Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects[*]

Kirill S. Evdokimov[†]
Princeton University

Michal Kolesár[‡]
Princeton University

January 25, 2018

Please click here for the latest version

Abstract

We study inference in an instrumental variables model with heterogeneous treatment effects and possibly many instruments and/or covariates. In this case two-step estimators such as the two-stage least squares (TSLS) or versions of the jackknife instrumental variables (JIV) estimator estimate a particular weighted average of the local average treatment effects. The weights in these estimands depend on the first-stage coefficients, and either the sample or population variability of the covariates and instruments, depending on whether they are treated as fixed (conditioned upon) or random. We give new asymptotic variance formulas for the TSLS and JIV estimators, and propose consistent estimators of these variances. The heterogeneity of the treatment effects generally increases the asymptotic variance. Moreover, when the treatment effects are heterogeneous, the conditional asymptotic variance is smaller than the unconditional one. Our results are also useful when the treatment effects are constant, because they provide the asymptotic distribution and valid standard errors for the estimators that are robust to the presence of many covariates.

Keywords: heterogeneous treatment effects, LATE, instrumental variables, jackknife, high-dimensional data.

[*] We thank participants at various conferences and seminars for helpful comments and suggestions. Evdokimov gratefully acknowledges financial support by the NSF. All errors are our own.

[†] Email: kevdokim@princeton.edu

[‡] Email: mkolesar@princeton.edu

1 Introduction

Empirical researchers are typically careful to interpret instrumental variables (IV) regressions as estimating a weighted average of local average treatment effects (LATEs), i.e., treatment effects specific to the individuals whose treatment status is affected by the instrument, see Imbens and Angrist (1994) and Heckman and Vytlacil (1999). When it comes to inference, however, they revert to standard errors that assume homogeneity of treatment effects, which are in general invalid in the LATE framework: they are generally too small relative to the actual sampling variability of the estimator. Oftentimes, inference is further complicated by the fact that the number of instruments is relatively large, and that one may also need to include a large number of control variables in order to ensure the instruments' validity.

This paper considers the problem of inference in the LATE framework, with a particular focus on the case of many instruments and/or covariates. We make three main contributions.

First, the paper points out the difference between conditional (on the realizations of instruments and covariates) and unconditional inference in the LATE framework. When the treatment effects are homogeneous, the two approaches to inference are indistinguishable from the practical point of view: they suggest identical formulas for the standard errors. The paper shows that this is no longer the case when the treatment effects are heterogeneous. The standard errors for the conditional inference are smaller than for the unconditional one. The reason is that the unconditional inference additionally accounts for the sampling variability of the conditional estimand.

Second, the paper studies the conditional and unconditional estimands of the TSLS and jackknife IV estimators. In particular, the paper investigates when can the conditional and unconditional estimands be guaranteed to be convex combinations of individual LATEs. One interesting result is that in the presence of many covariates, the unconditional estimand of the TSLS generally differs from the estimand given by Imbens and Angrist (1994), although the estimand is similar and generally is still a convex combination of individual LATEs.

Third, the paper derives the asymptotic distribution and provides valid standard errors for these estimators. Our large sample theory allows for both the number of instruments and the number of covariates to increase with the sample size, while also allowing for the heterogeneity of the treatment effects. Thus, for example, the paper provides what appears to be the first valid inference approach in the settings such as Aizer and Doyle (2015) or Angrist and Krueger (1991) with many instruments.[1]

As a by-product, the paper provides new asymptotic theory results that can be useful for analysing estimators and inference procedures in the presence of high-dimensional observables and possible treatment effect heterogeneity.

When the treatment effects are homogeneous, the IV model is defined by a moment condition that equals zero when the parameter β on the endogenous variable corresponds to the average treatment effect. On the other hand, when the treatment effects are heterogeneous, so that different pairs of in-

[1] The only previously available results on the distribution of the TSLS-like estimators in the LATE framework were obtained for the unconditional inference on the TSLS estimator with a finite number of instruments and covariates, see Imbens and Angrist (1994).

struments identify different LATEs, the IV model is misspecified in that there exists no single parameter that satisfies the moment condition. Valid inference in this case therefore requires a proper definition of the estimand of interest.

We define the unconditional estimand as the appropriate probability limit of the estimator. When the number of instruments and covariates is finite, Imbens and Angrist (1994) show that the unconditional TSLS estimand is given by a weighted average of LATEs, with the weights reflecting the strength (variability) of the instrument pair that defines the LATE. Kolesár (2013) shows that the unconditional estimand for other two-step estimators such as several versions of the jackknife IV estimator is the same, but that the unconditional estimands of minimum distance estimators such as the limited information maximum likelihood estimator is different, and cannot in general be guaranteed to lie inside the convex hull of the LATEs.

Another possibility is to define the estimand as the quantity obtained if the reduced-form errors were set to zero, which we refer to as the conditional estimand since the estimand conditions on the realized values of the instruments and covariates. We show that the conditional estimand of TSLS and jackknife estimators can also be interpreted as a weighted average of LATEs, but the weights now depend on the sample, rather than population variability of the instruments. As a result, it is difficult to guarantee that the weights are positive in a given sample, and so far we have only been able to guarantee that the weights are positive in finite samples when the covariates comprise only indicator variables. When the design is balanced in a certain precise sense and the number of instruments is of a smaller order than the sample size, the weights can be shown to be positive with probability approaching one for a wide range of settings.

As a consequence of the distinction between the conditional β_{C} and unconditional β_{U} estimands for an estimator $\hat{\beta}$, we show that the conditional (on the instruments and covariates) asymptotic variance of $\hat{\beta}$ is smaller than its unconditional asymptotic variance. The unconditional asymptotic variance is larger because it needs to take into account the variability of the conditional estimand due to the variability in the weights of the LATEs. More precisely, the unconditional asymptotic variance (the asymptotic variance of $\hat{\beta} - \beta_{\mathrm{U}}$) is given by the sum of the conditional asymptotic variance (the asymptotic variance of $\hat{\beta} - \beta_{\mathrm{C}}$), and the asymptotic variance of the conditional estimand (asymptotic variance of $\beta_{\mathrm{C}} - \beta_{\mathrm{U}}$). When the treatment effects are homogeneous, all LATEs are the same, and the variability of the weights due to the sampling variation in the instruments and covariates does not enter the asymptotic distribution. In this case, the two estimands coincide and the (unconditional) variance of β_{C} is zero. Otherwise, however, the distinction matters. That the distinction between the conditional and unconditional estimands can lead to the conditional asymptotic variance being lower than the unconditional one has been previously noted by Abadie et al. (2014) in the context of misspecified linear regression. It is worth noting that in our problem both the conditional and unconditional estimands can be of interest for causal inference.

We show that the conditional asymptotic variance can be decomposed into a sum of three terms. The first term corresponds to the usual heteroskedasticity-robust asymptotic variance expression under homogeneous treatment effects and standard asymptotics found in econometrics textbooks. The second term accounts for the variability of the treatment effect between individuals and equals zero when the treatment effects are homogeneous. It is in general positive, so that the standard errors are generally larger when the treatment effects are heterogeneous. The third term accounts for the presence of many instruments, and disappears when the number of instruments K grows more slowly than the strength of the instruments as measured by $\ddot{r}_n$, a version of the concentration parameter defined below.

The literature on inference for the two-step estimators in the presence of heterogeneous treatment effects is limited. Imbens and Angrist (1994) derive the (unconditional) asymptotic distribution of the TSLS estimator with finite number of instruments and covariates.[2] Formally, the problem can also be seen as inference in a misspecified IV/GMM model. The first to provide standard errors in this model were Maasoumi and Phillips (1982) for homoskedastic errors, and Hall and Inoue (2003) for general GMM estimators, see also Lee (2017). Carneiro et al. (2011) consider inference on the marginal treatment and policy effect estimators, but these are substantively and statistically different estimators from the two-step estimators considered in this paper. Kitagawa (2015) and Evdokimov and Lee (2013) develop tests of instrument validity when the treatment effects can be heterogeneous.

Our asymptotic analysis builds on the many instruments and many weak instruments literature due to Kunitomo (1980), Morimune (1983), Bekker (1994) and Chao and Swanson (2005). Our distributional results, in particular, build on those in Newey and Windmeijer (2009) and Chao et al. (2012). This literature is focused on the case in which the treatment effects are homogeneous, so that the IV moment condition holds, the number of covariates L is fixed, but the number of instruments K may grow with the sample size n. In contrast, we allow for the treatment effects to be heterogeneous, and the number of covariates to grow with the sample size. This is important in practice, since, as we argue in Section 2 below, in many empirical settings in which the number of instruments is large, the number of covariates is typically also large. Although an increasing number of covariates L has been previously considered in Anatolyev (2013) and Kolesár et al. (2015), these papers assume that the reduced-form errors are homoskedastic, which is unlikely when the treatment effects are heterogeneous, and impossible when the treatment is binary.

Consistency of the estimators of the asymptotic variance proposed in the many instruments literature typically relies on the fact that the number of parameters in the IV moment condition is fixed, and that, under homogeneous treatment effects, they can be estimated at the same rate as the rate of convergence of $\hat{\beta}$. This allows one to estimate the error in the IV moment condition, usually referred to as the "structural error" ϵ_i at a fast enough rate so that replacing the estimated structural error in the asymptotic variance formula with ϵ_i does not matter in large samples. When the treatment effects are heterogeneous and/or the number of covariates L grows with the sample size, this is no longer the case, and naïve plug-in estimators of the asymptotic variance are asymptotically biased upward. The feasible standard error formulas that we propose jackknifes the naïve plug-in estimator to remove this bias.

[2] Note that the heteroskedasticity-robust standard errors cannot account for the heterogeneity of treatment effects.

Although our asymptotic theory applies to a large class of two-step estimators, we focus on several specific estimators. In particular, we consider a version of the jackknife estimator proposed in Ackerberg and Devereux (2009), called ijive1, as well as a related estimator ijive2, which differs from ijive1 in that it does not rescale the first-stage predictor after removing the influence of own observation. This difference is similar to the difference between the jive1 estimator studied in Phillips and Hale (1977), Blomquist and Dahlberg (1999), and Angrist et al. (1999), and the jive2 estimator of Angrist et al. (1999). Ackerberg and Devereux (2009) have shown, however, using bias expansions similar to those in Nagar (1959), that these two estimators are biased when the number of covariates is large, just as the tsls estimator is biased when the number of instruments is large. See also Davidson and MacKinnon (2006) and other papers in the same issue. We also consider ujive estimator introduced in Kolesár (2013). Our consistency theorems show that a similar conclusion obtains under the many instrument asymptotic sequence that we consider. No inference procedures were previously available for the estimators robust to the presence of many covariates, and our paper fills this gap.

A potential criticism of our results is that, since the definition of the estimands depends on the estimator, the particular weighting of the local average treatment effects that it implies may not be policy-relevant. In the settings with a fixed number of strong instruments, a viable option is to report the lates separately and leave it up to the reader to choose their preferred weighting. Alternatively, one can use the marginal treatment effects framework of Heckman and Vytlacil (1999, 2005) to derive weights that are more policy-relevant, and build a confidence interval for an estimand that uses such weighting. However, Evdokimov and Lee (2013) point out that the valid confidence intervals for the weighted averages of lates with weights that do not shrink to zero for irrelevant instruments (e.g., equal- or census-weighted average of lates, smallest or largest late), generally are trivial $(-\infty, \infty)$, unless some additional restrictions (e.g., bounds on the support of the outcome variable) are introduced into the model. The presence of a single unidentified late implies that such weighted average is also unidentified. In practice, in many empirical settings, such as in Section 2 below, in which the instruments correspond to group indicators, the identification strength or group size at least for some instruments may be too small to accurately estimate every individual late. Then, any weighting scheme that ex ante puts a positive weight on a particular late will lead to uninformative inference if that particular late turns out to be very imprecisely estimated. Therefore, in such cases, one may have to choose a less ambitious goal of providing a confidence interval for some weighted average of lates that puts small (zero) weight on the lates corresponding to the weak (irrelevant) instruments, such as the tsls and jiv estimators. Importantly, our asymptotic theory results consider a broad class of estimators, and can be used to derive the asymptotic properties of other estimators besides those explicitly considered in the paper.

Whether or not a data-driven weighting of the lates is policy relevant, the tsls and jiv estimators are routinely reported in empirical studies. It is important to accompany such estimates with an accurate measure of their variability, which our standard errors provide.

The remainder of this paper is organized as follows. Section 2 motivates and explains our analysis and results in the empirically important simple special case in which the instruments and covariates correspond to group indicators. Section 3 sets up the general model and notation. Section 4 discusses the causal interpretation of the conditional and unconditional estimands. Section 5 presents our large-sample theory. Proofs are relegated to the Appendix.

2 Example: dummies as instruments

This section illustrates the main issues in a simplified setup. We are interested in the effect of a binary treatment variable X_i on an outcome Y_i, where $i = 1, \dots, n$ indexes individuals. The vector of exogenous covariates W_i has dimension L, and consists of group dummies: $W_{i,\ell} = \mathbb{I}\{G_i = \ell\}$ is the indicator that individual i belongs to group ℓ, where $G_i \in \{1, \dots, L\}$ denotes the group that the individual belongs to. For each individual, we have available an instrument S_i that takes on $M+1$ possible values in each group. We label the possible values in group ℓ by $s_{\ell 0}, \dots, s_{\ell M}$. The vector of instruments Z_i has dimension $K = ML$ and consists of indicators for the possible values, $Z_{i,\ell m} = \mathbb{I}\{S_i = s_{\ell m}\}$, with the indicator for the value $s_{\ell 0}$ in each group omitted: $Z_i = (Z_{i,11}, \dots, Z_{i,1M}, Z_{i,21}, \dots, Z_{i,LM})$.

This setup arises in many empirical applications. For example, in the returns to schooling study of Angrist and Krueger (1991), G_i corresponds to state of birth and the instruments are interactions between quarter of birth and state of birth, so that $S_i = s_{\ell m}$ if an individual is born in state ℓ and quarter $m-1$. Aizer and Doyle (2015), who study the effects of juvenile incarceration on adult recidivism, use the fact that conditional on a juvenile's neighborhood G_i, the judge assigned to their case is effectively random: here $S_i = s_{\ell m}$ if an individual is from neighborhood ℓ and is assigned the mth judge out of $M+1$ possible judges overseeing that neighborhood's cases (for simplicity, in this example we assume that number of judges is the same in each neighborhood). Similarly, Dobbie and Song (2015) use random assignment of bankruptcy filings to judges within each bankruptcy office to study the effect of Chapter 13 bankruptcy protection on subsequent outcomes. Silver (2016), who is interested in the effects of a physician's work pace on patient outcomes, uses the fact that by virtue of quasi-random assignment to work shifts, conditional on physician fixed effects G_i, a physician's peer group S_i is effectively randomly assigned.

The first-stage regression is given by

$$X_i = \sum_{\ell=1}^{L} \sum_{m=1}^{M} Z_{i,\ell m} \pi_{\ell m} + \sum_{\ell=1}^{L} W_{i,\ell} \psi_\ell + \eta_i, \tag{1}$$

where, by definition of regression, $E[\eta_i \mid G_i, S_i] = 0$. The reduced-form outcome equation is given by

$$Y_i = \sum_{\ell=1}^{L} \sum_{m=1}^{M} Z_{i,\ell m} \pi_{Y,\ell m} + \sum_{\ell=1}^{L} \mathbb{I}\{G_i = \ell\} \psi_{Y,\ell} + \zeta_i, \tag{2}$$

where, again by definition of regression, $E[\zeta_i \mid G_i, S_i] = 0$.

We assume that within each group, S_i is as good as randomly assigned and only affects the outcome through their effect on the treatment. We also assume that the instrument has a monotone effect on the treatment, so that $P(X_i(s_{\ell m}) > X_i(s_{\ell m'}) \mid G_i = \ell)$ equals either zero or one for all pairs m, m' and all ℓ, where $X_i(s)$ denotes the potential treatment when an individual is assigned $S_i = s$. This assumption implies that $\pi_{\ell m}$ corresponds to the fraction of "compliers", the subset of individuals in group ℓ who change their treatment status when their instrument changes from $s_{\ell 0}$ to $s_{\ell m}$. As shown in Imbens and Angrist (1994) these assumptions further imply that the ratio $\pi_{Y,\ell m}/\pi_{\ell m}$ can the interpreted as an average treatment effect for this subset of the population, $\beta_{\ell m0}$, also called a local average treatment effect (LATE), defined as

$$\beta_{\ell mm'} := E[Y_i(1) - Y_i(0) \mid X_i(s_{\ell m}) \neq X_i(s_{\ell m'}), G_i = \ell].$$

Here $Y_i(x)$ denotes the potential outcome corresponding to treatment status x. If individuals do not select into treatment based on expected gains from treatment, then all LATEs are the same and equal the average treatment effect, $\beta_{\ell mm'} = \text{ATE} := E[Y_i(1) - Y_i(0)]$ for all ℓ and all pairs m, m', and the regressions (1)–(2) reduce to the standard IV model, which assumes that $\pi_{Y,\ell m} = \text{ATE} \cdot \pi_{\ell m}$. Our goal, however, is to explicitly allow for the possibility that the LATEs may vary.

The two-stage least squares estimator can be obtained by first "projecting out" the effect of the exogenous regressors W_i by constructing the residuals $\ddot{Y}_i$, $\ddot{X}_i$ and $\ddot{Z}_i$ from the regression of Y_i, X_i and Z_i on W_i. One then constructs a single instrument $\widehat{R}_{\text{TSLS},i}$ as the predictor from the first stage regression of $\ddot{X}_i$ on $\ddot{Z}_i$. The two-stage least squares estimator $\hat{\beta}_{\text{TSLS}}$ is obtained as the IV estimator in the regression of $\ddot{Y}_i$ on $\ddot{X}_i$ that uses $\widehat{R}_{\text{TSLS},i}$ as a single instrument:

$$\hat{\beta}_{\text{TSLS}} = \frac{\sum_{i=1}^{n} \widehat{R}_{\text{TSLS},i} \ddot{Y}_i}{\sum_{i=1}^{n} \widehat{R}_{\text{TSLS},i} \ddot{X}_i}. \tag{3}$$

Because the exogenous covariates are group dummies, projecting out their effect is equivalent to subtracting group means from each variable: $\ddot{Y}_i = Y_i - n_{G_i}^{-1} \sum_{j \colon G_j = G_i} Y_i$, where n_{G_i} is the number of individuals in group G_i, and similarly for $\ddot{X}_i$ and $\ddot{Z}_i$. The predicted value $\widehat{R}_{\text{TSLS},i}$ is then given by the difference between the sample mean of $\ddot{X}_i$ for individuals in group G_i with instrument value equal to S_i, and the overall sample mean of $\ddot{X}_i$ in group G_i:

$$\widehat{R}_{\text{TSLS},i} = \frac{1}{n_{S_i}} \sum_{j \colon S_j = S_i} \ddot{X}_j - \frac{1}{n_{G_i}} \sum_{j \colon G_j = G_i} \ddot{X}_j = \frac{1}{n_{S_i}} \sum_{j \colon S_j = S_i} X_j - \frac{1}{n_{G_i}} \sum_{j \colon G_j = G_i} X_j, \tag{4}$$

where n_{S_i} is the number of individuals with instrument value equal to S_i.

One can think of the first-stage predictor as estimating the signal $\ddot{R}_i = \sum_{m=1}^{M} (Z_{i,G_i m} - n_{G_i m}/n_{G_i}) \pi_{G_i m}$: we have $\widehat{R}_{\text{TSLS},i} = \ddot{R}_i$ if the first-stage errors η_i are identically zero. The strength of the signal measures how fast the variance of $\hat{\beta}_{\text{TSLS}}$ shrinks with the sample size: we show in Section 5 below that it is of the order $1/\ddot{r}_n$, where $\ddot{r}_n = \sum_{i=1}^{n} \ddot{R}_i^2$ is a version of the concentration parameter. When the instruments are strong, $\ddot{r}_n$ grows as fast as the sample size n, but it may grow more slowly if the instruments are weaker. We require that $\ddot{r}_n \to \infty$ as $n \to \infty$, ruling out the Staiger and Stock (1997) weak instrument asymptotics under which $\ddot{r}_n$ is bounded.

We now consider the estimands. Assume, without loss of generality, that the instruments are ordered so that changing the instrument from $s_{\ell m}$ to $s_{\ell,m+1}$ (weakly) increases the treatment probability. Then $\pi_{\ell m} \geq \pi_{\ell,m-1}$ for all m and ℓ, where we define $\pi_{\ell 0} := 0$. If the reduced-form errors η_i and ζ_i were zero, then it follows from Lemma 4.1 below that the TSLS estimator would equal a weighted average of LATEs,

$$\beta_{\mathrm{C}} = \sum_{\ell=1}^{L} \sum_{m=1}^{M} \frac{\hat{\omega}_{\ell m}}{\sum_{\ell'=1}^{L} \sum_{m'=1}^{M} \hat{\omega}_{\ell' m'}} \beta_{\ell m, m-1}, \tag{5}$$

where the weights $\hat{\omega}_{\ell m}$ are all positive and given by

$$\hat{\omega}_{\ell m} = \frac{n_\ell}{n}(\pi_{\ell m} - \pi_{\ell,m-1}) \sum_{k=m}^{M} \frac{n_{\ell k}}{n_\ell} \left(\pi_{\ell k} - \sum_{m'=1}^{M} \frac{n_{\ell m'}}{n_\ell} \pi_{\ell m'} \right).$$

We call β_{C} conditional estimand, because it conditions on the realizations of the instruments and covariates (we keep this dependence implicit in the notation). Furthermore, under standard asymptotics which hold K, L, and the coefficients π and π_Y fixed as $n \to \infty$, $\hat{\beta}_{\text{TSLS}}$ converges to a weighted average of LATEs

$$\beta_{\mathrm{U}} = \sum_{\ell=1}^{L} \sum_{m=1}^{M} \frac{\omega_{\ell m}}{\sum_{\ell'=1}^{L} \sum_{m'=1}^{M} \omega_{\ell' m'}} \beta_{\ell m, m-1},$$

where the weights $\omega_{\ell m}$ replace the sample fractions n_ℓ / n and $n_{\ell m}/n_\ell$ in (5) with population probabilities $p_\ell = P(G_i = \ell)$ and $p_{s_{\ell m}} = P(S_i = s_{\ell m} \mid G_i = \ell)$, so $\omega_{\ell s} = p_\ell(\pi_{\ell m} - \pi_{\ell,m-1}) \sum_{k=m}^{M} p_{s_{\ell k}}(\pi_{\ell k} - \sum_{m'=1}^{M} p_{s_{\ell m'}} \pi_{\ell m'})$. We refer to β_{U} as the unconditional estimand. If the LATEs are all equal to the ATE, the weighting does not matter, and both β_{U} and β_{C} collapse to the ATE. Furthermore, the usual standard error formula can be used to construct asymptotically valid confidence intervals (CIs). Otherwise, however, $\beta_{\mathrm{U}} \neq \beta_{\mathrm{C}}$, and one has to choose whether one wants to report CIs for β_{C} or CIs for β_{U}. It follows from our results in Section 5 below that the usual standard error formula does not deliver valid CIs for either estimand, and that the CIs for the unconditional estimand will always be wider: the asymptotic variance of $\hat{\beta}_{\text{TSLS}} - \beta_{\mathrm{U}}$ can be written as the sum of the sampling variance of $\hat{\beta}_{\text{TSLS}} - \beta_{\mathrm{C}}$ and the variance of the conditional estimand, $\beta_{\mathrm{C}} - \beta_{\mathrm{U}}$.

A further problem complicating inference is that, as has been documented in the many instruments literature, the TSLS estimator is biased (even when the treatment effects are homogeneous), with the bias increasing with the number of instruments K. To see this, note that under regularity conditions, we can approximate $\hat{\beta}_{\text{TSLS}}$ by taking expectation of the numerator and denominator conditional on all instruments $Z = (Z_1, \ldots, Z_n)'$ and all covariates $W = (W_1, \ldots, W_n)'$:

$$\hat{\beta}_{\text{TSLS}} = \frac{\sum_{i=1}^{n} E[\widehat{R}_{\text{TSLS},i} \ddot{Y}_i \mid Z, W]}{\sum_{i=1}^{n} E[\widehat{R}_{\text{TSLS},i} \ddot{X}_i \mid Z, W]} + o_P(1).$$

To evaluate this expression, decompose the first-stage predictor into a signal and a noise component: $\widehat{R}_{\textsc{tsls},i} = \ddot{R}_i + (\frac{1}{n_{S_i}} \sum_{j\colon S_j=S_i} \eta_j - \frac{1}{n_{G_i}} \sum_{j\colon G_j=G_i} \eta_j)$. Using the identities $\sum_{i=1}^n E[X_i \ddot{R}_i \mid Z, W] = \ddot{r}_n$, $\beta_{\mathrm{C}} = \sum_{i=1}^n E[\ddot{R}_i Y_i \mid Z, W]/\ddot{r}_n$, and $\sum_{i=1}^n \widehat{R}_{\textsc{tsls},i}\ddot{Y}_i = \sum_{i=1}^n \widehat{R}_{\textsc{tsls},i} Y_i$, it follows that

$$\hat{\beta}_{\textsc{tsls}} = \beta_{\mathrm{C}} + \frac{\sum_{\ell=1}^{L}\sum_{m=0}^{M} \sigma_{\eta\nu,\ell m}(1 - n_{\ell m}/n_{\ell})}{\ddot{r}_n + \sum_{\ell=1}^{L}\sum_{m=0}^{M} \sigma^2_{\eta,\ell m}(1 - n_{\ell m}/n_{\ell})} + o_P(1), \tag{6}$$

where $\sigma_{\eta\nu,\ell m} = E[\eta_i(\zeta_i - \eta_i\beta_{\mathrm{C}}) \mid S_i = s_{\ell m}, G_i = \ell]$ measures the conditional covariance between η_i and $\nu_i = \zeta_i - \eta_i\beta_{\mathrm{C}}$, and $\sigma^2_{\eta,\ell m} = E[\eta_i^2 \mid S_i = s_{\ell m}, G_i = \ell]$. The second summand corresponds to the TSLS bias: it can be seen that in general, it is of the order $\sum_{\ell=1}^{L}\sum_{m=0}^{M}(1 - n_{\ell m}/n_{\ell})/\ddot{r}_n = LM/\ddot{r}_n = K/\ddot{r}_n$. The bias can thus be substantial if the number of instruments K is large relative to the concentration parameter $\ddot{r}_n$. Standard asymptotics, which assume that K is fixed, fail to capture this bias. In our asymptotics, we follow the many weak instruments literature and allow K to grow with the sample size. These asymptotics capture the fact that, in order for the bias to be asymptotically negligible relative to the standard deviation (which is of the order $\ddot{r}_n^{-1/2}$), we need $K^2/\ddot{r}_n$ to converge to zero, which is not an attractive assumption in most of the empirical applications discussed above.

The TSLS bias arises because the predictor $\widehat{R}_{\textsc{tsls},i}$ for observation i is constructed using its own observation, causing Y_i and X_i to be correlated with the noise component of $\widehat{R}_{\textsc{tsls},i}$. To deal with this problem we use a jackknifed version of the TSLS estimator proposed by Ackerberg and Devereux (2009), called the improved jackknife IV estimator: we remove the contribution of $\ddot{X}_i$ from the first-stage predictor $\widehat{R}_{\textsc{tsls},i}$, which, as can be seen from equation (4), is given by $D_i\ddot{X}_i$, with $D_i = (1/n_{S_i} - 1/n_{G_i})$, and rescale the weights on the remaining observations:

$$\hat{\beta}_{\textsc{ijive1}} = \frac{\sum_{i=1}^n \widehat{R}_{\textsc{ijive1},i}\ddot{Y}_i}{\sum_{i=1}^n \widehat{R}_{\textsc{ijive1},i}\ddot{X}_i}, \qquad \widehat{R}_{\textsc{ijive1},i} = (1 - D_i)^{-1}\left(\widehat{R}_{\textsc{tsls},i} - D_i\ddot{X}_i\right).$$

We also study a similar estimator that does not use the rescaling $(1 - D_i)^{-1}$ (which we call IJIVE2). Importantly, IJIVE1 differs from the original jackknife IV estimator (JIVE1) of Phillips and Hale (1977) (see also Angrist et al., 1999), which implements the jackknife correction *first* and *then* partials out the effect of the exogenous covariates (in contrast to IJIVE1, which partials out the effect of the exogenous covariates first). This leads to the estimator that uses, as a first-stage predictor, the sample average of X_j among observations j in group G_i and the same value of the instrument as observation i, with observation i excluded:

$$\hat{\beta}_{\textsc{jive1}} = \frac{\sum_{i=1}^n \widehat{R}_{\textsc{jive1},i}\ddot{Y}_i}{\sum_{i=1}^n \widehat{R}_{\textsc{jive1},i}\ddot{X}_i}, \qquad \widehat{R}_{\textsc{jive1},i} = \frac{1}{n_{S_i} - 1}\sum_{j\colon j\neq i, S_j = S_i} X_j.$$

Finally, we also study a version of the jackknife IV estimator proposed in Kolesár (2013), called UJIVE, which is given by

$$\hat{\beta}_{\text{UJIVE}} = \frac{\sum_{i=1}^{n} \widehat{R}_{\text{UJIVE},i} Y_i}{\sum_{i=1}^{n} \widehat{R}_{\text{UJIVE},i} X_i}, \qquad \widehat{R}_{\text{UJIVE},i} = \frac{1}{n_{G_i} - 1} \sum_{j \neq i\colon G_j = G_i} X_j - \frac{1}{n_{G_i} - 1} \sum_{j \neq i\colon G_j = G_i} X_j.$$

The first-stage predictor $\widehat{R}_{\text{UJIVE},i}$ is similar to the first-stage predictor of JIVE1, except it also partials out the effect of the exogenous covariates by subtracting off the sample average of X_j among observations j in group G_i, with observation i excluded. Because it never uses the treatment status of observation i, the error in this first-stage prediction will be uncorrelated with Y_i and X_i. Furthermore, it only partials out the effect of covariates in the first stage, but not the second stage (by replacing $\ddot{Y}_i$ and $\ddot{X}_i$ with Y_i and X_i). This ensures that the own-observation bias is not reintroduced in the second stage.

Using arguments similar to the derivation of (6), one can show that

$$\begin{aligned} \hat{\beta}_{\text{IJIVE1}} &= \beta_{\text{C}} + \frac{\sum_{\ell=1}^{L} \sum_{m=0}^{M} \sigma_{\eta\nu,\ell m} b_{\ell m}}{\ddot{r}_n + \sum_{\ell=1}^{L} \sum_{m=0}^{M} \sigma_{\eta,\ell m}^2 b_{\ell m}} + o_P(1), \qquad b_{\ell m} = \frac{(1 - n_{\ell m}/n_\ell)/n_\ell}{1 - 1/n_{\ell m} - 1/n_\ell}, \\ \hat{\beta}_{\text{JIVE1}} &= \beta_{\text{C}} - \frac{\sum_{\ell=1}^{L} \sum_{m=0}^{M} \sigma_{\eta\nu,\ell m} n_{\ell m}/n_\ell}{\ddot{r}_n - \sum_{\ell=1}^{L} \sum_{m=0}^{M} \sigma_{\eta,\ell m}^2 n_{\ell m}/n_\ell} + o_P(1), \end{aligned}$$

and

$$\hat{\beta}_{\text{UJIVE}} = \tilde{\beta}_{\text{C}} + o_P(1),$$

where $\tilde{\beta}_{\text{C}}$ is the same estimand as β_{C}, except that the weights $\hat{\omega}_{\ell m}$ are multiplied by $n_\ell/(n_\ell - 1)$. Therefore, if the conditional covariances $\sigma_{\eta\nu,\ell m}$ all have the same sign, the sign of the bias of JIVE1 is the opposite of that of IJIVE1 and TSLS. It can be seen that the JIVE1 bias is of the order $\sum_{\ell=1}^{L} \sum_{m=0}^{M} \frac{n_{\ell m}}{n_\ell \ddot{r}_n} = L/\ddot{r}_n$. Therefore, for the bias to be asymptotically negligible relative to the standard deviation of $\hat{\beta}_{\text{JIVE1}}$, we need $L^2/\ddot{r}_n$ to converge to zero. This is guaranteed under the many instrument asymptotics of Bekker (1994) and Chao et al. (2012), which treats L as fixed. Our theory permits L to increase with the sample size, which allows us to better capture the behavior of JIVE1 in the empirical applications discussed above, in which the number of covariates is tied to the number of instruments.

In comparison, the bias of IJIVE1 is of the order $\sum_{\ell=1}^{L} \sum_{m=0}^{M} b_{\ell m}$. If we assume that in large samples we have at least two observations for each possible value of $s_{\ell m}$, then the denominator of $b_{\ell m}$ is bounded, and the bias can be seen to be of the order $\ddot{r}_n^{-1} \sum_{\ell=1}^{L} M/n_\ell$. Since $n_\ell = \sum_{m=0}^{M} n_{\ell m} \geq (M+1) \min_m n_{\ell m}$, it follows that the bias is bounded by $M/(M+1) \cdot \ddot{r}_n^{-1} L / \min_m n_{\ell m}$. If the design is very unbalanced, so that the number of people assigned instrument $s_{\ell m}$ for some s and m can be thought of as fixed, then we would need $L^2/\ddot{r}_n$ to converge to zero to make sure that the bias is asymptotically negligible, which is the same rate as for JIVE1. Under a balanced design, however, when a comparable number of individuals are assigned each instrument value, so that $1/\min_m n_{\ell m}$ is proportional to $(M+1)L/n$, and the bias is negligible if $\frac{K^2 L^2}{n^2 \ddot{r}_n}$ converges to zero, which is a much weaker requirement than that for JIVE1 or TSLS.

The unconditional estimand is the limit of the conditional estimand, but we need to be careful about defining this limit. It turns out that when the number of covariates and/or instruments is relatively large, the estimands of IJIVE1, IJIVE2, and UJIVE can be asymptotically different, and can differ from the estimand in Imbens and Angrist (1994).

Consider the above example with $M = 1$. In this case the expressions simplify and we can express the conditional estimands as

$$\beta^G_{\mathrm{C}} = \sum_{l=1}^{L} \frac{\hat{\omega}^G_l}{\sum_{l=1}^{L} \hat{\omega}^G_l} \beta_l,$$

where β_l is the LATE that corresponds to the binary instrument in group l.

$$\begin{aligned}
\hat{\omega}^{\text{IJIVE1}}_l &\equiv n_l s^2_{\ddot{R}|l} = \hat{\omega}^{\text{TSLS}}_l = \hat{\omega}^{\text{JIVE1}}_l, \\
\hat{\omega}^{\text{IJIVE2}}_l &\equiv n_l s^2_{\ddot{R}|l} \left(1 - \widehat{\kappa}_{\ddot{R}|l}/n_l\right), \\
\hat{\omega}^{\text{UJIVE}}_l &\equiv n_l s^2_{\ddot{R}|l} \cdot \frac{n_l}{n_l - 1},
\end{aligned}$$

where $s^2_{\ddot{R}|l} = \frac{1}{n_l}\sum_{i:\,G_i=l} \ddot{R}^2_{il}$ is a sample variance estimator for the individuals in group l, $\widehat{\kappa}_{\ddot{R}|l} \equiv \frac{1}{n_l}\sum_{i:\,G_i=l} \ddot{R}^4_{il} \Big/ s^4_{\ddot{R}|l}$ is the kurtosis estimator. We can also write $s^2_{\ddot{R}|l} = \pi_l^2 s^2_{\ddot{Z}|l}$, where $s^2_{\ddot{Z}|l} \equiv \frac{1}{n_l}\sum_{i:\,G_i=l} \ddot{Z}^2_{il}$.

Correspondingly, the unconditional estimands turn out to be

$$\beta^G_U = \sum_{l=1}^{L} \frac{\omega^G_l}{\sum_{\ell=1}^{L} \omega^G_\ell} \beta_l,$$

with

$$\begin{aligned}
\omega^{\text{IJIVE1}}_l &= (np_l - 1)\, \sigma^2_{\widetilde{R}|l} = \omega^{\text{TSLS}}_l = \omega^{\text{JIVE1}}_l, \\
\omega^{\text{IJIVE2}}_l &= \left(np_l - 1 - \kappa_{\widetilde{R}|l}\right) \sigma^2_{\widetilde{R}|l}, \\
\omega^{\text{UJIVE}}_l &= np_l \sigma^2_{\widetilde{R}|l},
\end{aligned}$$

where $\kappa_{\widetilde{R}|l}$ is the population kurtosis of $\widetilde{R}$ in group l.

We show that the difference between these weights in general cannot be ignored. When the number of groups $L \gtrsim \sqrt{n}$, the seemingly negligible difference between the weights can accumulate and lead to asymptotically non-negligible difference in estimands.

When the treatment effects are heterogeneous, the three estimators correspond to different estimands, and the choice of the estimator affects not only the statistical properties such as bias, but also the interpretation of the corresponding estimand. We discuss this in more general settings below. Here, we note that the weights for all three estimators are non-negative. The appeal of the first estimand is that it is a "natural" TSLS estimand. The IJIVE1 estimator allows unbiased estimation of this estimand in the presence of many instruments and covariates. On the other hand, if we formally write down the estimand from Imbens and Angrist (1994), it coincides with the estimand of UJIVE:

$$\beta_{\text{U,IA94}} = \sum_{l=1}^{L} \frac{p_l \sigma^2_{\widetilde{R}|l}}{\sum_{\ell=1}^{L} p_l \sigma^2_{\widetilde{R}|l}} \beta_l.$$

As we will see, this property of UJIVE holds generally. Finally, the estimand of IJIVE2 does not seem to have any particular appeal, hence we do not study it in detail below.

Inference

For inference on the conditional estimand β_{C}, we show in Theorem 5.5 below that under the rate conditions on the rate of growth of K and L above, and if $(K+L)/n \to 0$, one can consistently estimate the asymptotic variance of the discussed estimators by

$$\begin{aligned}\widehat{\mathcal{V}}_{\text{cond}} = \frac{J(\ddot{X}, \ddot{X}, \hat{\sigma}^2_{\nu})}{\left(\sum_{i=1}^{n} \widehat{R}_{\text{IJIVE1},i} \ddot{X}_i\right)^2} &+ \frac{J(\ddot{Y} - \ddot{X}\hat{\beta}_{\text{IJIVE1}}, \ddot{Y} - \ddot{X}\hat{\beta}_{\text{IJIVE1}}, \hat{\sigma}^2_{\eta}) + 2J(\ddot{Y} - \ddot{X}\hat{\beta}_{\text{IJIVE1}}, \ddot{X}, \hat{\sigma}_{\nu\eta})}{\left(\sum_{i=1}^{n} \widehat{R}_{\text{IJIVE1},i} \ddot{X}_i\right)^2} \\ &+ \frac{\sum_{i \neq j} [(H_{\ddot{Z}})^2_{ij} \hat{\sigma}^2_{\eta,j} \hat{\sigma}^2_{\nu,i} + (H_{\ddot{Z}})_{ij} (H_{\ddot{Z}})_{ji} \hat{\sigma}_{\nu\eta,i} \hat{\sigma}_{\nu\eta,j}]}{\left(\sum_{i=1}^{n} \widehat{R}_{\text{IJIVE1},i} \ddot{X}_i\right)^2}\end{aligned}$$

where $H_{\ddot{Z}} = \ddot{Z}(\ddot{Z}'\ddot{Z})^{-1}\ddot{Z}'$ is the projection matrix of the instruments with the covariates partialled out,

$$J(A, B, C) = \sum_{i \neq j \neq k} A_i B_j C_k (H_{\ddot{Z}})_{ik} (H_{\ddot{Z}})_{jk},$$

and $\hat{\sigma}^2_{\nu}$, $\hat{\sigma}^2_{\eta}$, and $\hat{\sigma}_{\nu\eta}$ are estimators of $E[(\zeta_i - \eta_i \beta_{\text{C}})^2 \mid Z_i, W_i]$, $E[(\zeta_i - \eta_i \beta_{\text{C}})^2 \mid Z_i, W_i]$, and $E[(\zeta_i - \eta_i \beta_{\text{C}})^2 \mid Z_i, W_i]$ based on the reduced-form residuals. $J(\cdot, \cdot, \cdot)$ is a jackknife estimator of the variance components: removing the terms for which $i = j$ is necessary to ensure that the variance estimator remains asymptotically unbiased even as the number of instruments and covariates increases with the sample size. The variance estimator has three components: the first component estimates the "usual" asymptotic variance formula that obtains under homogeneous treatment effects and standard asymptotics, the second term accounts for treatment effect heterogeneity, and the third term accounts for the presence of many instruments. For unconditional inference, a consistent estimator of the asymptotic variance has an additional component that reflects the variability of the weights in the conditional estimand when the instruments and covariates are resampled:

$$\widehat{\mathcal{V}}_{\text{uncond}} = \widehat{\mathcal{V}}_{\text{cond}} + \frac{J(\ddot{Y} - \ddot{X}\hat{\beta}, \ddot{Y} - \ddot{X}\hat{\beta}, \widehat{R}^2_{\text{TSLS}})}{\left(\sum_{i=1}^{n} \widehat{R}_{\text{IJIVE1},i} \ddot{X}_i\right)^2}.$$

3 General model and estimators

3.1 Reduced form and notation

There is a sample of individuals $i = 1, \ldots, n$. For each individual i, we observe a vector of exogenous variables W_i with dimension L, and a vector of instruments Z_i with dimension K. Associated with every possible value z of the instrument is a scalar potential treatment $X_i(z)$. We denote the observed treatment by $X_i = X_i(Z_i)$. Associated with every value x of the treatment is a scalar potential outcome $Y_i(x)$. We denote the observed outcome by $Y_i = Y_i(X_i)$. Thus, for each individual we observe the tuple (Y_i, X_i, Z_i, W_i).

Let $R_i = E[X_i \mid Z_i, W_i]$ and $R_{Y,i} = E[Y_i \mid Z_i, W_i]$ denote the reduced-form conditional expectations. We assume that these conditional expectations are linear in the instruments and covariates, so that we can write the first-stage regression as

$$X_i = R_i + \eta_i, \qquad R_i = Z_i'\pi + W_i'\psi, \qquad E[\eta_i \mid Z_i, W_i] = 0, \tag{7}$$

and the reduced-form outcome regression as

$$Y_i = R_{Y,i} + \zeta_i, \qquad R_{Y,i} = Z_i'\pi_Y + W_i'\psi_Y, \qquad E[\zeta_i \mid Z_i, W_i] = 0. \tag{8}$$

In order to ensure that controlling for the covariates linearly is as good as conditioning on them, we also assume that the conditional expectation of Z_i is linear in W_i,

$$E[Z_i \mid W_i] = \Gamma W_i. \tag{9}$$

This assumption is not necessarily restrictive since the setup allows for Z_i to be constructed by interacting an original instrument with the covariates. It also holds trivially in models in which the covariates are discrete and saturated, so that W_i consists of dummy variables, as in Section 2. If the instrument is randomly assigned, W_i only needs to include the constant.

Let Y, X, R, and R_Y denote the vectors with ith element equal to Y_i, X_i, R_i, and $R_{Y,i}$, respectively, and let Z and W denote matrices with the ith row given by Z_i' and W_i', respectively. We denote the right-hand side variables collectively by $Q_i \equiv (Z_i', W_i')'$, and let Q denote the corresponding matrix. For a pair of random variables A_i, B_i that are mean zero conditional on Q, we use the notation $\sigma_{AB,i} = E[A_i B_i \mid Q]$ to denote their conditional covariance, and $\sigma^2_{A,i} = E[A_i^2 \mid Q]$ to denote the conditional variance. For any random vectors A_i, B_i, let $\Sigma_{AB} \equiv E[A_i B_i']$ and $\hat{\Sigma}_{AB} \equiv n^{-1} \sum_{i=1}^n A_i B_i'$. Let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the smallest and largest eigenvalues of a matrix M.

Since we will allow for triangular array asymptotics in which the distribution of the random variables may change with the sample size, the random variables as well as the regression coefficients π, ψ, π_Y, ψ_Y, and Γ are all indexed by n. To prevent notational clutter, we keep this dependence implicit.

For any matrix A, let $H_A = A(A'A)^{-1}A$ denote the projection (hat) matrix, and for any matrix B, let $\ddot{B} = B - H_W B$ denote the residuals after "partialling out" the effect of the covariates W. We denote the population analog by $\widetilde{B}_i = B_i - E[B_i \mid W_i]$. Thus, for instance $\ddot{R}_i = \ddot{Z}_i'\pi$, and $\widetilde{R}_i = \widetilde{Z}_i'\pi$, where $\ddot{Z}_i = Z_i - Z'W(W'W)^{-1}W_i$, and $\widetilde{Z}_i = Z_i - \Gamma W_i$.

3.2 Estimators and estimands

The two-stage least squares estimator can be written as

$$\hat{\beta}_{\text{TSLS}} = \frac{\ddot{Y}'\widehat{R}_{\text{TSLS}}}{\ddot{X}'\widehat{R}_{\text{TSLS}}}, \qquad \widehat{R}_{\text{TSLS}} = H_{\ddot{Z}}\ddot{X}.$$

Here $\widehat{R}_{\text{TSLS}}$ is the first-stage predictor of $\ddot{X}$ based on a linear regression of $\ddot{X}$ on $\ddot{Z}$, and can be thought of as an estimator of $\ddot{R} = \ddot{Z}\pi$. As explained in Section 2, this estimator does not perform well when the strength of the instruments, as measured by a version of the concentration parameter

$$\ddot{r}_n = \sum_{i=1}^{n} \ddot{R}_i^2 = \sum_{i=1}^{n} (\ddot{Z}_i'\pi)^2,$$

relative to their number K is small. The second estimator that we consider is the JIVE1 estimator studied in Phillips and Hale (1977), Angrist et al. (1999), and Blomquist and Dahlberg (1999), given by

$$\hat{\beta}_{\text{JIVE1}} = \frac{\ddot{Y}'\widehat{R}_{\text{JIVE1}}}{\ddot{X}'\widehat{R}_{\text{JIVE1}}}, \qquad \widehat{R}_{\text{JIVE1},i} = \frac{(H_Q X)_i - (H_Q)_{ii}X_i}{1 - (H_Q)_{ii}}.$$

As we argued in Section 2, and as we will show formally below, when the number of covariates L is large, the JIVE1 estimator does not perform well. The related estimator JIVE2, proposed by Angrist et al. (1999), can be shown to behave similarly. The third estimator that we study is the IJIVE1 estimator proposed by Ackerberg and Devereux (2009) that partials out the effect of the covariates first before implementing the jackknife correction,

$$\hat{\beta}_{\text{IJIVE1}} = \frac{\ddot{Y}'\widehat{R}_{\text{IJIVE1}}}{\ddot{X}'\widehat{R}_{\text{IJIVE1}}}, \qquad \widehat{R}_{\text{IJIVE1},i} = \frac{(H_{\ddot{Z}}\ddot{X})_i - (H_{\ddot{Z}})_{ii}\ddot{X}_i}{1 - (H_{\ddot{Z}})_{ii}}.$$

As we will show below, this way of implementing the jackknife correction yields better performance in settings with many covariates. Fourth, we study the related estimator that does not rescale the first stage predictor after removing the contribution of the own observation. We refer to this estimator as IJIVE2, and it is defined as

$$\hat{\beta}_{\text{IJIVE2}} = \frac{\ddot{Y}'\widehat{R}_{\text{IJIVE2}}}{\ddot{X}'\widehat{R}_{\text{IJIVE2}}}, \qquad \widehat{R}_{\text{IJIVE2},i} = (H_{\ddot{Z}}\ddot{X})_i - (H_{\ddot{Z}})_{ii}\ddot{X}_i.$$

Finally, we study a version of the jackknife IV estimator proposed in Kolesár (2013), which only partials out the effect of the covariates when constructing the first-stage predictor, and does not partial out their effect on the treatment X or outcome Y,

$$\hat{\beta}_{\text{UJIVE}} = \frac{Y'\widehat{R}_{\text{UJIVE}}}{X'\widehat{R}_{\text{UJIVE}}}, \qquad \widehat{R}_{\text{UJIVE},i} = \frac{(H_Q X)_i - (H_Q)_{ii} X_i}{1-(H_Q)_{ii}} - \frac{(H_W X)_i - (H_W)_{ii} X_i}{1-(H_W)_{ii}}.$$

Let

$$\beta_{\text{U,IA94}} = \frac{E\left[\widetilde{R}_{Yi}\widetilde{R}_i\right]}{E\left[\widetilde{R}_i^2\right]}$$

denote the probability limit of $\hat{\beta}_{\text{TSLS}}$ under standard asymptotics, as in Imbens and Angrist (1994). We show that the unconditional estimands of the estimators we consider are given by

$$\beta_{\text{U,TSLS}} = \beta_{\text{U,JIVE1}} = \beta_{\text{U,IJIVE1}} = \frac{E\left[\widetilde{R}_{Yi}\widetilde{R}_i\left(1-\frac{1}{n}W_i'\Sigma_{WW}^{-1}W_i\right)\right]}{E\left[\widetilde{R}_i^2\left(1-\frac{1}{n}W_i'\Sigma_{WW}^{-1}W_i\right)\right]}, \tag{10}$$

$$\beta_{\text{U,UJIVE}} = \beta_{\text{U,IA94}}. \tag{11}$$

We define the conditional estimand of an estimator $\hat{\beta}$ as the quantity that would obtain if the reduced-form errors η_i and ζ_i were zero for all i. For TSLS, JIVE1, and IJIVE1, this leads to the same estimand, given by

$$\beta_{\text{C,TSLS}} = \beta_{\text{C,JIVE1}} = \beta_{\text{C,IJIVE1}} = \frac{\frac{1}{n}\sum_{i=1}^n \pi_Y'\ddot{Z}_i\ddot{Z}_i'\pi}{\frac{1}{n}\sum_{i=1}^n \pi'\ddot{Z}_i\ddot{Z}_i'\pi} = \frac{\frac{1}{n}\sum_{i=1}^n \ddot{R}_{Yi}\ddot{R}_i}{\frac{1}{n}\sum_{i=1}^n \ddot{R}_i^2},$$

so that, relative to $\beta_{\text{U,IA94}}$, the population expectation is replaced by a sample average, and the population errors $\widetilde{Z}_i$ are replaced by sample residuals $\ddot{Z}_i$. For IJIVE2, the conditional estimand is different, due to the lack of rescaling:

$$\beta_{\text{C,IJIVE2}} = \frac{\frac{1}{n}\sum_{i=1}^n \pi_Y'\ddot{Z}_i(1-(H_{\ddot{Z}})_{ii})\ddot{Z}_i'\pi}{\frac{1}{n}\sum_{i=1}^n \pi'\ddot{Z}_i(1-(H_{\ddot{Z}})_{ii})\ddot{Z}_i'\pi}.$$

Finally, for UJIVE, the estimand is given by

$$\beta_{\text{C,UJIVE}} = \frac{\frac{1}{n}\sum_{i=1}^n \pi_Y'\ddot{Z}_i(1-(H_W)_{ii})^{-1}\ddot{Z}_i'\pi}{\frac{1}{n}\sum_{i=1}^n \pi'\ddot{Z}_i(1-(H_W)_{ii})^{-1}\ddot{Z}_i'\pi}.$$

The conditional estimand is implicitly indexed by n. Similarly, under the triangular array asymptotics that we consider in this paper, the unconditional estimand also depends on n.

Our aim is to provide valid standard errors for the estimators considered. Making the dependence on n explicit and letting P_n denote the probability measure at sample size n, for an estimator $\hat{\beta}_n$ with conditional and unconditional estimands $\beta_{\text{C},n}$ and $\beta_{\text{U},n}$, we provide standard errors $\widehat{se}_{\text{C},n}$ and $\widehat{se}_{\text{U},n}$, such that, under suitable restrictions on P_n, for a given confidence level $1-\alpha$,

$$\lim_n P_n(|\hat{\beta}_n - \beta_{\mathrm{U},n}| \leq z_{1-\alpha/2}\widehat{se}_{\mathrm{U},n}) = 1-\alpha,$$

and

$$\lim_n P_n(|\hat{\beta}_n - \beta_{\mathrm{C},n}| \leq z_{1-\alpha/2}\widehat{se}_{\mathrm{C},n}) = 1-\alpha,$$

where z_β denotes the β quantile of a standard normal distribution. Before presenting our asymptotic theory, we first discuss causal interpretation of the estimands in the next section.

4 Causal interpretation of estimands

4.1 Conditional estimand

For clarity of exposition, in this section only, we assume that the treatment X_i is binary, and that the instruments Z_i are discrete. The results can be extended to multivalued and continuous treatments by applying the results in Angrist and Imbens (1995) and Angrist et al. (2000), and to continuous instruments by embedding the analysis in the marginal treatment effects framework of Heckman and Vytlacil (1999, 2005).

We split the covariates into two groups, $W_i = (V_i, T_i)$, with T_i possibly absent and V_i corresponding to a vector of L_V group dummies, $V_{ig} = \mathbb{I}\{G_i = g\}$, $g = 1, \dots, L_V$, and $\sum_{g=1}^{L_V} V_{ig} = 1$. If $L_V = 1$, then the group dummies are absent, and V_i corresponds to the intercept. To further simplify the analysis, we assume that the support of the distribution of Z_i conditional on W_i depends only on G_i. Let $\mathcal{Z}_g = \{z_0^g, \dots, z_{J_g}^g\}$ denote the support of Z_i conditional on $G_i = g$. We assume without loss of generality that the support is ordered so that $(z_k^g - z_j^g)'\pi \geq 0$ whenever $k \geq j$. Here T_i are unrestricted controls that enter the model linearly, such as demographic controls. The setup covers cases discussed in Section 2, in which the support of the instrument, such as a judge assignment or an indicator for being born in a particular state in a particular quarter, depends on the group V_i that an individual i belongs to, a neighborhood or a state indicator.

We assume that the instruments are valid in the sense that they are independent of the potential outcomes and potential treatments conditional on the covariates. We also assume that the monotonicity assumption of Imbens and Angrist (1994) holds:

Assumption 1 (LATE model).

(i) (Independence) $\{Y_i(x), X_i(z)\}_{x\in\{0,1\}, z\in\mathcal{Z}_{G_i}} \perp\!\!\!\perp Z_i \mid G_i, T_i$;

(ii) (Monotonicity) For all g and all $z, z' \in \mathcal{Z}_g$, either $P(X_i(z) \geq X_i(z') \mid T_i, G_i = g) = 1$ a.s., or $P(X_i(z) \geq X_i(z') \mid T_i, G_i = g) = 0$ a.s.

For $k > j$, define

$$\alpha(z_k^g, z_j^g) = \frac{(z_k^g - z_j^g)'\pi_Y}{(z_k^g - z_j^g)'\pi},$$

with the convention that $\alpha(z_k^g, z_j^g) = 0$ if $(z_k^g - z_j^g)'\pi = 0$. For any $z_j^g, z_k^g \in \mathcal{Z}_g$ with $k > j$, it follows from Assumption 1 and the results in Imbens and Angrist (1994) that $\alpha(z_k^g, z_j^g)$ corresponds to a local average treatment effect,

$$E[Y_i(1) - Y_i(0) \mid X_i(z_j^g) > X_i(z_k^g), G_i = g, T_i] = \alpha(z_k^g, z_j^g).$$

Due to the linearity assumption on the reduced form given by Equation (9), the covariates do not affect the LATEs directly, only through the support $\mathcal{Z}_g$, which determines for which pairs of instruments z and z' the quantity $\alpha(z, z')$ corresponds to a LATE.

Lemma 4.1. *Consider the reduced form given in equations* (7)–(9)*, and suppose that Assumption 1 holds. Then*

(i)

$$\beta_{C,\text{TSLS}} = \beta_{C,\text{JIVE1}} = \beta_{C,\text{IJIVE1}} = \sum_{g=1}^{L_V} \sum_{j=1}^{J_g} \frac{\hat{\omega}_{gj} \alpha(z_j^g, z_{j-1}^g)}{\sum_{m=1}^{L_V} \sum_{k=1}^{J_g} \hat{\omega}_{mk}},$$

where

$$\hat{\omega}_{gj} = \pi'(z_j^g - z_{j-1}^g) \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{G_i = g\} \mathbb{I}\left\{Z_i \geq z_j^g\right\} \ddot{R}_i. \tag{12}$$

For IJIVE2*, the same conclusion holds with $\ddot{R}_i$ in the definition of $\hat{\omega}_{gj}$ in the equation* (12) *replaced by* $(1 - (H_{\ddot{Z}})_{ii})\ddot{R}_i + e_i' H_W \operatorname{diag}(H_{\ddot{Z}})\ddot{R}$.

(ii) If the only covariates are group dummies, then $\ddot{R}_i = R_i - n_{G_i}^{-1} \sum_{j=1}^{n} \mathbb{I}\{G_j = G_i\} R_j$, where $n_{G_i} = \sum_{j=1}^{n} \mathbb{I}\{G_j = G_i\}$, and the weights $\hat{\omega}_{gj}$ in equation (12) *are positive. Furthermore, in this case the conclusion in Part (i) also holds for* UJIVE*, with the weights $\hat{\omega}_{gj}$ replaced by $\frac{n_{G_i}}{n_{G_i}-1}\hat{\omega}_{gj}$.*

The weights for different LATEs by the conditional estimand are sample analogs of the unconditional weights given below. Unfortunately, we have been unable to give a general condition on the covariates T_i that guarantee positive weights.

4.2 Unconditional Estimand

The estimand given in Imbens and Angrist (1994) and the TSLS estimand have a similar structure of a weighted average of LATEs, but the weights can differ in the presence of many covariates:

$$\beta_{\text{U,TSLS}} = \frac{E\left[\widetilde{R}_{Yi}\widetilde{R}_i\left(1 - \frac{1}{n}W_i'\Sigma_{WW}^{-1}W_i\right)\right]}{E\left[\widetilde{R}_i^2\left(1 - \frac{1}{n}W_i'\Sigma_{WW}^{-1}W_i\right)\right]}, \quad \beta_{\text{U,IA94}} = \frac{E\left[\widetilde{R}_{Yi}\widetilde{R}_i\right]}{E\left[\widetilde{R}_i^2\right]}.$$

Denote $\sigma_{\widetilde{R}}^2(w) \equiv E\left[\widetilde{R}_i^2 | W_i = w\right]$, and for all w with $\sigma_{\widetilde{R}}^2(w) > 0$ let

$$\beta_W(w) \equiv \frac{E\left[\widetilde{R}_{Yi}\widetilde{R}_i | W_i = w\right]}{E\left[\widetilde{R}_i^2 | W_i = w\right]},$$

denote the LATE(or the weighted average of LATEs) conditional on covariates, which is interpreted as in Imbens and Angrist (1994). Set $\beta_W(w) = 0$ for w with $\sigma^2_{\widetilde{R}}(w) = 0$. Then $E\left[\widetilde{R}_{Yi}\widetilde{R}_i|W_i = w\right] = \beta_W(w)\,\sigma^2_{\widetilde{R}}(w)$, and we can write

$$\beta_{\text{U,IA94}} = \int \beta_W(w)\,\frac{\sigma^2_{\widetilde{R}}(w)}{\int \sigma^2_{\widetilde{R}}(w)\,dF_W(w)}dF_W(w)\,.$$

The estimator and the estimand put more weight on the w that have higher variance of the instrument $\sigma^2_{\widetilde{R}}(w)$ and higher density (probability mass function) of W at w. The TSLS estimand can be written as

$$\beta_{\text{U,TSLS}} = \int \beta_W(w)\,\frac{\upsilon^2(w)}{\int \upsilon^2(w)\,dF_W(w)}, \text{ where } \upsilon^2(w) \equiv \sigma^2_{\widetilde{R}}(w)\left(1 - \frac{1}{n}w'\Sigma^{-1}_{WW}w\right).$$

When the covariates have bounded support, $\lambda_{\max}(\Sigma_{WW})/\lambda_{\min}(\Sigma_{WW}) \leq C$ (balanced design), and $L = o(n)$, the weights of the TSLS (IJIVE1) estimand are non-negative, because $\sup_{w\in Support(W)}$ $w'\Sigma^{-1}_{WW}w = o(n)$.

The term $\frac{1}{n}W_i'\Sigma^{-1}_{WW}W_i$ in the unconditional estimand of TSLS appears because, instead of using the population variance $\sigma^2_{\widetilde{R}}(w)$ of $\widetilde{R}_i$ as the weight, the TSLS estimator uses the variance of the sample projection residual $\ddot{R}_i$, which is approximated by $\upsilon^2(w)$.

5 Large sample theory

The weakest condition on the strength of identification and the number of instruments and covariates we consider is

Assumption 2. The error terms (ν_i, η_i) are independent across i, conditionally on Q, and

(i) $(K + L)/n < C$ for some $C < 1$.

(ii) As $n \to \infty$, $\ddot{r}_n \to \infty$ and $\sum_{i=1}^n \ddot{R}^2_{Y,i}/\ddot{r}_n$ is bounded a.s.

(iii) $K/\ddot{r}^2_n \overset{a.s.}{\to} 0$.

Part (i) rules out the case in which the number of instruments and covariates is larger than the sample size. Part (ii) prevents Staiger and Stock (1997)-type asymptotics by requiring $\ddot{r}_n$ to diverge to ∞ (we will show below that $\ddot{r}_n$ determines the rate of convergence). Assuming that elements of $E_n[\ddot{R}^2_{Y,i}]$ are of the same order essentially requires that the LATEs are bounded, which holds automatically if the treatment effects are constant. This condition can be replaced by the assumption that $\sum_{i=1}^n \ddot{R}^2_{\Delta,i}/\ddot{r}_n$ is bounded a.s., where $\ddot{R}_\Delta = \ddot{R}_Y - \ddot{R}\beta_{\text{C}}$. However, since β_{C} depends on the estimator, this assumption is somewhat awkward. Part (iii) of the assumption is needed in order to ensure that the conditional variance of each estimator vanishes with the sample size.

To control the asymptotic bias of the estimators, and to construct standard errors that consistently estimate the asymptotic standard deviation of the estimators, we will need to further restrict the rate conditions on K and L, as explained further below.

Assumption 3.

(i) $E[\eta_i \mid Q] = 0$ and $E[\zeta_i \mid Q] = 0$, $E[\nu_i^2 + \eta_i^2 \mid Q]$ is bounded, and $|\operatorname{corr}(\zeta_i, \eta_i \mid Q)|$ is bounded away from one. Furthermore, $\sigma^2_{\zeta,i}$ is bounded away from zero.

(ii) $E[\nu_i^4 + \eta_i^4 \mid Q]$ is bounded.

Part (i) will be needed for consistency, and to make sure that the asymptotic covariance matrix is not degenerate. Part (ii) is needed for asymptotic normality, and also to derive the probability limits of inconsistent estimators.

The following assumption is used to establish the unconditional asymptotic results. Let $\widetilde{r}_n = nE[\widetilde{R}_i^2]$ denote the population analog of $\ddot{r}_n$.

Assumption 4. The observed data (Y_i, X_i, Z_i, W_i) is i.i.d., and

(i) $(K+L)\log(K+L)/n \to 0$.

(ii) $\widetilde{r}_n \to \infty$ and $E\left[\widetilde{R}_{Y,i}^{2(1+\delta)} + \widetilde{R}_i^{2(1+\delta)}\right] / E\left[\widetilde{R}_i^2\right]^{1+\delta}$ is bounded.

(iii) $K/\widetilde{r}_n^2 \to 0$.

(iv) $\lambda_{\max}(E[Q_iQ_i'])/\lambda_{\min}(E[Q_iQ_i'])$ is bounded.

(v) (Simple Sufficient Condition) $\|Q_i\|^2/E[\|Q_i\|^2]$ is bounded.

Parts (i)–(iii) are population analogs of Assumption 2(i)-(iii). Part (iv) is a balanced design assumption. Part (v) is used for the analysis of large-dimensional random matrices in the definitions of the estimators. It in particular allows for the covariates and/or instruments to be spline functions of some underlying low-dimensional variables. This condition can be relaxed. Assumption 4 in particular ensures that the conditional assumptions made above and below are satisfied w.p.a.1 unconditionally.

5.1 Consistency

To state the consistency results for the conditional estimand, note that each estimator that we consider can be written as $\hat{\beta}_G = \sum_{i,j} Y_i G_{ij} X_j / \sum_{i,j} X_i G_{ij} X_j$ for some matrix G. In particular,

$$G_{\text{TSLS}} = H_{\ddot{Z}}, \tag{13a}$$

$$G_{\text{IJIVE1}} = (I - H_W)(I - \operatorname{diag}(H_{\ddot{Z}}))^{-1}(H_{\ddot{Z}} - \operatorname{diag}(H_{\ddot{Z}}))(I - H_W), \tag{13b}$$

$$G_{\text{IJIVE2}} = H_{\ddot{Z}} - (I - H_W)\operatorname{diag}(H_{\ddot{Z}})(I - H_W), \tag{13c}$$

$$G_{\text{JIVE1}} = (I - H_W)(I - \operatorname{diag}(H_Q))^{-1}(H_Q - \operatorname{diag}(H_Q)), \tag{13d}$$

$$G_{\text{UJIVE}} = (I - \operatorname{diag}(H_Q))^{-1}(H_Q - \operatorname{diag}(H_Q)) - (I - \operatorname{diag}(H_W))^{-1}(H_W - \operatorname{diag}(H_W)). \tag{13e}$$

Under Assumptions 2 and 3, appropriately scaled sums in the numerator and denominator of $\hat{\beta}_G$, will converge to their conditional expectations, so that $\hat{\beta}_G - \frac{E[Y'GX|Q]}{E[X'GX|Q]} = o_P(1)$. We can write this as

$$\hat{\beta}_G - \beta_{\text{C},G} - \text{bias}(\hat{\beta}_G) \xrightarrow{p} 0, \tag{14}$$

where

$$\text{bias}(\hat{\beta}_G) = \frac{\sum_i G_{ii}\sigma_{\nu,\eta,i}}{\sum_i R_i(GR)_i + \sum_i G_{ii}\sigma^2_{\eta,i}} \tag{15}$$

is the conditional asymptotic bias of the estimator. Here $\nu_i = \zeta_i - \eta_i\beta_{\text{C},G}$, and $\beta_{\text{C},G} = R_Y'GR/R'GR$ is the conditional estimand. The diagonal elements G_{ii} of the matrix G exactly capture the bias that arises because the first-stage predictor of the treatment for individual i puts weight G_{ii} on the individual's observed treatment, accounting for the effect of partialling out the exogenous covariates. For the estimators we consider, $E_n[R_i(GR)_i] = \ddot{r}_n/n$, so that $\text{bias}(\hat{\beta}_G) = O(\sum_i|G_{ii}|/\ddot{r}_n)$. We will therefore need to control the diagonal elements of G_{ii} to ensure that there is no bias. Since for UJIVE $G_{ii} = 0$, its bias is zero.

Theorem 5.1. *Suppose Assumption 2 and Assumption 3 (i) hold.*

1. *If $K/\ddot{r}_n \to 0$, then $\hat{\beta}_{\text{TSLS}} = \beta_{C,\text{TSLS}} + o_P(1)$, where $\beta_{C,\text{TSLS}} = \frac{\ddot{R}_Y'\ddot{R}}{\ddot{R}'\ddot{R}}$. If $K/\ddot{r}_n$ is bounded and Assumption 3 (ii) holds, then $\hat{\beta}_{\text{TSLS}} = \beta_{C,\text{TSLS}} + \text{bias}(\hat{\beta}_{\text{TSLS}}) + o_P(1)$.*
2. *Suppose that for some $C < 1$, $\max_i(H_Q)_{ii} \leq C$. If $L/\ddot{r}_n \to 0$, then $\hat{\beta}_{\text{JIVE1}} = \beta_{C,\text{TSLS}} + o_P(1)$. If instead $L/\ddot{r}_n$ is bounded, Assumption 3 (ii) holds, and $\text{bias}(\hat{\beta}_{\text{JIVE1}})$ is bounded, then $\hat{\beta}_{\text{JIVE1}} = \beta_{C,\text{TSLS}} + \text{bias}(\hat{\beta}_{\text{JIVE1}}) + o_P(1)$.*
3. *Suppose that for some $C < 1$, $\max_i(H_{\ddot{Z}})_{ii} \leq C$. If $L\max_i(H_{\ddot{Z}})_{ii}/\ddot{r}_n \to 0$, then $\hat{\beta}_{\text{IJIVE1}} = \beta_{C,\text{TSLS}} + o_P(1)$. If instead $L\max_i(H_{\ddot{Z}})_{ii}/\ddot{r}_n$ is bounded, Assumption 3 (ii) holds, and $\text{bias}(\hat{\beta}_{\text{IJIVE1}})$ is bounded, then $\hat{\beta}_{\text{IJIVE1}} = \beta_{C,\text{TSLS}} + \text{bias}(\hat{\beta}_{\text{IJIVE1}}) + o_P(1)$.*
4. *Suppose that for some $C < 1$, $\max_i(H_{\ddot{Z}})_{ii} \leq C$. If $L\max_i(H_{\ddot{Z}})_{ii}/\ddot{r}_n \to 0$, then $\hat{\beta}_{\text{IJIVE2}} = \beta_{C,\text{IJIVE2}} + o_P(1)$, where $\beta_{C,\text{IJIVE2}} = \frac{\ddot{R}'(I-D_{\ddot{Z}})\ddot{R}_Y}{\ddot{R}'(I-D_{\ddot{Z}})\ddot{R}}$. If instead $L\max_i(H_{\ddot{Z}})_{ii}/\ddot{r}_n$ is bounded, Assumption 3 (ii) holds, and $\text{bias}(\hat{\beta}_{\text{IJIVE2}})$ is bounded, then $\hat{\beta}_{\text{IJIVE2}} = \beta_{C,\text{IJIVE2}} + \text{bias}(\hat{\beta}_{\text{IJIVE2}}) + o_P(1)$.*
5. *Suppose that for some $C < 1$, $\max_i(H_{\ddot{Z}})_{ii} \leq C$, that $\max_i(H_W)_{ii} \to 0$ a.s., $\max_i(|R_i| + |R_{Y,i}|)$ is bounded a.s., and that $L/\ddot{r}_n$ is bounded. Then $\hat{\beta}_{\text{UJIVE}} = \beta_{C,\text{UJIVE}} + o_P(1)$.*

The rate conditions given in the theorem control the bias of each estimator. For the jackknife estimators, G_{ii} may be negative, so that the denominator, scaled by n, in equation (15) may converge to zero even as $R'GR \to \infty$, so that the bias would grow unbounded. In order to prevent this, the theorem assumes directly that the bias is bounded. In general, the proof of the theorem shows that the bias of TSLS is of the order $K/\ddot{r}_n$, that of JIVE1 is of the order $L/\ddot{r}_n$, while the bias of IJIVE1 and IJIVE2 is of the order $L\max_i(H_{\ddot{Z}})_{ii}/\ddot{r}_n$. The term $\max_i(H_{\ddot{Z}})_{ii}$ measures the balance of the design: if the design is balanced, so that $\max_i (H_{\ddot{Z}})_{ii}$ is proportional to K/n, then the bias of IJIVE1 and IJIVE2 remains negligible under a much weaker condition on the rate of growth of L as that of JIVE1. For the asymptotic normality results and inference, we will therefore concentrate on TSLS, IJIVE1, and UJIVE for brevity.

Theorem 5.2. *Suppose Assumption 3 (i) and Assumption 4 hold. Then* $\ddot{r}_n/\widetilde{r}_n \xrightarrow{p} 1$, $\beta_{C,G} = \beta_{U,G} + o_P(1)$, $\hat{\beta} = \beta_{U,G} + o_P(1)$, *and* $\beta_{U,G} = \beta_{U,IA94} + o_P(1)$, *under the following conditions:*

Estimator	*Conditions*
TSLS	$K/\tilde{r}_n \to 0$
JIVE1	$L/\tilde{r}_n \to 0$
IJIVE1	$LK/(\tilde{r}_n n) \to 0$
UJIVE	$L/\tilde{r}_n^2 \to 0$

The theorem establishes that the conditional estimand converges to the unconditional one, and that the conditional regularity assumptions made by Theorem 5.1 hold with probability approaching one, and hence the estimators are consistent unconditionally.

5.2 Asymptotic normality

For the asymptotic normality, we will need to ensure no single observation has too much influence on the strength of identification:

Assumption 5. $\sum_i \ddot{R}_i^4/\ddot{r}_n^2 \xrightarrow{a.s.} 0$ and $\sum_i \ddot{R}_{Y,i}^4/\ddot{r}_n^2 \xrightarrow{a.s.} 0$.

This assumption is equivalent to Assumption 5 in Chao et al. (2012). It is needed to verify the Lindeberg condition in showing asymptotic normality of the estimators.

To state the asymptotic normality results, given a particular conditional or unconditional estimand β, let $\ddot{R}_\Delta = \ddot{R}_Y - \ddot{R}\beta$, and let $\nu_i = \zeta_i - \eta_i \beta$. Under constant treatment effects, $\ddot{R}_\Delta = 0$, and ν_i can be interpreted as the structural error.

Theorem 5.3. *Suppose that Assumptions 2, 3 and 5 hold.*

1. *If* $K^2/\ddot{r}_n \to 0$, *then*
$$\left(\frac{\mathcal{V}_C}{\ddot{r}_n}\right)^{-1/2} (\hat{\beta}_{\text{TSLS}} - \beta_{C,\text{TSLS}}) \xrightarrow{d} \mathcal{N}(0,1),$$
where
$$\mathcal{V}_C = \frac{1}{\ddot{r}_n} \sum_i [\ddot{R}_i^2 \sigma_{\nu,i}^2 + \sigma_{\eta,i}^2 \ddot{R}_{\Delta,i} (\beta_{C,\text{TSLS}})^2 + 2\sigma_{\nu\eta,i} \ddot{R}_i \ddot{R}_{\Delta,i}].$$

2. *Suppose further that* $L \max_i (H_{\ddot{Z}})_{ii}/\sqrt{\ddot{r}_n} \xrightarrow{a.s.} 0$, $\max_i (H_{\ddot{Z}})_{ii} \xrightarrow{a.s.} 0$, *and that* $K/\ddot{r}_n$ *is bounded, and let*
$$\mathcal{V}_{MW} = \frac{1}{\ddot{r}_n} \sum_{i \neq j} [(H_{\ddot{Z}})_{ij}^2 \sigma_{\eta,j}^2 \sigma_{\nu,i}^2 + (H_{\ddot{Z}})_{ij} (H_{\ddot{Z}})_{ji} \sigma_{\nu\eta,i} \sigma_{\nu\eta,j}].$$

Then

$$\left(\frac{\mathcal{V}_C + \mathcal{V}_{MW}}{\ddot{r}_n}\right)^{-1/2} (\hat{\beta}_{\text{IJIVE1}} - \beta_{C,\text{TSLS}}) \xrightarrow{d} \mathcal{N}(0,1).$$

If instead $K/\ddot{r}_n \to \infty$, then the above holds with $H_{\ddot{Z}}$ in the definition of $\mathcal{V}_{MW}$ replaced by G_{IJIVE1}.

3. *Suppose that $(L+K)/\ddot{r}_n$ is bounded, $\max_i (H_Q)_{ii} \to 0$, and $\max_i(|R_i| + |R_{Y,i}|)$ is bounded a.s. Then*

$$\left(\frac{\mathcal{V}_C + \mathcal{V}_{MW}}{\ddot{r}_n}\right)^{-1/2} (\hat{\beta}_{\text{UJIVE}} - \beta_{C,\text{UJIVE}}) \xrightarrow{d} \mathcal{N}(0,1).$$

Before discussing this result, it is useful to state the corresponding unconditional inference result.

Assumption 6.

(i) $E\left[\widetilde{R}_{Yi}^4 + \widetilde{R}_i^4 \middle| W_i\right]^{1/2} / E\left[\widetilde{R}_i^2\right] \le C$ a.s., and $E\left[\{(\widetilde{R}_{Yi}^2 + \widetilde{R}_i^2)\widetilde{R}_i^2\}^{1+\delta}\right] / E\left[\widetilde{R}_i^2\right]^{1+\delta} \le C$ for some $C > 0$.

(ii) $L^4 \log^2 L = o\left(n^3\right)$, and $\lambda_{\psi,n} = o\left(n^3\right)$, where $\lambda_{\psi,n} \equiv E\left[(\psi_i'\psi_j)^4\right]$, $\psi_i \equiv \Sigma_{WW}^{-1/2} W_i$.

Theorem 5.4. *Suppose Assumptions 3, 4, and 6(i) hold. Then, under the additional restrictions listed below,*

$$\left(\frac{\Omega_C + \Omega_E + \Omega_{MW}}{\widetilde{r}_n}\right)^{-1/2} (\hat{\beta}_G - \beta_{U,G}) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\begin{aligned}
\Omega_C &= \frac{1}{E[\widetilde{R}_i^2]} E[(\widetilde{R}_i \nu_i + \widetilde{R}_{\Delta,i}\eta_i)^2], \\
\Omega_E &= \frac{1}{E[\widetilde{R}_i^2]} E[\widetilde{R}_i^2 \widetilde{R}_{\Delta,i}^2], \\
\Omega_{MW} &= \frac{1}{\widetilde{r}_n} \operatorname{tr}\left(E[\nu_i^2 \widetilde{Z}_i \Sigma_{\widetilde{Z}\widetilde{Z}}^{-1} \widetilde{Z}_i'] E[\eta_i^2 \widetilde{Z}_i \Sigma_{\widetilde{Z}\widetilde{Z}}^{-1} \widetilde{Z}_i'] + E[\nu_i \eta_i \widetilde{Z}_i \Sigma_{\widetilde{Z}\widetilde{Z}}^{-1} \widetilde{Z}_i']^2 \right).
\end{aligned}$$

These results hold under the following assumptions:

1. *For $\hat{\beta}_{\text{TSLS}}$, if $K/\widetilde{r}_n \to 0$ and Assumption 6(ii) holds, with $\beta_{U,\text{TSLS}}$ defined in equation (10).*

2. *For $\hat{\beta}_{\text{IJIVE1}}$, if $K/\widetilde{r}_n$ is bounded and Assumption 6(ii) holds, with $\beta_{U,\text{IJIVE1}} = \beta_{U,\text{TSLS}}$ defined in equation (10).*

3. *For $\hat{\beta}_{\text{UJIVE}}$, if $(K+L)/\widetilde{r}_n$ and $|R_i| + |R_{Yi}|$ are bounded, with $\beta_{U,\text{UJIVE}}$ defined in equation (11).*

Let us first discuss the form of the asymptotic variance. The terms Ω_{C}, Ω_{MW}, and Ω_{E} are population analogs of $\mathcal{V}_{\text{C}}$, $\mathcal{V}_{\text{MW}}$, and $\mathcal{V}_{\text{E}}$. The term Ω_{MW} corresponds to the contribution to the asymptotic variance coming from many instruments. Under homoskedasticity, it simplifies to $K/\widetilde{r}_n \cdot (\sigma_\eta^2 \sigma_\nu^2 + \sigma_{\nu\eta}^2)$. It has the same form whether or not there is treatment effect heterogeneity, except that ν_i cannot in general be interpreted as the structural error. When the number of instruments grows slowly enough so that $K/\ddot{r}_n \to 0$, this term is negligible relative to $\mathcal{V}_{\mathrm{C}}$. This happens, in particular, under the standard asymptotics that hold the distribution of the data fixed as $n \to \infty$. For TSLS the condition $K/\ddot{r}_n \to 0$ is needed for consistency, so that the many instrument term is always of smaller order. If $K/\ddot{r}_n \to \infty$ then in general the many instruments term $\mathcal{V}_{\mathrm{MW}}$ dominates, the rate convergence is slower than $1/\ddot{r}_n^{1/2}$, and the asymptotic variances of different estimators may differ. On the other hand, if $K/\ddot{r}_n$ is bounded, the asymptotic variance for IJIVE1 and UJIVE is the same.

The term Ω_{E} accounts for the variability of the conditional estimand β_{C}. As a part of the proof of the theorem, we show that $\widetilde{r}_n^{1/2}(\beta_{\mathrm{C}} - \beta_{\mathrm{U}}) \xrightarrow{d} \mathcal{N}(0, \Omega_{\mathrm{E}})$. Theorem 5.4 effectively shows that this result also obtains under the many instrument asymptotics, and that, in addition, the term $\beta_{\mathrm{C}} - \beta_{\mathrm{U}}$ is asymptotically independent of the term $\hat{\beta} - \beta_{\mathrm{C}}$.

The term $\mathcal{V}_{\mathrm{C}}$ corresponds to the asymptotic variance of $\ddot{R}_i \nu_i + \ddot{R}_{\Delta,i}\eta_i$. The first term of $\mathcal{V}_{\mathrm{C}}$, $\sum_i \ddot{R}_i^2 \sigma_{\nu,i}^2$, accounts for the variance of $\ddot{R}_i \nu_i$, and corresponds to the standard asymptotic variance for TSLS: it is the only term present under the standard asymptotics and the assumption that the treatment effects are constant. The term $\ddot{R}_{\Delta,i}\eta_i$ corresponds to the uncertainty due to the treatment effects being different for different individuals. Typically, this uncertainty increases the asymptotic variance, i.e., typically $\mathcal{V}_{\mathrm{C}} \geq \sum_i \ddot{R}_i^2 \sigma_{\nu,i}^2$. Let us make a few remarks about the regularity and rate conditions:

Remark 1. The conditions $K^2/\ddot{r}_n \to 0$ for TSLS and $L^2 \max_i (H_{\ddot{Z}})_{ii}/\sqrt{\ddot{r}_n} \xrightarrow{a.s.} 0$ for IJIVE1 estimators in Theorem 5.3 ensure that the conditional bias of the estimator is negligible relative to its standard deviation. If these conditions are relaxed to $K^2/\ddot{r}_n$ and $L^2 \max_i (H_{\ddot{Z}})_{ii}/\sqrt{\ddot{r}_n}$ being bounded, then it follows from the proof of the theorem that the estimators will remain asymptotically normal, but one needs to subtract from $\hat{\beta}$ the conditional bias in addition to the conditional estimand, since the bias is no longer asymptotically negligible (see also Lemma D.5 in the appendix). However, it is unclear how to do inference in this case as it is unclear how one could properly center the confidence intervals.

Remark 2. Note that the estimands of TSLS and IJIVE1 differ from $\beta_{\mathrm{U,IA94}}$, while $\beta_{\mathrm{U,UJIVE}} = \beta_{\mathrm{U,IA94}}$. The difference between the estimands is potentially non-negligible when $\sqrt{\widetilde{r}_n} E\left[\widetilde{R}_{Yi}\widetilde{R}_i \frac{1}{n} W_i' \Sigma_{WW}^{-1} W_i\right] \simeq \frac{1}{n^2}\widetilde{r}_n^{3/2} L \not\to 0$. When the instruments are strong, the condition is $L/\sqrt{n} \not\to 0$.

Remark 3. As a part of the proof of Theorem 5.4 we show that $(\mathcal{V}_{\mathrm{C}} + \mathcal{V}_{\mathrm{MW}})/(\Omega_{\mathrm{C}} + \Omega_{\mathrm{MW}}) \xrightarrow{p} 1$.

Remark 4. Assumption 6 (ii) is used to derive the unconditional distribution of IJIVE1 in Theorem 5.4. We can view $E\left[(\psi_i'\psi_j)^4\right] \Big/ E\left[\|\psi_i\|^2\right]^4 \simeq \lambda_{\psi,n}/L^4$ as a measure of orthogonality of the independent random vectors ψ_i and ψ_j. Random vectors in high-dimensional spaces tend to be nearly orthogonal, and the rate at which $E\left[(\psi_i'\psi_j)^4\right]$ grows with L reflects the dependence structure of the components of the vector ψ_i. For example, $\lambda_{\psi,n} \simeq L^2$ when the components ψ_{il} are independent across l, $E\left[\psi_{il}\right] = 0$, and $E\left[|\psi_{il}|^4\right] \leq C$. When W_i are (appropriately rescaled) draws from Multinomial$(p_1, \ldots, p_L)$, satisfying the balance condition $\max_{l \leq L} p_l / \min_{l \leq L} p_l \leq C < \infty$, we have $\lambda_{\psi,n} \simeq L^3$.

Remark 5. If $\|\psi_i\| \leq \zeta_0(L)$ for a nonrandom function $\zeta_0(L)$, then $E\left[(\psi_i'\psi_j)^4\right] \lesssim \min\{\zeta_0(L)^4 L$, $\zeta_0(L)^2 E[\|\psi_i\|^4]\}$. An important case is $W_i \equiv \varphi^L(\mathcal{W}_i)$ for some low-dimensional observed variables $\mathcal{W}_i$, whose effect we are modelling nonparametrically, and $\varphi^L(\cdot)$ is a vector of some basis functions scaled to satisfy Assumption 4. For example, $\zeta_0\left(L\right) \leq C\sqrt{L}$ for splines when $\mathcal{W}$ has compact support, and hence $E\left[(\psi_i'\psi_j)^4\right] \lesssim L^3$.

Remark 6. The condition $\max_i (H_{\ddot{Z}})_{ii} \overset{a.s.}{\to} 0$ in Theorem 5.3 is a balance condition on the design. It requires that $K/n \to 0$. It follows from the proof of the theorem that the condition may be replaced by weaker regularity conditions that, in particular, allow K to grow as fast as n. In that case, one also needs to replace $\ddot{R}_i$ by $(GR)_i$ and $\ddot{R}_{\Delta,i}$ by $(G'R_{\Delta})_i$ in the expression for $\mathcal{V}_{\mathrm{C}}$, and replace $H_{\ddot{Z}}$ by G in the expression of $\mathcal{V}_{\mathrm{MW}}$. A sufficient weaker regularity condition is that L is constant, and that the treatment effects are homogeneous, in which case the result is similar to that for JIVE1 in Chao et al. (2012). Since we require the balance condition $\max_i (H_{\ddot{Z}})_{ii} \overset{a.s.}{\to} 0$ in order to construct a consistent standard error estimator, we impose it already in Theorem 5.3 as it allows us to state the results in a more unified way.

5.3 Inference

To define the standard error estimator that we consider, let $\hat{\eta} = X - H_Q X$ and $\hat{\zeta} = Y - H_Q Y$ denote residuals from the reduced-form regressions. We use plug-in estimators of σ_ν, $\sigma_{\nu\eta}$, and σ^2_η to estimate the variance components $\mathcal{V}_{\mathrm{C}}$, $\mathcal{V}_{\mathrm{E}}$, and $\mathcal{V}_{\mathrm{MW}}$:

$$\hat{\sigma}^2_{\nu,i} = (\hat{\zeta}_i - \hat{\eta}_i\hat{\beta})^2, \qquad \hat{\sigma}_{\nu\eta,i} = (\hat{\zeta}_i - \hat{\eta}_i\hat{\beta})\hat{\eta}_i, \qquad \hat{\sigma}^2_{\eta,i} = \hat{\eta}^2_i.$$

Rather than using a plug-in estimator for $\ddot{R}_i$, and $\ddot{R}_{\Delta,i}$ in the expression for $\mathcal{V}_{\mathrm{C}}$ and $\mathcal{V}_{\mathrm{E}}$, we use the following jackknife estimators

$$\widehat{\mathcal{V}}_{\mathrm{C}} = \frac{1}{\widehat{r}_{n,\textsc{ijive1}}}\left(J(\ddot{X}, \ddot{X}, \hat{\sigma}^2_\nu) + J(\ddot{Y} - \ddot{X}\hat{\beta}, \ddot{Y} - \ddot{X}\hat{\beta}, \hat{\sigma}^2_\eta) + 2J(\ddot{Y} - \ddot{X}\hat{\beta}, \ddot{X}, \hat{\sigma}_{\nu\eta})\right),$$

and

$$\widehat{\mathcal{V}}_{\mathrm{E}} = \frac{1}{\widehat{r}_{n,\textsc{ijive1}}} J(\ddot{Y} - \ddot{X}\hat{\beta}, \ddot{Y} - \ddot{X}\hat{\beta}, \widehat{R}^2_{\textsc{tsls}}),$$

where $\widehat{r}_{n,\textsc{ijive1}} = \sum_i \ddot{X}_i \widehat{R}_{\textsc{ijive1},i}$ and

$$J(A, B, C) = \sum_{i \neq j \neq k} A_i B_j C_k (H_{\ddot{Z}})_{ik} (H_{\ddot{Z}})_{jk}.$$

The "jackknifing" in the definition of J removes the asymptotic bias of the estimators. Here $\widehat{r}_{n,\textsc{ijive1}}$ is an estimator of $\ddot{r}_n$. For $\mathcal{V}_{MW}$, we use the estimator

$$\widehat{\mathcal{V}}_{\mathrm{MW}} = \frac{1}{\widehat{r}_{n,\textsc{ijive1}}} \sum_{i \neq j} [(H_{\ddot{Z}})^2_{ij} \hat{\sigma}^2_{\eta,j} \hat{\sigma}^2_{\nu,i} + (H_{\ddot{Z}})_{ij} (H_{\ddot{Z}})_{ji} \hat{\sigma}_{\nu\eta,i} \hat{\sigma}_{\nu\eta,j}].$$

The standard errors for the conditional and unconditional estimands are given by

$$\widehat{se}_{\mathrm{C},n} = \sqrt{(\widehat{\mathcal{V}}_{\mathrm{C}} + \widehat{\mathcal{V}}_{\mathrm{MW}})/\,\widehat{r}_{n,\textsc{ijive1}}},$$

$$\widehat{se}_{\mathrm{U},n} = \sqrt{(\widehat{\mathcal{V}}_{\mathrm{C}} + \widehat{\mathcal{V}}_{\mathrm{MW}} + \widehat{\mathcal{V}}_{\mathrm{E}})/\,\widehat{r}_{n,\textsc{ijive1}}}.$$

To show the consistency of $\widehat{se}_{\mathrm{C},n}$, we strengthen Assumption 5 to

Assumption 7. $\max_i |\ddot{R}_i| + \max_i |\ddot{R}_{Y,i}|$ are bounded a.s.

This assumption is similar to Assumption 6 in Chao et al. (2012).

Theorem 5.5. *Suppose that Assumptions 2, 3 and 7 hold. Suppose further that* $\max_i (H_Q)_{ii} \overset{a.s.}{\to} 0$*, and that* $(K+L)/\ddot{r}_n$ *is bounded a.s. Then,*

$$\widehat{se}^2_{C,n} = (\mathcal{V}_C + \mathcal{V}_{MW})/\ddot{r}_n + o_P(1/\ddot{r}_n).$$

The additional balance condition $\max_i (H_Q)_{ii} \overset{a.s.}{\to} 0$ that we impose is essential in proving the theorem. It implies that $(K+L)/n \to 0$, and ensures that bias induced by estimating the variance of the reduced-form errors is asymptotically negligible. Cattaneo et al. (2016) show that a similar balance condition is needed for the Eicker-Huber-White standard errors to be consistent in linear regression. When the treatment effects are homogeneous and when the number of covariates L is fixed, one can estimate the terms σ_ν and $\sigma_{\nu\eta}$ at a faster rate, and this condition is not needed. Cattaneo et al. (2016) also suggest an alternative estimator that does not require this condition. It is unclear however whether one can adapt their estimator to the current setting since the variance expression contains products of second moments of the reduced-form errors, rather than just second moments.

Relative to the asymptotic normality result, we need also to rule out the case in which K or L may grow faster than the concentration parameter. This is sufficient to ensure that the error in estimating the standard errors is negligible.

Assumption 8. $|\widetilde{R}_i| + |\widetilde{R}_{Y,i}|$ are bounded.

Theorem 5.6. *Suppose the conditions of Theorem 5.4 and Assumption 8 hold . Then,*

$$\widehat{se}^2_{U,n} = (\Omega_C + \Omega_{MW} + \Omega_E)/\widetilde{r}_n + o_P(1/\widetilde{r}_n).$$

For unconditional inference, the balance condition $\max_i (H_Q)_{ii} \to 0$ holds in large samples under the i.i.d. sampling and the rate conditions imposed by Assumption 6, and therefore does not need to be made explicit.

Appendices

The appendix is organized as follows. Appendix A contains general results and bounds for the estimators considered in the paper used throughout the rest of the appendix. Appendix B proves the Lemma in Section 4. Appendices C and E prove the conditional and unconditional results in Section 5, respectively. Appendices D and F contain auxiliary results used in Appendices C and E.

Below, w.p.a.1 stands for "with probability approaching 1 as $n \to \infty$". We write $a \prec b$ if there exists a constant C such that $a \leq b$. We write $a \preceq_{\text{a.s.}} b$ or $a \prec_{w.p.a.1} b$ if $a \prec b$ almost surely or w.p.a.1. Let $\|a\|_2$ or simply $\|a\|$ denote the Euclidean (ℓ_2) norm of a vector, and let $\|A\|_F$ denote the Frobenius norm of a matrix, and $\|A\|_2$ or $\|A\|_\lambda$ the spectral norm.

Appendix A Properties of estimators considered

It will be useful to collect some properties of the estimators that we consider, which we will use throughout the proof. The estimators we consider in this paper have the general form

$$\hat{\beta}_G = \frac{\sum_{i,j} Y_i G_{ij} X_j}{\sum_{i,j} X_i G_{ij} X_j}, \tag{16}$$

with the matrix G for different estimators given in equation (13). Observe that

$$\begin{aligned} G_{\text{TSLS}}(R_Y, R) = G_{\text{JIVE1}}(R_Y, R) &= G_{\text{IJIVE1}}(R_Y, R) = (\ddot{R}_Y, \ddot{R}), \\ G_{\text{IJIVE2}}(R_Y, R) &= (I - D_{\ddot{Z}} + H_W D_{\ddot{Z}})(\ddot{R}_Y, \ddot{R}), \\ G_{\text{UJIVE}}(R_Y, R) &= (I - D_W)^{-1}(\ddot{R}_Y, \ddot{R}). \end{aligned} \tag{17}$$

We now collect several useful bounds. First, we bound the norms $\|GR\|$, $\|G'R\|$, $\|GR_\Delta\|$, and $\|G'R_\Delta\|$. Using the triangle inequality, and the fact that for any projection matrix P and a vector a, $\|Pa\|_2 \leq \|a\|_2$, and $0 \leq P_{ii} \leq 1$, we obtain the bounds

$$\begin{aligned} \|G_{\text{TSLS}}R\|_2 = \|G_{\text{JIVE1}}R\|_2 = \|G_{\text{IJIVE1}}R\|_2 &\leq \ddot{r}_n^{1/2}, \\ \|G_{\text{IJIVE2}}R\|_2 &\leq 2\ddot{r}_n^{1/2}, \\ \|G_{\text{UJIVE}}R\|_2 &\leq \max_i (1 - (H_W)_{ii})^{-1} \ddot{r}_n^{1/2}, \\ \|G_{\text{TSLS}}R_Y\|_2 = \|G_{\text{JIVE1}}R_Y\|_2 = \|G_{\text{IJIVE1}}R_Y\|_2 &\leq \|\ddot{R}_Y\|_2, \\ \|G_{\text{IJIVE2}}R_Y\|_2 &\leq 2\|\ddot{R}_Y\|_2, \\ \|G_{\text{UJIVE}}R_Y\|_2 &\leq \max_i (1 - (H_W)_{ii})^{-1} \|\ddot{R}_Y\|_2. \end{aligned} \tag{18}$$

Furthermore,

$$G'_{\text{TSLS}}(R_Y, R) = (\ddot{R}, \ddot{R}_Y)$$

$$
\begin{aligned}
G_{\text{IJIVE2}}(R_Y, R) &= (I - D_{\ddot{Z}} + H_W D_{\ddot{Z}})(\ddot{R}_Y, \ddot{R}) \\
G'_{\text{IJIVE1}}(R_Y, R) &= (\ddot{R}_Y, \ddot{R}) - (I - H_Q) D_{\ddot{Z}} (I - D_{\ddot{Z}})^{-1} (\ddot{R}_Y, \ddot{R}), \\
G'_{\text{JIVE1}}(R_Y, R) &= (H_Q - D_Q)(I - D_Q)^{-1} (\ddot{R}_Y, \ddot{R}) \\
G'_{\text{UJIVE}}(R_Y, R) &= (\ddot{R}_Y, \ddot{R}) + \left[(I - H_W) D_W (I - D_W)^{-1} - (I - H_Q) D_Q (I - D_Q)^{-1} \right] (R_Y, R).
\end{aligned}
$$

By similar arguments as above,

$$
\begin{aligned}
\|G'_{\text{TSLS}} R\|_2 &\leq \ddot{r}_n^{1/2}, \qquad G'_{\text{TSLS}} R_Y{}_2 \leq \|R_Y\|_2, \\
\|G'_{\text{IJIVE2}} R\|_2 &\leq 2 \ddot{r}_n^{1/2}, \qquad G'_{\text{IJIVE2}} R_Y{}_2 \leq 2 \|R_Y\|_2, \\
\|G'_{\text{JIVE1}} R\|_2 &\leq 2 \max_i (1 - H_Q)_{ii}^{-1} \ddot{r}_n^{1/2}, \qquad \|G'_{\text{JIVE1}} R_Y\|_2 \leq 2 \max_i (1 - H_Q)_{ii}^{-1} \|\ddot{R}_Y\|_2, \\
\|G'_{\text{IJIVE1}} R\|_2 &\leq 2 \max_i (1 - H_{\ddot{Z}})_{ii}^{-1} \ddot{r}_n^{1/2} \qquad \|G'_{\text{IJIVE1}} R_Y\|_2 \leq 2 \max_i (1 - H_{\ddot{Z}})_{ii}^{-1} \|\ddot{R}_Y\|_2, \\
\|G'_{\text{UJIVE}} R\|_2 &\leq 2 \max_i \frac{(H_Q)_{ii}^{1/2}}{1 - (H_Q)_{ii}} \max_i |R_i| (L + K)^{1/2} + \ddot{r}_n^{1/2} \\
\|G'_{\text{UJIVE}} R_Y\|_2 &\leq 2 \max_i \frac{(H_Q)_{ii}^{1/2}}{1 - (H_Q)_{ii}} \max_i |R_{Y,i}| (L + K)^{1/2} + \|\ddot{R}_Y\|_2
\end{aligned}
\tag{19}
$$

where the last two lines follow since

$$
\begin{aligned}
\|G'_{\text{UJIVE}} R\|_2 &\leq \|\ddot{R} - G_{\text{UJIVE}} R\|_2 + \|\ddot{R}\|_2 \\
&\leq \max_i \frac{(H_W)_{ii}^{1/2}}{1 - (H_W)_{ii}} \|D_W^{1/2} R\|_2 + \max_i \frac{(H_Q)_{ii}^{1/2}}{1 - (H_Q)_{ii}} \|D_Q^{1/2} R\|_2 + \|\ddot{R}\|_2 \\
&\leq \max_i \frac{(H_W)_{ii}^{1/2}}{1 - (H_W)_{ii}} \max_i |R_i| L^{1/2} + \max_i \frac{(H_Q)_{ii}^{1/2}}{1 - (H_Q)_{ii}} \max_i |R_i| (L + K)^{1/2} + \|\ddot{R}\|_2 \\
&\leq 2 \max_i \frac{(H_Q)_{ii}^{1/2}}{1 - (H_Q)_{ii}} \max_i |R_i| (L + K)^{1/2} + \|\ddot{R}\|_2.
\end{aligned}
$$

Similar argument applies to the bound for $\|G'_{\text{UJIVE}} R_Y\|_2$.

Second, we bound the norm $\|G\|_F$. Using the triangle inequality, and the fact that for any projection matrix P and matrix A, $\|PA\|_F \leq \|A\|_F$,

$$
\begin{aligned}
\|G_{\text{TSLS}}\|_F &= K^{1/2} \\
\|G_{\text{IJIVE2}}\|_F &\leq \|H_{\ddot{Z}} - D_{\ddot{Z}}\|_F \leq K^{1/2} \\
\|G_{\text{IJIVE1}}\|_F &\leq \|(I - D_{\ddot{Z}})^{-1} (H_{\ddot{Z}} - D_{\ddot{Z}})\|_F \leq \max_i (1 - (H_{\ddot{Z}})_{ii})^{-1} K^{1/2} \\
\|G_{\text{JIVE1}}\|_F &\leq \|(I - D_Q)^{-1} (H_Q - D_Q)\|_F \leq \max_i (1 - (H_Q)_{ii})^{-1} (K + L)^{1/2} \\
\|G_{\text{UJIVE}}\|_F &\leq 2 \max_i (1 - (H_Q)_{ii})^{-1} (K + L)^{1/2}
\end{aligned}
\tag{20}
$$

Third, we bound the sum $\sum_i |G_{ii}|$. It follows by direct calculation that the diagonal elements G_{ii} for different estimators are given by

$$\begin{aligned}
G_{\text{TSLS},ii} &= (H_{\ddot{Z}})_{ii} \\
G_{\text{IJIVE2},ii} &= 2(H_W)_{ii}(H_{\ddot{Z}})_{ii} - \sum_{j=1}^{n}(H_W)_{ij}^2(H_{\ddot{Z}})_{jj}, \\
G_{\text{IJIVE1},ii} &= 2\frac{(H_W)_{ii}(H_{\ddot{Z}})_{ii}}{1-(H_{\ddot{Z}})_{ii}} - \sum_{j=1}^{n}\frac{(H_W)_{ij}^2(H_{\ddot{Z}})_{jj}}{1-(H_{\ddot{Z}})_{jj}} - (H_W(I-D_{\ddot{Z}})^{-1}D_{\ddot{Z}}H_{\ddot{Z}})_{ii}, \\
(G_{\text{JIVE1}})_{ii} &= \frac{(H_W)_{ii}(H_Q)_{ii}}{1-(H_Q)_{ii}} - (H_W(I-D_Q)^{-1}H_Q)_{ii} \\
(G_{\text{UJIVE}})_{ii} &= 0.
\end{aligned}$$

We now bound the sum $\sum_i |G_{ii}|$. Observe that by the Cauchy-Schwarz inequality, P is a projection matrix, and A a square matrix, $\sum_i |(PA)_i| \leq \|P\|_F \|PA\|_F$. Using this observation along with the triangle inequality, we obtain the bounds

$$\begin{aligned}
\sum_i |G_{\text{TSLS},ii}| &= K, \\
\sum_i |G_{\text{IJIVE2},ii}| &\leq 3\sum_i (H_W)_{ii}(H_{\ddot{Z}})_{ii} \leq L\max_i (H_{\ddot{Z}})_{ii}, \\
\sum_i |G_{\text{IJIVE1},ii}| &\leq 3\sum_i (H_W)_{ii}\frac{(H_{\ddot{Z}})_{ii}}{1-(H_{\ddot{Z}})_{ii}} + \|H_W\|_F\|H_W(I-D_{\ddot{Z}})^{-1}D_{\ddot{Z}}\|_F \\
&\leq 4L\max_i \frac{(H_{\ddot{Z}})_{ii}}{1-(H_{\ddot{Z}})_{ii}}, \\
\sum_i |(G_{\text{JIVE1}})_{ii}| &\leq 2L\max_i (1-(H_W)_{ii})^{-1}, \\
\sum_i |(G_{\text{UJIVE}})_{ii}| &= 0.
\end{aligned} \tag{21}$$

Finally, we bound the norm $\|G'G\|_F$. To this end, let $M = (I - H_W)$. By triangle inequality and arguments as above,

$$\begin{aligned}
\|G'_{\text{TSLS}}G_{\text{TSLS}}\|_F &= \|H_{\ddot{Z}}\|_F = K^{1/2}, \\
\|G'_{\text{IJIVE2}}G_{\text{IJIVE2}}\|_F &= \|H_{\ddot{Z}} - H_{\ddot{Z}}D_{\ddot{Z}}M - MD_{\ddot{Z}}H_{\ddot{Z}} + MD_{\ddot{Z}}MD_{\ddot{Z}}M\|_F \\
&\leq 3\|H_{\ddot{Z}}\|_F + \|MD_{\ddot{Z}}\|_F \leq 4K^{1/2}, \\
\|G'_{\text{IJIVE1}}G_{\text{IJIVE1}}\|_F &= \|M(H_{\ddot{Z}} - D_{\ddot{Z}})(I - D_{\ddot{Z}})^{-1}M(I - D_{\ddot{Z}})^{-1}(H_{\ddot{Z}} - D_{\ddot{Z}})M\|_F \\
&\leq \|(H_{\ddot{Z}} - 2D_{\ddot{Z}}H_{\ddot{Z}} + D^2_{\ddot{Z}})(I - D_{\ddot{Z}})^{-1}M(I - D_{\ddot{Z}})^{-1}\|_F \\
&\leq \max_i(1 - (D_{\ddot{Z}})_{ii})^{-2}\|(H_{\ddot{Z}} - 2D_{\ddot{Z}}H_{\ddot{Z}} + D^2_{\ddot{Z}})\|_F \\
&\leq 4K^{1/2}(\max_i(1 - (H_{\ddot{Z}})_{ii})^{-1})^2, \\
\|G'_{\text{JIVE1}}G_{\text{JIVE1}}\|_F &= \|(H_Q - 2D_QH_Q + D^2_Q)(I - D_Q)^{-1}M(I - D_Q)^{-1}\|_F \qquad (22) \\
&\leq \max_i(1 - (H_Q)_{ii})^{-2}\|(H_Q - 2D_QH_Q + D^2_Q)\|_F \\
&\leq 4\sqrt{K+L}(\max_i(1 - (H_Q)_{ii})^{-1})^2, \\
\|G'_{\text{UJIVE}}G_{\text{UJIVE}}\|_F &\leq \max_i(1 - (H_Q)_{ii})^{-2}\|H_Q - 2D_QH_Q + D^2_Q\|_F \\
&\qquad \max_i(1 - (H_W)_{ii})^{-2}\|H_W - 2D_WH_W + D^2_W\|_F \\
&\qquad \max_i(1 - (H_Q)_{ii})^{-2}\|H_W - H_QD_W - D_QH_W + D_WD_Q\|_F \\
&\leq 12\sqrt{K+L}\max_i(1 - (H_Q)_{ii})^{-2}.
\end{aligned}$$

Appendix B Proof of Lemma 4.1

We prove the results for a general class of estimators of the form given in Equation (16). We assume that $(G(R_Y, R))_i = (G(R_Y, R))_j$ whenever $z_i = z_j$ and $w_i = w_j$, which holds for all estimators considered in the statement of the Lemma 4.1. Let $A_i = (GR)_i$. Then we can write $A_i = A(Z_i, V_i, T_i)$. Also, let $\Delta^m(j) = (z^m_j - z^m_{j-1})'\pi$.

Using the definition of $\alpha(\cdot, \cdot)$ and recursion, we have

$$\pi'_Y z^m_k = \pi'_Y z^m_0 + \sum_{j=1}^{k} \alpha(z^m_j, z^m_{j-1})\Delta^m(j).$$

Therefore, we can write the numerator of the conditional estimand, $R_Y' GR$, as

$$\begin{aligned}\sum_i R_{Y,i}A_i &= \psi_Y' W'GR + \sum_{g,t}\sum_{k=0}^{J_g}\sum_{i=1}^{n} \mathbb{I}\left\{G_i = g, T_i = t, Z_i = z_k^g\right\} \pi_Y' z_k^g A(z_k^g, g, t) \\ &= \psi_Y' W'GR + \sum_g \pi_Y' z_0^g \sum_i \mathbb{I}\left\{G_i = g\right\} A_i \\ &\quad + \sum_{g=1}^{L_V}\sum_{k=1}^{J_g}\sum_{j=1}^{k} \alpha(z_j^g, z_{j-1}^g)\Delta^g(j) \sum_{i=1}^{n} \mathbb{I}\left\{G_i = g, Z_i = z_k^g\right\} A_i\end{aligned}$$

Changing the order of summation and rearranging then yields

$$\begin{aligned}\sum_i R_{Y,i}A_i &= \psi_Y' W'GR + \sum_{g=1}^{L_V} \pi_Y' z_0^g \sum_i \mathbb{I}\left\{G_i = g\right\} A_i \\ &\quad + \sum_{g=1}^{L_V}\sum_{j=1}^{J_g} \alpha(z_j^g, z_{j-1}^g)\Delta^g(j) \sum_{k=j}^{J_g}\sum_{i=1}^{n} \mathbb{I}\left\{G_i = g, Z_i = z_k^g\right\} A_i \\ &= \psi_Y' W'GR + \sum_{g=1}^{L_V} \pi_Y' z_0^g \sum_i \mathbb{I}\left\{G_i = g\right\} A_i \\ &\quad + \sum_{g=1}^{L_V}\sum_{j=1}^{J_g} \alpha(z_j^g, z_{j-1}^g)\Delta^g(j) \sum_i \mathbb{I}\left\{G_i = g, Z_i \geq z_j^g\right\} A_i.\end{aligned}$$

By similar arguments, we can write the denominator as

$$\sum_i R_i A_i = \psi' W'GR + \sum_{g=1}^{L_V} \pi' z_0^g \sum_i \mathbb{I}\left\{G_i = g\right\} A_i + \sum_{g=1}^{L_V}\sum_{j=1}^{J_g} \Delta^g(j) \sum_i \mathbb{I}\left\{G_i = g, Z_i \geq z_j^g\right\} A_i.$$

Note that $W'GR = 0$ implies $\sum_{i=1}^{n} \mathbb{I}\left\{G_i = m\right\} A_i = 0$ for all m. This condition holds for all estimators considered in the statement of Lemma 4.1, except UJIVE. Under this condition, the conditional estimand equals

$$\frac{\sum_{g=1}^{L_V}\sum_{j=1}^{J_g} \Delta^g(j)\frac{1}{n}\sum_i \mathbb{I}\left\{G_i = g, Z_i \geq z_j^g\right\} A_i \alpha(z_j^g, z_{j-1}^g)}{\sum_{g=1}^{L_V}\sum_{j=1}^{J_g} \Delta^g(j)\frac{1}{n}\sum_i \mathbb{I}\left\{G_i = g, Z_i \geq z_j^g\right\} A_i}. \tag{23}$$

Since by equation (17) $A_i = \ddot{R}_i$ for TSLS, JIVE1, and IJIVE1, and $A_i = (1-(H_{\ddot{Z}})_{ii})\ddot{R}_i + e_i' H_W \operatorname{diag}(H_{\ddot{Z}})\ddot{R}$ for IJIVE2, Part (i) follows. The first statement in Part (ii) is immediate. It therefore remains to show the result for UJIVE, for which $A_i = (1 - (H_W)_{ii})^{-1}\ddot{R}_i$, which, if the only covariates are group dummies can be written as $A_i = \frac{n_{G_i}}{n_{G_i}-1}(R_i - n_{G_i}^{-1}\sum_{j=1}^{n} \mathbb{I}\left\{G_j = G_i\right\} R_j)$. This implies $W' G_{\text{UJIVE}} R = 0$, which in turn implies (23), which yields the result.

Appendix C Proofs of conditional results

This section proves Theorem 5.1, Theorem 5.3, and Theorem 5.5. We prove a.s. convergence results below, but if the relevant assumptions are only assumed to hold w.p.a.1, the results and proofs below will hold w.p.a.1.

C.1 Proof of Theorem 5.1

To prove the result, we apply Lemma D.4 to each estimator. Condition (i) of Lemma D.4 holds for all estimators by Assumption 3 (i). Next, it follows from equation (18) and Assumption 2 (ii) that for all estimators $\|GR\|_2$ and $\|GR_Y\|_2$ are of the order $\ddot{r}_n^{1/2}$. Similarly, it follows from equation (19), Assumption 2 (ii), and assumptions of the Theorem that $\|G'R\|_2$ and $\|G'R_Y\|_2$ are also of the order $\ddot{r}_n^{1/2}$, and of the order $\sqrt{L+K}+\ddot{r}_n^{1/2}$ for UJIVE, so that condition (ii) of Lemma D.4 holds for all estimators. Furthermore, it follows from equation (20), Assumption 2 (iii), and assumptions of the Theorem that $\|G\|_F/\ddot{r}_n \overset{a.s.}{\to} 0$ for all estimators, so that condition (iii) of Lemma D.4 holds also. Now, for estimators other than IJIVE2 and UJIVE, $R'GR = \ddot{R}'\ddot{R}$, and $R_Y'GR = \ddot{R}_Y'\ddot{R}$. Condition (iv) of Lemma D.4 therefore holds for these estimators by Assumption 2 (ii) and the Cauchy-Schwarz inequality. For IJIVE2, $R'G_{\text{IJIVE2}}R = \ddot{R}'(I - D_{\ddot{Z}})\ddot{R}$, and $R_Y'G_{\text{IJIVE2}}R = \ddot{R}_Y'(I - D_{\ddot{Z}})\ddot{R}$, so that Assumption 2 (ii) holds by similar arguments and the fact that $(H_{\ddot{Z}})_{ii} \leq C < 1$ by assumption. For UJIVE, it follows from (17) that

$$|R_Y'G_{\text{UJIVE}}R - \ddot{R}_Y'\ddot{R}| = |R_Y D_W (I - D_W)^{-1}\ddot{R}| \leq \max_i |R_{Y,i}| \max_i \frac{(H_W)_{ii}^{1/2}}{1-(H_W)_{ii}} \sum_i |(H_W)_{ii}^{1/2}\ddot{R}_i|$$
$$\leq \max_i |R_{Y,i}| \max_i \frac{(H_W)_{ii}^{1/2}}{1-(H_W)_{ii}} L^{1/2}\ddot{r}_n^{1/2},$$

which is of the order $o(\ddot{r}_n)$ almost surely by assumption of the theorem. By analogous argument,

$$|R_Y'G_{\text{UJIVE}}R - \ddot{R}_Y'\ddot{R}| \leq \max_i |R_i| \max_i \frac{(H_W)_{ii}^{1/2}}{1-(H_W)_{ii}} L^{1/2}\ddot{r}_n^{1/2}. \tag{24}$$

Hence, by the preceding argument, Assumption 2 (ii) holds for UJIVE as well. Finally, condition (v) of Lemma D.4 holds by the bounds in Equation (21). If the estimator is inconsistent, we also need to make sure that $\ddot{r}_n/(\ddot{r}_n + \sum_i G_{ii}\sigma_{\eta,i}^2)$ is bounded for each estimator. Since $G_{ii,\text{TSLS}} \geq 0$, this holds trivially for TSLS. For other estimators, it follows by the assumption that the bias is bounded.

C.2 Proof of Theorem 5.3

To prove the result, we apply Lemma D.5 to each estimator. Condition (i) of Lemma D.5 holds for each estimator by Assumption 3 (i)–(ii), Condition (ii) of Lemma D.5 follows since by equation (21) and assumption of the theorem, $\sum_i |G_{ii}|/\ddot{r}_n^{1/2} \overset{a.s.}{\to} 0$ for all estimators. We have $\|G_{\text{TSLS}}R\|_2^2 = \|G_{\text{JIVE1}}R\|_2^2 =$

$\ddot{r}_n$. Since $G_{\text{IJIVE2}}R = \ddot{R} - (I - H_W)D_{\ddot{Z}}\ddot{R}$, we have, by the triangle inequality

$$\|G_{\text{IJIVE2}}R\|_2 \geq \|\ddot{R}\| - \|(I - H_W)D_{\ddot{Z}}\ddot{R}\|_2 \geq \|\ddot{R}\| - \|D_{\ddot{Z}}\ddot{R}\|_2 \geq (1 - \max_i (H_{\ddot{Z}})_{ii})\ddot{r}_n^{1/2}.$$

For UJIVE,

$$\|G_{\text{UJIVE}}R\|_2^2 = \sum_i (1 - H_W)_{ii}^{-2}\ddot{R}_i^2 \geq \sum_i \ddot{R}_i^2 = \ddot{r}_n.$$

Thus, condition (iii) of Lemma D.5 holds for all estimators. Next, note that for TSLS and IJIVE1 $GR = \ddot{R}$, so that $\sum_i (GR)_i^4/\ddot{r}_n^2 = \sum_i \ddot{R}_i^4/\ddot{r}_n^2 \overset{a.s.}{\to} 0$ by Assumption 5. For IJIVE2,

$$G_{\text{IJIVE2}}R = (I - (I - H_W)D_{\ddot{Z}})\ddot{R}.$$

Since for a projection matrix P, $\|Pa\|_2 \leq \|a\|$, we obtain the bound $\|G_{\text{IJIVE2}}R - \ddot{R}\|_2 \leq \max_i (H_{\ddot{Z}})_{ii}\ddot{r}_n^{1/2}$. Consequently, by Loève's c_r-inequality, and the ℓ_p-norm inequality $\|a\|_4 \leq \|a\|_2$, $\sum_i (G_{\text{IJIVE2}}R)_i^4/\ddot{r}_n^2 \leq 8\sum_i \ddot{R}_i^4/\ddot{r}_n^2 + 8\|G_{\text{IJIVE2}}R - \ddot{R}\|_2^4/\ddot{r}_n^2 \overset{a.s.}{\to} 0$. For UJIVE, observe that $\sum_i (G_{\text{UJIVE}}R)_i^4/\ddot{r}_n^2 \leq \max_i (1 - (H_W)_{ii})^{-4}\sum_i \ddot{R}_i^4/\ddot{r}_n^2 \overset{a.s.}{\to} 0$ by Assumption 5 and assumption of the theorem. Since G_{IJIVE2} and G_{TSLS} are symmetric, the same argument, together with Assumption 2 (ii), implies that $\sum_i (G'_{\text{TSLS}}R_\Delta)_i^4/\ddot{r}_n^2 \overset{a.s.}{\to} 0$, and $\sum_i (G'_{\text{IJIVE2}}R_\Delta)_i^4/\ddot{r}_n^2 \overset{a.s.}{\to} 0$. For IJIVE1,

$$G'_{\text{IJIVE1}}R_\Delta = \ddot{R}_\Delta - (I - H_W - H_{\ddot{Z}})D_{\ddot{Z}}(I - D_{\ddot{Z}})^{-1}\ddot{R}_\Delta,$$

so that $\|G'_{\text{IJIVE1}}R_\Delta - \ddot{R}_\Delta\|_2 \leq \max_i (H_{\ddot{Z}})_{ii}/(1 - (H_{\ddot{Z}})_{ii})\|\ddot{R}_\Delta\|_2$. Thus, by the previous arguments, we have $\sum_i (G'_{\text{IJIVE1}}R_\Delta)_i^4/\ddot{r}_n^2 \leq 8\sum_i \ddot{R}_{\Delta,i}^4/\ddot{r}_n^2 + 8\|G'_{\text{IJIVE2}}R_\Delta - \ddot{R}_\Delta\|_2^4/\ddot{r}_n^2 \overset{a.s.}{\to} 0$. For UJIVE, using arguments as in (19),

$$\begin{aligned}\|G'_{\text{UJIVE}}R\|_4/\ddot{r}_n &\leq \|\ddot{R}\|_4/\ddot{r}_n^{1/2} + \|G'_{\text{UJIVE}}R - \ddot{R}\|_4/\ddot{r}_n^{1/2} \leq o_{a.s.}(1) + \|G'_{\text{UJIVE}}R - \ddot{R}\|_2/\ddot{r}_n^{1/2} \\ &\leq o_{a.s.}(1) + 2\max_i \frac{(H_Q)_{ii}^{1/2}}{(1 - H_Q)_{ii}} \max_i |R_i|(L + K)^{1/2}/\ddot{r}_n^{1/2}.\end{aligned}$$

which converges to zero almost surely by the assumption of the theorem. Condition (iv) of Lemma D.5 therefore holds for all estimators. Finally, condition (v) of Lemma D.5 follows from equation (22) and Assumption 2 (iii). Therefore, for all estimators, equation (31) holds. To complete the proof, it remains to show that for IJIVE1 IJIVE2, and UJIVE, the variance expression $\widetilde{\mathcal{V}}_G/(R'GR)^2$ given in equation (32), is asymptotically equivalent to

$$\frac{\mathcal{V}_{\mathrm{C}} + \mathcal{V}_{\mathrm{MW},G}}{\ddot{r}_n}, \quad \mathcal{V}_{\mathrm{MW},G} = \frac{1}{\ddot{r}_n}\sum_{i \neq j}[G_{ij}^2\sigma_{\eta,j}^2\sigma_{\nu,i}^2 + G_{ij}G_{ji}\sigma_{\nu\eta,i}\sigma_{\nu\eta,j}],$$

and that if $K/\ddot{r}_n$ is bounded, we can also replace $\mathcal{V}_{\mathrm{MW},G}$ in the display above by $\mathcal{V}_{\mathrm{MW}}$, in the sense that $(\mathcal{V}_{\mathrm{C}} + \mathcal{V}_{\mathrm{MW},G})/\ddot{r}_n \cdot (R'GR)^2/\widetilde{\mathcal{V}}_G \overset{p}{\to} 1$, and, in the latter case, $(\mathcal{V}_{\mathrm{C}} + \mathcal{V}_{\mathrm{MW}})/\ddot{r}_n \cdot (R'GR)^2/\widetilde{\mathcal{V}}_G \overset{p}{\to} 1$. Since $\widetilde{\mathcal{V}}_G/\ddot{r}_n$ is bounded away from zero by proof of Lemma D.5, this is equivalent to showing that:

$(R'GR)/\ddot{r}_n \xrightarrow{p} 1$; and $\widetilde{\mathcal{V}}_G/\ddot{r}_n - (\mathcal{V}_{\mathrm{C}} + \mathcal{V}_{\mathrm{MW},G}) = o_P(1)$, or, if $K/\ddot{r}_n$ is bounded $\widetilde{\mathcal{V}}_G/\ddot{r}_n - (\mathcal{V}_{\mathrm{C}} + \mathcal{V}_{\mathrm{MW}}) = o_P(1)$. The first condition holds trivially for IJIVE1. For IJIVE2, it follows from the fact that $|R'G_{\text{IJIVE2}}R - \ddot{r}_n|/\ddot{r}_n = \ddot{R}'D_{\ddot{Z}}\ddot{R}/\ddot{r}_n \leq \max_i (H_{\ddot{Z}})_{ii} \xrightarrow{a.s.} 0$. For UJIVE, it follows from (24).

Since the terms $\sigma^2_{\nu,i}, \sigma^2_{\eta,i}$, and $\sigma_{\nu\eta,i}$ are bounded, the second condition holds if we can show that for IJIVE1, IJIVE2, and UJIVE, $\|\ddot{R} - GR\|_2/\ddot{r}_n^{1/2} \xrightarrow{p} 0$, $\|\ddot{R}_\Delta - G'R_\Delta\|_2/\ddot{r}_n^{1/2} \xrightarrow{p} 0$, and, if $K/\ddot{r}_n$ is bounded, $\sum_{i\neq j}((H_{\ddot{Z}})_{ij} - G_{ij})^2/\ddot{r}_n = o_P(1)$. The first two convergence results have been shown to hold earlier in this proof, so it remains to verify that $\sum_{i\neq j}((H_{\ddot{Z}})_{ij} - G_{ij})^2/\ddot{r}_n = o_P(1)$. Letting $M = I - H_W$, for IJIVE1, the left-hand side can be bounded by

$$\begin{aligned}\|H_{\ddot{Z}} - G_{\text{IJIVE1}}\|_F^2/\ddot{r}_n &= \|M(I - D_{\ddot{Z}})^{-1}D_{\ddot{Z}}H_{\ddot{Z}} - M(I - D_{\ddot{Z}})^{-1}D_{\ddot{Z}}M\|^2/\ddot{r}_n \\ &\leq 4\|(I - D_{\ddot{Z}})^{-1}D_{\ddot{Z}}\|_F^2/\ddot{r}_n \leq 4\max_i \frac{(H_{\ddot{Z}})_{ii}}{(1 - (H_{\ddot{Z}})_{ii})^2}\frac{K}{\ddot{r}_n},\end{aligned}$$

which converges to zero by assumption. For IJIVE2, the left-hand side is bounded by

$$\|G - H_{\ddot{Z}}\|_F^2/\ddot{r}_n = \|(I - H_W)D_{\ddot{Z}}(I - H_W)\|_F^2/\ddot{r}_n \leq \|D_{\ddot{Z}}\|_F^2/\ddot{r}_n \leq \max_i (H_{\ddot{Z}})_{ii}\frac{K}{\ddot{r}_n} \xrightarrow{a.s.} 0.$$

For UJIVE,

$$\begin{aligned}\|H_{\ddot{Z}} - G_{\text{UJIVE}}\|_F^2/\ddot{r}_n &= \|(I - D_Q)^{-1}D_Q(I - H_Q) + (I - D_W)^{-1}D_W(I - H_W)\|_F^2/\ddot{r}_n \\ &\leq \frac{2}{\ddot{r}_n}\|(I - D_Q)^{-1}D_Q\|_F^2 + \frac{2}{\ddot{r}_n}\|(I - D_W)^{-1}D_W\|_F^2 \\ &\leq \max_i \frac{(H_Q)_{ii}}{(1 - (H_Q)_{ii})^2}\frac{(K + L)}{\ddot{r}_n} + \max_i \frac{(H_W)_{ii}}{(1 - (H_W)_{ii})^2}\frac{L}{\ddot{r}_n} \xrightarrow{a.s.} 0,\end{aligned}$$

which completes the proof.

C.3 Proof of Theorem 5.5

Put $M = I - H_Q$, and $\delta = \hat{\beta} - \beta_{\mathrm{C}}$, and $\ddot{Y}_\Delta = \ddot{Y} - \ddot{X}\beta$. Then

$$\hat{\sigma}^2_{\nu,i} = ((M\nu)_i + (M\eta)_i\delta)^2, \qquad \hat{\sigma}_{\nu\eta,i} = ((M\nu)_i + (M\eta)_i\delta)(M\eta)_i, \qquad \hat{\sigma}^2_{\eta,i} = (M\eta)_i^2,$$

and $\ddot{Y} - \ddot{X}\hat{\beta} = \ddot{Y}_\Delta + \ddot{X}\delta$. Since J is linear in its arguments, it follows by plugging in these expressions into the definition of $\widehat{\mathcal{V}}_{\mathrm{C}}$ that that variance estimator can be decomposed as we can therefore decompose elements of the variance estimator as

$$\begin{aligned}\widehat{\mathcal{V}}_{\mathrm{C}} = \hat{r}_n^{-1}\big\{&J(\ddot{X}, \ddot{X}, (M\nu) \odot (M\nu)) + J(\ddot{Y}_\Delta, \ddot{Y}_\Delta, (M\eta) \odot (M\eta)) + J(\ddot{Y}_\Delta, \ddot{X}, (M\nu) \odot (M\eta)) \\ &+ 3\delta\left(\delta J(\ddot{X}, \ddot{X}, (M\eta) \odot (M\eta)) + J(\ddot{X}, \ddot{X}, (M\nu) \odot (M\eta)) + J(\ddot{Y}_\Delta, \ddot{X}, (M\eta) \odot (M\eta))\right)\big\},\end{aligned}$$

where $\odot$ denotes element-wise (Hadamard) product.

By Theorem 5.1 and Lemma D.10 below, applied to each term, since $\delta = o_P(1)$,

$$\widehat{\mathcal{V}}_{\mathrm{C}} = \ddot{r}_n \mathcal{V}_{\mathrm{C}} + o_P(\ddot{r}_n).$$

Similarly, letting $S(a, b) = \hat{r}_n^{-1} \sum_{i \neq j} (H_{\ddot{Z}})_{ij}^2 a_j b_i$, we can write

$$\begin{aligned}\widehat{\mathcal{V}}_{\mathrm{MW}} = S(\hat{\sigma}_{\eta}^2, \hat{\sigma}_{\nu}^2) + S(\hat{\sigma}_{\eta\eta}^2, \hat{\sigma}_{\eta\nu}^2) \\ &= S((M\nu) \odot (M\nu), (M\eta) \odot (M\eta)) + S((M\nu) \odot (M\eta), (M\nu) \odot (M\eta)) \\ &\quad + 4\delta S((M\nu) \odot (M\eta), (M\eta) \odot (M\eta)) + 2\delta^2 S((M\eta) \odot (M\eta), (M\eta) \odot (M\eta)).\end{aligned}$$

By Lemma D.11 below, applied to each term,

$$\widehat{\mathcal{V}}_{\mathrm{MW}} = \ddot{r}_n \mathcal{V}_{\mathrm{MW}} + o_P(\ddot{r}_n).$$

Appendix D Auxiliary results

D.1 Auxiliary results for quadratic forms

For the results in this subsection, let

$$Q = u't + s'v + \sum_{i \neq j} P_{ij} u_i v_j,$$

where, conditional on on some set of variables $\mathcal{Z}_n$, the matrix $P \in \mathbb{R}^{n \times n}$ and vectors s, t are non-random, and the elements (u_i, v_i) of vectors u, v are mean zero, and independent across i. Let $E_{\mathcal{Z}_n}$ denote the expectation conditional on $\mathcal{Z}_n$. We will prove a law of large numbers and a central limit theorem for Q.

Lemma D.1. *Suppose that conditional on $\mathcal{Z}_n$, the second moments of (u_i, v_i) are bounded a.s. Suppose further that $\|t\|_2 + \|s\|_2 + \|P\|_F \overset{a.s.}{\to} 0$. Then $Q = o_P(1)$.*

Proof. Since $E_{\mathcal{Z}_n}[Q] = 0$, its variance is given by

$$\operatorname{var}(Q \mid \mathcal{Z}_n) = \sum_i E_{\mathcal{Z}_n}[u_i t_i + v_i s_i]^2 + \sum_{i \neq j} \left(P_{ij}^2 E_{\mathcal{Z}_n}[u_i^2 v_j^2] + P_{ij} P_{ji} E_{\mathcal{Z}_n}[u_i v_i v_j u_j] \right).$$

Since the second moments are bounded, it follows that

$$\operatorname{var}(Q \mid \mathcal{Z}_n) \preceq_{\text{a.s.}} \|t\|_2^2 + \|s\|_2^2 + \sum_{i \neq j} P_{ij}^2 \leq \|t\|_2^2 + \|s\|_2^2 + \|P\|_F^2.$$

Since the right-hand side converges to zero almost surely by assumption, the result follows by Markov inequality and dominated convergence theorem. □

Lemma D.2. *Suppose that, conditional on $\mathcal{Z}_n$, the fourth moments of (u_i, v_i) are bounded a.s. Suppose further that:*

1. $\operatorname{var}(Q \mid \mathcal{Z}_n)^{-1/2}$ *is bounded a.s.*
2. $\sum_{i=1}^n t_i^4 + \sum_{i=1}^n s_i^4 \overset{a.s.}{\to} 0$
3. $\|P_L P_L'\|_F + \|P_U' P_U\|_F \overset{a.s.}{\to} 0$*, where P_L is a lower-triangular matrix with elements $P_{L,ij} = P_{ij}\mathbb{I}\{i > j\}$ and P_U is an upper-triangular matrix with elements $P_{U,ij} = P_{ij}\mathbb{I}\{i < j\}$.*

Then

$$\operatorname{var}(Q \mid \mathcal{Z}_n)^{-1/2} Q \overset{d}{\to} \mathcal{N}(0,1).$$

Proof. Let $B = \operatorname{var}(Q \mid \mathcal{Z}_n)^{-1/2}$. Then we can write $BQ = \sum_{i=1}^n B y_i$, where

$$y_i = u_i t_i + v_i s_i + u_i \sum_{j=1}^{i-1} P_{ij} v_j + v_i \sum_{j=1}^{i-1} P_{ji} u_j = u_i t_i + v_i s_i + u_i (P_L v)_i + v_i (P_U' u)_i.$$

Conditional on $\mathcal{Z}_n$, By_i is a martingale difference array with respect to the filtration $\mathcal{F}_{in} = \sigma(u_1, v_1, \dots, u_{i-1}, v_{i-1})$. Since B is bounded by assumption, by the martingale central limit theorem, if for some $\epsilon > 0$,

$$\sum_{i=1}^n E_{\mathcal{Z}_n}[|y_i|^{2+\epsilon}] \overset{a.s.}{\to} 0, \tag{25}$$

and if the conditional variance converges to one, $P(|\sum_{i=1}^n E[B^2 y_i^2 \mid \mathcal{F}_{i,n}, \mathcal{Z}_n] - 1| > \eta \mid \mathcal{Z}_n) \overset{a.s.}{\to} 0$ for any η, then conditional on $\mathcal{Z}_n$, $\sum_{i=1}^n B y_i$ converges in distribution to $\mathcal{N}(0,1)$, and the result will follow by the dominated convergence theorem.

By Loève's c_r-inequality, if

$$\sum_{i=1}^n E_{\mathcal{Z}_n}[u_i^4] t_i^4 + \sum_{i=1}^n E_{\mathcal{Z}_n}[v_i^4] s_i^4 \overset{a.s.}{\to} 0, \tag{26}$$

and if

$$\sum_{i=1}^n E_{\mathcal{Z}_n} u_i^4 (P_L v)_i^4 + \sum_{i=1}^n E_{\mathcal{Z}_n} v_i^4 (P_U' u)_i^4 \overset{a.s.}{\to} 0, \tag{27}$$

then (25) holds with $\epsilon = 2$. Now, equation (26) follows from condition 2. To verify equation (27), note that the first sum can be bounded as

$$\begin{aligned}\sum_{i=1}^n E_{\mathcal{Z}_n} u_i^4 (P_L v)_i^4 &= \sum_{i=1}^n E_{\mathcal{Z}_n}[u_i^4] E_{\mathcal{Z}_n}[(P_L v)_i^4] \preceq_{\text{a.s.}} \sum_{i=1}^n E_{\mathcal{Z}_n}[(P_L v)_i^4] \\ &= \sum_{i=1}^n \sum_{j=1}^n P_{L,ij}^4 E_{\mathcal{Z}_n}[v_j^4] + 3 \sum_{i=1}^n \sum_{j \neq k} P_{L,ij}^2 P_{L,ik}^2 E_{\mathcal{Z}_n}[v_j^2] E_{\mathcal{Z}_n}[v_k^2] \preceq_{\text{a.s.}} \sum_{i,j,k} P_{L,ij}^2 P_{L,ik}^2.\end{aligned}$$

Now,

$$\sum_{i,j,k} P_{L,ij}^2 P_{L,ik}^2 = \sum_{i=1}^n (P_L P_L')_{ii}^2 \leq \sum_{i=1}^n \sum_{j=1}^n (P_L P_L')_{ij}^2 = \|P_L P_L'\|_F^2.$$

By a symmetric argument, the second sum in (27) is of the order $\|P_U' P_U\|_F^2$, so that (27) holds by condition 3.

It remains to show convergence of the conditional variance. Let $W_i = u_i t_i + v_i s_i$, and let $X_i = u_i(P_L v)_i + v_i(P_U' u_i)$. Since $\operatorname{var}(BQ \mid \mathcal{Z}_n) = B^2 \sum_{i=1}^n E_{\mathcal{Z}_n}[W_i^2] + B^2 \sum_{i=1}^n E_{\mathcal{Z}_n}[X_i^2] = 1$, and since $E_{\mathcal{Z}_n}[W_i^2] = E[W_i^2 \mid \mathcal{Z}_n, \mathcal{F}_{in}]$, we have

$$\sum_{i=1}^n E[B^2 y_i^2 \mid \mathcal{F}_{in}, \mathcal{Z}_n] - 1 = \sum_{i=1}^n B^2 \left(E[X_i^2 \mid \mathcal{F}_{in}, \mathcal{Z}_n] - E[X_i^2 \mid \mathcal{Z}_n]\right) + 2B^2 \sum_{i=1}^n E[W_i X_i \mid \mathcal{F}_{in}, \mathcal{Z}_n]. \tag{28}$$

We show that both of the terms on the right-hand side converge to zero. The second sum can be written as

$$B^2 \sum_{i=1}^n E[W_i X_i \mid \mathcal{F}_{in}, \mathcal{Z}_n] = B^2 \sum_{i=1}^n (P_L v)_i E_{\mathcal{Z}_n}[W_i u_i] + B^2 \sum_{i=1}^n (P_U' u)_i E_{\mathcal{Z}_n}[W_i v_i] = \delta_u' P_L v + \delta_v' P_U' u,$$

where $\delta_{u,i} = B^2 E_{\mathcal{Z}_n}[W_i u_i]$ and $B^2 \delta_{v,i} = E_{\mathcal{Z}_n}[W_i v_i]$. Now, by Cauchy-Schwarz inequality and boundedness of second moments of v_i,

$$E_{\mathcal{Z}_n}[(\delta_u' P_L v)^2] \preceq_{\text{a.s.}} \delta_u' P_L P_L' \delta_u \leq \|\delta_u\|_2^2 \|P_L P_L'\|_2 \leq \|\delta_u\|_2^2 \|P_L P_L'\|_F \preceq_{\text{a.s.}} \|P_L P_L'\|_F$$

where the last inequality follows because by Cauchy-Schwarz inequality, $\delta_{u,i}^2 \leq B^4 E_{\mathcal{Z}_n}[W_i^2] E_{\mathcal{Z}_n}[u_i^2]$, so that $\|\delta_u\|_2^2 \preceq_{\text{a.s.}} B^2 \sum_i E_{\mathcal{Z}_n}[W_i^2] \leq \operatorname{var}(BQ) \leq 1$. By similar arguments, $E_{\mathcal{Z}_n}[(\delta_v P_U' u)^2] \preceq_{\text{a.s.}} \|P_U' P_U\|_F$. Thus, by condition 3 and Markov inequality, the second term in (28) a.s. converges to zero conditionally on $\mathcal{Z}_n$. The first term in (28) can be decomposed as

$$\sum_{i=1}^n \sigma_{u,i}^2 [(P_L v)_i^2 - E_{\mathcal{Z}_n}(P_L v)_i^2] + \sum_{i=1}^n \sigma_{v,i}^2 [(P_U' u)_i^2 - E_{\mathcal{Z}_n}(P_U' v)_i^2] \\ + 2 \sum_{i=1}^n \sigma_{vu,i} [(P_L v)_i (P_U' u)_i - E_{\mathcal{Z}_n}(P_L v)_i (P_U' u)_i].$$

Let D_u denote a diagonal matrix with elements $D_{u,ii} = \sigma_{u,i}^2$, and let $T = P_L' D_u P_L$. Then the first sum in the preceding display equals $v'Tv - E_{\mathcal{Z}_n}[v'Tv]$. The variance of this term can be bounded as

$$\begin{aligned} \operatorname{var}(v'Tv \mid \mathcal{Z}_n) &= \sum_i T_{ii}^2 \left(E_{\mathcal{Z}_n}[v_i^4] - 3E_{\mathcal{Z}_n}[v_i^2]^2\right) + \sum_i \sum_j (T_{ij}^2 + T_{ij} T_{ji}) E_{\mathcal{Z}_n}[v_i^2 v_j^2] \\ &\preceq_{\text{a.s.}} \sum_i \sum_j T_{ij}^2 = \|P_L' D_u P_L\|_F^2 \preceq_{\text{a.s.}} \|P_L P_L'\|_F^2. \end{aligned}$$

By similar arguments, the conditional variance of the second sum is of the order $\|P_U' P_U\|_F^2$ and the conditional variance of the third term is of the order $\|P_L P_L'\|_F^2 + \|P_U' P_U\|_F^2$. Thus, by condition 3 and Markov inequality, the first term in (28) a.s. converges to zero conditionally on $\mathcal{Z}_n$, which concludes the proof. □

The following result generalizes Lemma B.2 in Chao et al. (2012), and is used to verify condition 3 of Lemma D.2.

Lemma D.3. *Let $P = P_n$ be a sequence of random square matrices such that $\operatorname{tr}(P'PP'P) \overset{a.s.}{\to} 0$. Then $\|LL'\|_F^2 + \|UU'\|_F^2 \overset{a.s.}{\to} 0$, where $(L)_{ij} = P_{ij}\mathbb{I}\{i > j\}$ and $(U)_{ij} = P_{ij}\mathbb{I}\{j > i\}$.*

Proof. Note first that $\sum_{i,j,k} P_{ij}^2 P_{ik}^2 = \sum_i (PP')_{ii}^2 \leq \sum_{i,j}(PP')_{ij}^2 = \operatorname{tr}(PP'PP')$, and $\sum_{i,j,k} P_{ji}^2 P_{ki}^2 = \sum_i (P'P)_{ii}^2 \leq \sum_{i,j}(P'P)_{ij}^2 = \operatorname{tr}(PP'PP')$. Similarly, $\sum_{i,j} P_{ij}^4 \leq \sum_{i,j,k} P_{ij}^2 P_{ik}^2 \leq \operatorname{tr}(PP'PP')$. Using these observations, we get the bound

$$\begin{aligned}\|LL'\|_F^2 + \|UU'\|_F^2 &- 4\sum_{i<j<k<\ell}(P_{ik}P_{i\ell}P_{jk}P_{j\ell} + P_{ki}P_{\ell i}P_{kj}P_{\ell j}) \\ &= \sum_{i<j} P_{ij}^4 + 2\sum_{i<j<k}(P_{ij}^2 P_{ik}^2 + P_{ik}^2 P_{jk}^2) + \sum_{i<j} P_{ji}^4 + 2\sum_{i<j<k}(P_{ji}^2 P_{ki}^2 + P_{ki}^2 P_{kj}^2) \\ &\leq 6\operatorname{tr}(P'PP'P) \overset{a.s.}{\to} 0. \qquad (29)\end{aligned}$$

It therefore suffices to show that $\sum_{i<j<k<\ell}(P_{ik}P_{i\ell}P_{jk}P_{j\ell} + P_{ki}P_{\ell i}P_{kj}P_{\ell j}) \overset{a.s.}{\to} 0$. To that end, let $D = \operatorname{diag}(P)$, and observe first that by triangle inequality, we have,

$$\|(P-D)'(P-D)\|_F \leq \|P'P\|_F + \|D^2\|_F + 2\|DP\|_F \overset{a.s.}{\to} 0, \qquad (30)$$

since $\|D^2\|_F^2 = \sum_i P_{ii}^4 \leq \sum_i (PP')_{ii}^2 \overset{a.s.}{\to} 0$, and $\|DP\|_F^2 = \sum_i \sum_j P_{ij}^2 P_{ii}^2 \leq \sum_i \sum_j P_{ij}^2 \sum_k P_{ik}^2 = \sum_i (PP')_{ii}^2 \overset{a.s.}{\to} 0$. On the other hand, expanding the left-hand side and using the same argument as in (29) yields

$$\begin{aligned}\|(P-D)'(P-D)\|_F^2 &- 4S_n \\ &= \sum_{i<j}(P_{ji}^4 + P_{ij}^4) + 2\sum_{i<j<k}\left(P_{ki}^2 P_{kj}^2 + P_{ki}^2 P_{ji}^2 + P_{ik}^2 P_{jk}^2 + P_{ik}^2 P_{ij}^2 + P_{ji}^2 P_{jk}^2 + P_{ij}^2 P_{kj}^2\right) \overset{a.s.}{\to} 0,\end{aligned}$$

where $S_n = \sum_{i<j<k<\ell}(P_{ji}P_{ki}P_{k\ell}P_{j\ell} + P_{ji}P_{\ell i}P_{jk}P_{\ell k} + P_{ki}P_{\ell i}P_{kj}P_{\ell j} + P_{ij}P_{ik}P_{\ell k}P_{\ell j} + P_{ij}P_{i\ell}P_{kj}P_{k\ell} + P_{ik}P_{i\ell}P_{jk}P_{j\ell})$. Combining this with (30) yields that $S_n = O(G_n)$.

Let ϵ_i denote random random variables that, conditional on P, are independent, and have mean zero and unit variance. Define

$$\Delta_2 = \sum_{i<j<k}(P_{ij}P_{ik}\epsilon_j\epsilon_k + P_{ij}P_{jk}\epsilon_i\epsilon_k) \qquad \tilde{\Delta}_2 = \sum_{i<j<k}(P_{ji}P_{ki}\epsilon_j\epsilon_k + P_{ij}P_{jk}\epsilon_i\epsilon_k)$$

$$\Delta_3 = \sum_{i<j<k} P_{ik}P_{jk}\epsilon_i\epsilon_j \qquad\qquad \tilde{\Delta}_3 = \sum_{i<j<k} P_{ki}P_{kj}\epsilon_i\epsilon_j$$

and let $\Delta_1 = \Delta_2 + \Delta_3$ and $\tilde{\Delta}_1 = \tilde{\Delta}_2 + \tilde{\Delta}_3$. Observe that

$$\begin{aligned}
E[\Delta_3^2 \mid P] &= 2\sum_{i<j<k<\ell} P_{ik}P_{jk}P_{i\ell}P_{j\ell} + \sum_{i<j<k} P_{ik}^2 P_{jk}^2 \\
E[\tilde{\Delta}_3^2 \mid P] &= 2\sum_{i<j<k<\ell} P_{ki}P_{kj}P_{\ell i}P_{\ell j} + \sum_{i<j<k} P_{ki}^2 P_{kj}^2 \\
E[\Delta_2^2 + \tilde{\Delta}_2^2 \mid P] &= \sum_{i<j<k}(P_{ij}^2 P_{ik}^2 + P_{ij}^2 P_{jk}^2) + \sum_{i<j<k}(P_{ji}^2 P_{ki}^2 + P_{ji}^2 P_{kj}^2) + 2S_n \\
E[\Delta_1^2 + \tilde{\Delta}_1^2 \mid P] &= \|(P-D)'(P-D)\|_F^2 - \sum_{i<j}(P_{ij}^4 + P_{ji}^4) \overset{a.s.}{\to} 0.
\end{aligned}$$

Hence, $E[\Delta_3^2 + \tilde{\Delta}_3^2] - 2\sum_{i<j<k<\ell}(P_{ik}P_{jk}P_{i\ell}P_{j\ell} + P_{ki}P_{kj}P_{\ell i}P_{\ell j}) \overset{a.s.}{\to} 0$. On the other hand, by Loève's c_r-inequality,

$$E[\Delta_3^2 + \tilde{\Delta}_3^2] \leq 2E[\Delta_1^2 + \Delta_2^2 + \tilde{\Delta}_1^2 + \tilde{\Delta}_2^2] \preceq_{\text{a.s.}} 2S_n \overset{a.s.}{\to} 0,$$

so that $2\sum_{i<j<k<\ell}(P_{ik}P_{jk}P_{i\ell}P_{j\ell} + P_{ki}P_{kj}P_{\ell i}P_{\ell j}) \overset{a.s.}{\to} 0$, which proves the result. □

D.2 High-level theorems for conditional inference

Lemma D.4. *Consider an estimator $\hat{\beta}_G$. Suppose that*

(i) Conditional on Q, the reduced-form errors ζ_i and η_i are mean zero, independent across i, and with bounded second moments.

(ii) $(\|GR\|_2 + \|G'R\|_2 + \|GR_Y\|_2 + \|G'R_Y\|_2)/\ddot{r}_n \overset{a.s.}{\to} 0$.

(iii) $\|G\|_F/\ddot{r}_n \overset{a.s.}{\to} 0$.

(iv) $\ddot{r}_n/R'GR$ and $R_Y'GR/\ddot{r}_n$ are bounded a.s.

(v) $\sum_i |G_{ii}|/\ddot{r}_n \overset{a.s.}{\to} 0$

Then $\hat{\beta}_G \overset{p}{\to} \beta_G$. Furthermore, if condition (v) is replaced with the assumption that $\sum_i |G_{ii}|/\ddot{r}_n$, $\ddot{r}_n/(R'GR + \sum_i G_{ii}\sigma_{\eta,i}^2)$, and $E[\zeta_i^4 + \eta_i^4 \mid Q_i]$ are all bounded a.s., then $\hat{\beta}_G = \beta_G + \text{bias}_G + o_P(1)$, where

$$\text{bias}_G = \frac{\sum_i G_{ii}\sigma_{\nu(\beta_G)\eta,i}}{R'GR + \sum_i G_{ii}\sigma_{\eta,i}^2}.$$

Proof. By Lemma D.1 with $\mathcal{Z}_n = (Q_1, \ldots, Q_n)$, and $P = G/\ddot{r}_n$

$$Y'(G/\ddot{r}_n)X = R_Y'GR/\ddot{r}_n + \sum_{i=1}^n G_{ii}\zeta_i\eta_i/\ddot{r}_n + o_P(1),$$

$$X'(G/\ddot{r}_n)X = R'GR/\ddot{r}_n + \sum_{i=1}^{n} G_{ii}\eta_i^2/\ddot{r}_n + o_P(1).$$

If condition (v) holds, then by Markov inequality, for any $\epsilon > 0$, $P(|\sum_i G_{ii}\zeta_i\eta_i/\ddot{r}_n| > \epsilon \mid Q) \leq \sum_i |G_{ii}| E[|\zeta_i\eta_i| \mid Q]/\epsilon\ddot{r}_n \preceq_{\text{a.s.}} \sum_i |G_{ii}|/\epsilon\ddot{r}_n \xrightarrow{a.s.} 0$, so that by the dominated convergence theorem, $\sum_i G_{ii}\zeta_i\eta_i/\ddot{r}_n \xrightarrow{p} 0$. By similar arguments, $\sum_i G_{ii}\eta_i^2/\ddot{r}_n \xrightarrow{p} 0$. Thus, Condition (iv), $\hat{\beta}_G = (R_Y'GR/\ddot{r}_n + o_P(1))/(R'GR/\ddot{r}_n + o_P(1)) = \beta_G(1 + o_P(1)) + o_P(1) = \beta_G + o_P(1)$.

Otherwise, since $\operatorname{var}(\sum_i G_{ii}\eta_i^2/\ddot{r}_n \mid Q) \preceq_{\text{a.s.}} \sum_i G_{ii}^2/\ddot{r}_n^2 \leq \|G\|_F^2/\ddot{r}_n^2 \xrightarrow{a.s.} 0$, it follows by Markov inequality and dominated convergence theorem that $\sum_i G_{ii}\eta_i^2 = \sum_i G_{ii}\sigma_{\eta,i}^2 + o_P(1)$. By similar arguments $\sum_i G_{ii}\eta_i\zeta_i = \sum_i G_{ii}\sigma_{\zeta\eta,i} + o_P(1)$, and the result follows by the same argument as above.

□

Lemma D.5. *Consider an estimator of the form $\hat{\beta}_G$. Suppose that Conditions (i)–(iv) of Lemma D.4 hold, and that*

(i) $E[\nu_i^4 + \eta_i^4 \mid Q]$ is bounded, $|\operatorname{corr}(\nu_i, \eta_i \mid Q)|$ is bounded away from one, and $\sigma_{\nu,i}^2$ is bounded away from zero

(ii) $\sum_i |G_{ii}|/\ddot{r}_n^{1/2}$ is bounded a.s. and $\sum_i G_{ii}^2/\ddot{r}_n \xrightarrow{a.s.} 0$

(iii) $\ddot{r}_n/\|GR\|^2$ is bounded a.s.

(iv) $\sum_i (GR)_i^4/\ddot{r}_n^2 \xrightarrow{a.s.} 0$ and $\sum_i (G'R_\Delta)_i^4/\ddot{r}_n^2 \xrightarrow{a.s.} 0$

(v) $\|G'G\|_F/\ddot{r}_n \xrightarrow{a.s.} 0$.

Then

$$\frac{\hat{\beta}_G - \beta_G - \operatorname{bias}_G}{\sqrt{\widetilde{\mathcal{V}}_G/R'GR}} \xrightarrow{d} \mathcal{N}(0,1), \tag{31}$$

where

$$\widetilde{\mathcal{V}}_G = \sum_i [(G'R)_i^2\sigma_{\nu,i}^2 + \sigma_{\eta,i}^2(G'R_\Delta)_i^2 + 2\sigma_{\nu\eta,i}(G'R)_i(G'R_\Delta)_i] + \sum_{i\neq j}[G_{ij}^2\sigma_{\eta,j}^2\sigma_{\nu,i}^2 + G_{ij}G_{ji}\sigma_{\nu\eta,i}\sigma_{\nu\eta,j}]. \tag{32}$$

Condition (ii) ensures that $\sqrt{\ddot{r}_n}\operatorname{bias}_G$ is bounded. Conditions (i) and (iii) ensure that $\widetilde{\mathcal{V}}_G/\ddot{r}_n$ is bounded away from zero.

Proof of Lemma D.5. It follows from Lemma D.4 that $X'GX/R'GR = 1 + o_P(\ddot{r}_n)$. It follows from condition (ii) that $\sum_i G_{ii}\sigma_{\eta_i} = o_P(\ddot{r}_n)$. Therefore,

$$\frac{R'GR}{\ddot{r}_n^{1/2}}\left(\hat{\beta}_G - \beta_G - \operatorname{bias}_G\right) = \frac{(R_\Delta + \nu)'GX/\sqrt{\ddot{r}_n}}{X'GX/R'GR} - \frac{\sum_i G_{ii}\sigma_{\nu(\beta_G)\eta,i}/\sqrt{\ddot{r}_n}}{1 + \sum_i G_{ii}\sigma_{\eta,i}^2/R'GR}$$

$$= (R_\Delta + \nu)'GX/\sqrt{\ddot{r}_n}\,(1 + o_P(1)) - \sum_i G_{ii}\sigma_{\nu(\beta_G)\eta,i}/\sqrt{\ddot{r}_n}(1 + o_P(1)).$$

Furthermore, if follows from condition (ii), Markov inequality, and dominated convergence theorem that $\sum_i \nu_i\eta_i G_{ii}/\ddot{r}_n^{1/2} = \sum_i G_{ii}\sigma_{\nu\eta,i}/\ddot{r}_n^{1/2} + o_P(1)$. Thus,

$$\frac{\hat{\beta}_G - \beta_G - \text{bias}_G}{\widetilde{\mathcal{V}}_G^{1/2}/R'GR} = \frac{(R_\Delta' G\eta + \nu' GR + \sum_{i\neq j}\nu_i\eta_j G_{ij})/\sqrt{\ddot{r}_n}}{\sqrt{\widetilde{\mathcal{V}}_G/\ddot{r}_n}}\,(1 + o_P(1)) + \frac{\sum_i \sigma_{\eta\nu,i}G_{ii}/\sqrt{\ddot{r}_n}}{\sqrt{\widetilde{\mathcal{V}}_G/\ddot{r}_n}}\cdot o_P(1). \tag{33}$$

Since $\sum_i \sigma_{\eta\nu,i}G_{ii}/\sqrt{\ddot{r}_n}$ is bounded by condition (ii), to verify the claim, it suffices to show that the first term converges to a standard normal random variable, and that $\ddot{r}_n/\widetilde{\mathcal{V}}_G$ is bounded a.s. To that end, write $\widetilde{\mathcal{V}}_G$ as

$$\begin{aligned}
\widetilde{\mathcal{V}}_G &= \sum_i \operatorname{var}((G'R)_i\nu_i + (G'R_\Delta)_i\eta_i \mid Q) + \frac{1}{2}\sum_{i\neq j} E[(G_{ij}\nu_i\eta_j + G_{ji}\nu_j\eta_i)^2 \mid Q] \\
&\geq \sum_i \operatorname{var}((G'R)_i\nu_i + (G'R_\Delta)_i\eta_i \mid Q) \\
&\geq \sum_i (1 - |\operatorname{corr}((G'R)_i\nu_i, (G'R_\Delta)_i\eta_i \mid Q)|)(G'R)_i^2\sigma_{\nu,i}^2,
\end{aligned}$$

where the last line uses the fact that for any two random variables A and B,

$$\begin{aligned}
\operatorname{var}(A+B) &\geq \operatorname{var}(A) + \operatorname{var}(B) - 2\operatorname{var}(A)^{1/2}\operatorname{var}(B)^{1/2}|\operatorname{corr}(A,B)| \\
&\geq \operatorname{var}(A) + \operatorname{var}(B) - (\operatorname{var}(A) + \operatorname{var}(B))|\operatorname{corr}(A,B)| \geq \operatorname{var}(A)(1 - |\operatorname{corr}(A,B)|).
\end{aligned}$$

Since $\operatorname{corr}(\nu_i, \eta_i \mid Q)$ is bounded away from 1 and $\sigma_{\nu,i}^2$ is bounded away from zero a.s., it follows that, a.s,

$$\ddot{r}_n/\widetilde{\mathcal{V}}_G \preceq_{\text{a.s.}} O(\ddot{r}_n/\sum_i (G'R)_i^2) = O(1).$$

by Condition (iii). It remains to show that the first term in (33) converges to a standard normal random variable. To this end, we apply Lemma D.2 with $P = G/\sqrt{\ddot{r}_n}$, and $t = GR$ and $s = G'R_\Delta$. Condition 1 of Lemma D.2 holds since $\widetilde{\mathcal{V}}_G/\ddot{r}_n^{1/2}$ is bounded away from zero. Condition 2 of Lemma D.2 holds by Condition (iv). Finally, condition 3 holds by Lemma D.3. □

D.3 Lemmata for proving consistency of standard errors

First we introduce some notation that is used throughout the section. Let $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \in \mathbb{R}^n$ denote random vectors such that, conditional on $Q = (Z, W)$, $(\epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i}, \epsilon_{4i})$ are mean zero with bounded fourth moments, and the vectors $\{(\epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i}, \epsilon_{4i})\}_{i=1}^n$ are independent across i. Let $\sigma_{ab,i} = E[\epsilon_{ai}\epsilon_{bi} \mid Q]$, $\sigma_{abc,i} = E[\epsilon_{ai}\epsilon_{bi}\epsilon_{ci} \mid Q]$, and $\sigma_{abcd,i} = E[\epsilon_{ai}\epsilon_{bi}\epsilon_{ci}\epsilon_{di} \mid Q]$. Also, put $D_{ab} = \operatorname{diag}(\sigma_{ab})$, and similarly for D_{abc} and D_{abcd}, let $N = I - H_W$ and $M = I - H_Q$, and write $E_Q[\cdot]$ as a shorthand for $E[\cdot \mid Q]$.

Throughout the subsection, we use the inequality

$$\big(\textstyle\sum_{i=1}^{k} a_k\big)^2 \leq k \textstyle\sum_{i=1}^{k} a_k^2. \tag{34}$$

Lemma D.6. *Let $\{d_{ijk}\}_{i,j,k=1}^{n}$ be a sequence that is non-random conditional on Q. Then*

$$\sum_{i \neq j \neq k} d_{ijk}\epsilon_{1i}\epsilon_{2j}\epsilon_{3k}\epsilon_{4k} = O_P\left(\sqrt{\sum_{i,j,k} d_{ijk}^2 + \sum_{i,j}\Big(\sum_{k} d_{ijk}\sigma_{34k}\Big)^2}\right).$$

Proof. We will show that

$$A := E_Q\left(\sum_{i \neq j \neq k} d_{ijk}\epsilon_{1i}\epsilon_{2j}\epsilon_{3k}\epsilon_{4k}\right)^2 \preceq_{\text{a.s.}} \sum_{i,j,k} d_{ijk}^2 + \sum_{i,j}\left(\sum_{k} d_{ijk}\sigma_{34k}\right)^2.$$

The result will then follow by Markov inequality and dominated convergence theorem. Evaluating the expectation yields

$$\begin{aligned}
A = &\sum_{i \neq j \neq k \neq \ell} \left[d_{ijk}d_{ij\ell}\sigma_{11i}\sigma_{22j}\sigma_{34k}\sigma_{34\ell} + d_{ijk}d_{ji\ell}\sigma_{12i}\sigma_{12j}\sigma_{34k}\sigma_{34\ell}\right] \\
&+ \sum_{i \neq j \neq k} d_{ijk}\left[d_{ijk}\sigma_{11i}\sigma_{22j}\sigma_{3344k} + d_{ikj}\sigma_{11i}\sigma_{234j}\sigma_{234k} + d_{jik}\sigma_{12i}\sigma_{12j}\sigma_{3344k}\right] \\
&+ \sum_{i \neq j \neq k} d_{ijk}\left[d_{kij}\sigma_{12i}\sigma_{234j}\sigma_{134k} + d_{jki}\sigma_{134i}\sigma_{12j}\sigma_{234k} + d_{kji}\sigma_{134i}\sigma_{22j}\sigma_{134k}\right] \\
\preceq_{\text{a.s.}} &\sum_{i \neq j \neq k \neq \ell} d_{ijk}d_{ij\ell}\sigma_{11i}\sigma_{22j}\sigma_{34k}\sigma_{34\ell} + \sum_{i \neq j \neq k \neq \ell} d_{ijk}d_{ji\ell}\sigma_{12i}\sigma_{12j}\sigma_{34k}\sigma_{34\ell} + \sum_{i,j,k} d_{ijk}^2.
\end{aligned}$$

Let $c_{ijk} = \mathbb{I}\{i \neq j \neq k\}\, d_{ijk}$. The second term can then be bounded as

$$\begin{aligned}
&\sum_{i \neq j \neq k \neq \ell} d_{ijk}d_{ji\ell}\sigma_{12i}\sigma_{12j}\sigma_{34k}\sigma_{34\ell} \\
&\quad = \sum_{i,j} \sigma_{12i}\sigma_{12j}\left(\sum_{k} c_{ijk}\sigma_{34k}\right)\left(\sum_{\ell} c_{ji\ell}\sigma_{34\ell}\right) - \sum_{i \neq j \neq k} c_{ijk}c_{jik}\sigma_{12i}\sigma_{12j}\sigma_{34k}\sigma_{34k} \\
&\qquad\qquad \preceq_{\text{a.s.}} \sum_{i,j}\left(\sum_{k} c_{ijk}\sigma_{34k}\right)^2 + \sum_{i,j,k} d_{ijk}^2 \leq \sum_{i,j}\left(\sum_{k\colon k \neq i,j} d_{ijk}\sigma_{34k}\right)^2 + \sum_{i,j,k} d_{ijk}^2.
\end{aligned}$$

Since by (34),

$$\sum_{i,j}\left(\sum_{k\colon k \neq i,j} d_{ijk}\sigma_{34k}\right)^2 \leq 3\sum_{i,j}\left(\sum_{k} d_{ijk}\sigma_{34k}\right)^2 + 3\sum_{i,j}\left(d_{ijj}\sigma_{34j}\right)^2 + 3\sum_{i,j}\left(d_{iji}\sigma_{34i}\right)^2$$

$$\preceq_{\text{a.s.}} \sum_{i,j}\left(\sum_k d_{ijk}\sigma_{34k}\right)^2 + \sum_{i,j,k} d_{ijk}^2,$$

it follows that

$$\sum_{i\neq j\neq k\neq \ell} d_{ijk}d_{ji\ell}\sigma_{12i}\sigma_{12j}\sigma_{34k}\sigma_{34\ell} \preceq_{\text{a.s.}} \sum_{i,j}\left(\sum_k d_{ijk}\sigma_{34k}\right)^2 + \sum_{i,j,k} d_{ijk}^2.$$

By a symmetric argument,

$$\sum_{i\neq j\neq k\neq \ell} d_{ijk}d_{ij\ell}\sigma_{11i}\sigma_{22j}\sigma_{34k}\sigma_{34\ell} \preceq_{\text{a.s.}} \sum_{i,j}\left(\sum_k d_{ijk}\sigma_{34k}\right)^2 + \sum_{i,j,k} d_{ijk}^2,$$

which proves the result. □

Lemma D.7. *Let $\{d_{ij}\}_{i,j=1}^n$ be a sequence that is non-random conditional on Q. Then*

$$\sum_{i\neq j} d_{ij}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j} = O_P\left(\sqrt{\sum_{i,j} d_{ij}^2 + \sum_i \left(\sum_j d_{ij}\sigma_{23j}\right)^2}\right).$$

Proof. To prove the claim, we will show that $E_Q(\sum_{i\neq j} d_{ij}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j})^2 \preceq_{\text{a.s.}} \sum_{i,j} d_{ij}^2 + \sum_i \left(\sum_j d_{ij}\sigma_{23j}\right)^2$. The claim will then follow by Markov inequality and dominated convergence theorem. This expectation can be decomposed as

$$\begin{aligned} E_Q\left(\sum_{j\neq i} d_{ij}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j}\right)^2 &= \sum_{i\neq j}(\sigma_{123j}d_{ij}d_{ji}\sigma_{123i} + d_{ij}^2\sigma_{2233j}\sigma_{11i}) + \sum_{i\neq j\neq k} d_{ik}d_{ij}\sigma_{23k}\sigma_{23j}\sigma_{11i} \\ &\preceq_{\text{a.s.}} \sum_{i,j} d_{ij}^2 + \sum_{i\neq j\neq k} d_{ik}d_{ij}\sigma_{23k}\sigma_{23j}\sigma_{11i}. \end{aligned}$$

Let $c_{ij} = \mathbb{I}\{i\neq j\}\, d_{ij}$. The second term can then be decomposed as

$$\begin{aligned} \sum_{i\neq j\neq k} d_{ik}d_{ij}\sigma_{23k}\sigma_{23j}\sigma_{11i} &= \sum_{i,j,k} c_{ik}c_{ij}\sigma_{23k}\sigma_{23j}\sigma_{11i} - \sum_{i,j} c_{ij}^2\sigma_{23j}^2\sigma_{11i} \\ &= \sum_i \sigma_{11i}\Big(\sum_j c_{ij}\sigma_{23j}\Big)^2 - \sum_{i,j} c_{ij}^2\sigma_{23j}^2\sigma_{11i} \preceq_{\text{a.s.}} \sum_i\Big(\sum_j c_{ij}\sigma_{23j}\Big)^2 + \sum_{i,j} d_{ij}^2. \end{aligned}$$

The claim of the Lemma then follows from applying the inequality (34) to get the bound

$$\sum_i\Big(\sum_j c_{ij}\sigma_{23j}\Big)^2 \leq 2\sum_i\Big(\sum_j d_{ij}\sigma_{23j}\Big)^2 + 2\sum_i d_{ii}^2\sigma_{23i}^2 \preceq_{\text{a.s.}} \sum_i\Big(\sum_j d_{ij}\sigma_{23j}\Big)^2 + \sum_{i,j} d_{ij}^2.$$

□

Lemma D.8. *For any projection matrices P and R,*

$$E_Q[(P\epsilon_1)_i(R\epsilon_2)_i(P\epsilon_1)_j(R\epsilon_2)_j] \leq C\sqrt{P_{ii}P_{jj}R_{ii}R_{jj}},$$

where $C = \sup_i \sigma_{1122i} + \sup_i \sigma_{11i} \sup_j \sigma_{22j} + \sup_i \sigma_{12i}^2$.

Proof. Evaluating the expectation yields

$$\begin{aligned} E_Q[(P\epsilon_1)_i(R\epsilon_2)_i(P\epsilon_1)_j(R\epsilon_2)_j] &= \sum_k P_{ik}P_{jk}R_{ik}R_{jk}(\sigma_{1122k} - \sigma_{11k}\sigma_{22k} - 2\sigma_{12k}^2) \\ &+ \sum_{k,\ell} P_{ik}P_{j\ell}R_{ik}R_{j\ell}\sigma_{12k}\sigma_{12\ell} + \sum_{k,\ell} P_{ik}P_{jk}R_{i\ell}R_{j\ell}\sigma_{11k}\sigma_{22\ell} + \sum_{k,\ell} P_{ik}P_{jk}R_{i\ell}R_{j\ell}\sigma_{12k}\sigma_{12\ell}. \end{aligned}$$

By Cauchy-Schwarz inequality,

$$\sum_{k,\ell}|P_{ik}P_{j\ell}R_{ik}R_{j\ell}| \leq \left(\sum_{k,\ell} P_{ik}^2P_{j\ell}^2 \sum_{k,\ell} R_{ik}^2R_{j\ell}^2\right)^{1/2} = \sqrt{P_{ii}P_{jj}R_{ii}R_{jj}},$$

and similarly $\sum_{k,\ell}|P_{ik}P_{jk}R_{i\ell}R_{j\ell}| \leq \sqrt{P_{ii}P_{jj}R_{ii}R_{jj}}$ and $\sum_{k,\ell}|P_{ik}P_{jk}R_{i\ell}R_{j\ell}| \leq \sqrt{P_{ii}P_{jj}R_{ii}R_{jj}}$. Thus,

$$E_Q[(P\epsilon_1)_i(R\epsilon_2)_i(P\epsilon_1)_j(R\epsilon_2)_j] \leq 4C\sqrt{P_{ii}P_{jj}R_{ii}R_{jj}},$$

which proves the result. □

Lemma D.9. *Let $\mu_1, \mu_2 \in \mathbb{R}^n$ and $\mathcal{H}$ denote vectors and a projection matrix that are non-random conditional on Q. Let $M = I - \mathcal{H}$ and $\hat{\sigma}_{34k} = (M\epsilon_3)_k(M\epsilon_4)_k$, and suppose*

(i) $\|\mu_1\|_2^2/\ddot{r}_n$, $\|\mu_1\|_2^2/\ddot{r}_n$, $\|\mu_1\|_\infty$, $\|\mu_2\|_\infty$, *and* $(K+L)/\ddot{r}_n$ *is bounded a.s.*

(ii) $\max_i \mathcal{H}_{ii} \overset{a.s.}{\to} 0$.

Then $\sum_k \mu_{1k}\mu_{2k}\hat{\sigma}_{34k} = \sum_k \sigma_{34k}\mu_{1k}\mu_{2k} + o_P(\ddot{r}_n)$.

Proof. First we bound

$$\begin{aligned} \operatorname{var}(\sum_k \hat{\sigma}_{34k}\mu_{1k}\mu_{2k} \mid Q) &= \sum_{k,\ell} \mu_{1k}\mu_{2k}\mu_{1\ell}\mu_{2\ell} \sum_a M_{ak}^2 M_{a\ell}^2(\sigma_{3344a} - \sigma_{34a}^2) \\ &+ \sum_{k,\ell} \mu_{1k}\mu_{2k}\mu_{1\ell}\mu_{2\ell} \sum_{a\neq b} M_{ak}M_{bk}M_{a\ell}M_{b\ell}(\sigma_{33a}\sigma_{44b} + \sigma_{34a}\sigma_{34b}). \end{aligned} \tag{35}$$

Since $M_{ak}^2 = \mathbb{I}\{a = k\}(1 - (2\mathcal{H})_{aa}) + \mathcal{H}_{ak}^2$, it follows that the first term in (35) of the order

$$\sum_{k,\ell}|\mu_{1k}\mu_{2k}\mu_{1\ell}\mu_{2\ell}| \sum_a M_{ak}^2M_{a\ell}^2 \preceq_{\text{a.s.}} \sum_{k,\ell,a} \mathcal{H}_{ak}^2\mathcal{H}_{a\ell}^2 + \sum_{k,\ell} \mathcal{H}_{k\ell}^2 + \sum_k \mu_{1k}^2\mu_{2k}^2 \preceq_{\text{a.s.}} \ddot{r}_n.$$

Now, $\sum_{a,b} M_{ak}M_{bk}M_{a\ell}M_{b\ell}\sigma_{33a}\sigma_{44b} = (MD_{33}M)_{k\ell}(MD_{44}M)_{k\ell} \leq (MD_{33}M)^2_{k\ell} + (MD_{44}M)^2_{k\ell}$. Furthermore, note that, letting $a_k = \mu_{1k}\mu_{2k}$,

$$\begin{aligned}\sum_{k,\ell} a_k a_\ell (MD_{33}M)^2_{k\ell} &= \sum_k a_k^2 \sigma^2_{33k} + \sum_{k,\ell} a_k a_\ell \mathcal{H}^2_{k\ell}(\sigma^2_{33k} + \sigma^2_{33\ell}) + \sum_{k,\ell} a_k a_\ell (\mathcal{H}D_{33}\mathcal{H})^2_{k\ell} \\ &\preceq_{\text{a.s.}} \sum_k \mu^2_{1k}\mu^2_{2k} + \sum_{k,\ell}\mathcal{H}^2_{kl} + \|\mathcal{H}D_{33}\mathcal{H}\|^2_F \preceq_{\text{a.s.}} \ddot{r}_n + 2K.\end{aligned}$$

Therefore, the second term in (35) is of the order $\ddot{r}_n$. Thus, by Markov inequality and dominated convergence theorem,

$$\begin{aligned}\sum_k \mu_{1k}\mu_{2k}\hat{\sigma}_{34k} - \sum_k \mu_{1k}\mu_{2k}\sigma_{34k} &= \sum_k \mu_{1k}\mu_{2k} E_Q[\hat{\sigma}_{34k} - \sigma_{34k}] + o_P(\ddot{r}_n) \\ &= \sum_k \mu_{1k}\mu_{2k} \sum_i \mathcal{H}^2_{ik}\sigma_{34i} - 2\sum_k \mu_{1k}\mu_{2k}\sigma_{34k}\mathcal{H}_{kk} + o_P(\ddot{r}_n) \\ &\preceq_{\text{a.s.}} \max_i \mathcal{H}_{ii}\ddot{r}_n + o_P(\ddot{r}_n) = o_P(\ddot{r}_n),\end{aligned}$$

as claimed. □

Lemma D.10. *Let $\mu_1, \mu_2 \in \mathbb{R}^n$ denote vectors that are non-random conditional on Q. Put $\hat{\sigma}_{34k} = (M\epsilon_3)_k(M\epsilon_4)_k$, and consider a variance estimator of the form*

$$\hat{\Omega} = J(\mu_1 + (I - H_W)\epsilon_1, \mu_2 + (I - H_W)\epsilon_2, \hat{\sigma}_{34}) \tag{36}$$

such that

(i) $\sum_j (H_{\ddot{Z}})_{jk}\mu_{2j} = \mu_{2k}$ *and* $\sum_j (H_{\ddot{Z}})_{jk}\mu_{1j} = \mu_{1k}$*,*

(ii) $\|\mu_1\|^2_2/\ddot{r}_n$ *and* $\|\mu_1\|^2_2/\ddot{r}_n$ *are bounded a.s.*

(iii) $\max_i (H_Q)_{ii} \overset{a.s.}{\to} 0$.

(iv) $\|\mu_1\|_\infty$ *and* $\|\mu_2\|_\infty$ *are bounded a.s.*

(v) $(K + L)/\ddot{r}_n$ *is bounded a.s.*

Then $\hat{\Omega} = \sum_k E[\epsilon_{3k}\epsilon_{4k} \mid Q]\mu_{1k}\mu_{2k} + o_P(\ddot{r}_n)$*.*

Proof. *1.* To prove the Lemma, it will be convenient to decompose the right-hand side of (36). Write

$$\hat{\Omega} = \sum_k \hat{\sigma}_{34k} \sum_{i\colon i \neq k} \sum_{j\colon j \neq i,k} a_{ijk},$$

where $a_{ijk} = (H_{\ddot{Z}})_{ik}(H_{\ddot{Z}})_{jk}(\mu_{1i} + (N\epsilon_1)_i)(\mu_{2i} + (N\epsilon_2)_i)$. We can write

$$\begin{aligned}\sum_{i\colon i\neq k}\sum_{j\colon j\neq i,k} a_{ijk} &= \sum_{i,j} a_{ijk} - \sum_j a_{kjk} - \sum_i a_{iik} - \sum_i a_{ikk} + 2a_{kkk} \\ &= \mu_{1k}\mu_{2k} + \sum_{\ell=1}^{5} T_{\ell k},\end{aligned} \tag{37}$$

where

$$\begin{aligned}
T_{1k} &= -\sum_i (H_{\ddot{Z}})_{ik}^2 \mu_{1i}\mu_{2i} - 2(H_{\ddot{Z}})_{kk}(1-(H_{\ddot{Z}})_{kk})\mu_{1k}\mu_{2k},\\
T_{2k} &= \mu_{1k}\left[(1-(H_{\ddot{Z}})_{kk})(H_{\ddot{Z}}\epsilon_2)_k - (H_{\ddot{Z}})_{kk}(1-2(H_{\ddot{Z}})_{kk})(N\epsilon_2)_k\right],\\
T_{3k} &= \mu_{2k}\left[(1-(H_{\ddot{Z}})_{kk})(H_{\ddot{Z}}\epsilon_1)_k - (H_{\ddot{Z}})_{kk}(1-2(H_{\ddot{Z}})_{kk})(N\epsilon_1)_k\right],\\
T_{4k} &= (H_{\ddot{Z}}\epsilon_1)_k(H_{\ddot{Z}}\epsilon_2)_k - (H_{\ddot{Z}})_{kk}(N\epsilon_1)_k(H_{\ddot{Z}}\epsilon_2)_k - (H_{\ddot{Z}})_{kk}(H_{\ddot{Z}}\epsilon_1)_k(N\epsilon_2)_k\\
&\quad + 2(H_{\ddot{Z}})_{kk}^2(N\epsilon_1)_k(N\epsilon_2)_k - \sum_i (H_{\ddot{Z}})_{ik}^2(N\epsilon_2)_i(N\epsilon_1)_i,\\
T_{5k} &= -\sum_i (H_{\ddot{Z}})_{ik}^2(\mu_{1i}(N\epsilon_2)_i + \mu_{2i}(N\epsilon_1)_i).
\end{aligned}$$

Therefore, equation (36) can be written as

$$\hat{\Omega} = \sum_k \hat{\sigma}_{34k}\mu_{1k}\mu_{2k} + \sum_k(\hat{\sigma}_{34k} - \epsilon_{3k}\epsilon_{4k})\sum_{\ell=1}^{5} T_{\ell k} + \sum_k \epsilon_{3k}\epsilon_{4k}\sum_{\ell=1}^{5} T_{\ell k}. \tag{38}$$

By Lemma D.9 with $\mathcal{H} = H_Q$, $\sum_k \hat{\sigma}_{34k}\mu_{1k}\mu_{2k} = \sum_k \sigma_{34k}\mu_{1k}\mu_{2k} + o_P(\ddot{r}_n)$. To prove the assertion of the Lemma, we will show that the remaining terms are of order $o_P(\ddot{r}_n)$.
2. Consider the second term in (38). It follows from the Cauchy-Schwarz inequality and the inequality (34) that

$$(E_Q\sum_k(\hat{\sigma}_{34k} - \epsilon_{3k}\epsilon_{4k})\sum_{\ell=1}^{5} T_{\ell k})^2 \leq 5\sum_m E_Q(\hat{\sigma}_{34m} - \epsilon_{3m}\epsilon_{4m})^2\sum_{\ell=1}^{5}\sum_k E_Q T_{\ell k}^2.$$

We now show that the right-hand side is of the order $o(\ddot{r}_n^2)$. By (34), $(\hat{\sigma}_{34k} - \epsilon_{3k}\epsilon_{4k})^2 \leq 3\epsilon_{3k}^2(H_Q\epsilon_4)_k^2 + 3(H_Q\epsilon_3)_k^2\epsilon_{4k}^2 + 3(H_Q\epsilon_3)_k^2(H_Q\epsilon_4)_k^2$, so that by Lemma D.8,

$$\sum_k E_Q(\hat{\sigma}_{34k} - \epsilon_{3k}\epsilon_{4k})^2 \preceq_{\text{a.s.}} 2\sum_k (H_Q)_{kk} + \sum_k (H_Q)_{kk}^2 \preceq_{\text{a.s.}} K + L. \tag{39}$$

Therefore, to prove the claim, we need to show for $\ell = 1, \dots, 5$, $\sum_k E_Q T_{1k}^2 = o(\ddot{r}_n)$. Using the inequality (34), and the assumptions of the Lemma yields

$$\begin{aligned}\sum_k T_{1k}^2 &\preceq_{\text{a.s.}} \sum_{i,j,k}(H_{\ddot{Z}})_{ik}^2(H_{\ddot{Z}})_{jk}^2 + \sum_k (H_{\ddot{Z}})_{kk}^2\mu_{1k}^2\mu_{2k}^2 \preceq_{\text{a.s.}} \max_i (H_{\ddot{Z}})_{ii}\ddot{r}_n,\\ E_Q\sum_k T_{2k}^2 &\preceq_{\text{a.s.}} \sum_k \mu_{1k}^2 E_Q\left[(H_{\ddot{Z}}\epsilon_2)_k^2 - (H_{\ddot{Z}})_{kk}^2(N\epsilon_2)_k^2\right] \preceq_{\text{a.s.}} \max_i (H_{\ddot{Z}})_{ii}\ddot{r}_n,\end{aligned}$$

and, by a symmetric argument $E_Q\sum_k T_{3k}^2 \preceq_{\text{a.s.}} \max_i (H_{\ddot{Z}})_{ii}\ddot{r}_n$. To bound the term $\sum_k T_{4k}^2$, first observe that by Lemma D.8,

$$\begin{aligned}\sum_k E_Q\left(\sum_i (H_{\ddot{Z}})_{ik}^2(N\epsilon_2)_i(N\epsilon_1)_i\right)^2 &= \sum_{i,j,k}(H_{\ddot{Z}})_{ik}^2(H_{\ddot{Z}})_{jk}^2 E_Q(N\epsilon_2)_i(N\epsilon_1)_i(N\epsilon_2)_j(N\epsilon_1)_j\\ &\preceq_{\text{a.s.}} \sum_{i,j,k}(H_{\ddot{Z}})_{ik}^2(H_{\ddot{Z}})_{jk}^2 N_{ii}N_{jj} \le \max_i (H_{\ddot{Z}})_{ii}K.\end{aligned} \tag{40}$$

By (40) and Lemma D.8,

$$E_Q\sum_k T_{4k}^2 \preceq_{\text{a.s.}} \sum_k\left[(H_{\ddot{Z}})_{kk}^2 + (H_{\ddot{Z}})_{kk}^3N_{kk} + (H_{\ddot{Z}})_{kk}^4N_{kk}^2\right] + \max_i (H_{\ddot{Z}})_{ii}K \preceq_{\text{a.s.}} \max_i (H_{\ddot{Z}})_{ii}K.$$

Finally, to bound the term $\sum_k E_Q T_{5k}^2$, we first need a preliminary result. Let A denote the matrix with elements $A_{ik} = (H_{\ddot{Z}})_{ik}^2\mu_{1i}$. Then

$$E_Q\sum_k\left(\sum_i (H_{\ddot{Z}})_{ik}^2\mu_{1i}(N\epsilon_2)_i\right)^2 = \sum_\ell (A'N)_{k\ell}^2\sigma_{22,\ell}^2 \preceq_{\text{a.s.}} \|A'N\|_F^2 \le \|A\|_F^2 \le \ddot{r}_n \max_i (H_{\ddot{Z}})_{ii}. \tag{41}$$

By (41),

$$\sum_k T_{5k}^2 \le 2\sum_k E_Q\left(\textstyle\sum_i (H_{\ddot{Z}})_{ik}^2\mu_{1i}(N\epsilon_2)_i\right)^2 + 2\textstyle\sum_k E_Q\left(\sum_i (H_{\ddot{Z}})_{ik}^2\mu_{2i}(N\epsilon_1)_i\right)^2 \preceq \max_i (H_{\ddot{Z}})_{ii}\ddot{r}_n.$$

By Markov inequality and dominated convergence theorem, the second term in (38) is of the order $o(\ddot{r}_n)$.

3a. To finish the proof of the Lemma, it remains to show that the third term in (38) is of the order $o(\ddot{r}_n)$, for which it suffices to show that

$$\sum_k \epsilon_{3k}\epsilon_{4k}T_{\ell k} = o_P(\ddot{r}_n), \tag{42}$$

for $\ell = 1, \dots, 5$. To show that (42) holds for $\ell = 1$, note that by triangle inequality,

$$E_Q\sum_k |\epsilon_{3k}\epsilon_{4k}T_{1k}| \preceq 2\max_i (H_{\ddot{Z}})_{ii}\sum_i |\mu_{1k}\mu_{2k}| + \sum_{i,k}(H_{\ddot{Z}})_{ik}^2|\mu_{1k}\mu_{2k}| \preceq \max_i (H_{\ddot{Z}})_{ii}\ddot{r}_n.$$

By Markov inequality, equation (42) therefore holds for $\ell = 1$. To show (42) for $\ell = 2$, write

$$\sum_k \epsilon_{3k}\epsilon_{4k}T_{2k} = \sum_{i \neq k} f_{ik}\epsilon_{2i}\epsilon_{3k}\epsilon_{4k} + \sum_k f_{kk}\epsilon_{2k}\epsilon_{3k}\epsilon_{4k}$$

where $d_{ij} = d_{ij} + \tilde{d}_{ij}$, $d_{ij} = \mu_{1j}(1 - (H_{\ddot{Z}})_{jj})(H_{\ddot{Z}})_{ij}$, and $\tilde{d}_{ij} = -\mu_{1j}(H_{\ddot{Z}})_{jj}(1 - 2(H_{\ddot{Z}})_{jj})N_{ij}$. Note that $\sum_{i,j}(d_{ij}^2 + \tilde{d}_{ij}^2) \leq 2\ddot{r}_n \max_i(H_{\ddot{Z}})_{ii}$, and that $\sum_i(\sum_j d_{ij}\sigma_{34j})^2 + \sum_i(\sum_j \tilde{d}_{ij}\sigma_{34j})^2 \preceq_{\text{a.s.}} \|H_{\ddot{Z}}\mu_1\|_2^2 + \|N\mu_1\|_2^2 \preceq_{\text{a.s.}} \ddot{r}_n$. Therefore, by Lemma D.7, the first term is of the order $O_P(\ddot{r}_n^{1/2})$. The expectation of the second term can be bounded as

$$\begin{aligned} E_Q|\sum_k f_{kk}\epsilon_{2k}\epsilon_{3k}\epsilon_{4k}| \preceq_{\text{a.s.}} \sum_k |f_{kk}| &= \sum_k |\mu_{1k}||(H_{\ddot{Z}})_{kk}^2 + (H_{\ddot{Z}})_{kk}(H_W)_{kk}(1 - 2(H_{\ddot{Z}})_{kk})| \\ &\preceq_{\text{a.s.}} \sum_k ((H_{\ddot{Z}})_{kk}^2 + (H_{\ddot{Z}})_{kk}(H_W)_{kk}) \leq \max_i (H_Q)_{ii}K. \end{aligned}$$

so that by Markov inequality and dominated convergence theorem, $\sum_k f_{kk}\epsilon_{2k}\epsilon_{3k}\epsilon_{4k} = o_P(\ddot{r}_n)$, so that (42) holds for $\ell = 2$, and by a symmetric argument, for $\ell = 3$ also.
3b. To show (42) for $\ell = 4$, write $\sum_k \epsilon_{3k}\epsilon_{4k}T_{4k} = \sum_{i,j,k} d_{ijk}\epsilon_{1i}\epsilon_{2j}\epsilon_{3k}\epsilon_{4k}$, where

$$\begin{aligned} d_{ijk} = \mathbb{I}\{i \neq j\}\,& (H_{\ddot{Z}})_{ik}(H_{\ddot{Z}})_{jk} - (H_{\ddot{Z}})_{kk}(H_{\ddot{Z}})_{jk}N_{ik} - (H_{\ddot{Z}})_{kk}(H_{\ddot{Z}})_{ik}N_{jk} \\ &+ 2(H_{\ddot{Z}})_{kk}^2 N_{ik}N_{jk} + \left[(H_{\ddot{Z}})_{ik}^2 + (H_{\ddot{Z}})_{jk}^2\right](H_W)_{ij} - \sum_\ell (H_{\ddot{Z}})_{\ell k}^2 (H_W)_{\ell i}(H_W)_{\ell j}. \end{aligned}$$

We can therefore decompose this term as

$$\begin{aligned} \sum_k \epsilon_{3k}\epsilon_{4k}T_{4k} = \sum_i d_{iii}\epsilon_{1i}\epsilon_{2i}\epsilon_{3i}\epsilon_{4i} &+ \sum_{i \neq j} d_{iij}\epsilon_{1i}\epsilon_{2i}\epsilon_{3j}\epsilon_{4j} \\ &+ \sum_{i \neq j} d_{ijj}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j}\epsilon_{4j} + \sum_{i \neq j} d_{iji}\epsilon_{1i}\epsilon_{2j}\epsilon_{3i}\epsilon_{4i} + \sum_{i \neq j \neq k} d_{ijk}\epsilon_{1i}\epsilon_{2j}\epsilon_{3k}\epsilon_{4k}. \end{aligned} \tag{43}$$

We will show that all five terms in (43) are of the order $o_P(\ddot{r}_n)$. Since

$$d_{iii} = 2(H_{\ddot{Z}})_{ii}^2(H_W)_{ii}^2 - \sum_\ell (H_{\ddot{Z}})_{\ell i}^2 (H_W)_{\ell i}^2,$$

by triangle inequality,

$$E_Q|\textstyle\sum_i d_{iii}\epsilon_{1i}\epsilon_{2i}\epsilon_{3i}\epsilon_{4i}| \preceq_{\text{a.s.}} \sum_i |d_{iii}| \leq 2\sum_i (H_{\ddot{Z}})_{ii}^2(H_W)_{ii}^2 + \sum_{i,\ell}(H_{\ddot{Z}})_{\ell i}^2(H_W)_{\ell i}^2 \leq \max_i (H_W)_{ii}K,$$

so that by Markov inequality, the first term in (43) is of the order $o_P(\ddot{r}_n)$. Similarly, by triangle inequality, and the inequality $|2ab| \leq a^2 + b^2$,

$$\begin{aligned}
\sum_{i\neq j}|d_{iij}| &= \sum_{i\neq j}|2(H_{\ddot{Z}})_{jj}(H_{\ddot{Z}})_{ij}(H_W)_{ij} - 2(H_{\ddot{Z}})^2_{jj}(H_W)^2_{ij} + 2(H_W)_{ii}(H_{\ddot{Z}})^2_{ij} - \sum_{\ell}(H_{\ddot{Z}})^2_{\ell j}(H_W)^2_{\ell i}| \\
&\leq \sum_{i,j}(H_{\ddot{Z}})_{jj}\left[(H_W)^2_{ij} + (H_{\ddot{Z}})^2_{ij}\right] + 2\sum_{i}\left[(H_{\ddot{Z}})^2_{ii} + (H_{\ddot{Z}})_{ii}\right](H_W)_{ii} + \sum_{\ell}(H_{\ddot{Z}})^2_{\ell\ell} \\
&\leq 2K\max_i(H_{\ddot{Z}})_{ii} + K\max_i(H_W)_{ii} + 4L\max_i(H_{\ddot{Z}})_{ii}
\end{aligned}$$

Therefore, by triangle inequality

$$E_Q|\textstyle\sum_{i\neq j} d_{iij}\epsilon_{1i}\epsilon_{2i}\epsilon_{3j}\epsilon_{4j}| \preceq_{\text{a.s.}} \sum_{i\neq j}|d_{iij}| \leq 4\ddot{r}_n \max_i(H_Q)_{ii},$$

so that by Markov inequality, the second term in (43) is of the order $o_P(\ddot{r}_n)$ also. To bound the third term in (43), decompose it as

$$\begin{aligned}
\sum_{i\neq j} d_{ijj}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j}\epsilon_{4j} &= \sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}(H_W)_{ij}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j}\epsilon_{4j} \\
&\quad + \sum_{i\neq j}\left[2(H_{\ddot{Z}})^2_{jj}(H_W)_{jj}(H_W)_{ij} + (H_{\ddot{Z}})_{jj}(H_W)_{jj}(H_{\ddot{Z}})_{ij}\right]\epsilon_{1i}\epsilon_{2j}\epsilon_{3j}\epsilon_{4j} \\
&\quad - \sum_{i\neq j}\sum_{\ell}(H_{\ddot{Z}})^2_{\ell j}(H_W)_{\ell i}(H_W)_{\ell j}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j}\epsilon_{4j}. \qquad (44)
\end{aligned}$$

By triangle inequality

$$E_Q|\textstyle\sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}(H_W)_{ij}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j}\epsilon_{4j}| \preceq_{\text{a.s.}} \sum_{i,j}|(H_{\ddot{Z}})^2_{ij}(H_W)_{ij}| \leq K\max_{i,j}(H_W)_{ij},$$

so that the first term in (44) is of the order $o_P(\ddot{r}_n)$. Next, note that for any vector a, and a projection matrix P, by Cauchy-Schwarz inequality,

$$E_Q|\sum_{i\neq j} a_j P_{ij}\epsilon_{3j}\epsilon_{4j}\epsilon_{1i}\epsilon_{2j}|^2 \leq \sum_j a_j^2\sigma_{3344j}\cdot\sum_j\sigma_{22j}\sum_{i\colon i\neq j}P^2_{ij}\sigma_{11i} \preceq_{\text{a.s.}} \|a\|_2^2\|P\|_F^2.$$

Applying this to two summands in the second term in (44), with $a = 2(H_{\ddot{Z}})^2_{jj}(H_W)_{jj}$ and $P = (H_W)_{ij}$, and $a = (H_{\ddot{Z}})_{jj}(H_W)_{jj}$ and $P = (H_{\ddot{Z}})_{ij}$, and combining the result with Markov inequality implies that the second term in (43) is also of the order $o_P(\ddot{r}_n)$. Finally, by Cauchy-Schwarz and triangle inequalities, the expected value of the third term in (44) can be bounded as

$$\begin{aligned}
&E_Q\left|\sum_{\ell}\sum_{i\neq j}(H_{\ddot{Z}})^2_{\ell j}(H_W)_{\ell i}(H_W)_{\ell j}\epsilon_{1i}\epsilon_{2j}\epsilon_{3j}\epsilon_{4j}\right| \\
&\qquad \leq \sum_{\ell}\left(\sum_j(H_W)^2_{\ell j}\sigma_{3344j}\right)^{1/2}\left(\sum_j\sigma_{22j}(H_{\ddot{Z}})^4_{\ell j}\sum_{i\colon i\neq j}(H_W)^2_{\ell i}\sigma_{11i}\right)^{1/2}
\end{aligned}$$

$$\preceq_{\text{a.s.}} \sum_{\ell} (H_W)_{\ell\ell}^{1/2} \left(\sum_j (H_{\ddot{Z}})_{\ell j}^2 (H_W)_{\ell\ell} \right)^{1/2} \leq \max_i (H_{\ddot{Z}})_{\ell\ell}^{1/2} L.$$

Thus, by Markov inequality, the third term in (44) is also of the order $o_P(\ddot{r}_n)$, which shows that the third term in (43) is of the order $o_P(\ddot{r}_n)$. Since $d_{iji} = d_{jii}$ if $i \neq j$, the fourth term in (43) is of the order $o_P(\ddot{r}_n)$ by a similar argument. Next, we show that the last term in (43) is of the order $o_P(\ddot{r}_n)$. By Lemma D.6,

$$\sum_{i \neq j \neq k} d_{ijk}^2 \epsilon_{1i} \epsilon_{2j} \epsilon_{3k} \epsilon_{4k} = O_p\Big(\sqrt{\textstyle\sum_{i,j,k} d_{ijk}^2 + \sum_{i,j} \left(\sum_k d_{ijk} \sigma_{34k} \right)^2}\Big). \tag{45}$$

Note that by (34),

$$\begin{aligned} \sum_{i,j,k} d_{ijk}^2 \preceq{}& \sum_{i,j,k} (H_{\ddot{Z}})_{ik}^2 (H_{\ddot{Z}})_{jk}^2 + \sum_{i,j,k} (H_{\ddot{Z}})_{kk}^2 (H_{\ddot{Z}})_{jk}^2 N_{ik}^2 + \sum_{i,j,k} (H_{\ddot{Z}})_{kk}^4 N_{ik}^2 N_{jk}^2 + \sum_{i,j,k} (H_{\ddot{Z}})_{jk}^4 (H_W)_{ij}^2 \\ &+ \sum_{i,j,k,\ell,m} (H_{\ddot{Z}})_{\ell k}^2 (H_W)_{\ell i} (H_W)_{\ell j} (H_{\ddot{Z}})_{mk}^2 (H_W)_{mi} (H_W)_{mj} \\ \leq{}& 3 \sum_k (H_{\ddot{Z}})_{kk}^2 + \sum_j (H_{\ddot{Z}})_{jj} (H_W)_{jj} + \sum_{k,\ell,m} (H_{\ddot{Z}})_{\ell k}^2 (H_{\ddot{Z}})_{mk}^2 (H_W)_{m\ell} (H_W)_{\ell m} \\ \leq{}& 4 \max_i (H_Q)_{ii} K + \sum_{k,\ell,m} (H_{\ddot{Z}})_{\ell k}^2 (H_{\ddot{Z}})_{mk}^2 \leq 5 \max_i (H_Q)_{ii} K. \end{aligned}$$

To bound the term in equation (45), $\sum_{i,j} \left(\sum_k d_{ijk} \sigma_{34k} \right)^2$, write $d_{ijk} = \sum_{a=1}^8 d_{ijk}^a$, where $d_{ijk}^1 = -\mathbb{I}\{i = j\} (H_{\ddot{Z}})_{ik}^2$, $d_{ijk}^2 = (H_{\ddot{Z}})_{ik} (H_{\ddot{Z}})_{jk}$, $d_{ijk}^3 = -(H_{\ddot{Z}})_{kk} (H_{\ddot{Z}})_{jk} N_{ik}$, $d_{ijk}^4 = d_{jik}^3$, $d_{ijk}^5 = 2(H_{\ddot{Z}})_{kk}^2 \cdot N_{ik} N_{jk}$, $d_{ijk}^6 = (H_{\ddot{Z}})_{ik}^2 (H_W)_{ij}$, $d_{ijk}^7 = d_{jik}^6$, and $d_{ijk}^8 = -\sum_{\ell} (H_{\ddot{Z}})_{\ell k}^2 (H_W)_{\ell i} (H_W)_{\ell j}$. Note that for any projection matrices P, R, and a vector a,

$$\sum_{i,j} \left(\sum_k a_k P_{ik} R_{jk} \right)^2 = \|P \operatorname{diag}(a) R\|_F^2 \leq \|\operatorname{diag}(a)\|_F^2 = \sum_i a_i^2.$$

Hence,

$$\sum_{i,j} \left[(\sum_k d_{ijk}^3 \sigma_{34k})^2 + (\sum_k d_{ijk}^4 \sigma_{34k})^2 + (\sum_k d_{ijk}^5 \sigma_{34k})^2 \right] \leq \sum_i \sigma_{34i}^2 \left(2(H_{\ddot{Z}})_{ii}^2 + 4(H_{\ddot{Z}})_{ii}^4 \right) \preceq_{\text{a.s.}} K.$$

Furthermore,

$$\begin{aligned} &\sum_{i,j} (\sum_k d_{ijk}^1 \sigma_{34k})^2 \preceq \sum_i (H_{\ddot{Z}})_{ii}^2 \leq K \\ &\sum_{i,j} (\sum_k d_{ijk}^2 \sigma_{34k})^2 \preceq \sum_{i,j} (\sum_k d_{ijk}^2)^2 = \sum_{ij} (H_{\ddot{Z}})_{ij}^2 = K \end{aligned}$$

$$
\begin{aligned}
\sum_{i,j}(\sum_k d^8_{ijk}\sigma_{34k})^2 &= \sum_{k,\ell,i,j} \sigma_{34k}\sigma_{34i}(H_{\ddot{Z}})^2_{\ell k}(H_{\ddot{Z}})^2_{ij}(H_W)^2_{j\ell} \preceq_{\text{a.s.}} \sum_{k,\ell,i,j} (H_{\ddot{Z}})^2_{\ell k}(H_{\ddot{Z}})^2_{ij}(H_W)^2_{j\ell} \\
&\leq K \\
\sum_{i,j}(\sum_k d^6_{ijk}\sigma_{34k})^2 &= \sum_{i,j}(H_W)^2_{ij}(\sum_k(H_{\ddot{Z}})^2_{ik}\sigma_{34k})^2 \preceq_{\text{a.s.}} \sum_{i,j}(H_W)^2_{ij}(\sum_k(H_{\ddot{Z}})^2_{ik})^2 \leq K,
\end{aligned}
$$

and $\sum_{i,j}(\sum_k d^7_{ijk}\sigma_{34k})^2 \preceq_{\text{a.s.}} K$ by a similar argument. Thus, by (34),

$$
\sum_{i,j}(\sum_k d_{ijk}\sigma_{34k})^2 \preceq_{\text{a.s.}} \sum_{a=1}\sum_{i,j}(\sum_k d^a_{ijk}\sigma_{34k})^2 \preceq_{\text{a.s.}} K,
$$

so that by (45), the last term in (43) is of the order $o_P(\ddot{r}_n)$ as claimed.
3c. To complete the proof, it remains to show that (42) holds for $\ell = 5$. We have

$$
\sum_k \epsilon_{3k}\epsilon_{4k}T_{5k} = -\sum_{i,k}\epsilon_{3k}\epsilon_{4k}(H_{\ddot{Z}})^2_{ik}\mu_{1i}(N\epsilon_2)_i - \sum_{i,k}\epsilon_{3k}\epsilon_{4k}(H_{\ddot{Z}})^2_{ik}\mu_{1i}(N\epsilon_2)_i
$$

We will show that the first term in the above display is of the order $o_P(\ddot{r}_n)$; the proof the the second term is of the order $o_P(\ddot{r}_n)$ follows by a similar argument. To this end, letting $A_{ik} = (H_{\ddot{Z}})^2_{ik}\mu_{1i}$, we can write

$$
\sum_{i,k}\epsilon_{3k}\epsilon_{4k}(H_{\ddot{Z}})^2_{ik}\mu_{1i}(N\epsilon_2)_i = \sum_i(A_{ii} - (H_WA)_{ii})\epsilon_{3i}\epsilon_{4i}\epsilon_{2i} + \sum_{i\neq j}(NA)_{ij}\epsilon_{2i}\epsilon_{3j}\epsilon_{4j}. \tag{46}
$$

The expected value of first term in (46) can be bounded as

$$
\begin{aligned}
E_Q|\textstyle\sum_i(A_{ii} - (H_WA)_{ii})\epsilon_{3i}\epsilon_{4i}\epsilon_{2i}| &\preceq_{\text{a.s.}} \textstyle\sum_i|A_{ii}| + \sum_i|(H_WA)_{ii}| \\
&\preceq_{\text{a.s.}} \sum_i(H_Z)^2_{ii} + \|A\|_F\|H_W\|_F \preceq_{\text{a.s.}} \ddot{r}_n \max_i(H_{\ddot{Z}})_{ii}.
\end{aligned}
$$

Thus, by Markov inequality, the first term in (46) is of the order $o_P(\ddot{r}_n)$. Note that $\|NA\|^2_F \leq \|A\|^2_F \preceq_{\text{a.s.}} K$, and that $\|NA\sigma_{34}\|^2_2 \leq \|A\sigma_{34}\|^2_2 \preceq_{\text{a.s.}} K$, so that by Lemma D.7, the second term in (46) is also of the order $o_P(\ddot{r}_n)$, so that (42) holds for $\ell = 5$, which proves the result. □

Lemma D.11. *Put* $\hat{\sigma}_{34k} = (M\epsilon_3)_k(M\epsilon_4)_k$ *and* $\hat{\sigma}_{12k} = (M\epsilon_1)_k(M\epsilon_2)_k$. *Then, if* $\max_i(H_Q)_{ii} \to 0$ *and* $(K+L)/\ddot{r}_n = O(1)$ *a.s.,*

$$
\sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}\hat{\sigma}_{12,j}\hat{\sigma}_{34,i} = \sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}E[\epsilon_{3k}\epsilon_{4k} \mid Q]E[\epsilon_{1k}\epsilon_{2k} \mid Q] + o_P(\ddot{r}_n). \tag{47}
$$

Proof. Decompose the left-hand side of (47) as

$$
\sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}\hat{\sigma}_{12,j}\hat{\sigma}_{34,i} = \sum_i\left[\hat{\sigma}_{34i} - \epsilon_{3i}\epsilon_{4i}\right]T_i + \sum_i[\hat{\sigma}_{12i} - \epsilon_{1i}\epsilon_{2i}]S_i + \sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}\epsilon_{3i}\epsilon_{4i}\epsilon_{1j}\epsilon_{2j}, \tag{48}
$$

where $T_i = \sum_{j\colon j\neq i}(H_{\ddot{Z}})^2_{ij}(M\epsilon_1)_j(M\epsilon_2)_j$, and $S_i = \sum_{j\colon j\neq i}(H_{\ddot{Z}})^2_{ij}\epsilon_{3j}\epsilon_{4j}$. The conditional variance of the third term in (48) satisfies

$$
\begin{aligned}
\operatorname{var}&\left(\sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}\epsilon_{3i}\epsilon_{4i}\epsilon_{1j}\epsilon_{2j} \mid Q\right)\\
&= \sum_{i\neq j}(H_{\ddot{Z}})^4_{ij}\widetilde{\mathcal{V}}_{1ijk} + \sum_{i\neq j\neq k}(H_{\ddot{Z}})^2_{ij}(H_{\ddot{Z}})^2_{ik}\widetilde{\mathcal{V}}_{2ijk} + \sum_{i\neq j\neq k}(H_{\ddot{Z}})^2_{ij}(H_{\ddot{Z}})^2_{jk}\widetilde{\mathcal{V}}_{3ijk}\\
&\preceq_{\text{a.s.}} \sum_{i\neq j}(H_{\ddot{Z}})^4_{ij} + \sum_{i\neq j\neq k}(H_{\ddot{Z}})^2_{ij}(H_{\ddot{Z}})^2_{ik} + \sum_{i\neq j\neq k}(H_{\ddot{Z}})^2_{ij}(H_{\ddot{Z}})^2_{jk} \leq 3K\max_i (H_{\ddot{Z}})_{ii}.
\end{aligned}
$$

where

$$
\begin{aligned}
\widetilde{\mathcal{V}}_{1ijk} &= \sigma_{3344i}\sigma_{1122j} + \sigma_{1234i}\sigma_{1234j} - \sigma^2_{34i}\sigma^2_{12j} - \sigma_{34i}\sigma_{12j}\sigma_{34j}\sigma_{12i},\\
\widetilde{\mathcal{V}}_{2ijk} &= \sigma_{3344i}\sigma_{12j}\sigma_{12k} + \sigma_{1234i}\sigma_{12j}\sigma_{34k} - \sigma_{34i}\sigma_{12j}\sigma_{34k}\sigma_{12i} - \sigma_{34i}\sigma_{12j}\sigma_{34i}\sigma_{12k},\\
\widetilde{\mathcal{V}}_{3ijk} &= \left(\sigma_{34i}\sigma_{1122j}\sigma_{34k} + \sigma_{1234j}\sigma_{34i}\sigma_{12k} - \sigma_{34i}\sigma_{12j}\sigma_{34k}\sigma_{12j} - \sigma_{34i}\sigma_{12j}\sigma_{34j}\sigma_{12k}\right).
\end{aligned}
$$

Therefore, by Markov inequality,

$$
\sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}\epsilon_{3i}\epsilon_{4i}\epsilon_{1j}\epsilon_{2j} = E_Q\sum_{i\neq j}(H_{\ddot{Z}})^2_{ij}\epsilon_{3i}\epsilon_{4i}\epsilon_{1j}\epsilon_{2j} + o_P(\ddot{r}_n) = \sum_{i\neq j}\sigma_{34i}\sigma_{12j}(H_{\ddot{Z}})^2_{ij} + o_P(\ddot{r}_n).
$$

To prove the claim of the Lemma, it therefore suffices to show that the first and second terms in (48) are of the order $o_P(\ddot{r}_n)$. To that end, note that by Cauchy-Schwarz inequality,

$$
(E_Q|\sum_i\left[\hat{\sigma}_{34i} - \epsilon_{3i}\epsilon_{4i}\right]T_i|)^2 \leq \sum_i\left[\hat{\sigma}_{34i} - \epsilon_{3i}\epsilon_{4i}\right]^2 \cdot \sum_i E_Q T_i^2 \tag{49}
$$

If we can show that the right-hand side is of smaller order than $\ddot{r}_n^2$, then it follows by Markov inequality that the first term in (48) is order the order $o_P(\ddot{r}_n)$. It follows from equation (39) in the proof of Lemma D.10 that $E_Q\sum_i\left[\hat{\sigma}_{34i} - \epsilon_{3i}\epsilon_{4i}\right]^2 \preceq_{\text{a.s.}} K + L$. By Lemma D.8,

$$
\begin{aligned}
\sum_i E_Q T_i^2 &= \sum_i\sum_{j\colon j\neq i}(H_{\ddot{Z}})^2_{ij}\sum_{k\colon k\neq i}(H_{\ddot{Z}})^2_{ik}E_Q(M\epsilon_1)_j(M\epsilon_2)_j(M\epsilon_1)_k(M\epsilon_2)_k\\
&\preceq_{\text{a.s.}} \sum_{i,j,k}(H_{\ddot{Z}})^2_{ij}(H_{\ddot{Z}})^2_{ik} \leq K\max_i (H_{\ddot{Z}})_{ii},
\end{aligned}
$$

so that the right-hand side of (49) is of the order $o(\ddot{r}_n^2)$ as claimed. By similar arguments,

$$
E_Q|\sum_{i\neq j}(H_{\ddot{Z}})_{ij}\epsilon_{3i}\epsilon_{4j}S_i|^2 \leq \sum_i E_Q S_i^2 \cdot \sum_i E_Q[\hat{\sigma}_{12i} - \epsilon_{1i}\epsilon_{2i}] \preceq_{\text{a.s.}} (K+L)\sum_i E_Q S_i^2.
$$

Since

$$E_Q S_i^2 = \sum_i \sum_{j:\, j\neq i} \sum_{k:\, k\neq i} (H_{\ddot{Z}})_{ij}(H_{\ddot{Z}})_{ik} E_Q \epsilon_{3j}\epsilon_{4j}\epsilon_{3k}\epsilon_{4k} \preceq_{\text{a.s.}} \sum_{i,j,k} (H_{\ddot{Z}})_{ij}(H_{\ddot{Z}})_{ik} \leq K \max_i (H_{\ddot{Z}})_{ii},$$

it follows by Markov inequality that the second term in (48) is also of the order $o_P(\ddot{r}_n)$. □

Appendix E Proofs of unconditional results

Let $\varrho \equiv E\left[\widetilde{R}_i^2\right]$ and $\psi_i \equiv \Sigma_{WW}^{-1/2} W_i$.

Proof of Theorem 5.2: *1.* First we show that the conditions of the theorem ensure that the conditions of Theorem 5.1 are satisfied w.p.a.1. By Assumption 4 (v), $\|Q_i\| \lesssim \sqrt{K+L}$. Since $\widetilde{Z}_i \equiv Z_i - E\left[Z_i | W_i\right]$, Assumption 4 (iv) implies that $\|W_i\| \lesssim \sqrt{L}$, $\left\|\widetilde{Z}_i\right\| \lesssim \sqrt{K}$, and that the eigenvalues of $\Sigma_{\widetilde{Z}\widetilde{Z}} \equiv E\left[\widetilde{Z}_i \widetilde{Z}_i'\right]$ are uniformly bounded from above and away from zero. Then, $\left\|\hat{\Sigma}_{QQ}\right\|_\lambda + \left\|\hat{\Sigma}_{QQ}^{-1}\right\|_\lambda + \left\|\hat{\Sigma}_{WW}\right\|_\lambda + \left\|\hat{\Sigma}_{WW}^{-1}\right\|_\lambda + \left\|\hat{\Sigma}_{\widetilde{Z}\widetilde{Z}}\right\|_\lambda + \left\|\hat{\Sigma}_{\widetilde{Z}\widetilde{Z}}^{-1}\right\|_\lambda \lesssim 1$ by Lemma F.5 and Assumption 4 (i). Also, $\hat{\Sigma}_{\ddot{Z}\ddot{Z}} = \frac{1}{n}\ddot{Z}'\ddot{Z} = \frac{1}{n}\widetilde{Z}'\left(I - H_W\right)\widetilde{Z} = \hat{\Sigma}_{\widetilde{Z}\widetilde{Z}} - E_n\left[\widetilde{Z}_i W_i'\right]\hat{\Sigma}_{WW}^{-1} E_n\left[\widetilde{Z}_i W_i\right]$, so

$$\left\|\hat{\Sigma}_{\ddot{Z}\ddot{Z}} - \hat{\Sigma}_{\widetilde{Z}\widetilde{Z}}\right\|_\lambda = O_p\left(\left\|E_n\left[\widetilde{Z}_i W_i\right]\right\|_\lambda^2\right) = O_p\left(\frac{1}{n}(K+L)\log(K+L)\right), \tag{50}$$

which in particular implies that

$$\left\|\hat{\Sigma}_{\ddot{Z}\ddot{Z}}\right\|_\lambda + \left\|\hat{\Sigma}_{\ddot{Z}\ddot{Z}}^{-1}\right\|_\lambda \lesssim 1. \tag{51}$$

Thus, $\max_{i\leq n} H_{\ddot{Z},ii} \lesssim K/n$, $\max_{i\leq n} H_{W,ii} \lesssim L/n$, and $\max_{i\leq n} H_{Q,ii} \lesssim \left(K+L\right)/n$ w.p.a.1.
2. Next, we show that for each of the estimators,

$$\frac{1}{\widetilde{r}_n} R_A' G R = \frac{1}{\varrho} E\left[\widetilde{R}_{Ai}\widetilde{R}_i\right] + o_p\left(1\right). \tag{52}$$

Equation (52) implies that $\ddot{r}_n/\widetilde{r}_n = 1 + o_P\left(1\right)$, and that $\beta_{\text{C,G}} = \beta_{\text{U,IA94}} + o_P\left(1\right)$.
(i) For TSLS, IJIVE1, and JIVE1 we have: $R_A G_{\text{TSLS}} R = \ddot{R}_A'\ddot{R} = \widetilde{R}_A'\left(I - H_W\right)\widetilde{R} = \widetilde{R}_A'\widetilde{R} - \widetilde{R}_A' H_W \widetilde{R}$. Note that $\lambda_{\max}\left(\hat{\Sigma}_{WW}^{-1}\right) \lesssim 1$ w.p.a.1, $E\left[\widetilde{R}_A W\right] = 0$, $E\left[\left\|\widetilde{R}_{Ai} W_i\right\|^2\right] \lesssim E\left[\widetilde{R}_{Ai}^2\right] L \lesssim \varrho L$, and $\left|\widetilde{R}_A' H_W \widetilde{R}\right| \leq \widetilde{R}_A' H_W \widetilde{R}_A + \widetilde{R}' H_W \widetilde{R}$. For any $A \in \{X, Y, \Delta\}$,

$$\frac{1}{\widetilde{r}_n}\widetilde{R}_A' H_W \widetilde{R}_A = \frac{1}{\varrho} E_n\left[\widetilde{R}_{Ai} W_i'\right]\hat{\Sigma}_{WW}^{-1} E_n\left[W_i \widetilde{R}_{Ai}\right] = O_p\left(\frac{1}{\varrho}\left\|E_n\left[\widetilde{R}_{Ai} W_i'\right]\right\|^2\right) = O_p\left(\frac{L}{n}\right) = o_p\left(1\right).$$

Then equation (52) holds, since $\frac{1}{\widetilde{r}_n}\widetilde{R}_A'\widetilde{R} = \frac{E\left[\widetilde{R}_{Ai}\widetilde{R}_i\right]}{E\left[\widetilde{R}_i^2\right]} + o_P\left(1\right)$ by Assumption 4 (ii) and the LLN.
(ii) For IJIVE2, $R_A' G_{\text{IJIVE2}} R = \ddot{R}_A'\ddot{R} - \ddot{R}_A' D_{\ddot{Z}}\ddot{R}$. Since $\left|\ddot{R}_A' D_{\ddot{Z}}\ddot{R}\right| \leq \max_{i\leq n} H_{\ddot{Z},ii}\left(\ddot{R}_A'\ddot{R}_A + \ddot{R}'\ddot{R}\right)$ the conclusion follows from the arguments for IJIVE1 above.
(iii) For UJIVE, equation (52) follows from Lemma E.3, which shows that $\frac{1}{\widetilde{r}_n} R_A G_{\text{UJIVE}} R = \frac{1}{\widetilde{r}_n}\widetilde{R}_A'\widetilde{R} + o_p\left(1\right)$. Remember that $E\left[\left(\zeta, \eta\right) G_{\text{UJIVE}}\eta \middle| W, Z\right] = 0$, since $G_{ii} = 0$. We have $\frac{1}{\widetilde{r}_n}\left\|\left(\zeta, \eta\right) G_{\text{UJIVE}} R\right\| =$

$\frac{1}{\widetilde{r}_n} O_P\left(\|G_{\text{UJIVE}} R\|\right) = O_P\left(\widetilde{r}_n^{-1/2}\right) = o_P\left(1\right)$, since $\|G_{\text{UJIVE}} R\| = \left\|\left(I - D_W\right)^{-2} \ddot{R}\right\| = \left|\ddot{R}' \left(I - D_W\right)^{-2} \ddot{R}\right|^{1/2}$ $\lesssim_{\text{w.p.a.1}} \left|\ddot{R}'\ddot{R}\right|^{1/2} = \ddot{r}_n^{1/2}$. Then, using part *(i)* we have $\frac{1}{\widetilde{r}_n} \left\|\left(\zeta, \eta\right) G_{\text{UJIVE}} R\right\| = O_P\left(\frac{1}{\widetilde{r}_n}\sqrt{\widetilde{r}_n + \frac{L}{n}}\right) = o_P\left(1\right)$. Also, $\frac{1}{\widetilde{r}_n} \left\|\left(\zeta, \eta\right) G_{\text{UJIVE}} \eta\right\| = o_P\left(1\right)$ by Lemma D.1 and equations (20), since $\|G_{\text{UJIVE}}\|_F \lesssim_{\text{w.p.a.1}} \left(K + L\right)^{1/2} = o\left(\widetilde{r}_n\right)$.

This verifies Assumption 2 (ii) and (iii), that $\ddot{r}_n/\tilde{r} = 1 + o_P\left(1\right)$, and that $\beta_{\text{C,G}} - \beta_{\text{U,IA94}} = o_P\left(1\right)$ for all of the considered estimators. □

Proof of Theorem 5.4:
1. First, we show that Assumption 5 holds. Indeed, by Lemma F.7, $\max_{i \le n} \widetilde{r}_n^{-1/2} \left|\ddot{R}_i - \widetilde{R}_i\right| = o_P\left(1\right)$, and hence, $\left|E_n\left[\ddot{R}_i^4\right] - E_n\left[\widetilde{R}_i^4\right]\right| \lesssim \max_{i \le n}\left|\ddot{R}_i - \widetilde{R}_i\right| \times E_n\left[1 + \left|\widetilde{R}_i\right|^3\right] = o_p\left(\sqrt{\widetilde{r}_n}\right)$ by the LLN. Thus, by Theorem 5.2 the conditions of Theorem 5.3 hold w.p.a.1. Let

$$\kappa_n = \left(\mathcal{V}_{\text{C}} + \mathcal{V}_{\text{MW}}\right)^{-1/2}, \alpha_n = \left(\Omega_{\text{C}} + \Omega_{\text{MW}}\right)^{-1/2}.$$

By Lemma F.1, $\kappa_n/\alpha_n \xrightarrow{p} 1$. Then, the conclusion follows from Lemma F.3 with $A = \hat{\beta}$, $\beta_{\text{C},n} = \beta_{\text{C},G}$, $\beta_{\text{U}} = \beta_{\text{U},G}$, and $\sigma_\beta^2 = \Omega_{\text{E}}$. □

Lemma E.1. *Suppose* $R_A' G R = \widetilde{R}_A' \widetilde{R} + S_{A,n} + o_p\left(\widetilde{r}_n^{1/2}\right)$, *where* $E\left[S_{A,n}\right] = o\left(\widetilde{r}_n\right)$, *and* $V\left[S_{A,n}\right] = o\left(\widetilde{r}_n\right)$, *for* $A \in \{X, Y\}$. *Suppose* $E\left[\left\{\left(\widetilde{R}_{Yi}^2 + \widetilde{R}_i^2\right)\widetilde{R}_i^2\right\}^{1+\delta}\right] \lesssim E\left[\widetilde{R}_i^2\right]^{1+\delta}$. *Then, for any* $s \in \mathbb{R}$ *and* $c \in \mathbb{R}$,

$$\left|E\left[\exp\left\{\mathrm{i}s\sqrt{\widetilde{r}_n}\left(\beta_{C,G} - \beta_{U,G}\right)\right\}\right] \to e^{-s^2\Omega_{\text{E}}/2}\right| \to 0, \text{ and}$$
$$P\left(\sqrt{\widetilde{r}_n}\left(\beta_{C,G} - \beta_{U,G}\right) < c\right) - P\left(N\left(0, \Omega_E\right) < c\right) \to 0, \text{ where}$$
$$\beta_{U,G} \equiv \frac{E\left[\widetilde{R}_{Yi}\widetilde{R}_i\right] + E\left[S_{Y,n}\right]/r_n}{E\left[\widetilde{R}_i^2\right] + E\left[S_{X,n}\right]/r_n}, \Omega_E = \frac{E\left[(\widetilde{R}_{\Delta i} - \beta_{U,G}\widetilde{R}_i)^2\widetilde{R}_i^2\right]}{E\left[\widetilde{R}_i^2\right]}.$$

Proof. Since $S_{A,n} = E\left[S_{A,n}\right] + o_p\left(\widetilde{r}_n^{1/2}\right)$, $\widetilde{R}_Y'\widetilde{R} = O_p\left(\widetilde{r}_n\right)$, and $\widetilde{R}'\widetilde{R} = \widetilde{r}_n + O_p\left(\widetilde{r}_n^{1/2}\right)$, we have

$$\begin{aligned}\beta_{\text{C,G}} &= \frac{\widetilde{R}_Y'\widetilde{R}/r_n + E\left[S_{Y,n}\right]/r_n + o_p\left(\widetilde{r}_n^{-1/2}\right)}{\widetilde{R}'\widetilde{R}/r_n + E\left[S_{X,n}\right]/r_n + o_p\left(\widetilde{r}_n^{-1/2}\right)} \\ &= \beta_{\text{U,G}} - \left(\widetilde{R}_Y - \beta_{\text{U,G}}\widetilde{R}\right)'\widetilde{R}/r_n + E\left[S_{Y,n} - \beta_{\text{U,G}}S_{X,n}\right]/r_n + o_p\left(\widetilde{r}_n^{-1/2}\right).\end{aligned}$$

The conclusion of the Lemma now follows from

$$E\left[\left(\widetilde{R}_Y - \beta_{\text{U,G}}\widetilde{R}\right)'\widetilde{R}/r_n + E\left[S_{Y,n} - \beta_{\text{U,G}}S_{X,n}\right]/r_n\right] = 0,$$

Lyapunov CLT, and

$$V\left[\widetilde{r}_n^{-1/2}\left(\widetilde{R}_Y-\beta_{\text{U,G}}\widetilde{R}\right)'\widetilde{R}\right]=\frac{E\left[\left(\widetilde{R}_{\Delta i}-\beta_{\text{U,G}}\widetilde{R}_i\right)^2\widetilde{R}_i^2\right]}{E\left[\widetilde{R}_i^2\right]}=\frac{E\left[\left(\widetilde{R}_{Yi}-\beta_{\text{U,IA94}}\widetilde{R}_i\right)^2\widetilde{R}_i^2\right]}{E\left[\widetilde{R}_i^2\right]}\left(1+o\left(1\right)\right).$$

□

E.1 Unconditional Expansions of Estimators

Lemma E.2. *Suppose*

(i) $L^4\log^2 L=o\left(n^3\right)$.

(ii) $\|W_i\|\leq C\sqrt{L}$.

(iii) $E\left[\left.\widetilde{R}_{Yi}^2+\widetilde{R}_i^2\right|W_i\right]+E\left[\left.\widetilde{R}_{Yi}^4+\widetilde{R}_i^4\right|W_i\right]^{1/2}\leq CE\left[\widetilde{R}_i^2\right]$ *a.s.*

(iv) $E\left[\left\{\left(\widetilde{R}_{Yi}^2+\widetilde{R}_i^2\right)\widetilde{R}_i^2\right\}^{1+\delta}\right]\lesssim E\left[\widetilde{R}_i^2\right]^{1+\delta}$.

(v) $\lambda_{\psi,n}/n^3=o\left(1\right)$.

Then,

$$R_A'G_{\text{IJIVE1}}R=R_A'G_{\text{TSLS}}R=\widetilde{R}_A'\widetilde{R}+S_{A,n}+o_p\left(\widetilde{r}_n^{1/2}\right),\text{ with } S_{A,n}\equiv E\left[\widetilde{R}_{Ai}\widetilde{R}_i\left\|\psi_i\right\|^2\right].$$

Proof. 1. Write $R_Y'G_{\text{TSLS}}R=\widetilde{R}_Y'\left(I-H_W\right)\widetilde{R}=\widetilde{R}_Y'\widetilde{R}-\widetilde{R}_Y'D_W\widetilde{R}-\widetilde{R}_Y'\left(H_W-D_W\right)\widetilde{R}$. We will show that

$$\widetilde{R}_Y'D_W\widetilde{R}=E\left[\widetilde{R}_{Yi}\widetilde{R}_i\left\|\psi_i\right\|^2\right]+o_p\left(\widetilde{r}_n^{1/2}\right),\tag{53}$$

$$\widetilde{R}_Y'\left(H_W-D_W\right)\widetilde{R}=o_p\left(\widetilde{r}_n^{1/2}\right).\tag{54}$$

2. We use Lemma D.1 to establish equation (54), taking $\mathcal{Z}_n=W$, $P=\widetilde{r}_n^{-1/2}\left(H_W-D_W\right)$, $u=E\left[\widetilde{R}_i^2\right]^{-1/2}\widetilde{R}_Y$, $v=E\left[\widetilde{R}_i^2\right]^{-1/2}\widetilde{R}$, and $s=t=0$. By the triangle inequality $\|H_W-D_W\|_F\leq 2\sqrt{L}$, so $\widetilde{r}_n^{-1/2}\widetilde{R}_Y'\left(H_W-D_W\right)\widetilde{R}=O_p\left(\sqrt{E\left[\widetilde{R}_i^2\right]L/n}\right)=o_p\left(1\right)$.

3. Let $\rho_i\equiv E\left[\left.\widetilde{R}_{Yi}\widetilde{R}_i\right|W_i\right]$, then $E\left[\left.\widetilde{R}_Y'D_W\widetilde{R}\right|W\right]=\sum_i\rho_iH_{W,ii}$ and

$$\begin{aligned}V\left[\left.\widetilde{R}_Y'D_W\widetilde{R}\right|W\right]&=\sum_i V\left[\left.\widetilde{R}_{Yi}\widetilde{R}_i\right|W_i\right]H_{W,ii}^2\\&\leq\sum_i\left(E\left[\left.\widetilde{R}_{Yi}^2\widetilde{R}_i^2\right|W_i\right]+E\left[\left.\widetilde{R}_{Yi}\widetilde{R}_i\right|W_i\right]^2\right)H_{W,ii}^2\\&\lesssim\sup_w\left(E\left[\left.\widetilde{R}_{Yi}^4+\widetilde{R}_i^4\right|W_i=w\right]+E\left[\left.\widetilde{R}_{Yi}^2+\widetilde{R}_i^2\right|W_i=w\right]^2\right)\sum_i H_{W,ii}^2\\&\lesssim E\left[\widetilde{R}_i^2\right]^2\sum_i H_{W,ii}^2,\end{aligned}$$

where the last equality follows by condition (iii). Condition (ii) and Lemma F.5 imply that w.p.a.1

$$\sum_i H_{W,ii}^2 \leq \left(\max_{i \leq n} \frac{1}{n} W_i' \hat{\Sigma}_{WW}^{-1} W_i \right) \sum_i H_{W,ii} \leq CL^2/n.$$

Hence $\widetilde{r}_n^{-1} V \left[\widetilde{R}_Y' D_W \widetilde{R} \middle| W \right] \lesssim E\left[\widetilde{R}_i^2\right] L^2/n^2 = o\left(1\right)$ and

$$\widetilde{R}_Y' D_W \widetilde{R} = \sum_i \rho_i H_{W,ii} + o_p\left(\widetilde{r}_n^{1/2}\right). \tag{55}$$

4. Note that $H_W = H_\psi$, and let $S \equiv 2I - \hat{\Sigma}_{\psi\psi}$ be an approximate inverse of $\hat{\Sigma}_{\psi\psi}$. Then $\left\|\hat{\Sigma}_{\psi\psi}^{-1} - S\right\|_\lambda \leq \left\|I - \hat{\Sigma}_{\psi\psi}\right\|_\lambda^2 \left\|\hat{\Sigma}_{\psi\psi}^{-1}\right\| = O_p\left(L \log L/n\right)$, where the last equality follows by Lemma F.5. Then

$$\sum_i \rho_i H_{W,ii} = \sum_i \rho_i H_{\psi,ii} = \frac{1}{n} \sum_i \rho_i \psi_i' \hat{\Sigma}_{\psi\psi}^{-1} \psi_i = \frac{1}{n} \sum_i \rho_i \psi_i' \psi_i + A_{1n} - A_{2n}, \tag{56}$$

where

$$A_{1n} \equiv \frac{1}{n} \sum_i \rho_i \psi_i' \left(\hat{\Sigma}_{\psi\psi}^{-1} - S\right) \psi_i, \qquad A_{2n} \equiv \frac{1}{n} \sum_i \rho_i \psi_i' \left(I - \hat{\Sigma}_{\psi\psi}\right) \psi_i.$$

Here

$$\begin{aligned} \left|A_{1n}\right| &\leq \left\|\hat{\Sigma}_{\psi\psi}^{-1} - S\right\|_\lambda \frac{1}{n} \sum_i \left|\rho_i\right| \left\|\psi_i\right\|^2 = \left\|\hat{\Sigma}_{\psi\psi}^{-1} - S\right\|_\lambda \left(E\left[\left|\rho_i\right| \left\|\psi_i\right\|^2\right] + o_p\left(1\right)\right) \\ &= O_p\left(\frac{L \log L}{n} \left(E\left[\widetilde{R}_i^2\right] L\right)\right), \end{aligned} \tag{57}$$

where the last equality holds because $E\left[\left|\widetilde{R}_{Yi} \widetilde{R}_i\right| \left\|\psi_i\right\|^2\right] \lesssim E\left[\widetilde{R}_i^2\right] L$ by conditions (ii) and (iii). Thus, by condition (i), $\widetilde{r}_n^{-1/2} \left|A_{1n}\right| = O_p\left(E\left[\widetilde{R}_i^2\right]^{1/2} \frac{L^2 \log L}{n^{3/2}}\right) = o_p\left(1\right)$.

Next, let $u_{ij} \equiv \rho_i \left\|\psi_i\right\|^2 + \rho_j \left\|\psi_j\right\|^2 - \left(\rho_i + \rho_j\right) \left(\psi_i' \psi_j\right)^2$. Then

$$\begin{aligned} A_{2n} &= \frac{1}{n^2} \sum_i \sum_j \left\{ \rho_i \left\|\psi_i\right\|^2 - \rho_i \left(\psi_i' \psi_j\right)^2 \right\} \\ &= \frac{1}{n} E_n\left[\rho_i \left(\left\|\psi_i\right\|^2 - \left\|\psi_i\right\|^4\right)\right] + \frac{1}{n^2} \sum_{i<j} u_{ij} = O_p\left(E\left[\widetilde{R}_i^2\right] \frac{L^2}{n}\right) + \frac{1}{n^2} \sum_{i<j} u_{ij}, \end{aligned}$$

where the last equality makes use of conditions (ii) and (iii). For the U-statistic term we have $E\left[u_{ij}\right] = 0$,

$$\begin{aligned} E[u_{ij}|\psi_j] &= E\left[\rho_i \left\|\psi_i\right\|^2\right] + \rho_j \left\|\psi_j\right\|^2 - \psi_j' E\left[\rho_i \psi_i \psi_i'\right] \psi_j - \rho_j \left\|\psi_j\right\|^2 \\ &= E\left[\rho_i \left\|\psi_i\right\|^2\right] - \psi_j' E\left[\rho_i \psi_i \psi_i'\right] \psi_j, \\ V\left[u_{ij}\right] &= E\left[u_{ij}^2\right] \leq 4E\left[\rho_i^2 \left(\left\|\psi_i\right\|^2 - \left(\psi_i' \psi_j\right)^2\right)^2\right] = 4E\left[\rho_i^2 \left(\left\|\psi_i\right\|^4 + \left(\psi_i' \psi_j\right)^4\right)\right] \\ &\lesssim E\left[\widetilde{R}_i^2\right]^2 \left(E\left[\left\|\psi_i\right\|^4\right] + E\left[\left(\psi_i' \psi_j\right)^4\right]\right) \lesssim E\left[\widetilde{R}_i^2\right]^2 \left(L^2 + \lambda_{\psi,n}\right), \end{aligned}$$

and

$$V\left[E[u_{ij}|\psi_j]\right] \lesssim E\left[\widetilde{R}_i^2\right]^2 L^2.$$

By the formula for the variance of a U-statistic we have

$$V\left[\frac{1}{n^2}\sum_{i<j} u_{ij}\right] \lesssim \frac{1}{n} V\left[E[u_{ij}|\psi_j]\right] + \frac{1}{n^2} V\left[u_{ij}\right] \lesssim E\left[\widetilde{R}_i^2\right]^2 \left(\frac{1}{n}L^2 + \frac{1}{n^2}\lambda_{\psi,n}\right).$$

Combining these we have

$$A_{2n} = O_p\left(E\left[\widetilde{R}_i^2\right]\frac{L^2}{n} + E\left[\widetilde{R}_i^2\right]\frac{L}{\sqrt{n}} + \frac{1}{n}\lambda_{\psi,n}^{1/2}\right) = O_p\left(\widetilde{r}_n^{1/2}\left\{\frac{L^4+\lambda_{\psi,n}}{n^3}\right\}^{1/2}\right) = o_p\left(\widetilde{r}_n^{1/2}\right), \quad (58)$$

where the last equality follows by conditions (i) and (v).
5. Combining equations (55)-(58) we obtain

$$\widetilde{R}_Y' D_W \widetilde{R} = \frac{1}{n}\sum_i \rho_i \left\|\psi_i\right\|^2 + o_p\left(\widetilde{r}_n^{1/2}\right).$$

Here $E_n\left[\rho_i \left\|\psi_i\right\|^2\right] = E\left[\rho_i \left\|\psi_i\right\|^2\right] + O_p\left(E\left[\widetilde{R}_i^2\right] L/\sqrt{n}\right) = E\left[\rho_i \left\|\psi_i\right\|^2\right] + o_p\left(\widetilde{r}_n^{1/2}\right)$, and hence equation (53) holds, which concludes the proof. □

Lemma E.3. *Suppose* $1/C \leq \lambda_{\min}\left(E\left[\psi_i\psi_i'\right]\right) \leq \lambda_{\max}\left(E\left[\psi_i\psi_i'\right]\right) \leq C$, $\|\psi_i\| \leq C\sqrt{L}$, *with* $L\log L = o\left(n\right)$, W *includes the constant,* $\max_{i\leq n} E\left[\left.\widetilde{R}_{Yi}^2 + \widetilde{R}_i^2\right| W_i\right] \leq C\widetilde{r}_n/n$ *and* $E\left[R_{Yi}^2 + R_i^2\right] \leq C$. *Then* $\left(R_Y, R\right)' G_{\text{UJIVE}} R = \left(\widetilde{R}_Y, \widetilde{R}\right)' \widetilde{R} + o_p\left(\sqrt{\widetilde{r}_n}\right)$.

Proof. By the invariance of the estimators to invertible linear transformations we can w.l.o.g. take $W_i = \psi_i$. The conditions of the Lemma imply that $\max_{i\leq n}\left(H_W\right)_{ii} = O_p\left(L/n\right) = o_p\left(1\right)$, and $\left\|E_n\left[W_iW_i'\right]\right\|_\lambda \leq C$. It is sufficient to consider only $R_Y = \widetilde{R}_Y + W\theta_Y$:

$$\begin{aligned} R_Y' G_{\text{UJIVE}} R &= R_Y'\left(I - D_W\right)^{-1}\ddot{R} \\ &= \ddot{R}_Y'\ddot{R} + \widetilde{R}_Y' D_W\left(I - D_W\right)^{-1}\ddot{R} + \theta_Y' W' D_W\left(I - D_W\right)^{-1}\ddot{R} \\ &= T_1 + T_2 + T_3. \end{aligned}$$

Here, $n^{-1/2}\left\|\left(H_W - D_W\right)\right\|_F = o_p\left(1\right)$, hence by Lemma D.1 with $\mathcal{Z}_n = W$ we have $\widetilde{r}_n^{-1/2}\widetilde{R}_Y'\left(H_W - D_W\right)\widetilde{R} = o_P\left(1\right)$, and hence

$$T_1 = \ddot{R}_Y'\ddot{R} = \widetilde{R}_Y'\left(I - D_W\right)\widetilde{R} + o_p\left(\sqrt{\widetilde{r}_n}\right).$$

Likewise, since $n^{-1/2}\left\|D_W\left(I - D_W\right)^{-1}\left(H_W - D_W\right)\right\|_F = o_p\left(1\right)$ we have

$$T_2 = \widetilde{R}_Y' D_W\left(I - D_W\right)^{-1}\left(I - H_W\right)\widetilde{R} = \widetilde{R}_Y' D_W\left(I - D_W\right)^{-1}\left(I - D_W\right)\widetilde{R} + o_p\left(\sqrt{\widetilde{r}_n}\right)$$

$$= \widetilde{R}_Y' D_W \widetilde{R} + o_p\left(\sqrt{\widetilde{r}_n}\right).$$

Thus, $T_1 + T_2 = \widetilde{R}_Y' \widetilde{R} + o_p\left(\sqrt{\widetilde{r}_n}\right)$.

Consider $T_3 = \theta_Y' W' D_W \left(I - D_W\right)^{-1} H_W \widetilde{R}$. Note that $E\left[T_3 | W\right] = 0$ and

$$\begin{aligned} E\left[T_3^2 \middle| W\right] &\lesssim \frac{\widetilde{r}_n}{n} \theta_Y' W' D_W \left(I - D_W\right)^{-1} H_W \left(I - D_W\right)^{-1} D_W W \theta_Y \lesssim \frac{\widetilde{r}_n}{n} \theta_Y' W' D_W H_W D_W W \theta_Y \\ &= \frac{\widetilde{r}_n}{n^2} \theta_Y' W' D_W W \hat{\Sigma}_{WW}^{-1} W' D_W W \theta_Y \lesssim \frac{\widetilde{r}_n}{n^2} \left(1 + o_p\left(1\right)\right) \theta_Y' \left(W' D_W W\right)^2 \theta_Y. \end{aligned}$$

Here $W' D_W W \leq n \max_{i \leq n} \left(H_W\right)_{ii} \cdot E_n\left[W_i W_i'\right]$, and $\|\theta_Y\| \leq C$ since $E\left[R_{Yi}^2\right] \leq C$. Then,

$$\begin{aligned} E\left[T_3^2 \middle| W\right] &\lesssim \frac{\widetilde{r}_n}{n^2} \left(1 + o_p\left(1\right)\right) \left(n \cdot \max_{i \leq n} \left(H_W\right)_{ii}\right)^2 \theta_Y' E_n\left[W_i W_i'\right]^2 \theta_Y \\ &\lesssim \widetilde{r}_n \left(1 + o_p\left(1\right)\right) \cdot \left(\max_i \left(H_W\right)_{ii}\right)^2. \end{aligned}$$

Thus, $T_3 = o_p\left(\sqrt{\widetilde{r}_n}\right)$ if $\max_{i \leq n} H_{W,ii} \to 0$, which completes the proof. □

Appendix F Auxiliary proofs for unconditional results

Lemma F.1. *Suppose Assumption 4 holds,* $E\left[\widetilde{R}_i^4 + \widetilde{R}_{\Delta i}^4\right] \lesssim \varrho^2$, $E\left[\eta_i^2 + \nu_i^2 \middle| Q_i\right] \leq C$, *and* $|corr\left(\eta_i, \nu_i | Q_i\right)| \leq C < 1$. *Then*

$$\left(\frac{\widetilde{\mathcal{V}}_C + \widetilde{\mathcal{V}}_{MW}}{\widetilde{r}_n}\right)^{-1} \left(\Omega_C + \Omega_{MW}\right) \xrightarrow{p} 1,$$

where

$$\begin{aligned} \widetilde{\mathcal{V}}_C &= \sum_i [\ddot{R}_i^2 \sigma_{\nu,i}^2 + \ddot{R}_{\Delta i}^2 \sigma_{\eta,i}^2 + 2 \ddot{R}_i \ddot{R}_{\Delta i} \sigma_{\nu\eta,i}], \\ \widetilde{\mathcal{V}}_{MW} &= \sum_{i \neq j} [(H_{\ddot{Z}})_{ij}^2 \sigma_{\eta,j}^2 \sigma_{\nu,i}^2 + (H_{\ddot{Z}})_{ij} (H_{\ddot{Z}})_{ji} \sigma_{\nu\eta,i} \sigma_{\nu\eta,j}], \\ \Omega_C &= \frac{1}{\varrho} E[(\widetilde{R}_i \nu_i + \widetilde{R}_{\Delta i} \eta_i)^2], \\ \Omega_{MW} &= \frac{1}{\widetilde{r}_n} \operatorname{tr}\left(E[\nu_i^2 g_i g_i'] E[\eta_i^2 g_i g_i'] + E[\nu_i \eta_i g_i g_i']^2\right), \quad g_i \equiv E[\widetilde{Z}_i \widetilde{Z}_i']^{-1/2} \widetilde{Z}_i. \end{aligned}$$

Proof. First, consider $\widetilde{\mathcal{V}}_{\mathrm{C}}$:

$$\begin{aligned} &\frac{1}{\widetilde{r}_n} \widetilde{\mathcal{V}}_{\mathrm{C}} - \frac{1}{\widetilde{r}_n} \sum_i E[(\widetilde{R}_i \nu_i + \widetilde{R}_{\Delta i} \eta_i)^2 | Q] \\ &\qquad = \frac{1}{\widetilde{r}_n} \sum_i [\left(\ddot{R}_i^2 - \widetilde{R}_i^2\right) \sigma_{\nu,i}^2 + \left(\ddot{R}_{\Delta i}^2 - \widetilde{R}_{\Delta i}^2\right) \sigma_{\eta,i}^2 + 2\left(\ddot{R}_i \ddot{R}_{\Delta i} - \widetilde{R}_i \widetilde{R}_{\Delta i}\right) \sigma_{\nu\eta,i}]. \end{aligned}$$

Note that

$$\begin{aligned}\left\|\ddot{R}-\widetilde{R}\right\|^2 &= \left\|-H_W\widetilde{R}\right\|^2 = O_p\left(\varrho L\right)\\ \left\|\ddot{R}+\widetilde{R}\right\|^2 &= \left\|\left(2I-H_W\right)\widetilde{R}\right\|^2 = 4\left\|\widetilde{R}\right\|^2 - 3\widetilde{R}'H_W\widetilde{R} = 4\varrho + o_P\left(\varrho\right).\end{aligned}$$

Then, for any bounded nonrandom a_i, using the above bounds we have

$$E_n\left[\left(\ddot{R}_i^2-\widetilde{R}_i^2\right)a_i\right]^2 \leq E_n\left[\left(\ddot{R}_i-\widetilde{R}_i\right)^2\right]E_n\left[\left(\ddot{R}_i+\widetilde{R}_i\right)^2a_i^2\right] \lesssim \frac{1}{n}\left\|\ddot{R}-\widetilde{R}\right\|^2 E_n\left[\left(\ddot{R}_i+\widetilde{R}_i\right)^2\right] = O_p\left(\varrho^2\frac{L}{n}\right),$$

and hence

$$\frac{1}{\widetilde{r}_n}\sum_i\left(\ddot{R}_i^2-\widetilde{R}_i^2\right)\sigma_{\nu,i}^2 = O_p\left(\sqrt{\frac{L}{n}}\right) = o_p\left(1\right).$$

Likewise, $\frac{1}{\widetilde{r}_n}\sum_i\left(\ddot{R}_{\Delta i}^2-\widetilde{R}_{\Delta i}^2\right)\sigma_{\eta,i}^2 = o_p\left(1\right)$. Since $2\left(\ddot{R}_i\ddot{R}_{\Delta i}-\widetilde{R}_i\widetilde{R}_{\Delta i}\right) = \left(\ddot{R}_i+\ddot{R}_{\Delta i}\right)^2-\left(\widetilde{R}_i+\widetilde{R}_{\Delta i}\right)^2-\{\ddot{R}_i^2-\widetilde{R}_i^2\}+\{\ddot{R}_{\Delta i}^2-\widetilde{R}_{\Delta i}^2\}$, by the same arguments we have $\frac{1}{\widetilde{r}_n}\sum_i\sigma_{\nu\eta,i}[\ddot{R}_i\ddot{R}_{\Delta i}-\widetilde{R}_i\widetilde{R}_{\Delta i}] = o_p\left(1\right)$. Thus, we have shown that $\frac{1}{\widetilde{r}_n}\widetilde{\mathcal{V}}_{\mathrm{C}} = \frac{1}{\widetilde{r}_n}\sum_{i=1}^n E[(\widetilde{R}_i\nu_i+\widetilde{R}_{\Delta i}\eta_i)^2|Q_i] + o_p\left(1\right)$. Since $V[E[(\widetilde{R}_i\nu_i+\widetilde{R}_{\Delta i}\eta_i)^2|Q_i]] \lesssim E\left[\widetilde{R}_i^4+\widetilde{R}_{\Delta i}^4\right] \lesssim \varrho^2$ we have

$$\frac{1}{\widetilde{r}_n}\widetilde{\mathcal{V}}_{\mathrm{C}} = \Omega_{\mathrm{C}} + o_p\left(1\right).$$

Here $1 \lesssim \Omega_{\mathrm{C}} \lesssim 1$. Then, by Lemma F.4

$$\frac{1}{\widetilde{r}_n}\left(\widetilde{\mathcal{V}}_{\mathrm{C}}+\widetilde{\mathcal{V}}_{\mathrm{MW}}\right) - \left(\Omega_{\mathrm{C}}+\Omega_{\mathrm{MW}}\right) = o_p\left(1\right).$$

□

We make use of the following simple lemma.

Lemma F.2. *Suppose r.v.* $\zeta_n \equiv E\left[\left|A_n\right|\middle|\,\mathcal{Z}_n\right]$ *satisfies (i)* $\zeta_n \to 0$ *w.p.a.1, and (ii)* ζ_n *is uniformly bounded by a constant. Then* $E\left[\zeta_n\right] \to 0$.

Proof. W.l.o.g. $0 \leq \zeta_n \leq 1$. Suppose $E\left[\zeta_n\right] \not\to 0$, i.e., $\exists\varepsilon > 0 : E\left[\zeta_n\right] \geq \varepsilon$ for all large n. Together with condition (ii) this implies that $P\left[\zeta_n \geq \varepsilon\right] \geq \varepsilon$, which contradicts condition (i). □

Lemma F.3. *Suppose* κ_n *and* $\beta_{C,n}$ *are measurable w.r.t.* $\mathcal{Z}_n$*;* α_n, σ_β, β_U *are nonrandom,* σ_β *and* β_U *are bounded, and, as* $n \to \infty$,

(i) For all $s \in \mathbb{R}$, $E\left[\exp\left\{\mathrm{i}s\kappa_n(A-\beta_{C,n})\right\}\middle|\,\mathcal{Z}_n\right] - e^{-s^2/2} \to 0$ *w.p.a.1.*

(ii) $\kappa_n/\alpha_n = 1 + o_p(1)$ *and* $\alpha_n \to \infty$.

(iii) For all $s \in \mathbb{R}$, $E\left[\exp\left\{\mathrm{i}s\alpha_n\left(\beta_{C,n}-\beta_U\right)\right\}\right] - e^{-s^2\sigma_\beta^2/2} \to 0$.

Then,

$$\left(1+\sigma_{\beta}^{2}\right)^{-1/2}\alpha_{n}\left(A-\beta_{U}\right)\rightarrow_{d}N\left(0,1\right).$$

Proof. Fix any $s\in\mathbb{R}$, and let $\Delta_{\mathcal{Z}_n}\left(s\right)\equiv\left|E\left[e^{\mathrm{i}s\kappa_n(A-\beta_{\mathrm{C},n})}|\mathcal{Z}_n\right]-E\left[e^{\mathrm{i}s\alpha_n(A-\beta_{\mathrm{C},n})}|\mathcal{Z}_n\right]\right|$. Then $\Delta_{\mathcal{Z}_n}\left(s\right)=\left|E\left[\left(e^{\mathrm{i}s(\kappa_n-\alpha_n)(A-\beta_{\mathrm{C},n})}-1\right)e^{\mathrm{i}s\alpha_n(A-\beta_{\mathrm{C},n})}\right|\mathcal{Z}_n\right]\Big|\ \leq\ E\left[\left|e^{\mathrm{i}s(1-\alpha_n/\kappa_n)\kappa_n(A-\beta_{\mathrm{C},n})}-1\right|\Big|\,\mathcal{Z}_n\right]\ =\ o_P\left(1\right).$ Since characteristic functions are bounded, we have

$$\left|E\left[e^{\mathrm{i}s\alpha_n(A-\beta_{\mathrm{C},n})}|\mathcal{Z}_n\right]-e^{-s^2/2}\right|\leq\min\left\{2,\left|E\left[e^{\mathrm{i}s\kappa_n(A-\beta_{\mathrm{C},n})}|\mathcal{Z}_n\right]-e^{-s^2/2}\right|+\Delta_{\mathcal{Z}_n}\left(s\right)\right\}=\min\left\{2,o_p\left(1\right)\right\},$$

and hence $E\left[e^{\mathrm{i}s\alpha_n(A-\beta_{\mathrm{C},n})}\right]-e^{-s^2/2}=o\left(1\right)$, so by (iii), $E\left[e^{\mathrm{i}s\alpha_n(A-\beta_{\mathrm{U}})}\right]=E\left[e^{-s^2/2}e^{\mathrm{i}s\alpha_n(\beta_{\mathrm{C},n}-\beta_{\mathrm{U}})}\right]+o\left(1\right)=e^{-s^2(1+\sigma_\beta^2)/2}+o\left(1\right)$. □

Lemma F.4. *Suppose*

(i) $\exists C:\sup_q E\left[\eta_i^4+\nu_i^4|\,Q_i=q\right]\leq C$.

(ii) $E\left[\left(\eta_i,\nu_i\right)|Q_i\right]=0$.

(iii) $\exists C>0:1/C\leq\lambda_{\min}\left(\Sigma_{QQ}\right)\leq\lambda_{\max}\left(\Sigma_{QQ}\right)\leq C$.

(iv) $\left\|\hat{\Sigma}_{QQ}-\Sigma_{QQ}\right\|_{\lambda}=o_p\left(1\right)$.

(v) $K+L=O\left(\widetilde{r}_n\right)$.

Then

$$\frac{1}{\widetilde{r}_n}\left(\widetilde{\mathcal{V}}_{MW}-\widetilde{V}_{MW}\right)=o_p\left(1\right),$$

where

$$\widetilde{\mathcal{V}}_{MW}=\sum_{i\neq j}[(H_{\ddot{Z}})_{ij}^2\sigma_{\eta,j}^2\sigma_{\nu,i}^2+(H_{\ddot{Z}})_{ij}(H_{\ddot{Z}})_{ji}\sigma_{\nu\eta,i}\sigma_{\nu\eta,j}],$$
$$\widetilde{V}_{MW}=\mathrm{tr}\left(E\left[g_ig_i'\nu_i^2\right]E\left[g_ig_i'\eta_i^2\right]+E\left[g_ig_i'\nu_i\eta_i\right]^2\right),\ g_i\equiv\Sigma_{\widetilde{Z}\widetilde{Z}}^{-1/2}\widetilde{Z}_i.$$

Proof. 1. Consider

$$\widetilde{\mathcal{V}}_{ab}\equiv\sum_{i\neq j}H_{\ddot{Z},ij}^2a_ib_j,\text{ and }\widetilde{\mathcal{V}}_{ab,2}\equiv\sum_{i,j}H_{\ddot{Z},ij}^2a_ib_j,$$

for some bounded sequences a_i and b_i. First, by equation (50), $H_{\ddot{Z},ii}=\frac{1}{n}\ddot{Z}_i'\hat{\Sigma}_{\ddot{Z}\ddot{Z}}^{-1}\ddot{Z}_i\lesssim\frac{1}{n}\left\|\ddot{Z}_i\right\|^2\lesssim\frac{\kappa}{n}$ w.p.a.1. Then $\sum_i H_{\ddot{Z},ii}^2\left|a_i\right|\lesssim\frac{\kappa}{n}\sum_i H_{\ddot{Z},ii}=\frac{\kappa}{n}K$ w.p.a.1, and hence $\left|\widetilde{\mathcal{V}}_{ab,2}-\widetilde{\mathcal{V}}_{ab}\right|\equiv\left|\sum_i H_{\ddot{Z},ii}^2a_ib_i\right|\lesssim O_p\left(\frac{\kappa}{n}K\right)=o_p\left(\widetilde{r}_n\right)$. Thus,

$$\left|\widetilde{\mathcal{V}}_{\mathrm{MW},2}-\widetilde{\mathcal{V}}_{\mathrm{MW}}\right|=o_p\left(\widetilde{r}_n\right),$$

where, denoting $\hat{g}_i\equiv\hat{\Sigma}_{\ddot{Z}\ddot{Z}}^{-1/2}\ddot{Z}_i$,

$$\widetilde{\mathcal{V}}_{\mathrm{MW},2}\equiv\sum_{i,j}H_{\ddot{Z},ij}^2\left(\sigma_{\nu,i}^2\sigma_{\eta,j}^2+\sigma_{\nu\eta,i}\sigma_{\nu\eta,j}\right)=\frac{1}{n^2}\sum_{i,j}\mathrm{tr}\left(\hat{\Sigma}_{\ddot{Z}\ddot{Z}}^{-1}\ddot{Z}_i\ddot{Z}_i'\hat{\Sigma}_{\ddot{Z}\ddot{Z}}^{-1}\ddot{Z}_j\ddot{Z}_j'\left(\sigma_{\nu,i}^2\sigma_{\eta,j}^2+\sigma_{\nu\eta,i}\sigma_{\nu\eta,j}\right)\right)$$

$$
\begin{aligned}
&= \operatorname{tr}\left\{\hat{\Sigma}^{-1}_{\ddot{Z}\ddot{Z}} E_n\left[\ddot{Z}_i\ddot{Z}_i'\sigma^2_{\nu,i}\right]\hat{\Sigma}^{-1}_{\ddot{Z}\ddot{Z}} E_n\left[\ddot{Z}_j\ddot{Z}_j'\sigma^2_{\eta,j}\right]\right\} + \operatorname{tr}\left\{\hat{\Sigma}^{-1}_{\ddot{Z}\ddot{Z}} E_n\left[\ddot{Z}_j\ddot{Z}_j'\sigma_{\nu\eta,j}\right]\hat{\Sigma}^{-1}_{\ddot{Z}\ddot{Z}} E_n\left[\ddot{Z}_i\ddot{Z}_i'\sigma_{\nu\eta,i}\right]\right\} \\
&= \operatorname{tr}\left\{E_n\left[\hat{g}_i\hat{g}_i'\sigma^2_{\nu,i}\right]E_n\left[\hat{g}_j\hat{g}_j'\sigma^2_{\eta,j}\right]\right\} + \operatorname{tr}\left\{\left(E_n\left[\hat{g}_i\hat{g}_i'\sigma_{\nu\eta,i}\right]\right)^2\right\}.
\end{aligned}
$$

2. For a random variable ξ_i with $\sup_q E\left[\xi_i^2\right|Q_i = q]$ uniformly bounded, let $\mu_{\xi i} \equiv E\left[\xi_i|Q_i\right]$. Then

$$
-I_{K+L} \lesssim -\sup_q E\left[|\xi_i|\right| Q_i = q] E\left[Q_iQ_i'\right] \lesssim E\left[Q_iQ_i'\xi_i\right] \lesssim \sup_q E\left[|\xi_i|\right| Q_i = q] E\left[Q_iQ_i'\right] \lesssim I_{K+L},
$$

where the inequalities are in the matrix sense. Thus, $\left\|E\left[Q_iQ_i'\xi_i\right]\right\|_\lambda \lesssim 1$, $\left\|E\left[Q_iQ_i'\xi_i\right]\right\|_F \lesssim \sqrt{K+L}$, and by Lemmas F.5 and F.6,

$$
\begin{aligned}
\left\|E_n\left[\ddot{Z}_i\ddot{Z}_i'\mu_{\xi i}\right] - E\left[\widetilde{Z}_i\widetilde{Z}_i'\xi_i\right]\right\|_F &\leq O_p(1)\left\|E_n\left[Q_iQ_i'\mu_{\xi i}\right] - E\left[Q_iQ_i'\xi_i\right]\right\|_F + o_p\left(\sqrt{K+L}\right) \\
&\leq O_p\left(n^{-1/2}(K+L)\right) + o_p\left(\sqrt{K+L}\right) = o_p\left(\sqrt{K+L}\right). \quad (59)
\end{aligned}
$$

Therefore, $\left\|E_n\left[\ddot{Z}_i\ddot{Z}_i'\mu_{\xi i}\right]\right\|_F = O_p\left(\sqrt{K+L}\right)$. Also, by Lemma F.6 and conditions (iii) and (iv),

$$
\left\|\hat{\Sigma}^{-1}_{\ddot{Z}\ddot{Z}}\right\|_\lambda \leq C \text{ w.p.a.1.} \tag{60}
$$

Then

$$
\left\|E_n\left[\hat{g}_i\hat{g}_i'\mu_{\xi i}\right] - E\left[g_ig_i'\xi_i\right]\right\|_F \leq S_{1,n} + S_{2,n}, \quad \text{where} \tag{61}
$$

$$
\begin{aligned}
S_{1,n} &\equiv \left\|\hat{\Sigma}^{-1/2}_{\ddot{Z}\ddot{Z}} E\left[\widetilde{Z}_i\widetilde{Z}_i'\xi_i\right]\hat{\Sigma}^{-1/2}_{\ddot{Z}\ddot{Z}} - \Sigma^{-1/2}_{\widetilde{Z}\widetilde{Z}} E\left[\widetilde{Z}_i\widetilde{Z}_i'\xi_i\right]\Sigma^{-1/2}_{\widetilde{Z}\widetilde{Z}}\right\|_F, \\
S_{2,n} &\equiv \left\|\hat{\Sigma}^{-1/2}_{\ddot{Z}\ddot{Z}}\left(E_n\left[\ddot{Z}_i\ddot{Z}_i'\mu_{\xi i}\right] - E\left[\widetilde{Z}_i\widetilde{Z}_i'\xi_i\right]\right)\hat{\Sigma}^{-1/2}_{\ddot{Z}\ddot{Z}}\right\|_F.
\end{aligned}
$$

Note that $\left\|E\left[g_ig_i'\xi_i\right]\right\|_\lambda \lesssim 1$.

For symmetric $K \times K$ matrices A, $\hat{A}$, S, the following inequality holds

$$
\left\|\hat{A}'S\hat{A} - A'SA\right\|_F \leq \left(2\left\|A\right\|_\lambda + \left\|\hat{A} - A\right\|_\lambda\right)\sqrt{K}\left\|S\right\|_\lambda\left\|\hat{A} - A\right\|_\lambda.
$$

Taking $\hat{A} = \hat{\Sigma}^{-1/2}_{\ddot{Z}\ddot{Z}}$, $A = \Sigma^{-1/2}_{\widetilde{Z}\widetilde{Z}}$, and $S = E\left[\widetilde{Z}_i\widetilde{Z}_i'\xi_i\right]$ and applying the inequality to term $S_{1,n}$ we obtain that

$$
S_{1,n} \lesssim_{\text{w.p.a.1}} \sqrt{K}\left\|\hat{\Sigma}^{-1/2}_{\ddot{Z}\ddot{Z}} - \Sigma^{-1/2}_{\widetilde{Z}\widetilde{Z}}\right\|_\lambda = o_p\left(\sqrt{K}\right), \tag{62}
$$

where we use equation (60) and condition (iii).

Since $\|AB\|_F \leq \|A\|_\lambda \|B\|_F$ for any symmetric A, B. Using equations (59) and (60) we obtain

$$
S_{2,n} \leq \left\|\hat{\Sigma}^{-1/2}_{\ddot{Z}\ddot{Z}}\right\|^2_\lambda \left\|E_n\left[\ddot{Z}_i\ddot{Z}_i'\mu_{\xi i}\right] - E\left[\widetilde{Z}_i\widetilde{Z}_i'\xi_i\right]\right\|_F = o_p\left(\sqrt{K+L}\right). \tag{63}
$$

Together, equations (59), (61)–(63) imply

$$\left\|E_n\left[\hat{g}_i\hat{g}_i'\mu_{\xi i}\right]-E\left[g_ig_i'\xi_i\right]\right\|_F=o_p\left(\sqrt{K+L}\right). \tag{64}$$

3. For any conformable real matrices

$$\operatorname{tr}\left\{\hat{M}_1\hat{M}_2-M_1M_2\right\}\leq\left\|\hat{M}_1-M_1\right\|_F\left(\left\|M_2\right\|_F+\left\|\hat{M}_2-M_2\right\|_F\right)+\left\|M_1\right\|_F\left\|\hat{M}_2-M_2\right\|_F.$$

Thus,

$$\begin{aligned}&\left|\operatorname{tr}\left\{\left(E_n\left[\hat{g}_i\hat{g}_i'\sigma^2_{\nu,i}\right]\right)\left(E_n\left[\hat{g}_j\hat{g}_j'\sigma^2_{\eta,j}\right]\right)-E\left[g_ig_i'\nu_i^2\right]E\left[g_ig_i'\eta_i^2\right]\right\}\right|\\&\leq\left\|E_n\left[\hat{g}_i\hat{g}_i'\sigma^2_{\nu,i}\right]-E\left[g_ig_i'\nu_i^2\right]\right\|_F\left(\left\|E\left[g_ig_i'\eta_i^2\right]\right\|_F+\left\|E_n\left[\hat{g}_j\hat{g}_j'\sigma^2_{\eta,j}\right]-E\left[g_ig_i'\eta_i^2\right]\right\|_F\right)\\&+\left\|E_n\left[\hat{g}_j\hat{g}_j'\sigma^2_{\eta,j}\right]-E\left[g_ig_i'\eta_i^2\right]\right\|_F\left\|E\left[g_ig_i'\nu_i^2\right]\right\|_F\end{aligned} \tag{65}$$

and

$$\begin{aligned}&\left|\operatorname{tr}\left\{E_n\left[\hat{g}_i\hat{g}_i'\sigma_{\nu\eta,i}\right]-E\left[g_ig_i'\nu_i\eta_i\right]^2\right\}\right|\\&\qquad\leq2\left\|E_n\left[\hat{g}_i\hat{g}_i'\sigma_{\nu\eta,i}\right]-E\left[g_ig_i'\nu_i\eta_i\right]\right\|_F\left\|E\left[g_ig_i'\nu_i\eta_i\right]\right\|_F+\left\|E_n\left[\hat{g}_i\hat{g}_i'\sigma_{\nu\eta,i}\right]-E\left[g_ig_i'\nu_i\eta_i\right]\right\|_F^2.\end{aligned}$$

Since $\left\|E\left[g_ig_i'\eta_i^2\right]\right\|_F\lesssim\sqrt{K}$, $\left\|E\left[g_ig_i'\nu_i^2\right]\right\|_F\lesssim\sqrt{K}$, $\left\|E\left[g_ig_i'\nu_i\eta_i\right]\right\|_F\lesssim\sqrt{K}$, and $K+L\lesssim\widetilde{r}_n$ we have

$$\frac{1}{\widetilde{r}_n}\left(\widetilde{\mathcal{V}}_{\text{MW,2}}-\widetilde{V}_{\text{MW}}\right)=o_p\left(\frac{1}{\widetilde{r}_n}\sqrt{K\left(K+L\right)}\right)=o_p\left(1\right).$$

□

Lemma F.5. *Let $A_i\in\mathbb{R}^{m_a}$, $B_i\in\mathbb{R}^{m_b}$ be i.i.d. (for each n) vectors with m_a and m_b allowed to change with n. Suppose for some C and all n, $\lambda_{\max}\left(E\left[A_iA_i'\right]\right)\leq C$ and $\lambda_{\max}\left(E\left[B_iB_i'\right]\right)\leq C$. Then*

$$\left\|\hat{\Sigma}_{AB}-\Sigma_{AB}\right\|_F=O_P\left(E\left[\left\|A_i\right\|^2\left\|B_i\right\|^2\right]^{1/2}/\sqrt{n}\right). \tag{66}$$

If in addition (iii) $\exists C:\left\|A_i\right\|\leq C\sqrt{m_a}$, $\left\|B_i\right\|\leq C\sqrt{m_b}$ *for all n, and $m\log\left(m\right)/n=o\left(1\right)$ for $m\equiv m_A+m_B$ then*

$$\begin{aligned}\left\|\hat{\Sigma}_{AB}-\Sigma_{AB}\right\|_\lambda&=O_P\left(\sqrt{m\log\left(m\right)/n}\right), &(67)\\\left\|\hat{\Sigma}_{AB}-\Sigma_{AB}\right\|_\lambda&\overset{a.s.}{\to}0. &(68)\end{aligned}$$

Proof. Equation (66) is easily verified by a direct calculation. Equation (67) follows from Theorem 1.6 in Tropp (2012). □

Lemma F.6. *Suppose ξ_i and $\hat{\xi}_i$ are some scalar random variables, and* *(i)* $\varphi_{\hat{\theta}_Z,\lambda} = o\left(1\right)$.

(ii) $\exists C > 0 : 1/C \leq \lambda_{\min}\left(\Sigma_{QQ}\right) \leq \lambda_{\max}\left(\Sigma_{QQ}\right) \leq C$.

Then

$$\left\|E_n\left[\ddot{Z}_i\hat{\xi}_i\right] - E\left[\widetilde{Z}_i\xi_i\right]\right\| = O_p\left(1\right)\left\|E_n\left[Q_i\hat{\xi}_i\right] - E\left[Q_i\xi_i\right]\right\| + o_p\left(1\right)\left\|E\left[Q_i\xi_i\right]\right\|,$$
$$\left\|E_n\left[\ddot{Z}_i\ddot{Z}_i'\hat{\xi}_i\right] - E\left[\widetilde{Z}_i\widetilde{Z}_i'\xi_i\right]\right\|_N = O_p\left(1\right)\left\|E_n\left[Q_iQ_i'\hat{\xi}_i\right] - E\left[Q_iQ_i'\xi_i\right]\right\|_N + o_p\left(1\right)\left\|E\left[Q_iQ_i'\xi_i\right]\right\|_N,$$

where norm $\left\|\cdot\right\|_N$ can be Frobenius or spectral norm.

Proof. We prove only the second statement, the proof of the first statement is analogous. Since $\ddot{Z}_i = \left(I_K, -\hat{\theta}_Z'\right)Q_i$, and $\widetilde{Z}_i' = \left(I_K, -\theta_Z'\right)Q_i$ write

$$\begin{aligned}&\left\|E_n\left[\ddot{Z}_i\ddot{Z}_i'\hat{\xi}_i\right] - E\left[\widetilde{Z}_i\widetilde{Z}_i'\xi_i\right]\right\|_N \\ &= \left\|\left(I_K, -\hat{\theta}_Z'\right)E_n\left[Q_iQ_i'\hat{\xi}_i\right]\left(I_K, -\hat{\theta}_Z'\right)' - \left(I_K, -\theta_Z\right)'E\left[Q_iQ_i'\xi_i\right]\left(I_K, -\theta_Z'\right)'\right\|_N \leq T_1 + T_2,\end{aligned} \tag{69}$$

where

$$\begin{aligned}T_1 &\equiv \left\|\left(I_K, -\hat{\theta}_Z'\right)\left(E_n\left[Q_iQ_i'\hat{\xi}_i\right] - E\left[Q_iQ_i'\xi_i\right]\right)\left(I_K, -\hat{\theta}_Z'\right)'\right\|_N, \\ T_2 &\equiv \left\|\left(I_K, -\hat{\theta}_Z'\right)E\left[Q_iQ_i'\xi_i\right]\left(I_K, -\hat{\theta}_Z'\right)' - \left(I_K, -\theta_Z'\right)E\left[Q_iQ_i'\xi_i\right]\left(I_K, -\theta_Z'\right)'\right\|_N.\end{aligned}$$

Consider term T_1. Since $\left\|AB\right\|_N \leq \left\|A\right\|_\lambda\left\|B\right\|_N$, we have

$$T_1 \leq \left\|E_n\left[Q_iQ_i'\hat{\xi}_i\right] - E\left[Q_iQ_i'\xi_i\right]\right\|_N\left\|\left(I_K, -\hat{\theta}_Z'\right)\right\|_\lambda^2.$$

By the triangle inequality, $\left\|\left(I_K, -\hat{\theta}_Z'\right)\right\|_\lambda \leq \left\|\left(I_K, -\theta_Z'\right)\right\|_\lambda + \left\|\hat{\theta}_Z - \theta_Z\right\|_\lambda \leq \left(1 + \left\|\theta_Z\right\|_\lambda\right) + \left\|\hat{\theta}_Z - \theta_Z\right\|_\lambda$.

Since $Z_i' = W_i'\theta_Z + \widetilde{Z}_i'$ with $E\left[W_i\widetilde{Z}_i'\right] = 0$ we have $E\left[Z_iZ_i'\right] = \theta_Z'E\left[W_iW_i'\right]\theta_Z + E\left[\widetilde{Z}_i\widetilde{Z}_i'\right] \geq \theta_Z'\Sigma_{WW}\theta_Z \geq \lambda_{\min}\left(\Sigma_{WW}\right)\theta_Z'\theta_Z$, where the inequalities are in the matrix sense. From condition (ii) it follows that $\lambda_{\max}\left(E\left[Z_iZ_i'\right]\right) \leq C$ and $\lambda_{\min}\left(\Sigma_{WW}\right) \geq C > 0$, and hence the above implies that $\left\|\theta_Z\right\|_\lambda \leq C$. Thus,

$$\left\|\left(I_K, -\theta_Z'\right)\right\|_\lambda \leq C, \tag{70}$$

and hence by condition (i), $T_1 = O_p\left(1\right)\left\|E_n\left[Q_iQ_i'\hat{\xi}_i\right] - E\left[Q_iQ_i'\xi_i\right]\right\|_N$.

Next, consider T_2 in equation (69). For matrices $\hat{A} = \left(I_K, -\hat{\theta}_Z'\right)'$, $A' = \left(I_K, -\theta_Z'\right)'$, and $S = E\left[Q_iQ_i'\xi_i\right]$ we have

$$\begin{aligned}T_2 = \left\|\hat{A}'S\hat{A} - A'SA\right\|_N &\leq \left(2\left\|A\right\|_\lambda + \left\|\hat{A} - A\right\|_\lambda\right)\left\|S\right\|_N\left\|\hat{A} - A\right\|_\lambda \\ &= \left(2\left\|\left(I_K, -\theta_Z'\right)\right\|_\lambda + \left\|\hat{\theta}_Z - \theta_Z\right\|_\lambda\right)\left\|E\left[Q_iQ_i'\xi_i\right]\right\|_N\left\|\hat{\theta}_Z - \theta_Z\right\|_\lambda.\end{aligned}$$

Then, by condition (i) and equation (70), $T_2 = o_p\left(1\right)\left\|E\left[Q_iQ_i'\xi_i\right]\right\|_N$, which concludes the proof. □

Lemma F.7. *Let $A \in \{X, Y, \Delta\}$, and suppose $\|W_i\| \le C\sqrt{L}$, $L \log L = o(n)$, and $E\left[\|W_i\|^2 \widetilde{R}_{Ai}^2\right] \le CL\varrho$. Then*

$$\max_{i\le n} \frac{1}{\sqrt{\widetilde{r}_n}} \left|\ddot{R}_{Ai} - \widetilde{R}_{Ai}\right| = O_p\left(\frac{L}{n}\right) = o_p(1).$$

Proof. From $\ddot{R} - \widetilde{R} = -H_W \widetilde{R}$ it follows that

$$\begin{aligned}\max_{i\le n}\left|\ddot{R}_{Ai}-\widetilde{R}_{Ai}\right| &= \max_{i\le n}\left|\sum_j H_{ij}\widetilde{R}_{Aj}\right| = \max_{i\le n}\left|W_i'\hat{\Sigma}_{WW}^{-1}E_n\left[W_i\widetilde{R}_{Ai}\right]\right| \\ &\le C\sqrt{L}\left\|\hat{\Sigma}_{WW}^{-1}E_n\left[W_i\widetilde{R}_{Ai}\right]\right\| \lesssim (1+o_p(1))\sqrt{L}\left\|E_n\left[W_i\widetilde{R}_{Ai}\right]\right\| = O_p\left(\frac{L\sqrt{\varrho}}{\sqrt{n}}\right).\end{aligned}$$

and hence $\max_{i\le n} \frac{1}{\sqrt{\widetilde{r}_n}}\left|\ddot{R}_{Ai}-\widetilde{R}_{Ai}\right| = O_p\left(\frac{L}{n}\right) = o_p(1)$. □

References

Abadie, A., G. W. Imbens, and F. Zheng (2014): "Inference for Misspecified Models With Fixed Regressors," *Journal of the American Statistical Association*, 109, 1601–1614.

Ackerberg, D. A. and P. J. Devereux (2009): "Improved Jive estimators for overidentified linear models with and without heteroskedasticity," *Review of Economics and Statistics*, 91, 351–362.

Aizer, A. and J. J. J. Doyle (2015): "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges," *The Quarterly Journal of Economics*, 130, 759–803.

Anatolyev, S. (2013): "Instrumental variables estimation and inference in the presence of many exogenous regressors," *The Econometrics Journal*, 16, 27–72.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *Review of Economic Studies*, 67, 499–527.

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife instrumental variables estimation," *Journal of Applied Econometrics*, 14, 57–67.

Angrist, J. D. and A. B. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics*, 106, 979–1014.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62, 657–681.

Blomquist, S. and M. Dahlberg (1999): "Small sample properties of LIML and jackknife IV estimators: experiments with weak instruments," *Journal of Applied Econometrics*, 14, 69–88.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–2781.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2016): "Treatment Effects with Many Covariates and Heteroskedasticity," Working paper, University of Michigan.

Chao, J. C. and N. R. Swanson (2005): "Consistent estimation with a large number of weak instruments," *Econometrica*, 73, 1673–1692.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," *Econometric Theory*, 12, 42–86.

Davidson, R. and J. G. MacKinnon (2006): "Reply to Ackerberg and Devereux and Blomquist and Dahlberg on 'The case against JIVE'," *Journal of Applied Econometrics*, 21, 843–844.

Dobbie, W. and J. Song (2015): "Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection," *American Economic Review*, 105, 1272–1311.

Eagleson, G. K. (1975): "Martingale Convergence to Mixtures of Infinitely Divisible Laws," *The Annals of Probability*, 3, 557–562.

Evdokimov, K. S. and D. Lee (2013): "Diagnostics for Exclusion Restrictions in Instrumental Variables Estimation," Working paper, Princeton University.

Hall, A. R. and A. Inoue (2003): "The large sample behaviour of the generalized method of moments estimator in misspecified models," *Journal of Econometrics*, 114, 361–394.

Heckman, J. J. and E. J. Vytlacil (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the National Academy of Sciences of the United States of America*, 96, 4730–4734.

——— (2005): "Structural equations, treatment effects and econometric policy evaluation," *Econometrica*, 73, 669–738.

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

Kolesár, M. (2013): "Estimation in instrumental variables models with heterogeneous treatment effects," Working Paper, Princeton University.

Kolesár, M., R. Chetty, J. Friedman, E. L. Glaeser, and G. W. Imbens (2015): "Identification and Inference with Many Invalid Instruments," *Journal of Business & Economic Statistics*, 33, 474–484.

Kunitomo, N. (1980): "Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations," *Journal of the American Statistical Association*, 75, 693–700.

Lee, S. (2017): "A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs," *Journal of Business & Economic Statistics*, 1–11.

Maasoumi, E. and P. C. Phillips (1982): "On the behavior of inconsistent instrumental variable estimators," *Journal of Econometrics*, 19, 183–201.

Morimune, K. (1983): "Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size," *Econometrica*, 51, 821–841.

Nagar, A. L. (1959): "The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations," *Econometrica*, 27, 575–595.

Newey, W. K. and F. Windmeijer (2009): "Generalized Method of Moments With Many Weak Moment Conditions," *Econometrica*, 77, 687–719.

Phillips, G. D. A. and C. Hale (1977): "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems," *International Economic Review*, 18, 219–228.

Silver, D. (2016): "Haste or Waste? Peer Pressure and the Distribution of Marginal Returns to Health Care," Working Paper, University of California, Berkeley.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Tropp, J. A. (2012): "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12, 389–434.