTE and monotone decreasing MTE that I propose later in Section 4.

1.2.1 Setup

I consider a nonparametric instrumental variables model with binary treatment $D \in \{0, 1\}$. Let the observed outcome be Y; let the potential outcomes be $Y(0)$ when an individual is untreated and $Y(1)$ when an individual is treated. As usual with a binary treatment, the observed outcome is $Y = Y(0)$ for untreated individuals, and $Y = Y(1)$ for treated individuals, or equivalently

$$\begin{aligned} Y = Y(D) &= DY(1) + (1 - D)Y(0) \\ &= Y(0) + \tau D, \end{aligned} \tag{1}$$

where $\tau \equiv Y(1) - Y(0)$ is the individual treatment effect.

In general, the treated potential outcome $Y(1)$ is identified only for the treatment group, and the untreated potential outcome $Y(0)$ only for the control group. Hence, the individual treatment effect τ is not identified without additional assumptions. Multiple causal estimands are averages of τ; e.g., the average treatment effect (ATE) is $\mathbb{E}[\tau]$. In many applications of interest, treatment D is confounded, so a simple comparison of mean outcomes in treated and control groups does not identify a valid causal parameter. One way to overcome treatment confoundedness is to employ an instrumental variable Z that is exogenous and helps one to identify causal effects. Let X be observed covariates.

Assumption 1.

1.1 $D = \mathbb{1}\{p(Z, X) \geq U\}$

1.2 $U \perp\!\!\!\perp Z \mid X$

1.3 $\mathbb{E}[Y(d) \mid X, Z, U] = \mathbb{E}[Y(d) \mid X, U]$ for $d \in \{0, 1\}$

1.4 $U \sim \text{Uniform}[0, 1]$

1.5 $|\text{Supp}(Z)| \geq 3$

Assumptions 1.1-1.4 are standard in the literature (see e.g. Brinch et al., 2017), while Assumption 1.5 is an additional requirement for my test. Assumption 1.1 formally introduces a threshold-crossing treatment participation mechanism. After instrument $Z = z$ is assigned to an individual, she selects into treatment if and only if the incentive z (conditional on $X = x$) is strong enough that the induced treatment probability $p(z, x)$ exceeds an individual-specific resistance to treatment $U = u$ (formally, U is the quantile of resistance to treatment). Assumption 1.2 implies that, conditional on $X = x$, all individuals can be ordered according to their U, and if everyone gets the same incentive z^* (hence the same treatment probability $p(z^*, x)$), only individuals with $U \leq p(z^*, x)$ will select themselves into treatment. Changing Z (and hence $p(Z, x)$) will either induce additional individuals into treatment or induce some individuals out of treatment, depending on the $p(Z, X)$. Assumption 1.2 requires assigned instrument Z to be conditionally independent of the resistance to treatment U. Assumption 1.3 is an exclusion restriction – potential outcomes are conditionally mean independent of instrument Z, i.e. subpopulations with different Z do not differ in mean potential outcomes, and in average treatment effects: $\mathbb{E}[\tau|X, Z, U] = \mathbb{E}[\tau|X, U]$. One case when Assumption 1.2 and Assumption 1.3 are ensured is when Z is randomized. Further, Assumptions $1.1 - 1.3$ are equivalent to the well-known IV monotonicity assumption (Vytlacil, 2002), i.e. if an instrument induces someone into (out of) treatment, it must not induce anyone out of (into) treatment . Assumption 1.4 is a common normalization (see Mogstad and Torgovitsky, 2024).

Assumption 1.5 is crucial for this paper. It requires the support of the instrumental variables Z to have at least 3 distinct values. It permits continuous Z and multivalued Z but effectively rules out binary Z. The intuition for this additional requirement is that my test effectively compares various LATEs (see Section 1.2.3 for a concrete discussion of what they are and how they are identified), while binary Z identifies only one LATE. With multivalued or continuous Z, multiple LATEs are identified and I can compare them with each other. Thus, binary Z does not provide sufficient variation for the construction of my test.

1.2.2 Marginal treatment effects

Marginal treatment effects (MTE) are a useful way to represent treatment effect heterogeneity. MTE are defined (see e.g. Mogstad and Torgovitsky, 2024) as ATE for subpopulations with the same margin of participation $U = u$ (conditional on covariates $X = x$):

$$\text{MTE}(x, u) \equiv \mathbb{E}[\tau | X = x, U = u]. \tag{2}$$

The thought experiment behind MTE is that (conditional on X) we can order individuals by their resistance to treatment U from lowest (i.e. individuals that are the easiest to induce into treatment) to highest (i.e. individuals that are the hardest to induce into treatment). Before conducting an experiment, we are agnostic about unobservable U, so we can not predict the expected treatment effect based on U. At the same time, we know observables X and can predict the expected treatment effect based on X. For the majority of the paper, I suppress the dependence of MTE on observed X for the clarity of exposition; hence the analysis is implicitly conditional on X.

The average treatment effects might be different across different margins of participation (i.e. MTE is non-constant), which would be the evidence supporting the "selection on gains" hypothesis – that individuals select into treatment depending on their (perceived) treatment effects. Statistically, "selection on gains" implies that individual resistance to treatment U and treatment effects τ are dependent: e.g. individuals with higher perceived treatment effects may have smaller resistance to treatment U, i.e. they are easier to induce into treatment.

Alternatively, if the MTE curve is constant in u, the average treatment effects do not vary between different margins of participation: there are no systematic differences in average treatment effects between individuals that are more or less eager to select themselves into treatment. In this case, there is no evidence in favor of "selection on gains". Instead, the endogenous selection bias of the OLS estimand comes from the correlation of untreated outcomes $Y(0)$ with realized treatment status D:

$$\frac{\mathbb{E}[YD]}{\mathbb{E}[D^2]} = \bar{\tau} + \frac{\mathbb{E}[Y(0)D]}{\mathbb{E}[D^2]} + \frac{\mathbb{E}[(\tau - \bar{\tau})D^2]}{\mathbb{E}[D^2]},$$

and under no selection on gains, the last term is 0, so only the second term contributes to the endogenous selection bias.

While the definition of MTE in (2) provides a useful interpretation as the average treatment effect conditional on the resistance to treatment U, it is not instructive for the construction of MTE as U is inherently unobservable. Heckman and Vytlacil (1999) provide an alternative representation of MTE as a derivative of the regression function of Y given propensity score $p(Z) = p$ [2]

$$\text{MTE}(u) = \left(\frac{d}{dp} \mathbb{E}[Y | p(Z) = p] \right) \Bigg|_{p=u}. \tag{3}$$

[2] Note that under Assumption 1, the regression of Y given Z (*reduced-form regression*) is the same as the regression of Y given $p(Z)$: $\mathbb{E}[Y|Z] = \mathbb{E}[Y|p(Z)]$ (see Lemma A1 in Appendix).

The expression in (3) implies that the MTE curve is identified by the derivative of the regression function of Y given propensity score $p(Z) = p$. In particular, MTE is identified at every point if $p(Z)$ has full support. As noted in the discussion after Assumption 1, I do not require the full support of $p(Z)$ because I do not need to identify the MTE at every point. For the remainder of the paper, I will use $p = p(z)$ for generic $Z = z$ and $p_i \equiv p(Z_i)$ for brevity.

Example 1. (Linear reduced-form regression) If the reduced-form regression function is linear in treatment probability p, i.e. $\mathbb{E}[Y|Z = z] = a+bp(z)$ with $b \neq 0$, then representation (3) straightforwardly implies constant MTE:

$$\text{MTE}(u) = \left(\frac{d}{dp}(a + bp)\right)\bigg|_{p=u} = b.$$

Under the linear reduced-form regression, the MTE curve is constant, and there is no treatment effect heterogeneity across different margins of participation u. □

Example 2. (Quadratic reduced-form regression) If the reduced-form regression function is quadratic in treatment probability p, i.e. $\mathbb{E}[Y|Z = z] = a + bp(z) + cp^2(z)$ with $c \neq 0$, then representation (3) straightforwardly implies linear MTE:

$$\text{MTE}(u) = \left(\frac{d}{dp}(a + bp + cp^2)\right)\bigg|_{p=u} = b + 2cu.$$

Under quadratic reduced-form regression, the MTE is linear, and its slope is constant in the resistance to treatment $U = u$. Suppose that $c > 0$, i.e. the MTE is upward-sloping. In this case, the individuals with higher resistance to treatment U (more difficult to induce into treatment) have higher treatment effects on average. Suppose that $c < 0$, i.e. the MTE is downward-sloping. Then individuals with lower resistance to treatment U (less difficult to induce into treatment) have lower treatment effects on average. □

Unlike in Examples 1 and 2, if the functional form of the reduced-form regression $\mathbb{E}[Y|Z = z]$ is unknown, then one could use (3) and directly obtain the MTE by estimating regression derivatives. Alternatively, one could try to approximate the unknown MTE by some basis functions (Shea and Torgovitsky, 2021).

Alternatively, in this paper I propose two procedures that test the shape of the MTE curve without directly estimating it. In Section 3, I use the regression derivative representation in (3) to identify the average slope of the MTE curve, and then, in Section 4, I propose two tests: a test of constant MTE and a test of monotone decreasing MTE.

1.2.3 MTE and causal parameters

One appealing feature of MTE is that it is an important building block of many causal parameters (Heckman and Vytlacil, 2005). Some relevant average causal effects can be represented as

$$\mathcal{T} = \int_0^1 \text{MTE}(p)\omega^{\mathcal{T}}(p)dp, \tag{4}$$

i.e. $\mathcal{T}$ is a certain average of MTE, with weights $\omega^{\mathcal{T}}(p)$. Some notable examples include ATE and average treatment effect on the treated (ATT). ATE is $\mathcal{T}$ in (4) with weights $\omega^{\mathcal{T}}(p) = 1$, i.e. ATE $= \int_0^1 \text{MTE}(p)dp$; ATT is $\mathcal{T}$ with weights $\omega^{\mathcal{T}}(p) = \frac{1-F(p)}{\int_0^1 (1-F(t))dt}$, where F is the cdf of p, i.e. ATT $= \int_0^1 \frac{1-F(p)}{\int_0^1 (1-F(t))dt}\text{MTE}(p)dp$. See Heckman and Vytlacil (2005) for more examples of $\mathcal{T}$ and corresponding $\omega^{\mathcal{T}}(p)$.

While multiple causal parameters are average MTE, not many of them can be identified with the usual methods. In IV models, the usual two-stage least squares estimator identifies one particular average causal effect, the local average treatment effect (LATE, Imbens and Angrist, 1994), the average treatment effect for *compliers* – the subpopulation whose treatment status is changed by a given instrument. It is plausible that compliers have treatment effects that are on average different from those of other individuals, so LATE is different from other effects, e.g. ATE. Hence, the external validity and policy relevance of LATE is not straightforward.

One case when LATE is policy relevant is under no (average) treatment effect heterogeneity, which corresponds to constant MTE, i.e. $\text{MTE}(u) = C$ for some constant C. In this case, average effects are constant over margins of participation, and the manipulation of treatment probabilities will not change average effects. In particular, from (4),

$$\int_0^1 \text{MTE}(p)\omega^{\mathcal{T}}(p)dp = C\int_0^1 \omega^{\mathcal{T}}(p)dp = C$$

for any proper (non-negative, integrates to one) weighting function $\omega^{\mathcal{T}}(p)$. Hence, a test of no (average) treatment effect heterogeneity (i.e. test of constant MTE) that I develop in Section 4 provides a useful tool that can signal the policy relevance of LATE.

1.2.4 MTE and welfare maximization

One use of the MTE curve is to inform policy on optimal allocation of incentive Z (Liu, 2022)[3]. Fundamentally, individual resistance to treatment U is unobservable (and in general not point identified), so the only possible MTE-based policy is essentially uniform across margins of participation u. Suppose for a moment that no covariates X are available. There is no way to determine where on the MTE curve individuals are because $\{U_i\}_i$ are unobservable, and the only policy possible is to set a common cutoff z^*, such that individuals with $U \leq p(z^*)$ select themselves into treatment, according to the treatment selection mechanism in Assumption 1.2. Thus, the actionability of the MTE curve is limited by threshold-setting policies.

[3]Sasaki and Ura (2024) use the MTE framework to derive optimal allocation of treatment D directly. However, in some applications it might be infeasible or undesirable to allocate treatment, hence I focus on the optimal allocation of incentive Z that induces individuals to self-select into or out of treatment.

If there are additional covariates X, the same argument applies to separate covariate-specific MTE curves. E.g. if X is a set of dummy variables for each US state, then we can consider separate MTE curves for each state; within each of these *covariate-specific* MTE curves there is no way to distinguish individuals with different U (same as in the absence of X). Hence, the policy is still limited by cutoff policies that now depend on observed covariates, i.e. $z^*(X)$.

Existing methods are based on the fact that multiple causal estimands can be represented as various averages of MTE (Heckman et al., 1999, Heckman et al., 2006) and rely on an estimated MTE curve. In particular, the empirical welfare criterion in Liu (2022) is

$$\hat{W}_n(\underline{z}, \bar{z}) = \frac{1}{n}\sum_{i=1}^{n}\int_{\hat{p}(X_i,\underline{z})}^{\hat{p}(X_i,\bar{z})} \widehat{\text{MTE}}(u, X_i)du,$$

and the maximization of $\hat{W}_n$ essentially requires comparisons of various integrals of MTE. Hence, while allowing for general shapes on the MTE curve, the method of Liu (2022) relies on functionals of estimated MTE. However, I show that for certain shapes of the MTE curve, the optimal policy can be derived without estimation of the curve itself.

The easiest example is when the MTE curve is constant as in Figure 1. It is straightforward that optimal policy z^* in Figure 1 is given either by maximum incentive $\arg\max_z p(z)$ (Panel A) or minimum incentive $\arg\min_z p(z)$ (Panel B). The former case corresponds to a positive average treatment effect for every margin of participation u, and hence it is optimal to induce everyone into treatment – set the maximum incentive possible. The latter case corresponds to a negative average treatment effect for every u, and hence it is optimal to induce everyone out of treatment by setting the minimum incentive possible. In practice it is straightforward to distinguish between the two cases. Under the null of constant MTE, the regression of outcome on treatment probabilities is linear $Y = a + bp + e$ by (3) (as in Example 1), and can be estimated by OLS. Then a straightforward t-test of $b \geq 0$ against $b < 0$ provides evidence in favor of non-negative constant MTE (Panel A, optimal $z^* = \arg\max_z p(z)$) against negative constant MTE (Panel B, optimal $z^* = \arg\min_z p(z)$)[4].

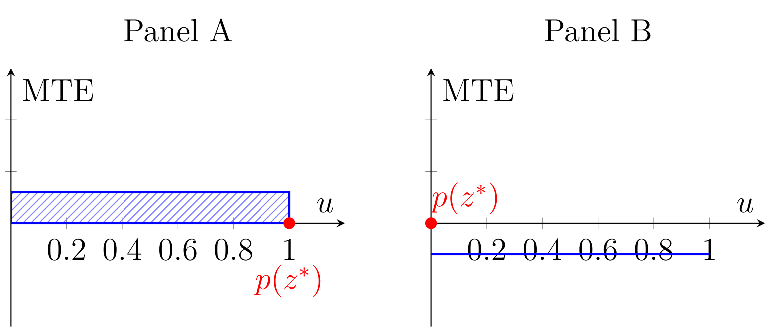

Figure 1: MTE curve example under no selection on gains

If the MTE curve is not constant as in Figure 1, the optimal policy can still be derived in a simple way. If the MTE curve is monotone (Figure 2), the optimal policy corresponds

[4]If treatment probabilities p_i are not observed, one can apply a two-step procedure, similar to the one discussed in Section 5; the corresponding variance estimator uses estimated $\hat{p}_i$

to zeros of the MTE curve (in line with the results of Chen and Xie, 2022). In particular, if the MTE curve is monotone decreasing as in Figure 2, Panel A), then it is optimal to induce into treatment individuals with low resistance to treatment U, namely $U \leq 0.6$, and induce out of treatment individuals with higher resistance to treatment, $U > 0.6$. Hence, the optimal cutoff policy is z^* such that $p(z^*) = 0.6$ – at the intersection of the MTE curve with the u-axis. However, if the MTE curve is monotone increasing (as in adverse selection), there is no simple optimal policy available. To see this, notice that in Figure 2, Panel B, it is desirable to induce into treatment those individuals with $U \in (0.5, 1]$, for whom the average treatment effects are positive, and induce out of treatment those individuals with $U \in [0, 0.5)$, for whom the average treatment effects are negative. The difficulty is that individual resistance to treatment U is unobservable, so the cutoff policy that induces into treatment those individuals with positive average effects must also induce into treatment those individuals with negative effects. There are two candidates for optimal policy: induce as many as possible and set z^* equal to $\arg\max_z p(z)$ or induce as few as possible and set z^* equal to $\arg\min_z p(z)$.[5] Which one of them is truly optimal depends on the area of positive effects $|\int_{0.5}^{1} \text{MTE}(u)du|$ relative to the area of negative effects $|\int_{0}^{0.5} \text{MTE}(u)du|$, and hence requires the estimation of the MTE curve.

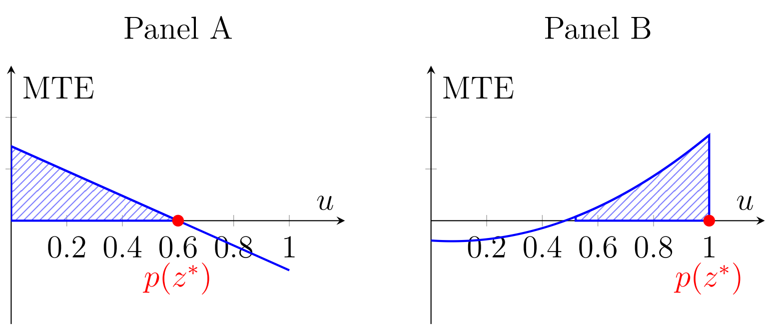

Figure 2: Monotone MTE curve example

Further, if the MTE curve is not monotone, the optimal policy requires the estimation of the curve. Figure 3 shows two MTE curves that have the same intervals of monotonicity but different optimal policies. The general principle for optimal policy setting in non-monotone settings is to compare the areas of positive effects to the areas of negative effects and set cutoffs at the zeros of MTE or extreme points of the support – 0 and 1. In both Panels of Figure 3, 1 is not an optimal policy (dominated by 0.8), 0 is not an optimal policy (dominated by 0.3), and 0.5 is not an optimal policy (dominated by 0.8), so in fact only 0.3 and 0.9 are valid candidates. The decision rule depends on the integrals of MTE:

$$z^* = \begin{cases} z : p(z) = 0.8 \text{ if } \int_0^{0.3} \text{MTE}(u)du + \int_{0.5}^{0.8} \text{MTE}(u)du \geq |\int_{0.3}^{0.5} \text{MTE}(u)du|, \\ z : p(z) = 0.3 \text{ otherwise,} \end{cases}$$

[5] A simple argument is that for every cutoff $p' \in (\min_z p(z), \max_z p(z))$ with a corresponding policy $z' : p(z') = p'$ there exists an alternative cutoff $p'' > p'$ such that the corresponding policy $z'' : p(z'') = p''$ induces more individuals into treatment, and the average treatment effect of these additionally treated (under z'' relative to under z') is bigger than the average treatment effect at any margin of participation u that is treated (under z') due to monotone increasing MTE. Thus, z'' yields a higher total effect than z'.

and thus the *feasible* decision rule requires an estimate $\widehat{\text{MTE}}(u, x)$.

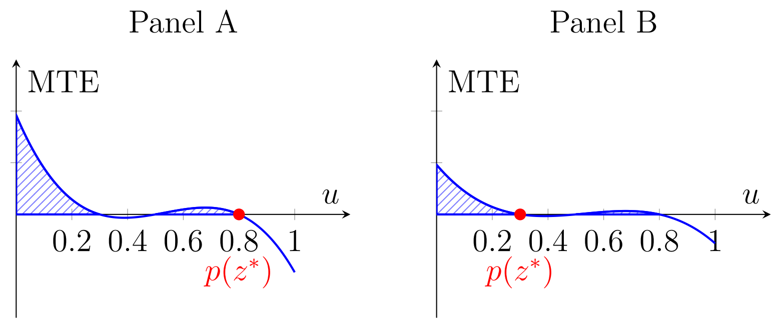

Figure 3: Non-monotone MTE curve example

Example 3. (U-shaped MTE) Brinch et al. (2017) recover a U-shaped MTE curve with negative average treatment effects for non-extreme margins. Without the use of covariates, the optimal policy based on such an MTE curve is either to treat everyone with $u \leq u^1$, where u^1 is the first zero of the MTE curve, or to treat everyone. Which of these two options is optimal is governed by the relative size of $|\int_{u^1}^{u^2} \text{MTE}(u)du|$ and $|\int_{u^2}^{1} \text{MTE}(u)du|$, and therefore requires the identification of the whole MTE curve. $\square$

From the discussion above, the constant MTE and monotone decreasing MTE have direct implications for optimal incentive assignment. In Section 4, I propose nonparametric adaptive testing procedures that check for constant MTE and for monotone decreasing MTE, i.e. tests for the "simple optimal policy": if the tests suggest that the MTE is constant (no selection on gains) or monotone decreasing, then the optimal policy can be easily determined; if the tests reject, then the MTE needs to be estimated for the determination of the optimal policy.

1.3 Identification by projection

In this section, I show how least squares projections can identify average regression derivatives. This identification result is the foundation of the testing procedure I propose in Section 4. I start with a well-known statistical result that the linear projection coefficient identifies the average regression derivative. Second, I extend this result to quadratic projection and average second-order derivatives. Third, I show how this result implies identification of the average MTE slope.

1.3.1 Identification of regression derivatives by projections

It has been well-established in the literature that the linear projection coefficient identifies an average regression slope (Angrist and Krueger, 1999). In this paper, I extend this result by showing that the *quadratic* projection coefficient is an average regression curvature (i.e. second regression derivative).

Let the unknown reduced-form regression be $m(p) \equiv \mathbb{E}[Y|p(Z) = p]$, and let f be the probability density function of p. Consider an infeasible[6] linear projection of observed outcome Y on treatment propensities p:

$$Y = \alpha_L + \beta_L p + e. \tag{5}$$

By usual arguments (see e.g. Angrist and Pischke, 2009), the population projection coefficient of Y on p is

$$\beta_L = \frac{\text{Cov}(Y, p)}{\text{Var}(p)} = \frac{\mathbb{E}[m(p)(p - \mathbb{E}[p])]}{\mathbb{E}[p(p - \mathbb{E}[p])]}. \tag{6}$$

While the true regression $m(p)$ might be non-linear and is usually unknown, projection (5) provides the best linear approximation and is straightforward to estimate and do inference on. Additionally, β_L identifies the average derivative of the unknown regression $m(p)$.

Lemma 1 (*Angrist and Pischke, 2009*). Assume reduced-form regression $m(p)$ is differentiable, and $\text{Var}(p) > 0$. Then

$$\beta_L = \int_0^1 m'(t)\omega(t)dt, \tag{7}$$

with weights

$$\begin{aligned} \omega(t) &= \frac{\int_t^1 (p - \mathbb{E}[p]) f(p) dp}{\mathbb{E}[p(p - \mathbb{E}[p])]} \\ &= \frac{\mathbb{E}[p - \mathbb{E}[p]|p \geq t] \Pr\{p \geq t\}}{\mathbb{E}[p(p - \mathbb{E}[p])]} \geq 0. \end{aligned} \tag{8}$$

Proof. See Angrist and Pischke (2009), Section 3.5.

Remark 1. (Average MTE identification by projection) Notice that from (3) and the definition of $m(p)$,

$$m'(p) \equiv \frac{d\mathbb{E}[Y|p(Z) = p]}{dp} = \text{MTE}(p). \tag{9}$$

Combine (7) and (9) to obtain

$$\beta_L = \int_0^1 \text{MTE}(p)\omega(p)dp,$$

i.e. linear projection (5) identifies average MTE, and the weights are given by (8) in Lemma 1.

[6]for the majority of the paper, I implicitly discuss infeasible projections; except for section 5, where I discuss the feasible version of my tests

Remark 2. (LATE identification by projection) As noted in Section 2.3, multiple causal parameters are certain averages of MTE. However, unlike other causal estimands (e.g. ATE or ATT), LATE is point identified by conventional methods. Namely, LATE is identified by the two-stage least squares estimand (Imbens and Angrist, 1994), which is essentially a coefficient of a linear projection of outcome on treatment probabilities.

Now consider a quadratic projection of Y on p

$$Y = \alpha + \beta p + \gamma p^2 + \varepsilon. \tag{10}$$

Let $\mathbb{L}[Y|X]$ be the population projection of Y on some variables X and a constant. β and γ are ratios of covariances to variances of orthogonalized variables (p and p^2, respectively, Angrist and Pischke, 2009, chapter 3.1):

$$\beta = \frac{\mathrm{Cov}(Y, p - \mathbb{L}[p|p^2])}{\mathrm{Var}(p - \mathbb{L}[p|p^2])} = \frac{\mathrm{Cov}(m(p), p - \mathbb{L}[p|p^2])}{\mathrm{Var}(p - \mathbb{L}[p|p^2])}, \tag{11}$$

$$\gamma = \frac{\mathrm{Cov}(Y, p^2 - \mathbb{L}[p^2|p])}{\mathrm{Var}(p^2 - \mathbb{L}[p^2|p])} = \frac{\mathrm{Cov}(m(p), p^2 - \mathbb{L}[p^2|p])}{\mathrm{Var}(p^2 - \mathbb{L}[p^2|p])}. \tag{12}$$

Example 2. (Quadratic reduced-form regression, *continued*) If the true reduced-form regression function is quadratic in p (i.e. $\mathbb{E}[Y|Z=z] = a + bp(z) + cp^2(z)$), then this regression coincides with the quadratic projection, $\mathbb{E}[Y|p] = \mathbb{L}[Y|p, p^2]$, and $\gamma = c$. As shown in Example 2, the MTE in this case is linear: $\mathrm{MTE}(u) = b + 2cu = b + 2\gamma u$, and γ is (proportional to) the constant MTE slope. □

If the reduced-form regression is not quadratic or linear in p (see Examples 1 and 2), the MTE slope is not constant. However, the coefficient γ of the quadratic projection of outcome on treatment probabilities (10) still identifies the *average* MTE slope.

Proposition 1. Assume reduced-form regression $m(p)$ is twice differentiable. Let $\widetilde{p^2} \equiv p^2 - \mathbb{L}[p^2|p]$, and assume $\mathrm{Var}(\widetilde{p^2}) > 0$. Then

$$\gamma = \int_0^1 m''(t)\omega(t)dt, \tag{13}$$

with weights

$$\omega(t) = \frac{\mathbb{E}\left[\widetilde{p^2}(p-t)|p \geq t\right] \Pr\{p \geq t\}}{\mathbb{E}[p^2\widetilde{p^2}]} \geq 0. \tag{14}$$

Proof. See Appendix.

The intuition is that higher order polynomials in projection reflect higher order properties of the regression curve. At the same time, inclusion of lower order polynomials plays the orthogonalizing role, and removes lower order effects. E.g. inclusion of p removes first order effects from γ; without p, γ would not have a clean interpretation of the average second derivative. By arguments similar to the proof of Lemma 1, it can be shown that the linear coefficient β in (11) is a weighted slope of the reduced-form regression $m(p)$.

Weights $\omega(t)$ are nonnegative, and they are the average approximation errors $p^2 - \mathbb{L}[p^2|p]$, i.e. the difference between the quadratic function p^2 and its linear projection on p. Notice that for $m''(t)$, only larger observations, $p \geq t$, influence the weights $\omega(t)$. Further, the approximation errors are inversely rescaled by the distance to t: so closer (in p) observations have less influence on $\omega(t)$.

While Proposition 1 provides the average regression derivative interpretation to the *unweighted* projection coefficient γ in (10), it is useful to provide a similar interpretation to the coefficient of a *weighted* projection. Proposition 2 below formalizes the interpretation of a quadratic coefficient in a *weighted* projection as the average second regression derivative. Later, in Section 4, I apply Proposition 2 to kernel-weighted projections in order to construct a test statistic of the null hypothesis of no treatment effect heterogeneity.

Fix parameter s and consider a weighted projection of Y on p with some weights $K(p,s)$. Denote the $K(p,s)$-weighted projection of A on B and a constant as $\mathbb{L}_s[A|B]$. The coefficients of this weighted projection are

$$(\alpha(s), \beta(s), \gamma(s)) = \underset{a(s),b(s),c(s)}{\arg\min}\ \mathbb{E}[K(p,s)\left(Y - a(s) - b(s)p - c(s)p^2\right)^2]. \tag{15}$$

Proposition 2 Assume reduced-form regression $m(p)$ is twice differentiable. Let $\widetilde{p^2} \equiv p^2 - \mathbb{L}_s[p^2|p]$ and assume $\mathrm{Var}(\widetilde{p^2}) > 0$. Then

$$\gamma(s) = \int_0^1 \omega(t,s) m''(t) dt,$$

with weights

$$\omega(t,s) = \frac{\mathbb{E}[\widetilde{p^2}(p-t)K(p,s)|p \geq t]\Pr\{p \geq t\}}{\mathbb{E}[p^2\widetilde{p^2}K(p,s)]} \geq 0. \tag{16}$$

Proof. See Appendix.

Proposition 2 extends the result of Proposition 1 to weighted projections, i.e. it provides the interpretation of the average second regression derivatives to the quadratic coefficient. One useful example of weighted projections that I use in this paper is the kernel-weighted projection, $K(p,s)$ a kernel function K with parameters s, evaluated at point p. Hence, by Proposition 2, the kernel-weighted projection estimand is a proper average second regression derivative, and can be tested as a moment restriction.

Weights $\omega(t,s)$ in (16) are nonnegative, and are similar to the weights $\omega(t)$ in (14), i.e. they are the average approximation errors $\widetilde{p^2} = p^2 - \mathbb{L}[p^2|p]$ for $p \geq t$ inversely weighted by the distance to t. However, the weights $\omega(t, s)$ in (16) are further rescaled by the weighting function $K(p, s)$. For $K(p, s)$ a kernel function, this additional rescaling gives proportionally more weight to observations close to t. The degree of this adjustment depends on the kernel and the bandwidth.

1.3.2 MTE identification by projections

In this subsection, I show how Propositions 1 and 2 help to identify the slope of the MTE curve. By (3), the MTE curve is a derivative of the unknown reduced-form regression: $\text{MTE}(p) = m'(p)$. Therefore, if the reduced-form regression $m(p)$ is smooth enough, the *slope* of the MTE curve is given by $m''(p)$. Proposition 1 directly implies that the average slope of the MTE curve is identified by quadratic projection of outcome Y on treatment probabilities p.

Corollary 1. Suppose Assumption 1 is satisfied, the reduced-form regression $m(p)$ is twice differentiable, and p has at least 3 distinct points. A quadratic projection of outcome Y on treatment propensity p is the average slope of the MTE curve

$$\int_0^1 \omega(u)\text{MTE}'(u)du,$$

and the weights $\omega(u)$ are given by (14).

Proof. See Appendix.

Corollary 1 shows that the shape of the MTE curve can be identified by a straightforward projection. Knowledge of its shape provides guidance on both the optimal incentive assignment (see Section 2.4), and the effects heterogeneity, which is indicative of global validity of local effects (see Section 2.3).

While Corollary 1 is a global result, i.e. it identifies the average slope over the whole support of p, it can be applied locally by manipulating the weights $\omega(t)$. E.g. one possibility is to split the support of p by quartiles and run the quadratic projection for each of four subintervals defined by quartiles. The result will be the average MTE slope for each such subinterval. Another possibility is to reweigh observations to obtain another average MTE slope. One particular reweighting scheme I consider is kernel weighting, and Corollary 2 extends the identification result in Corollary 1 to weighted projections.

Corollary 2. Suppose Assumption 1 is satisfied, the reduced-form regression $m(p)$ is twice differentiable, and p has at least 3 distinct points. Let $K(p, s)$ be a kernel function with parameter s. The $K(p, s)$-weighted quadratic projection of outcome Y on treatment propensity p is the average slope of the MTE curve

$$\int_0^1 \omega(u, s)\text{MTE}'(u)du,$$

and the weights $\omega(u, s)$ are given by (16).

Proof. See Appendix.

One important advantage of Corollaries 1 and 2 is that they identify the (average) slope of the MTE curve without estimating the curve. Another important advantage is that Corollaries 1 and 2 do not require a continuous instrumental variable and a full support of treatment probabilities.

Remark 3. (Smooth $m(p)$ with discrete Z) Propositions 1 and 2 require a differentiable reduced-form regression function $m(p)$. In practice, however, the distribution of p might be discrete (if available instruments are discrete) and hence $m(p) \equiv \mathbb{E}[Y|p]$ is identified only at some points. It turns out that the distribution of p matters only for the weights $\omega(t)$ (or $\omega(t,s)$), while the smoothness of $m(p)$ (which might be unidentified at some points) is crucial for Propositions 1 and 2 to hold for either continuous or discrete distribution of p. In the case of the discrete support of p, one might think about hypothetical experiments that change p to values other than those retrieved from the data. These hypothetical experiments would fully describe the MTE curve, even though MTE will be identified only at some points.

1.4 Adaptive nonparametric tests

In this Section, I propose two tests: one is a test of treatment effect heterogeneity in the MTE framework, and the other is a test of monotone decreasing MTE. My proposed tests are based on Proposition 2 and Corollary 2, derived in Section 3; the inference is based on bootstrap. I discuss how to use the tests for optimal incentive assignment and interpretation of causal effects. Finally, in Section 4.5, I suggest a multiple testing procedure that detects the observable subpopulations (defined by covariates) with non-trivial treatment effect heterogeneity.

1.4.1 Null hypothesis

1.4.1.1 Heterogeneity test

treatment effect homogeneity is equivalent to constant MTE curve

$$\text{MTE}(u) = \text{const} \quad a.s. \tag{17}$$

For twice differentiable $m(p)$, (17) is equivalent to

$$m''(p) = 0 \quad a.s. \tag{18}$$

$m''(p)$ is not always nonparametrically identified at every point, so testing (18) directly might be complicated. However, a straightforward testable implication of (18) is that every weighted average MTE slope $m''(p)$ is zero. For some parameter $s \in \mathcal{S}$ and a weighting function $\Delta(p,s)$, I propose the null hypothesis of equality of average slope to zero

$$\mathbb{H}_0 : \mathbb{E}[\Delta(p,s)m''(p)] = 0, \text{ for all } s \in \mathcal{S},$$
$$\mathbb{H}_a : \mathbb{E}[\Delta(p,s)m''(p)] \neq 0, \text{ for some } s \in \mathcal{S}.$$

In contrast to the second regression derivative $m''(p)$, the *average* second regression derivative $\mathbb{E}[\Delta(p,s)m''(p)]$ is nonparametrically identified by kernel-weighted projections (by Corollary 2) with a two-dimensional parameter s. Specifically, $s \in \mathcal{S} = \{(\mu, h)\}$ with μ being a location parameter and h being a scale parameter (bandwidth). $\mu \in \{p_1, ..., p_n\}$, where p_i is observed (or estimated – see Section 5) propensity score, and h is some fixed bandwidth, e.g. $h \in \{0.1, 0.2, 0.3\}$. By construction, it might be that $|\mathcal{S}| > n$, i.e. the number of tested moment conditions can be larger than the sample size. Thus, in constructing my test, I follow the literature on testing many moment inequalities.

1.4.1.2 Monotone decreasing MTE test

As noted in Section 2.4, it might be useful for optimal incentives assignment to know if the MTE curve is monotone decreasing. In this case, the MTE is characterized by non-positive slope

$$m''(p) \leq 0 \quad a.s. \tag{19}$$

Similarly to the argument in Section 4.1.1, a straightforward testable implication of (19) is that every weighted average MTE slope $m''(p)$ is non-positive. For some parameter $s \in \mathcal{S}$, I propose the null hypothesis of non-positivity of average slope

$$\begin{aligned} \mathbb{H}_0^m &: \mathbb{E}[\Delta(p,s)m''(p)] \leq 0, \text{ for all } s \in \mathcal{S}, \\ \mathbb{H}_a^m &: \mathbb{E}[\Delta(p,s)m''(p)] > 0, \text{ for some } s \in \mathcal{S}. \end{aligned}$$

As noted in Section 4.1.1, the *average* second regression derivative $\mathbb{E}[\Delta(p,s)m''(p)]$ is nonparametrically identified by kernel-weighted projections for $s \in \mathcal{S} = \{(\mu, h)\}$ with $\mu \in \{p_1, ..., p_n\}$, and h is some fixed bandwidth.

1.4.2 Test statistics

1.4.2.1 Test function by quadratic projection

In line with $\mathbb{H}_0$ (and $\mathbb{H}_0^m$), let test function $b(s)$ be an average MTE slope with weights $\Delta(p,s)$, where s is some parameter that defines the weighting scheme:

$$b(s) = \mathbb{E}[\Delta(p,s)m''(p)] = \int_0^1 \Delta(s,t)m''(t)f(t)dt. \tag{20}$$

Under the null of no heterogeneity, $b(s) = 0$ for every weighting scheme s. Test function $b(s)$ is an average, which is common in the nonparametric testing literature (see e.g. Chetverikov, 2019). If the second regression derivative $m''(p)$ is observed, the test function $b(s)$ in (20) is straightforward to estimate. However, usually $m''(p)$ is not observed. In this case, one could estimate $m''(p)$ at every point and then estimate $b(s)$ using first-step estimates $\widehat{m''}(p)$. Instead of estimating $m''(p)$, I suggest computing multiple weighted quadratic projections, each corresponding to every s, and use the quadratic coefficients $\gamma(s)$ of each of them as test functions $b(s)$.

I focus on the kernel weighting scheme with a two-dimensional parameter s. Namely, for $s = (\mu, h)$ and some kernel $K(p, s)$, by Proposition 2, $\gamma(s)$ identifies average $m''(p)$

$$\begin{aligned}\gamma(s) &= \int_0^1 \frac{\mathbb{E}[(p^2 - (a_0 + b_0 p))(p-t)K(p,s)|p \geq t]\Pr\{p \geq t\}}{\mathbb{E}[p^2 \tilde{p^2} K(p,s)]} m''(t)dt, \\ &= \mathbb{E}[\Delta(p,s)m''(p)] \end{aligned} \tag{21}$$

with

$$\Delta(t,s) = \frac{\mathbb{E}[(p^2 - (a_0 + b_0 p))(p-t)K(p,s)|p \geq t]\Pr\{p \geq t\}}{\mathbb{E}[p^2 \tilde{p^2} K(p,s)]f(t)}$$

if $f(t) \neq 0$, and 0 otherwise. A straightforward benefit of using a weighted projection coefficient $\gamma(s)$ as a test function is that it is easy to compute using the regular weighted least squares formula (see B. Hansen, 2022):

$$\begin{pmatrix}\alpha(s)\\ \beta(s)\\ \gamma(s)\end{pmatrix} = \left(\mathbb{E}\left[K(p,s)\begin{pmatrix}1\\ p\\ p^2\end{pmatrix}\begin{pmatrix}1\\ p\\ p^2\end{pmatrix}'\right]\right)^{-1} \mathbb{E}\left[YK(p,s)\begin{pmatrix}1\\ p\\ p^2\end{pmatrix}\right]. \tag{22}$$

Another benefit of using a weighted projection coefficient $\gamma(s)$ as a test function is that it is straightforward to obtain an estimate $\widehat{\gamma(s)}$ by the sample analog of (22) Finally, the inference on $\gamma(s)$ is straightforward, and the variance $V(s)$ of $b(s)$ is the usual weighted least squares variance (also straightforwardly estimable).

Example 4.1. (Test function weights with rectangular kernel) Let K be a rectangular kernel, i.e. $K(p, \mu, h) = \frac{1}{2}$ if $|p - \mu| \leq h$ and 0 otherwise. Weighting by a rectangular kernel gives the same weight to all observations with p within the distance h from location parameter μ. As $\omega(t, s)$ in (16) reweighs approximation errors asymmetrically (only for $p \geq t$), additional weighting by a kernel function will also be asymmetric. □

Example 4.2. (Test function weights with triangular kernel) Let K be a triangular kernel, i.e. $K(p, \mu, h) = \max\{(1 - \frac{|p-\mu|}{h\sqrt{6}})/\sqrt{6}, 0\}$. Weighting by a triangular kernel gives maximum weight to observations with $p = \mu$, and the weight decays linearly with distance $|p - \mu|$. As in Example 4.1, weighting by a kernel function will be asymmetric by the form of $\omega(t, s)$. □

1.4.2.2 Test statistics and adaptivity

In this paper, I focus on the maximum test statistic in line with the literature (see Chetverikov, 2018, Chernozhukov et al., 2019)

$$T_n = \max_{s \in \mathcal{S}} \sqrt{n}|\widehat{\gamma(s)}|/\sqrt{\widehat{V(s)}}. \tag{23}$$

The intuition of using a two-dimensional parameter $s = (\mu, h)$ is to make the test adaptive to both the location (μ) and the scale (h) of deviations from the null $\mathbb{H}_0$ (similar to the test of Chetverikov, 2018). By optimization over $\mathcal{S}$, T_n captures deviations of the average slope $b(s)$ from the null of the zero slope by inspecting different locations μ and smoothing parameters h. Larger deviations from the null (i.e. larger $|b(s)|$) that are identified with better precision (i.e. smaller $V(s)$) indicate stronger evidence against the null hypothesis of the constant MTE curve, hence the decision rule is to reject the null if

$$T_n > c_n. \tag{24}$$

I discuss how to construct critical value c_n in Section 4.3.

The adaptivity of (23) means that the test statistic T_n automatically adapts to various alternatives. To see this, consider an example MTE curve in Figure 4. There are two deviations from the null of constant MTE $\mathbb{H}_0$: a steady downward slope on $[u_1, u_2]$, and a sharp drop on $[u_3, u_4]$. In order to detect these violations of the null, one needs to test locations in these intervals; hence the importance of location adaptivity. Further, in order to detect the sharp drop on $[u_3, u_4]$, one needs to adapt not only to the location but also to the shape of the deviation. Using large bandwidth even for locations in $[u_3, u_4]$ might overlook this deviation by using a lot of points that do not contradict the null of constant MTE, but a smaller bandwidth will help to detect this deviation by more heavily relying on points that violate the null. Hence the test statistic in (23) is more powerful against deviations similar to the one on $[u_3, u_4]$ when using smaller bandwidth h.

As for the deviation from the null hypothesis $\mathbb{H}_0$ on $[u_1, u_2]$, it is not necessary to have a smaller bandwidth in order to detect it, but having a larger bandwidth is useful for better precision. With a larger bandwidth, variance $V(s)$ is smaller and the pivotized test function $|b(s)|/V(s)$ is larger; hence the test statistic in (23) is more powerful against deviations similar to the one on $[u_1, u_2]$ when using larger bandwidth h. Thus, optimizing over both larger and smaller bandwidth h provides the adaptivity of the test statistic T_n (23) to the shape of various alternatives.

The algorithm is:

Adaptive test of heterogeneity

1. Let $\{p_i\}_{i=1}^n$ be a collection of treatment probabilities; for every $\mu \in \{p_1, ..., p_n\}$ run a kernel-weighted (with bandwidth h) quadratic projection, obtain quadratic coefficients $\gamma(\mu, h) = b(\mu, h)$. Repeat for every $\mu \in \{p_1, ..., p_n\}$ and $h \in \{0.1, 0.2, 0.3\}$
2. Construct $T_n = \max_{\{s=(\mu,h)\}} \sqrt{n}|\widehat{\gamma(s)}|/\sqrt{\widehat{V(s)}}$
3. Reject the null of no heterogeneity if $T_n > c_n$

Finally, the same test function $b(s)$ can be used to construct the test of monotone decreasing MTE from Section 4.1.2. The test statistic T_n in (23) penalizes deviations from zero slope, $b(s) \neq 0$ (i.e. constant MTE). One way to modify T_n is to penalize only deviations

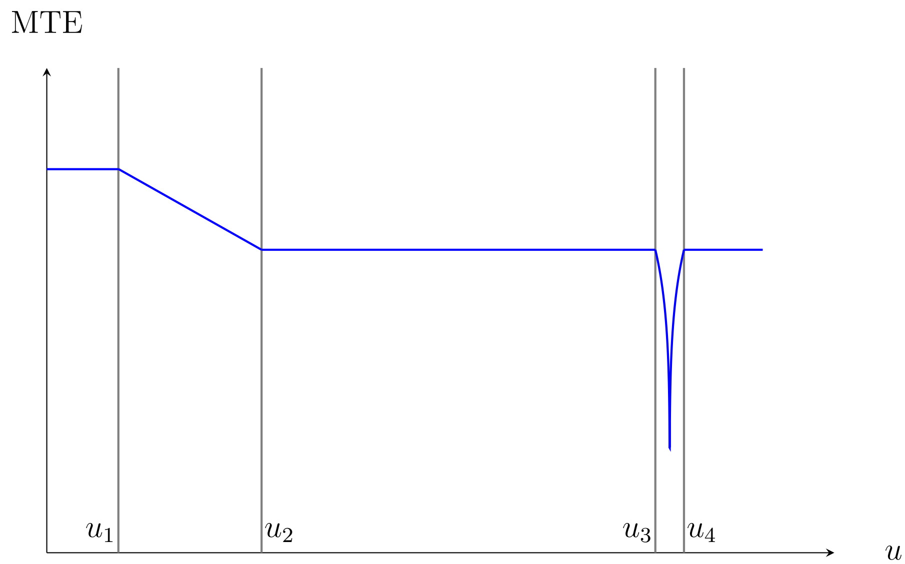

Figure 4: Adaptivity example

from non-positive slope, $b(s) > 0$ (i.e. monotone decreasing MTE):

$$T_n^m = \max_{s \in \mathcal{S}} \sqrt{n}\widehat{\gamma(s)}/\sqrt{\widehat{V(s)}}. \tag{25}$$

Similarly to T_n, the test statistic T_n^m is adaptive to the location and shape of deviations from the null hypothesis $\mathbb{H}_0^m$. Larger T_n^m indicates larger and more precisely identified deviations from the non-positive slope; hence the decision rule is to reject $\mathbb{H}_0^m$ if

$$T_n^m > c_n^m. \tag{26}$$

The algorithm is:

Adaptive test of monotone decreasing MTE

1. Let $\{p_i\}_{i=1}^n$ be a collection of treatment probabilities; for every $\mu \in \{p_1, ..., p_n\}$ run a kernel-weighted (with bandwidth h) quadratic projection, obtain quadratic coefficients $\gamma(\mu, h) = b(\mu, h)$. Repeat for every $\mu \in \{p_1, ..., p_n\}$ and $h \in \{0.1, 0.2, 0.3\}$
2. Construct $T_n^m = \max_{\{s=(\mu,h)\}} \sqrt{n}\widehat{\gamma(s)}/\sqrt{\widehat{V(s)}}$
3. Reject the null of monotone decreasing MTE if $T_n^m > c_n^m$

1.4.3 Critical values and asymptotic validity

It remains to discuss how to construct critical values for the decision rules (24) and (26). I follow the literature on testing many moment inequalities (Chernozhukov et al., 2017, Chernozhukov et al., 2019, Y. Bai et al., 2022); namely, I propose bootstrap-based procedures for the construction of c_n and c_n^m and show their validity.

In Section 4.2.1, by (21), the test function $b(s)$ is set to be the kernel-weighted least squares estimand $\gamma(s)$. Hence, by construction of the weighted least squares estimand in (22), the test function is linear in the outcome Y:

$$\begin{aligned} b(s) = \gamma(s) &= C'\mathbb{E}\left[YK(p,s) \begin{pmatrix} 1 \\ p \\ p^2 \end{pmatrix} \right] \\ &\equiv \mathbb{E}\left[M_s \right] \end{aligned} \tag{27}$$

for some constant vector C, where the moment function M_s is

$$M_s \equiv M(Y,p,s) = YK(p,s)C' \begin{pmatrix} 1 \\ p \\ p^2 \end{pmatrix}. \tag{28}$$

Notice that for fixed s, M_s is a function of (Y,p) only, so if the data $\{Y_i,p_i\}_i$ is a random sample, then the moment function M_s is i.i.d. across observations i, but can be dependent across test parameters s. Thus, $\gamma(s)$ is an expectation of i.i.d. terms, and therefore I can apply results from Chernozhukov et al. (2019) for the construction of critical values. Let $\mathbb{E}_n[.]$ denote a sample mean, i.e. $\mathbb{E}_n[Q] = \sum_{i=1}^n Q_i/n$ for some random variable Q, and let $\mathbb{E}^*[.]$ denote a mean over the bootstrap distribution.

The algorithm for the construction of critical values for the heterogeneity test is:

Algorithm B1

1. Generate a bootstrap sample $\{Y_i^*, p_i^*\}_i$
2. Construct the test statistic
$$W = \max_s \frac{\sqrt{n}|\gamma^*(s) - \widehat{\gamma(s)}|}{\sqrt{\widehat{V(s)}}}$$
3. Calculate $c_n(\alpha)$ as $1-\alpha$ quantile of W

where $\gamma^*(s) = \mathbb{E}^*[M_s^*] = \mathbb{E}^*[M(Y^*,p^*,s)]$, i.e. bootstrap weighted projection quadratic coefficient.

Similarly, the algorithm for the construction of critical values for the test of the monotone decreasing MTE is:

Algorithm B2

1. Generate a bootstrap sample $\{Y_i^*, p_i^*\}_i$
2. Construct the test statistic

$$W^m = \max_s \frac{\sqrt{n}(\gamma^*(s) - \widehat{\gamma(s)})}{\sqrt{\widehat{V(s)}}}$$

3. Calculate $c_n^m(\alpha)$ as $1-\alpha$ quantile of W^m

For the proof of the validity of the proposed bootstrap algorithms, I follow Chernozhukov et al. (2019) and use the standardized moment function $\tilde{M}_s$:

$$\tilde{M}_s = \frac{M_s - \gamma(s)}{\sqrt{V(s)}}.$$

Notice that $\tilde{M}_s$ is i.i.d. over i, $\mathbb{E}[\tilde{M}_s] = 0$, $\mathrm{Var}(\tilde{M}_s) = 1$. Let $A_{n,s} = \max_s(\mathbb{E}[|\tilde{M}_{1,s}|^k])^{1/k}$, $B_n = (\mathbb{E}[\max_s \tilde{M}_{1,s}^4])^{1/4}$. Intuitively, $A_{n,s}$ and B_n contain information about heavy-tailedness of the distribution of the standardized moment function $\tilde{M}$.

Proposition 3. (Validity of B1 and B2 methods) Suppose that there exist constants $c_1 \in (0, 1/2)$ and $c_2 \geq 0$ such that

$$(\max\{A_{n,3}^3, A_{n,4}^2, B_n\})^2 \log^{7/2}(|\mathcal{S}|n) \leq c_2 n^{1/2-c_1}.$$

Then there exists $c_3 > 0$ and $c_4 > 0$ such that under $\mathbb{H}_0$,

$$\Pr\{T_n > c_n(\alpha)\} \leq \alpha + c_3 n^{-c_4}$$

and there exists $c_5 > 0$ and $c_6 > 0$ such that under $\mathbb{H}_0^m$,

$$\Pr\{T_n^m > c_n^m(\alpha)\} \leq \alpha + c_5 n^{-c_6}.$$

This bound holds uniformly over all distributions of data (Y, p) such that $(\max\{A_{n,3}^3, A_{n,4}^2, B_n\})^2 \log^{7/2}(|\mathcal{S}|n) \leq c_2 n^{1/2-c_1}$, $\mathbb{E}[M_s^2] < \infty$ and $\mathrm{Var}(M_s) > 0$ for all s.

Proof. See in Appendix.

Proposition 3 shows that bootstrap procedures B1 and B2 provide critical values that asymptotically control size α (i.e. the probability of the incorrect rejection of the null $\mathbb{H}_0$ or $\mathbb{H}_0^m$) uniformly over different distributions of data.

Remark 4. (Connection to local quadratic regression) The algorithms of my adaptive tests require the construction of kernel-weighted quadratic projections; while technically similar to local quadratic regressions, there are important differences. First, there is a fundamental difference in motivation. Coefficients of local quadratic regression are usually interpreted as "regression derivative *at a point*", and this interpretation is based on the Taylor theorem ("function in the neighborhood of some point is well approximated by a polynomial"); my result, in contrast, provides the interpretation of an *average* regression derivative. Instead of the regression derivative identification at every point, I focus on the *average* regression derivative identification. It turns out that for my tests this *average* derivative interpretation is sufficient. Second, and related to the previous point, bandwidth plays different roles in my weighted projections and in local quadratic regressions. For the latter, a bandwidth is required to be small so that one can invoke the Taylor theorem and fill-in asymptotics. For the former, a bandwidth is a tool for adaptivity towards various alternatives ("smoothness parameter"), and it is not required to be asymptotically small.

Remark 5. (Alternative testing procedures) The test of $m''(p) = 0$ (or $m''(p) \leq 0$) is essentially a *specification test*. There are a number of approaches how one can conduct it, and these approaches have different power properties. For concreteness, in this paper, I follow Chernozhukov et al. (2019) and use the maximum of moment conditions as the test statistic T_n (or T_n^m). The procedure of Chernozhukov et al. (2019) is minimax optimal with respect to the ℓ^∞-norm. Alternatively, one can use the procedure of Horowitz and Spokoiny (2001), which is minimax optimal with respect to the ℓ^2-norm. Different testing procedures then have different power properties, and their performance depends on the alternative (namely, the distance between the null and the alternative).

Remark 6. (IV-like estimands) From (27) and (28), $M_s = M(p, Y, s) = \mathcal{M}_s(p)Y$ for some function $\mathcal{M}_s$, hence the kernel-weighted projection estimand $\mathbb{E}[M_s]$ is an *IV-like estimand* (Mogstad et al., 2018) for every s. Mogstad et al. (2018) use the MTE analysis in partially identified models to construct an identified set for causal parameters of interest (e.g. ATE). In particular, they use multiple IV-like estimands to identify the set of structural functions that are compatible with the data, and use this set to bound causal parameters of interest under (average) treatment effect heterogeneity. While in this paper I also use multiple IV-like estimands, the goal here is different – instead of targeting the identified set for some average causal effect, I want to test for treatment effect heterogeneity, i.e. for the equality of various average causal effects. In this sense, my paper is complimentary to the paper of Mogstad et al. (2018): one can use my test in order to check if LATE is actually policy relevant, and if it is rejected by my test, one way to proceed is to construct (maybe using the same IV-like estimands – weighted projections) the identified set for the average causal effect of interest.

1.4.4 Implications and extensions

1.4.4.1 Implications for optimal welfare

As discussed in Section 2.4, constant MTE and monotone decreasing MTE provide simple guidance to optimal policy assignment. The test in (23) is a simple and powerful procedure that provides evidence in favor of or against the null of constant MTE. If it does not reject the null of no selection on gains, the optimal policy is given either by $\arg\max_z p(z)$ or $\arg\min_z p(z)$ in line with the discussion in Section 2.4. In this case, it is left to determine which of two candidate policies should be chosen. Figure 1 suggests that it depends on whether the MTE is positive or negative, and it can be easily checked. If the null of constant MTE is not rejected, one can run a linear regression of Y on p (which is equivalent to constant MTE by (3)), and t-test test the slope coefficient (i.e. MTE) $\beta \geq 0$ vs $\beta < 0$; if no rejection, the MTE is non-negative for all margins of participation u, and the optimal policy is $\arg\max_z p(z)$ – induce as many as possible into treatment; if rejection, the MTE is negative for all u, optimal policy is $\arg\max_z p(z)$ – induce as many as possible out of treatment.

If constant MTE is unlikely to hold in a particular application, there is still a possibility of a simple optimal policy if MTE is monotone decreasing, which can be checked by (25). If the test rejects, optimal policy assignment would require estimation of MTE and explicitly solving the welfare optimization problem (Liu, 2022). If the test does not reject, it is necessary to find the intersection of the MTE curve with the u-axis, u^*, and set z^* such that $p(z^*) = u^*$. (3) implies that the intersection of the MTE curve with the u-axis is given by $m'(p) = 0$, which can also be found by local linear regression.

1.4.4.2 Redirecting power

Sometimes a researcher might be interested in a particular average treatment effect (e.g. a policy relevant treatment effect (see Heckman and Vytlacil, 2001 and Heckman and Vytlacil, 2005), i.e. not the whole MTE curve but a fragment of it on a certain proper subinterval $[u_1, u_2]$. In this case, it might be desirable to modify T_n in (23) in order to make a test more powerful against deviations of the null on $u \in [u_1, u_2]$, possibly at the expense of power against other alternatives. Namely, let $\Omega(\mu, h) = 1$ for every bandwidth h and location parameters $\mu \in [u_1, u_2] \cap \{p_1, ..., p_n\}$, and $\Omega(\mu, h) = 0$ for $\mu \notin [u_1, u_2] \cap \{p_1, ..., p_n\}$; then

$$T_n(\Omega) = \max_{s \in \mathcal{S}} \sqrt{n}\Omega(s)|\widehat{\gamma(s)}|/\sqrt{\widehat{V(s)}}.$$

Under the decision rule similar to (24) – reject if $T_n(\Omega) > c_n^r$ – the critical values c_n^r can be constructed similarly to Algorithm B1 (see Appendix D for the algorithm).

1.4.5 Detection of subpopulations with nontrivial heterogeneity

Finally, I propose a procedure designed for detection of observable subpopulations (defined by covariates) with nontrivial heterogeneity. The heterogeneity test in Section 4.1-4.3 is implicitly conditional on covariates; however, covariates can provide the foundation for covariate-specific, and hence individualized, optimal incentive assignment (more efficient than unconditional common policy). Further, in the case where the null of no heterogeneity $\mathbb{H}_0$ is rejected for the whole sample, maybe there are subpopulations for which the null of no heterogeneity is true, and (in line with the discussion in Section 2.3) conditional LATE for such subpopulations is indeed policy relevant.

I propose to apply the heterogeneity test developed in Sections 4.1-4.3 for each value of covariates x, collect bootstrap p-values $\rho(x)$ for each x, and reject the null of no heterogeneity for subpopulations with

$$\rho(x) < c^{MHT}. \tag{29}$$

The decision rule in (29) involves running a heterogeneity test for each value of the covariates (e.g. if the only covariate is race, then it involves a separate test for each race). If there are multiple distinct values of covariates (hence multiple tests conducted), there is a problem of multiple hypotheses testing, which is especially important when the number of such values (and tests) is large. Ignoring multiplicity of hypotheses leads to multiple incorrect inferences.

One way to mitigate the problems of multiple hypotheses testing is to construct c^{MHT} in such a way that controls the false discovery rate (FDR), i.e. the expected proportion of false negatives (Efron, 2007). Compared to the alternative (i.e. familywise error rate), FDR is less conservative, and is better suited for detection of non-trivial effects. I propose to construct c^{MHT} by applying the sequential Bonferroni-type approach of Benjamini and Hochberg (1995). One of the most important requirements of their approach is that test statistics are independent. In my setting, this requirement is satisfied automatically because test statistics are computed using different subsamples, defined by non-overlapping covariate cells. The details of the procedure are given below.

The algorithm for the detection of observed subpopulations with nontrivial treatment effect heterogeneity is:

Detection of non-trivial heterogeneity algorithm

1. Let $\mathcal{X}$ be the support of covariates X, i.e. $X \in \mathcal{X}$, and let $\mathcal{X}$ be finite
2. For every $x^* \in \mathcal{X}$, apply the heterogeneity test from Sections 4.1-4.3; collect p-value $\rho(x^*)$
3. Obtain c^{MHT} by Benjamini and Hochberg (1995) procedure:
   - 3.1 Order p-values $\rho(x^*)$: $\rho(x_1^*) \leq ... \leq \rho(x_{|\mathcal{X}|}^*)$, where $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$
   - 3.2 Let k^* be the largest i for which $\rho(x_i^*) \leq \frac{i}{|\mathcal{X}|}\overline{\text{FDR}}$, where $\overline{\text{FDR}}$ is the prespecified level at which FDR is to be controlled
   - 3.3 Set $c^{MHT} = \frac{k^*}{|\mathcal{X}|}\overline{\text{FDR}}$
4. Apply decision rule (29) and find the subset $\mathcal{X}_A \subset \mathcal{X}$, for which (29) rejects

Based on this procedure, one can argue the policy relevance of conditional LATEs (see Section 2.3) and construct optimal incentive assignments (see Section 2.4) conditional on observed characteristics.

1.5 Feasible test

So far in the paper, I have treated treatment probabilities $p = p(z)$ as observed. However, usually they stem from individual choice decisions and are unobservable, and in this sense the test is infeasible. In this case, one can set apply a two-step procedure, in which p is estimated using available instrumental variables Z. Then, in the second stage one would use estimated $\hat{p}$ and $\hat{p}^2$ instead of p, obtain the feasible estimate $\widehat{\gamma(s)}_F$ and run the tests proposed in Section 4 with this feasible $\widehat{\gamma(s)}_F$. Proposition 4 provides high-level conditions under which a non-parametric estimate of p yields the feasible projection coefficient that has the same probability limit as the feasible projection coefficient, which uses the true p.

Proposition 4. (Feasible estimator converges to infeasible estimand) Let $p_0 \equiv p_0(z) = \mathbb{E}[D|Z = z]$ be the true regression function; $D \in \{0, 1\}$. Let $\mathcal{P}_n$ be the set of estimators of p_0. Denote

$$(\alpha(s), \beta(s), \gamma(s)) = \underset{a(s),b(s),c(s)}{\arg\min}\ \mathbb{E}[K(p_0, s)\left(Y - a(s) - b(s)p_0 - c(s)p_0^2\right)^2] \tag{30}$$

$$(\hat{\alpha}_F(s), \hat{\beta}_F(s), \hat{\gamma}_F(s)) = \underset{a(s),b(s),c(s)}{\arg\min}\ \mathbb{E}_n[K(\hat{p}, s)\left(Y - a(s) - b(s)\hat{p} - c(s)\hat{p}^2\right)^2] \tag{31}$$

for some $\hat{p} \in \mathcal{P}_n$. Suppose for some constant C

$$(i)\ \sup_{p\in\mathcal{P}_n} \mathbb{E}[p^{2k}] \leq C \text{ for } k \in \{1, 2, 3, 4\}$$

$$(ii)\ \sup_{p\in\mathcal{P}_n} \left(\mathbb{E}[(p - p_0)^2]\right)^{1/2} \to 0$$

$$(iii)\ \sup_{p\in\mathcal{P}_n} \mathbb{E}[Y^2 p^{2k}] \leq C \text{ for } k \in \{1, 2, 3, 4\}$$

$$(iv)\ (\mathbb{E}[Y^2])^{1/2} \leq C.$$

Then $(\hat{\alpha}_F(s), \hat{\beta}_F(s), \hat{\gamma}_F(s)) \xrightarrow{p} (\alpha(s), \beta(s), \gamma(s))$.

Proof. See Appendix.

The conditions of Proposition 4 impose restrictions on the estimator of p and on the moments of Y. (i) requires the powers of the estimator to be bounded, (ii) requires the estimator to concentrate around true p, and (iii) and (iv) impose restrictions on joint moments of the estimator and the outcome Y.

The most crucial condition is (ii), which essentially requires that the mean squared error converges to 0. While it is possible to guarantee for $p(z)$ that depends on a small-dimensional parameter (e.g. with finite support of p, i.e. finite support of the instrumental variable Z). Alternatively, when condition (ii) is not satisfied, one can apply a sample-splitting algorithm that might help with non-trivial mean squared error.

As a simple example, consider first a feasible version of the linear projection in (5):

$$Y_i = \alpha_{L,F} + \beta_{L,F}\hat{p}_i + e_i^F, \tag{32}$$

where $\hat{p}_i$ are estimated in the first step with some approximation error u_i. In general, this approximation error is correlated with the projection residuals e_i^F: $\mathbb{E}[u_i e_i^F] \neq 0$. In this case, $\beta_{L,F}$ in (32) can be different from infeasible estimand β_L in (5). One can deal with this problem by using one half of the sample to estimate $\hat{p}_i = \hat{p}(Z_i)$ and impute them to the other half of the sample based on observable Z. Then one would run (32) for the other half of the sample; it is straightforward that in this case the first step estimation noise is uncorrelated with the projection residuals as the first step was estimated on a different dataset than the second step.

This sampling-splitting idea can be applied to the quadratic projection as well. Let the sample be split into four batches. Use three batches to estimate $\hat{p}(Z_i)$ by projecting D on Z. Let three estimators be $\hat{p}^{(1)}(Z_i)$, $\hat{p}^{(2)}(Z_i)$ and $\hat{p}^{(3)}(Z_i)$. All three use different batches of data, so they are independent estimators of p_i. Use the fourth batch to run the projection

$$Y_i = \alpha_F + \beta_F \hat{p}_i + \gamma_F \hat{p}_i^2 + \varepsilon_i^F.$$

Apply the first estimator to impute $\hat{p}_i = \hat{p}^{(1)}(Z_i)$; apply the second and third estimators to impute $\widehat{p^2}_i = \hat{p}^{(2)}(Z_i)\hat{p}^{(3)}(Z_i)$. Finally, run a projection of outcome Y on imputed $\hat{p}_i$ and $\widehat{p^2}_i$ (and a constant), using the fourth batch of data. In this case, first step estimation errors do not correlate with each other or with the projection residual ε_i^F because different subsamples are used.

1.6 Simulations

In this section, I illustrate the performance of my proposed heterogeneity test by a simulation study. I use a data generating process calibrated to Walters (2018) [7]. In his paper, Walters (2018) studies the effects of attending a charter school, and pays special attention to the issues of selection. For the sake of tractability, I focus on the population of applicants to charter schools (hence I treat the application decisions as given) that have received an offer from a charter school. For this population, I model the treatment (charter school attendance) selection mechanism as

$$D_i(Z_i) = \mathbb{1}\{\lambda \tau_i + (1-\lambda)\xi_i - Z_i \geq 0\} \tag{33}$$

where τ_i is the individual treatment effect that combines functional form assumptions and parameter estimates from Walters (2018) with calibrated data, ξ_i is an individual preference shock, and instrumental variable Z_i is the differential distance to the closest charter school versus the closest public school. According to treatment assignment model (33), i selects into treatment if the combination of treatment effect τ_i and preference shock ξ_i offsets the disutility of distance to a charter school Z_i.

Parameter λ governs selection on gains, and under constant MTE, $\lambda = 0$, i.e. the treatment effect τ_i does not influence the selection into treatment. By shifting λ, I model the selection on treatment effects: for $\lambda = 0$, individuals select into treatment independent of their treatment effect τ_i, purely based only on their individual shock ξ_i and the disutility of the

[7] See Appendix C for details on the simulation design

differential distance to a charter school Z_i. As λ increases, the impact of τ_i on the selection decision increases. Intuitively, for higher λ, my proposed test should be more powerful – it is easier to detect the violations of the null hypothesis of no heterogeneity $\mathbb{H}_0$ if the impact of treatment effect τ_i on the selection decision is stronger.

The resulting dataset contains 1600 observations. The IV is a continuous variable (distance to school), which provides 1600 distinct values of (estimated) treatment probabilities. Thus, with a "default" set of bandwidths $\{0.1, 0.2, 0.3\}$, the tests proposed in Section 4 require estimation of $|\mathcal{S}| = 1600 \times 3 = 4800$ weighted projections. Even though the computation time is not restrictive, for this simulation study I need to repeat the test multiple times, so I reduce the dimension of $|\mathcal{S}|$. In particular, instead of using all observed $\hat{p}$ as location parameters μ, I use only certain quantiles of $\hat{p}$ as μ. Different rows of Table 2 represent different choices of quantiles: row 1 uses 10 10% quantiles, row 2 uses 20 5% quantiles, row 3 uses 40 2.5% quantiles.

Table 2 shows the actual rejection rates of the heterogeneity test. Each column represents a distinct value of λ, i.e. the degree of the deviation from the null of no heterogeneity. For $\lambda = 0$ (Column 1), there is no selection on gains, the null is true, hence the test should reject in 5% of all such datasets. In line with this argument, the figures in Column 1 are close to 0.05 for all values of $|\mathcal{S}|$, which indicates the correct size control (in line with Proposition 3).

Columns 2-6 represent various deviations from the null hypothesis of no heterogeneity; higher λ corresponds to stronger such deviations, and it is reasonable to expect that stronger deviations are more easily detectable by the test. Indeed, as λ increases, the test rejects more often; hence it becomes more powerful.

Columns 2 and 3 indicate low power against alternatives that are close to the null hypothesis. One possible reason is that by decreasing the number of possible location points μ, we reduce the simulation intensity and also reduce the amount of information we use for the construction of the test statistic. At the same time, increasing the number of location parameters does not improve power. One possible reason is that increasing the number of evaluated test functions (i.e. increasing $|\mathcal{S}|$) that are not far from the null mechanically increases the critical value but does not affect the test statistic. Under small to moderate deviations from the null, adding moment conditions to the evaluated set $\mathcal{S}$ might be unhelpful, which is indicated by the small differences between the lines of Table 2.

	$\lambda = 0.00$	$\lambda = 0.20$	$\lambda = 0.40$	$\lambda = 0.60$	$\lambda = 0.80$	$\lambda = 1.00$
$\lvert\mathcal{S}\rvert = 30$	0.06	0.11	0.19	0.56	0.97	1.00
$\lvert\mathcal{S}\rvert = 60$	0.06	0.16	0.17	0.46	0.98	1.00
$\lvert\mathcal{S}\rvert = 120$	0.06	0.09	0.10	0.48	0.96	1.00

Table 1: Actual rejection rates

1.7 Application

In this Section, I apply the tests proposed in Section 4 to the data of Ashraf et al. (2010). In their paper, Ashraf et al. (2010) study the effect of prices on product use, namely a home water purification solution. Ashraf et al. (2010) run a field experiment, in which prices of the solution were randomized among households. For the purpose of illustration, I choose the attitude towards water purification as the outcome of interest. LATE of buying a solution on the attitude towards water purification is identified by two-stage least squares.

For every point of observed p, and $h \in \{0.1, 0.2, 0.3\}$, I construct $\widehat{b(s)}$ and corresponding $\widehat{V(s)}$. All $\widehat{b(s)}$ are negative, which suggests a decreasing MTE curve, consistent with the idea of people with low U_i having higher treatment effects on average. The test statistic is $T_n = -1.36$, and the 95th quantile of the bootstrap distribution is $c_n^m(0.05) = 2.46$, hence the test does not reject monotone decreasing MTE.

1.8 Conclusion

In this paper I propose a test of treatment effect heterogeneity in IV models and a test of monotone decreasing MTE curve. First, I show a novel statistical result on the identification of an average second regression derivative via a quadratic projection. I further show that this result extends to the case of weighted projections. Second, I show how my statistical result helps to identify the average slope of the MTE curve. Third, I propose two separate tests of constant MTE and monotone decreasing MTE that are based on the many moment inequalities literature. These tests have implications for the interpretation of causal parameters, in particular on the external validity of LATE, and for welfare maximization. Finally, I suggest a multiple hypothesis testing procedure that aims at the detection of observable subpopulations with nontrivial heterogeneity, and illustrate the tests with calibrated simulations and real data.

1.A Appendix: auxiliary results

Lemma A1. Suppose Assumption 1 is satisfied. Then $\mathbb{E}[Y|Z,X] = \mathbb{E}[Y|p(Z,X),X]$.

Proof.

$$\begin{aligned}
\mathbb{E}[Y|Z,X] &= \mathbb{E}[DY(1)|Z,X] + \mathbb{E}[(1-D)Y(0)|Z,X] \\
&= \mathbb{E}[\mathbb{E}[DY(1)|D,Z,X]|Z,X] + \mathbb{E}[\mathbb{E}[(1-D)Y(0)|D,Z,X]|Z,X] \\
&= \mathbb{E}[\mathbb{E}[Y(1)|D=1,Z,X]\mathbb{1}_{\{D=1\}}|Z,X] \\
&\quad + \mathbb{E}[\mathbb{E}[Y(0)|D=0,Z,X]\mathbb{1}_{\{D=0\}}|Z,X] \\
&= \Pr\{D=1|Z,X\}\mathbb{E}[\mathbb{E}[Y(1)|D=1,Z,X]|Z,X] \\
&\quad + \Pr\{D=0|Z,X\}\mathbb{E}[\mathbb{E}[Y(0)|D=0,Z,X]|Z,X] \\
&= p(Z,X)\mathbb{E}[\mathbb{E}[Y(1)|D=1,Z,X]|Z,X] \\
&\quad + (1-p(Z,X))\mathbb{E}[\mathbb{E}[Y(0)|D=0,Z,X]|Z,X],
\end{aligned}$$

where the first equality follows from the definition of observed Y in (1), the second equality follows by LIME, the third equality follows from D being binary, the fourth equality follows from the pull-out property, and the fifth equality follows from the definition of propensity score $p(Z, X)$. Further, notice that event $\{D = 1\}$ is equivalent to the event $\{p(Z, X) \geq U\}$ by Assumption 1.1, and similarly $\{D = 0\}$ is equivalent to $\{p(Z, X) \leq U\}$. Thus,

$$\mathbb{E}[Y(1)|D = 1, Z, X] = \mathbb{E}[Y(1)|U \leq p(Z, X), Z, X]$$

and

$$\mathbb{E}[Y(0)|D = 0, Z, X] = \mathbb{E}[Y(0)|U \geq p(Z, X), Z, X].$$

Combine with Assumption 1.3 to obtain

$$\mathbb{E}[Y(1)|D = 1, Z, X] = \mathbb{E}[Y(1)|U \leq p(Z, X), X]$$

and

$$\mathbb{E}[Y(0)|D = 0, Z, X] = \mathbb{E}[Y(0)|U \geq p(Z, X), X].$$

Finally, notice that the inner expectation depends on Z only through $p(Z, X)$, then by the pull-out property,

$$\mathbb{E}[\mathbb{E}[Y(1)|U \leq p(Z, X), X]|Z, X] = \mathbb{E}[\mathbb{E}[Y(1)|U \leq p(Z, X), X]|p(Z, X), X]$$

and

$$\mathbb{E}[\mathbb{E}[Y(0)|U \geq p(Z, X), X]|Z, X] = \mathbb{E}[\mathbb{E}[Y(0)|U \geq p(Z, X), X]|p(Z, X), X].$$

Combine all the above to obtain

$$\begin{aligned}
\mathbb{E}[Y|Z, X] &= p(Z, X)\mathbb{E}[\mathbb{E}[Y(1)|D = 1, Z, X]|Z, X] \\
&\quad + (1 - p(Z, X))\mathbb{E}[\mathbb{E}[Y(0)|D = 0, Z, X]|Z, X] \\
&= p(Z, X)\mathbb{E}[\mathbb{E}[Y(1)|U \leq p(Z, X), X]|p(Z, X), X] \\
&\quad + (1 - p(Z, X))\mathbb{E}[\mathbb{E}[Y(0)|U \geq p(Z, X), X]|p(Z, X), X].
\end{aligned}$$

Now consider $\mathbb{E}[Y|p(Z, X), X]$. Similarly to the argument for $\mathbb{E}[Y|Z, X]$ at the beginning of the proof and using the fact that (Z, X) is finer than $(p(Z, X), X)$,

$$\begin{aligned}
\mathbb{E}[Y|p(Z, X), X] &= \mathbb{E}[DY(1)|p(Z, X), X] + \mathbb{E}[(1 - D)Y(0)|p(Z, X), X] \\
&= \mathbb{E}[\mathbb{E}[DY(1)|D, p(Z, X), Z, X]|p(Z, X), X] + \mathbb{E}[\mathbb{E}[(1 - D)Y(0)|D, p(Z, X), Z, X]|p(Z, X) \\
&= \mathbb{E}[\mathbb{E}[DY(1)|D, Z, X]|p(Z, X), X] + \mathbb{E}[\mathbb{E}[(1 - D)Y(0)|D, Z, X]|p(Z, X), X] \\
&= \mathbb{E}[\mathbb{E}[Y(1)|D = 1, Z, X]\mathbb{1}_{\{D=1\}}|p(Z, X), X] \\
&\quad + \mathbb{E}[\mathbb{E}[Y(0)|D = 0, Z, X]\mathbb{1}_{\{D=0\}}|p(Z, X), X] \\
&= p(Z, X)\mathbb{E}[\mathbb{E}[Y(1)|D = 1, Z, X]|p(Z, X), X] \\
&\quad + (1 - p(Z, X))\mathbb{E}[\mathbb{E}[Y(0)|D = 0, Z, X]|p(Z, X), X] \\
&= p(Z, X)\mathbb{E}[\mathbb{E}[Y(1)|U \leq p(Z, X), X]|p(Z, X), X] \\
&\quad + (1 - p(Z, X))\mathbb{E}[\mathbb{E}[Y(0)|U \geq p(Z, X), X]|p(Z, X), X],
\end{aligned}$$

which provides the conclusion. $\square$

Lemma A2. Assume conditions of Proposition 1 are satisfied. Then

$$\beta = \int_0^1 m'(t)\omega_\beta(t)dt.$$

Proof. By the fundamental theorem of calculus, the reduced-form regression is

$$m(p) = \lim_{p\to 0} m(p) + \int_0^p m'(t)dt.$$

Plug this expression into (11):

$$\begin{aligned}
\beta &= \frac{\mathbb{E}[m(p_i)(p_i - \mathbb{L}[p_i|p_i^2])]}{\operatorname{Var}(p_i - \mathbb{L}[p_i|p_i^2])} \\
&= \frac{\mathbb{E}[(\lim_{p\to 0} m(p) + \int_0^{p_i} m'(t)dt)\,(p_i - \mathbb{L}[p_i|p_i^2])]}{\operatorname{Var}(p_i - \mathbb{L}[p_i|p_i^2])} \\
&= \frac{\mathbb{E}[\int_0^{p_i} m'(t)dt(p_i - \mathbb{L}[p_i|p_i^2])]}{\operatorname{Var}(p_i - \mathbb{L}[p_i|p_i^2])} \\
&= \int_0^1 m'(t)\frac{\mathbb{E}[p - \mathbb{L}[p|p^2]|p \geq t]\Pr\{p \geq t\}}{\operatorname{Var}(p_i - \mathbb{L}[p_i|p_i^2])}dt.
\end{aligned}$$

$\square$

Lemma A3. For some non-degenerate random variable X, let $b = \frac{\operatorname{Cov}(X^2,X)}{\operatorname{Var}(X)}$ and $a = \mathbb{E}[X^2] - b\mathbb{E}[X]$. Then X^2 and a linear projection $\mathbb{E}^*[X^2|X] \equiv a + bX$ have two intersection points.

Proof. First, I consider unbounded X, i.e. $\operatorname{Supp}(X) = (-\infty, \infty)$, and let $x \in \operatorname{Supp}(X)$ be its generic element. I show that $\exists\ x_0 \in \operatorname{Supp}(X)$ such that $x_0^2 < a + bx_0$ is equivalent to x^2 and $a + bx$ having two intersection points. Then I show that $x^2 \geq a + bx\ \forall\ x \in \operatorname{Supp}(X)$ contradicts the projection FOC.

Suppose $\operatorname{Supp}(X) = (-\infty, \infty)$ and $\exists\ x_0 \in \operatorname{Supp}(X)$ such that $x_0^2 < a + bx_0$. Consider $\ell(x) = x^2 - (a+bx)$. Notice that $\ell(x)$ is a polynomial and hence a continuous function. Further, $\ell(x_0) < 0$ and $\lim_{x\to\infty} \ell(x) = \lim_{x\to -\infty} \ell(x) = \infty$. Then $\exists\ x_-$ such that $\ell(x) > 0 \forall\ x < x_-$ and $\exists x_+$ such that $\ell(x) > 0 \forall\ x > x_+$. By the Intermediate Value Theorem, $\exists\ \underline{x} \in (x_-, x_0)$ such that $\ell(\underline{x}) = 0 = \underline{x}^2 - (a + b\underline{x})$. Similarly, $\exists\ \bar{x} \in (x_0, x_+)$ such that $\ell(\bar{x}) = 0 = \bar{x}^2 - (a + b\bar{x})$. Finally, notice that $\ell(x)$ is a quadratic polynomial that has at most two real roots, hence there are at most two intersection points of x^2 and $a + bx$. Therefore, $\exists\ x_0 \in \operatorname{Supp}(X)$ such that $x_0^2 < a + bx_0$ implies that there are two intersection points of x^2 and $a + bx$.

Now I show that the converse is true. Suppose $\operatorname{Supp}(X) = (-\infty, \infty)$ and $\exists\ (x_-, x_+) \in \operatorname{Supp}(X)$ such that $x_- < x_+$, $x_-^2 = a + bx_-$ and $x_+^2 = a + bx_+$. x^2 is a strictly convex function of x, then the epigraph of x^2, i.e. $\chi \equiv \{(x, x^2) | x^2 \geq a + bx\}$, is strictly convex. $(x_-, x_-^2) \in \chi$, $(x_+, x_+^2) \in \chi$ and $x_- \neq x_+$ then there exists $x_0 \in (x_-, x_+)$ and, by the definition of a strictly convex set, $(x_0, x_0^2) \in \text{int}(\chi)$, where $\text{int}(\chi)$ is an interior of χ. Equivalently, $\exists\ x_0 \in \text{Supp}(X)$ such that $x_0^2 > a + bx_0$. Combining all above, for $\text{Supp}(X) = (-\infty, \infty)$,

$$\exists\ x_0 \in \text{Supp}(X) \text{ such that } x_0^2 < a+bx_0 \Leftrightarrow \text{there are two intersection points of } x^2 \text{ and } a+bx.$$

Now I show that the above is implied by the projection FOC. Suppose towards contradiction that $x^2 \leq a + bx\ \forall\ x \in \text{Supp}(X)$. Then there is at most one point x_0 such that $x_0 = a + bx_0$ (if there were more than one, their linear combination would be in the epigraph of x^2, violating $x^2 \leq a + bx\ \forall\ x \in \text{Supp}(X)$). Then by $\text{Supp}(X) = (-\infty, \infty)$, $\mathbb{E}[x^2 - a + bx] < 0$. However, by the projection FOC, $\mathbb{E}[x^2 - (a + bx)] = 0$, contradiction. Therefore, the projection FOC implies the existence of two intersection points of x^2 and the linear projection $a + bx$.

Second, let the support of X be compact: $\text{Supp}(X) = [L_X, U_X]$. The Intermediate Value Theorem is impossible to invoke straightaway. Instead, notice that for $x_1, x_2 \in \text{Supp}(X)$

$$\begin{aligned}
\text{Var}(X) &= \frac{1}{2}\mathbb{E}[(x_1 - x_2)^2], \\
\text{Var}(X^2) &= \frac{1}{2}\mathbb{E}[(x_1^2 - x_2^2)^2], \\
\text{Cov}(X, X^2) &= \frac{1}{2}\mathbb{E}[(x_1 - x_2)(x_1^2 - x_2^2)], \\
\text{Cov}(X, X^3) &= \frac{1}{2}\mathbb{E}[(x_1 - x_2)(x_1^3 - x_2^3)]
\end{aligned}$$

because Var() and Cov() are second-order U-statistics (Vaart, 1998, Chapter 12, **hoeffding1948**). Below, I show that $U_X^2 > a + bU_X$ and $L_X^2 > a + bL_X$, which implies that there are two intersection points of x^2 and $a + bx$.

Notice that for any $x \in \text{Supp}(X)$ and using the U-statistic representation

$$\begin{aligned}
a + bx - x^2 = {} & \mathbb{E}[X^2] - x^2 + \frac{\text{Cov}(X, X^2)}{\text{Var}(X)}(x - \mathbb{E}[X]) \\
\propto {} & -\text{Var}(X)(x^2 - \mathbb{E}[X^2]) + \text{Cov}(X, X^2)(x - \mathbb{E}[X]) \\
= {} & -\text{Var}(X^2) - x^2\text{Var}(X) + x\text{Cov}(X, X^2) + \text{Cov}(X, X^3) \\
= {} & -\frac{1}{2}\mathbb{E}\left[\left(x_1^2 - x_2^2\right)^2 + x^2\left(x_1 - x_2\right)^2 - x\left(x_1 - x_2\right)\left(x_1^2 - x_2^2\right) - \left(x_1 - x_2\right)\left(x_1^3 - x_2^3\right)\right] \\
= {} & -\frac{1}{2}\mathbb{E}\left[\left(x_1 - x_2\right)^2\left(\left(x_1 + x_2\right)^2 + x^2 - x\left(x_1 + x_2\right) - \left(x_1^2 + x_1x_2 + x_2^2\right)\right)\right] \\
= {} & -\frac{1}{2}\mathbb{E}\left[\left(x_1 - x_2\right)^2\left(x^2 - x\left(x_1 + x_2\right) + x_1x_2\right)\right] \\
= {} & -\frac{1}{2}\mathbb{E}\left[\left(x_1 - x_2\right)^2\left(x - x_1\right)\left(x - x_2\right)\right].
\end{aligned}$$

$\left(x_1 - x_2\right)^2 \geq 0 \forall x_1, x_2 \in \text{Supp}(X)$. Further, $(U_X - x) \geq 0 \forall x \in \text{Supp}(X)$ and $(L_X - x) \leq 0 \forall x \in \text{Supp}(X)$; then $\left(U_X - x_1\right)\left(U_X - x_2\right) \geq 0 \forall x, x_2 \in \text{Supp}(X)$ and $\left(L_X - x_1\right)\left(L_X - x_2\right) \geq$

$0 \forall x, x_2 \in \text{Supp}(X)$. Then

$$a + bU_X - U_X^2 \leq 0, \tag{34}$$
$$a + bL_X - L_X^2 \leq 0, \tag{35}$$

i.e. x^2 is larger than the projection line $a + bp$ at the endpoints of the bounded $\text{Supp}(X)$. Now I show that there are points in $\text{Supp}(X)$ where x^2 is *strictly smaller* than the projection line $a + bp$ that implies that there are two intersection points of x^2 and $a + bp$, which would conclude the proof for the bounded $\text{Supp}(X)$. Suppose towards contradiction that such that $a + bx - x^2 \leq 0 \forall x \in \text{Supp}(X)$. Then $\mathbb{E}[a + bx - x^2] < 0$ because X is not degenerate. The FOC of the projection problem is $\mathbb{E}[x^2 - (a + bx)] = 0$ – contradiction. Then $\exists x_0 \in \text{Supp}(X)$ such that $a + bx_0 > x_0^2$. Then either the endpoints (L_X, U_X) are the intersection points of x^2 and $a + bx$ if (34) and (35) hold with equality. If (34) is strict and (35) is equality, then by the Intermediate Value Theorem, $\exists \bar{x} \in (x_0, U_X)$ such that $a + b\bar{x} - \bar{x}^2 = 0$. If (34) is equality and (35) is strict, then by the Intermediate Value Theorem, $\exists \underline{x} \in (L_X, x_0)$ such that $a + b\underline{x} - \underline{x}^2 = 0$. Finally, if both (34) and (35) hold with strict inequality, $\exists \bar{x} \in (x_0, U_X)$ such that $a + b\bar{x} - \bar{x}^2 = 0$ and $\exists \underline{x} \in (L_X, x_0)$ such that $a + b\underline{x} - \underline{x}^2 = 0$.

□

Lemma A4. Assume conditions of Proposition 1 are satisfied. Then

$$\omega(t) = \frac{\mathbb{E}\left[\widetilde{p^2}(p - t)|p \geq t\right] \Pr\{p \geq t\}}{\mathbb{E}[p^2 \widetilde{p^2}]} \geq 0 \quad a.s.$$

Proof. From Proposition 1, the weights are

$$\omega(t) = \frac{1}{\mathbb{E}[p^2 \widetilde{p^2}]} \mathbb{E}[(p^2 - (a + bp))(p - t)|p \geq t] \Pr\{p \geq t\},$$

where (a, b) are the coefficients of projection of p^2 on p. For $t = 1$, $\omega(1) = \mathbb{E}[(p^2 - (a + bp))(p - 1)|p \geq 1]/\mathbb{E}[p^2 \widetilde{p^2}] \Pr\{p \geq t\} = 0$ trivially. For every $t \in [0, 1)$, notice that $\mathbb{E}[(p^2 - (a + bp))(p - t)|p \geq t] \Pr\{p \geq t\}/\mathbb{E}[p^2 \widetilde{p^2}] \geq 0$ if and only if $\varphi(t) \equiv \mathbb{E}[(p^2 - (a + bp))(p - t)|p \geq t] \Pr\{p \geq t\} \geq 0$. Denote $\varepsilon = \varepsilon(p) = p^2 - (a + bp)$ and rewrite

$$\varphi(t) = \mathbb{E}[\varepsilon(p - t)|p \geq t] \Pr\{p \geq t\} = \int_t^1 \varepsilon(p - t) f(p) dp = \int_t^1 \varepsilon p f(p) dp - t \int_t^1 \varepsilon f(p) dp.$$

The derivative of $\varphi(t)$ is

$$\varphi'(t) = -[\varepsilon p f(p)]_{p=t} - \int_t^1 \varepsilon f(p) dp + t[\varepsilon f(p)]_{p=t} = -\int_t^1 \varepsilon f(p) dp,$$

and the second derivative of $\varphi(t)$ is

$$\varphi''(t) = [\varepsilon f(p)]_{p=t} = (t^2 - (a + bt)) f(t).$$

Notice that $\varphi''(t)$ is zero at the intersections of p^2 and the projection line $a + bp$. By Lemma A3, there are always two such intersection points. Denote these intersection points as t_1 and t_2, $t_1 < t_2$, so that $\varphi''(t_1) = \varphi''(t_2) = 0$. $\varphi''(t) \geq 0$ for $t \in [0, t_1] \cup (t_2, t]$, $\varphi''(t) \leq 0$ for $t \in (t_1, t_2]$.

At $t = 0$,

$$\varphi'(0) = -\int_0^1 \varepsilon f(p) dp = -\mathbb{E}[\varepsilon] = 0$$

by FOC of the projection coefficients (a, b). Together $\varphi'(0)$ and $\varphi''(t) \geq 0$ for $t \in [0, t_1]$ imply that φ' is monotone increasing on $[0, t_1]$, then $\varphi'(t) \geq \varphi'(0) \geq 0$ for all $t \in [0, t_1]$, and then $\omega(t) \geq 0$ for all $t \in [0, t_1]$.

Now consider $t \in (t_2, 1]$. Recall that $p = t_2$ is the intersection of p^2 with projection $a + bp$, then $p^2 - (a + bp) \geq 0$ for all $p \geq t$, for all $t \in [t_2, 1]$. Further, $p - t \geq 0$ for all $p \geq t$, for all $t \in [0, 1]$. Then $\omega(t) = \mathbb{E}[(p^2 - (a + bp))(p - t)|p \geq t] \geq 0$ for all $t \in (t_2, 1]$.

It is left to show $\omega(t) \geq 0$ for all $t \in (t_1, t_2]$. Recall that $\varphi''(t) \leq 0$ for all $t \in (t_1, t_2]$, hence φ is concave on $(t_1, t_2]$ and $\varphi(t) \geq \min\{\varphi(t_1), \varphi(t_2)\}$ for all $t \in [t_1, t_2]$. From the argument above, $\varphi(t_1) \geq 0$ and $\varphi(t_2) \geq 0$, hence $\varphi(t) \geq 0$ for all $t \in [t_1, t_2]$. It implies $\omega(t) \geq 0$ for all $t \in (t_1, t_2)$. □

1.B Appendix: proofs of main results

Proof of Proposition 1. By the fundamental theorem of calculus, the reduced-form regression is

$$m(p) = \lim_{p \to 0} m(p) + \int_0^p m'(t) dt. \tag{36}$$

Application of the fundamental theorem of calculus to m' yields

$$m'(p) = \lim_{p \to 0} m'(p) + \int_0^p m''(t) dt. \tag{37}$$

Plug (37) into (36) to obtain

$$\begin{aligned} m(p_i) &= \lim_{p \to 0} m(p) + \int_0^{p_i} \lim_{\tau \to 0} m'(\tau) d\tau + \int_0^{p_i} \int_0^{\tau} m''(t) dt d\tau \\ &= c_1 + c_2 p_i + \int_0^{p_i} \int_0^{\tau} m''(t) dt d\tau, \end{aligned} \tag{38}$$

where $c_1 = \lim_{p \to 0} m(p)$ and $c_2 = \lim_{p \to 0} m'(p)$ are constants. Intuitively, both (36) and (38) (partly) linearize $m(p_i)$: the first term is a constant, the second term is linear in p_i, the third is potentially of higher order or nonlinear in p_i. Plug (38) into (12):

$$\begin{aligned}\gamma &= \frac{\mathbb{E}[m(p_i)\widetilde{p^2}_i]}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\\ &= \frac{\mathbb{E}[(c_1 + c_2 p_i + \int_0^{p_i}\int_0^{\tau} m''(t)dtd\tau)\widetilde{p^2}_i]}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\\ &= \frac{\mathbb{E}[\widetilde{p^2}_i \int_0^{p_i}\int_0^{\tau} m''(t)dtd\tau}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\\ &= \frac{1}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\int_0^1\int_0^p\int_0^{\tau} m''(t)\widetilde{p^2}dtd\tau dF(p)\\ &= \frac{1}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\int_0^1\int_t^1\int_t^p m''(t)\widetilde{p^2}d\tau dF(p)dt\\ &= \int_0^1 m''(t)\omega(t)dt,\end{aligned}$$

where the third line follows from orthogonality between p_i and $\widetilde{p^2}_i$, the fifth line follows from changing the order of integrations and weights $\omega(t)$ are

$$\begin{aligned}\omega(t) &= \frac{1}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\int_t^1\int_t^p \left(p^2 - \mathbb{L}[p^2|p]\right) d\tau dF(p)\\ &= \frac{1}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\int_t^1\int_t^p \left(p^2 - (a_0 + b_0 p)\right) d\tau dF(p)\\ &= \frac{1}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\int_t^1 \left(p^2 - (a_0 + b_0 p)\right)\left(\int_t^p d\tau\right) dF(p)\\ &= \frac{1}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\int_t^1 \left(p^2 - (a_0 + b_0 p)\right)(p-t)\, dF(p)\\ &= \frac{1}{\mathbb{E}[p_i^2\widetilde{p^2}_i]}\mathbb{E}\left[(p^2 - (a_0 + b_0 p))(p-t)|p \geq t\right]\Pr\{p \geq t\},\end{aligned}$$

(a_0, b_0) are coefficients of projection of p^2 on p and a constant, $\mathbb{L}[p^2|p] = a_0 + b_0 p$:

$$\begin{aligned}b_0 &= \frac{\mathrm{Cov}(p^2, p)}{\mathrm{Var}(p)},\\ a_0 &= \mathbb{E}[p^2] - b_0\mathbb{E}[p].\end{aligned}$$

By Lemma A4, $\omega(t) \geq 0$. $\square$

Proof of Proposition 2. Let $f_{Y,p}$ be the joint density function of Y and p, denote $g_s(y,p) = f_{Y,p}(y,p)K(p,s)$ and $G_s = \int g_s(y,p)d(y,p)$. Let $\mathbb{E}_s[.]$ denote the expectation with respect to the joint probability density function $g_s(y,p)/G_s$. The optimization problem in (15) is equivalent to

$$\min_{a(s),b(s),c(s)} \mathbb{E}[K_i(s)\left(Y_i - a(s) - b(s)p_i - c(s)p_i^2\right)^2/G_s]. \tag{39}$$

The objective function in (39) is equivalent to

$$\begin{aligned}
&\mathbb{E}[K_i(s)\left(Y_i - a(s) - b(s)p_i - c(s)p_i^2\right)/G_s] \\
=& \int \left(y - a(s) - b(s)p - c(s)p^2\right)^2 K(p,s) f_{Y,p}(y,p)/G_s d(y,p) \\
=& \int \left(y - a(s) - b(s)p - c(s)p^2\right)^2 g_s(y,p)/G_s d(y,p).
\end{aligned}$$

Note that $g_s(y,p)/G_s \geq 0$ for every s, and $\int g_s(y,p)/G_s dp = 1$ by construction, hence $g_s(y,p)/G_s$ is a proper probability density function. Then (15) can be interpreted as an unweighted quadratic projection when the joint density function of (Y,p) is $g_s(y,p)/G_s$ and by Proposition 1,

$$\gamma(s) = \int_0^1 \omega(t,s) m''(t) dt,$$

where

$$\begin{aligned}
\omega(t,s) &= \frac{1}{\mathbb{E}_{gs}[p_i^2 \widetilde{p^2}_i]} \mathbb{E}_{gs}\left[(p^2 - (a_0 + b_0 p))(p-t)|p \geq t\right] \Pr_{gs}\{p \geq t\} \\
&= \frac{1}{\mathbb{E}_{gs}[p_i^2 \widetilde{p^2}_i]} \mathbb{E}_{gs}\left[(p^2 - (a_0 + b_0 p))(p-t)|p \geq t\right] \Pr_{gs}\{p \geq t\},
\end{aligned}$$

that are non-negative by Lemma A4. Let $\mathbb{L}[p^2|p] = a_0 + b_0 p$:

$$\begin{aligned}
b_0 &= \frac{\mathrm{Cov}(p^2, p)}{\mathrm{Var}(p)}, \\
a_0 &= \mathbb{E}[p^2] - b_0 \mathbb{E}[p].
\end{aligned}$$

Consider the numerator of the weighting function $\omega(t,s)$:

$$\begin{aligned}
&\mathbb{E}_{gs}\left[(p^2 - (a_0 + b_0 p))(p-t)|p \geq t\right] \Pr_{gs}\{p \geq t\} \\
=& \int_{\mathbb{Y}} \int_t^1 (p^2 - (a_0 + b_0 p))(p-t) g_s(y,p)/G_s d(y,p) \\
=& \int_t^1 (p^2 - (a_0 + b_0 p))(p-t) K(p,s)/G_s \left(\int_{\mathbb{Y}} f_{Y,p}(y,p) dy\right) dp \\
=& \int_t^1 (p^2 - (a_0 + b_0 p))(p-t) K(p,s)/G_s f(p) dp \\
=& \mathbb{E}[(p^2 - (a_0 + b_0 p))(p-t) K(p,s)|p \geq t] \Pr\{p \geq t\}/G_s.
\end{aligned}$$

The denominator is

$$\begin{aligned}
\mathbb{E}_{gs}\left[p^2 \tilde{p^2}\right] &= \int_{\mathbb{Y}} \int_0^1 p^2 \tilde{p^2} g_s(y,p)/G_s d(y,p) \\
&= \int_0^1 p^2 \tilde{p^2} K(p,s)/G_s \left(\int_{\mathbb{Y}} f_{Y,p}(y,p) dy\right) dp \\
&= \int_0^1 p^2 \tilde{p^2} K(p,s)/G_s f(p) dp \\
&= \mathbb{E}[p^2 \tilde{p^2} K(p,s)]/G_s.
\end{aligned}$$

Combine to obtain

$$\omega(t,s) = \frac{1}{\mathbb{E}[p^2\tilde{p^2}K(p,s)]}\mathbb{E}[(p^2-(a_0+b_0p))(p-t)K(p,s)|p\geq t]\Pr\{p\geq t\}.$$

□

Proof of Corollary 1. Consider a quadratic projection of outcome Y on treatment probabilities p: $Y = \alpha + \beta p + \gamma p^2 + e$. If reduced-form regression $m(p) = \mathbb{E}[Y|p]$ is twice differentiable, then by Proposition 1, γ is an average second derivative of $m(p)$. By (3), $\text{MTE}(u) = dm(p)/dp|_{p=u} \equiv m'(u)$. Then the slope of MTE at point u is $m''(u)$, hence $\gamma = \int_0^1 \omega(u)\text{MTE}(u)du$, the average slope of the MTE curve with the weights $\omega(u)$ given by Proposition 1 in (14). □

Proof of Corollary 2. Similarly to the proof of Corollary 1, consider a weighted quadratic projection of outcome Y on treatment probabilities p: $Y = \alpha(s) + \beta(s)p + \gamma(s)p^2 + e^s$ with some weights $K(p,s)$. If reduced-form regression $m(p) = \mathbb{E}[Y|p]$ is twice differentiable, then by Proposition 2, $\gamma(s)$ is an average second derivative of $m(p)$. By (3), $\text{MTE}(u) = dm(p)/dp|_{p=u} \equiv m'(u)$. Then the slope of MTE at point u is $m''(u)$, hence $\gamma(s) = \int_0^1 \omega(u,s)\text{MTE}(u)du$, the average slope of the MTE curve with the weights $\omega(u,s)$ given by Proposition 2 in (16). □

Proof of Proposition 3. By the argument in Section 4.2, the moment function $M_{i,s}$ is i.i.d. across i, then $\gamma(s)$ is an expectation of i.i.d. terms.

Then show that conditions of Theorem 4.3 in Chernozhukov et al. (2019) are satisfied for two test statistics T_n and T_n^m and two algorithms B1 and B2. Consider first T_n^m and algorithm B2.

$$\begin{aligned} W^m &= \max_s \frac{\sqrt{n}(\gamma^*(s)-\gamma(s))}{\sqrt{V(s)}} \\ &= \max_s \frac{\sqrt{n}(\mathbb{E}_n[M_{i,s}^*]-\gamma(s))}{\sqrt{V(s)}}, \end{aligned}$$

which corresponds to algorithm EB in Chernozhukov et al. (2019) with the choice set dimension $p = |\mathcal{S}|$, choice variable $j = s$ and bootstrap moment function $X_{i,s}^* = M_{i,s}^*$. Further, by the statement of Proposition 3, $(\max(\{A_{n,3}^3, A_{n,4}^2, B_n\})^2 \log^{7/2}(|\mathcal{S}|n) \leq c_2 n^{1/2-c_1}$ for some constants $c_1 \in (0,1/2)$ and $c_2 \geq 0$. Hence by Theorem 4.3 in Chernozhukov et al. (2019), there exists $c_5 > 0$ and $c_6 > 0$ such that under $\mathbb{H}_0^m$,

$$\Pr\{T_n^m > c_n^m(\alpha)\} \leq \alpha + c_5 n^{-c_6}.$$

Now first the test statistic T_n and algorithm B1. The test statistic is

$$T_n = \max_s \frac{|b(s)|}{\sqrt{V(s)}} = \max\left\{\max_s \frac{b(s)}{\sqrt{V(s)}}, \max_s \frac{-b(s)}{\sqrt{V(s)}}\right\}.$$

Let $\breve{s} \in \breve{\mathcal{S}} = \mathcal{S} \times \{0,1\}$. Further, let $\breve{b}(\breve{s}) = \breve{b}(s,j) = (-1)^j b(s)$, where $j \in \{0,1\}$, so that $\breve{b}(s,0) = b(s)$, $\breve{b}(s,1) = -b(s)$. Notice that $\mathrm{Var}(\breve{b}(\breve{s})) = \breve{V}(\breve{s}) = \mathrm{Var}(b(s)) = V(s)$. Hence,

$$T_n = \max_s \frac{|b(s)|}{\sqrt{V(s)}} = \max_{\breve{s}} \frac{\breve{b}(\breve{s})}{\sqrt{V(s)}} = \max_{\breve{s}} \frac{\breve{b}(\breve{s})}{\sqrt{\breve{V}(\breve{s})}}.$$

Similarly, let $\breve{\gamma}(s,j) = (-1)^j \gamma(s)$ for $j \in \{0,1\}$. Notice that $\breve{\gamma}(\breve{s}) = \breve{b}(\breve{s})$. Finally, let the $\breve{M}_{i,\breve{s}} = \breve{M}_{i,s,j} = (-1)^j M_{i,s}$ for $j \in \{0,1\}$, and let $\breve{M}^*_{i,\breve{s}}$ be the bootstrap version of $\breve{M}_{i,\breve{s}}$. The bootstrap test statistic in B1 is

$$\begin{aligned} W &= \max_s \frac{\sqrt{n}|\gamma^*(s) - \gamma(s)|}{\sqrt{V(s)}} \\ &= \max\left\{\max_s \frac{\sqrt{n}(\gamma^*(s) - \gamma(s))}{\sqrt{V(s)}}, \max_s \frac{\sqrt{n}(-\gamma^*(s) - (-\gamma(s)))}{\sqrt{V(s)}}\right\} \\ &= \max_{\breve{s}} \frac{\sqrt{n}(\mathbb{E}_n[\breve{M}^*_{i,\breve{s}}] - \breve{\gamma}(\breve{s}))}{\sqrt{\breve{V}(\breve{s})}}, \end{aligned}$$

which corresponds to algorithm EB in Chernozhukov et al. (2019) with the choice set dimension $p = |\breve{\mathcal{S}}|$, choice variable $j = \breve{s}$ and bootstrap moment function $X^*_{i,s} = \breve{M}^*_{i,s}$.

Further, note that for every $\breve{s} = (s,j)$,

$$\begin{aligned} \left|\frac{\breve{M}_{1,\breve{s}} - \breve{\gamma}(\breve{s})}{\sqrt{\breve{V}(\breve{s})}}\right| &= \left|\frac{(-1)^j M_{1,s} - (-1)^j \gamma(s)}{\sqrt{V(s)}}\right| \\ &= \left|\frac{M_{1,s} - \gamma(s)}{\sqrt{V(s)}}\right| \\ &= |\tilde{M}_{i,s}|, \end{aligned}$$

then

$$\max_{\breve{s}} \left(\mathbb{E}\left[\left|\frac{\breve{M}_{1,\breve{s}} - \breve{\gamma}(\breve{s})}{\sqrt{\breve{V}(\breve{s})}}\right|^k\right]\right)^{1/k} = \max_s \left(\mathbb{E}\left[|\tilde{M}_{i,s}|^k\right]\right)^{1/k} = A_{n,k}.$$

Moreover,

$$\begin{aligned}
\left(\frac{\breve{M}_{1,\breve{s}} - \breve{\gamma}(\breve{s})}{\sqrt{\breve{V}(\breve{s})}}\right)^4 &= \left(\frac{\breve{M}_{1,\breve{s}} - \breve{\gamma}(\breve{s})}{\sqrt{\breve{V}(\breve{s})}}\right)^4 \\
&= \left(\frac{(-1)^j M_{1,s} - (-1)^j \gamma(s)}{\sqrt{V(s)}}\right)^4 \\
&= \left(\frac{M_{1,s} - \gamma(s)}{\sqrt{V(s)}}\right)^4 = B_n.
\end{aligned}$$

Then by the statement of Proposition 3, $(\max\{A_{n,3}^3, A_{n,4}^2, B_n\})^2 \log^{7/2}(|\mathcal{S}|n) \leq c_2 n^{1/2-c_1}$ for some constants $c_1 \in (0, 1/2)$ and $c_2 \geq 0$. Hence by Theorem 4.3 in Chernozhukov et al. (2019), there exists $c_3 > 0$ and $c_4 > 0$ such that under $\mathbb{H}_0$,

$$\Pr\{T_n > c_n(\alpha)\} \leq \alpha + c_3 n^{-c_4}.$$

$\square$

Proof of Proposition 4. In the first part of the proof, I show that

$$\mathbb{E}_n[K(\hat{p}, s)\hat{p}] \xrightarrow{p} \mathbb{E}[K(p_0, s)p_0], \tag{40}$$
$$\mathbb{E}_n[K(\hat{p}, s)\hat{p}^2] \xrightarrow{p} \mathbb{E}[K(p_0, s)p_0^2], \tag{41}$$
$$\mathbb{E}_n[K(\hat{p}, s)\hat{p}^3] \xrightarrow{p} \mathbb{E}[K(p_0, s)p_0^3], \tag{42}$$
$$\mathbb{E}_n[K(\hat{p}, s)\hat{p}^4] \xrightarrow{p} \mathbb{E}[K(p_0, s)p_0^4], \tag{43}$$
$$\mathbb{E}_n[YK(\hat{p}, s)\hat{p}] \xrightarrow{p} \mathbb{E}[YK(p_0, s)p_0], \tag{44}$$
$$\mathbb{E}_n[YK(\hat{p}, s)\hat{p}^2] \xrightarrow{p} \mathbb{E}[YK(p_0, s)p_0^2]. \tag{45}$$

First, consider $\mathbb{E}_n[K(\hat{p}, s)\hat{p}] - \mathbb{E}[K(p_0, s)p_0]$. By triangle inequality,

$$\begin{aligned}
\Pr\{|\mathbb{E}_n[K(\hat{p}, s)\hat{p}] - \mathbb{E}[K(p_0, s)p_0]| \geq 4\varepsilon\} \leq\ & \Pr\{|\mathbb{E}_n[K(\hat{p}, s)\hat{p}] - \mathbb{E}[K(\hat{p}, s)\hat{p}]| \geq 2\varepsilon\} \\
& + \Pr\{|\mathbb{E}[K(\hat{p}, s)\hat{p}] - \mathbb{E}[K(p_0, s)p_0]| \geq 2\varepsilon\} \\
\leq\ & \Pr\{|\mathbb{E}_n[K(\hat{p}, s)\hat{p}] - \mathbb{E}[K(\hat{p}, s)\hat{p}]| \geq 2\varepsilon\} \\
& + \Pr\{|\mathbb{E}[K(\hat{p}, s)\hat{p}] - \mathbb{E}[K(\hat{p}, s)p_0]| \geq \varepsilon\} \\
& + \Pr\{|\mathbb{E}[K(\hat{p}, s)p_0] - \mathbb{E}[K(p_0, s)p_0]| \geq \varepsilon\}
\end{aligned} \tag{46}$$

Notice that for fixed h, kernel function $K(p, s) \equiv K(p, \mu, h) \leq C$. Consider the first term in

(46):

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}_n[K(\hat{p},s)\hat{p}] - \mathbb{E}[K(\hat{p},s)\hat{p}])^2] &= \frac{1}{n}(\mathbb{E}[K(\hat{p},s)\hat{p}])^2 \\
&\leq \frac{1}{n}\mathbb{E}[(K(\hat{p},s)\hat{p})^2] \\
&\leq \sup_{p\in\mathcal{P}_n} \frac{1}{n}\mathbb{E}[(K(p,s)p)^2] \\
&\leq C \sup_{p\in\mathcal{P}_n} \frac{1}{n}\mathbb{E}[p^2] \leq C/n \to 0.
\end{aligned}$$

Therefore, by Markov's inequality, $\mathbb{E}_n[K(\hat{p},s)\hat{p}] - \mathbb{E}[K(\hat{p},s)\hat{p}] = o_p(1)$, i.e. $\Pr\{|\mathbb{E}_n[K(\hat{p},s)\hat{p}] - \mathbb{E}[K(\hat{p},s)\hat{p}]| \geq 2\varepsilon\} \to 0$.

Now consider the second term in (46):

$$\begin{aligned}
|\mathbb{E}[K(\hat{p},s)\hat{p}] - \mathbb{E}[K(\hat{p},s)p_0]| &\leq \left(\mathbb{E}[K^2(\hat{p},s)]\mathbb{E}[(\hat{p}-p_0)^2]\right)^{1/2} \\
&\leq C \sup_{p\in\mathcal{P}_n} \left(\mathbb{E}[(p-p_0)^2]\right)^{1/2} \xrightarrow{p} 0.
\end{aligned}$$

Finally, consider the third term in (46):

$$\begin{aligned}
|\mathbb{E}[K(\hat{p},s)p_0] - \mathbb{E}[K(p_0,s)p_0]| &\leq \left(\mathbb{E}[p_0^2]\mathbb{E}[(K(\hat{p},s) - K(p_0,s))^2]\right)^{1/2} \\
&\leq C\left(\mathbb{E}[(\hat{p}-p_0)^2]\right)^{1/2} \\
&\leq C \sup_{p\in\mathcal{P}_n} \left(\mathbb{E}[(p-p_0)^2]\right)^{1/2} \xrightarrow{p} 0,
\end{aligned}$$

where the second line uses $K(p,s)$ being Lipschitz continuous in p. Hence, (40) follows by triangle inequality.

Now consider $\mathbb{E}_n[K(\hat{p},s)\hat{p}^2] - \mathbb{E}[K(p_0,s)p_0^2]$. By triangle inequality,

$$\begin{aligned}
\Pr\{|\mathbb{E}_n[K(\hat{p},s)\hat{p}^2] - \mathbb{E}[K(p_0,s)p_0^2]| \geq 4\varepsilon\} \leq\ & \Pr\{|\mathbb{E}_n[K(\hat{p},s)\hat{p}^2] - \mathbb{E}[K(\hat{p},s)\hat{p}^2]| \geq 2\varepsilon\} \\
&+ \Pr\{|\mathbb{E}[K(\hat{p},s)\hat{p}^2] - \mathbb{E}[K(\hat{p},s)p_0^2]| \geq \varepsilon\} \\
&+ \Pr\{|\mathbb{E}[K(\hat{p},s)p_0^2] - \mathbb{E}[K(p_0,s)p_0^2]| \geq \varepsilon\}.
\end{aligned}$$

Similarly, I show that all three terms above are $o_p(1)$, hence (41) would follow by triangle inequality.

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}_n[K(\hat{p},s)\hat{p}^2] - \mathbb{E}[K(\hat{p},s)\hat{p}^2])^2] &\leq \frac{1}{n}\mathbb{E}[(K(\hat{p},s)\hat{p}^2)^2] \\
&\leq \sup_{p\in\mathcal{P}_n} \frac{1}{n}\mathbb{E}[(K(p,s)p^2)^2] \\
&\leq C \sup_{p\in\mathcal{P}_n} \frac{1}{n}\mathbb{E}[p^4] \leq C/n \to 0.
\end{aligned}$$

Therefore, by Markov's inequality, $\mathbb{E}_n[K(\hat{p},s)\hat{p}^2]-\mathbb{E}[K(\hat{p},s)\hat{p}^2]=o_p(1)$. Further,

$$\begin{aligned}
|\mathbb{E}[K(\hat{p},s)\hat{p}^2]-\mathbb{E}[K(\hat{p},s)p_0^2]| &= |\mathbb{E}[K(\hat{p},s)(\hat{p}^2-p_0^2)]| \\
&= |\mathbb{E}[K(\hat{p},s)(\hat{p}-p_0)(\hat{p}+p_0)]| \\
&\leq \left(\mathbb{E}[K^2(\hat{p},s)(\hat{p}+p_0)^2]\mathbb{E}[(\hat{p}-p_0)^2]\right)^{1/2} \\
&\leq C\max\left\{\sup_{p\in\mathcal{P}_n}\left(\mathbb{E}[p^2]\right)^{1/2},1\right\}\sup_{p\in\mathcal{P}_n}\left(\mathbb{E}[(p-p_0)^2]\right)^{1/2}\xrightarrow{p}0.
\end{aligned}$$

Finally,

$$\begin{aligned}
|\mathbb{E}[K(\hat{p},s)p_0^2]-\mathbb{E}[K(p_0,s)p_0^2]| &\leq \left(\mathbb{E}[p_0^4]\mathbb{E}[(K(\hat{p},s)-K(p_0,s))^2]\right)^{1/2} \\
&\leq C\left(\mathbb{E}[(\hat{p}-p_0)^2]\right)^{1/2} \\
&\leq C\sup_{p\in\mathcal{P}_n}\left(\mathbb{E}[(p-p_0)^2]\right)^{1/2}\xrightarrow{p}0.
\end{aligned}$$

Hence, (41) follows by triangle inequality.

For (42) and (43) the argument is similar to the above.

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}_n[K(\hat{p},s)\hat{p}^3]-\mathbb{E}[K(\hat{p},s)\hat{p}^3])^2] &\leq \frac{1}{n}\mathbb{E}[(K(\hat{p},s)\hat{p}^3)^2] \\
&\leq C\sup_{p\in\mathcal{P}_n}\frac{1}{n}\mathbb{E}[p^6]\leq C/n\to 0.
\end{aligned}$$

Therefore, by Markov's inequality, $\mathbb{E}_n[K(\hat{p},s)\hat{p}^3]-\mathbb{E}[K(\hat{p},s)\hat{p}^3]=o_p(1)$. Further,

$$\begin{aligned}
|\mathbb{E}[K(\hat{p},s)\hat{p}^3]-\mathbb{E}[K(\hat{p},s)p_0^3]| &= |\mathbb{E}[K(\hat{p},s)(\hat{p}^3-p_0^3)]| \\
&= |\mathbb{E}[K(\hat{p},s)(\hat{p}-p_0)(\hat{p}^2+p_0^2+\hat{p}p_0)]| \\
&\leq C\sup_{p\in\mathcal{P}_n}\left(\mathbb{E}[(p-p_0)^2]\right)^{1/2}\xrightarrow{p}0.
\end{aligned}$$

Finally,

$$\begin{aligned}
|\mathbb{E}[K(\hat{p},s)p_0^3]-\mathbb{E}[K(p_0,s)p_0^3]| &\leq C\left(\mathbb{E}[p_0^6]\right)^{1/2}\left(\mathbb{E}[(\hat{p}-p_0)^2]\right)^{1/2} \\
&\leq C\sup_{p\in\mathcal{P}_n}\left(\mathbb{E}[(p-p_0)^2]\right)^{1/2}\xrightarrow{p}0.
\end{aligned}$$

Hence, (42) follows by triangle inequality. Further,

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}_n[K(\hat{p},s)\hat{p}^4]-\mathbb{E}[K(\hat{p},s)\hat{p}^2])^4] &\leq \frac{1}{n}\mathbb{E}[(K(\hat{p},s)\hat{p}^4)^2] \\
&\leq C\sup_{p\in\mathcal{P}_n}\frac{1}{n}\mathbb{E}[p^8]\leq C/n\to 0.
\end{aligned}$$

Therefore, by Markov's inequality, $\mathbb{E}_n[K(\hat{p},s)\hat{p}^4] - \mathbb{E}[K(\hat{p},s)\hat{p}^4] = o_p(1)$. Further,

$$\begin{aligned} |\mathbb{E}[K(\hat{p},s)\hat{p}^4] - \mathbb{E}[K(\hat{p},s)p_0^4]| &= |\mathbb{E}[K(\hat{p},s)(\hat{p}^4 - p_0^4)]| \\ &= |\mathbb{E}[K(\hat{p},s)(\hat{p}-p_0)(\hat{p}^3 + p_0^3 + \hat{p}^2 p_0 + \hat{p}p_0^2)]| \\ &\leq C \sup_{p\in\mathcal{P}_n} \left(\mathbb{E}[(p-p_0)^2]\right)^{1/2} \xrightarrow{p} 0. \end{aligned}$$

Finally,

$$\begin{aligned} |\mathbb{E}[K(\hat{p},s)p_0^4] - \mathbb{E}[K(p_0,s)p_0^4]| &\leq C \left(\mathbb{E}[p_0^8]\right)^{1/2} \left(\mathbb{E}[(\hat{p}-p)^2]\right)^{1/2} \\ &\leq C \sup_{p\in\mathcal{P}_n} \left(\mathbb{E}[(p-p_0)^2]\right)^{1/2} \xrightarrow{p} 0. \end{aligned}$$

Hence, (43) follows by triangle inequality.

Now consider $\mathbb{E}_n[YK(\hat{p},s)\hat{p}] - \mathbb{E}[YK(p_0,s)p_0]$. By triangle inequality,

$$\begin{aligned} \Pr\{|\mathbb{E}_n[YK(\hat{p},s)\hat{p}] - \mathbb{E}[YK(p_0,s)p_0]| \geq 4\varepsilon\} \leq\ & \Pr\{|\mathbb{E}_n[YK(\hat{p},s)\hat{p}] - \mathbb{E}[YK(\hat{p},s)\hat{p}]| \geq 2\varepsilon\} \\ &+ \Pr\{|\mathbb{E}[YK(\hat{p},s)\hat{p}] - \mathbb{E}[YK(\hat{p},s)p_0]| \geq \varepsilon\} \\ &+ \Pr\{|\mathbb{E}[YK(\hat{p},s)p_0] - \mathbb{E}[YK(p_0,s)p_0]| \geq \varepsilon\} \end{aligned} \tag{47}$$

Consider the first term in (47)

$$\begin{aligned} \mathbb{E}[(\mathbb{E}_n[YK(\hat{p},s)\hat{p}] - \mathbb{E}[YK(\hat{p},s)\hat{p}])^2] &= \frac{1}{n}(\mathbb{E}[YK(\hat{p},s)\hat{p}])^2 \\ &\leq \frac{1}{n}\mathbb{E}[Y^2K(\hat{p},s)^2\hat{p}^2] \\ &\leq \sup_{p\in\mathcal{P}_n} \frac{1}{n}\mathbb{E}[Y^2K(p,s)^2p^2] \\ &\leq C \sup_{p\in\mathcal{P}_n} \frac{1}{n}\mathbb{E}[Y^2p^2] \to 0. \end{aligned}$$

Therefore, by Markov's inequality, $\mathbb{E}_n[YK(\hat{p},s)\hat{p}] - \mathbb{E}[YK(\hat{p},s)\hat{p}] = o_p(1)$.

Now consider the second term in (46):

$$\begin{aligned} |\mathbb{E}[YK(\hat{p},s)\hat{p}] - \mathbb{E}[YK(\hat{p},s)p_0]| &= |\mathbb{E}[YK(\hat{p},s)(\hat{p}-p_0)]| \\ &\leq (\mathbb{E}[Y^2K^2(\hat{p},s)])^{1/2} \sup_{p\in\mathcal{P}_n}(\mathbb{E}[(p-p_0)^2])^{1/2} \\ &\leq C\mathbb{E}[Y^2])^{1/2} \sup_{p\in\mathcal{P}_n}(\mathbb{E}[(p-p_0)^2])^{1/2} \\ &\leq C \sup_{p\in\mathcal{P}_n}(\mathbb{E}[(p-p_0)^2])^{1/2} \xrightarrow{p} 0. \end{aligned}$$

Finally, consider the third term in (46):

$$\begin{aligned} |\mathbb{E}[YK(\hat{p},s)p_0] - \mathbb{E}[YK(p_0,s)p_0]| &= |\mathbb{E}[Yp_0(K(\hat{p},s) - K(p_0,s))]| \\ &\leq (\mathbb{E}[Y^2p_0^2])^{1/2} \sup_{p\in\mathcal{P}_n}(\mathbb{E}[(K(p,s)-K(p_0,s))^2])^{1/2} \\ &\leq C(\mathbb{E}[Y^2])^{1/2} \sup_{p\in\mathcal{P}_n}(\mathbb{E}[(p-p_0)^2])^{1/2} \xrightarrow{p} 0. \end{aligned}$$

Hence, (44) follows by triangle inequality.

By the similar argument,

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}_n[YK(\hat{p},s)\hat{p}^2] - \mathbb{E}[YK(\hat{p},s)\hat{p}^2])^2] &= \frac{1}{n}(\mathbb{E}[YK(\hat{p},s)\hat{p}^2])^2 \\
&\leq \frac{1}{n}\mathbb{E}[Y^2K(\hat{p},s)^2\hat{p}^4] \\
&\leq \sup_{p\in\mathcal{P}_n} \frac{1}{n}\mathbb{E}[Y^2K(p,s)^2p^4] \\
&\leq C \sup_{p\in\mathcal{P}_n} \frac{1}{n}\mathbb{E}[Y^2p^4] \to 0.
\end{aligned}$$

Therefore, by Markov's inequality, $\mathbb{E}_n[YK(\hat{p},s)\hat{p}] - \mathbb{E}[YK(\hat{p},s)\hat{p}^2] = o_p(1)$. Further,

$$\begin{aligned}
|\mathbb{E}[YK(\hat{p},s)\hat{p}^2] - \mathbb{E}[YK(\hat{p},s)p_0^2]| &= |\mathbb{E}[YK(\hat{p},s)(\hat{p}-p_0)(\hat{p}+p_0)]| \\
&\leq C\mathbb{E}[Y^2])^{1/2} \sup_{p\in\mathcal{P}_n} (\mathbb{E}[(p-p_0)^2])^{1/2} \\
&\leq C \sup_{p\in\mathcal{P}_n} (\mathbb{E}[(p-p_0)^2])^{1/2} \xrightarrow{p} 0.
\end{aligned}$$

Finally,

$$\begin{aligned}
|\mathbb{E}[YK(\hat{p},s)p^2] - \mathbb{E}[YK(p_0,s)p_0^2]| &= |\mathbb{E}[Yp_0^2(K(\hat{p},s) - K(p_0,s))]| \\
&\leq (\mathbb{E}[Y^2p_0^4])^{1/2} \sup_{p\in\mathcal{P}_n} (\mathbb{E}[(K(p,s) - K(p_0,s))^2])^{1/2} \\
&\leq C(\mathbb{E}[Y^2])^{1/2} \sup_{p\in\mathcal{P}_n} (\mathbb{E}[(p-p_0)^2])^{1/2} \xrightarrow{p} 0.
\end{aligned}$$

Hence, (45) follows by triangle inequality.

In the second part of the proof, notice that (30) and (31) are WLS estimators and have closed-form solutions

$$\begin{pmatrix}\alpha(s)\\ \beta(s)\\ \gamma(s)\end{pmatrix} = \left(\mathbb{E}\left[K(p_0,s)\begin{pmatrix}1\\ p_0\\ p_0^2\end{pmatrix}\begin{pmatrix}1\\ p_0\\ p_0^2\end{pmatrix}'\right]\right)^{-1} \mathbb{E}\left[YK(p_0,s)\begin{pmatrix}1\\ p_0\\ p_0^2\end{pmatrix}\right]$$

and

$$\begin{pmatrix}\hat{\alpha}_F(s)\\ \hat{\beta}_F(s)\\ \hat{\gamma}_F(s)\end{pmatrix} = \left(\mathbb{E}_n\left[K(\hat{p},s)\begin{pmatrix}1\\ \hat{p}\\ \hat{p}^2\end{pmatrix}\begin{pmatrix}1\\ \hat{p}\\ \hat{p}^2\end{pmatrix}'\right]\right)^{-1} \mathbb{E}_n\left[YK(\hat{p},s)\begin{pmatrix}1\\ \hat{p}\\ \hat{p}^2\end{pmatrix}\right].$$

From (40)-(45), by the Slutsky theorem and Mann-Wald theorem,

$$\begin{aligned}
\begin{pmatrix}\hat{\alpha}_F(s)\\ \hat{\beta}_F(s)\\ \hat{\gamma}_F(s)\end{pmatrix} &= \left(\mathbb{E}_n\left[K(\hat{p},s)\begin{pmatrix}1\\ \hat{p}\\ \hat{p}^2\end{pmatrix}\begin{pmatrix}1\\ \hat{p}\\ \hat{p}^2\end{pmatrix}'\right]\right)^{-1} \mathbb{E}_n\left[YK(\hat{p},s)\begin{pmatrix}1\\ \hat{p}\\ \hat{p}^2\end{pmatrix}\right] \\
&\xrightarrow{p} \left(\mathbb{E}\left[K(p_0,s)\begin{pmatrix}1\\ p_0\\ p_0^2\end{pmatrix}\begin{pmatrix}1\\ p_0\\ p_0^2\end{pmatrix}'\right]\right)^{-1} \mathbb{E}\left[YK(p_0,s)\begin{pmatrix}1\\ p_0\\ p_0^2\end{pmatrix}\right] = \begin{pmatrix}\alpha(s)\\ \beta(s)\\ \gamma(s)\end{pmatrix}.
\end{aligned}$$

□

1.C Appendix: details of the simulation design

I will focus on the population of charter school applicants. The IV is the differential distance to a charter school

$$Z_i = \text{distance to closest charter school} - \text{distance to closest public school.}$$

Conditional on covariates, this differential distance is a valid IV (Walters, 2018). Below I discuss the data generating process in detail. All references to tables in this section implicitly refer to tables in Walters (2018).

First, I simulate covariates X (Female, Black, Hispanic, Subsidized lunch, Special education, Limited English proficiency, Value-added of closest district school, 4th grade math score, 4th grade reading score) and instrument Z and by matching Table 1 (means and standard deviations of variables).

Second, I simulate individual treatment effects and untreated potential outcome, according to

$$\begin{aligned} \tau_i = Y_i(1) - Y_i(0) &= \hat{\mu}_1 - \hat{\mu}_0 + X_i'(\hat{\lambda}_x^1 - \hat{\lambda}_x^0) + \theta_i(\hat{\lambda}_\theta^1 - \hat{\lambda}_\theta^0), \\ Y_i(0) &= \hat{\mu}_0 + X_i'\hat{\lambda}_x^0) + \theta_i\hat{\lambda}_\theta^0, \end{aligned}$$

where $(\hat{\mu}_0, \hat{\mu}_1, \hat{\lambda}_x^1 - \hat{\lambda}_x^0, \hat{\lambda}_\theta^1 - \hat{\lambda}_\theta^0)$ are taken from Table 7 (Estimates of Charter School Effects on Eighth-Grade Test Scores), and θ_i is an individual preference parameter that is distributed as a mixture of two normals ($\mathcal{N}(1.641, 0.140^2)$ with probability 0.417, and $\mathcal{N}(-1.174, 0.093^2)$ with probability 0.583 – according to Table 6, Distribution of unobservable preferences).

Third, I simulate treatment indicators D in a way that allows me to adjust the data generating process to various deviations from the null $\mathbb{H}_0$. I assume that the selection mechanism is

$$D_i(Z_i) = \mathbb{1}\{\lambda\tau_i + (1-\lambda)\xi_i - Z_i \geq 0\}.$$

ξ_i is a an individual preference shock, which is independent of all other variables and $\xi_i = 2 + \xi_i^0$, where ξ_i^0 is a standard logistic variable, and shift 2 is needed to make $\mathbb{E}[\xi_i] \approx \mathbb{E}[\tau_i]$.

Finally, it remains to simulate outcomes Y (8th grade test scores) by combining simulated treatment indicators D, potential untreated outcome $Y(0)$ and treatment effects τ, according to the outcome equation (1)

$$Y_i = Y_i(0) + \tau_i D_i.$$

1.D Appendix: redirection of power

The algorithm for the construction of critical values c_n^r is

Algorithm B3

1. generate a bootstrap sample $\{Y_i^*, p_i^*\}_i$
2. construct the test statistic

$$W^r = \max_s \frac{\sqrt{n}\Omega(s)|\gamma^*(s) - \widehat{\gamma(s)}|}{\sqrt{\widehat{V(s)}}}$$

3. calculate $c_n^r(\alpha)$ as $1 - \alpha$ quantile of W^r

2 Many Instruments Estimation and Inference under Clustered Dependence

2.1 Introduction

It is well known that two-stage least squares (2SLS) estimation is not suitable in a many instrument setting (Bekker, 1994).[8] Along with the bias-corrected version of 2SLS (Donald and Newey, 2001), a number of judicious approaches have been proposed in order to mitigate the negative consequences of high numerosity of instrumental variables when data are independent. One such approach is the limited information maximum likelihood estimator (LIML) (C. Hansen et al., 2008, van Hasselt, 2010), which, however, ceases to be consistent under conditional heteroskedasticity and requires corrections (Hausman et al., 2012). Another solution, which, in contrast, is robust to conditional heteroskedasticity, rests on the *jackknife*, also known as *leave-one-out*, idea. The *jackknife instrumental variables estimator (JIVE)* (Angrist et al., 1999, Blomquist and Dahlberg, 1999) is consistent and asymptotically normal under many instruments and independent data (Chao et al., 2012).

In clustered environments, the situation gets more complicated. In some settings, the clustered dependence can render JIVE inconsistent, while in others, JIVE remains consistent, but asymptotic inferences gets distorted. Furthermore, the asymptotic bias and inference distortions may be huge, especially when clusters are large. In this paper, we propose parameter estimation and inference tools that extend existing analogous tools under data independence to the cluster structure. There are, to our knowledge, four papers that tackle clustered dependence under many possibly weak instruments, namely Hausman et al. (2011), Chao et al. (2023), Ligtenberg and Woutersen (2024) and Ligtenberg (2025). In an early work, Hausman et al. (2011) suggest computationally intensive penalization methods, the main goal being to achieve the existence of moments. Later, Carrasco (2012) noted that the existence of moments and consistency of their procedure are not simultaneously possible. Another stab is Chao et al. (2023), who propose filtering out the individual effects by a within-type transformation and subsequently applying the jackknife methodology. However, this approach is tied to the classical one-way error component model structure of regression errors, and may not work for more general within-cluster dependencies. Indeed, Ligtenberg and Woutersen (2024) show that the approach of Chao et al. (2023) can be unreliable when cluster effects do not follow the additive error component structure.

We follow another approach to tackle the problem, which is computationally attractive and is not tied to the rigid error component model structure, allowing a more general setup, where the structural and reduced-form errors may be conditionally heteroskedastic and within-cluster heterocorrelated in an arbitrary unknown form. The proposed estimator is robust to all the mentioned features – many possibly weak instruments, conditional heteroskedasticity and clustered dependence – and combines two ideas. The first idea is familiar from the literature: to replace leave-one-out by its natural extension – the leave-cluster-out (LCO) machinery, which takes the leave-one-out idea of JIVE from the level of observations to

[8] This Chapter is joint work with Stanislav Anatolyev. Section 2.C contains material from a separate paper already submitted to a journal.

that of clusters. Similarly to Ligtenberg and Woutersen (2024) and Ligtenberg (2025), instead of jackknifing, we use its natural extension – leave-cluster-out methodology, applied to the instrument projection matrix. In the same way leave-one-out neutralizes the effect of endogeneity for the same observation in the influence function, leave-cluster-out removes the impact of endogeneity for all observations in the same cluster. Thus, leave-cluster-out allows one to dispose of the very source of problems – intra-cluster dependence in the influence function of the structural parameter's estimator.

While the leave-cluster-out component is not new and has been utilized in other contexts such as regressions with many regressors and bootstrapping, the second idea built-in in our estimator is more fresh: reweighting the observations by the inverse geometric average of corresponding cluster sizes. The main purpose of this reweighting is to soften the requirement on the growth rate of cluster sizes, and, as a consequence, to allow for larger clusters. The presence of large clusters distorts inference due to the unrestricted within-cluster dependence (see MacKinnon et al., 2023 and references therein). Moreover, even in the regression context, the cluster sizes are usually required to be by orders of magnitude smaller than the sample size (e.g. B. E. Hansen and Lee, 2019), while in the many weak IV settings, these requirements typically become even stronger. In particular, Ligtenberg (2025) has to impose the condition $n_{max}^6/K \to 0$, where n_{max} is the size of the biggest cluster, and K is a number of instruments, which asymptotically grows with a rate no larger than the sample size, and is strictly smaller when instruments are weak. Such rates are very stringent, and their practical implication is that the clusters need to be really small. The reweighting allows considerable reduction in the restrictiveness of the assumptions on cluster sizes and, therefore, larger clusters. In particular, our rate assumption reads $n_{max}/K \to 0$, which is much weaker, and hence implies a possibility of a much larger biggest cluster in practice. Technically, in our estimator, a smaller weight is assigned to observations from larger clusters, and thus the asymptotic order of the influence function becomes smaller. In the absence of this reweighting, unconstrained within-cluster dependence increases the complexity of the asymptotic variance due to increasing cluster sizes and non-negligible covariances terms. Weighting by inverse cluster sizes has been explored in the literature on randomized experiments (Cox, 1956, Athey and Imbens, 2017, Bugni et al., 2024). However, to the best of our knowledge, it has not been explicitly applied to IV settings and more generally to the asymptotic analysis of bilinear forms.

Averaging over clusters and constructing a least squares estimator of a regression using cluster-specific averages is a well-known *between estimator* used in panel data regressions (see Arellano, 2003). It exploits solely the variation between different clusters (or cross-sectional units in panel settings). In contrast, we do not normalize pairs of clusters by the product of cluster sizes, which would correspond to cluster-pairs averaging, and would be a direct extension of the between estimator to bilinear forms. Instead, we normalize pairs of clusters by the geometric average of corresponding cluster sizes, which is smaller than the product of cluster sizes. Thus, we still assign relatively larger weights to pairs of larger clusters, and therefore, in the construction of the estimator, we use not only the between-cluster-pair variation, but also the within-cluster-pair variation.

Taken together, leaving clusters out and reweighting by cluster sizes give rise to the *weighted leave-cluster-out IV (WLCOIV)* estimator. Then, we develop the associated WLCOIV variance estimator, which is a weighted cluster extension of the Chao et al. (2012)

variance formula, and also work towards the vectorization of the formula for the WLCOIV variance estimator to reduce the computational burden. Both the WLCOIV estimator and its vectorized form are computationally simple and allow for general structures of intra-cluster correlations.

Subsequently, we set out a formal asymptotic framework with the presence of many weak instruments, an increasing number of clusters, and possibly increasing cluster sizes (with a liberal rate!). Within this framework, we prove a central limit theorem for the influence function embedded in the WLCOIV estimator, thus establishing its asymptotic normality, and prove the consistency of the associated WLCOIV variance estimator. We closely follow the steps of Chao et al. (2012), who provided asymptotic analysis of JIVE under many instruments in a random sampling environment, extending their results to the clustering setup with reweighting.

Finally, we run a number of Monte-Carlo experiments in a setup featuring various cluster patterns, various kinds of instrument numerosity, and various degrees of instrument weakness. We verify the ability of the t-statistic based on WLCOIV to control size in moderately small samples, which turns out to be good except when the instruments are too weak.

The remainder of the paper is structured as follows. In Section 2, we set up the model and discuss limitations of JIVE in terms of asymptotic bias and validity of inference. In Section 3, we propose the WLCOIV estimator and develop its asymptotic variance estimator. In Section 4, we present a formal asymptotic theory of WLCOIV estimation and inference. In Section 5, we present the design and results of simulations. Finally, Section 6 concludes.

2.2 Model

2.2.1 Setup

We consider the following IV model:

$$y = X\beta + \varepsilon,$$
$$X = \Upsilon + U,$$

where y is an n-vector of the outcome variable, X is an $n \times L$ vector of explanatory variables, some of which are endogenous and some are allowed to be exogenous, β is the coefficient vector of interest, ε is an n-vector of structural errors, U is an $n \times L$ matrix of reduced-form errors, and Υ is an $n \times L$ matrix of first stage reduced form. Finally, Υ is the true reduced form and it is not observed; however, we assume that we observe an $n \times K$ matrix of instruments Z that asymptotically linearly approximate Υ well (see Assumption 4). Hence, our model allows for the nonlinear structure of the first stage. In a special case, our setup nests a linear IV model with $\Upsilon = Z\Gamma$, where Γ is a vector of coefficients. We conduct our analysis conditionally on $\mathcal{Z} = (Z, \Upsilon)$ in the spirit of Chao et al. (2012).

We allow unobservables (ε, U) to be dependent within a cluster, but require independence across clusters. Let the population be partitioned into G mutually exclusive clusters with possibly heterogeneous and asymptotically increasing sizes n_g, $g \in \{1, ..., G\}$. Let $g(i)$

be the cluster of observation i, i.e. $g(i) \in \{1, ..., G\}$, and its size be $n_{g(i)}$. In particular, we allow $\mathbb{E}[a_i b_j|\mathcal{Z}] \neq 0$ for $a \in \{\varepsilon, U\}$, $b \in \{\varepsilon, U\}$ if i and j are in the same cluster, i.e. $g(i) = g(j)$. Throughout the paper we assume without loss of generality that observations $1, ..., n$ are ordered so that they are grouped by cluster. We denote stacked-by-cluster vectors and matrices $A_{[g]} = (A_i)_{i \in g}$ for $A \in \{\varepsilon, U, X, Z\}$. Let $P = Z\left(Z'Z\right)^{-1} Z'$ be the projection matrix onto the column space of observed instruments Z. Let p_{ij} denote the (i, j)th element of P, and let $P_{[g,h]}$ denote the (g, h)th block of P, which corresponds to observations from clusters g and h.

2.2.2 Three cases against JIVE

We begin by pointing out the futility of relying on JIVE in the clustered data setup, expanding on the title of Davidson and MacKinnon (2006)'s article. For simplicity of exposition, in this section we presume that identification is strong. Let $\dot{P}$ be a modification of the projection matrix P with the main diagonal removed; i.e. $\dot{p}_{ij} = p_{ij}$ for $i \neq j$, and $\dot{p}_{ij} = 0$ for $i = j$. The JIVE estimator is

$$\hat{\beta}_{JIVE} = \left(X'\dot{P}X\right)^{-1} X'\dot{P}Y = \left(\sum_{i \neq j} p_{ij} X_i X_j'\right)^{-1} \sum_{i \neq j} p_{ij} X_i Y_j,$$

where $\sum_{i \neq j}$ is shortcut notation for $\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n}$. Then,

$$\sqrt{n}\left(\hat{\beta}_{JIVE} - \beta\right) = \left(H_{JIVE} + o_p(1)\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \neq j} p_{ij} X_i \varepsilon_j,$$

where

$$H_{JIVE} = \text{plim} \frac{1}{n} \sum_{i \neq j} p_{ij} X_i X_j'$$

is assumed to be non-singular.

The JIVE, by dropping the terms related to correlations of endogenous regressors X with the structural error ε for the same unit, removes the adverse effects of endogeneity when many instruments are present, but gives up on full utilization of information corresponding to that unit. While this information is possible to utilize under strong restrictions like conditional homoskedasticity, which is what bias-corrected 2SLS and LIML do, it appears impossible to accomplish in a non-restricted environment. Thus, in return JIVE buys the robustness to conditional heteroskedasticity under many instruments and random sampling.

We consider expansions of the JIVE influence function into linear and quadratic components

$$\frac{1}{\sqrt{n}} \sum_{i \neq j} p_{ij} X_i \varepsilon_j = \frac{1}{\sqrt{n}} \sum_{i} \left(1 - p_{ii}\right) \Upsilon_i \varepsilon_i + \frac{1}{\sqrt{n}} \sum_{i \neq j} p_{ij} U_i \varepsilon_j \equiv A_{1n}^{JIVE} + A_{2n}^{JIVE}.$$

Example 1. Let the matrix of regressors X be a vector ($L = 1$), the cluster sizes be equal, with $n = n_g G$, and suppose that all the errors are conditionally homoskedastic and within-cluster equicorrelated, i.e. $\mathbb{E}\left[U_i U_j|\mathcal{Z}\right] = \sigma_U^2 \mathbb{1}\{g(i) = g(j)\}$, $\mathbb{E}\left[\varepsilon_i \varepsilon_j|\mathcal{Z}\right] = \sigma_\varepsilon^2 \mathbb{1}\{g(i) = g(j)\}$ and $\mathbb{E}\left[U_i \varepsilon_j|\mathcal{Z}\right] = \sigma_{U\varepsilon} \mathbb{1}\{g(i) = g(j)\}$ for some constants σ_U^2, σ_ε^2 and $\sigma_{U\varepsilon}$. Let instruments be cluster indicators, $K = G$, then Z_i is a G-vector with 1 at $g(i)^{th}$ coordinate and 0's elsewhere. Clearly, $p_{ij} = 1/n_g$ if both $i, j \in g$, otherwise $p_{ij} = 0$. Then, in particular,

$$\mathbb{E}\left[A_{2n}^{JIVE}|\mathcal{Z}\right] = \frac{G}{\sqrt{n}} \sum_{i \in g \neq j \in g} n_g^{-1} \sigma_{U\varepsilon} = \frac{n - G}{\sqrt{n}} \sigma_{U\varepsilon} \sim \sqrt{n},$$

whereas $H_{JIVE} > 0$. Hence, the JIVE is asymptotically biased when cluster indicators are used as instruments. □

The case of group indicator instruments is an important empirical setup, for example, in judge designs (see Chyn et al., 2025 for a review). The key feature of the asymptotic behavior of JIVE is the biasedness of the quadratic part (A_{2n}^{JIVE}), which arises from a lot of within-cluster correlations among the structural (ε) and reduced-form (U) errors falling into the same cluster, and corresponding weights p_{ij} not vanishing fast enough asymptotically because of the instrument design. We expand this discussion after the next, conceptually different, example.

Example 2. Let the matrix of regressors X be a vector ($L = 1$), the cluster sizes be equal, with $n = n_g G$, and suppose that all the errors are conditionally homoskedastic, within-cluster equicorrelated, i.e. $\mathbb{E}\left[U_i U_j|\mathcal{Z}\right] = \sigma_U^2 \mathbb{1}\{g(i) = g(j)\}$, $\mathbb{E}\left[\varepsilon_i \varepsilon_j|\mathcal{Z}\right] = \sigma_\varepsilon^2 \mathbb{1}\{g(i) = g(j)\}$ and $\mathbb{E}\left[U_i \varepsilon_j|\mathcal{Z}\right] = \sigma_{U\varepsilon} \mathbb{1}\{g(i) = g(j)\}$ for some constants $\sigma_{U\varepsilon}$, σ_U^2 and σ_ε^2. Let the elements of Z except the first one be i.i.d. standard normal, and the first one equal 1. Then, the off-diagonal elements of P, by symmetry, possess the same marginal distribution. The properties of the linear component in the expansion for the JIVE are the following :

$$\mathbb{E}\left[A_{1n}^{JIVE}|\mathcal{Z}\right] = 0, \qquad \mathbb{V}\left(A_{1n}^{JIVE}|\mathcal{Z}\right) \sim n_g,$$

and the properties of the quadratic component are:

$$\mathbb{E}\left[A_{2n}^{JIVE}|\mathcal{Z}\right] \sim \frac{n_g}{\sqrt{n}}, \qquad \mathbb{V}\left(A_{2n}^{JIVE}|\mathcal{Z}\right) \sim n_g^2$$

(see Appendix Q1–Q2 for derivations). Hence, clustering induces a (higher-order) bias into the quadratic component, leading to a bias for the JIVE of order $n^{-1/2} n_g/\sqrt{n} = G^{-1}$, which is of smaller order than n^{-1} when clusters grow asymptotically. Furthermore, the linear component A_{1n}^{JIVE} and the quadratic component A_{2n}^{JIVE} are in general correlated (see Appendix Q3). Thus, in this example, the JIVE is consistent, under some restriction on the rate of growth of n_g in case it does increase asymptotically. The JIVE influence function is not clean but is contaminated by various factors that complicate modification and/or pivotization of the JIVE. □

In the two examples above, the orders of the bias term are so different that they resulted in JIVE inconsistency in one case and its consistency in the other. The JIVE (in)consistency depends largely on the instrument design. Consider a more general conditionally heteroskedastic environment, and denote $\sigma_{ij}^{eu} = \mathbb{E}\left[e_i U_j | \mathcal{Z}\right]$ for i and j within the same cluster. Then, a necessary condition for JIVE consistency is

$$\sum_{i=1}^{n} \sum_{\substack{j \neq i \\ g(j)=g(i)}} p_{ij} \sigma_{ij}^{u\varepsilon} = o(n). \tag{48}$$

This does not hold in Example 1, where p_{ij}'s for $i, j \in g$ don't vanish fast enough and $\sigma_{ij}^{u\varepsilon}$ does not depend on (i, j), so that $\sum_{i \neq j, g(j)=g(i)} p_{ij} = O(n)$. However, this condition does hold in Example 2. Heuristically, the instrument design in Example 1 results in all non-zero elements in $\dot{P}$, although individually small, being of the same sign, so en mass they weigh much higher than an analogous sum of elements of $\dot{P}$ in Example 2, as these tend to be distributed around an asymptotically zero mean.[9]

Even when the JIVE is consistent, the use of the standard JIVE variance estimator without accounting for clustering can lead to biases and, as a result, to inferential distortions when the clusters are asymptotically increasing, beyond those that could be expected when a wrong variance estimator is used. The pivotization of JIVE is unable to keep up with asymptotically increasing cluster size.

To summarize, the three cases against JIVE are: (1) in some setups, JIVE is inconsistent; (2) when JIVE is consistent, under-independence JIVE inference is invalid; (3) even when JIVE is consistent, the linear and quadratic terms in the JIVE influence function are correlated and the quadratic term has a non-zero (higher-order) bias, which greatly complicates asymptotic derivations and pivotization. These properties of JIVE motivate one to deviate from such a construct instead of attempting to "save" it, which is arguably impossible if one does not impose strong restrictions on the within-cluster correlation structure, and resort to the leave-cluster-out machinery instead.

2.3 Weighted leave-cluster-out IV

2.3.1 Leaving cluster out

As an extension of the leave-one-out idea, consider the "intermediate" *leave-cluster-out IV (LCOIV)* estimator $\hat{\beta}_{LCOIV}$. Let $\mathring{P}$ be a modification of the projection matrix P with the diagonal blocks (corresponding to pairs of indices from the same clusters) removed; i.e. $\mathring{p}_{ij} = p_{ij}$ for (i, j) such that $g(i) \neq g(j)$, and $\mathring{p}_{ij} = 0$ for (i, j) such that $g(i) = g(j)$. Thus, $\hat{\beta}_{LCOIV}$ excludes the terms that correspond to observations within the same cluster:

$$\hat{\beta}_{LCOIV} = \left(X' \mathring{P} X\right)^{-1} X' \mathring{P} Y = \left(\sum_{\substack{i,j \\ g(i) \neq g(j)}} p_{ij} X_i X_j' \right)^{-1} \sum_{\substack{i,j \\ g(i) \neq g(j)}} p_{ij} X_i y_j, \tag{49}$$

[9] The sum of elements of $\dot{P}$ equals $\sum_{j \neq i} p_{ij} = \sum_{i,j} p_{ij} - \sum_i p_{ii} = n - K$, so their average is $(n - K) / \left(n^2 - n\right)$, which asymptotically ofs of order n^{-1}.

where $\sum_{\substack{i,j \\ g(i)\neq g(j)}} = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j:g(j)\neq g(i)}}^{n}$, a sum over all i and j such that ith cluster is different from jth cluster. Then,

$$\sqrt{n}\left(\hat{\beta}_{LCOIV} - \beta\right) = \left(H_{LCOIV} + o_p\left(1\right)\right)^{-1} \frac{1}{\sqrt{n}} \sum_{\substack{i,j \\ g(i)\neq g(j)}} p_{ij} X_i \varepsilon_j,$$

where

$$H_{LCOIV} = \text{plim} \frac{1}{n} \sum_{\substack{i,j \\ g(i)\neq g(j)}} p_{ij} X_i X_j'$$

is assumed to be non-singular.

As mentioned in Section 2.2, in some situations JIVE is inconsistent, and this inconsistency originates from the correlation in pairs of observations that belong to the same cluster. These observations are omitted in the construction of the LCOIV estimator and therefore (48) is satisfied; as a result, the correlations within the cluster do not bias the estimates.

Example 2 *(continued)*. For the example with random instruments, the LCOIV estimator, out of $n^2 - n = n\left(n-1\right)$ off-diagonal elements, removes $G\left(n/G\right)^2 - n = n\left(n/G - 1\right)$ more elements, the proportion of $\left(n/G - 1\right) / \left(n-1\right)$, which is asymptotically inversely proportional to G. □

The influence function of the LCOIV estimator contains two terms:

$$\frac{1}{\sqrt{n}} \sum_{\substack{i,j \\ g(i)\neq g(j)}} p_{ij} X_i \varepsilon_j = \frac{1}{\sqrt{n}} \sum_{\substack{i,j \\ g(i)\neq g(j)}} p_{ij} \Upsilon_i \varepsilon_j + \frac{1}{\sqrt{n}} \sum_{\substack{i,j \\ g(i)\neq g(j)}} p_{ij} U_i \varepsilon_j \equiv A_{1G}^{LCOIV} + A_{2G}^{LCOIV}.$$

The orders of these two terms differ sharply from those for JIVE, hence resulting in different asymptotic properties of the LCOIV. In particular, leave-cluster-out removes any bias in A_{2n}^{LCOIV}:

$$\mathbb{E}\left[A_{2n}^{LCOIV} | \mathcal{Z}\right] = \frac{1}{\sqrt{n}} \sum_{\substack{i,j \\ g(j)\neq g(i)}} p_{ij} \mathbb{E}\left[U_i \varepsilon_j | \mathcal{Z}\right] = 0,$$

because $\mathbb{E}\left[U_i \varepsilon_j | \mathcal{Z}\right] = 0$ whenever i and j belong to different clusters. Also, leave-cluster-out makes the two components A_{1n}^{LCOIV} and A_{2n}^{LCOIV} uncorrelated:

$$\mathbb{C}\left(A_{1n}^{LCOIV}, A_{2n}^{LCOIV} | \mathcal{Z}\right) = \frac{1}{n} \sum_{\substack{i,j \\ g(j)\neq g(i)}} p_{ij} \Upsilon_i \sum_{\substack{k,\ell \\ g(\ell)\neq g(k)}} p_{k\ell} \mathbb{E}\left[\varepsilon_j U_k \varepsilon_\ell | \mathcal{Z}\right] = 0,$$

because $\mathbb{E}\left[\varepsilon_j U_k \varepsilon_\ell | \mathcal{Z}\right] = 0$ whenever k and ℓ belong to different clusters, regardless of which cluster j belongs to. This is a very desirable feature that is not shared by JIVE, which greatly facilitates the derivation of proper inference tools.

To summarize, LCOIV goes further than JIVE and drops, along the terms related to correlations of endogenous regressors with the structural errors for the same observation, the terms related to such correlations for the other observations in the same cluster. In exchange, the estimation gains robustness to arbitrary patterns of within-cluster conditional heterocorrelatedness.

2.3.2 Weighting by cluster sizes

In addition, on top of the leaving-cluster-out idea, we also weight pairs of observations by the geometric mean of the corresponding cluster sizes. This reweighting aligns the amount of information coming from observations in different clusters, so that larger clusters still get more weight than smaller clusters, though the difference is smaller than in the unweighted case, where larger clusters receive larger weights proportionally to their size. Both with and without this reweighting, the variance of the estimator grows with the largest cluster size; however, for the weighted version this growth is much slower.

The main purpose of this reweighting is to soften the requirement on the growth rate of cluster sizes, and, as a consequence, to allow for larger clusters. The reweighting allows considerable reduction of the restrictiveness of the assumptions on cluster sizes and therefore allows for larger clusters. Our estimator mechanically assigns smaller weights to observations from larger clusters, and thus the asymptotic order of the influence function becomes smaller, while in the absence of such reweighting, unconstrained within-cluster dependence increases the complexity of the asymptotic variance due to increasing cluster sizes and non-negligible covariances.

Taken together, leaving clusters out and reweighting by cluster sizes lead to the *weighted leave-cluster-out IV (WLCOIV)* estimator:

$$\hat{\beta}_{WLCOIV} = \left(X'\breve{P}X\right)^{-1} X'\breve{P}Y,$$

where $\breve{P}$ is a reweighted leave-cluster-out projection matrix with a typical element $\breve{p}_{ij} = \mathring{p}_{ij}/\sqrt{n_{g(i)}n_{g(j)}}$, i.e. $\breve{p}_{ij} = p_{ij}/\sqrt{n_{g(i)}n_{g(j)}}$ if $g(i) \neq g(j)$ and $\breve{p}_{ij} = 0$ if $g(i) = g(j)$. Then,

$$\hat{\beta}_{WLCOIV} = \left(\sum_{\substack{i,j \\ g(i)\neq g(j)}} p_{ij} X_i X_j' / \sqrt{n_{g(i)}n_{g(j)}} \right)^{-1} \sum_{\substack{i,j \\ g(i)\neq g(j)}} p_{ij} X_i y_j / \sqrt{n_{g(i)}n_{g(j)}}.$$

Thus, WLCOIV combines both leave-cluster-out methodology with weighting the observations on X and y by an inverse square root of the size of the cluster, to which these observations belong.

Due to the reweighting, the WLCOIV numerator is a sum of pair-of-clusters-specific averages, and the contribution of each pair of clusters to the asymptotic variance is of the same order (i.e. independent of cluster sizes). Hence, this weighting scheme allows for larger clusters to be accommodated by our framework. In particular, this normalization by cluster sizes allows us to relax the typical assumptions for the asymptotic theory we formally lay out in Section 4.

2.3.3 Asymptotic variance estimation

The population variance of $\hat{\beta}_{WLCOIV}$ has a sandwich form

$V_G = H_{WLCOIV}^{-1} \Sigma H_{WLCOIV}^{-1}.$

The population variance Σ of the non-normalized WLCOIV influence function is, by definition,

$$\mathbb{E}\left[\Big(\sum_{\substack{i,j \\ g(i)\neq g(j)}} \breve{p}_{ij} X_i \varepsilon_j \Big)^2 | \mathcal{Z}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \sum_{\substack{j=1 \\ g(j)\neq g(i)}}^{n} \sum_{k=1}^{n} \sum_{\substack{\ell=1 \\ g(\ell)\neq g(k)}}^{n} \breve{p}_{ij}\breve{p}_{k\ell} X_i X_k' \varepsilon_j \varepsilon_\ell | \mathcal{Z}\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} \sum_{\substack{j=1 \\ g(j)\neq g(i)}}^{n} \sum_{\substack{k=1 \\ g(k)=g(j)}}^{n} \sum_{\substack{\ell=1 \\ g(\ell)=g(i)}}^{n} \breve{p}_{ij}\breve{p}_{k\ell} X_i X_k' \varepsilon_j \varepsilon_\ell \right.$$

$$\left. + \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{\substack{j=1 \\ g(j)\notin \{g(i),g(k)\}}}^{n} \sum_{\substack{\ell=1 \\ g(\ell)=g(j)}}^{n} \breve{p}_{ij}\breve{p}_{k\ell} X_i X_k' \varepsilon_j \varepsilon_\ell | \mathcal{Z}\right].$$

It consists of two parts, the first representing within-cluster association between the reduced-form errors U (embedded in X) and structural errors ε, while the second one – among the reduced-form errors U (embedded in X) and among the structural errors ε. Depending on the strength of the instruments, the two parts can be of the same order asymptotically, or the second part might dominate. The population variance above can be straightforwardly estimated by its sample analog

$$\hat{\Sigma} = \sum_{i} \sum_{\substack{j \\ g(j)\neq g(i)}} \sum_{\substack{k \\ g(k)=g(j)}} \sum_{\substack{\ell \\ g(\ell)=g(i)}} \breve{p}_{ij}\breve{p}_{k\ell} X_i X_k' \hat{\varepsilon}_j \hat{\varepsilon}_\ell \tag{50}$$

$$+ \sum_{i} \sum_{k} \sum_{\substack{j \\ g(j)\notin \{g(i),g(k)\}}} \sum_{\substack{\ell \\ g(\ell)=g(j)}} \breve{p}_{ij}\breve{p}_{k\ell} X_i X_k' \hat{\varepsilon}_j \hat{\varepsilon}_\ell, \tag{51}$$

where $\hat{\varepsilon}_i$ is a residual $\hat{\varepsilon}_i = y_i - X_i'\hat{\beta}_{WLCOIV}$. The "bun" matrix H_{WLCOIV} can be estimated by

$$\hat{H}_{WLCOIV} = \sum_{\substack{i,j \\ g(i)\neq g(j)}} \breve{p}_{ij} X_i X_j'. \tag{52}$$

Thus, the variance estimator of $\hat{\beta}_{WLCOIV}$ is

$$\hat{V}_G = \hat{H}_{WLCOIV}^{-1} \hat{\Sigma} \hat{H}_{WLCOIV}^{-1}. \tag{53}$$

While the influence function variance estimator (50)-(51) and the WLCOIV variance estimator (53) can be calculated directly, for larger G the computational burden grows rapidly. Hence, similarly to Chao et al. (2012), we propose a vectorized variance formula that has higher computational efficiency. Using $\sum_{\substack{i\in g \\ j\in h}} \breve{p}_{ij} X_i X_j' = X_{[g]}' \breve{P}_{[g,h]} X_{[h]}$, we obtain

$$\hat{H}_{WLCOIV} = X'\breve{P}X - \sum_{g} X_{[g]}' \breve{P}_{[g,g]} X_{[g]},$$

which is more computationally efficient than the formula in (52). Similarly, the variance estimator for the influence function can be represented by

$$\begin{aligned}\hat{\Sigma} &= \sum_{\substack{g,h \\ g \neq h}} X_{[g]}' \breve{P}_{[g,h]} \hat{\varepsilon}_{[h]} \hat{\varepsilon}_{[g]}' \breve{P}_{[g,h]} X_{[h]} + \sum_{\substack{g,h,g' \\ g' \notin \{g,h\}}} X_{[g]}' \breve{P}_{[g,g']} \hat{\varepsilon}_{[g']} \hat{\varepsilon}_{[g']}' \breve{P}_{[g',h]} X_{[h]} \\ &= \sum_{g} \frac{\bar{X}_{[g]}' \hat{\varepsilon}_{[g]} \hat{\varepsilon}_{[g]}' \bar{X}_{[g]} - X_{[g]}' P_{[g,g]} \hat{\varepsilon}_{[g]} \hat{\varepsilon}_{[g]}' \bar{X}_{[g]} / \sqrt{n_g} - \bar{X}_{[g]}' \hat{\varepsilon}_{[g]} \hat{\varepsilon}_{[g]}' P_{[g,g]} X_{[g]} / \sqrt{n_g}}{n_g},\end{aligned}$$

where $\bar{X}_{[g]} = \sum_{h=1}^{G} P_{[g,h]} X_{[h]} / \sqrt{n_h}$. These vectorized formulae replace multiple summations across observations by summations across clusters, which usually results in considerably fewer summands. As a result, the computation time can be reduced substantially.

2.4 Asymptotic theory

We combine several asymptotic frameworks, including many instruments asymptotics (see Bekker, 1994), many weak instrument asymptotics (see Chao and Swanson, 2005), and many clusters asymptotics (see e.g. B. E. Hansen and Lee, 2019). The majority of proofs extend the cross-sectional many weak instruments methodology of Chao et al. (2012) to the case of many weak instruments under clustering. Fundamentally, we require the number of clusters to grow with the sample size and we allow for heterogeneous unbounded cluster sizes, weak identification, and the number of instruments growing with sample size.

The structure of this section is as follows: we gradually introduce the framework and assumptions and state our results. All proofs are in the appendix. Theorem 1 shows consistency. Theorems 2-3 postulate asymptotic normality of $\hat{\beta}_{WLCOIV}$ under strong identification and under weak identification. In Section 2.3.3, Theorem 4 provides the consistency of our variance estimator.

Assumption 1. $G \to \infty$, $K = K_G \to \infty$. As $n \to \infty$, $\max_{g=1,\dots,G} \frac{n_g}{K} \to 0$. Z includes among its columns a vector of ones, $\text{rank}(Z) = K$, and for some, $C < 1$ $p_{ii} \leq C$ $\forall$ i a.s.n; $||P_{[g,g]}|| \leq C < 1$ $\forall$ g a.s.

Assumption 1 formalizes our asymptotic framework. Specifically, Assumption 1 requires the number of clusters G and the number of instruments K to grow with the sample size. We do not require cluster sizes n_g to be of the same size, and allow for their growth but restrict the growth rate relative to the number of instruments K. Notice that in few instruments cases, the largest cluster has the growth rate limited by the sample size (see B. E. Hansen and Lee, 2019), while here it is limited by the number of instruments. Moreover, Assumption 1 mildly restricts the instrumental design such that Z contains a constant term and the diagonal of the projection matrix $P = Z(Z'Z)^{-1}Z'$ is bounded away from 1 a.s. for large sample size n. The latter is a usual assumption that is made in the many instruments literature (see e.g. Chao et al., 2012) to ensure well-defined design. The former can be relaxed such that columns of Z span the unity vector (see e.g. Kolesár, 2013), and is not restrictive with many instruments. Finally, Assumption 1 imposes certain regularity conditions on diagonal blocks of P that correspond to clusters. In particular, we require that the spectral norms of these blocks are bounded away from 1. Notice that $||P|| = 1$, by the properties of P, and $p_{ii} < 1$ by assumption are two limiting cases with $n_g = n$ and $n_g = 1$, respectively. Then one might think about $||P_{[g,g]}|| < 1$ as a reasonable extension to intermediate cases. $||P_{[g,g]}|| < 1$ further implies that the norm of the corresponding residual-maker matrix is also bounded by 1: $||M_{[g,g]}|| = ||I_{n_g} - P_{[g,g]}|| \leq 1 \; \forall \; g$ a.s. by the eigenvalue interlacing theorem.

While Assumption 1 provides high-level conditions on the projection matrix P, in Appendix C we characterize the joint asymptotic behavior of diagonal and off-diagonal elements of P in terms of low-level conditions. Specifically, we show that under the rotated i.i.d. assumption for elements of Z, the rate of convergence is $\sqrt{n}$, and the limiting distribution is multivariate centered Gaussian. The formulas for the asymptotic variances and covariances are expressed as functions of α and moments of elements of Z. The instrumental tools in deriving the asymptotic results are the Woodbury matrix identity, a central limit theorem for quadratic forms, and elements of the random matrix theory such as the Marchenko-Pastur law. These results in Appendix C provide novel insights into the structure of large projection matrices.

Assumption 2. Let $i \in g$. $\Upsilon_i = S_G z_i/\sqrt{n}$ where $S_G = \tilde{S}_G \text{diag}(\mu_{1G}, ..., \mu_{LG})$, $\tilde{S}_G$ is $L \times L$ and bounded, and the smallest eigenvalue of $\tilde{S}_G \tilde{S}_G'$ is bounded away from zero. Also, for each j, either $\mu_{jG} = \sqrt{n}$ or $\mu_{jG}/\sqrt{n} \to 0$, $r_G = (\min_{1 \leq j \leq G} \mu_{jG})^2 \to \infty$, $\sqrt{K}/r_G \to 0$. Also, there is $C > 0$ such that $\sum_{g=1}^G \left|\left|z_{[g]} z_{[g]}'/n_g\right|\right|/n \leq C$ and $\lambda_{\min}\left(\sum_{i=1}^n z_i z_i'/n\right) \geq 1/C$ a.s.n.

Assumption 2 defines the structure of the reduced form Υ and somewhat follows a similar assumption to C. Hansen et al. (2008) and Chao et al. (2012). Most importantly, Assumption 2 introduces the S_G structure that governs the identification strength for every variable in the structural equation, and therefore convergence rates of the WLCOIV estimator for different structural parameters. The fastest convergence rate is conventional semiparametric $\sqrt{n}$, but more generally it is allowed to be slower up to some $\sqrt{r_G}$. r_G governs the strength of identification and is a version of the concentration parameter. Indeed, r_G simplifies to the concentration parameter in the linear case, i.e. $\Upsilon = Z\Pi$. The ratio of the number of instruments and the concentration parameter is required to be restricted by $\sqrt{K}/r_G \to \infty$, as in the cross-sectional setup (Chao et al., 2012). Essentially our inverse-geometric-mean-cluster-size weighting allows us to rely on the cross-sectional assumptions without strengthening them. Moreover, Assumption 2 imposes regularity conditions on the reduced form. Specifically, it bounds the variance of the explained part of the reduced form both from above and below.

Assumption 3. There is a constant, C, such that conditional on $\mathcal{Z} = (\Upsilon, Z)$, the observations $(\varepsilon_{[g]}, U_{[g]})(g = 1, ..., G)$ are independent, with $\mathbb{E}[\varepsilon_i|\mathcal{Z}] = 0$, $\mathbb{E}[U_i|\mathcal{Z}] = 0 \; \forall \; i$, $\sup_i \mathbb{E}[\varepsilon_i^2|\mathcal{Z}] < C$, and $\sup_i \mathbb{E}[||U_i||^2|\mathcal{Z}] \leq C$ a.s.

Assumption 3 formally defines the usual cluster structure by requiring independence of unobservable errors between clusters, while allowing for their unrestricted dependence within clusters. Further, Assumption 3 imposes standard regularity conditions on the variance of error terms, namely mean zero and finite variances.

Assumption 4. There is a π_K such that $\sum_i ||z_i/\sqrt{n_{g(i)}} - Z_i\pi_K||^2/n \to 0$ a.s.

Assumption 4 ensures that unknown and potentially non-linear reduced form is well approximated by a linear combination of instruments Z. This assumption is analogous to the corresponding cross-sectional assumption. Assumption 4 is automatically satisfied under linear reduced form. Under Assumptions 1-4, the WLCOIV estimator is consistent.

Theorem 1. Suppose Assumptions 1-4 are satisfied. Then, $r_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta) \xrightarrow{p} 0$.

Proof in the Appendix.

Assumption 5. There is a constant, $C > 0$, such that $\sum_{g=1}^{G} ||z_{[g]}/\sqrt{n_g}||^4/n^2 \to 0$, $\sup_i \mathbb{E}[\varepsilon_i^4|\mathcal{Z}] \le C$, and $\sup_i \mathbb{E}[||U_i||^4|\mathcal{Z}] \le C$ a.s.

Assumption 5 further imposes some additional regularity conditions on the higher moments of the reduced form and errors, required for the asymptotic normality.

Under Assumptions 1-5 and additional regularity conditions, the WLCOIV estimator is asymptotically normal. Theorem 2 formalizes this result for the case of strong identification (bounded K/r_G). For the case of weak identification ($K/r_G \to \infty$), Theorem 3 formalizes the asymptotic normality of $\hat{\beta}_{WLCOIV}$. Under strong identification, similar to the results for cross-sectional data in Chao et al. (2012), the asymptotic variance of the linear and the quadratic terms of the WLCOIV influence function are asymptotically balanced, i.e. of the same order. Let

$$\Psi_G = S_G^{-1} \sum_{\substack{g,g' \\ g' \neq g}} \sum_{\substack{i,k \in g \\ j,\ell \in g'}} (\mathbb{E}[U_i U_k'|\mathcal{Z}]\mathbb{E}[\varepsilon_j \varepsilon_\ell|\mathcal{Z}] + \mathbb{E}[U_i \varepsilon_k|\mathcal{Z}]\mathbb{E}[\varepsilon_j U_\ell'|\mathcal{Z}])p_{ij}p_{k\ell}S_G^{-1\prime}/(n_g n_{g'}),$$

$$\Omega_G = \sum_g z_{[g]}'\mathbb{E}[\varepsilon_{[g]}\varepsilon_{[g]}'|\mathcal{Z}]z_{[g]}/n_g^2.$$

Theorem 2. Suppose that Assumptions 1-5 are satisfied,

$$\lambda_{\min}\left(\sum_g z_{[g]}'(I_{n_g} - P_{[g,g]})\mathbb{E}\left[\varepsilon_{[g]}\varepsilon_{[g]}'|\mathcal{Z}\right](I_{n_g} - P_{[g,g]})z_{[g]}\right)/n \ge C > 0 \quad a.s.,$$

and K/r_G is bounded. Then V_G is nonsingular a.s.n, and

$$V_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta) \xrightarrow{d} \mathcal{N}(0, I_L).$$

Proof in the Appendix.

Under weak identification, the quadratic term dominates the asymptotic variance. Let L_G be a sequence of $K \times L$ matrices.

Theorem 3. Suppose that Assumptions 1-5 are satisfied and $K/r_G \to \infty$. Suppose further that L_G is bounded and there is $C > 0$ such that $\lambda_{\min}(L_G V_G^* L_G') \geq C$ a.s.n. Then

$$(L_G V_G^* L_G')^{-1/2} L_G \sqrt{r_G/K} S_G' (\hat{\beta}_{WLCOIV} - \beta) \xrightarrow{d} \mathcal{N}(0, I_L).$$

Proof in the Appendix.

Notice that under strong identification, by Theorem 2, the asymptotic variance of the normalized WLCOIV is

$$V_G = H_{WLCOIV}^{-1}(\Omega_G + \Psi_G) H_{WLCOIV}^{-1},$$

i.e. the linear and quadratic terms are in balance asymptotically and are of the same order. On the contrary, under weak identification, by Theorem 3,

$$V_G^* = H_{WLCOIV}^{-1}(r_G/K)\Psi_G H_{WLCOIV}^{-1},$$

i.e. the quadratic term dominates asymptotically.

Assumption 6. For any $\eta_{[g]}$ such that $\mathbb{E}[||\eta_{[g]}||^2 | \mathcal{Z}] \leq C n_g^2$, and for conformable unit vectors e_{t_1} and e_{t_2}

$$\begin{aligned} &\sum_{h \neq g \neq m} e_{t_1}' S_G^{-1} X_{[h]}' P_{[h,g]} \eta_{[g]} P_{[g,m]} X_{[g]} S_G^{-1\prime} e_{t_2} / \sqrt{n_h n_m n_g^2} \\ &- \mathbb{E}\left[\sum_{h \neq g \neq m} e_{t_1}' S_G^{-1} X_{[h]}' P_{[h,g]} \eta_{[g]} P_{[g,m]} X_{[g]} S_G^{-1\prime} e_{t_2} / \sqrt{n_h n_m n_g^2} | \mathcal{Z} \right] \xrightarrow{p} 0. \end{aligned}$$

Assumption 6 requires that the expected square of the linear term is close to its expectation in the sense of convergence in probability. This additional assumption allows us to prove the consistency of the variance estimators under both strong and weak identification in Theorem 4. Armed with Theorems 2-4, one can use pivotized t-statistics for inference on coefficients β.

Theorem 4. Suppose that Assumptions 1–6 are satisfied. If $n_{\max} K / r_G$ is bounded, then $S_G' \hat{V} S_G - V_G \xrightarrow{p} 0$. Also if $n_{\max} K / r_G \to \infty$, then $r_G / n_{\max} S_G' \hat{V} S_G / K - V_G^* \xrightarrow{p} 0$.

Proof in the Appendix.

2.5 Simulations

In this section, we illustrate the performance of our proposed WLCOIV estimator and the validity of inference based on it. We fix the nominal size at 5% and run a number of experiments. Table (2) reports the actual size of the WLCOIV-based t-tests. Different columns correspond to different identification strengths, starting from the "very weak IV case" in column 1 ($r_G = 4$) to the "very strong IV case" in column 4 ($r_G = 64$). We also vary total sample size: $N \in \{200, 500\}$; and homogeneous cluster sizes: $n_g \equiv \text{const} \in \{2, 4, 10\}$.

For simplicity, let $A_{gi} = A_i$ where $i \in g$ for $A \in \{Z, u, \varepsilon\}$. The setup of our simulations incorporates the additive error component clustered structure of instrumental variables:

$$Z_{gi} = \zeta_g + \zeta_{gi}, \qquad \zeta_g \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad \zeta_{gi} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Further, our setup naturally incorporates clustering of error terms:

$$\begin{aligned} u_{gi} &= v_{gi} + v_{gi}^{u} + \xi_g, & \varepsilon_{gi} &= v_{gi} + v_{gi}^{\varepsilon} + \xi_g, \\ \xi_g &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \zeta_g^2\right), & v_{gi} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \zeta_{gi}^2\right), \\ v_{gi}^{u} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), & v_{gi}^{\varepsilon} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \end{aligned}$$

The common component $\xi_g + v_{gi}$ introduces both clustering and endogeneity by inducing non-zero correlation of errors terms in the structural equation and the first stage. Finally, $\mathbb{V}(\xi_g)$ and $\mathbb{V}(v_{gi})$ vary across cluster and individuals, respectively, thus further introducing heteroskedasticity.

Column 4 in Table (2) indicates that for very strong identification, WLCOIV-based inference provides approximately accurate inferences even in relatively small samples ($N = 200$) with relatively large clusters ($n_g = 10$). As the strength of identification decreases, the size control is still reasonable with slightly liberal inferences in many cases (columns 2-3). For the case of very weak instruments, the WLCOIV-based inferences become overly liberal, resulting in the actual rejection rates of between 15% and 25%.

		(1)	(2)	(3)	(4)
		$r_G = 4$	$r_G = 16$	$r_G = 32$	$r_G = 64$
$N = 200$	$n_g = 2$	0.171	0.068	0.062	0.054
	$n_g = 4$	0.209	0.069	0.065	0.077
	$n_g = 10$	0.250	0.099	0.089	0.065
$N = 500$	$n_g = 2$	0.153	0.048	0.054	0.040
	$n_g = 4$	0.168	0.055	0.062	0.049
	$n_g = 10$	0.210	0.062	0.064	0.055

Table 2: Actual rejection rates

2.6 Application

In this section we illustrate the performance of LCOIV and WLCOIV estimators with a classical empirical example and compare it to 2SLS and JIVE. Specifically, we apply our estimators to two sets of instruments from the seminal returns to education study by Angrist and Krueger (1991). In one of the setups we use quarter of birth dummies as instruments for education (30 instruments in total). In another one, we augment the set of instruments by state of birth dummies and their interactions with quarter of birth (180 instruments in total). We treat states as clusters. Table 3 presents the results. WLCOIV provides a much larger effect than 2SLS, and the effect is quite precise, and stable. JIVE is very unstable in this setup and provides point estimates that differ dramatically in these two setups. Standard errors for WLCOIV and 2SLS are robust to clustering, but WLCOIV standard errors are also robust to many weak instruments. Overall, WLCOIV is precise and gives stable point estimates. Finally, notice that LCOIV and WLCOIV give similar point estimates in both designs. The standard error of the WLCOIV are slightly smaller with 180 instruments than the standard error of LCOIV, while the converse is true with 30 instruments.

Instruments set	2SLS	JIVE	LCOIV	WLCOIV
30 instruments	0.0600	0.0860	0.1300	0.1369
			(0.0210)	(0.0301)
180 instruments	0.0635	0.1329	0.1113	0.1101
			(0.0254)	(0.0180)

Table 3: Angrist and Krueger (1991) re-estimation of $\hat{\beta}$; s.e. in parentheses; 2SLS and JIVE standard errors are not robust for this setting

2.7 Conclusion

In this paper we propose a novel weighted leave-cluster-out – WLCOIV – estimator, suitable for many weak instruments under clustering. We formally derive its asymptotic properties under clustering and the possible presence of many weak instruments. We show that the WLCOIV estimator is consistent and asymptotically normal. Our procedure is robust to heterogeneous clusters that can grow with sample size. We suggest a consistent estimator the asymptotic variance for the WLCOIV estimator such that the inferences based on a pivotized t-statistic are valid. Additionally, we demonstrate the performance of WLCOIV in the context of Angrist and Krueger (1991). We find that WLCOIV is precise compared to competing estimators.

2.Q Appendix: derivations for Example 2

Q1: expectations. Obviously, for the linear part,

$$E\left[A_{1n}^{JIVE}\right] = 0.$$

For the quadratic part, note that due to the presence of a constant among the instruments,

$$\sum_i \sum_{j \neq i} p_{ij} = \sum_i \sum_j p_{ij} - \sum_i p_{ii} = n - \ell.$$

Due to the symmetry across off-diagonal elements,

$$\mathbb{E}\left[\sum_i \sum_{\substack{j\neq i\\ g(j)=g(i)}} p_{ij}\right] = \frac{G\left(n_g^2-n_g\right)}{n^2-n}\mathbb{E}\left[\sum_i\sum_{j\neq i} p_{ij}\right] \sim n_g,$$

where $G\left(n_g^2-n_g\right)$ is the number of off-diagonal elements in the double sum of interest, and n^2-n is the total number of off-diagonal elements in P. Therefore,

$$\mathbb{E}\left[A_{2n}^{JIVE}\right] = \frac{\sigma_{U\varepsilon}}{\sqrt{n}}\mathbb{E}\left[\sum_i \sum_{\substack{j\neq i\\ g(j)=g(i)}} p_{ij}\right] \sim \frac{n_g}{\sqrt{n}}.$$

Q2: variances.

For the linear part, the conditional variance is

$$\begin{aligned}
\mathbb{V}\left(A_{1n}^{JIVE}|\mathcal{Z}\right) &= \mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_i(1-p_{ii})\Upsilon_i\varepsilon_i\right)\left(\frac{1}{\sqrt{n}}\sum_j(1-p_{jj})\Upsilon_j\varepsilon_j\right)|\mathcal{Z}\right]\\
&= \frac{1}{n}\sum_i \sum_{\substack{j\neq i\\ g(j)=g(i)}} (1-p_{ii})(1-p_{jj})\Upsilon_i\Upsilon_j\mathbb{E}\left[e_ie_j\right] + \frac{1}{n}\sum_i(1-p_{ii})^2\Upsilon_i^2\mathbb{E}\left[e_i^2\right]\\
&\sim \frac{1}{n}n(n_g-1) + \frac{1}{n}n \sim n_g.
\end{aligned}$$

For the quadratic part, the conditional variance $\mathbb{V}\left(A_{2n}^{JIVE}|\mathcal{Z}\right)$ has the asymptotic order of the following sums:

$$\sum_i \sum_{\substack{j\\ g(j)\neq g(i)}} \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\\ g(\ell)=g(j)}} m_{ijk\ell}p_{ij}p_{k\ell} \quad \text{and} \quad \sum_i \sum_{\substack{j\neq i\\ g(j)=g(i)}} \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\neq k\\ g(\ell)=g(j)}} m_{ijk\ell}p_{ij}p_{k\ell},$$

where $m_{ijk\ell}$ is a four-tuple error moment, which is uniformly bounded in absolute value by, say, C_m. Orders of sums of both types can be bounded from above:

$$\begin{aligned}
\left|\sum_i \sum_{\substack{j\\ g(j)\neq g(i)}} \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\\ g(\ell)=g(j)}} m_{ijk\ell}p_{ij}p_{k\ell}\right| &\leq \left(\sum_i \sum_{\substack{j\\ g(j)\neq g(i)}} \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\\ g(\ell)=g(j)}} C_m p_{ij}^2\right)^{1/2}\\
&\quad\times\left(\sum_i \sum_{\substack{j\\ g(j)\neq g(i)}} \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\\ g(\ell)=g(j)}} C_m p_{k\ell}^2\right)^{1/2}\\
&\leq C_m\left(n_g^2\sum_i\sum_j p_{ij}^2\right)^{1/2}\left(n_g\sum_k n_g\sum_\ell p_{k\ell}^2\right)^{1/2}\\
&= C_m\left(n_g^2\ell\right)^{1/2}\left(n_g^2\ell\right)^{1/2}\\
&\sim nn_g^2,
\end{aligned}$$

and in a similar vein, though less tightly,

$$\begin{aligned}\left|\sum_i \sum_{\substack{j\neq i\\ g(j)=g(i)}} \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\neq k\\ g(\ell)=g(j)}} C_m p_{ij} p_{k\ell}\right| &\leq \left(\sum_i \sum_{\substack{j\neq i\\ g(j)=g(i)}} \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\neq k\\ g(\ell)=g(j)}} C_m p_{ij}^2\right)^{1/2} \\ &\quad \times\left(\sum_i \sum_{\substack{j\neq i\\ g(j)=g(i)}} \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\neq k\\ g(\ell)=g(j)}} C_m p_{k\ell}^2\right)^{1/2} \\ &\sim nn_g^2.\end{aligned}$$

Hence,

$$\mathbb{V}\left(A_{2n}^{JIVE}|\mathcal{Z}\right) \sim n_g^2.$$

Q3: covariance. For the covariance in the general heteroskedastic case,

$$\mathbb{C}\left(A_{1n}^{JIVE}, A_{2n}^{JIVE}|\mathcal{Z}\right) = \frac{1}{n}\sum_i \sum_{\substack{k\\ g(k)=g(i)}} \sum_{\substack{\ell\neq k\\ g(\ell)=g(i)}} E\left[\varepsilon_i U_k \varepsilon_\ell\right](1-p_{ii})\Upsilon_i p_{k\ell} \neq 0.$$

2.A Appendix: useful lemmas and main results

In our derivations we follow closely the proof techniques and the structure of Chao et al. (2012). We contribute by extending the arguments of Chao et al. (2012) to the multidimensional (clusters) case with additional (inverse cluster-size) reweighting. Let CS, T and M mean Cauchy-Schwartz, triangle and Markov inequalities, respectively.

Lemma A1. If, conditional on $\mathcal{Z} = (\Upsilon, Z)$, $(W_{[g]}, Y_{[g]})(g = 1, ..., G)$ are independent, W_i and Y_i are scalars, $W_{[g]} = (W_i)_{i\in g}$, $Y_{[g]} = (Y_i)_{i\in g}$, and P is a symmetric, idempotent matrix of rank K, then for $\breve{w}_i = \mathbb{E}[W_i/\sqrt{n_{g(i)}}|\mathcal{Z}]$, $\breve{y}_i = \mathbb{E}[Y_i/\sqrt{n_{g(i)}}|\mathcal{Z}]$, $\breve{w} = \mathbb{E}[(W_1/\sqrt{n_{g(1)}}, ..., W_n/\sqrt{n_{g(n)}})'|\mathcal{Z}]$ and $\breve{y} = \mathbb{E}[(Y_1/\sqrt{n_{g(1)}}, ..., Y_n/\sqrt{n_{g(n)}})'|\mathcal{Z}]$, $\bar{\sigma}_{Y_n} = \max_i \sqrt{\mathbb{V}(Y_i|\mathcal{Z})}$, $\bar{\sigma}_{W_n} = \max_i \sqrt{\mathbb{V}(W_i|\mathcal{Z})}$, and $D_G = \bar{\sigma}_{W_n}^2 \bar{\sigma}_{Y_n}^2 K + \breve{y}'\breve{y}\bar{\sigma}_{W_n}^2 + \breve{w}'\breve{w}\bar{\sigma}_{Y_n}^2$, there exists $C > 0$ such that

$$\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)} p_{ij} W_i Y_j/\sqrt{n_{g(i)}\ n_{g(j)}} - \sum_{g(i)\neq g(j)} p_{ij}\breve{w}_i\breve{y}_j\right)^2 |\mathcal{Z}\right] \leq CD_G \quad a.s.$$

Proof. Let $\tilde{W}_i = W_i - \mathbb{E}[W_i|\mathcal{Z}]$ and $\tilde{Y}_i = Y_i - \mathbb{E}[Y_i|\mathcal{Z}]$. Note that

$$\sum_{g(i)\neq g(j)} \frac{p_{ij} W_i Y_j}{\sqrt{n_{g(i)}\ n_{g(j)}}} - \sum_{g(i)\neq g(j)} p_{ij}\breve{w}_i\breve{y}_j = \sum_{g(i)\neq g(j)} \frac{p_{ij}\tilde{W}_i\tilde{Y}_j}{\sqrt{n_{g(i)}\ n_{g(j)}}} + \sum_{g(i)\neq g(j)} \frac{p_{ij}\tilde{W}_i\breve{y}_j}{\sqrt{n_{g(i)}}} + \sum_{g(i)\neq g(j)} \frac{p_{ij}\breve{w}_i\tilde{Y}_j}{\sqrt{n_{g(j)}}}.$$

By CS and $\sum_j p_{ij}^2 = p_{ii}$,

$$
\begin{aligned}
\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)} \frac{p_{ij}\tilde{W}_i\tilde{Y}_j}{\sqrt{n_{g(i)}\ n_{g(j)}}}\right)^2 |\mathcal{Z}\right] =& \sum_{\mathcal{I}} \frac{p_{ij}p_{k\ell}}{\sqrt{n_{g(i)}\ n_{g(j)}\ n_{g(k)}\ n_{g(\ell)}}} \\
&\times \left(\mathbb{E}\left[\tilde{W}_i\tilde{W}_k|\mathcal{Z}\right]\mathbb{E}\left[\tilde{Y}_j\tilde{Y}_\ell|\mathcal{Z}\right] + \mathbb{E}\left[\tilde{W}_i\tilde{Y}_k|\mathcal{Z}\right]\mathbb{E}\left[\tilde{Y}_j\tilde{W}_\ell|\mathcal{Z}\right]\right) \\
&\leq 2\bar{\sigma}_{W_n}^2\bar{\sigma}_{Y_n}^2 \sum_{\mathcal{I}} \frac{\left|p_{ij}p_{k\ell}\right|}{n_{g(i)}\ n_{g(j)}} \\
&\leq 2\bar{\sigma}_{W_n}^2\bar{\sigma}_{Y_n}^2 \sqrt{\sum_{\mathcal{I}} \frac{p_{ij}^2}{n_{g(i)}\ n_{g(j)}}}\sqrt{\sum_{\mathcal{I}} \frac{p_{k\ell}^2}{n_{g(i)}\ n_{g(j)}}} \\
&\leq 2\bar{\sigma}_{W_n}^2\bar{\sigma}_{Y_n}^2 \sqrt{\sum_{i,j} p_{ij}^2}\sqrt{\sum_{k,\ell} p_{k\ell}^2} \\
&= 2\bar{\sigma}_{W_n}^2\bar{\sigma}_{Y_n}^2 \sum_i p_{ii} \\
&\leq 2\bar{\sigma}_{W_n}^2\bar{\sigma}_{Y_n}^2 K,
\end{aligned}
$$

where $\mathcal{I} = \{(i,j,k,\ell)|g(i) = g(k) \neq g(j) = g(\ell)\}$. Also by c_r inequality,

$$
\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)} \frac{p_{ij}\tilde{W}_i\breve{y}_j}{\sqrt{n_{g(i)}}}\right)^2 |\mathcal{Z}\right] \leq C\left(\mathbb{E}\left[\left(\sum_{i,j} \frac{p_{ij}\tilde{W}_i\breve{y}_j}{\sqrt{n_{g(i)}}}\right)^2 |\mathcal{Z}\right] + \mathbb{E}\left[\left(\sum_{g(i)=g(j)} \frac{p_{ij}\tilde{W}_i\breve{y}_j}{\sqrt{n_{g(i)}}}\right)^2 |\mathcal{Z}\right]\right)
$$

Now let $\breve{X} = (\tilde{X}_1/\sqrt{n_{g(1)}}, ..., \tilde{X}_n/\sqrt{n_{g(n)}})'$ and $\breve{X}(g) = (\tilde{X}_i/\sqrt{n_g})_{i\in g}$ for $X \in \{W, Y\}$, $\breve{x}(g) = \left(\mathbb{E}[X_i|\mathcal{Z}]/\sqrt{n(g(i))}\right)_{i\in g}$ for $X \in \{W, Y\}$, $\Sigma_{X_G} = \mathbb{E}\left[\breve{X}\breve{X}'|\mathcal{Z}\right]$ for $X \in \{W, Y\}$, $\Sigma_{X_{G(g)}} = \mathbb{E}\left[\breve{X}_{[g]}\breve{X}'_{[g]}|\mathcal{Z}\right]$ for $X \in \{W, Y\}$. Notice that by independence across g, Σ_{X_G} is block-diagonal with blocks $\{\Sigma_{X_{n(g)}}\}_g$, hence $||\Sigma_{X_G}|| = \max_g ||\Sigma_{X_{G(g)}}||$ for $X \in \{W, Y\}$. Consider two terms above separately:

$$\mathbb{E}\left[\left(\sum_{i,j}\frac{p_{ij}\tilde{W}_i\breve{y}_j}{\sqrt{n_{g(i)}}}\right)^2|\mathcal{Z}\right]=\mathbb{E}\left[\left(\breve{y}'P\breve{W}\right)^2|\mathcal{Z}\right]=\breve{y}'P\mathbb{E}\left[\breve{W}\breve{W}'|\mathcal{Z}\right]P\breve{y}$$

$$\begin{aligned}
&\leq \breve{y}'P\breve{y}\left\|\Sigma_{W_G}\right\|=\breve{y}'\breve{y}\max_g\left\|\Sigma_{W_G(g)}\right\| \\
&\leq \breve{y}'\breve{y}\max_g\left\|\Sigma_{W_G(g)}\right\|_F=\breve{y}'\breve{y}\max_g\sqrt{\sum_{i,j\in g}\Sigma^2_{W_G(g),ij}} \\
&\leq \breve{y}'\breve{y}\max_g\sqrt{n_g^2\left(\max_{i\in g}\mathbb{V}(W_i/\sqrt{n_g}|\mathcal{Z})\right)^2} \\
&\leq \breve{y}'\breve{y}\bar{\sigma}^2_{W_n},
\end{aligned}$$

$$\mathbb{E}\left[\left(\sum_{g(i)=g(j)}\frac{p_{ij}\tilde{W}_i\breve{y}_j}{\sqrt{n_{g(i)}}}\right)^2|\mathcal{Z}\right]=\mathbb{E}\left[\left(\sum_{i,j\in g}\frac{p_{ij}\tilde{W}_i\breve{y}_j}{\sqrt{n_g}}\right)^2|\mathcal{Z}\right]=\mathbb{E}\left[\left(\sum_g\breve{y}(g)'P_{[g,g]}\breve{W}(g)\right)^2|\mathcal{Z}\right]$$

$$\begin{aligned}
&\leq \sum_g\breve{y}(g)'P^2_{[g,g]}\breve{y}(g)\left\|\mathbb{E}\left[\breve{W}(g)\breve{W}(g)'|\mathcal{Z}\right]\right\| \\
&\leq \sum_g\breve{y}(g)'\breve{y}(g)\left\|\Sigma_{W_G(g)}\right\|\cdot\left\|P_{[g,g]}\right\|^2 \\
&\leq \bar{\sigma}^2_{W_n}\sum_g\breve{y}(g)'\breve{y}(g)\leq\breve{y}'\breve{y}\bar{\sigma}^2_{W_n}.
\end{aligned}$$

Hence,

$$\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)}p_{ij}\tilde{W}_i\breve{y}_j/\sqrt{n_{g(i)}}\right)^2|\mathcal{Z}\right]\leq C\breve{y}'\breve{y}\bar{\sigma}^2_{W_n}.$$

Similarly, interchanging roles of Y_i and W_j,

$$\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)}p_{ij}\breve{w}_i\tilde{Y}_j/\sqrt{n_{g(j)}}\right)^2|\mathcal{Z}\right]\leq C\breve{w}'\breve{w}\bar{\sigma}^2_{Y_n}.$$

Then by c_r-inequality

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)}\frac{p_{ij}W_iY_j}{\sqrt{n_{g(i)}\ n_{g(j)}}}-\sum_{g(i)\neq g(j)}p_{ij}\breve{w}_i\breve{y}_j\right)^2\right]&\leq C\left(\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)}\frac{p_{ij}\tilde{W}_i\tilde{Y}_j}{\sqrt{n_{g(i)}\ n_{g(j)}}}\right)^2|\mathcal{Z}\right]\right. \\
&\quad+\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)}\frac{p_{ij}\tilde{W}_i\breve{y}_j}{\sqrt{n_{g(i)}}}\right)^2|\mathcal{Z}\right] \\
&\quad\left.+\mathbb{E}\left[\left(\sum_{g(i)\neq g(j)}\frac{p_{ij}\breve{w}_i\tilde{Y}_j}{\sqrt{n_{g(j)}}}\right)^2|\mathcal{Z}\right]\right) \\
&\leq C\left(\bar{\sigma}^2_{W_n}\bar{\sigma}^2_{Y_n}K+\breve{y}'\breve{y}\bar{\sigma}^2_{W_n}+\breve{w}'\breve{w}\bar{\sigma}^2_{Y_n}\right) \\
&\leq CD_G.
\end{aligned}$$

$\square$

Lemma A2. Suppose that, conditional on $\mathcal{Z}$, the following conditions hold a.s.:

(i) $P = P(\mathcal{Z})$ is a symmetric, idempotent matrix with $\text{rank}(P) = K$ and $p_{ii} \leq C < 1$;

(ii) $(W_{gG}, U_{[g]}, \varepsilon_{[g]})_g$ are independent, and $D_G = \sum_{g=1}^{G} \mathbb{E}[W'_{gG}W_{gG}|\mathcal{Z}]$ satisfies $||D_G|| \leq C$ a.s.n;

(iii) $\mathbb{E}[W_{gG}|\mathcal{Z}] = 0$, $\mathbb{E}[U_{[g]}|\mathcal{Z}] = 0$, $\mathbb{E}[\varepsilon_{[g]}|\mathcal{Z}] = 0$, and there exists a constant C such that $\mathbb{E}[||U_i||^4|\mathcal{Z}] \leq C$, $\mathbb{E}[||\varepsilon_i||^4|\mathcal{Z}] \leq C$

(iv) $\sum_{g=1}^{G} \mathbb{E}\left[||W_{gG}||^4|\mathcal{Z}\right] \to 0$;

(v) $K \to \infty$ as $G \to \infty$ and $\frac{\max_g n_g}{K} \to 0$.

Then for

$$\bar{\Sigma}_G \equiv \sum_{\substack{g,g' \\ g' \neq g}} \sum_{\substack{i,k \in g \\ j,\ell \in g'}} p_{ij}p_{k\ell} \left(\mathbb{E}[U_iU'_k|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}] + \mathbb{E}[U_i\varepsilon_k|\mathcal{Z}]\mathbb{E}[\varepsilon_jU'_\ell|\mathcal{Z}]\right) / \left(Kn_gn_{g'}\right)$$

and any sequences c_{1G} and c_{2G} depending on $\mathcal{Z}$ of conformable vectors with $||c_{1G}|| \leq C$, $||c_{2G}|| \leq C$, and $\Xi_G = c'_{1G}D_Gc_{1G} + c'_{2G}\bar{\Sigma}_Gc_{2G} > 1/C$ a.s.n, it follows that

$$Y_G = \Xi_G^{-1/2} \left(c'_{1G} \sum_g W_{gG} + c'_{2G} \sum_{\substack{g,g' \\ g' \neq g}} \sum_{\substack{i \in g \\ j \in g'}} U_ip_{ij}\varepsilon_j / \sqrt{Kn_gn_{g'}} \right) \xrightarrow{d} N(0,1), \qquad a.s.;$$

i.e., $\Pr(Y_G \leq y|\mathcal{Z}) \xrightarrow{a.s.} \Phi(y)$ for all y.

Proof. Let $b_{1G} = c_{1G}\Xi_G^{-1/2}$ and $b_{2G} = c_{2G}\Xi_G^{-1/2}$. $||b_{1G}|| \leq ||c_{1G}||\,||\Xi_G^{-1/2}|| \leq C$ by hypothesis. Similarly, $||b_{2G}|| \leq C$. Let $w_{gG} = b'_{1G}W_{gG}$ and $u_i = b'_{2G}U_i$. Then, $Y_G = w_{1G} + \sum_{g=2}^{G} y_{gG}$, $y_{gG} = w_{gG} + \bar{y}_{gG}$, $\bar{y}_{gG} = \sum_{g'<g}\sum_{i\in g}\sum_{j\in g'}(u_ip_{ij}\varepsilon_j + u_jp_{ji}\varepsilon_i)/\sqrt{Kn_gn_{g'}}$.

Also, $\mathbb{E}[||w_{1G}||^4|\mathcal{Z}] \leq \sum_g \mathbb{E}[||w_{gG}||^4|\mathcal{Z}] \leq C\mathbb{E}[||W_{gG}||^4|\mathcal{Z}] \to 0$ a.s., so by a conditional version of M, we deduce that for any $v > 0$, $P(|w_{1G}| \geq v|\mathcal{Z}) \to 0$. Moreover, note that $\sup_n \mathbb{E}[|P(|w_{1G}| \geq v|\mathcal{Z})|^2] < \infty$. It follows that, by a version of the dominated convergence theorem, as given by the Corollary to Theorem 25.12 of Billingsley (1986), it follows that $P(|w_{1G}| \geq v) = \mathbb{E}[P(|w_{1G}| \geq v|\mathcal{Z})] \to 0$; i.e. $w_{1G} \xrightarrow{p} 0$ unconditionally. Hence, $Y_G = \sum_{i=2}^{G} y_{gG} + o_p(1)$.

Let $\mathcal{X}_g = (W'_{gG}, \{U'_i, \varepsilon_i\}_{i\in g})'$ for $g = 1, ..., G$. Define the σ-fields $F_{g,G} = \sigma(\mathcal{X}_1, ..., \mathcal{X}_g)$ for $g = 1, ..., G$. By construction, $F_{g-1,G} \subseteq F_{g,G}$. Conditional on $\mathcal{Z}$, $\{y_{gG}, \mathcal{F}_{g,G}, 1 \leq g \leq G, G \geq 2\}$ is a martingale difference array.

Now we need to verify that $\{y_{gG}, \mathcal{F}_{g,G}, 1 \leq g \leq G, G \geq 2\}$ is square integrable, conditional on $\mathcal{Z}$, i.e. $\mathbb{E}[y_{gG}^2|\mathcal{Z}] < \infty$. First note that $\mathbb{E}[w_{gG}\bar{y}_{gG}|\mathcal{Z}] = 0$ a.s. Then $\mathbb{E}[y_{gG}^2|\mathcal{Z}] = \mathbb{E}[w_{gG}^2|\mathcal{Z}] + \mathbb{E}[\bar{y}_{gG}^2|\mathcal{Z}]$. Using Lemmas B1 and B2,

$$
\begin{aligned}
\mathbb{E}[\bar{y}_{gG}^2|\mathcal{Z}] &\leq \frac{1}{Kn_g}\mathbb{E}\left[\sum_{g'<g}\frac{(u'_{[g]}P_{[g,g']}\varepsilon_{[g']})^2 + 2|u_{[g']}P_{[g,g']}\varepsilon_{[g]}|\cdot|u_{[g]}P_{[g,g']}\varepsilon_{[g']}| + (u'_{[g']}P_{[g,g']}\varepsilon_{[g]})^2}{n_{g'}}\right] \\
&\leq \frac{C}{K}\sum_{g'<g}\lambda_{\max}(P_{[g,g']}P'_{[g,g']}) \\
&\leq \frac{C}{K}\sum_{g'=1}^{G}\operatorname{tr}(P_{[g,g']}P'_{[g,g']}) \\
&= \frac{C}{K}\sum_{i\in g}p_{ii} \\
&\leq \frac{C}{K}n_{\max}, \\
\mathbb{E}[w_{gG}^2|\mathcal{Z}] &\leq \sqrt{\mathbb{E}[w_{gG}^4|\mathcal{Z}]} \to 0 \text{ a.s.}
\end{aligned}
$$

Hence, $\mathbb{E}[y_{gG}^2|Z] < \infty$.

Conditional variance of y_{gG} is

$$
\begin{aligned}
s_G^2(\mathcal{Z}) &= \mathbb{E}\left[\left(\sum_{g=2}^{G}y_{gG}\right)^2|\mathcal{Z}\right] = \sum_{g=2}^{G}\mathbb{E}\left[w_{gG}^2\right] + \sum_{g=2}^{G}\mathbb{E}\left[\bar{y}_{gG}^2|\mathcal{Z}\right] \\
&= b'_{1G}D_Gb_{1G} - \mathbb{E}[w_{1G}] + b'_{2G}\bar{\Sigma}_Gb_{2G} + o_{a.s.}(1) \\
&= \Xi_G^{-1/2}\left(c'_{1G}D_Gc_{1G} + c'_{2G}\bar{\Sigma}_Gc_{2G}\right)\Xi_G^{-1/2} + o_{a.s.}(1) \\
&= \Xi_G^{-1/2}\Xi_G\Xi_G^{-1/2} + o_{a.s.}(1) = 1 + o_{a.s.}(1) \xrightarrow{a.s.} 1,
\end{aligned}
$$

Thus, $s_G^2(\mathcal{Z})$ is bounded and bounded away from 0 a.s.

Next, we show that Lindeberg's condition holds. By c_r inequality, $\sum_{g=2}^{G}\mathbb{E}[y_{gG}^4|\mathcal{Z}] \leq C\sum_{g=2}^{G}\mathbb{E}\left[w_{gG}^4|\mathcal{Z}\right] + C\sum_{g=2}^{G}\mathbb{E}\left[\bar{y}_{gG}^4|\mathcal{Z}\right]$. The first term is

$$
\sum_{g=2}^{G}\mathbb{E}\left[w_{gG}^4|\mathcal{Z}\right] = \sum_{g=2}^{G}\mathbb{E}\left[||b'_{1G}W_{gG}||^4|\mathcal{Z}\right] \leq C\sum_{g=2}^{G}\mathbb{E}\left[||W_{gG}||^4|\mathcal{Z}\right] \to 0.
$$

The second term is

$$\sum_{g=2}^{G} \mathbb{E}\left[\bar{y}_{gG}^4|\mathcal{Z}\right] = \mathbb{E}\left[\sum_{g=2}^{G}\sum_{g_1<g}\left(u'_{[g]}P_{[g,g_1]}\varepsilon_{[g_1]}\right)^4/(n_g^2 n_{g_1}^2 K^2)\right] + \mathbb{E}\left[\sum_{g=2}^{G}\sum_{g_1<g}\left(\varepsilon'_{[g]}P_{[g,g_1]}u_{[g_1]}\right)^4/(n_g^2 n_{g_1}^2 K^2)\right]$$
$$+ 3\mathbb{E}\left[\sum_{g=2}^{G}\sum_{\substack{g_1,g_2<g\\ g_1\neq g_2}}\left(u'_{[g]}P_{[g,g_1]}\varepsilon_{[g_1]}\right)^2\left(u'_{[g]}P_{[g,g_2]}\varepsilon_{[g_2]}\right)^2/(n_g^2 n_{g_1} n_{g_2} K^2)\right]$$
$$+ 3\mathbb{E}\left[\sum_{g=2}^{G}\sum_{\substack{g_1,g_2<g\\ g_1\neq g_2}}\left(\varepsilon'_{[g]}P_{[g,g_1]}u_{[g_1]}\right)^2\left(\varepsilon'_{[g]}P_{[g,g_2]}u_{[g_2]}\right)^2/(n_g^2 n_{g_1} n_{g_2} K^2)\right].$$

By Lemma B2, $||P_{[g,h]}|| \leq 1$ a.s. Using Lemmas B1, B2, and c_r inequality, the first summand of the second term is

$$\mathbb{E}\left[\sum_{g=2}^{G}\sum_{g_1<g}\left(u'_{[g]}P_{[g,g_1]}\varepsilon_{[g_1]}\right)^4/(n_g^2 n_{g_1}^2 K^2)\right] \leq \mathbb{E}\left[\sum_{g=2}^{G}\sum_{g_1<g}\left(u'_{[g]}u_{[g]}\right)^2\left(\varepsilon'_{[g_1]}\varepsilon_{[g_1]}\right)^2||P_{[g,g_1]}P_{[g_1,g]}||^2/(n_g^2 n_{g_1}^2 K$$
$$\leq C\sum_{g=2}^{G}\sum_{g_1<g}\mathbb{E}\left[\left(\sum_{i\in g}u_i^2\right)^2/n_g^2\right]\mathbb{E}\left[\left(\sum_{j\in g_1}\varepsilon_j^2\right)^2/n_{g_1}^2\right]$$
$$\times\lambda_{\max}\left(P_{[g,g_1]}P_{[g_1,g]}\right)/K^2$$
$$\leq C\sum_{g=2}^{G}\sum_{g_1}\mathbb{E}\left[\sum_{i\in g}u_i^4/n_g\right]\mathbb{E}\left[\sum_{j\in g_1}\varepsilon_j^4/n_{g_1}\right]$$
$$\times\lambda_{\max}\left(P_{[g,g_1]}P_{[g_1,g]}\right)/K^2$$
$$\leq C\sum_{g=2}^{G}\mathrm{tr}\left(\sum_{g_1}P_{[g,g_1]}P_{[g_1,g]}\right) = C\sum_{g=2}^{G}\mathrm{tr}\left(P_{gg}\right) \leq C/K.$$

Similarly, exchanging the roles of ε and u, the second summand is

$$\mathbb{E}\left[\sum_{g=2}^{G}\sum_{g_1<g}\left(\varepsilon'_{[g]}P_{[g,g_1]}u_{[g_1]}\right)^4/(n_g^2 n_{g_1}^2 K^2)\right] \leq C/K.$$

The third summand is

$$
\begin{aligned}
&3\mathbb{E}\left[\sum_{g=2}^{G}\sum_{\substack{g_1,g_2<g\\ g_1\neq g_2}}\left(u_{[g]}'P_{[g,g_1]}\varepsilon_{[g_1]}\right)^2\left(u_{[g]}'P_{[g,g_2]}\varepsilon_{[g_2]}\right)^2/(n_g^2 n_{g_1}n_{g_2}K^2)\right]\\
&\quad\leq 3\mathbb{E}\left[\sum_{g=2}^{G}\sum_{\substack{g_1,g_2<g\\ g_1\neq g_2}}\left(u_{[g]}'u_{[g]}\right)^2\left(\varepsilon_{[g_1]}'\varepsilon_{[g_1]}\right)\left(\varepsilon_{[g_2]}'\varepsilon_{[g_2]}\right)||P_{[g,g_1]}||^2||P_{[g,g_2]}||^2/(n_g^2 n_{g_1}n_{g_2}K^2)\right]\\
&\quad\leq C\sum_{g=2}^{G}\sum_{g_1<g}||P_{[g,g_1]}||^2\sum_{g_2<g}||P_{[g,g_2]}||^2/K^2\\
&\quad\leq C\sum_{g=2}^{G}\left(\sum_{g_1<g}||P_{[g,g_1]}||_F^2\right)^2/K^2\\
&\quad= C\sum_{g=2}^{G}\left(\sum_{g_1}\sum_{\substack{i\in g\\ j\in g_1}}p_{ij}^2\right)^2/K^2\\
&\quad\leq C\sum_{g=2}^{G}\left(\sum_{i\in g}p_{ii}\right)^2/K^2\\
&\quad\leq Cn_{\max}\sum_{g=2}^{G}\sum_{i\in g}p_{ii}^2\leq Cn_{\max}/K.
\end{aligned}
$$

Similarly, exchanging the roles of ε and u, the fourth summand is

$$
3\mathbb{E}\left[\sum_{g=2}^{G}\sum_{\substack{g_1,g_2<g\\ g_1\neq g_2}}\left(\varepsilon_{[g]}'P_{[g,g_1]}u_{[g_1]}\right)^2\left(\varepsilon_{[g]}'P_{[g,g_2]}u_{[g_2]}\right)^2/(n_g^2 n_{g_1}n_{g_2}K^2)\right]\leq Cn_{\max}/K.
$$

Then, a.s. $\sum_{g=2}^{G}\mathbb{E}\left[\bar{y}_{gG}^4|\mathcal{Z}\right]\to 0$, and $\sum_{g=2}^{G}\mathbb{E}[y_{gG}^4|\mathcal{Z}]\to 0$.

To apply the martingale central limit theorem (Lemma B3), it suffices to show that for any $v>0$

$$
P\left(\left|\sum_{g=2}^{G}\mathbb{E}\left[y_{gG}^2|\mathcal{X}_1,...,\mathcal{X}_{g-1},\mathcal{Z}\right]-s_G^2(\mathcal{Z})\right|\geq v|\mathcal{Z}\right)\to 0.
$$

As noted above, $\mathbb{E}[w_{gG}\bar{y}_{gG}|\mathcal{Z}]=0$ a.s., thus we can write

$$
\begin{aligned}
\sum_{g=2}^{G}\mathbb{E}\left[y_{gG}^2|\mathcal{X}_1,...,\mathcal{X}_{g-1},\mathcal{Z}\right]-s_G^2(\mathcal{Z}) &= \sum_{g=2}^{G}\left(\mathbb{E}\left[w_{gG}^2|\mathcal{X}_1,...,\mathcal{X}_{g-1},\mathcal{Z}\right]-\mathbb{E}\left[w_{gG}^2|\mathcal{Z}\right]\right)\\
&\quad+2\sum_{g=2}^{G}\mathbb{E}\left[w_{gG}\bar{y}_{gG}|\mathcal{X}_1,...,\mathcal{X}_{g-1},\mathcal{Z}\right]+\sum_{g=2}^{G}\left(\mathbb{E}\left[\bar{y}_{gG}^2|\mathcal{X}_1,...,\mathcal{X}_{g-1},\mathcal{Z}\right]-\mathbb{E}\left[\bar{y}_{gG}^2|\mathcal{Z}\right]\right).
\end{aligned}
$$

By independence of $(W_{gG})_g$ conditional on $\mathcal{Z}$, $\mathbb{E}\left[w_{gG}^2|\mathcal{X}_1,...,\mathcal{X}_{g-1},\mathcal{Z}\right] = \mathbb{E}\left[w_{gG}^2|\mathcal{Z}\right]$ a.s., so the first term is 0. Next note that $\mathbb{E}\left[w_{gG}\bar{y}_{gG}|\mathcal{X}_1,...,\mathcal{X}_{g-1},\mathcal{Z}\right] = \sum_{g'<g}\sum_{\substack{i\in g\\ j\in g'}}\varepsilon_j p_{ij}\mathbb{E}\left[w_{gG}u_i|\mathcal{Z}\right]/\sqrt{Kn_g n_{g'}}+$ $\sum_{g'<g}\sum_{\substack{i\in g\\ j\in g'}}u_j p_{ij}\mathbb{E}\left[w_{gG}\varepsilon_i|\mathcal{Z}\right]/\sqrt{Kn_g n_{g'}}$.

Let $\breve{\delta}_i = \breve{\delta}_i(\mathcal{Z}) = \mathbb{E}\left[w_{g(i)G}u_i|\mathcal{Z}\right]/\sqrt{n_{g(i)}}$ and consider the first term, $\sum_{g'<g}\sum_{\substack{i\in g\\ j\in g'}}\varepsilon_j p_{ij}\breve{\delta}_i/\sqrt{Kn_{g'}}$. Let $\bar{P}$ be a matrix with (i,j)th entry $\bar{p}_{ij} = p_{ij}$ if $g(i) < g(j)$ and 0 otherwise. Let for $i \in g$, $\breve{\varepsilon}_i = \varepsilon_i/\sqrt{g(i)}$. Then $\sum_g\sum_{g'<g}\sum_{\substack{i\in g\\ j\in g'}}\varepsilon_j p_{ij}\breve{\delta}_i/\sqrt{Kn_{g'}} = \breve{\delta}'\bar{P}'\breve{\varepsilon}/\sqrt{K}$. By CS, $\breve{\delta}'\breve{\delta} = \sum_g\sum_{i\in g}\left(\mathbb{E}\left[w_{gG}u_i|\mathcal{Z}\right]\right)^2/n_g \le C\sum_g\mathbb{E}\left[w_{gG}^2|\mathcal{Z}\right] \le C$ a.s.

Further,

$$\begin{aligned}
||\bar{P}'\bar{P}||_F^2 &= \mathrm{tr}(\bar{P}'\bar{P}\bar{P}'\bar{P})\\
&= \sum_{i,j,k,\ell} p_{ij}\mathbb{1}\{i>j, g(i)\neq g(j)\}p_{jk}\mathbb{1}\{j<k, g(j)\neq g(k)\}\\
&\quad\times p_{k\ell}\mathbb{1}\{k>\ell, g(k)\neq g(\ell)\}p_{\ell i}\mathbb{1}\{\ell<i, g(\ell)\neq g(i)\}\\
&= \sum_{\substack{i>j\\ g(i)\neq g(j)}} p_{ij}p_{ji}p_{ij}p_{ji} + \sum_{\substack{i>k>j\\ g(i)\neq g(j), g(k)\neq g(j)}} p_{ij}p_{jk}p_{kj}p_{ji}\\
&\quad+\sum_{\substack{k>i>j\\ g(i)\neq g(j), g(k)\neq g(j)}} p_{ij}p_{jk}p_{kj}p_{ji} + \sum_{\substack{i>j>\ell\\ g(i)\neq g(j), g(i)\neq g(\ell)}} p_{ij}p_{ji}p_{i\ell}p_{\ell i} + \sum_{\substack{i>\ell>j\\ g(i)\neq g(j), g(i)\neq g(\ell)}} p_{ij}p_{ji}p_{i\ell}p_{\ell i}\\
&\quad+\sum_{\substack{i>k>j>\ell\\ g(i)\neq g(j), g(k)\neq g(\ell)}} p_{ij}p_{jk}p_{k\ell}p_{\ell i} + \sum_{\substack{k>i>j>\ell\\ g(i)\neq g(j), g(k)\neq g(\ell)}} p_{ij}p_{jk}p_{k\ell}p_{\ell i}\\
&\quad+\sum_{\substack{i>k>\ell>j\\ g(i)\neq g(j), g(k)\neq g(\ell)}} p_{ij}p_{jk}p_{k\ell}p_{\ell i} + \sum_{\substack{k>i>\ell>j\\ g(i)\neq g(j), g(k)\neq g(\ell)}} p_{ij}p_{jk}p_{k\ell}p_{\ell i}\\
&= \sum_{\substack{i>j\\ g(i)\neq g(j)}} p_{ij}^4 + 2\sum_{\substack{i>k>j\\ g(i)\neq g(j), g(k)\neq g(j)}} p_{ij}^2p_{jk}^2 + 2\sum_{\substack{i>j>\ell\\ g(i)\neq g(j), g(i)\neq g(\ell)}} p_{ij}^2p_{i\ell}^2\\
&\quad+4\sum_{\substack{i>k>j>\ell\\ g(i)\neq g(j), g(k)\neq g(\ell)}} p_{ij}p_{jk}p_{k\ell}p_{\ell i}\\
&\le \sum_{i>j} p_{ij}^4 + 2\sum_{i>k>j}(p_{ij}^2p_{jk}^2 + p_{ij}^2p_{i\ell}^2) + 4\sum_{\substack{i>k>j>\ell\\ g(i)\neq g(j), g(k)\neq g(\ell)}} p_{ij}p_{jk}p_{k\ell}p_{\ell i}\\
&= \sum_{i>j} p_{ij}^4 + 2\sum_{i>k>j}(p_{ij}^2p_{jk}^2 + p_{ij}^2p_{i\ell}^2) + 4\sum_{i>k>j>\ell} p_{ij}p_{jk}p_{k\ell}p_{\ell i} - 4\sum_{\substack{i>k>j>\ell\\ g(i)=g(j), g(k)=g(\ell)}} p_{ij}p_{jk}p_{k\ell}p_{\ell i}\\
&\le CK + 4|\sum_g\sum_{\substack{i,k,j,\ell\in g\\ i>k>j>\ell}} p_{ij}p_{jk}p_{k\ell}p_{\ell i}|.
\end{aligned}$$

Also

$$\sum_g\sum_{\substack{i,k,j,\ell\in g\\ i>k>j>\ell}} p_{ij}p_{jk}p_{k\ell}p_{\ell i} \le \mathrm{tr}(P_{[g,g]}^4).$$

Then

$$\mathbb{E}\left[\left(\breve{\delta}'\bar{P}'\breve{\varepsilon}/\sqrt{K}\right)^2|\mathcal{Z}\right] = \breve{\delta}'\bar{P}'\mathbb{E}\left[\breve{\varepsilon}\breve{\varepsilon}'|\mathcal{Z}\right]\bar{P}\breve{\delta}/K \leq C\breve{\delta}'\bar{P}'\bar{P}\breve{\delta}/K$$
$$\leq \breve{\delta}'\breve{\delta}||\bar{P}'\bar{P}||/K \leq C/\sqrt{K} \to 0 \quad \text{a.s.}$$

By M we have for any $v > 0$, $P\left(\left|\breve{\delta}'\bar{P}\breve{\varepsilon}/\sqrt{K}\right| \geq v|\mathcal{Z}\right) \to 0$ a.s. Exchanging the roles of u and ε, we have $\sum_{\substack{g \\ g'<g}}\sum_{\substack{i\in g \\ j\in g'}} u_j p_{ij}\mathbb{E}\left[w_{gG}\varepsilon_i|\mathcal{Z}\right]/\sqrt{Kn_g n_{g'}} \to 0$ a.s. Therefore, it follows by T that, for any $v > 0$, $P\left(\left|\sum_{g=2}^{G}\mathbb{E}\left[w_{gG}\bar{y}_{gG}|\mathcal{X}_1, ..., \mathcal{X}_{g-1}, \mathcal{Z}\right]\right| \geq v|\mathcal{Z}\right) \to 0$ a.s.

It remains only to show that, for any $v > 0$,

$$P\left(\left|\sum_{g=2}^{G}\left(\mathbb{E}\left[\bar{y}_{gG}^2|\mathcal{X}_1, ..., \mathcal{X}_{g-1}, \mathcal{Z}\right] - \mathbb{E}\left[\bar{y}_{gG}^2|\mathcal{Z}\right]\right)\right| \geq v|\mathcal{Z}\right) \to 0 \quad \text{a.s.}$$

Now write

$$
\begin{aligned}
&\sum_{g=2}^{G}\left(\mathbb{E}\left[\bar{y}_{gG}^2|\mathcal{X}_1,...,\mathcal{X}_{g-1},\mathcal{Z}\right]-\mathbb{E}\left[\bar{y}_{gG}^2|\mathcal{Z}\right]\right)\\
&=\frac{1}{K}\left(\sum_{g=2}^{G}\sum_{\substack{g'<g\\h<g}}\sum_{\substack{i,k\in g\\j\in g'\\\ell\in h}}\frac{p_{ij}p_{k\ell}}{n_g\sqrt{n_{g'}n_h}}\left(\mathbb{E}[u_iu_k|\mathcal{Z}]\varepsilon_j\varepsilon_\ell+\mathbb{E}[\varepsilon_iu_k|\mathcal{Z}]u_j\varepsilon_\ell+\mathbb{E}[u_i\varepsilon_k|\mathcal{Z}]\varepsilon_ju_\ell+\mathbb{E}[\varepsilon_i\varepsilon_k|\mathcal{Z}]u_ju_\ell\right)\right.\\
&\quad\left.-\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_gn_{g'}}\left(\mathbb{E}[u_iu_k|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]+2\mathbb{E}[\varepsilon_iu_k|\mathcal{Z}]\mathbb{E}[u_j\varepsilon_\ell|\mathcal{Z}]+\mathbb{E}[\varepsilon_i\varepsilon_k|\mathcal{Z}]\mathbb{E}[u_ju_\ell|\mathcal{Z}]\right)\right)\\
&=\frac{1}{K}\left(\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_gn_{g'}}\left(\mathbb{E}[u_iu_k|\mathcal{Z}]\varepsilon_j\varepsilon_\ell+\mathbb{E}[\varepsilon_iu_k|\mathcal{Z}]u_j\varepsilon_\ell+\mathbb{E}[u_i\varepsilon_k|\mathcal{Z}]\varepsilon_ju_\ell+\mathbb{E}[\varepsilon_i\varepsilon_k|\mathcal{Z}]u_ju_\ell\right)\right.\\
&\quad+2\sum_{g=3}^{G}\sum_{h<g'<g}\sum_{\substack{i,k\in g\\j\in g'\\\ell\in h}}\frac{p_{ij}p_{k\ell}}{n_g\sqrt{n_{g'}n_h}}\left(\mathbb{E}[u_iu_k|\mathcal{Z}]\varepsilon_j\varepsilon_\ell+\mathbb{E}[\varepsilon_iu_k|\mathcal{Z}]u_j\varepsilon_\ell+\mathbb{E}[u_i\varepsilon_k|\mathcal{Z}]\varepsilon_ju_\ell+\mathbb{E}[\varepsilon_i\varepsilon_k|\mathcal{Z}]u_ju_\ell\right)\\
&\quad\left.-\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_gn_{g'}}\left(\mathbb{E}[u_iu_k|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]+2\mathbb{E}[\varepsilon_iu_k|\mathcal{Z}]\mathbb{E}[u_j\varepsilon_\ell|\mathcal{Z}]+\mathbb{E}[\varepsilon_i\varepsilon_k|\mathcal{Z}]\mathbb{E}[u_ju_\ell|\mathcal{Z}]\right)\right)\\
&=\frac{1}{K}\left(2\sum_{g=3}^{G}\sum_{h<g'<g}\sum_{\substack{i,k\in g\\j\in g'\\\ell\in h}}\frac{p_{ij}p_{k\ell}}{n_g\sqrt{n_{g'}n_h}}\left(\mathbb{E}[u_iu_k|\mathcal{Z}]\varepsilon_j\varepsilon_\ell+\mathbb{E}[\varepsilon_iu_k|\mathcal{Z}]u_j\varepsilon_\ell+\mathbb{E}[u_i\varepsilon_k|\mathcal{Z}]\varepsilon_ju_\ell+\mathbb{E}[\varepsilon_i\varepsilon_k|\mathcal{Z}]u_ju_\ell\right)\right.\\
&\quad+\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_gn_{g'}}\mathbb{E}[u_iu_k|\mathcal{Z}]\left(\varepsilon_j\varepsilon_\ell-\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)\\
&\quad+\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_gn_{g'}}\mathbb{E}[\varepsilon_iu_k|\mathcal{Z}]\left(u_j\varepsilon_\ell-\mathbb{E}[u_j\varepsilon_\ell|\mathcal{Z}]\right)\\
&\quad+\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_gn_{g'}}\mathbb{E}[u_i\varepsilon_k|\mathcal{Z}]\left(\varepsilon_ju_\ell-\mathbb{E}[\varepsilon_ju_\ell|\mathcal{Z}]\right)\\
&\quad\left.+\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_gn_{g'}}\mathbb{E}[\varepsilon_i\varepsilon_k|\mathcal{Z}]\left(u_ju_\ell-\mathbb{E}[u_ju_\ell|\mathcal{Z}]\right)\right).
\end{aligned}
$$

By applying Lemma B4, we show that each term converges to 0 in probability.

By the martingale central limit theorem (Lemma B3), as $G \to \infty$, $P(Y_G \leq y|\mathcal{Z}) \to \Phi(y)$ a.s., for every real number y, where $\Phi(y)$ denotes denotes the c.d.f. of a standard normal distribution. Moreover, notice that $\sup_G \mathbb{E}[|P(Y_G \leq y|\mathcal{Z})|^2] < \infty$. Hence, by the Corollary to Theorem 25.12 of Billingsley (1986), it follows that $P(Y_G \leq y) = \mathbb{E}[P(Y_G \leq y|\mathcal{Z})] \to \mathbb{E}[\Phi(y)] = \Phi(y)$ unconditionally. $\square$

Lemma A4. Suppose that there is $C > 0$ such that, conditional on $\mathcal{Z}$, $(W_{[g]}, Y_{[g]})(g = 1, ..., G)$ are independent, W_i and Y_i are scalars with $|\mathbb{E}[W_i|\mathcal{Z}]| \leq C/\sqrt{n}, |\mathbb{E}[Y_i|\mathcal{Z}]| \leq C/\sqrt{n}, \mathbb{E}[||\eta_{[g]}||^2|\mathcal{Z}] \leq Cn_g^2, \mathbb{V}[W_i|\mathcal{Z}] \leq C/r_G$, and $\mathbb{V}[Y_i|\mathcal{Z}] \leq C/r_G$ and $\sqrt{K}/r_G \to 0$. Then

$$A_G = \mathbb{E}\left[\sum_{h \neq g \neq m} W_{[h]} P_{[h,g]} \eta_{[g]} P_{[g,m]} Y_{[m]} / \sqrt{n_h n_m n_g^2} |\mathcal{Z}\right] = O_p(1).$$

Proof. Let $\bar{w}_{[g]} = \mathbb{E}[W_{[g]}|\mathcal{Z}]/\sqrt{n_g}$, $\tilde{W}_{[g]} = W_{[g]}/\sqrt{n_g} - \bar{w}_{[g]}$, $\bar{y}_{[g]} = \mathbb{E}[Y_{[g]}|\mathcal{Z}]/\sqrt{n_g}$, $\tilde{Y}_{[g]} = Y_{[g]}/\sqrt{n_g} - \bar{y}_{[g]}$, $\bar{\eta}_{[g]} = \mathbb{E}[\eta_{[g]}|\mathcal{Z}]$, $\tilde{\eta}_{[g]} = \eta_{[g]} - \bar{\eta}_{[g]}$. Further, a.s. $\bar{\mu}_W^2 = \max_i \bar{w}_i^2 \leq C/n$, $\bar{\mu}_Y^2 = \max_i \bar{y}_i^2 \leq C/n$, $\bar{\mu}_\eta^2 = \max_i \bar{\eta}_i^2 \leq C$, $\bar{\sigma}_W^2 = \max_{i \leq n} \mathbb{V}[W_i|\mathcal{Z}] \leq C/r_G$, $\bar{\sigma}_Y^2 = \max_{i \leq n} \mathbb{V}[Y_i|\mathcal{Z}] \leq C/r_G$, and $\bar{\sigma}_\eta^2 = \max_g \mathbb{E}[||\tilde{\eta}_{[g]}||^2|\mathcal{Z}]/n_g^2 \leq C$. Also, let $\breve{y}_i = \sum_j P_{ij}\bar{y}_j/\sqrt{n_{g(j)}}$, $\breve{w}_i = \sum_j P_{ij}\bar{w}_j/\sqrt{n_{g(j)}}$. Note that

$$\sum_i \breve{y}_i^2 = \breve{y}'\breve{y} = \bar{y}'P\bar{y} \leq \bar{y}'\bar{y} = \sum_i \bar{y}_i^2 \leq C.$$

And similarly, $\sum_i \breve{w}_i^2 \leq \sum_i \bar{w}_i^2 \leq C$. By the argument similar to the beginning of the proof of Theorem 4,

$$\begin{aligned}
A_G = &\sum_{g \neq h} \sum_{k \notin \{g,h\}} \bar{w}'_{[g]} P_{[g,k]} \bar{\eta}_{[k]}/n_k P_{[k,h]} \bar{y}_{[h]} \\
= &\sum_g \left(\breve{w}'_{[g]}\bar{\eta}_{[g]}/n_g \breve{y}_{[g]} - \bar{w}'_{[g]} P_{[g,g]}\bar{\eta}_{[g]}/n_g \breve{y}_{[g]} - \breve{w}'_{[g]}\bar{\eta}_{[g]}/n_g P_{[g,g]}\bar{y}_{[g]} + 2\bar{w}'_{[g]}P_{[g,g]}\bar{\eta}_{[g]}/n_g P_{[g,g]}\bar{y}_{[g]}\right) \\
&- \sum_{g,h} \bar{w}'_{[g]} P_{[g,h]} \bar{\eta}_{[h]}/n_h P_{[h,g]} \bar{y}_{[g]}.
\end{aligned}$$

Notice that $||\bar{\eta}_{[g]}||^2 \leq \mathbb{E}[||\eta_{[g]}||^2|\mathcal{Z}] \leq Cn_g^2$. By T, CS, and $||\bar{\eta}_{[g]}|| \leq Cn_g$,

$$|\sum_g \breve{w}_{[g]}\bar{\eta}_{[g]}/n_g \breve{y}_{[g]}| \leq \sum_g ||\bar{\eta}_{[g]}||/n_g \sqrt{\breve{w}'_{[g]}\breve{w}_{[g]}} \sqrt{\breve{y}'_{[g]}\breve{y}_{[g]}} \leq C,$$

$$|\sum_g \bar{w}_{[g]} P_{[g,g]}\bar{\eta}_{[g]}\breve{y}_{[g]}|/n_g \leq \sum_g ||\bar{\eta}_{[g]}||/n_g \sqrt{\bar{w}'_{[g]}P_{[g,g]}^2\bar{w}_{[g]}} \sqrt{\breve{y}'_{[g]}\breve{y}_{[g]}} \leq C,$$

and

$$|\sum_g \breve{w}_{[g]}\bar{\eta}_{[g]} P_{[g,g]}\bar{y}_{[g]}|/n_g \leq \sum_g ||\bar{\eta}_{[g]}||/n_g \sqrt{\breve{w}'_{[g]}\breve{w}_{[g]}} \sqrt{\bar{y}'_{[g]}P_{[g,g]}^2\bar{y}_{[g]}} \leq C.$$

By Lemma B1, $|\sum_{g,h} \bar{w}'_{[g]} P_{[g,h]} \bar{\eta}_{[h]}/n_h P_{[h,g]}\bar{y}_{[g]}| \leq C/n \sum_{g,h} ||P_{[g,h]}||^2 \leq CK/n \leq C$, and $|\sum_g \bar{w}'_{[g]} P_{[g,g]}\bar{\eta}_{[g]}/n_g P_{[g,g]}\bar{y}_{[g]}| \leq CG/n \leq C$. Thus, $|A_G| \leq C$ holds by T, and the conclusion follows by M. $\square$

Lemma A5. If Assumptions 1-3 are satisfied, then

$$(i)\ S_G^{-1}\hat{H}S_G^{-1} = \sum_{\substack{i,j\\ g(i)\neq g(j)}} p_{ij}z_iz_j'/(n\sqrt{n_{g(i)}\ n_{g(j)}}) + o_p(1),$$

$$(ii)\ S_G^{-1}\sum_{\substack{i,j\\ g(i)\neq g(j)}} p_{ij}X_i\varepsilon_j/\sqrt{n_{g(i)}\ n_{g(j)}} = O_p\left(1+\sqrt{K/r_G}\right).$$

Proof. Let e_k be kth unit vector and let $Y_i = e_k'S_G^{-1}X_i = e_k'z_i/\sqrt{n} + e_k'S_G^{-1}U_i$ and $W_i = e_\ell'S_G^{-1}X_i$ for some (k,ℓ). By Assumption 2, $\lambda_{\min}(S_G) \geq C\sqrt{r_G}$, then $||S_G^{-1}|| \leq C/\sqrt{r_G}$. It follows that a.s.

$$\mathbb{E}[Y_i|\mathcal{Z}] = e_k'z_i/\sqrt{n}, \quad \mathbb{V}(Y_i|\mathcal{Z}) = e_k'S_G^{-1}\mathbb{E}[U_iU_i']S_G^{-1\prime}e_k \leq C||S_G^{-1}||^2 \leq C/r_G,$$
$$\mathbb{E}[W_i|\mathcal{Z}] = e_\ell'z_i/\sqrt{n}, \quad \mathbb{V}(W_i|\mathcal{Z}) = e_\ell'S_G^{-1}\mathbb{E}[U_iU_i']S_G^{-1\prime}e_\ell \leq C||S_G^{-1}||^2 \leq C/r_G.$$

Denote $z_{ik} = e_k'z_i$ and let $z_{[g]k} = (e_k'z_i)_{i\in g}$. By Assumption 2, $\sum_g ||z_{[g]}'z_{[g]}/n_g||/n$ is bounded a.s., then every entry of it is bounded a.s.: $\sum_g z_{[g]k}'z_{[g]k}/n_g/n \leq C$ for every $k \in \{1,...,L\}$ a.s.. Further, a.s.

$$\bar{\sigma}_{W_n}\bar{\sigma}_{Y_n}\sqrt{K} \leq C\sqrt{K}/r_G \to 0,$$
$$\sqrt{\breve{y}'\breve{y}\bar{\sigma}_{W_n}^2} \leq C\sqrt{\sum_g z_{[g]k}'z_{[g]k}/n_g/n}/\sqrt{r_G} \leq C/\sqrt{r_G} \to 0,$$
$$\sqrt{\breve{w}'\breve{w}\bar{\sigma}_{Y_n}^2} \leq C\sqrt{\sum_g z_{[g]k}'z_{[g]k}/n_g/n}/\sqrt{r_G} \leq C/\sqrt{r_G} \to 0.$$

Because

$$e_k'S_G^{-1}\hat{H}S_G^{-1}e_\ell = e_k'S_G^{-1}\sum_{g(i)\neq g(j)}\frac{p_{ij}X_iX_j'}{\sqrt{n_{g(i)}\ n_{g(j)}}}S_G^{-1}e_\ell = \sum_{g(i)\neq g(j)}\frac{p_{ij}Y_iW_j}{\sqrt{n_{g(i)}\ n_{g(j)}}},$$

and $p_{ij}\bar{y}_i\bar{w}_j/\sqrt{n_{g(i)}\ n_{g(j)}} = p_{ij}e_k'z_iz_j'e_\ell/(n\sqrt{n_{g(i)}\ n_{g(j)}})$, applying Lemma A1 and the conditional version of M, we deduce that for any $v > 0$ and $A_n = \{|e_k'S_G^{-1}\hat{H}S_G^{-1\prime}e_\ell/\sqrt{n_{g(i)}\ n_{g(j)}} - \sum_{g(i)\neq g(j)} p_{ij}e_k'z_iz_j'e_\ell/(n\sqrt{n_{g(i)}\ n_{g(j)}})| \geq v\}$, $P(A_n|\mathcal{Z}) \to 0$ a.s. By the dominated convergence theorem, $P(A_n) = \mathbb{E}[P(A_n|\mathcal{Z})] \to 0$.

The preceding argument establishes the first conclusion for the (k,ℓ)th element. Doing so for every element of $S_G^{-1}\hat{H}S_G^{-1}$ completes the proof of the first conclusion.

For the second conclusion, let $Y_i = e_k'S_G^{-1}X_i = e_k'z_i/\sqrt{n}+e_k'S_G^{-1}U_i$ for some k as before and $W_i = \varepsilon_i$. Note that a.s. $\mathbb{E}[W_i|\mathcal{Z}] = 0$ and $\mathbb{V}(W_i|\mathcal{Z}) \leq C$ by Assumption 3. Then by Lemma A1,

$$\mathbb{E}\left[\left(e_k'S_G^{-1}\sum_{g(i)\neq g(j)}\frac{p_{ij}X_i\varepsilon_j}{\sqrt{n_{g(i)}\ n_{g(j)}}}\right)^2|\mathcal{Z}\right] \leq CK/r_G + C.$$

The conclusion then follows by M. $\square$

Lemma A6. If Assumptions 1–4 are satisfied, then

$$S_G^{-1}\hat{H}S_G^{-1\prime} = H_G + o_p(1).$$

Proof. Denote $I_{[g]}$ to be $n_g \times n_g$ identity matrix. Using Lemma A5,

$$\begin{aligned}
S_G^{-1}\hat{H}S_G^{-1\prime} = & \sum_{\substack{i,j \\ g(i)\neq g(j)}} p_{ij} z_i z_j' / \sqrt{n^2 \ n_{g(i)} \ n_{g(j)}} + o_p(1) \\
= & \sum_{g,h} z_{[g]}' P_{[g,h]} z_{[h]} / \sqrt{n^2 n_g n_h} - \sum_g z_{[g]}' P_{[g,g]} z_{[g]} / (n n_g) + o_p(1) \\
= & \sum_g z_{[g]}' / \sqrt{n_g} \left(\sum_h P_{[g,h]} z_{[h]} / \sqrt{n_h} - z_{[g]} / \sqrt{n_g} \right) / n \\
& + \sum_g z_{[g]}' (I_{[g]} - P_{[g,g]}) z_{[g]} / (n n_g) + o_p(1).
\end{aligned}$$

It suffices to show that $\sum_g z_{[g]}'/\sqrt{n_g}\left(\sum_h P_{[g,h]} z_{[h]}/\sqrt{n_h} - z_{[g]}/\sqrt{n_g}\right)/n = o(1)$ a.s. Let $\bar{z}_{[g]} = \sum_h P_{[g,h]} z_{[h]}/\sqrt{n_h}$ and $\bar{z}_i = \sum_j p_{ij} z_j/\sqrt{n_{g(j)}}$. Denote $\dot{z} = (z_1'/\sqrt{n_{g(1)}}, ..., z_n'/\sqrt{n_{g(n)}})'$ Using Assumption 4,

$$\begin{aligned}
\sum_g ||\bar{z}_{[g]} - z_{[g]}/\sqrt{n_g}||^2/n = & \sum_i ||z_i/\sqrt{n_{g(i)}} - \bar{z}_i||_F^2/n \\
& = ||(I-P)\dot{z}||^2/n = \mathrm{tr}(\dot{z}'(I-P)\dot{z}/n) \\
& = \mathrm{tr}((\dot{z} - Z\pi_K)'(I-P)(\dot{z} - Z\pi_K)/n) \leq \mathrm{tr}((\dot{z} - Z\pi_K)'(\dot{z} - Z\pi_K)/n) \\
& = \sum_i ||z_i/\sqrt{n_{g(i)}} - Z_i'\pi_K||^2/n \to 0 \ \text{ a.s.}
\end{aligned}$$

Further, by Assumption 2, $\sum_g \left|\left|z_{[g]}/\sqrt{n_g}\right|\right|^2/n \leq C$. By CS, it follows that

$$\begin{aligned}
\left|\left|\sum_g z_{[g]}'/\sqrt{n_g}\left(\bar{z}_{[g]} - z_{[g]}/\sqrt{n_g}\right)/n\right|\right| & \leq \sum_g \left|\left|z_{[g]}/\sqrt{n_g}\right|\right| \times \left|\left|\bar{z}_{[g]} - z_{[g]}/\sqrt{n_g}\right|\right|/n \\
& \leq \sqrt{\sum_g \left|\left|z_{[g]}/\sqrt{n_g}\right|\right|^2/n}\sqrt{\sum_g \left|\left|\bar{z}_{[g]} - z_{[g]}/\sqrt{n_g}\right|\right|^2/n} \to 0 \ \text{ a.s.}
\end{aligned}$$

$\square$

Proof of Theorem 1. First note that that by Assumption 2, $\lambda_{\min}(S_G S_G'/r_G) \geq \lambda_{\min}(\tilde{S}_G \tilde{S}_G') \geq C$, we have

$$\left|\left|S_G'(\hat{\beta}_{WLCOIV} - \beta)/\sqrt{r_G}\right|\right| \geq \sqrt{\lambda_{\min}(S_G S_G'/r_G)}\left|\left|\hat{\beta}_{WLCOIV} - \beta\right|\right| \geq C\left|\left|\hat{\beta}_{WLCOIV} - \beta\right|\right|.$$

Therefore, $S_G'(\hat{\beta}_{WLCOIV} - \beta)/\sqrt{r_G} \xrightarrow{p} 0$ implies $\hat{\beta}_{WLCOIV} \xrightarrow{p} \beta$.

Lemma A6 and Assumption 2 imply that $\lambda_{\min}(S_G^{-1}\hat{H}S_G^{-1}) \geq C > 0$ w.p.a. 1, hence $\left(S_G^{-1}\hat{H}S_G^{-1\prime}\right)^{-1} = O_p(1)$. By Lemma A5,

$$r_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta) = \left(S_G^{-1}\hat{H}S_G^{-1\prime}\right)^{-1} S_G^{-1} \sum_{g(i)\neq g(j)} X_i p_{ij}\varepsilon_j/\sqrt{n_{g(i)}\ n_{g(j)}\ r_G} = O_p(1)o_p(1) \xrightarrow{p} 0,$$

i.e. for any $v > 0$ $\Pr(||r_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta)|| \geq v|\mathcal{Z}) \to 0$. By the dominated convergence theorem,

$$\Pr(||r_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta)|| \geq v) = \mathbb{E}[\Pr(||r_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta)|| \geq v|\mathcal{Z})] \to 0,$$

hence $r_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta) \xrightarrow{p} 0$. $\square$

Proof of Theorem 2. Define

$$Y_G = \sum_{g=1}^{G} z_{[g]}'(I_{[g]} - P_{[g,g]})\varepsilon_{[g]}/\sqrt{nn_g^2} + S_G^{-1}\sum_{\substack{g,g'\\g'\neq g}}\sum_{\substack{i\in g\\j\in g'}} U_i\varepsilon_j p_{ij}/\sqrt{n_g n_{g'}}.$$

By Assumptions 2-4 and similarly to the proof of Lemma A6,

$$\begin{aligned}\mathbb{E}\left[\left|\left|\sum_{g=1}^{G}(z_{[g]}/\sqrt{n_g} - \bar{z}_{[g]})'\varepsilon_{[g]}/\sqrt{nn_g}\right|\right|^2|\mathcal{Z}\right] &= \sum_{g=1}^{G}\left|\left|z_{[g]}/\sqrt{n_g} - \bar{z}_{[g]}\right|\right|^2\mathbb{E}\left[\left|\left|\varepsilon_{[g]}\right|\right|^2|\mathcal{Z}\right]/(nn_g)\\ &\leq C\sum_{g=1}^{G}\left|\left|z_{[g]}/\sqrt{n_g} - \bar{z}_{[g]}\right|\right|^2/n \to 0 \ \text{ a.s.}\end{aligned}$$

Then by M,

$$\begin{aligned}S_G^{-1}\sum_{\substack{g,g'\\g'\neq g}}\sum_{\substack{i\in g\\j\in g'}}\frac{X_i\varepsilon_j p_{ij}}{\sqrt{n_g n_{g'}}} - Y_G &= \sum_{\substack{g,g'\\g'\neq g}}\frac{z_{[g]}'P_{[g,g']}\varepsilon_{[g']}}{\sqrt{nn_g n_{g'}}} - \sum_g\frac{z_{[g]}'\varepsilon_{[g]}}{\sqrt{nn_g^2}} + \sum_g\frac{z_{[g]}'P_{[g,g]}\varepsilon_{[g]}}{\sqrt{nn_g^2}}\\ &= \sum_{g,g'}\frac{z_{[g]}'P_{[g,g']}\varepsilon_{[g']}}{\sqrt{nn_g n_{g'}}} - \sum_g\frac{z_{[g]}'\varepsilon_{[g]}}{\sqrt{nn_g^2}}\\ &= \sum_{g=1}^{G}(\bar{z}_{[g]} - z_{[g]}/\sqrt{n_g})'\varepsilon_{[g]}/\sqrt{nn_g} \xrightarrow{p} 0.\end{aligned}$$

Let $\Gamma_G = \mathbb{V}(Y_G|\mathcal{Z})$, so

$$\begin{aligned}\Gamma_G = &\sum_{g=1}^{G} z_{[g]}'(I_{[g]} - P_{[g,g]})\mathbb{E}\left[\varepsilon_{[g]}\varepsilon_{[g]}'\right](I_{[g]} - P_{[g,g]})z_{[g]}/(nn_g^2)\\ &+ S_G^{-1}\sum_{\substack{g,g'\\g'\neq g}}\sum_{\substack{i,k\in g\\j,\ell\in g'}}(\mathbb{E}[U_iU_k'|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}] + \mathbb{E}[U_i\varepsilon_k|\mathcal{Z}]\mathbb{E}[\varepsilon_jU_\ell'|\mathcal{Z}])p_{ij}p_{k\ell}S_G^{-1\prime}/(n_g n_{g'}).\end{aligned}$$

$||S_G^{-1}|| \le C/\sqrt{r_G}$ by Assumption 2, and, using CS,

$$\sum_{g,h}\sum_{\substack{i,k\in g\\ j,\ell\in h}} |p_{ij}p_{k\ell}|/(n_g n_h) = \sum_{g,h}\left(\sum_{\substack{i\in g\\ j\in h}} \frac{|p_{ij}|}{\sqrt{n_g n_h}}\right)^2 \le \sum_{g,h}\sum_{\substack{i\in g\\ j\in h}} p_{ij}^2 = \sum_g \sum_{i\in g} p_{ii} = K.$$

By Assumption 1 and Assumption 3,

$$\begin{aligned}
&\left|\left|\sum_{g=1}^G z_{[g]}'(I_{[g]} - P_{[g,g]})\mathbb{E}\left[\varepsilon_{[g]}\varepsilon_{[g]}'|\mathcal{Z}\right](I_{[g]} - P_{[g,g]})z_{[g]}/(nn_g^2)\right|\right| \\
&\le \sum_g^G ||z_{[g]}||^2/n_g||I_{[g]} - P_{[g,g]}||^2||\mathbb{E}[\varepsilon_{[g]}\varepsilon_{[g]}'|\mathcal{Z}]||/n_g/n \\
&\le C\sum_g^G ||z_{[g]}||^2/n_g/n \le C.
\end{aligned}$$

Then by boundedness of K/r_G, $||\Gamma_G|| \le C$ a.s.n.

By Assumption 2, $\lambda_{\min}(\Gamma_G) \ge C > 0$ a.s.n, and $||\Gamma_G^{-1}|| \le C$ a.s.n.

Let α be an $L \times 1$ nonzero vector. Let $W_{gG} = z_{[g]}'(I_{[g]} - P_{[g,g]})\varepsilon_{[g]}/\sqrt{nn_g^2}$, $c_{1G} = \Gamma_G^{-1/2}\alpha$, and $c_{2G} = \sqrt{K}S_G^{-1}\Gamma_G^{-1/2}\alpha$. Condition (i) of Lemma A2 is satisfied.

Condition (ii) of Lemma A2 is satisfied by Assumptions 1-3:

$$\begin{aligned}
&\left|\left|\sum_{g=1}^G \mathbb{E}[W_{gG}'W_{gG}|\mathcal{Z}]\right|\right| = \left|\left|\sum_{g=1}^G z_{[g]}'(I_{[g]} - P_{[g,g]})\mathbb{E}\left[\varepsilon_{[g]}\varepsilon_{[g]}'|\mathcal{Z}\right](I_{[g]} - P_{[g,g]})z_{[g]}/(nn_g^2)\right|\right| \\
&\le C\sum_{g=1}^G \left|\left|z_{[g]}\right|\right|^2/n_g\left|\left|I_{[g]} - P_{[g,g]}\right|\right|^2/n \le C\sum_{g=1}^G \left|\left|z_{[g]}'z_{[g]}/n_g\right|\right|/n \le C.
\end{aligned}$$

Condition (iii) of Lemma A2 is satisfied by Assumptions 3 and 5.

Condition (iv) of Lemma A2 is satisfied by Assumptions 3 and 5:

$$\begin{aligned}
\sum_g \mathbb{E}[||W_{gG}||^4|\mathcal{Z}] &\le C\sum_g ||z_{[g]}/\sqrt{n_g}||^4||I_{[g]} - P_{[g,g]}||^4\mathbb{E}[||\varepsilon_{[g]}||^4|\mathcal{Z}]/n_g^2/n^2 \\
&\le C\sum_g ||z_{[g]}/\sqrt{n_g}||^4/n^2 \to 0 \quad \text{a.s.}
\end{aligned}$$

Condition (v) of Lemma A2 is satisfied by Assumption 1.

Further note that $||c_{1G}|| \le C$ and $||c_{2G}|| \le C$ a.s.n. Also by construction

$$\Xi_G = \mathbb{V}\left(c_{1G}'\sum_g W_{gG} + c_{2G}'\sum_{\substack{g,g'\\ g'\neq g}}\sum_{\substack{i\in g\\ j\in g'}} U_i\varepsilon_j p_{ij}/\sqrt{Kn_g n_{g'}}|\mathcal{Z}\right) = \mathbb{V}\left(\alpha'\Gamma_G^{-1/2}{}'Y_G|\mathcal{Z}\right) = \alpha'\alpha.$$

By Lemma A2, it follows that

$$(\alpha'\alpha)^{-1/2}\alpha'\Gamma_G^{-1/2}Y_G \xrightarrow{d} \mathcal{N}(0,1) \quad \text{a.s..}$$

By the Cramér-Wold device, $\Gamma_G^{-1/2}Y_G \xrightarrow{d} \mathcal{N}(0, I_L)$ a.s.

It follows that $\Gamma_G^{-1/2}Y_G = O_p(1)$. From the argument above, $||\Gamma_G|| \leq C$ a.s.n., then $||\Gamma_G^{1/2}|| \leq C$ a.s.n.; by Assumption 2, $||H_G|| \leq C$ and $\lambda_{\min}(\Gamma_G) \geq C > 0$, it follows that $V_G = H_G^{-1}\Gamma_G H_G^{-1}$, $\lambda_{\min}(V_G) \geq C > 0$ a.s.n., then $||V_G^{-1/2}|| \leq C$ a.s.n. Then $\Gamma_G^{1/2} = O_p(1)$ and $V_G^{-1/2} = O_p(1)$. Let $B_G = V_G^{-1/2}H_G^{-1}\Gamma_G^{1/2}$. Using Assumption 2, it follows that $B_G\Gamma_G^{-1/2} = V_G^{-1/2}H_G^{-1} = O_p(1)$ and $V_G^{-1/2}Y_G = V_G^{-1/2}\Gamma_G^{1/2}\Gamma_G^{-1/2}Y_G = O_p(1)$. Combining it with Lemma A6, Mann-Wald theorem, and the definition of $\hat{\beta}_{WLCOIV}$,

$$\begin{aligned}
V_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta) &= V_G^{-1/2}(S_G^{-1}\hat{H}S_G^{-1\prime})^{-1}S_G^{-1}\sum_{g(i)\neq g(j)} p_{ij}X_i\varepsilon_j/\sqrt{n_{g(i)}n_{g(j)}} \\
&= V_G^{-1/2}(H_G^{-1} + o_p(1))\Gamma_G^{1/2}\Gamma_G^{-1/2}(Y_G + o_p(1)) \\
&= B_G\Gamma_G^{-1/2}Y_G + o_p(1).
\end{aligned}$$

Note that $B_G'B_G = I_L$ and B_G is a function of $\mathcal{Z}$ only. Then using the Slutsky theorem,

$$V_G^{-1/2}S_G'(\hat{\beta}_{WLCOIV} - \beta) = B_G\Gamma_G^{-1/2}Y_G + o_p(1) \xrightarrow{d} \mathcal{N}(0, I_L).$$

□

Proof of Theorem 3. Notice that $\sqrt{r_G/K}\sum_g z_{[g]}'M_{[g,g]}\varepsilon_{[g]}/\sqrt{nn_g^2} \xrightarrow{p} 0$. Then for

$$Y_G = \sqrt{r_G}S_G^{-1}\sum_{\substack{g,g'\\ g'\neq g}}\sum_{\substack{i\in g\\ j\in g'}} U_i\varepsilon_j p_{ij}/\sqrt{Kn_g n_{g'}},$$

we have, similar to the proof of Theorem 2, $\sqrt{r_G/K}S_G^{-1}\sum_{\substack{g,g'\\ g'\neq g}} X_{[g]}'P_{g,g'}\varepsilon_{[g']}/\sqrt{n_g n_{g'}} = Y_G + o_p(1)$. Further,

$$\Gamma_G = \mathbb{V}(Y_G|\mathcal{Z}) = r_G S_G^{-1}\sum_{\substack{g,g'\\ g'\neq g}}\sum_{\substack{i,k\in g\\ j,\ell\in g'}}(\mathbb{E}[U_iU_k'|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}] + \mathbb{E}[U_i\varepsilon_k|\mathcal{Z}]\mathbb{E}[\varepsilon_jU_\ell'|\mathcal{Z}])p_{ij}p_{k\ell}S_G^{-1\prime}/K.$$

Similar to the proof of Theorem 2, by Assumptions 2-3, $\|\Gamma_G\| \leq C(r_G/K)(K/r_G) \leq C$ a.s.n. Let $\bar{L}_G$ be any sequence of bounded matrices such that $\lambda_{\min}(\bar{L}_G\Gamma_G\bar{L}_G') \geq C > 0$ a.s.n and let $\bar{Y}_G = (\bar{L}_G\Gamma_G\bar{L}_G')^{-1/2}\bar{L}_GY_G$. Let α be a nonzero vector and apply Lemma A2 with $W_{gG} = 0$, $c_{1G} = 0$, and $c_{2G} = \alpha'(\bar{L}_G\Gamma_G\bar{L}_G')^{-1/2}\bar{L}_G\sqrt{r_G}S_G^{-1}$.

We have $\mathbb{V}\left(c_{2G}'\sum_{g\neq g'}\sum_{\substack{i\in g\\ j\in g'}} U_i\varepsilon_jp_{ij}/\sqrt{Kn_gn_{g'}}|\mathcal{Z}\right) = \alpha'\alpha > 0$ by construction, and the other hypotheses of Lemma A2 can be verified as in the proof of Theorem 2.

By the conclusion of Lemma A2, $\alpha'\bar{Y}_G \xrightarrow{d} \mathcal{N}(0, \alpha'\alpha)$ a.s., and by the Cramér–Wold device, $\bar{Y}_G \xrightarrow{d} \mathcal{N}(0, I_L)$ a.s.

Let L_G be specified as in the statement such that $\lambda_{\min}(L_G V_G^* L_G') \geq C > 0$ a.s.n. Let $\bar{L}_G = L_G H_G^{-1}$, so $L_G V_G^* L_G' = \bar{L}_G \Gamma_G \bar{L}_G'$. Note that $\|(\bar{L}_G \Gamma_G \bar{L}_G')^{-1/2}\| \leq C$ and $\|\Gamma_G^{1/2}\| \leq C$ a.s.n. By Lemma A6 and Mann-Wald Theorem, $(S_G^{-1}\hat{H}S_G^{-1'})^{-1} = H_G^{-1} + o_p(1)$. Therefore, we have

$$(\bar{L}_G \Gamma_G \bar{L}_G')^{-1/2} L_G (S_G^{-1}\hat{H}S_G^{-1'})^{-1} = (\bar{L}_G \Gamma_G \bar{L}_G')^{-1/2} L_G (H_G^{-1} + o_p(1)).$$

Using $Y_G = O_p(1)$ by M, and boundedness of L_G, $\bar{L}_G$ and $(\bar{L}_G \Gamma_G \bar{L}_G')^{-1/2}$, we have

$$\begin{aligned}
&(L_G V_G^* L_G')^{-1/2} L_G \sqrt{r_G/K} S_G' (\hat{\beta}_{WLCOIV} - \beta) \\
&\quad = (\bar{L}_G \Gamma_G \bar{L}_G')^{-1/2} L_G (S_G^{-1}\hat{H}S_G^{-1'})^{-1} \sqrt{r_G/K} S_G^{-1} \sum_{\substack{g,g' \\ g' \neq g}} X_{[g]} P_{[g,g]} \varepsilon_{[g]} / \sqrt{n_g n_{g'}} \\
&\quad = (\bar{L}_G \Gamma_G \bar{L}_G')^{-1/2} L_G (H^{-1} + o_p(1))(Y_G + o_p(1)) \\
&\quad = \left((\bar{L}_G \Gamma_G \bar{L}_G')^{-1/2} \bar{L}_G + o_p(1) \right) (Y_G + o_p(1)) = \bar{Y}_G + o_p(1) \xrightarrow{d} \mathcal{N}(0, I_L).
\end{aligned}$$

□

In order to prove Theorem 4, we introduce the following notation:

$$\begin{aligned}
\dot{X}_{[g]} &= S_G^{-1} X_{[g]} / \sqrt{n_g} \\
\hat{\Sigma}_1 &= \sum_{g \neq h \neq g'} \dot{X}_{[g]} P_{[g,h]} \hat{\varepsilon}_{[h]} \hat{\varepsilon}_{[h]}' / n_h P_{[h,g']} \dot{X}_{[g']}' \\
\hat{\Sigma}_2 &= \sum_{g \neq h} \dot{X}_{[g]} P_{[g,h]} \hat{\varepsilon}_{[h]} \hat{\varepsilon}_{[h]}' / n_h P_{[h,g]} \dot{X}_{[g]}' + \sum_{g \neq h} \dot{X}_{[g]} P_{[g,h]} \hat{\varepsilon}_{[h]} \hat{\varepsilon}_{[g]}' / \sqrt{n_h n_g} P_{[g,h]} \dot{X}_{[h]}' \\
\dot{\Sigma}_1 &= \sum_{g \neq h \neq g'} \dot{X}_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon_{[h]}' / n_h P_{[h,g']} \dot{X}_{[g']}' \\
\dot{\Sigma}_2 &= \sum_{g \neq h} \dot{X}_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon_{[h]}' / n_h P_{[h,g]} \dot{X}_{[g]}' + \sum_{g \neq h} \dot{X}_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon_{[g]}' / \sqrt{n_h n_g} P_{[g,h]} \dot{X}_{[h]}'.
\end{aligned}$$

Lemma A7. If Assumptions 1-6 are satisfied, then $\hat{\Sigma}_1 - \dot{\Sigma}_1 = o_p(1)$ and $\hat{\Sigma}_2 - \dot{\Sigma}_2 = o_p(K/r_G)$.

Proof. To show the first conclusion, we use Lemma A4.

$$\begin{aligned}
\hat{\Sigma}_1 - \dot{\Sigma}_1 &= \sum_{g \neq h \neq m} \dot{X}_{[g]}' P_{[g,h]} \left(\hat{\varepsilon}_h \hat{\varepsilon}_h' - \varepsilon_h \varepsilon_h' \right) / n_h P_{[h,m]} \dot{X}_{[m]} \\
&= \sum_{g \neq h \neq m} \dot{X}_{[g]}' P_{[g,h]} X_{[h]} / \sqrt{n_h} \left(\hat{\beta} - \beta \right) \left(\hat{\beta} - \beta \right)' X_{[h]}' / \sqrt{n_h} P_{[h,m]} \dot{X}_{[m]} \\
&\quad - \sum_{g \neq h \neq m} \dot{X}_{[g]}' P_{[g,h]} \varepsilon_{[h]} / \sqrt{n_h} \left(\hat{\beta} - \beta \right)' X_{[h]}' / \sqrt{n_h} P_{[h,m]} \dot{X}_{[m]} \\
&\quad - \sum_{g \neq h \neq m} \dot{X}_{[g]}' P_{[g,h]} X_{[h]} / \sqrt{n_h} \left(\hat{\beta} - \beta \right) \varepsilon_{[h]}' / \sqrt{n_h} P_{[h,m]} \dot{X}_{[m]} \\
&= \hat{\Delta}_G \sum_{g \neq h \neq k} \dot{X}_{[g]}' P_{[g,k]} \left(\eta_{[k]}^{(1)} - \eta_{[k]}^{(2)} - \eta_{[k]}^{(3)} \right) / n_k P_{[k,h]} \dot{X}_{[h]},
\end{aligned}$$

where $\eta^{(1)}_{[k]} = X_{[k]}X'_{[k]}$, $\eta^{(2)}_{[k]} = \varepsilon_{[k]}1'_L X'_{[k]}$, $\eta^{(3)}_{[k]} = X_{[k]}1_L\varepsilon'_{[k]}$, 1_L is an L-vector of ones, and $\hat{\Delta}_G$ is a sequence converging to 0 in probability by $\hat{\beta} - \beta \xrightarrow{p} 0$.

Note that $S_G/\sqrt{n}$ is bounded, so by CS, $||\Upsilon_{[g]}|| = ||S_G z_{[g]}/\sqrt{n}|| \leq C\sqrt{n_g}$. Then by CS,

$$\begin{aligned}
\mathbb{E}[||\eta^{(2)}_{[k]}||^2|\mathcal{Z}] &= \mathbb{E}[||\eta^{(3)}_{[k]}||^2|\mathcal{Z}] \leq \mathbb{E}[||\Upsilon_{[k]}1_L\varepsilon'_{[k]}||^2|\mathcal{Z}] + \mathbb{E}[||U_{[k]}1_L\varepsilon'_{[k]}||^2|\mathcal{Z}] \\
&\leq C\mathbb{E}[||\Upsilon_{[k]}||^4|\mathcal{Z}]\mathbb{E}[||\varepsilon_{[k]}||^4|\mathcal{Z}] + C\mathbb{E}[||U_{[k]}||^2|\mathcal{Z}]\mathbb{E}[||\varepsilon_{[k]}||^2|\mathcal{Z}] \leq Cn_k^2, \\
\mathbb{E}[||\eta^{(1)}_{[k]}||^2|\mathcal{Z}] &\leq C\mathbb{E}[||\Upsilon_{[k]}||^4|\mathcal{Z}] + C\mathbb{E}[||U_{[k]}||^4|\mathcal{Z}] \leq Cn_k^2,
\end{aligned}$$

$$\mathbb{E}[|| \left(\eta^{(1)}_{[k]} - \eta^{(2)}_{[k]} - \eta^{(3)}_{[k]}\right)/n_k||^2|\mathcal{Z}] \leq \ Cn_k^2/n_k^2 = C.$$

Then by Lemma A4,

$$\hat{\Sigma}_1 - \dot{\Sigma}_1 = \hat{\Delta}_G O_p(1) = o_p(1)O_p(1) = o_p(1),$$

which provides the first conclusion.

Let $d_{[g]} = ||\Upsilon_{[g]}|| + ||\varepsilon_{[g]}|| + ||U_{[g]}||$, $\hat{A} = 1 + ||\hat{\beta}||$, $\hat{B} = ||\hat{\beta} - \beta||$. By the conclusion of Theorem 1, we have $\hat{A} = O_p(1)$ and $\hat{B} \xrightarrow{p} 0$. Hence, $||X_{[g]}|| \leq ||\Upsilon_{[g]}|| + ||U_{[g]}|| \leq d_{[g]}$, $\dot{X}_{[g]} \leq Cr_G^{-1/2}d_{[g]}/\sqrt{n_g}$, $||\hat{\varepsilon}_{[g]} - \varepsilon_{[g]}|| = ||X_{[g]}(\hat{\beta} - \beta)|| \leq Cd_{[g]}\hat{B}$, $||\hat{\varepsilon}_{[g]}|| = ||X_{[g]}(\beta - \hat{\beta}) + \varepsilon_{[g]}|| \leq Cd_{[g]}(1 + \hat{B}) \leq Cd_{[g]}\hat{A}$, $||\hat{\varepsilon}_{[g]}\hat{\varepsilon}'_{[g]} - \varepsilon_{[g]}\varepsilon'_{[g]}|| \leq \left(||\varepsilon_{[g]}|| + ||\hat{\varepsilon}_{[g]}||\right)||\hat{\varepsilon}_{[g]} - \varepsilon_{[g]}|| \leq Cd_{[g]}(1 + \hat{A})d_{[g]}\hat{B} \leq Cd^2_{[g]}\hat{A}\hat{B}$, $||\hat{\varepsilon}_{[h]}\hat{\varepsilon}'_{[g]} - \varepsilon_{[h]}\varepsilon'_{[g]}|| \leq ||\hat{\varepsilon}_{[g]}||\ ||\hat{\varepsilon}_{[h]} - \varepsilon_{[h]}|| + ||\varepsilon_{[h]}||\ ||\hat{\varepsilon}_{[g]} - \varepsilon_{[g]}|| \leq Cd_{[g]}d_{[h]}(1 + \hat{A})\hat{B} \leq Cd_{[g]}d_{[h]}\hat{A}\hat{B}$.

Also note that because $\mathbb{E}[d^2_{[g]}|\mathcal{Z}] \leq Cn_g$,

$$\mathbb{E}[\sum_{g\neq h}||P_{[g,h]}||^2 d^2_{[g]}/n_g d^2_{[h]}/n_h r_G^{-1}|\mathcal{Z}] \leq Cr_G^{-1}\sum_{g,h}||P_{[g,h]}||^2 = Cr_G^{-1}\sum_g \mathrm{tr}(P_{[g,g]}) = CK/r_G,$$

so $\sum_{g\neq h}||P_{[g,h]}||^2 d^2_{[g]}/n_g d^2_{[h]}/n_h r_G^{-1} = O_p(K/r_G)$ by M. Then it follows that

$$\begin{aligned}
|| \sum_{g\neq h}\dot{X}'_{[g]}P_{[g,h]}\left(\hat{\varepsilon}_{[h]}\hat{\varepsilon}'_{[h]} - \varepsilon_{[h]}\varepsilon'_{[h]}\right)/n_h P_{[h,g]}\dot{X}_{[g]}|| &\leq \sum_{g\neq h}||P_{[g,h]}||^2\ ||\dot{X}_{[g]}||^2||\hat{\varepsilon}_{[h]}\hat{\varepsilon}'_{[h]} - \varepsilon_{[h]}\varepsilon'_{[h]}||/n_h \\
&\leq Cr_G^{-1}\sum_{g\neq h}||P_{[g,h]}||^2 d^2_{[g]}/n_g d^2_{[h]}/n_h\hat{A}\hat{B} = o_p(K/r_G).
\end{aligned}$$

We also have

$$\begin{aligned}
&|| \sum_{g\neq h}\left(\dot{X}'_{[g]}P_{[g,h]}\hat{\varepsilon}_{[h]}\hat{\varepsilon}'_{[g]}P_{[g,h]}\dot{X}_{[h]}\ - \dot{X}'_{[g]}P_{[g,h]}\varepsilon_{[h]}\varepsilon'_{[g]}P_{[g,h]}\dot{X}_{[h]}\right)||/\sqrt{n_h n_g} \\
&\qquad\leq \sum_{g\neq h}||P_{[g,h]}||^2\ ||\dot{X}_{[g]}||\ ||\dot{X}_{[h]}||\ ||\hat{\varepsilon}_{[h]}\hat{\varepsilon}'_{[g]} - \varepsilon_{[h]}\varepsilon'_{[g]}||/\sqrt{n_h n_g} \\
&\qquad\leq Cr_G^{-1}\sum_{g\neq h}||P_{[g,h]}||^2 d^2_{[g]}/n_g d^2_{[h]}/n_h\hat{A}\hat{B} = o_p(K/r_G).
\end{aligned}$$

The second conclusion follows from T. □

Lemma A8. If Assumptions 1-6 are satisfied, then

$$\begin{aligned}
\dot{\Sigma}_1 &= \sum_{g\neq h\neq k} z_{[g]}'/\sqrt{n_g}P_{[g,k]}\mathbb{E}[\varepsilon_{[k]}\varepsilon_{[k]}'|\mathcal{Z}]/n_k P_{[k,h]}z_{[h]}/\sqrt{n_h}/n + o_p(1),\\
\dot{\Sigma}_2 &= \sum_{g\neq h} z_{[g]}'/\sqrt{n_g}P_{[g,h]}\mathbb{E}[\varepsilon_{[h]}\varepsilon_{[h]}'|\mathcal{Z}]/n_h P_{[h,g]}z_{[g]}/\sqrt{n_g}/n\\
&\quad + S_G^{-1}\sum_{g\neq g'}\sum_{\substack{i,k\in g\\ j,\ell\in g'}}(\mathbb{E}[U_iU_k'|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}] + \mathbb{E}[U_i\varepsilon_k|\mathcal{Z}]\mathbb{E}[\varepsilon_jU_\ell'|\mathcal{Z}])p_{ij}p_{k\ell}S_G^{-1\prime} + o_p(n_{\max}K/r_G).
\end{aligned}$$

Proof. To prove the first conclusion, we apply Lemma A4 with $W_{[g]}/\sqrt{n_g}$ equal to a column of $\dot{X}_{[g]}$, $Y_{[h]}/\sqrt{n_h}$ equal to a column of $\dot{X}_{[h]}$, and $\eta_{[k]} = \varepsilon_{[k]}\varepsilon_{[k]}'$.

To prove the second conclusion, write

$$
\begin{aligned}
\dot{\Sigma}_2 - & \sum_{g \neq h} z'_{[g]} P_{[g,h]} \mathbb{E}[\varepsilon_{[h]} \varepsilon'_{[h]} | \mathcal{Z}] P_{[h,g]} z_{[g]} / (n n_g n_h) \\
& - S_G^{-1} \sum_{\substack{g,h \\ h \neq g}} \sum_{\substack{i,k \in g \\ j,\ell \in h}} p_{ij} p_{k\ell} \left(\mathbb{E}[U_i U'_k | \mathcal{Z}] \mathbb{E}[\varepsilon_j \varepsilon_\ell | \mathcal{Z}] + \mathbb{E}[U_i \varepsilon_k | \mathcal{Z}] \mathbb{E}[\varepsilon_j U'_\ell | \mathcal{Z}] \right) S_G^{-1} / (n_g n_h) \\
= & \sum_{g \neq h} \dot{X}_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[h]} P_{[h,g]} \dot{X}'_{[g]} / (n_g n_h) + \sum_{g \neq h} \dot{X}_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[g]} P_{[g,h]} \dot{X}'_{[h]} / (n_g n_h) \\
& - \sum_{g \neq h} z'_{[g]} P_{[g,h]} \mathbb{E}[\varepsilon_{[h]} \varepsilon'_{[h]} | \mathcal{Z}] P_{[h,g]} z_{[g]} / (n n_g n_h) \\
& - S_G^{-1} \sum_{\substack{g,h \\ h \neq g}} \sum_{\substack{i,k \in g \\ j,\ell \in h}} p_{ij} p_{k\ell} \left(\mathbb{E}[U_i U'_k | \mathcal{Z}] \mathbb{E}[\varepsilon_j \varepsilon_\ell | \mathcal{Z}] + \mathbb{E}[U_i \varepsilon_k | \mathcal{Z}] \mathbb{E}[\varepsilon_j U'_\ell | \mathcal{Z}] \right) S_G^{-1} / (n_g n_h) \\
= & \sum_{g \neq h} z'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[h]} P_{[h,g]} z_{[g]} / (n n_g n_h) + \sum_{g \neq h} z'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[g]} P_{[g,h]} z_{[h]} / (n n_g n_h) \\
& + S_G^{-1} \sum_{g \neq h} U'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[h]} P_{[h,g]} z_{[g]} / (n n_g n_h) + S_G^{-1} \sum_{g \neq h} U'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[g]} P_{[g,h]} z_{[h]} / (n n_g n_h) \\
& + \sum_{g \neq h} z'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[h]} P_{[h,g]} U_{[g]} S_G^{-1} / (n n_g n_h) + \sum_{g \neq h} z'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[g]} P_{[g,h]} U_{[h]} S_G^{-1} / (n n_g n_h) \\
& + S_G^{-1} \sum_{\substack{g,h \\ h \neq g}} \sum_{\substack{i,k \in g \\ j,\ell \in h}} p_{ij} p_{k\ell} \left(U_i U'_k \varepsilon_j \varepsilon_\ell + U_i \varepsilon_k \varepsilon_j U'_\ell \right) S_G^{-1} / (n_g n_h) \\
& - \sum_{g \neq h} z'_{[g]} P_{[g,h]} \mathbb{E}[\varepsilon_{[h]} \varepsilon'_{[h]} | \mathcal{Z}] P_{[h,g]} z_{[g]} / (n n_g n_h) \\
& - S_G^{-1} \sum_{\substack{g,h \\ h \neq g}} \sum_{\substack{i,k \in g \\ j,\ell \in h}} p_{ij} p_{k\ell} \left(\mathbb{E}[U_i U'_k | \mathcal{Z}] \mathbb{E}[\varepsilon_j \varepsilon_\ell | \mathcal{Z}] + \mathbb{E}[U_i \varepsilon_k | \mathcal{Z}] \mathbb{E}[\varepsilon_j U'_\ell | \mathcal{Z}] \right) S_G^{-1} / (n_g n_h) \\
= & \sum_{g \neq h} z'_{[g]} P_{[g,h]} \left(\varepsilon_{[h]} \varepsilon'_{[h]} - \mathbb{E}[\varepsilon_{[h]} \varepsilon'_{[h]} | \mathcal{Z}] \right) P_{[h,g]} z_{[g]} / (n n_g n_h) \\
& + \sum_{g \neq h} z'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[g]} P_{[g,h]} z_{[h]} / (n n_g n_h) \\
& + S_G^{-1} \sum_{g \neq h} U'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[h]} P_{[h,g]} z_{[g]} / \left(\sqrt{n} n_g n_h \right) + S_G^{-1} \sum_{g \neq h} U'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[g]} P_{[g,h]} z_{[h]} / \left(\sqrt{n} n_g n_h \right) \\
& + \sum_{g \neq h} z'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[h]} P_{[h,g]} U_{[g]} S_G^{-1} / \left(\sqrt{n} n_g n_h \right) + \sum_{g \neq h} z'_{[g]} P_{[g,h]} \varepsilon_{[h]} \varepsilon'_{[g]} P_{[g,h]} U_{[h]} S_G^{-1} / \left(\sqrt{n} n_g n_h \right) \\
& + S_G^{-1} \sum_{\substack{g,h \\ h \neq g}} \sum_{\substack{i,k \in g \\ j,\ell \in h}} p_{ij} p_{k\ell} \left(U_i U'_k \varepsilon_j \varepsilon_\ell - \mathbb{E}[U_i U'_k | \mathcal{Z}] \mathbb{E}[\varepsilon_j \varepsilon_\ell | \mathcal{Z}] \right) S_G^{-1} / (n_g n_h) \\
& + S_G^{-1} \sum_{\substack{g,h \\ h \neq g}} \sum_{\substack{i,k \in g \\ j,\ell \in h}} p_{ij} p_{k\ell} \left(U_i \varepsilon_k \varepsilon_j U'_\ell - \mathbb{E}[U_i \varepsilon_k | \mathcal{Z}] \mathbb{E}[\varepsilon_j U'_\ell | \mathcal{Z}] \right) S_G^{-1} / (n_g n_h) = \sum_{i=1}^{8} \psi_i.
\end{aligned}
$$

It suffices to show that $\psi_i = o_p(K/r_G)$ for $i = 1, .., 8$.

$$\mathbb{E}[\psi_1^2|\mathcal{Z}] = \mathbb{E}\left[\left(\sum_{g\neq h} z'_{[g]}P_{[g,h]}\left(\varepsilon_{[h]}\varepsilon'_{[h]} - \mathbb{E}\left[\varepsilon_{[h]}\varepsilon'_{[h]}|\mathcal{Z}\right]\right)P_{[h,g]}z_{[g]}/(nn_gn_h)\right)^2|\mathcal{Z}\right]$$
$$= \sum_{\substack{g,g',h\\ h\notin\{g,g'\}}} \mathbb{E}\left[z'_{[g]}P_{[g,h]}\left(\varepsilon_{[h]}\varepsilon'_{[h]} - \mathbb{E}\left[\varepsilon_{[h]}\varepsilon'_{[h]}|\mathcal{Z}\right]\right)P_{[h,g]}z_{[g]}\right.$$
$$\left.\times\ z'_{[g']}P_{[g',h]}\left(\varepsilon_{[h]}\varepsilon'_{[h]} - \mathbb{E}\left[\varepsilon_{[h]}\varepsilon'_{[h]}|\mathcal{Z}\right]\right)P_{[h,g']}z_{[g']}|\mathcal{Z}\right]/(n\sqrt{n_gn_{g'}}n_h)^2$$
$$\leq \sum_{\substack{g,g',h\\ h\notin\{g,g'\}}} ||z_{[g]}||^2\ ||P_{[g,h]}||^2\mathbb{E}\left[||\varepsilon_{[h]}\varepsilon'_{[h]} - \mathbb{E}\left[\varepsilon_{[h]}\varepsilon'_{[h]}|\mathcal{Z}\right]||^2\ |\mathcal{Z}\right]||z_{[g']}||^2\ ||P_{[g',h]}||^2/(n\sqrt{n_gn_{g'}}n_h)^2$$
$$\leq \sum_{g,g',h} ||P_{[g,h]}||^2\mathbb{E}\left[||\varepsilon_{[h]}\varepsilon'_{[h]} - \mathbb{E}\left[\varepsilon_{[h]}\varepsilon'_{[h]}|\mathcal{Z}\right]||^2\ |\mathcal{Z}\right]||P_{[g',h]}||^2/(nn_h)^2$$
$$\leq C\sum_{g,g',h} ||P_{[g,h]}||^2\mathbb{E}\left[||\varepsilon_{[h]}||^4\ |\mathcal{Z}\right]||P_{[g',h]}||^2/(nn_h)^2$$
$$\leq C\sum_{g,g',h} ||P_{[g,h]}||^2\mathbb{E}\left[\sum_{i\in g}\varepsilon_i^4\ |\mathcal{Z}\right]||P_{[g',h]}||^2/(n^2n_h)$$
$$\leq C\sum_{g,g',h} ||P_{[g,h]}||^2||P_{[g',h]}||^2/n^2$$
$$\leq C\sum_h\left(\sum_g \mathrm{tr}(P_{[g,h]}P_{[h,g]})\right)^2/n^2 = C\sum_h\left(\mathrm{tr}(P_{[g,g]})\right)^2/n^2$$
$$\leq C\sum_h\left(\sum_g \mathrm{tr}(P_{[g,h]}P_{[h,g]})\right)^2/n^2 = Cn_{\max}K/n^2 \to 0,$$

then $\psi_1 \xrightarrow{p} 0$ by M. Notice that by Lemma B2 and Assumption 2, $||P_{[g,h]}|| \leq \sqrt{||P_{[g,g]}||\ ||P_{[h,h]}||} \leq 1$. Then $||P_{[g,h]}||^4 \leq ||P_{[g,h]}||^2$.

$$\mathbb{E}[\psi_2^2|\mathcal{Z}] = \mathbb{E}\left[\left(\sum_{g\neq h} z'_{[g]}P_{[g,h]}\varepsilon_{[h]}\varepsilon'_{[g]}P_{[g,h]}z_{[h]}/(nn_gn_h)\right)^2|\mathcal{Z}\right]$$
$$\leq C\sum_{g\neq h}\mathbb{E}\left[||z_{[g]}||^2\ ||P_{[g,h]}||^4\ ||\varepsilon_{[h]}||^2||\varepsilon_{[g]}||^2\ ||z_{[h]}||^2/(nn_gn_h)^2|\mathcal{Z}\right]$$
$$\leq C\sum_{g,h} ||P_{[g,h]}||^2/n^2$$
$$\leq C\sum_g \mathrm{tr}(P_{[g,g]})/n^2 = CK/n^2 \to 0,$$

then $\psi_2 \xrightarrow{p} 0$ by M.

$$
\begin{aligned}
\mathbb{E}\left[\psi_3^2|\mathcal{Z}\right] &= \mathbb{E}\left[\left(S_G^{-1}\sum_{g\neq h} U_{[g]}'P_{[g,h]}\varepsilon_{[h]}\varepsilon_{[h]}'P_{[h,g]}z_{[g]}/\left(\sqrt{n}n_g n_h\right)\right)^2|\mathcal{Z}\right] \\
&= S_G^{-1}\mathbb{E}\left[\sum_{\substack{g,h,h'\\ g\notin\{h,h'\}}} U_{[g]}'P_{[g,h]}\varepsilon_{[h]}\varepsilon_{[h]}'P_{[h,g]}z_{[g]}U_{[g]}'P_{[g,h']}\varepsilon_{[h']}\varepsilon_{[h']}'P_{[h',g]}z_{[g]}|\mathcal{Z}\right]/\left(\sqrt{n}n_g\sqrt{n_h}\sqrt{n_{h'}}\right)^2 S_G^{-1\prime} \\
&\quad + S_G^{-1}\mathbb{E}\left[\sum_{g\neq h} U_{[g]}'P_{[g,h]}\varepsilon_{[h]}\varepsilon_{[h]}'P_{[h,g]}z_{[g]}U_{[h]}'P_{[h,g]}\varepsilon_{[g]}\varepsilon_{[g]}'P_{[g,h]}z_{[h]}|\mathcal{Z}\right]/\left(\sqrt{n}n_g n_h\right)^2 S_G^{-1\prime} \\
&\le ||S_G^{-1}||^2\sum_{\substack{g,h,h'\\ g\notin\{h,h'\}}}\mathbb{E}\left[||U_{[g]}||^2|\mathcal{Z}\right]||P_{[g,h]}||^2||P_{[g,h']}||^2\mathbb{E}\left[||\varepsilon_{[h]}\varepsilon_{[h]}'||\ |\mathcal{Z}\right] \\
&\qquad \times\mathbb{E}\left[||\varepsilon_{[h']}\varepsilon_{[h']}'||\ |\mathcal{Z}\right]||z_{[g]}||^2/\left(\sqrt{n}n_g\sqrt{n_h}\sqrt{n_{h'}}\right)^2 \\
&\quad + ||S_G^{-1}||^2\sum_{g\neq h}\mathbb{E}\left[||U_{[g]}|||\mathcal{Z}\right]\mathbb{E}\left[||U_{[h]}|||\mathcal{Z}\right]||P_{[g,h]}||^4\mathbb{E}\left[||\varepsilon_{[h]}\varepsilon_{[h]}'||\ |\mathcal{Z}\right] \\
&\qquad \times\mathbb{E}\left[||\varepsilon_{[g]}\varepsilon_{[g]}'||\ |\mathcal{Z}\right]||z_{[g]}||\ ||z_{[h]}||/\left(\sqrt{n}n_g n_h\right)^2 \\
&\le r_G^{-1}n_{\max}K/n\to 0,
\end{aligned}
$$

then $\psi_3 \xrightarrow{p} 0$ by M. Similarly, $\psi_5 \xrightarrow{p} 0$.

$$
\begin{aligned}
\mathbb{E}\left[\psi_4^2|\mathcal{Z}\right] &= \mathbb{E}\left[\left(S_G^{-1}\sum_{g\neq h} U_{[g]}'P_{[g,h]}\varepsilon_{[h]}\varepsilon_{[g]}'P_{[g,h]}z_{[h]}/\left(\sqrt{n}n_g n_h\right)\right)^2|\mathcal{Z}\right] \\
&= S_G^{-1}\mathbb{E}\left[\sum_{\substack{g,h,g'\\ h\notin\{g,g'\}}} U_{[g]}'P_{[g,h]}\varepsilon_{[h]}\varepsilon_{[g]}'P_{[g,h]}z_{[h]}U_{[g']}'P_{[g',h]}\varepsilon_{[h]}\varepsilon_{[g']}'P_{[g',h]}z_{[h]}|\mathcal{Z}\right]/\left(\sqrt{n}n_h\sqrt{n_g}\sqrt{n_{g'}}\right)^2 S_G^{-1\prime} \\
&\quad + S_G^{-1}\mathbb{E}\left[\sum_{g\neq h} U_{[g]}'P_{[g,h]}\varepsilon_{[h]}\varepsilon_{[g]}'P_{[g,h]}z_{[h]}U_{[h]}'P_{[h,g]}\varepsilon_{[g]}\varepsilon_{[h]}'P_{[h,g]}z_{[g]}|\mathcal{Z}\right]/\left(\sqrt{n}n_g n_h\right)^2 S_G^{-1\prime} \\
&\le ||S_G^{-1}||^2\sum_{\substack{g,h,g'\\ h\notin\{g,g'\}}}\mathbb{E}\left[||U_{[g]}|||\mathcal{Z}\right]\mathbb{E}\left[||U_{[g']}|||\mathcal{Z}\right]||P_{[g,h]}||^2||P_{[g',h]}||^2\mathbb{E}\left[||\varepsilon_{[h]}||^2|\mathcal{Z}\right] \\
&\qquad \times\mathbb{E}\left[||\varepsilon_{[g]}||\ |\mathcal{Z}\right]\mathbb{E}\left[||\varepsilon_{[g']}||\ |\mathcal{Z}\right]||z_{[h]}||^2/\left(\sqrt{n}n_h\sqrt{n_g}\sqrt{n_{g'}}\right)^2 \\
&\quad + ||S_G^{-1}||^2\sum_{g\neq h}\mathbb{E}\left[||U_{[g]}|||\mathcal{Z}\right]\mathbb{E}\left[||U_{[h]}|||\mathcal{Z}\right]||P_{[g,h]}||^4\mathbb{E}\left[||\varepsilon_{[h]}\varepsilon_{[h]}'||\ |\mathcal{Z}\right] \\
&\qquad \times\mathbb{E}\left[||\varepsilon_{[g]}\varepsilon_{[g]}'||\ |\mathcal{Z}\right]||z_{[g]}||\ ||z_{[h]}||/\left(\sqrt{n}n_g n_h\right)^2 \\
&\le r_G^{-1}n_{\max}K/n\to 0,
\end{aligned}
$$

then $\psi_4 \xrightarrow{p} 0$ by M, and similarly $\psi_6 \xrightarrow{p} 0$. Notice that

$$
U_iU_k'\varepsilon_j\varepsilon_\ell - \mathbb{E}[U_iU_k'|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}] = \left(U_iU_k' - \mathbb{E}[U_iU_k'|\mathcal{Z}]\right)\varepsilon_j\varepsilon_\ell + \mathbb{E}[U_iU_k'|\mathcal{Z}]\left(\varepsilon_j\varepsilon_\ell - \mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)
$$

Let e_t be an L-vector with t'th coordinate equal to 1, and other coordinates are 0. Let $u_{it} = e_t' U_i$. Then

$$\begin{aligned}
\mathbb{E}\left[\left(e_{t_1}'\psi_7 e_{t_2}\right)^2 |\mathcal{Z}\right] &= \mathbb{E}\left[\left(S_G^{-1}\sum_{\substack{g,h\\h\neq g}}\sum_{\substack{i,k\in g\\j,\ell\in h}} p_{ij}p_{k\ell}\left(u_{it_1}u_{kt_2}\varepsilon_j\varepsilon_\ell - \mathbb{E}[u_{it_1}u_{kt_2}|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)S_G^{-1}/\left(n_g n_h\right)\right)^2 |\mathcal{Z}\right] \\
&\leq ||S_G^{-1}||^4 \sum_{\substack{g,h,h'\\g\neq h\neq h'}}\sum_{\substack{i,k,i',k'\in g\\j,\ell\in h\\j',\ell'\in h'}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}\mathbb{E}\left[\varepsilon_j\varepsilon_\ell|\mathcal{Z}\right]\mathbb{E}\left[\varepsilon_{j'}\varepsilon_{\ell'}|\mathcal{Z}\right]/\left(n_g^2 n_h n_{h'}\right) \\
&\quad\times \mathbb{E}\left[\left(u_{it_1}u_{kt_2} - \mathbb{E}[u_{it_1}u_{kt_2}|\mathcal{Z}]\right)\left(u_{i't_1}u_{k't_2} - \mathbb{E}[u_{i't_1}u_{k't_2}|\mathcal{Z}]\right)|\mathcal{Z}\right]| \\
&\quad + ||S_G^{-1}||^4 \sum_{\substack{g,h,g'\\g\neq h\neq g'}}\sum_{\substack{i,k,j',\ell'\in g\\j,\ell\in h\\i',k'\in g'}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}\mathbb{E}\left[\varepsilon_j\varepsilon_\ell|\mathcal{Z}\right]\mathbb{E}\left[u_{i't_1}u_{k't_2}|\mathcal{Z}\right]/\left(n_g^2 n_h n_{g'}\right) \\
&\quad\times \mathbb{E}\left[\left(u_{it_1}u_{kt_2} - \mathbb{E}[u_{it_1}u_{kt_2}|\mathcal{Z}]\right)\left(\varepsilon_{j'}\varepsilon_{\ell'} - \mathbb{E}[\varepsilon_{j'}\varepsilon_{\ell'}|\mathcal{Z}]\right)|\mathcal{Z}\right]| \\
&\quad + ||S_G^{-1}||^4 \sum_{\substack{g,h,h'\\g\neq h\neq h'}}\sum_{\substack{i,k\in g\\j,\ell,i',k'\in h\\j',\ell'\in h'}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}\mathbb{E}\left[\varepsilon_{j'}\varepsilon_{\ell'}|\mathcal{Z}\right]\mathbb{E}\left[u_{it_1}u_{kt_2}|\mathcal{Z}\right]/\left(n_g n_h^2 n_{h'}\right) \\
&\quad\times \mathbb{E}\left[\left(\varepsilon_j\varepsilon_\ell - \mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)\left(u_{i't_1}u_{k't_2} - \mathbb{E}[u_{i't_1}u_{k't_2}|\mathcal{Z}]\right)|\mathcal{Z}\right]| \\
&\quad + ||S_G^{-1}||^4 \sum_{\substack{g,h,g'\\g\neq h\neq g'}}\sum_{\substack{i,k\in g\\j,\ell,j',\ell'\in h\\i',k'\in g'}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}\mathbb{E}\left[u_{it_1}u_{kt_2}|\mathcal{Z}\right]\mathbb{E}\left[u_{i't_1}u_{k't_2}|\mathcal{Z}\right]/\left(n_g n_h^2 n_{g'}\right) \\
&\quad\times \mathbb{E}\left[\left(\varepsilon_j\varepsilon_\ell - \mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)\left(\varepsilon_{j'}\varepsilon_{\ell'} - \mathbb{E}[\varepsilon_{j'}\varepsilon_{\ell'}|\mathcal{Z}]\right)|\mathcal{Z}\right]| \\
&\quad + ||S_G^{-1}||^4 \sum_{\substack{g,h\\g\neq h}}\sum_{\substack{i,k,i',k'\in g\\j,\ell,j',\ell'\in h}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}/\left(n_g^2 n_h^2\right) \\
&\quad\times \left(\mathbb{E}\left[\varepsilon_j\varepsilon_\ell\varepsilon_{j'}\varepsilon_{\ell'}|\mathcal{Z}\right]\mathbb{E}\left[\left(u_{it_1}u_{kt_2} - \mathbb{E}[u_{it_1}u_{kt_2}|\mathcal{Z}]\right)\left(u_{i't_1}u_{k't_2} - \mathbb{E}[u_{i't_1}u_{k't_2}|\mathcal{Z}]\right)|\mathcal{Z}\right]\right. \\
&\qquad + \mathbb{E}\left[u_{it_1}u_{kt_2}|\mathcal{Z}\right]\mathbb{E}\left[u_{i't_1}u_{k't_2}|\mathcal{Z}\right]\mathbb{E}\left[\left(\varepsilon_j\varepsilon_\ell - \mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)\left(\varepsilon_{j'}\varepsilon_{\ell'} - \mathbb{E}[\varepsilon_{j'}\varepsilon_{\ell'}|\mathcal{Z}]\right)|\mathcal{Z}\right]| \\
&\qquad + \mathbb{E}\left[u_{it_1}u_{kt_2}|\mathcal{Z}\right]\mathbb{E}\left[\varepsilon_{i'}\varepsilon_{k'}|\mathcal{Z}\right]\mathbb{E}\left[\left(\varepsilon_j\varepsilon_\ell - \mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)\left(u_{j't_1}u_{\ell't_1} - \mathbb{E}[u_{j't_1}u_{\ell't_1}|\mathcal{Z}]\right)|\mathcal{Z}\right]| \\
&\qquad \left.+\mathbb{E}\left[u_{j't_1}u_{\ell't_2}|\mathcal{Z}\right]\mathbb{E}\left[\varepsilon_j\varepsilon_\ell|\mathcal{Z}\right]\mathbb{E}\left[\left(\varepsilon_{i'}\varepsilon_{k'} - \mathbb{E}[\varepsilon_{i'}\varepsilon_{k'}|\mathcal{Z}]\right)\left(u_{it_1}u_{kt_2} - \mathbb{E}[u_{it_1}u_{kt_2}|\mathcal{Z}]\right)|\mathcal{Z}\right]\right)| \\
&\leq Cr_G^{-2} \sum_{\substack{g,h,h'\\g\neq h\neq h'}}\sum_{\substack{i,k,i',k'\in g\\j,\ell\in h\\j',\ell'\in h'}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}|/\left(n_g^2 n_h n_{h'}\right) \\
&\quad + Cr_G^{-2}\sum_{\substack{g,h\\g\neq h}}\sum_{\substack{i,k,i',k'\in g\\j,\ell,j',\ell'\in h}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}|/\left(n_g^2 n_h^2\right).
\end{aligned}$$

Further, by CS

$$\begin{aligned}
\sum_{\substack{g,h,h' \\ g\neq h\neq h'}} \sum_{\substack{i,k,i',k'\in g \\ j,\ell\in h \\ j',\ell'\in h'}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}| / \left(n_g^2 n_h n_{h'}\right) &\leq \sum_g \left(\sum_h \left(\sum_{\substack{i\in g \\ j\in h}} |p_{ij}| / \sqrt{n_g n_h} \right)^2 \right)^2 \\
&\leq \sum_g \left(\sum_h \sum_{\substack{i\in g \\ j\in h}} p_{ij}^2 \right)^2 \\
&= \sum_g \left(\sum_{i\in g} p_{ii} \right)^2 \\
&\leq n_{\max} \sum_g \sum_{i\in g} p_{ii} = n_{\max} K
\end{aligned}$$

and

$$\begin{aligned}
\sum_{\substack{g,h \\ g\neq h}} \sum_{\substack{i,k,i',k'\in g \\ j,\ell,j',\ell'\in h}} |p_{ij}p_{k\ell}p_{i'j'}p_{k'\ell'}| / \left(n_g^2 n_h^2\right) &\leq \sum_g \sum_h \left(\sum_{\substack{i\in g \\ j\in h}} |p_{ij}| / \sqrt{n_g n_h} \right)^4 \\
&\leq \sum_g \sum_h \left(\sum_{\substack{i\in g \\ j\in h}} p_{ij}^2 \right)^2 \\
&\leq \sum_g \sum_h n_g n_h \sum_{\substack{i\in g \\ j\in h}} p_{ij}^2 \\
&\leq n_{\max}^2 \sum_g \sum_{i\in g} p_{ii} \leq n_{\max}^2 K.
\end{aligned}$$

Then

$$\mathbb{E}\left[\left(e_{t_1}' \psi_7 e_{t_2} \right)^2 | \mathcal{Z} \right] \leq C r_G^{-2} n_{\max}^2 K,$$

and $\left(e_{t_1}' \psi_7 e_{t_2} \right) = O_p(n_{\max}\sqrt{K}/r_G)$ by M. It establishes asymptotic order for (t_1, t_2)'th element of ψ_7; doing it for every element implies $\psi_7 = O_p(n_{\max}\sqrt{K}/r_G)$. Then $\psi_7 = o_p(n_{\max}K/r_G)$. Similarly, $\psi_8 = o_p(n_{\max}K/r_G)$. The second conclusion follows by T. $\square$

Proof of Theorem 4. Let $\hat{\Sigma}_{[g]} = \hat{\varepsilon}_{[g]}\hat{\varepsilon}'_{[g]}$. Let $\bar{z}_{[g]} = \sum_{h=1}^{G} P_{g,h} z_{[h]}/\sqrt{n_h}$. Notice that

$$\begin{aligned}
&\sum_g \left(\bar{z}'_{[g]}\hat{\Sigma}_{[g]}\bar{z}_{[g]} - z'_{[g]}P_{[g,g]}\hat{\Sigma}_{[g]}\bar{z}_{[g]}/\sqrt{n_g} - \bar{z}'_{[g]}\hat{\Sigma}_{[g]}P_{[g,g]}z_{[g]}/\sqrt{n_g}\right)/(nn_g)\\
=&\sum_g \left(\left(\sum_h P_{[g,h]}z_{[h]}\right)'\hat{\Sigma}_{[g]}\left(\sum_m P_{[g,m]}z_{[m]}\right) - z'_{[g]}P_{[g,g]}\hat{\Sigma}_{[g]}\left(\sum_h P_{[g,h]}z_{[h]}\right)\right.\\
&\left.\quad -\left(\sum_h P_{[g,h]}z_{[h]}\right)'\hat{\Sigma}_{[g]}P_{[g,g]}z_{[g]}\right)/(nn_g)\\
=&\sum_{g,h,m} z'_{[h]}P_{[h,g]}\hat{\Sigma}_{[g]}P_{[g,m]}z_{[m]}/(nn_g) - \sum_{g,h} z'_{[g]}P_{[g,g]}\hat{\Sigma}_{[g]}P_{[g,h]}z_{[h]}/(nn_g)\\
&\quad -\sum_{g,h} z'_{[h]}P_{[h,g]}\hat{\Sigma}_{[g]}P_{[g,g]}z_{[g]}/(nn_g)\\
=&\sum_{g\neq h\neq m} z'_{[h]}P_{[h,g]}\hat{\Sigma}_{[g]}P_{[g,m]}z_{[m]}/(nn_g) + \sum_{g\neq h} z'_{[g]}P_{[g,h]}\hat{\Sigma}_{[h]}P_{[h,g]}z_{[g]}/(nn_g)\\
&\quad -\sum_g z'_{[g]}P_{[g,g]}\hat{\Sigma}_{[g]}P_{[g,g]}z_{[g]}/(nn_g),
\end{aligned}$$

where the last equality follows from the proof of Theorem 4 in Chao et al. (2012).

Similar to the proof of Lemma A6 and using CS, a.s.

$$\begin{aligned}
\left\|\sum_g \left(\bar{z}'_{[g]}\hat{\Sigma}_{[g]}\bar{z}_{[g]} - z'_{[g]}\hat{\Sigma}_{[g]}z_{[g]}\right)/(nn_g)\right\| \le& \sum_g \left\|\hat{\Sigma}_{[g]}\right\|/n_g \left\|\bar{z}_{[g]} - z_{[g]}\right\|^2/n\\
&+2\sum_g \left\|\hat{\Sigma}_{[g]}\right\|/n_g\left\|z_{[g]}\right\|/\sqrt{n}\left\|\bar{z}_{[g]} - z_{[g]}\right\|/\sqrt{n}\\
\le& C\sum_g \left\|\bar{z}_{[g]} - z_{[g]}\right\|^2/n\\
&+C\sqrt{\sum_g \left\|z_{[g]}\right\|^2/n}\sqrt{\sum_g \left\|\bar{z}_{[g]} - z_{[g]}\right\|^2/n} \to 0.
\end{aligned}$$

Similarly, a.s.

$$\begin{aligned}
\left\|\sum_g \left(\bar{z}'_{[g]}\hat{\Sigma}_{[g]}P_{[g,g]}z_{[g]} - z'_{[g]}\hat{\Sigma}_{[g]}P_{[g,g]}z_{[g]}\right)/(nn_g)\right\| \le& \sum_g \left\|\hat{\Sigma}_{[g]}\right\|/n_g\left\|z_{[g]}\right\|\left\|P_{[g,g]}\right\|\left\|\bar{z}_{[g]} - z_{[g]}\right\|/n\\
\le& C\sqrt{\sum_g \left\|z_{[g]}\right\|^2/n}\sqrt{\sum_g \left\|\bar{z}_{[g]} - z_{[g]}\right\|^2/n} \to 0.\\
\left\|\sum_g \left(z'_{[g]}P_{[g,g]}\hat{\Sigma}_{[g]}\bar{z}_{[g]} - z'_{[g]}P_{[g,g]}\hat{\Sigma}_{[g]}z_{[g]}\right)/(nn_g)\right\| \le& C\sqrt{\sum_g \left\|z_{[g]}\right\|^2/n}\sqrt{\sum_g \left\|\bar{z}_{[g]} - z_{[g]}\right\|^2/n} \to 0.
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{g\neq h\neq m} z'_{[h]}P_{[h,g]}\hat{\Sigma}_{[g]}P_{[g,m]}z_{[m]}/(nn_g) =& \sum_g z'_{[g]}\left(I_{[g]} - P_{[g,g]}\right)\hat{\Sigma}_{[g]}\left(I_{[g]} - P_{[g,g]}\right)z_{[g]}/(nn_g)\\
&+\sum_{g\neq h} z'_{[g]}P_{[g,h]}\hat{\Sigma}_{[h]}P_{[h,g]}z_{[g]}/(nn_g) + o_{\text{a.s.}}(1).
\end{aligned}$$

From Lemmas A7-A8,

$$
\begin{aligned}
\hat{\Sigma}_1+\hat{\Sigma}_2 = & \sum_{g\neq h\neq m} z_{[h]}' P_{[h,g]}\hat{\Sigma}_{[g]} P_{[g,m]} z_{[m]}/\sqrt{n^2 n_g^2 n_h n_m} + \sum_{g\neq h} z_{[g]}' P_{[g,h]}\hat{\Sigma}_{[h]} P_{[h,g]} z_{[g]}/(n n_g n_h) \\
& + S_G^{-1} \sum_{g\neq g'} \sum_{\substack{i,k\in g\\ j,\ell\in g'}} (\mathbb{E}[U_i U_k'|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}] + \mathbb{E}[U_i\varepsilon_k|\mathcal{Z}]\mathbb{E}[\varepsilon_j U_\ell'|\mathcal{Z}]) p_{ij} p_{k\ell} S_G^{-1\prime}/(n_g n_{g'}) \\
& + o_p(1) + o_p(n_{\max}K/r_G) \\
= & \sum_g z_{[g]}' \left(I_{[g]} - P_{[g,g]}\right)\hat{\Sigma}_{[g]}\left(I_{[g]} - P_{[g,g]}\right) z_{[g]}/(n n_g) \\
& + S_G^{-1} \sum_{g\neq g'} \sum_{\substack{i,k\in g\\ j,\ell\in g'}} (\mathbb{E}[U_i U_k'|\mathcal{Z}]\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}] + \mathbb{E}[U_i\varepsilon_k|\mathcal{Z}]\mathbb{E}[\varepsilon_j U_\ell'|\mathcal{Z}]) p_{ij} p_{k\ell} S_G^{-1\prime} \\
& + o_p(1) + o_p(n_{\max}K/r_G) \\
= & \, \Omega_G + \Psi_G + o_p(1) + o_p(n_{\max}K/r_G).
\end{aligned}
$$

Suppose $n_{\max}K/r_G$ is bounded. H_G^{-1}, Ω_G and Ψ_G are bounded a.s.n. By Lemma A6,

$$
\begin{aligned}
S_G'\hat{V}S_G = & \left(S_G^{-1}\hat{H}_G S_G^{-1\prime}\right)^{-1}\left(\hat{\Sigma}_1+\hat{\Sigma}_2\right)\left(S_G^{-1}\hat{H}_G S_G^{-1\prime}\right)^{-1} \\
= & \left(H_G^{-1}+o_p(1)\right)(\Omega_G+\Psi_g+o_p(1))\left(H_G^{-1}+o_p(1)\right) = V_G + o_p(1),
\end{aligned}
$$

which provides the first conclusion.

Suppose now that $n_{\max}K/r_G \to \infty$. H_G^{-1}, Ω_G and $(r_G/(n_{\max}K))\Psi_G$ are bounded a.s.n. By Lemma A6,

$$
\begin{aligned}
S_G'\hat{V}S_G = & \left(S_G^{-1}\hat{H}_G S_G^{-1\prime}\right)^{-1}\left(\hat{\Sigma}_1+\hat{\Sigma}_2\right)\left(S_G^{-1}\hat{H}_G S_G^{-1\prime}\right)^{-1} \\
= & \left(H_G^{-1}+o_p(1)\right)((r_G/K)\Psi_g+o_p(1))\left(H_G^{-1}+o_p(1)\right) = V_G^* + o_p(1),
\end{aligned}
$$

which provides the second conclusion.

2.B Appendix: auxiliary results

Lemma B1. Let x and y be vectors, suppose there exists $C>0$ such that $||x||\leq C$ and $||y||\leq C$. Let A be a conformable matrix. Then $(x'Ay)^2 \leq x'xy'y||A'A|| \leq C||A'A||$.

Proof. By CS, $(x'Ay)^2 \leq x'xy'A'Ay$. Further, by Theorem 6, Chapter 11 in Magnus and Neudecker (1988), $y'A'Ay \leq ||A'A||y'y$. Then $(x'Ay)^2 \leq x'xy'y||A'A|| \leq C||A'A||$. □

Lemma B2. Consider $P = Z(Z'Z)^{-1}Z'$, and let $P_{[g,h]}$ be a submatrix of P. Then

1. $\sum_h \operatorname{tr}(P_{[g,h]}P_{[h,g]}) = \operatorname{tr}(P_{[g,g]})$,

2. $||P_{[g,h]}|| \leq \sqrt{||P_{[g,g]}||\ ||P_{[h,h]}||}$.

Proof. Notice that using the linearity of trace,

$$\begin{aligned}
\sum_h \mathrm{tr}(P_{[g,h]}P_{[h,g]}) &= \mathrm{tr}(\sum_h P_{[g,h]}P_{[h,g]}) \\
&= \mathrm{tr}(\sum_h Z'_{[g]}(Z'Z)^{-1}Z_{[h]}Z'_{[h]}(Z'Z)^{-1}Z_{[g]}) \\
&= \mathrm{tr}(Z'_{[g]}(Z'Z)^{-1}\sum_h Z_{[h]}Z'_{[h]}(Z'Z)^{-1}Z_{[g]}) \\
&= \mathrm{tr}(Z'_{[g]}(Z'Z)^{-1}(Z'Z)(Z'Z)^{-1}Z_{[g]}) \\
&= \mathrm{tr}(Z'_{[g]}(Z'Z)^{-1}Z_{[g]}) \\
&= \mathrm{tr}(P_{[g,g]}),
\end{aligned}$$

hence the first conclusion holds.

Then consider

$$\begin{aligned}
||P_{[g,h]}|| &= ||Z'_{[g]}(Z'Z)^{-1/2}(Z'Z)^{-1/2}Z_{[h]}|| \\
&\leq ||Z'_{[g]}(Z'Z)^{-1/2}||\ ||(Z'Z)^{-1/2}Z_{[h]}|| \\
&= \sqrt{||Z'_{[g]}(Z'Z)^{-1/2}(Z'Z)^{-1/2}Z_{[g]}||}\sqrt{||Z_{[h]}(Z'Z)^{-1/2}(Z'Z)^{-1/2}Z_{[h]}||} \\
&= \sqrt{||P_{[g,g]}||}\sqrt{||P_{[h,h]}||},
\end{aligned}$$

hence the second conclusion holds. □

Lemma B4. Suppose that

(a) $P = P(\mathcal{Z})$ is a symmetric, idempotent matrix with $\mathrm{rank}(P) = K$ and $p_{ii} \leq C < 1$;

(b) $(\varepsilon_{[g]}, u_{[g]})_{g=1}^G$ are independent conditional on $\mathcal{Z}$; *-$n_{\max}/K \to 0$;

(ii) $\{C_g\}_{g=1}^G$ is a partition of $\{1, ..., n\}$;

(c) there exists a constant C such that a.s., $\sup_i \mathbb{E}[u_i^4|\mathcal{Z}] \leq C$, $\sup_i \mathbb{E}[\varepsilon_i^4|\mathcal{Z}] \leq C$, and $\sup_i |\phi_i(\mathcal{Z})| \leq C$. Then a.s.

Then, a.s.,

(i) $\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=2}^G \sum_{g'<g} \sum_{\substack{i,k\in g \\ j,\ell\in g'}} \frac{p_{ij}p_{k\ell}}{n_g n_{g'}}\phi_{ik}\left(\varepsilon_j\varepsilon_\ell - \mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)\right)^2 |\mathcal{Z}\right] \to 0,$

(ii) $\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=2}^G \sum_{g'<g} \sum_{\substack{i,k\in g \\ j,\ell\in g'}} \frac{p_{ij}p_{k\ell}}{n_g n_{g'}}\phi_{ik}\left(u_j\varepsilon_\ell - \mathbb{E}[u_j\varepsilon_\ell|Z]\right)\right)^2 |\mathcal{Z}\right] \to 0,$

(iii) $\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\ j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_g n_{g'}}\phi_{ik}\left(\varepsilon_j u_\ell - \mathbb{E}[\varepsilon_j u_\ell|Z]\right)\right)^2 |\mathcal{Z}\right] \to 0,$

(iv) $\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\ j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_g n_{g'}}\phi_{ik}\left(u_j u_\ell - \mathbb{E}[u_j u_\ell|Z]\right)\right)^2 |\mathcal{Z}\right] \to 0,$

(v) $\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=3}^{G}\sum_{h<g'<g}\sum_{\substack{i,k\in g\\ j\in g'\\ \ell\in h}}\frac{p_{ij}p_{k\ell}}{n_g\sqrt{n_{g'}n_h}}\phi_{ik}\varepsilon_j\varepsilon_\ell\right)^2 |\mathcal{Z}\right] \to 0,$

(vi) $\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=3}^{G}\sum_{h<g'<g}\sum_{\substack{i,k\in g\\ j\in g'\\ \ell\in h}}\frac{p_{ij}p_{k\ell}}{n_g\sqrt{n_{g'}n_h}}\phi_{ik}u_j\varepsilon_\ell\right)^2 |\mathcal{Z}\right] \to 0,$

(vii) $\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=3}^{G}\sum_{h<g'<g}\sum_{\substack{i,k\in g\\ j\in g'\\ \ell\in h}}\frac{p_{ij}p_{k\ell}}{n_g\sqrt{n_{g'}n_h}}\phi_{ik}\varepsilon_j u_\ell\right)^2 |\mathcal{Z}\right] \to 0,$

(viii) $\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=3}^{G}\sum_{h<g'<g}\sum_{\substack{i,k\in g\\ j\in g'\\ \ell\in h}}\frac{p_{ij}p_{k\ell}}{n_g\sqrt{n_{g'}n_h}}\phi_{ik}u_j u_\ell\right)^2 |\mathcal{Z}\right] \to 0.$

Proof. By Hölder's inequality, for some (i,j,k,ℓ)

$$\begin{aligned}|\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_\ell] - \mathbb{E}[\varepsilon_i\varepsilon_j]\mathbb{E}[\varepsilon_k\varepsilon_\ell]| &\leq |\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_\ell]| + |\mathbb{E}[\varepsilon_i\varepsilon_j]|\ |\mathbb{E}[\varepsilon_k\varepsilon_\ell]| \\ &= \sup_i \mathbb{E}[\varepsilon_i^4] + \left(\sup_i \mathbb{E}[\varepsilon_i^2]\right)^2 \leq C.\end{aligned}$$

We show part (i) first:

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=2}^{G}\sum_{g'<g}\sum_{\substack{i,k\in g\\ j,\ell\in g'}}\frac{p_{ij}p_{k\ell}}{n_g n_{g'}}\phi_{ik}\left(\varepsilon_j\varepsilon_\ell-\mathbb{E}[\varepsilon_j\varepsilon_\ell|\mathcal{Z}]\right)\right)^2|\mathcal{Z}\right]\\
&=\frac{1}{K^2}\Bigg|\sum_{\substack{g,h\\ g'<\min\{g,h\}}}\sum_{\substack{i_1,k_1\in g\\ i_2,k_2\in h\\ j_1,\ell_1,j_2,\ell_2\in g'}}\frac{p_{i_1j_1}p_{k_1\ell_1}p_{i_2j_2}p_{k_2\ell_2}}{n_g n_{g'}^2 n_h}\phi_{i_1k_1}\phi_{i_2k_2}\\
&\qquad\qquad\times\left(\mathbb{E}[\varepsilon_{j_1}\varepsilon_{\ell_1}\varepsilon_{j_2}\varepsilon_{\ell_2}|\mathcal{Z}]-\mathbb{E}[\varepsilon_{j_1}\varepsilon_{\ell_1}|\mathcal{Z}]\mathbb{E}[\varepsilon_{j_2}\varepsilon_{\ell_2}|\mathcal{Z}]\right)\Bigg|\\
&\le\frac{C}{K^2}\sum_{g'}\left(\sum_{g>g'}\sum_{\substack{i_1,k_1\in g\\ j_1,\ell_1\in g'}}\frac{|p_{i_1j_1}p_{k_1\ell_1}|}{n_g n_{g'}}\right)\left(\sum_{h>g'}\sum_{\substack{i_2,k_2\in h\\ j_2,\ell_2\in g'}}\frac{|p_{i_2j_2}p_{k_2\ell_2}|}{n_h n_{g'}}\right)\\
&=\frac{C}{K^2}\sum_{g'}\left(\sum_{g>g'}\sum_{\substack{i_1\in g\\ j_1\in g'}}\frac{|p_{i_1j_1}|}{\sqrt{n_g n_{g'}}}\sum_{\substack{k_1\in g\\ \ell_1\in g'}}\frac{|p_{k_1\ell_1}|}{\sqrt{n_g n_{g'}}}\right)^2=\frac{C}{K^2}\sum_{g'}\left(\sum_{g>g'}\left(\sum_{\substack{i_1\in g\\ j_1\in g'}}\frac{|p_{i_1j_1}|}{\sqrt{n_g n_{g'}}}\right)^2\right)^2\\
&\le\frac{C}{K^2}\sum_{g'}\left(\sum_{g}\sum_{\substack{i_1\in g\\ j_1\in g'}}\frac{p_{i_1j_1}^2}{n_g n_{g'}}n_g n_{g'}\right)^2\le\frac{C}{K^2}\sum_{g}\left(\sum_{j\in g}p_{jj}\right)^2\\
&\le\frac{C}{K^2}\sum_{g}\sum_{j\in g}n_g p_{jj}^2\le\frac{Cn_{\max}}{K}\to 0.
\end{aligned}
$$

Parts (ii), (iii), and (iv) follow similarly by interchanging the roles of ε and u.

Now we show part (v):

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{K}\sum_{g=3}^{G}\sum_{h<g'<g}\sum_{\substack{i,k\in g\\ j\in g'\\ \ell\in h}}\frac{p_{ij}p_{k\ell}}{n_g\sqrt{n_{g'}n_h}}\phi_{ik}\varepsilon_j\varepsilon_\ell\right)^2|\mathcal{Z}\right]\\
&=\frac{1}{K^2}\left|\sum_{\substack{g,h\\ g_2<g_1<g,h}}\sum_{\substack{i_1,k_1\in g\\ i_2,k_2\in h\\ j_1,j_2\in g_1\\ \ell_1,\ell_2\in g_2}}\frac{1}{n_g n_{g_1} n_{g_2} n_h}p_{i_1j_1}p_{k_1\ell_1}p_{i_2j_2}p_{k_2\ell_2}\phi_{i_1k_1}\phi_{i_2k_2}\mathbb{E}\left[\varepsilon_{j_1}\varepsilon_{\ell_1}\varepsilon_{j_2}\varepsilon_{\ell_2}|\mathcal{Z}\right]\right|\\
&\leq\frac{C}{K^2}\sum_{\substack{g_2\\ g_1>g_2}}\frac{1}{n_{g_1}n_{g_2}}\left(\sum_{g>g_1}\sum_{\substack{i_1,k_1\in g\\ j_1\in g_1\\ \ell_1\in g_2}}\frac{1}{n_g}|p_{i_1j_1}p_{k_1\ell_1}|\right)\left(\sum_{h>g_1}\sum_{\substack{i_2,k_2\in h\\ j_2\in g_1\\ \ell_2\in g_2}}\frac{1}{n_h}|p_{i_2j_2}p_{k_2\ell_2}|\right)\\
&\leq\frac{C}{K^2}\sum_{\substack{g_2\\ g_1>g_2}}\frac{1}{n_{g_1}n_{g_2}}\sum_{g>g_1}\sum_{\substack{i_1\in g\\ j_1\in g_1}}\frac{1}{n_g}p_{i_1j_1}^2 n_g n_{g_2}\sum_{g>g_1}\sum_{\substack{k_1\in g\\ \ell_1\in g_2}}\frac{1}{n_g}p_{k_1\ell_1}^2 n_g n_{g_1}\\
&=\frac{C}{K^2}\sum_{\substack{g_2\\ g_1>g_2}}\sum_{g>g_1}\sum_{\substack{i_1\in g\\ j_1\in g_1}}p_{i_1j_1}^2\sum_{g>g_1}\sum_{\substack{k_1\in g\\ \ell_1\in g_2}}p_{k_1\ell_1}^2\\
&=\frac{C}{K^2}\sum_{\substack{g_2\\ g_1>g_2}}\left(\sum_{g>g_1}\sum_{\substack{i_1\in g\\ j_1\in g_1}}p_{i_1j_1}^2\right)^2\leq\frac{C}{K^2}\sum_{g}\left(\sum_{j\in g}p_{jj}\right)^2\leq\frac{Cn_{\max}}{K}\to 0.
\end{aligned}
$$

Parts (vi), (vii), and (viii) follow similarly by interchanging the roles of ε and u. □

2.C Appendix: Asymptotic Properties of Large Projection Matrices

2.C.1 Introduction

We consider entries of the $n\times n$ projection matrix

$$P = Z(Z'Z)^{-1}Z',$$

where the $n\times\ell$ matrix Z contains independent and identically distributed (i.i.d.) copies z_{ij}, $i=1,...,n$, $j=1,...,\ell$, of random variable z. Specifically, we are interested in joint asymptotic behavior of both diagonal and off-diagonal elements when the underlying dimension ℓ (corresponding to the dimensionality of n column vectors z_i in Z'), which is the rank of P, grows proportionately to the dimension n (corresponding to the sample size of ℓ-vectors z_i) with the aspect ratio $\lim \ell/n \equiv \alpha \in (0,1)$. Note that by "high-dimensional" we refer to the smaller dimension, ℓ, as the larger dimension, n, is always high in any asymptotic analysis with $n\to\infty$. Also note that P is invariant to non-singular linear transformations of Z, hence the results will cover all rotated i.i.d. designs as well, including some types of dependence.

Traditionally, in the context of random matrix theory and dimension asymptotics, one focuses on the behavior of (eigenvalues of) large covariance matrices; less frequent are studies of large correlation matrices, setting aside Wigner random matrices (see, for example, the monograph by Z. Bai and Silverstein, 2010). Investigation of projection matrices is much more rare, but the conclusions of such investigations are useful in the theory of regression models with many regressors (e.g., Cattaneo et al., 2018, Mikusheva and Sølvsten, 2025) and instrumental variables models with many instruments (e.g., van Hasselt, 2010), where high-dimensional projection matrices and their properties play an important role. In these setups, one is confronted with the dimension asymptotics, while projection matrices constitute an important building block of the surrounding asymptotic theories.

For large dimensional projection matrices, Anatolyev and Yaskov (2017) studied a probability limit of the diagonal elements in various circumstances. Under the rotated i.i.d. design, this probability limit is constant and equals α. Anatolyev and Smirnov (2024) determined that in this case, the off-diagonal elements are $\sqrt{n}$-consistent for zero and asymptotically Gaussian with a simple form of variance. The driving force of such a limit is the asymptotic normality of high-dimensional bilinear forms, and the formula for the variance rests on the Marchenko-Pastur law. In the present paper, we refine and extend these results by deriving the joint asymptotic distribution of an arbitrary number of diagonal and off-diagonal elements. In particular, we show that the diagonal elements are also asymptotically Gaussian, derive the formula for their asymptotic variance, and find out that all the elements under consideration, both diagonal and off-diagonal, are asymptotically independent.

Technically, the derivations proceed in the following way. First, using the celebrated Woodbury identity for an inverse of rank-k-perturbated matrices, we work out the formulas for quadratic and bilinear forms with respect to such an inverse.[10] On the basis of these, we figure out the connection between the elements of the projection matrix P and the elements of its leave-a-few-out analog, say P_{-}. This helps break the dependence between the high-dimensional covariance matrix in the "denominator" $Z'Z$ and its "numerator" $Z \cdots Z'$ due to the presence of common elements. Second, the asymptotic Gaussianity of the elements of P_{-} is derived from the central limit theorem for quadratic forms in i.i.d. random variables. The asymptotic variances and covariances are computed using the properties of partitioned matrices and the random matrix theory. Finally, the asymptotic Gaussianity of the elements of the leave-a-few-out projection matrix P_{-} is translated to that for the elements of P.

Finally, we apply our main result to figure out the asymptotic distributions of Frobenius and spectral norms of diagonal and off-diagonal blocks of P. These results may be useful for statistical theories for regression or instrumental variables models, where many regressors or many instruments are accompanied with data clustering (e.g., Chao et al., 2023). Another potentially useful domain of applications relates to many regressor models with limited time-series dependence (e.g., Mikusheva and Sølvsten, 2025), where non-leading diagonals of projection matrices are involved.

The section is structured as follows. Section C.2 lays out the setup, lists the notation used, and discusses the assumptions made. Section C.3 contains helpful intermediate results formulated in two auxiliary lemmas. Section C.4 contains the main result on the joint

[10]These formulas may also have independent value for matrix theories.

asymptotics of elements of projection matrices. Section C.5 describes implications for asymptotics of submatrix norms.

Some matrix notation used in this section follows: $\mathrm{dg}\left\{A_1, ..., A_m\right\}$ is a block-diagonal matrix of with blocks $A_1, ..., A_m$ on the diagonal, $\det\left(A\right)$ denotes the determinant of matrix A, $\mathrm{rk}\left(A\right)$ denotes the rank of matrix A, $\lambda_{\min}\left(A\right)$ and $\lambda_{\max}\left(A\right)$ denote minimal and maximal eigenvalues of symmetric matrix A, $\left[A\right]_{1:m}$ denotes the uper-left $m \times m$ submatrix of square symmetric matrix A, $\left[A\right]_{m_1:m_2,m_3:m_4}$ denotes the submatrix of A extracted from row m_1 to row m_2 and from column m_3 to column m_4. By $\|A\| = \sqrt{\mathrm{tr}\left(A'A\right)}$ we denote the Frobenius norm of matrix A, and by $\|A\|_S = \sqrt{\lambda_{\max}\left(A'A\right)}$ we denote its spectral norm. A $m \times m$ identity matrix is denoted I_m; $\mathrm{O}_{m_1 \times m_2}$ is $m_1 \times m_2$ zero matrix, 1_m is $m \times 1$ vector filled with ones, $\mathrm{e}_{m,j}$ is j^{th} unit vector of dimension m, and 0_m is $m \times 1$ zero vector. Finally, $\mathcal{P}_X$-a.s. means almost surely with respect to the probability measure of X.

2.C.2 Setup and assumptions

Ler us fix an arbitrary finite integer $k \geq 2$ such that all the elements of P of interest have indices not exceeding k. We will be considering the joint asymptotics of elements of the northwest $k \times k$ block of P, i.e. of $P_{1:k}$.

Stack all the diagonal elements of $P_{1:k}$ into the vector

$$\vec{p}_{\mathrm{dg}} = \left(p_{1,1}, p_{2,2}, ..., p_{k,k}\right)',$$

and the upper-off-diagonal elements into the vector

$$\vec{p}_{\mathrm{off}} = \left(p_{1,2}, p_{1,3}, ..., p_{1,k}, p_{2,3}, ..., p_{2,k}, ..., p_{k-1,k}\right)'.$$

We focus on derivation of the asymptotics for the vector

$$\vec{p} = \begin{pmatrix} \vec{p}_{\mathrm{dg}} \\ \vec{p}_{\mathrm{off}} \end{pmatrix}.$$

Assumption C.1.

Let the entries of Z be i.i.d. with zero mean, unit variance, and finite kurtosis $\kappa = E[z_{ij}^4]$.

Note that any fixed non-singular linear transformation $Z \mapsto ZC$, where C is $\ell \times \ell$, leaves P intact. Thus, the results will cover all rotated i.i.d. designs as well, and not only the ideal i.i.d. setup imposed in Assumption C.1. Rotation equivalence also means that some types of dependence are also allowed. In particular, our results hold for "transformed independence" design matrices Z, similar to, e.g., Athey et al. (2018, Assumption 3), which is important in regression setups under random sampling, where the feature matrix Z typically has independent rows but non-independent columns. However, rotation equivalence also means that time series dependence of similar types across rows is also allowed (cf. e.g., Z. Bai and Zhou, 2008).

The isotropy (i.e., equality of variances) imposed in Assumption C.1 is assumed for the sake of convenience, while imposing unit variance is a normalization and does not lose generality either. The existence of the fourth moment is necessary for application of the random matrix theory and the central limit theorem; the parameter κ will explicitly enter the asymptotic distribution of $\vec{p}$.

The next assumption introduces the asymptotic framework.

Assumption C.2.

As $n \to \infty$, we have $\ell \to \infty$ and $\ell/n \to \alpha \in (0,1)$ so that $\ell/n - \alpha = o\left(n^{-1/2}\right)$.

This is a standard assumption of dimension asymptotics, used in the random matrix theory, which is also often exploited in the analysis of regression models with many regressors and instrumental variables models with many instruments.

Among other things, Assumptions C.1 and C.2 guarantee that $Z'Z$ has full rank ℓ with a probability approaching one as $n \to \infty$,[11] and so do matrices of a similar structure, such as $Z'_{-}Z_{-}$ that will appear in the next subsection.

2.C.3 Relation of $\vec{p}$ to leave-k-out analog

The elements of $\vec{p}$ are bilinear or quadratic forms for individual z_i and/or z_j with respect to the matrix $(Z'Z)^{-1}$, which contains all individual observations on z and thus are dependent on z_i and/or z_j. In order to derive their asymptotic properties of elements of $\vec{p}$, one has to break this dependence. This is accomplished by application of the celebrated Woodbury identity for an inverse of rank-k-perturbated matrices, and then working out the formulas for quadratic and bilinear forms with respect to such an inverse. On the basis of these, we figure out the connection between the elements of $\vec{p}$ and the elements of its leave-a-few-out analog, say $\vec{p}^{\,-}$.

To this end, let us split $Z' = \begin{bmatrix} z_1 \; z_2 \; \cdots \; z_k \; Z'_{-} \end{bmatrix}$, where Z'_{-} contains columns of Z' from $(k+1)^{th}$ to n^{th}. Introduce the "leave-k-out quasi-projection" matrix

$$P_{-} = Z\left(Z'_{-}Z_{-}\right)^{-1} Z'$$

with diagonal elements $p^{-}_{s,s}$, $1 \leq s \leq n$, and off-diagonal elements $p^{-}_{s,t}$, $1 \leq s \neq t \leq n$. Similarly to $\vec{p}$, define the vectors

$$\begin{aligned} \vec{p}^{\,-}_{\mathrm{dg}} &= \left(p^{-}_{1,1}, p^{-}_{2,2}, \ldots, p^{-}_{k,k}\right)', \\ \vec{p}^{\,-}_{\mathrm{off}} &= \left(p^{-}_{1,2}, p^{-}_{1,3}, \ldots, p^{-}_{1,k}, p^{-}_{2,3}, \ldots, p^{-}_{2,k}, \ldots, p^{-}_{k-1,k}\right)', \end{aligned}$$

[11] See Anatolyev and Yaskov (2017, Lemma 3.1 implying Condition A).

and, finally,

$$\vec{p}^{\,-} = \begin{pmatrix} \vec{p}^{\,-}_{\text{dg}} \\ \vec{p}^{\,-}_{\text{off}} \end{pmatrix}.$$

The first key result is presented in Lemma C.1, which presents the formulas connecting the elements of the projection matrix P to the elements of the leave-k-out matrix P_{-}, and hence the vector $\vec{p}$ to the vector $\vec{p}^{\,-}$.

Lemma C.1.

The off-diagonal $(s,t)^{th}$ entry of P equals

$$p_{s,t} = \frac{(-1)^{s+t+1}\,\nabla_{s,t}}{\det\left(I_k + [P_{-}]_{1:k}\right)}, \tag{54a}$$

and the diagonal s^{th} entry of P equals

$$p_{s,s} = 1 - \frac{\nabla_{s,s}}{\det\left(I_k + [P_{-}]_{1:k}\right)}, \tag{54b}$$

where $\nabla_{i,j}$ is $(i,j)^{th}$ minor of the matrix $I_k + [P_{-}]_{1:k}$.

The result in Lemma 1 is an extended version of simpler schemes in deriving the probability limit of a diagonal element as in Anatolyev and Yaskov (2017): the relation of the i^{th} diagonal element $p_{i,i}$ to its leave-one-out version $p^{-}_{i,i}$ is

$$p_{i,i} = \frac{p^{-}_{i,i}}{1 + p^{-}_{i,i}},$$

and in deriving the asymptotic distribution of an off-diagonal element of P as in Anatolyev and Smirnov (2024): the relation of the $(i,j)^{th}$ off-diagonal element $p_{i,j}$ to its leave-two-out version $p^{-}_{i,j}$ is

$$p_{i,j} = \frac{p^{-}_{i,j}}{(1 + p^{-}_{i,i})(1 + p^{-}_{j,j}) - (p^{-}_{i,j})^2}.$$

It is easy to see that the latter formula is a special case of (54a) when $k = 2$. The more indices are "left-out," the more (exponentially) complex the relations (54a)–(54b) become.

The next key result is a connection of asymptotics of $\vec{p}$ to the asymptotics of $\vec{p}^{\,-}$. Define

$$\alpha_{-} = \frac{\alpha}{1-\alpha}.$$

Lemma C.2.

Under Assumptions C.1 and C.2, we have

$$\sqrt{n}\begin{pmatrix} \vec{p}_{\text{dg}} - \alpha 1_k \\ \vec{p}_{\text{off}} \end{pmatrix} = (1-\alpha)^2\,\sqrt{n}\begin{pmatrix} \vec{p}^{\,-}_{\text{dg}} - \alpha_{-} 1_k \\ \mp\,\vec{p}^{\,-}_{\text{off}} \end{pmatrix} + o_p\left(1\right),$$

where $\mp$ means a minus or plus sign depending on the position of an element in the vector.

Despite the complex mapping of $\vec{p}^-$ to $\vec{p}$ documented in Lemma 1, their asymptotic behaviors have this pretty simple connection. Lemma 2 helps translate the asymptotics for the elements of the leave-k-out analog P_- to that for the elements of the original projection matrix P.

2.C.4 Asymptotic distribution of $\vec{p}$

The asymptotics of $\vec{p}^-$, which we derive in the proof of Theorem 1 below, and utilization of the relationship between asymptotics of elements of $\vec{p}^-$ and $\vec{p}$ from the previous Section, result in the following asymptotics of $\vec{p}$.

Theorem C.1.

Under Assumptions C.1 and C.2, we have

$$\sqrt{n}\begin{pmatrix}\vec{p}_{\mathrm{dg}}-\alpha 1_k\\ \vec{p}_{\mathrm{off}}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0_{k(k+1)/2}, \alpha\left(1-\alpha\right)\begin{bmatrix}\omega I_k & \mathrm{O}_{k\times k(k-1)/2}\\ \mathrm{O}_{k(k-1)/2\times k} & I_{k(k-1)/2}\end{bmatrix}\right),$$

where

$$\omega = 2+\left(1-\alpha\right)\left(\kappa-3\right).$$

That is, any finite number of elements of the projection matrix are asymptotically Gaussian and asymptotically independent. The asymptotic variances exist in a closed form, and those for the diagonal elements differ from those for the off-diagonal elements by the factor of ω. Of course, $\omega > 0$, and, given α, the minimal value of ω is 2α. Note also that $\omega > 2$ for leptokurtic distributions, $\omega = 2$ for mesokurtic distributions, $\omega < 2$ for platykurtic distributions, and ω may be smaller than unity for very platykurtic ones.

2.C.5 Some implications

The results derived in this section may be helpful when the collective behavior of subsets of elements of the projection matrix is of interest; in particular, a relevant setup may presume a clustered structure (e.g., Chao et al., 2023). In that case, the statistical properties of norms of diagonal $q \times q$ blocks $[P]_{1:q}$ of the projection matrix P, as well as those of off-diagonal $q_1 \times q_2$ blocks $[P]_{1:q_1,q_1+1:q_1+q_2}$ that do not contain any diagonal elements of P, may be of interest.

The asymptotic Gaussianity and especially asymptotic independence allows one to come up with the asymptotics for the Frobenius norms of such submatrices.

Corollary C.1.

Under Assumptions C.1 and C.2, we have:

(i) For the diagonal $q \times q$ block $[P]_{1:q}$ of P, where $q \geq 2$,

$$\sqrt{n}\left(\left\| [P]_{1:q} \right\| - \sqrt{q}\alpha\right) \xrightarrow{d} \mathcal{N}\left(0, \alpha\left(1-\alpha\right)\omega\right).$$

(ii) For the off-diagonal $q_1 \times q_2$ block $[P]_{1:q_1,q_1+1:q_1+q_2}$ of P, where $q_1, q_2 \geq 2$,

$$\sqrt{n}\left\| [P]_{1:q_1,q_1+1:q_1+q_2} \right\| \xrightarrow{d} \sqrt{\alpha\left(1-\alpha\right)}\sqrt{\chi^2_{q_1 q_2}}.$$

Interestingly, given α, while the norm of diagonal blocks is centered at the value that is drifting with the block dimension q, the asymptotic variance does not depend on q. In fact, the diagonal elements asymptotically dominate the whole mass of the off-diagonal elements. The norm of the off-diagonal blocks, in contrast, is centered at zero, and has first order asymptotic mean and variance that are both increasing in both block dimensions q_1 and q_2.

The results of Theorem C.1 also imply the following asymptotics for the spectral norms of such submatrices.

Corollary C.2.

Under Assumptions C.1 and C.2, we have:

(i) For the diagonal $q \times q$ block $[P]_{1:q}$ of P,

$$\sqrt{n}\left(\left\| [P]_{1:q} \right\|_S - \alpha\right) \xrightarrow{d} \sqrt{\alpha\left(1-\alpha\right)\omega} \cdot \varsigma,$$

where the random variable ς has probability density $f_\varsigma\left(u\right) = q\Phi\left(u\right)^{q-1}\phi\left(u\right)$, where $\Phi\left(u\right)$ and $\phi\left(u\right)$ are cumulative distribution and probability density of the standard normal.

(ii) For the off-diagonal $q_1 \times q_2$ block $[P]_{1:q_1,q_1+1:q_1+q_2}$ of P,

$$\sqrt{n}\left\| [P]_{1:q_1,q_1+1:q_1+q_2} \right\|_S \xrightarrow{d} \sqrt{\alpha\left(1-\alpha\right)} \cdot \varkappa,$$

where the non-negative random variable $\varkappa$ has a probability density that is bounded from above as

$$f_\varkappa\left(u\right) \leq \frac{\sqrt{2\pi}}{\Gamma\left(q_1/2\right)\Gamma\left(q_2/2\right)}\left(\frac{u}{\sqrt{2}}\right)^{q_1+q_2-2} \exp\left(-\frac{u^2}{2}\right).$$

Note that the case of $q_1 = q_2 = 1$ is trivially covered by Corollary 2 in the light of Theorem 1, with ς being standard normal and $\varkappa$ being standard half-normal. Given α, while the spectral norm of diagonal blocks is centered at the value independent of the block dimension q, the first order asymptotic mean and variance are very slowly increasing functions of q. With the spectral norm, too, the diagonal elements asymptotically dominate the whole mass of the off-diagonal elements. The spectral norm of the off-diagonal blocks is centered at zero, and has positive first order asymptotic mean squared error that is increasing in the block dimensions q_1 and q_2.

2.C.6 Proofs

Lemma C.A1.

Let A be a $k \times k$ matrix, A_{ij} be its $(i,j)^{th}$ element, and $\nabla_{i,j}$ be its $(i,j)^{th}$ minor. Let $\bar{h} \neq i \leq k$. Then,

$$\sum_{j=1}^{k} (-1)^{i+j} \nabla_{i,j} A_{\bar{h},j} = 0.$$

Proof of Lemma C.A1.

Consider some permutation of k natural numbers $\sigma = (\sigma(1), ..., \sigma(k))$, with the sign of this permutation $\mathrm{sgn}(\sigma)$. For some $i \in 1, ..., k$, fix $\sigma(i) = j$ for some $j \in \{1, 2, ..., k\}$ (it may be that $i = j$). Then, σ can be represented as a composition of two sets of transpositions, denoted as σ_1 and σ_2, say. One, σ_1, is a permutation of k natural numbers such that $\sigma_1(1) < \sigma_1(2) < ... < \sigma_1(i-1) < \sigma_1(i+1) < ... < \sigma_1(k)$ and $\sigma_1(i) = j$. Then, σ_1 can be obtained from the identity permutation by $|j - i|$ transpositions[12]. The other, σ_2, is a permutation of $k-1$ natural numbers such that $\sigma_2(h) = \sigma(h)$ for all $h \neq i$, and $\sigma_2(i) = \sigma_1(i) = \sigma(i) = j$ is fixed. The sign of the composition is $\mathrm{sgn}(\sigma) = \mathrm{sgn}(\sigma_1)\mathrm{sgn}(\sigma_2) = (-1)^{i+j}\mathrm{sgn}(\sigma_2)$. Then, $\mathrm{sgn}(\sigma_2) = \mathrm{sgn}(\sigma)/(-1)^{i+j}$.

By definition of a determinant and using $\mathrm{sgn}(\sigma_2) = \mathrm{sgn}(\sigma)/(-1)^{i+j}$, we have

$$\begin{aligned} \nabla_{i,j} &= \sum_{\substack{\text{all permutations} \\ \sigma_2 = (\sigma_2(1), ..., \sigma_2(k)) \\ \text{s.t. } \sigma_2(i) = j}} A_{1,\sigma_2(1)} ... A_{i-1,\sigma_2(i-1)} A_{i+1,\sigma_2(i+1)} ... A_{k,\sigma_2(k)} \mathrm{sgn}(\sigma_2) \\ &= \sum_{\substack{\text{all permutations} \\ \sigma_2 = (\sigma_2(1), ..., \sigma_2(k)) \\ \text{s.t. } \sigma_2(i) = j}} A_{1,\sigma_2(1)} ... A_{i-1,\sigma_2(i-1)} A_{i+1,\sigma_2(i+1)} ... A_{k,\sigma_2(k)} \frac{\mathrm{sgn}(\sigma)}{(-1)^{i+j}}. \end{aligned}$$

Multiply both sides by $(-1)^{i+j} A_{\bar{h},j}$ for some $\bar{h} \neq i \leq k$ to obtain

$$(-1)^{i+j} \nabla_{i,j} A_{\bar{h},j} = \sum_{\substack{\text{all permutations} \\ \sigma_2 = (\sigma_2(1), ..., \sigma_2(k)) \\ \text{s.t. } \sigma_2(i) = j}} A_{1,\sigma_2(1)} ... A_{i-1,\sigma_2(i-1)} A_{\bar{h},j} A_{i+1,\sigma_2(i+1)} ... A_{k,\sigma_2(k)} \mathrm{sgn}(\sigma).$$

[12]Namely, $(i \;\; i+1)...(j-2 \;\; j-1)(j-1 \;\; j)$ if $j > i$; or $(i-1 \;\; i)...(j+1 \;\; j+2)(j \;\; j+1)$ if $j < i$.

Sum both sides across all j to obtain

$$\begin{aligned}
\sum_{j=1}^{k}(-1)^{i+j}\nabla_{i,j}A_{\bar{h},j} &= \sum_{j=1}^{k}\sum_{\substack{\text{all permutations}\\ \sigma_2=(\sigma_2(1),\ldots,\sigma_2(k))\\ \text{s.t. } \sigma_2(i)=j}} A_{1,\sigma_2(1)}\ldots A_{i-1,\sigma_2(i-1)}A_{\bar{h},j}A_{i+1,\sigma_2(i+1)}\ldots A_{k,\sigma_2(k)}\operatorname{sgn}(\sigma)\\
&= \sum_{\substack{\text{all permutations}\\ \sigma=(\sigma(1),\ldots,\sigma(k))}} A_{1,\sigma(1)}\ldots A_{i-1,\sigma(i-1)}A_{\bar{h},\sigma(i)}A_{i+1,\sigma(i+1)}\ldots A_{k,\sigma(k)}\operatorname{sgn}(\sigma)\\
&= \sum_{\substack{\text{all permutations}\\ \sigma=(\sigma(1),\ldots,\sigma(k))}} A_{\bar{h},\sigma(i)}\operatorname{sgn}(\sigma)\prod_{\substack{h=1\\ \text{s.t. } h\neq i}}^{k} A_{h,\sigma(h)}. \qquad (55)
\end{aligned}$$

For every permutation σ there exists another permutation $\tilde{\sigma}$ such that $\sigma(h)=\tilde{\sigma}(h)$ if $h\notin\{i,\bar{h}\}$, $\sigma(i)=\tilde{\sigma}(\bar{h})$, $\sigma(\bar{h})=\tilde{\sigma}(i)$. Then, $\operatorname{sgn}(\sigma)=-\operatorname{sgn}(\tilde{\sigma})$ and

$$\begin{aligned}
&A_{\bar{h},\sigma(i)}\operatorname{sgn}(\sigma)\prod_{\substack{h=1\\ \text{s.t. } h\neq i}}^{k} A_{h,\sigma(h)} + A_{\bar{h},\tilde{\sigma}(i)}\operatorname{sgn}(\tilde{\sigma})\prod_{\substack{h=1\\ \text{s.t. } h\neq i}}^{k} A_{h,\tilde{\sigma}(h)}\\
&= A_{\bar{h},\sigma(i)}A_{\bar{h},\sigma(\bar{h})}\operatorname{sgn}(\sigma)\prod_{\substack{h=1\\ \text{s.t. } h\notin\{i,\bar{h}\}}}^{k} A_{h,\sigma(h)} - A_{\bar{h},\sigma(\bar{h})}A_{\bar{h},\sigma(i)}\operatorname{sgn}(\sigma)\prod_{\substack{h=1\\ \text{s.t. } h\notin\{i,\bar{h}\}}}^{k} A_{h,\sigma(h)} = 0.
\end{aligned}$$

There is such $\tilde{\sigma}$ for every permutation σ, so all the terms in (55) cancel out, and the conclusion follows. □

Proof of Lemma C.1.

Define

$$H = I_k + [P_-]_{1:k}\,.$$

The Woodbury matrix identity for inversion of an updated matrix (e.g., Henderson and Searle, 1981) implies

$$(Z'Z)^{-1} = \left(Z_-'Z_- + \sum_{i=1}^{k} z_i z_i'\right)^{-1} = \left(Z_-'Z_-\right)^{-1} + \left(Z_-'Z_-\right)^{-1}\Omega\left(Z_-'Z_-\right)^{-1},$$

where

$$\Omega = \frac{1}{\det(H)}\sum_{i,j=1}^{k}(-1)^{i+j+1}\nabla_{i,j}z_i z_j'.$$

Take some $1\leq\bar{h}\neq\bar{f}\leq k$. The bilinear form for vectors $z_{\bar{h}}$ and $z_{\bar{f}}$ with respect to

$(Z'Z)^{-1}$ then reads

$$\begin{aligned}
z_{\bar{h}}' (Z'Z)^{-1} z_{\bar{f}} &= z_{\bar{h}}' \left(\left(Z_-' Z_- \right)^{-1} + \left(Z_-' Z_- \right)^{-1} \Omega \left(Z_-' Z_- \right)^{-1} \right) z_{\bar{f}} \\
&= \frac{1}{\det(H)} \left(z_{\bar{h}}' \left(Z_-' Z_- \right)^{-1} z_{\bar{f}} \det(H) \right. \\
&\qquad \left. + \sum_{i,j=1}^{k} (-1)^{i+j+1} \nabla_{i,j} z_{\bar{h}}' \left(Z_-' Z_- \right)^{-1} z_i z_j' \left(Z_-' Z_- \right)^{-1} z_{\bar{f}} \right) \\
&= \frac{1}{\det(H)} \left(H_{\bar{h},\bar{f}} \det(H) - \sum_{i,j=1}^{k} (-1)^{i+j} \nabla_{i,j} H_{\bar{h},i} H_{\bar{f},j} \right. \\
&\qquad \left. + \sum_{i=1}^{k} (-1)^{i+\bar{h}} \nabla_{i,\bar{h}} H_{\bar{f},i} + \sum_{i=1}^{k} (-1)^{i+\bar{f}} \nabla_{i,\bar{f}} H_{\bar{h},i} - (-1)^{\bar{h}+\bar{f}} \nabla_{\bar{h},\bar{f}} \right)
\end{aligned}$$

From Lemma C.A1,

$$\sum_{i=1}^{k} \nabla_{i,\bar{h}} (-1)^{i+\bar{h}} H_{\bar{f},i} = \sum_{i=1}^{k} \nabla_{i,\bar{f}} (-1)^{i+\bar{f}} H_{\bar{h},i} = 0.$$

Also, from Lemma C.A1 and using the cofactor expansion (see, e.g., Magnus and Neudecker, 1988, p.11),

$$\begin{aligned}
\sum_{i,j=1}^{k} (-1)^{i+j} \nabla_{i,j} H_{\bar{h},i} H_{\bar{f},j} &= \sum_{i=1}^{k} H_{\bar{h},i}^{k} \sum_{j=1}^{k} (-1)^{i+j} \nabla_{i,j} H_{\bar{f},j} \\
&= H_{\bar{h},\bar{f}} \sum_{j=1}^{k} (-1)^{\bar{f}+j} \nabla_{\bar{f},j} H_{\bar{f},j} \\
&= H_{\bar{h},\bar{f}} \det(H).
\end{aligned}$$

Combining these results, we obtain that

$$z_{\bar{h}}' (Z'Z)^{-1} z_{\bar{f}} = \frac{(-1)^{\bar{h}+\bar{f}+1} \nabla_{\bar{h},\bar{f}}}{\det(H)}.$$

Then, the first conclusion follows.

For the second conclusion, take some $1 \leq \bar{h} \leq k$. Similarly,

$$\begin{aligned}
z_{\bar{h}}' (Z'Z)^{-1} z_{\bar{h}} &= z_{\bar{h}}' \left(\left(Z_-' Z_- \right)^{-1} + \left(Z_-' Z_- \right)^{-1} \Omega \left(Z_-' Z_- \right)^{-1} \right) z_{\bar{h}} \\
&= \frac{1}{\det(H)} \left(\left(H_{\bar{h},\bar{h}} - 1 \right) \det(H) - \sum_{i,j=1}^{k} (-1)^{i+j} \nabla_{i,j} H_{\bar{h},i} H_{\bar{h},j} \right. \\
&\qquad \left. + 2 \sum_{i=1}^{k} (-1)^{i+\bar{h}} \nabla_{i,\bar{h}} H_{\bar{h},i} - \nabla_{\bar{h},\bar{h}} \right).
\end{aligned}$$

Similarly, from Lemma C.A1 and using the cofactor expansion, the second term in the parentheses is $-H_{\bar{h},\bar{h}}\det(H)$. However, the third term is non-zero – by the cofactor expansion, it is equal to $2\det(H)$. Combining these facts, we obtain that

$$z_{\bar{h}}'\left(Z'Z\right)^{-1}z_{\bar{h}} = 1 - \frac{\nabla_{\bar{h},\bar{h}}}{\det(H)},$$

and the second conclusion follows. □

Proof of Lemma C.2.

From Anatolyev and Yaskov (2017), $p_{s,s}^{-} \xrightarrow{p} \alpha_{-}$, and hence $1 + p_{s,s}^{-} \xrightarrow{p} 1/(1-\alpha)$ for all s. We will need the order of this convergence. In the course of the proof of Theorem 1 ahead, it will be shown that $E\left[p_{s,s}^{-}|Z_{-}\right] - \alpha_{-} = O(1/\sqrt{n})$ and $var\left(p_{s,s}^{-}|Z_{-}\right) = O(1/n)$.[13] Therefore, $p_{s,s}^{-} = \alpha_{-} + O_p(1/\sqrt{n})$, or $1 + p_{s,s}^{-} = 1/(1-\alpha) + O_p(1/\sqrt{n})$.

From Anatolyev and Smirnov (2024), $p_{s,t}^{-} = O_p(1/\sqrt{n}) \xrightarrow{p} 0$ for all pairs $s \neq t$. From the structure of P_{-} it follows that $I_k + [P_{-}]_{1:k} \xrightarrow{p} \mathrm{dg}\left\{1 + \mathrm{p}\lim p_{s,s}^{-}\right\}_{s=1}^{k} = \mathrm{dg}\left\{(1-\alpha)^{-1}\right\}_{i=1}^{k}$ and so

$$\det\left(I_k + [P_{-}]_{1:k}\right) \xrightarrow{p} (1-\alpha)^{-k}.$$

Similarly,

$$\nabla_{s,s} \xrightarrow{p} (1-\alpha)^{-(k-1)}.$$

To get a more refined statement, we utilize computation of a determinant via the first row (or column) expansion:

$$\begin{aligned}
\det\left(I_k + [P_{-}]_{1:k}\right) &= \det\begin{bmatrix} 1+p_{1,1}^{-} & p_{1,2}^{-} & \cdots & p_{1,k}^{-} \\ p_{1,2}^{-} & 1+p_{2,2}^{-} & \cdots & p_{2,k}^{-} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1,k}^{-} & p_{2,k}^{-} & \cdots & 1+p_{k,k}^{-} \end{bmatrix} \\
&= \prod_{i=1}^{k}\left(1+p_{i,i}^{-}\right) + o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

The order of the remainder here is such because all the non-leading terms in the expansion of the determinant are products of the type $\prod_i\left(1+p_{i,i}^{-}\right)\prod_{s\neq t}p_{s,t}^{-}$, where i runs over no more than $k-2$ indices, and (s,t) runs over at least two pairs $p_{s,t}^{-}$, deeming the whole remainder of order $\prod_i O_p(1)\prod_{s\neq t}O_p(1/\sqrt{n}) = o_p(1/\sqrt{n})$. Likewise,

$$\begin{aligned}
\nabla_{1,1} &= \det\begin{bmatrix} 1+p_{2,2}^{-} & \cdots & p_{2,k}^{-} \\ \vdots & \ddots & \vdots \\ p_{2,k}^{-} & \cdots & 1+p_{k,k}^{-} \end{bmatrix} \\
&= \prod_{i=2}^{k}\left(1+p_{i,i}^{-}\right) + o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

[13] See the derivation for the more complex object p_{γ}^{-}, which would coincide with $p_{s,s}^{-}$ for a particular choice of $\vec{\gamma}_{\mathrm{dg}} = \mathrm{e}_{k,s}$ and $\vec{\gamma}_{\mathrm{off}} = 0_{k(k-1)/2}$.

Then, from (54b), we have

$$\vec{p}_{\mathrm{dg}} - \alpha 1_k = \frac{1}{\det\left(I_k + [P_-]_{1:k}\right)} \left((1-\alpha)\det\left(I_k + [P_-]_{1:k}\right) - \begin{bmatrix} \nabla_{1,1} \\ \vdots \\ \nabla_{k,k} \end{bmatrix} \right).$$

Now, the first element in the numerator is

$$\begin{aligned}
(1-\alpha)\det\left(I_k + [P_-]_{1:k}\right) - \nabla_{1,1} &= \left((1-\alpha)\left(1+p_{1,1}^-\right) - 1\right)\prod_{i=2}^{k}\left(1+p_{i,i}^-\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&= (1-\alpha)\left(p_{1,1}^- - \alpha_-\right)\prod_{i=2}^{k}\left(\frac{1}{1-\alpha} + O_p\left(\frac{1}{\sqrt{n}}\right)\right) \\
&\qquad + o_p\left(\frac{1}{\sqrt{n}}\right) \\
&= (1-\alpha)^{-(k-2)}\left(p_{1,1}^- - \alpha_-\right) + o_p\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

so

$$\begin{aligned}
p_{1,1} - \alpha &= \frac{1}{(1-\alpha)^{-k} + o_p(1)}\left((1-\alpha)^{-(k-2)}\left(p_{1,1}^- - \alpha_-\right) + o_p\left(\frac{1}{\sqrt{n}}\right)\right) \\
&= (1-\alpha)^2\left(p_{1,1}^- - \alpha_-\right) + o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Similarly, one handles all the other $\nabla_{s,s}$ and $p_{s,s} - \alpha$. Stacking the results for all s and using the Slutsky Theorem, we get

$$\vec{p}_{\mathrm{dg}} - \alpha 1_k = (1-\alpha)^2\left(\vec{p}_{\mathrm{dg}}^{\,-} - \alpha_- 1_k\right) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{56}$$

For the off-diagonal elements, from (54a),

$$\vec{p}_{\mathrm{off}} = \begin{bmatrix} p_{1,2} \\ \vdots \\ p_{k-1,k} \end{bmatrix} = \frac{\mp 1}{\det\left(I_k + [P_-]_{1:k}\right)} \begin{bmatrix} \nabla_{1,2} \\ \vdots \\ \nabla_{k-1,k} \end{bmatrix}.$$

Let us look at, say, $\nabla_{1,2}$. Using computation of a determinant via the first row (or column) expansion,

$$\begin{aligned}
\nabla_{1,2} &= \det\begin{bmatrix}
p_{1,2}^- & p_{2,3}^- & \cdots & p_{2,k-1}^- & p_{2,k}^- \\
p_{1,3}^- & 1+p_{3,3}^- & \cdots & p_{3,k-1}^- & p_{3,k}^- \\
p_{1,4}^- & p_{3,4}^- & \ddots & p_{4,k-1}^- & p_{4,k}^- \\
\vdots & \vdots & \cdots & \ddots & \vdots \\
p_{1,k}^- & p_{3,k}^- & \cdots & p_{k-1,k}^- & 1+p_{k,k}^-
\end{bmatrix} \\
&= p_{1,2}^-\prod_{i=3}^{k}\left(1+p_{i,i}^-\right) + o_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

if $k \geq 3$. Again, the order of the remainder is such because all the non-leading terms in the expansion of the determinant are products of the type $\prod_i \left(1+p_{i,i}^{-}\right)\prod_{s\neq t} p_{s,t}^{-}$, where i runs over no more than $k-3$ indices, and (s,t) runs over at least two pairs $p_{s,t}^{-}$, deeming the whole remainder of order $\prod_i O_p(1)\prod_{s\neq t} O_p(1/\sqrt{n}) = o_p(1/\sqrt{n})$. Hence, when $k \geq 3$,

$$\begin{aligned}\nabla_{1,2} &= p_{1,2}^{-}\prod_{i=3}^{k}\left(\frac{1}{1-\alpha}+O_p\left(\frac{1}{\sqrt{n}}\right)\right)+o_p\left(\frac{1}{\sqrt{n}}\right)\\ &= (1-\alpha)^{-(k-2)}\,p_{1,2}^{-}+o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

When $k=2$, simply $\nabla_{1,2}=p_{1,2}^{-}$, which also fits the formula above. Similarly, one handles all the other $\nabla_{s,t}$. Stacking the results for all pairs s,t and using the Slutsky Theorem, we get

$$\vec{p}_{\text{off}} = \mp(1-\alpha)^2\,\vec{p}_{\text{off}}^{\,-}+o_p\left(\frac{1}{\sqrt{n}}\right). \tag{57}$$

Combining (56) and (57), we obtain

$$\sqrt{n}\begin{pmatrix}\vec{p}_{\text{dg}}-\alpha 1_k\\ \vec{p}_{\text{off}}\end{pmatrix} = (1-\alpha)^2\sqrt{n}\begin{pmatrix}\vec{p}_{\text{dg}}^{\,-}-\alpha_{-}1_k\\ \mp\vec{p}_{\text{off}}^{\,-}\end{pmatrix}+o_p(1). \tag{58}$$

$\square$

Proof of Theorem C.1.

Let us define $n_{-}=n-k$. Let us also denote for future use

$$\Xi_n = Z_{-}'Z_{-}/n_{-},$$

which is a sample covariance matrix, and has eigenvalues, say, $\{\lambda_j\}_{j=1,\ldots,\ell}$.

Let $\vec{\gamma}_{\text{dg}}=(\gamma_{1,1},\gamma_{2,2},\ldots,\gamma_{k,k})'\neq 0_k$ and $\vec{\gamma}_{\text{off}}=(\gamma_{1,2},\gamma_{1,3},\ldots,\gamma_{1,k},\gamma_{2,3},\ldots,\gamma_{2,k},\ldots,\gamma_{k-1,k})'\neq 0_{k(k-1)/2}$ be vectors of real numbers conformable to $\vec{p}_{\text{dg}}^{\,-}$ and $\vec{p}_{\text{off}}^{\,-}$, and let $\vec{\gamma}=\left(\vec{\gamma}_{\text{dg}}',\vec{\gamma}_{\text{off}}'\right)'$. Create a random variable

$$p_{\gamma}^{-}=\vec{\gamma}'\vec{p}^{\,-}=\vec{\gamma}_{\text{dg}}'\vec{p}_{\text{dg}}^{\,-}+\vec{\gamma}_{\text{off}}'\vec{p}_{\text{off}}^{\,-}.$$

Notice that

$$p_{\gamma}^{-}=\frac{1}{n_{-}}\vec{z}\,'\Psi_{\gamma}\vec{z},$$

where $\vec{z}=(z_1',z_2',\ldots,z_k')'$, and Ψ_γ is a $k\ell\times k\ell$ matrix that has the structure

$$\Psi_\gamma=\Gamma\otimes\Xi_n^{-1}$$

with

$$\Gamma=\begin{bmatrix}\gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,k-1} & \gamma_{1,k}\\ 0 & \gamma_{2,2} & \cdots & \gamma_{2,k-1} & \gamma_{2,k}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & \gamma_{k-1,k-1} & \gamma_{k-1,k}\\ 0 & 0 & \cdots & 0 & \gamma_{k,k}\end{bmatrix}.$$

The conditional mean of p_γ^-, using that $E\left[\vec{z}\vec{z}'|Z_-\right] = I_{k\ell}$, is

$$E\left[p_\gamma^-|Z_-\right] = \frac{1}{n_-}\mathrm{tr}\left(\Psi_\gamma\right) = \frac{1}{n_-}\mathrm{tr}\left(\Xi_n^{-1}\right)\sum_{i=1}^{k}\gamma_{i,i} \equiv \mu_n$$

The conditional variance of p_γ^-, using that $E\left[z_i z_j'|Z_-\right] = \mathrm{O}_{\ell\times\ell}$ if $1 \leq i \neq j \leq k$, is

$$var\left(p_\gamma^-|Z_-\right) = \frac{1}{n_-^2}\sum_{j=1}^{k}\left(\gamma_{j,j}^2 + \sum_{i=1}^{j}\gamma_{i,j}^2\right)\mathrm{tr}\left(\Xi_n^{-2}\right) + \frac{\kappa-3}{n_-^2}\sum_{i=1}^{k}\gamma_{i,i}^2\sum_{j=1}^{\ell}\left[\Xi_n^{-1}\right]_{jj}^2 \equiv \upsilon_n.$$

The squared spectral norm of Ψ_γ is $\left\|\Psi_\gamma\right\|_S^2 = \lambda_{\max}\left(\Gamma'\Gamma \otimes \Xi_n^{-2}\right) \leq \left\|\Gamma\right\|_S^2 \left(\min_{j=1,\ldots,\ell}\lambda_j\right)^{-2}$. The squared Frobenius norm of Ψ_γ is $\left\|\Psi_\gamma\right\|^2 = \mathrm{tr}\left(\Gamma'\Gamma \otimes \Xi_n^{-2}\right) = \left\|\Gamma\right\|^2 \sum_{j=1}^{\ell}\lambda_j\left(\Xi_n\right)^{-2}$, and note that $\left\|\Gamma\right\| \neq 0$ because $\vec{\gamma} \neq 0_{k(k+1)/2}$. Hence, the spectral and Frobenius norms of Ψ_γ relate as

$$\frac{\left\|\Psi_\gamma\right\|_S^2}{\left\|\Psi_\gamma\right\|^2} \leq \frac{\left\|\Gamma\right\|_S^2}{\left\|\Gamma\right\|^2}\frac{1}{\ell}\frac{\left(\min_{j=1,\ldots,\ell}\lambda_j\right)^{-2}}{\left(\max_{j=1,\ldots,\ell}\lambda_j\right)^{-2}} \overset{a.s.}{\to} 0.$$

due to the almost sure convergence of the extreme eigenvalues of $\min_{j=1,\ldots,\ell}\lambda_j$ and $\max_{j=1,\ldots,\ell}\lambda_j$ to $\underline{\alpha} \equiv (1-\sqrt{\alpha})^2$ and $\bar{\alpha} \equiv (1+\sqrt{\alpha})^2$, respectively (Z. Bai and Silverstein, 2010, Theorem 5.11). By Assumption C.1, $E[z_{ij}^4] < \infty$. Therefore, by the central limit theorem for quadratic forms of Bhansali et al. (2007, Theorem 2.1(ii)), we have that conditional on Z_-,

$$\upsilon_n^{-1/2}\left(p_\gamma^- - \mu_n\right) \overset{d}{\to} \mathcal{N}\left(0,1\right) \tag{59}$$

$\mathcal{P}_{Z_-}$-a.s.

Next, note that

$$\mu_n = \frac{1}{n_-}\sum_{j=1}^{\ell}\lambda_j^{-1}\sum_{i=1}^{k}\gamma_{i,i}.$$

Following Anatolyev and Smirnov (2024, proof of Theorem) and using the Marchenko-Pastur law (Bai and Silverstein, 2010, Theorem 3.6), we obtain

$$\frac{1}{\ell}\sum_{j=1}^{\ell}\lambda_j^{-1} \overset{a.s.}{\to} \int_{\underline{\alpha}}^{\bar{\alpha}}\frac{1}{x}\frac{1}{2\pi x\alpha}\sqrt{(\bar{\alpha}-x)(x-\underline{\alpha})}dx = \frac{1}{1-\alpha}.$$

Also, by the CLT for linear spectral statistics of large covariance matrices (e.g., Z. Bai and Silverstein, 2010, Theorem 9.10), almost sure asymptotic boundedness of the support (see Anatolyev and Smirnov, 2024, proof of Theorem), and the Delta method, we have

$$\frac{1}{\ell}\sum_{j=1}^{\ell}\lambda_j^{-1} - \frac{1}{1-\alpha} = O_p\left(\frac{1}{\ell}\right).$$

Further, $n_- \upsilon_n$ converges in probability to a non-zero constant, as will be shown below. These facts, together with the rate in Assumption C.2, imply that

$$
\begin{aligned}
\upsilon_n^{-1/2}\left(\mu_n - \alpha_- \sum_{i=1}^{k} \gamma_{i,i}\right) &= O_p\!\left(n_-^{1/2}\right)\left(\left(\frac{\ell}{n_-} - \alpha\right)\frac{1}{\ell}\sum_{j=1}^{\ell}\lambda_j^{-1} + \alpha\left(\frac{1}{\ell}\sum_{j=1}^{\ell}\lambda_j^{-1} - \frac{1}{1-\alpha}\right)\right) \\
&\qquad \times \sum_{i=1}^{k}\gamma_{i,i} \\
&= O_p\left(n^{1/2}\right)\left(o\left(\frac{1}{n^{1/2}}\right)O_p(1) + \alpha O_p\left(\frac{1}{\ell}\right)\right)O(1) \\
&= o_p(1).
\end{aligned}
$$

Combining this with (59), we obtain:

$$
\upsilon_n^{-1/2}\left(p_\gamma^- - \alpha_- \sum_{i=1}^{k}\gamma_{i,i}\right) \xrightarrow{d} \mathcal{N}(0,1) \tag{60}
$$

$\mathcal{P}_{Z_-}$-a.s. As the joint limit does not depend on Z_-, it is also an unconditional limit by the dominated convergence theorem.

Now we determine the limiting behavior of υ_n. Recall that

$$
n_-\upsilon_n = \frac{\ell}{n_-}\sum_{j=1}^{k}\left(\gamma_{j,j}^2 + \sum_{i=1}^{j}\gamma_{i,j}^2\right)\cdot V_{1n} + \frac{\ell}{n_-}(\kappa-3)\sum_{i=1}^{k}\gamma_{i,i}^2 \cdot V_{2n},
$$

where

$$
V_{1n} = \frac{\operatorname{tr}\left(\Xi_n^{-2}\right)}{\ell}, \quad V_{2n} = \frac{1}{\ell}\sum_{j=1}^{\ell}\left[\Xi_n^{-1}\right]_{jj}^2.
$$

First we compute the limit of V_{1n}:

$$
V_{1n} = \frac{\operatorname{tr}\left(\Xi_n^{-2}\right)}{\ell} = \frac{1}{\ell}\sum_{j=1}^{\ell}\lambda_j^{-2}.
$$

Following Anatolyev and Smirnov (2024, proof of Theorem) and using the Marchenko-Pastur law (Z. Bai and Silverstein, 2010, Theorem 3.6), we obtain

$$
V_{1n} = \frac{1}{\ell}\sum_{j=1}^{\ell}\lambda_j^{-2} \xrightarrow{a.s.} \int_{\underline{\alpha}}^{\bar{\alpha}} \frac{1}{x^2}\frac{1}{2\pi x\alpha}\sqrt{(\bar{\alpha}-x)(x-\underline{\alpha})}dx = \frac{1}{(1-\alpha)^3}.
$$

Second, we compute the limit of V_{2n}. By the matrix inversion formula for the $(1,1)$ element,

$$
\left[\Xi_n^{-1}\right]_{11} = \frac{(-1)^{1+1}\det\left([\Xi_n]_{2:\ell,2:\ell}\right)}{\det(\Xi_n)},
$$

where the numerator is a cofactor corresponding to the $(1,1)$ element. By the partitioned matrix determinant formula (e.g., Abadir and Magnus, 2005),

$$
\det(\Xi_n) = \frac{1}{n_-}\xi_1' M_{-1}\xi_1 \cdot \det\left([\Xi_n]_{2:\ell,2:\ell}\right),
$$

where $\xi_1 = (z_{1,k+1}, z_{1,k+2}, ..., z_{1,n})'$, the $n_- \times 1$ vector of (all but the first k) observations on the first regressor, is independent of the $n_- \times n_-$ projection matrix of rank $n_\ell \equiv n_- - (\ell - 1)$

$$M_{-1} = I_{n_-} - Z_{-2:\ell} \left(Z'_{-2:\ell} Z_{-2:\ell} \right)^{-1} Z'_{-2:\ell}$$

that contains only observations on the other regressors, collected in the $n_- \times (\ell - 1)$ regressor submatrix $Z_{-2:\ell}$. Then,

$$\left[\Xi_n^{-1} \right]_{11} = \left(\frac{1}{n_-} \xi_1' M_{-1} \xi_1 \right)^{-1}.$$

Now,

$$E \left[\frac{1}{n_-} \xi_1' M_{-1} \xi_1 | Z_{-2:\ell} \right] = \frac{\text{tr} \left(M_{-1} E \left[\xi_1 \xi_1' \right] \right)}{n_-} = \frac{\text{rk} \left(M_{-1} \right)}{n_-} = \frac{n_\ell}{n_-} \to 1 - \alpha$$

$\mathcal{P}_{Z_{-2:\ell}}$-a.s., and

$$\begin{aligned} var \left(\frac{1}{n_-} \xi_1' M_{-1} \xi_1 | Z_{-2:\ell} \right) &= \frac{1}{n_-^2} var \left(\sum_{i=k+1}^{n} \left[M_{-1} \right]_{i-k,i-k} z_{1,i}^2 + \sum_{i,j=k+1, i \neq j}^{n} \left[M_{-1} \right]_{i-k,j-k} z_{1,i} z_{1,j} | Z_{-2:\ell} \right) \\ &= \frac{1}{n_-^2} \sum_{i=k+1}^{n} \left[M_{-1} \right]_{i-k,i-k}^2 (\kappa - 1) + \frac{2}{n_-^2} \sum_{i,j=k+1, i \neq j}^{n} \left[M_{-1} \right]_{i-k,j-k}^2 \\ &= \frac{1}{n_-^2} \sum_{i=k+1}^{n} \left[M_{-1} \right]_{i-k,i-k} \left(2 + \left[M_{-1} \right]_{i-k,i-k} (\kappa - 3) \right) \to 0 \end{aligned}$$

$\mathcal{P}_{Z_{-2:\ell}}$-a.s. by the independence of $z_{1,i}$'s, their independence of $Z_{-2:\ell}$, and the properties of projection matrices. Hence,

$$\left[\Xi_n^{-1} \right]_{11} = \left(\frac{1}{n_-} \xi_1' M_{-1} \xi_1 \right)^{-1} \xrightarrow{p} (1 - \alpha)^{-1}$$

$\mathcal{P}_{Z_{-2:\ell}}$-a.s., and, by the dominated convergence theorem, unconditionally. The same can be derived also for $[\Xi_n^{-1}]_{jj}$ for all $j = 2, ..., \ell$. Hence, $[\Xi_n^{-1}]_{jj}^2 - (1 - \alpha)^{-2} = o_p(1)$ for all $j = 1, ..., \ell$, and thus $\ell^{-1} \sum_{i=1}^{\ell} [\Xi_n^{-1}]_{ii}^2 - (1 - \alpha)^{-2} = \ell^{-1} o_p(\ell) = o_p(1)$. As a result,

$$V_{2n} = \frac{1}{\ell} \sum_{i=1}^{\ell} \left[\Xi_n^{-1} \right]_{ii}^2 \xrightarrow{p} \frac{1}{(1 - \alpha)^2}.$$

Together, convergence of V_{1n} and V_{2n} implies

$$\begin{aligned} n_- \upsilon_n &\xrightarrow{p} \alpha \sum_{j=1}^{k} \left(\gamma_{j,j}^2 + \sum_{i=1}^{j} \gamma_{i,j}^2 \right) \frac{1}{(1 - \alpha)^3} + \alpha (\kappa - 3) \sum_{i=1}^{k} \gamma_{i,i}^2 \frac{1}{(1 - \alpha)^2} \\ &= \frac{\alpha_-}{(1 - \alpha)^2} \left(\sum_{j=1}^{k} \left(\gamma_{j,j}^2 + \sum_{i=1}^{j} \gamma_{i,j}^2 \right) + (\omega - 2) \sum_{i=1}^{k} \gamma_{i,i}^2 \right) = \frac{\alpha_-}{(1 - \alpha)^2} \left(\omega \sum_{i=1}^{k} \gamma_{i,i}^2 + \sum_{j=1}^{k} \sum_{i=1}^{j-1} \gamma_{i,j}^2 \right). \end{aligned}$$

To summarize, we have

$$\sqrt{n} \left(p_\gamma^- - \alpha_- \sum_{i=1}^{k} \gamma_{i,i} \right) \xrightarrow{d} \frac{\alpha_-^{1/2}}{1 - \alpha} \left(\omega \sum_{i=1}^{k} \gamma_{i,i}^2 + \sum_{j=1}^{k} \sum_{i=1}^{j-1} \gamma_{i,j}^2 \right)^{1/2} \mathcal{N}(0, 1). \qquad (61)$$

Finally, combining (61) with the result of Lemma C.2, we obtain

$$\sqrt{n}\left(p_\gamma - \alpha\sum_{i=1}^{k}\gamma_{i,i}\right) \xrightarrow{d} \alpha^{1/2}\left(1-\alpha\right)^{1/2}\left(\omega\sum_{i=1}^{k}\gamma_{i,i}^2 + \sum_{j=1}^{k}\sum_{i=1}^{j-1}\gamma_{i,j}^2\right)^{1/2}\mathcal{N}\left(0,1\right). \tag{62}$$

Now, if we premultiply the stated conclusion of the theorem

$$\sqrt{n}\begin{pmatrix}\vec{p}_{\mathrm{dg}} - \alpha 1_k \\ \vec{p}_{\mathrm{off}}\end{pmatrix} \xrightarrow{d} \alpha^{1/2}\left(1-\alpha\right)^{1/2}\mathcal{N}\left(0_{k(k+1)/2}, \mathrm{dg}\left\{\omega I_k, I_{k(k-1)/2}\right\}\right)$$

by $\vec{\gamma}\,'$, on both sides of it we will obtain exactly both sides of (62). By the Cramér–Wold Theorem, the conclusion follows. $\square$

Proof of Corollary C.1.

From Theorem C.1, setting $q = k$, we have

$$\sqrt{n}\begin{pmatrix}\vec{p}_{\mathrm{dg}} - \alpha 1_q \\ \vec{p}_{\mathrm{off}}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0_{q(q+1)/2}, \alpha\left(1-\alpha\right)\mathrm{dg}\left\{\omega I_q, I_{q(q-1)/2}\right\}\right).$$

Hence, $\left\|\,[P]_{1:q}\right\|^2 = \mathrm{tr}\left([P]_{1:q}^2\right) = \sum_{i=1}^{q} p_{i,i}^2 + \sum_{i=1}^{q}\sum_{j=1+i}^{q} p_{i,j}^2 \xrightarrow{p} q\alpha^2 + q(q-1)\cdot 0 = q\alpha^2$, and

$$\begin{aligned}\sqrt{n}\left(\left\|\,[P]_{1:q}\right\|^2 - q\alpha^2\right) &= \sum_{i=1}^{q}\left(p_{i,i}+\alpha\right)\sqrt{n}\left(p_{i,i}-\alpha\right) + 2\sum_{i=1}^{q}\sum_{j=1+i}^{q} p_{i,j}\sqrt{n}p_{i,j} \\ &\xrightarrow{d} \mathcal{N}\left(0, \sum_{i=1}^{q}\left(2\alpha\right)^2\alpha\left(1-\alpha\right)\omega + \sum_{i=1}^{q}\sum_{j=1+i}^{q} 0^2\alpha\left(1-\alpha\right)\right).\end{aligned}$$

This and the Slutsky Theorem lead to

$$\sqrt{n}\left(\left\|\,[P]_{1:q}\right\| - \sqrt{q}\alpha\right) = \frac{\sqrt{n}\left(\left\|\,[P]_{1:q}\right\|^2 - q\alpha^2\right)}{\left\|\,[P]_{1:q}\right\| + \sqrt{q}\alpha} \xrightarrow{d} \mathcal{N}\left(0, \frac{q\left(2\alpha\right)^2\alpha\left(1-\alpha\right)\omega}{\left(\sqrt{q\alpha^2}+\sqrt{q}\alpha\right)^2}\right),$$

and the first conclusion follows.

Next, from Theorem C.1, considering only the off-diagonal elements in the rectangle $1:q_1, q_1+1:q_1+q_2$, we have

$$\sqrt{n}\begin{bmatrix} p_{1,q_1+1} \\ \vdots \\ p_{q_1,q_1+q_2}\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0_{q(q+1)/2}, \alpha\left(1-\alpha\right)I_{q_1q_2}\right).$$

Hence,

$$\begin{aligned} n\left\|\,[P]_{1:q_1,q_1+1:q_1+q_2}\right\|^2 &= n\cdot\mathrm{tr}\left([P]_{1:q_1,q_1+1:q_1+q_2}'\,[P]_{1:q_1,q_1+1:q_1+q_2}\right) \\ &= \sum_{i=1}^{q_1}\sum_{j=q_1+1}^{q_1+q_2}\left(\sqrt{n}p_{i,j}\right)^2 \xrightarrow{d} \alpha\left(1-\alpha\right)\chi_{q_1q_2}^2,\end{aligned}$$

and the second conclusion follows. $\square$

Proof of Corollary C.2.

The squared spectral norm of $[P]_{1:q}$ is $\left\| [P]_{1:q} \right\|_S^2 = \lambda_{\max}\Big([P]_{1:q}^2 \Big)$. The characteristic equation for the eigenvalues of $[P]_{1:q}^2$ is $\det\Big([P]_{1:q}^2 - \lambda I_q \Big) = 0$, which, after the limit is taken with $\vec{p}_{\mathrm{dg}} = \alpha 1_q + o_p(1)$ and $\vec{p}_{\mathrm{off}} = o_p(1)$, gives

$$\det\left(\alpha^2 I_q - \lambda I_q\right) \xrightarrow{p} 0,$$

yielding that the limiting maximal eigenvalue of $\lambda_{\max}\Big([P]_{1:q}^2 \Big)$ is α^2. Next, using that $\vec{p}_{\mathrm{dg}} = O_p(1)$ and $\vec{p}_{\mathrm{off}} = O_p(1/\sqrt{n})$, the diagonal $(i,i)^{th}$ element of $[P]_{1:q}^2$ is $p_{i,i}^2 + \sum_{j=1, j\neq i}^{q} p_{i,j}^2 = \alpha^2 + O_p(1/n)$, and the off-diagonal $(i,j)^{th}$ element of $[P]_{1:q}^2$ is $\sum_{s=1}^{q} p_{i,s} p_{j,s} = p_{i,j}(p_{i,i} + p_{j,j}) + O_p(1/n) = O_p(1/\sqrt{n})$. The scaled characteristic equation is

$$\sqrt{n}\det\left(\mathrm{dg}\Big\{ [P]_{1:q} \Big\}^2 - \lambda I_q + \mathrm{dg}\left\{O_p(1/n)\right\} + \mathrm{offdg}\left\{O_p\left(1/\sqrt{n}\right)\right\}\right) = 0,$$

where offdg signifies a matrix with zeros on the main diagonal. Computation of the determinant via the first row (or column) expansion yields

$$\sqrt{n}\prod_{i=1}^{q}\left(p_{i,i}^2 - \lambda + O_p(1/n)\right) + o_p(1) = 0.$$

The order of the remainder here is such because all the non-leading terms in the expansion of the determinant are products containing at least two of $p_{s,t}$ with $s \neq t$ (cf. the proof of Lemma 2). Recentering around the limit α^2 and solving, we obtain

$$\sqrt{n}\Big(\lambda_{\max}([P]_{1:q}^2) - \alpha^2\Big) = \max_{i=1,\dots,q} \sqrt{n}\left(p_{i,i}^2 - \alpha^2\right) + o_p(1).$$

This and the Slutsky Theorem lead to

$$\sqrt{n}\Big(\left\| [P]_{1:q} \right\|_S - \alpha\Big) = \frac{\sqrt{n}\Big(\lambda_{\max}([P]_{1:q}^2) - \alpha^2\Big)}{\sqrt{\lambda_{\max}([P]_{1:q}^2)} + \alpha} = \frac{\max_{i=1,\dots,q}\sqrt{n}\left(p_{i,i}^2 - \alpha^2\right)}{\sqrt{\alpha^2} + \alpha} + o_p(1).$$

From Theorem C.1 and the Slutsky Theorem, we have

$$\sqrt{n}\left(p_{i,i}^2 - \alpha^2\right) = (p_{i,i} + \alpha)\sqrt{n}(p_{i,i} - \alpha) \xrightarrow{d} 2\alpha\sqrt{\alpha(1-\alpha)\omega}\,\mathcal{N}(0,1)_i$$

for all $i = 1, \dots, q$ asymptotically independently, so, therefore,

$$\sqrt{n}\Big(\| [P]_{1:q} \|_S - \alpha\Big) \xrightarrow{d} \sqrt{\alpha(1-\alpha)\omega}\max_{i=1,\dots,q}\mathcal{N}(0,1)_i.$$

The density for $\max_{i=1,\dots,q}\mathcal{N}(0,1)_i$ follows from the formula for the density of the maximal order statistics:

$$f_{\max_{i=1,\dots,q}\mathcal{N}(0,1)_i}(u) = q\Phi(u)^{q-1}\phi(u),$$

and the first conclusion follows.

Next, from Theorem C.1, $n\,[P]_{1:q_1,q_1+1:q_1+q_2}\,[P]'_{1:q_1,q_1+1:q_1+q_2} \xrightarrow{d} \alpha\,(1-\alpha)\,\mathcal{W}\,(q_1,q_2)\,,$ where $\mathcal{W}\,(q_1,q_2)$ is the standard Wishart distribution with dimension parameters q_1 and q_2. From Edelman (1988, Lemma 4.2), the probability density of the maximal eigenvalue of $\mathcal{W}\,(q_1,q_2)$-distributed random variable has the bound

$$f_{\lambda_{\max}(\mathcal{W}(q_1,q_2))}\,(u) \leq \frac{\sqrt{\pi}2^{(1-q_1-q_2)/2}}{\Gamma\,(q_1/2)\,\Gamma\,(q_2/2)}u^{(q_1+q_2-3)/2}\exp\left(-\frac{u}{2}\right).$$

Then, by the change of variable formula, the probability density of $\sqrt{\lambda_{\max}\,(\mathcal{W}\,(q_1,q_2))}$ is bounded by

$$f_{\sqrt{\lambda_{\max}(\mathcal{W}(q_1,q_2))}}\,(v) \leq \frac{\sqrt{\pi}2^{(3-q_1-q_2)/2}}{\Gamma\,(q_1/2)\,\Gamma\,(q_2/2)}v^{q_1+q_2-2}\exp\left(-\frac{v^2}{2}\right),$$

and the second conclusion follows. $\square$

References

Abadir, K. M., & Magnus, J. R. (2005). *Matrix algebra* (Vol. 1). Cambridge University Press.

Anatolyev, S., & Smirnov, M. (2024). Off-diagonal elements of projection matrices and dimension asymptotics. *Economics Letters*, *239*, 111761.

Anatolyev, S., & Yaskov, P. (2017). Asymptotics of diagonal elements of projection matrices under many instruments/regressors. *Econometric Theory*, *33*(3), 717–738.

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, *14*(1), 57–67.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444–455.

Angrist, J. D., & Krueger, A. B. (1999). Chapter 23 - empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.). Elsevier. https://doi.org/https://doi.org/10.1016/S1573-4463(99)03004-7

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arellano, M. (2003). *Panel data econometrics*. OUP Oxford.

Ashraf, N., Berry, J., & Shapiro, J. M. (2010). Can higher prices stimulate product use? evidence from a field experiment in zambia. *American Economic Review*, *100*(5), 2383–2413.

Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments* (pp. 73–140, Vol. 1). Elsevier.

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(4), 597–623.

Bai, Y., Santos, A., & Shaikh, A. M. (2022). A two-step method for testing many moment inequalities. *Journal of Business & Economic Statistics*, *40*(3), 1070–1080.

Bai, Z., & Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices* (Vol. 20). Springer.

Bai, Z., & Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 425–442.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, 657–681.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Bhansali, R. J., Giraitis, L., & Kokoszka, P. S. (2007). Convergence of quadratic forms with nonvanishing diagonal. *Statistics & probability letters*, *77*(7), 726–734.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica: Journal of the Econometric Society*, 1443–1458.

Billingsley, P. (1986). *Probability and measure*. John Wiley & Sons.

Björklund, A., & Moffitt, R. (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, *69*(1), 42–49. http://www.jstor.org/stable/1937899

Blomquist, S., & Dahlberg, M. (1999). Small sample properties of liml and jackknife iv estimators: Experiments with weak instruments. *Journal of Applied Econometrics*, *14*(1), 69–88.

Brinch, C. N., Mogstad, M., & Wiswall, M. (2017). Beyond late with a discrete instrument. *Journal of Political Economy*, *125*(4), 985–1039.

Bugni, F., Canay, I., Shaikh, A., & Tabord-Meehan, M. (2024). Inference for cluster randomized experiments with non-ignorable cluster sizes. https://arxiv.org/abs/2204.08356

Carrasco, M. (2012). A regularization approach to the many instruments problem [Thirtieth Anniversary of Generalized Method of Moments]. *Journal of Econometrics*, *170*(2), 383–398. https://doi.org/https://doi.org/10.1016/j.jeconom.2012.05.012

Cattaneo, M. D., Jansson, M., & Newey, W. K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, *113*(523), 1350–1361.

Chao, J. C., & Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, *73*(5), 1673–1692.

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K., & Woutersen, T. (2012). Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory*, *28*(1), 42–86. https://doi.org/10.1017/S0266466611000120

Chao, J. C., Swanson, N. R., & Woutersen, T. (2023). Jackknife estimation of a cluster-sample iv regression model with many weak instruments. *Journal of Econometrics*, *235*(2), 1747–1769. https://doi.org/https://doi.org/10.1016/j.jeconom.2022.12.011

Chen, Y.-C., & Xie, H. (2022). Personalized subsidy rules. *arXiv preprint arXiv:2202.13545*.

Chernozhukov, V., Chetverikov, D., & Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, *45*(4), 2309–2352.

Chernozhukov, V., Chetverikov, D., & Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, *86*(5), 1867–1900.

Chetverikov, D. (2018). Adaptive tests of conditional moment inequalities. *Econometric Theory*, *34*(1), 186–227.

Chetverikov, D. (2019). Testing regression monotonicity in econometric models. *Econometric Theory*, *35*(4), 729–776.

Chyn, E., Frandsen, B., & Leslie, E. (2025). Examiner and judge designs in economics: A practitioner's guide. *Journal of Economic Literature*, *63*(2), 401–439.

Cox, D. (1956). A note on weighted randomization. *The Annals of Mathematical Statistics*, 1144–1151.

Davidson, R., & MacKinnon, J. G. (2006). The case against jive. *Journal of Applied Econometrics*, *21*(6), 827–833.

Deaton, A. S. (2009, January). *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development* (Working Paper No. 14690). National Bureau of Economic Research. https://doi.org/10.3386/w14690

Donald, S. G., & Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, *69*(5), 1161–1191.

Edelman, A. (1988). Eigenvalues and condition numbers of random matrices. *SIAM journal on matrix analysis and applications*, *9*(4), 543–560.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics*, *35*(4), 1351–1377. Retrieved September 30, 2025, from http://www.jstor.org/stable/25464543

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Hansen, B. E., & Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of econometrics*, *210*(2), 268–290.

Hansen, C., Hausman, J., & Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, *26*(4), 398–422.

Hausman, J., Lewis, R., Menzel, K., & Newey, W. (2011). Properties of the cue estimator and a modification with moments. *Journal of Econometrics*, *165*(1), 45–57.

Hausman, J., Newey, W. K., Woutersen, T., Chao, J. C., & Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, *3*(2), 211–255.

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In *Handbook of labor economics* (pp. 1865–2097, Vol. 3). Elsevier.

Heckman, J. J., Schmierer, D., & Urzua, S. (2010). Testing the correlated random coefficient model. *Journal of econometrics*, *158*(2), 177–203.

Heckman, J. J., & Urzua, S. (2009, February). *Comparing iv with structural models: What simple iv can and cannot identify* (Working Paper No. 14706). National Bureau of Economic Research. https://doi.org/10.3386/w14706

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, *88*(3), 389–432. http://www.jstor.org/stable/40043006

Heckman, J. J., & Vytlacil, E. (2001). Policy-relevant treatment effects. *American Economic Review*, *91*(2), 107–111.

Heckman, J. J., & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, *73*(3), 669–738.

Heckman, J. J., & Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(8), 4730–4734. http://www.jstor.org/stable/47634

Henderson, H. V., & Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM review*, *23*(1), 53–60.

Horowitz, J. L., & Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, *69*(3), 599–631.

Imbens, G. W. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature*, *48*(2), 399–423. https://doi.org/10.1257/jel.48.2.399

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475. http://www.jstor.org/stable/2951620

Kolesár, M. (2013). *Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity* (Working Papers No. 2013-2). Princeton University. Economics Department. https://ideas.repec.org/p/pri/econom/2013-2.html

Li, R., & Nie, L. (2008). Efficient statistical inference procedures for partially nonlinear models and their applications. *Biometrics*, *64*(3), 904–911.

Ligtenberg, J. W. (2025). Inference in clustered iv models with many and weak instruments. https://arxiv.org/abs/2306.08559

Ligtenberg, J. W., & Woutersen, T. (2024). Multidimensional clustering in judge designs. *arXiv preprint arXiv:2406.09473*.

Liu, Y. (2022). Policy learning under endogeneity using instrumental variables. *arXiv preprint arXiv:2206.09883*.

MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, *232*(2), 272–299.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. Wiley.

Mikusheva, A., & Sølvsten, M. (2025). Linear regression with weak exogeneity. *Quantitative Economics*, *16*(2), 367–403.

Mogstad, M., Santos, A., & Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, *86*(5), 1589–1619.

Mogstad, M., & Torgovitsky, A. (2024). Chapter 1 - instrumental variables with unobserved heterogeneity in treatment effects. In C. Dustmann & T. Lemieux (Eds.). Elsevier. https://doi.org/https://doi.org/10.1016/bs.heslab.2024.11.003

Mogstad, M., Torgovitsky, A., & Walters, C. R. (2024). Policy evaluation with multiple instrumental variables. *Journal of Econometrics*, *243*(1-2), 105718.

Sasaki, Y., & Ura, T. (2024). Welfare analysis via marginal treatment effects. *Econometric Theory*, 1–24.

Shea, J., & Torgovitsky, A. (2021). Ivmte: An r package for implementing marginal treatment effect methods. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2020-01).

Vaart, A. W. v. d. (1998). *Asymptotic statistics*. Cambridge University Press.

van Hasselt, M. (2010). Many instruments asymptotic approximations under nonnormal error distributions. *Econometric Theory*, *26*(2), 633–645.

Voronin, A. (2025). Linear programming approach to partially identified econometric models. https://arxiv.org/abs/2503.14940

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, *70*(1), 331–341.

Walters, C. R. (2018). The demand for effective charter schools. *Journal of Political Economy*, *126*(6), 2179–2223.

Yitzhaki, S. (1996). On using linear regressions in welfare economics. *Journal of Business & Economic Statistics*, *14*(4), 478–486.