INSTRUMENTAL VARIABLES ESTIMATION WITH MANY WEAK INSTRUMENTS USING REGULARIZED JIVE

CHRISTIAN HANSEN AND DAMIAN KOZBUR

Abstract. We consider instrumental variables regression in a setting where the number of instruments is large and the first stage prediction signal is not necessarily sparse. In particular, we work with models where the number of available instruments may be larger than the sample size and consistent model selection in the first stage may not be possible. Such a situation may arise when there are many weak instruments. With many weak instruments, regularization or model selection in the first stage can lead to a large bias relative to standard errors. We propose a jackknife instrumental variables estimator (JIVE) with regularization at each jackknife iteration. We derive the limiting behavior for a ridge-regularized JIV estimator (RJIVE), verifying that the RJIVE is consistent and asymptotically normal under conditions which allow for more instruments than observations and do not require consistent model selection. We provide simulation results that demonstrate the proposed RJIVE performs favorably in terms of size of tests and risk properties relative to other many-weak instrument estimation strategies in high dimensional settings. We also apply the RJIVE to the Angrist and Krueger (1991) example where it performs favorably relative to other many-instrument robust procedures.

Key Words: ridge regression, high-dimensional models, endogeneity

1. Introduction

Instrumental variables (IV) regression is commonly used in economic research to calculate treatment effects for endogenous regressors. While the use of instrumental variables can aid in identification of structural effects, IV estimates of structural effects are often imprecise in practice as only variation in the endogenous variables induced by the instruments is used in estimating the treatment effect. One strategy to increasing the precision of IV estimates is to include many instruments in the hope that these will capture as much exogenous variation in the explanatory variable as possible. The use of many instruments can also be motivated by a desire to nonparametrically estimate the optimal instruments via series as in Newey (1990); see also Amemiya (1974) and Chamberlain (1987). In addition, the increasing availability of high-dimensional data makes it likely that applications where the number of potential instruments is similar to or larger than the number of observations will be increasingly common in applied work; see, for example, the empirical application in Belloni, Chen, Chernozhukov, and Hansen (2010) (BCCH hereafter). While the improvement in efficiency available from using many instruments is appealing, it is well-known that the usual GMM-type approaches to estimating structural

Date: First version: August 2012, this version December 7, 2012.

parameters using instrumental variables, which include IV and 2SLS, may have substantial bias when the number of instruments is not small relative to the sample size; see Bekker (1994) and Newey and Smith (2004). This bias results in poor performance of the usual asymptotic distribution in finite sample simulation experiments and theoretically leads to inconsistency of the 2SLS estimator in an asymptotic sequence where the number of instruments is smaller than but proportional to the sample size.

In this paper, we present an estimation and inference procedure that remains valid in the presence of very many instruments, allowing for more instruments than there are observations. The strategy we consider uses a jackknife to estimate first stage predictions of the endogenous variables. The chief innovation of our procedure is the use of ridge regression at each iteration of the jackknife. The use of ridge regression regularizes the problem and allows us to avoid extreme overfitting of the first-stage even when there are more instruments than observations in the sample. We provide asymptotic theory for the resulting regularized jackknife instrumental variables estimator (RJIVE), giving conditions under which the RJIVE is consistent and asymptotically normal. Importantly, the conditions we impose allow for the number of instruments, K, to be larger than the sample size, n, and do not require that the first-stage relationship between the endogenous variables and instruments be sparse. That is, we allow for very many instruments and do not impose that there is a low-dimensional set of variables that essentially captures the relationship between the instruments and endogenous variables among the set of instruments considered. That we do not assume sparsity allows us to consider scenarios where there are many instruments each of which has a small signal which is not well-separated from zero relative to sampling variation, i.e. scenarios with "many weak instruments." The presence of many weak instruments rules out consistent model selection and first stage estimation. Despite this, the use of the jackknife and regularization allows us to sufficiently avoid overfitting while simultaneously extracting sufficient signal in the first-stage to allow consistent estimation of the structural parameter of interest. In addition to providing theory for our proposal, we provide simulation evidence that suggests that the RJIVE performs well relative to other weak instrument robust procedures and other regularized IV estimators. We also present a brief empirical example.

Our work complements existing approaches to providing estimation and inference strategies that are robust to many instruments. One approach that has been considered is the use of "many-instrument" asymptotics, popularized by Bekker (1994). The goal of many-instrument asymptotics is to provide the approximate behavior of estimators under approximating sequences where the number of instruments is smaller than the sample size but $\frac{K}{n} \to \rho$ where $0 \leq \rho < 1$ and to find estimators that perform well under this approximation. While the traditional 2SLS estimator is inconsistent in this environment, other IV estimators including LIML, Fuller's (1977) modification of LIML (FULL), and jackknife instrumental variables (JIV) remain consistent and asymptotically normal; see Bekker (1994), Chao and Swanson (2005), Hansen, Hausman, and

Newey (2008), and Chao, Swanson, Hausman, Newey, and Woutersen (2012) (CSHNW hereafter).[1] Under the many-instrument sequence, the asymptotic variance of these estimators differs from the asymptotic variance under the usual asymptotics but can be consistently estimated. In simulations, these estimators perform relatively well when the number of instruments is an appreciable fraction of the sample size, and simulation evidence suggests that inference based on these estimators and the many-instrument asymptotic distribution controls size of tests far better than inference based on the usual asymptotic approximation that takes the number of instruments to be small relative to the sample size. A drawback of the many-instrument asymptotic approach is that it requires the number of instruments to be less than the sample size. Many instrument robust estimators also tend to perform poorly in simulations when $\frac{K}{n} \approx 1$; see BCCH. We contribute to this literature by considering cases where $K > n$ and regularization or instrument selection is necessary. Within this setting, we show that the RJIVE retains the desirable asymptotic features derived in CSHNW.

The RJIVE is also related to and complements other many-instrument estimation strategies that make use of first-stage regularization. There is a long history of using first-stage regularization to estimate IV models. One approach to first-stage regularization is using model selection to select a low-dimensional set of instruments for use in fitting the first stage. The idea of instrument selection goes back at least to Kloek and Mennes (1960) and Amemiya (1966). Recently, Bai and Ng (2009) considered using modern variable selection techniques for first-stage instrument selection, and BCCH provide a formal analysis of IV estimators with first-stage fit using methods for fitting high-dimensional-sparse models such as LASSO, e.g. Tibshirani (1996) or Bickel, Ritov, and Tsybakov (2009), or Boosting, e.g. Bühlmann (2006). See also Gautier and Tsybakov (2011) for a different sparsity-based approach to IV estimation related to BCCH. Related ideas also appear in Bai and Ng (2010), Kapetanios and Marcellino (2010), Kapetanios, Khalaf, and Marcellino (2011), and Caner (2009). A common condition in all of these approaches is that the first-stage is sparse; that is, there is a relatively small number of instruments contained within a known set of instruments that provide a good approximation to the relationship between the endogenous variables and instruments. The assumed sparsity in these approaches rules out many-weak-instruments where the signal provided by the instruments is not concentrated among a small number of variables. Donald and Newey (2001) consider a different style of variable selection procedure that minimizes higher-order asymptotic MSE which relies on *a priori* knowledge that allows one to order the instruments in terms of instrument strength. The regularization approach we consider allows us to relax the sparsity requirement and does not require *ex ante* knowledge about the ordering of instruments.

[1] Under many-instrument asymptotics, LIML and FULL are only consistent in the absence of heteroskedasticity. CSHNW provide a JIV estimator that remains consistent under many-instrument asymptotics in the presence of heteroskedasticity.

Our paper is also related to other shrinkage-based approaches to dealing with many instruments. Chamberlain and Imbens (2004) considers IV estimation with many instruments using a shrinkage estimator based on putting a Gaussian random coefficients structure over the first-stage coefficients in a homoskedastic setting which is closely related to using ridge regression in the first-stage. Okui (2010) uses ridge regression for estimating the first-stage regression in a homoskedastic framework where the instruments may be ordered in terms of relevance. Okui (2010) derives the asymptotic distribution of the resulting IV estimator and provides a method for choosing the ridge regression smoothing parameter that minimizes the higher-order asymptotic mean-squared-error (MSE) of the IV estimator. Perhaps the closest paper to our approach is Carrasco (2012) which is based on directly regularizing the inverse that appears in the definition of the 2SLS estimator; see also Carrasco and Tchuente Nguembu (2012). Carrasco (2012) considers three regularization schemes, including Tikhonov regularization which corresponds to ridge regression, and shows that the regularized estimators achieve the semi-parametric efficiency bound under some conditions. The theoretical development in Carrasco (2012) essentially relies on a sparse first-stage structure through restrictions on the covariance structure of the instruments and thus is complementary to the present paper.

2. A Dense High-Dimensional Instrumental Variables Model

In this section, we provide an intuitive discussion of the model we consider. The model is similar to a conventional linear instrumental variables model where interest focuses on structural parameters from a single equation. However, our setup differs from the traditional framework in two key respects. First, we do not assume that the number of instruments is smaller than the sample size and explicitly address the resulting need for regularization. Second, unlike other models that allow for a high-dimensional instrument set, we allow the relationship between the instrments and the endogenous variables to be dense. That is, we do not assume that the signal about the endogenous variables available in the instruments is contained in a small set of possibly unknown variables but allow for a diffuse signal in which all instruments may have individually small contributions that cannot be reliably distinguished from estimation noise.

2.1. The Model.

We consider a model which holds for all observations $i = 1, ..., n$ and all n given by

$$y_i = X_i'\delta_0 + \epsilon_i \tag{2.1}$$

$$X_i = \Upsilon_i + U_i \tag{2.2}$$

where y_i is a scalar outcome of interest, X_i is a G-dimensional treatment variable, and δ_0 is the G-dimensional structural effect of interest. In the model, $\mathrm{E}[U_i\epsilon_i] \neq 0$ which leads to endogeneity of X_i and inconsistency of the conventional regression of y on X. We assume that Υ captures the part of X that is orthogonal to ϵ; that is, we assume $\mathrm{E}[\Upsilon_i\epsilon_i] = 0$. Further, we assume that

$\text{Var}(\Upsilon_i) \neq 0$ and that $\text{E}[U_i|\Upsilon_i] = 0$. Thus, Υ is a valid instrument for X, and we refer to Υ as the optimal instrument. Estimation of δ_0 could be achieved by a straighforward application of classical instrumental variables methods if Υ were observed.

The assumption that one knows the optimal instrument, Υ, seems unrealistic in many situations. To capture this, we assume that y_i and X_i are observed but that Υ_i is not. Rather than assume that Υ is observed, we consider the case where estimation will be based on a K-dimensional instrument Z_i which provides a signal about Υ_i.[2] We focus on the case of an approximately linear signal, $\Upsilon_i \approx Z_i'\pi$, and note that this is a relatively weak restriction since the vector Z_i could consist of a dictionary of transformations of some more elementary variable. E.g. we could have $Z_i = \{p_{kK}(W_i)\}_{k=1}^K$ for some set of basis functions $\{p_{kK}(\cdot)\}$ such as orthogonal polynomials or splines and some set of exogenous variables W_i.

2.2. **High-Dimensional Instruments.** We are particularly interested in the case where the number of instruments in Z_i, K, is large relative to the number of observations in the data, n. For example, many instrumental variables may exist because the set of available instruments itself is high-dimensional as in BCCH or because one is interested in approximating Υ through the use of basis expansions as in Newey (1990), Hansen, Hausman, and Newey (2008), or CSHNW. With many instruments, regularization on the instrument set is desirable as it helps to avoid overfitting of the relationship between the instruments and endogenous variables. It is this overfitting that leads to inconsistency of the standard GMM estimator of δ_0. Estimation procedures that remain valid under many-instrument-asymptotics where $\frac{K}{N} \to \rho < 1$ such as LIML or JIV implicitly make use of regularization to avoid this overfitting. When K is larger than n, these strategies also become inconsistent, and it is clear that further dimension reduction or regularization of the instrument set is necessary for consistent estimation of δ_0 unless U_i and ϵ_i are uncorrelated.

As an illustration, consider the class of estimators used in Hansen, Hausman, and Newey (2008) which include all the so-called k-class estimators except for OLS:

$$\widehat{\delta} = (X'PX - \widehat{\alpha}X'X)^{-1}(X'PY - \widehat{\alpha}X'Y)$$

where $P = Z(Z'Z)^{-}Z'$, A^{-} denotes a generalized inverse of A, X is an $n \times G$ matrix formed by stacking the X_i, Z is an $n \times K$ matrix formed by stacking the Z_i, Y is an $n \times 1$ vector formed by stacking the y_i, and $\widehat{\alpha}$ is specified by the researcher. When $K \geq n$, $X'PX = X'X$ and $X'PY = X'Y$; so, for a fixed $\widehat{\alpha} \neq 1$,[3] $\widehat{\delta}$ reduces to the OLS estimator defined by the

[2]We ignore the presence of included exogenous variables that show up in both the structural equation (2.1) and first-stage equation (2.2). It would be straightforward to accommodate a known, fixed-dimensional vector of such variables. In this case, the variables in model (2.1)-(2.2) and instruments Z_i may be defined as residuals from projecting the observed variables onto the included exogenous variables. An interesting extension would be to consider estimation when the dimension of the included exogenous variables is large relative to the sample size in which case regularization would be needed over the effects of these variables as well.

[3]LIML uses a data-dependent $\widehat{\alpha}$ which is identically equal to one when $K \geq n$.

regression of Y on X and is inconsistent for estimating δ_0 unless $\mathrm{E}[X_i\epsilon_i] = 0$. The class of JIVE estimators considered in CSHNW require $(Z'Z)^{-1}$ in their construction and thus rely on $K \leq n$ in practice.[4] Further regularization relative to what is already implicit in conventional many-instrument-robust procedures may also be desirable in cases where $K < n$ as the additional regularization may produce estimators that are better-behaved in finite samples, especially when $\frac{K}{n}$ is close to one. Finally, note that these estimators may be written as a generic IV estimator

$$\widehat{\delta}_{IV} = (\widehat{\Upsilon}'X)(\widehat{\Upsilon}'Y) \tag{2.3}$$

for $\widehat{\Upsilon} = (P - \widehat{\alpha}I_n)X$ where I_n is the $n \times n$ identity matrix.

2.3. **Regularization with a Dense Signal.** There are, of course, many options available for performing regularization in cases where there are many more available instruments than endogenous variables. Perhaps the most obvious approach to regularization is to directly reduce the number of instruments either through intuition or some more formal mechanism. Any dimension reduction mechanism that makes use of only Z_i without regard to X_i will produce a set of instruments, $\widetilde{Z}_i$ that satisfies the exclusion restriction $\mathrm{E}[\widetilde{Z}_i\epsilon_i] = 0$ if the original instruments satisfied $\mathrm{E}[Z_i\epsilon_i] = 0$. Such approaches include, for example, selecting a set of instruments at random or performing a factor decomposition of Z_i and choosing the first few factors as instruments. The drawback of such approaches is that they may discard some or even all of the signal about the relationship between the instruments and endogenous variables if the correct set of factors is not chosen. Discarding this signal will result in a loss in efficiency and may result in a lack of identification if insufficient signal is maintained.

The desire to maintain the signal available in the instruments while simultaneously regularizing the estimation problem leads to the consideration of data-driven dimension reduction schemes that make use of the information in X_i as well as in the instrument set. A natural approach is to use a variable selection procedure to select a small dimensional set of instruments to then use in a conventional IV-based estimator. Bai and Ng (2009), BCCH , and Gautier and Tsybakov (2011) provide examples of this approach. The formal validity of this approach relies on the assumption that the signal, Υ, may be well-approximated by a sparse model in Z. That is, these approaches make use of a model where $\Upsilon_i = \widetilde{Z}_i'\Pi$ up to a small approximation error where $\widetilde{Z}_i$ is an s-dimensional set of the "relevant" instruments whose identities are unknown and estimated from the data. BCCH show that valid inference for δ_0 may be obtained after doing first-stage variable selection if $s^2/n \to 0$[5] and also show that this condition can be weakened to $s/n \to 0$ if a split-sample procedure is used. Intuitively, sparsity requires that the signal be concentrated among a small set of factors within the set of considered instruments. Since the

[4]Formally, CSHNW require that $(Z'Z/n)$ have minimum eigenvalue bounded away from 0 for n large enough.

[5]The formal condition on the growth rate of instruments is slightly more stringent but would involve the introduction of additional notation.

researcher may specify a very large set of potential instruments, sparsity seems like a reasonable assumption in many cases.[6]

Rather than select instruments, one could also estimate the first-stage relationship using other shrinkage devices. Carrasco (2012) and Carrasco and Tchuente Nguembu (2012) consider a variety of regularization devices designed to directly regularize the inverse of the covariance matrix of the instruments, $(Z'Z/n)$, that shows up in the definition of standard IV estimators. See also Okui (2010) and Chamberlain and Imbens (2004) who take similar approaches. While these approaches are similar to the approach that we take in this paper in that we also use shrinkage rather than variable selection, they differ by relying on conditions that effectively rule out the case of a dense signal and by using the full-sample in constructing the first-stage fit for each observation.[7]

In contrast to these approaches, the estimation method we present in Section 3 allows for a dense signal. For intuition, consider an array of models with only one endogenous variable that has an exactly linear and homoskedastic first-stage, so that $X_i = Z_i'\Pi_n + U_i$ where $\mathrm{E}[U_i^2|Z_i] = \sigma_U^2$. Recall that the concentration parameter

$$\mu_n^2 = \frac{n\Pi_n'\mathrm{E}[Z_iZ_i']\Pi_n}{\sigma_U^2}$$

is a measure of the information available in the instruments and determines the rate of convergence of IV estimators under the usual and many-instrument asymptotics in this case. The concentration parameter satisfies $\mu_n^2/n \geq C$ for some positive constant C under the usual asymptotics which have the dimension of Z_i and the coefficients Π_n fixed. Many-instrument asymptotics in which $\frac{K}{n} < 1$ relaxes this by allowing for cases where $\mu_n^2/n \to 0$ but $\sqrt{K}/\mu_n^2 \to 0$. The sparse model allows $K >> n$[8] but imposes that

$$\frac{n\Pi_n'\mathrm{E}[Z_iZ_i']\Pi_n}{\sigma_U^2} = \frac{n\widetilde{\Pi}_n'\mathrm{E}[\widetilde{Z}_i\widetilde{Z}_i']\widetilde{\Pi}_n}{\sigma_U^2} + O(s)$$

where $\widetilde{Z}_i$ are the relevant instruments, $\dim(\widetilde{Z}_i) = \dim(\widetilde{\Pi}_n) \leq s$ with $s = o(n)$, and $\mu_n^2/n \geq C$. That is, the sparse model allows for very many instruments but requires that the signal in the instruments be concentrated among a small set of instruments and that this signal is sufficiently strong that the usual asymptotics would apply if one knew the correct set of instruments $\widetilde{Z}_i$.[9]

In this paper, we impose a different set of assumptions on the first-stage signal. Our conditions allow for cases where the concentration parameter based on any s dimensional subset of the

[6]BCCH also consider augmenting the set of instruments to be selected over with the fitted value from a ridge-regression of X on Z to accommodate some violations of sparsity.

[7]Such conditions include that the amount of penalization be asymptotically negligible which requires $\frac{K}{n} \to \rho < 1$ and that the covariance operator be compact.

[8]For example, the conditions in BCCH only require that $\log(K) = o(n^{1/3})$.

[9]BCCH also proposes an inference method that remains valid which allows for $\mu_n^2/n \to 0$ or even $\mu_n^2 = 0$.

instruments, $\widetilde{Z}_i$, with $s = o(n)$, $\widetilde{\mu}_n^2 = \frac{n\widetilde{\Pi}_n' \mathrm{E}[\widetilde{Z}_i \widetilde{Z}_i'] \widetilde{\Pi}_n}{\sigma_U^2}$, satisfies $\widetilde{\mu}_n^2/n \to 0$ but $\mu_n^2/n \geq C$. This case allows for scenarios where there is signal available in a combination of the full set of instruments but there is not a small set of instruments which contains the majority of the signal. We refer to this process for the first-stage signal as a "dense first-stage signal" or more simply as a "dense signal." It seems that such cases may occur in practice. For example, the available instrument set may be a large set of dummy variables where there is no obvious reason one would believe the signal concentrates on a few categories and there is no natural way to aggregate the instrument set. Our proposed approach offers a simple, feasible estimation and inference option in cases such as this.

In the case of a dense first-stage signal, sparsity-based estimators may not perform well. As above, suppose for simplicity that X_i has an exact linear relationship with the instruments given by $X_i = \sum_j Z_{ij}\pi_j + U_i$ and the problem is to pick regressors Z_{ij} that have the strongest signals π_j (or collectively the strongest combined signal) to use as instruments. Valid instruments need to satisfy the exclusion condition

$$E[Z_{ij}\epsilon_i] = 0.$$

However, the exclusion restriction will not hold in general for a set of instruments selected from a variable selection procedure that uses X_i since

$$E[Z_{ij}U_i | Z_j \text{ selected }] \neq 0.$$

The failure of the IV exclusion occurs when there are model selection mistakes that select instruments for which the population value of the coefficient, π_j, on the instrument is 0. These model selection mistakes will occur when there is a high within-sample correlation between the instrument and the first-stage error U_i. Given the correlation between the structural error ϵ_i and U_i, these incorrectly selected instruments will then violate the exclusion restriction.

When consistent model selection is possible, for example when a strong sparse signal is present, the magnitude of the expectation $E[Z_{ij}U_i|Z_j \text{ selected}]$ will vanish rapidly enough so as not to affect estimation of the structural effect δ_0. However, with a dense signal this is no longer the case. Consistent variable selection is not feasible with a dense signal as the signals available in the individual instruments are not well-separated from being uninformative. The use of default choices of parameters involved in variable selection schemes such as those given in BCCH will often result in selection of no variables since the default choices are based on the assumption that the signal available in the informative variables is sufficiently strong to dominate the noise. As one relaxes the default parameters, one starts selecting a random set of instruments which consist of informative instruments and the uninformative instruments that are most highly correlated to the first-stage noise within sample. This in turn implies $E[Z_{ij}\epsilon_i|Z_j \text{ selected }] \neq 0$ provided U_i and ϵ_i are correlated. We demonstrate this bias and the resulting potential for poor performance of post-model-selection inference in the simulation study reported in Section 4. This problem is not reserved to strict variable selection devices but also applies to many other regularization

schemes, such as those in Carrasco (2012) and Carrasco and Tchuente Nguembu (2012). In simulations, we show that our approach performs well relative to these in a non-sparse setting.

To deal with the correlation between estimated instruments and first-stage errors induced by first-stage regularization, we propose using jackknife estimators $\widehat{\Upsilon}_i$ of Υ_i. Let $\phi(Z_i; X, Z)$ be a data-depentent rule for assigning predictions Υ_i based on Z_i. Then by the independence of (ϵ_i, U_i) across i, $\widehat{\Upsilon}_i = \phi(Z_i; X_{-i}, Z_{-i})$, where X_{-i} and Z_{-i} represent the data for all observations but observation i, is independent of ϵ_i. That is, the exclusion restriction

$$E[\widehat{\Upsilon}_i \epsilon_i] = E[\phi(Z_i; X_{-i}, Z_{-i})\epsilon_i] = 0$$

holds by construction if the conventional mean independence assumption, $E[\epsilon_i | Z_i] = 0$, is satisfied.[10] Then using

$$\tilde{\delta} = \left(\sum_{i=1}^{n} \widehat{\Upsilon}_i' X_i \right)^{-1} \sum_{i=1}^{n} \widehat{\Upsilon}_i' y_i$$

consistent estimates of δ_0 can be obtained provided that there is sufficient signal that the regularized leave-one-out-forecast gives informative predictions of Υ_i.

3. Ridge-Regularized JIVE

In this section, we present the details of the proposed regularized JIVE. We choose to work with regularization via the ridge regression though we expect that the results below would remain valid for the combination of the jackknife with other regularization schemes. We focus on ridge-regularization for several reasons. First, since our primary interest is in models for which sparsity in the first stage is not necessarily correct, it seems natural to impose regularization via pure shrinkage rather than a combination of selection and shrinkage as in LASSO or a pure-selection device. Second, the expression for the estimator of δ_0 when using ridge in the first stage are in closed form, simplifying the analysis of the theoretical properties. Third, the theory for the ridge estimator fits very nicely with the existing theory for JIVE with many instruments developed in CSHNW which allows us to present concise theoretical results by adapting arguments from CSHNW allowing for regularization.

3.1. The Ridge-Regularized JIVE.

Recall that the ridge regression coefficient estimate from the regression of the variable X onto the set of variables Z is given as

$$\widehat{\Pi} = \arg\min_{\Pi} \|X - Z\Pi\|_{2,I_n}^2 + \|\Pi\|_{2,\Lambda}^2$$

where we let $\|W\|_{2,A}^2 = W'A'AW$ for vector W and conformable, positive definite matrix A. That is, the ridge regression chooses parameters to minimize the within sample sum of squared

[10] A related approach is to maintain the validity of the exclusion restriction by sample splitting as described in BCCH, where instruments are selected with one half of the sample and the IV model is estimated using the second half. In initial simulations, we found that the RJIVE produced more efficient results.

residuals plus a penalty term for the weighted sum of squared regression coefficients. The penalty is designed so that the ridge regression favors models with small coefficients which helps to avoid overfitting. The ridge regression is extremely convenient as a shrinkage device since the ridge coefficient estimates are available in closed form:

$$\widehat{\Pi} = (Z'Z + \Lambda'\Lambda)^{-1}(Z'X). \tag{3.4}$$

From this regression, it is apparent that the impact of the penalty term is to stabilize the inverse of the sample covariance matrix of the regressors, $Z'Z$. Since $Z'Z$ is positive semi-definite, the addition of the positive definite matrix $\Lambda'\Lambda$ guarantees that the inverse in the definition of $\widehat{\Pi}$ is always well-defined. This behavior is in contrast to the usual OLS estimator $\widehat{\Pi}_{OLS} = (Z'Z)^{-1}(Z'X)$ which is ill-defined when $K > n$ since $Z'Z$ is singular by construction and may be poorly behaved far more generally since $Z'Z$ may be near-singular when the dimension of Z is large relative to n. Note that usual implementations of ridge regression set $\Lambda = \gamma^{1/2} I_K$ for a scalar penalty parameter γ. Our theory allows for the more general case but we follow this implemention in the simulation and empirical examples.[11]

In principle, one could use $\widehat{\Upsilon} = \widehat{\Pi}X$ for $\widehat{\Pi}$ in (3.4) in defining a regularized IV estimator which essentially corresponds to one of the regularization strategies pursued in Carrasco (2012). Carrasco (2012) provides conditions under which this approach is consistent and asymptotically normal for estimating δ_0. The drawback of this approach is that the estimated instrument $\widehat{\Upsilon}$ is correlated to the structural error ϵ by construction in finite samples. Under the conditions of Carrasco (2012) this correlation is asymptotically negligible and thus does not adversely affect the asymptotic properties of the regularized IV estimator. However, these conditions rule out the case of a dense signal with number of instruments greater than the sample size. In the following, we offer a complementary approach that applies in the case of a high-dimensional dense signal.

Define $\widehat{\Pi}^{\Lambda}_{-i} = (Z'Z - Z_i Z_i' + \Lambda'\Lambda)^{-1}(Z'X - Z_i X_i')$ which is the ridge regression coefficient from running a ridge regression of X on Z with regularization matrix Λ using all but the i^{th} observation. The leave-one-out estimator $\widehat{\Upsilon}_i$ for the value of the instrument for the i^{th} individual may then be defined as $\widehat{\Upsilon}_i = Z_i'\widehat{\Pi}^{\Lambda}_{-i}$. Using the constructed $\widehat{\Upsilon}_i$, we then define the ridge-regularized JIVE as

$$\tilde{\delta} = \left(\sum_{i=1}^{n} \widehat{\Pi}^{\Lambda}_{-i}{}' Z_i X_i'\right)^{-1} \sum_{i=1}^{n} \widehat{\Pi}^{\Lambda}_{-i}{}' Z_i y_i. \tag{3.5}$$

By using the sample excluding the i^{th} observation, the RJIVE breaks the correlation between $\widehat{\Upsilon}_i$ and ϵ_i in the case of a dense first-stage signal. The use of ridge in constructing this estimated

[11]This choice of penalty matrix would not generally be optimal in the case where consistent model selection is possible. However, feasible optimality of IV estimators with dense first-stage signal is an interesting question as estimation of the optimal instruments will generally not be possible in this scenario.

instrument also regularizes the problem allowing signal extraction from the large set of instruments while avoiding potential overfitting. The cost of this regularization is that some signal will be lost due to the imposed shrinkage. It seems that loss of first-stage signal will be generic in high-dimensional dense settings, but further exploration of this issue is warranted.

It is useful to relate the RJIVE to the formulation of the JIVE from CSHNW[12] since the estimators are quite similar and exploiting this similarity allows us to simplify the proofs and technical details of the RJIVE. Because estimates from ridge regression can be calculated by performing the augmented regression of $X^{aug} = \begin{pmatrix} X' & 0'_{G\times K} \end{pmatrix}'$ on $Z^{aug} = \begin{pmatrix} Z' & \Lambda \end{pmatrix}'$, results on recursive residuals as used in CSHNW continue to apply. That is,

$$\widehat{\Pi}^{\Lambda}_{-i}{}' Z_i = \left(X^{aug\prime} Z^{aug} (Z^{aug\prime} Z^{aug})^{-1} Z_i - P^{aug}_{ii} X_i\right) / (1 - P^{aug}_{ii}) = \sum_{j\neq i}^{n+K} P^{aug}_{ij} X^{aug}_j / (1 - P^{aug}_{ii})$$

where $P^{aug} := Z^{aug}(Z^{aug\prime} Z^{aug})^{-1} Z^{aug\prime}$ and P^{aug}_{ij} denotes the (i,j) element of P^{aug}. Also, the principal $n \times n$ submatrix of P^{aug} is $P^{\Lambda} = Z(Z'Z + \Lambda'\Lambda)^{-1} Z'$. Due to the fact that $X_j = 0$ for $j > n$, the expression for $\widetilde{\delta}$ can then be simplified to

$$\tilde{\delta} = \tilde{H}^{-1} \sum_{i\neq j} X_i P^{\Lambda}_{ij} (1 - P^{\Lambda}_{jj})^{-1} y_j$$

with

$$\tilde{H} = \sum_{i\neq j} X_i P^{\Lambda}_{ij} (1 - P^{\Lambda}_{jj})^{-1} X'_j. \tag{3.6}$$

Letting $\xi_i = (1 - P^{\Lambda}_{ii})^{-1} \epsilon_i$ and substituting $y_i = X'_i \delta_0 + \epsilon_i$, we finally have

$$\tilde{\delta} = \delta_0 + \tilde{H}^{-1} \sum_{i\neq j} X_i P^{\Lambda}_{ij} \xi_j.$$

In the following section, we provide conditions under which $\tilde{\delta}$ is consistent and asymptotically normal. Specifically, we provide conditions so that, after scaling, $\tilde{H}^{-1}$ converges in probability to a non-singular limit and a central limit theorem applies to $\sum_{i\neq j} X_i P^{\Lambda}_{ij} \xi_j$. Finally, we verify that the asymptotic variance of $\tilde{\delta}$ can be consistently estimated with

$$\tilde{V} = \tilde{H}^{-1} \tilde{\Sigma} \tilde{H}^{-1} \tag{3.7}$$

where

$$\tilde{\Sigma} = \sum_{i,j} \sum_{k\notin\{i,j\}} P^{\Lambda}_{ik} P^{\Lambda}_{jk} X_i X'_j \tilde{\xi}^2_k + \sum_{i\neq j} (P^{\Lambda}_{ij})^2 X_i \xi_i X'_j \xi_j$$

[12]CSHNW consider an alternative jackknife estimator which they call JIV2. We only consider the regularized analogue of JIV1 since its use is motivated by the dense signal problem and it exhibits better performance in simulation.

and

$$\tilde{\xi}_i = (1 - P^{\Lambda}_{ii})^{-1}(y_i - X_i'\tilde{\delta}).^{13}$$

A final consideration for implementing the RJIVE is selection of the penalty matrix Λ. We follow the usual approach for ridge regression by setting $\Lambda = \gamma^{1/2} I_n$ and further set $\gamma^{1/2} = CK^{1/2}$. In our simulations, we set the constant of proportionality C for each first-stage equation to the sample standard deviation of the element of X_i being considered. This practice is closely related to the penalty used in Dicker (2012).

3.2. **Asymptotic Properties of RJIVE.** In this section, we state the conditions under which we derive the properties of the RJIVE and state formal results. Proofs of theorems are provided in the appendix. Throughout, we work with an asymptotic approximation that considers an array of models where the number of instruments increases with the number of observations, n. Thus, all objects are implicitly indexed by n, but we suppress this indexing except where it would cause confusion for notational convenience. We let C be a generic constant whose value does not depend on n but may change with each use. We let a.s. denote almost surely and a.s.n. denote a.s. for n large enough.

Assumption 1. *$K \to \infty$. $\Lambda = \Lambda_n$ is a sequence of positive definite penalty matrices such that for $P^{\Lambda} := Z(Z'Z + \Lambda'\Lambda)^{-1}Z'$, $P^{\Lambda}_{ii} \leqslant C$ for some $C < 1$ and for all $i = 1, ..., n$, a.s.n.*

Assumption 2. *The optimal instrument Υ can be written $\Upsilon_i = S_n z_i/\sqrt{n}$ for some z_i and S_n where $S_n = \tilde{S}_n diag(\mu_{1n}, ..., \mu_{Gn})$, $\tilde{S}_n$ is $G \times G$ and bounded and the smallest eigenvalue of $\tilde{S}_n\tilde{S}_n'$ is bounded away from zero. Also, for each $1 \leqslant j \leqslant G$, either $\mu_{jn} = \sqrt{n}$ or $\mu_{jn}/\sqrt{n} \to 0$, $r_n = (\min_{1\leqslant j\leqslant G} \mu_{jn})^2 \to \infty$ and $\sqrt{K}/r_n \to 0$. Finally, there is $C > 0$ such that $\| \sum_{i=1}^n z_i z_i'/n\| \leqslant C$ where $\| \cdot \|$ denotes the Euclidean norm, and the smallest eigenvalue $\lambda_{min}(\sum_{i=1}^n z_i z_i'/n) \geqslant 1/C$ a.s.n.*

Assumption 3. *There is a constant, C, such that conditional on $\mathcal{Z} = (\Upsilon, Z, \Lambda)$, the observations $(\epsilon_1, U_1), ...(\epsilon_n, U_n)$ are independent with $E[\epsilon_i|\mathcal{Z}] = 0$ for all i, $E[U_i|\mathcal{Z}] = 0$ for all i, $\sup_i E[\epsilon_i^2|\mathcal{Z}] < C$, and $\sup_i E[\|U_i^2\||\mathcal{Z}] < C$, a.s.*

Assumption 1 places mild restrictions on the sequence of penalty matrices that are used to regularize the problem. The condition is akin to the usual full rank assumption on the matrix of

[13]For large datasets, the following expressions may be faster to compute:

$$\tilde{H} = X'P^{\Lambda}\tilde{X} - \sum_{i=1}^n X_i P^{\Lambda}_{ii}\tilde{X}_i'$$

$$\tilde{\Sigma} = \sum_{i=1}^n (\overline{X}_i\overline{X}_i' - X_i P^{\Lambda}_{ii}\overline{X}_i' - \overline{X}_i P^{\Lambda}_{ii}X_i')\tilde{\xi}_i^2 + \sum_{k=1}^K\sum_{l=1}^K \left(\sum_{i=1}^n \tilde{Z}_{ik}\tilde{Z}_{il}X_i\tilde{\xi}_i\right)\left(\sum_{j=1}^n Z_{jk}Z_{jl}X_i\tilde{\xi}_j\right)'$$

where $\overline{X} = P^{\Lambda}X$, $\tilde{X}_i = X_i/(1 - P^{\Lambda}_{ii})$ and $\tilde{Z} = Z(Z'Z + \Lambda'\Lambda)^{-1}$. It can be shown that the two expressions are numerically equivalent. For the non-penalized case, this was proven explicitly in CSHNW.

instruments Z but allows for more general behavior. Importantly, this condition allows for Z to be rank deficient and, as such, allows for $K > n$. The condition will be satisfied quite generally when the sequence of regularization matrices has minimum eigenvalue bounded away from zero and proportional to K and can generically be made true by imposing sufficient regularization.

Assumption 2 imposes restrictions on the strength of the first-stage signal available in the infeasible optimal instrument Υ. It is identical to Assumption 2 of CSHNW, and CSHNW provide detailed discussion of the classes of models for optimal instruments accommodated by this assumption. z_i defined in Assumption 2 is unobserved and of the same dimension as the infeasible optimal instrument for observation i, Υ_i. As such, z_i is best regarded simply as a rescaled version of this optimal instrument. The statement of the condition is general enough to allow for strong identification of first-stage relationships when $\mu_{jn} = \sqrt{n}$ for each $1 \leqslant j \leqslant G$, weak identification of first-stage relationships when $\mu_{jn}/\sqrt{n} \to 0$ for each $1 \leqslant j \leqslant G$, and situations in which some endogenous variables, those with $\mu_{jn} = \sqrt{n}$, have strong first-stage relationships and some endogenous variables, those with $\mu_{jn}/\sqrt{n} \to 0$, have weak first-stage relationships. The condition that $r_n \to \infty$ rules out the case of a small number of weak instruments as considered in, for example, Staiger and Stock (1997), Andrews and Stock (2007), and Stock, Wright, and Yogo (2002).

Assumption 3 places standard conditions on the error terms and formalizes the exclusion restrictions. The conditions on second moments impose bounded conditional heteroskedasticity, and we also assume independence of the errors across observations.

Next define

$$\overline{z}_i := Z_i(Z'Z + \Lambda'\Lambda)^{-1}Z'z = \sum_{j=1}^{n} P_{ij}^{\Lambda} z_j = [P^{\Lambda} z]_i$$

which can be interpreted as the predicted value of z_i after ridge-regularization. In addition, define the jackknife analog for prediction:

$$\widehat{z}_i := Z_i(Z'Z + \Lambda'\Lambda - Z_i Z_i')^{-1}(Z'z - Z_i z_i').$$

Recall that Z_i are the observed, feasible instruments that are used in estimation while z_i is an unobservable related to the optimal instrument.

Assumption 4. *There is a constant $C > 0$ such that $\lambda_{min}(\sum_{i=1}^{n} z_i \widehat{z}_i/n) \geqslant C$ a.s.n.*

This assumption is a relevance condition stating that the ridge-predicted value of the optimal instrument is related to the unobserved optimal instrument. This condition requires that the signal remaining after regularization is a non-vanishing fraction of the signal present before regularization.[14] Satisfaction of this condition will require both that there was signal in the

[14]This condition can be weakened to allow a regularized signal that is non-zero, but approaching zero at the cost of strengthening Assumption 2. We forego this generalization because it requires introduction of more notation and makes the proofs more cumbersome.

optimal instrument as in Assumption 2 and that the problem is not so high-dimensional that the amount of regularization required to guarantee stable behavior of the regularized estimator results in loss of all the signal available in the instruments. For example, Dicker (2012) shows that the ℓ^2-risk of the optimal ridge-estimator with Gaussian regressors in a dense model with $K/n \to \infty$ is asymptotically the same as the risk from the trivial estimator that sets all coefficients to 0 and that this trivial estimator is asymptotically minimax. Thus, the ridge-fit should be very similar to constant in this case and thus have no signal. Assumption 4 rules this and similar cases out.

In practice, the observable quantity $\tilde{H}$ defined in (3.6) gives a signal about the size of $\sum_{i=1}^n z_i \widehat{z}_i'/n$, and the two quantities are approximately equal in large enough samples. This gives the researcher a heuristic to assess the validity of Assumption 4. If it is observed that the eigenvalues of $\tilde{H}$ are small relative to the variability of X, then Assumption 4 may be questionable.

Assumptions 1-4 are sufficient for consistency of RJIVE.

Theorem 1. *Suppose that Assumptions 1-4 are satisfied. Then,* $r_n^{-1/2} S_n'(\tilde{\delta} - \delta_0) \xrightarrow{p} 0$ *and*

$$(\tilde{\delta} - \delta_0) \xrightarrow{p} 0.$$

To establish asymptotic normality and provide a consistent estimator of the asymptotic variance of $\widetilde{\delta}$, we impose two additional conditions.

Assumption 5. *There is a constant* $C > 0$ *such that* $\sum_{i=1}^n \|z_i\|^4/n^2 \to 0$, $\sum_{i=1}^n \|\overline{z}_i\|^4/n^2 \to 0$, $\sup_i E[\epsilon_i^4|\mathcal{Z}] < C$, *and* $\sup_i E[\|U_i\|^4|\mathcal{Z}] \leqslant C$ *a.s.*

Assumption 6. *There exists* $C > 0$ *such that a.s.n* $\sup_i \|\overline{z}_i\| \leqslant C$ *and* $\sup_i \|z_i\| \leqslant C$.

Assumption 5 is quite standard in the literature and assumes that various fourth moments are bounded. Assumption 6 imposes that the (appropriately rescaled) optimal instrument is bounded almost surely and that the regularized predictions of the rescaled optimal instruments are bounded almost surely. The latter condition seems quite reasonable since regularized predictions are generally biased towards the sample mean. Nevertheless, it is possible to exhibit a sequence of optimal instruments and a sequence of regularized projection matrices such that the regularized predictions grow without bound. Assumption 6 rules out such sequences.

The following notation will aid the discussion and proofs of the asymptotic normality results that follow. First, define $H_n := \sum_{i=1}^n z_i \widehat{z}_i'/n$. We will show that suitably normalized, the difference between $\tilde{H}$ and H_n vanishes in probability. In addition, as mentioned above, a central limit theorem will apply to the term $\sum_{i \neq j} X_i P_{ij}^{\Lambda} \xi_i$. The asymptotic variance of this term will decompose into the sum of

$$\Omega_n = \sum_i E[\xi_i^2|\mathcal{Z}](\overline{z}_i - P_{ii}^{\Lambda} z_i)(\overline{z}_i - P_{ii}^{\Lambda} z_i)'/n,$$

which corresponds to the usual limiting variance given a fixed number of instruments, and

$$\Psi_n = S_n^{-1} \sum_{i \neq j} {P_{ij}^{\Lambda}}^2 \left(E[U_i U_i'|\mathcal{Z}] E[\xi_i^2|\mathcal{Z}] + E[U_i \xi_i'|\mathcal{Z}] E[U_j' \xi_j^2|\mathcal{Z}] \right) {S_n^{-1}}'$$

which can be thought of as a correction for the presence of an increasing number of instruments. Finally, the asymptotic variance of $\tilde{\delta}$ will take the form

$$V_n = H_n^{-1}(\Omega_n + \Psi_n)H_n^{-1}.$$

Theorem 2. *Suppose that Assumptions 1-5 are satisfied, $\sigma_i^2 := E[\epsilon_i^2|\mathcal{Z}] \geqslant C > 0$ a.s. and K/r_n is bounded. Then V_n is nonsingular a.s.n and*

$$V_n^{-1/2} S_n'(\tilde{\delta} - \delta_0) \xrightarrow{d} N(0, I).$$

As in CSHNW, asymptotic variance matrices may be singular when $K/r_n \to \infty$. Such could arise when there are different strengths of identification for different elements of X. To accommodate the possibility of singularity, results are stated in terms of linear combinations of the RJIVE defined by a sequence of $\ell \times G$ matrices, L_n.

Theorem 3. *Suppose that Assumptions 1-5 are satisfied and $K/r_n \to \infty$. If L_n is a bounded sequence of $\ell \times G$ matrices such that $\lambda_{min}(L_n V_N^* L_n') \geqslant C$ a.s.n for some $C > 0$, then for $V_n^* := H_n^{-1}(r_n/K)\Psi_n H_n^{-1}$,*

$$(L_n V_n^* L_n')^{-1/2} L_n \sqrt{r_n/K} S_n'(\tilde{\delta} - \delta_0) \xrightarrow{d} N(0, I_\ell).$$

Theorems 2 and 3 provide asymptotic distributions under the case where the optimal instrument is strong and weak respectively. These results provide an interesting comparison to the asymptotic results in BCCH. BCCH consider the case where K may be much greater than n, allowing for $K = b_n \exp(n^{1/3})$ for a decreasing sequence b_n, and provide consistent and asymptotically normal estimators of δ_0 under the assumption that the first-stage signal is sparse. In our paper, we relax the condition that the first-stage is sparse but, to jointly satisfy Assumptions 1 and 4 in the dense case, implicitly impose stronger conditions on the rate of growth of K relative to n; see, e.g. Dicker (2012). Thus, we feel the two approaches are complements. The sparse model allows one to potentially consider far more instruments than in the case where one is unwilling to assume sparsity, and indeed sparsity seems more reasonable when one is allowed to consider a vast array of potential instruments. On the other hand, the results in this paper suggest that one can do without the sparse model assumption in the scenario when the number of available instruments is not very much greater than the number of observations which seems like a relevant scenario in practice.

Our final result verifies that the variance estimator given in 3.7 is consistent.

Theorem 4. *Under assumptions 1-6, if K/r_n is bounded, then*

$$S_n'\tilde{V}S_n - V_n \xrightarrow{p} 0.$$

If $K/r_n \to \infty$, then

$$r_n S_n'\tilde{V}Sn/K - V_n^* \xrightarrow{p} 0.$$

4. Simulation Study

The results in the previous sections suggest that RJIVE should have good estimation and inference propoerties provided the sample size n is large. We demonstrate the performance of our asymptotic approximation for RJIVE and provide a comparison with several other standard estimators using a simulation study based on the simple data generating process:

$$\begin{array}{c} y_i = x_i\delta_0 + \epsilon_i \\ x_i = Z_i'\Pi + U_i \end{array}, \quad (\epsilon_i, U_i) \sim N\left(0, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon U} \\ \sigma_{\epsilon U} & \sigma_U^2 \end{pmatrix}\right)$$

where the treatment variable x_i is scalar and the parameter of interest is $\delta_0 = 1$. We fix the sample size at $n = 100$ and the correlation between ϵ_i and U_i to $\rho = .6$. We consider various settings for the remaining parameters of the model.

First, we consider two different data-generating processes on the instruments. We are particularly motived by the performance of estimators in the presence of many categorical variables and therefore consider a binary instrument design. In the binary instrument design, all instruments are independently drawn with $\mathrm{P}(Z_{ij} = \frac{1}{2}) = \mathrm{P}(Z_{ij} = -\frac{1}{2}) = \frac{1}{2}$. The second design considers Gaussian instruments that are correlated with one another. Under the Gaussian instrument design, all instruments are drawn with mean 0 and variance $\mathrm{var}(Z_{ij}) = .3$. Dependence between instruments is given by $\mathrm{corr}(Z_{ij}, Z_{ik}) = .5^{|j-k|}$. In each design, we set the number of instruments to $K = 95$ and $K = 190$.

We also consider two different set of first-stage coefficients that are meant to generate dense and sparse first-stage relationships. In the dense case, the signal is determined with coefficient $\Pi = (\iota_{.4K}, 0_{.6K})'$ where ι_p is a $1 \times p$ vector of ones and 0_q is a $1 \times q$ vector of zeros. In the sparse case, we set $\Pi = (\iota_5, 0_{K-5})'$ so only the first five instruments are relevant. We alter the strength of the instruments by adjusting the noise σ_U^2 in the first stage regression. We measure instrument strength using the concentration parameter $\mu^2 = n\Pi' E[Z_i'Z_i]\Pi/\sigma_U^2$. We consider a weak and a strong first stage signal with $\mu^2 = 30$ and $\mu^2 = 150$ respectively. The remaining component of the covariance matrix of the errors is the variance of the structural error which we fix at $\sigma_\epsilon^2 = 2$.

In addition to RJIVE, we consider five alternative estimators for each setting. We report the Post-LASSO estimator described in BCCH, LASSO-JIVE which is an *ad hoc* modified version of the Post-LASSO described below, the shrinkage estimator of Carrasco (2012), the

standard JIVE without regularization, and two stage least squares. The Post-LASSO estimator is expected to perform well in the sparse design. In the dense design, LASSO is likely to select no instruments in many simulation replications when the penalty level is set according to usual recommendations motivated from sparse estimation which leaves the estimator undefined. To address this, we consider a second, Post-LASSO-like estimator, LASSO-JIVE with a liberal penalty level[15] that allows more instruments to be selected in the first stage model. To account for the fact that many instruments may be selected with the liberal penalty level, we then apply JIVE using the selected instruments. We calculate the Tychonov-regularized version of Carrasco's (2012) estimator.[16] JIVE is valid under many instruments and is an alternative to regularization estimators when $K < n$. Since JIVE is ill-defined with $K > n$, we proceed by selecting 95 instruments at random and performing JIVE with the selected instruments with $K = 190$. Finally, we consider 2SLS since it is the most common IV estimator found in the literature and provides a natural benchmark. Just as for JIVE, we randomly select a subset of 95 instruments for forming the 2SLS estimator when $K = 190$. We set the penalty matrix in RJIVE equal to $s_x\sqrt{K}I_K$ where s_x is the sample standard deviation of x which is recalculated at every simulation replication and I_K is the $K \times K$ identity matrix.

The results are based on 1500 simulations for each setting described above. For each estimator, we calculate the median bias, median absolute deviation, and rejection rate for a 5%-level test of $H_0 : \delta_0 = 1$. In many of the simulations, the Post-LASSO estimator is undefined as LASSO sets all coefficients to zero. In such a case, the null is not rejected. This is a conservative alternative to applying the Sup-score statistic described in BCCH. Median bias and median absolute deviation for Post-LASSO are calculated conditional on the LASSO producing at least one nonzero coefficient.[17] In all simulations, LASSO-JIVE selected at least one instruments with the median number of instruments selected ranging from 49 to 74.

The results for $K = 95$ are reported in Table 1. Panels A and B show results for a weak signal ($\mu^2 = 30$). For the weak signal, RJIVE and JIVE are the only estimators that produce reasonable rejection frequencies. They both show small bias relative to median absolute deviation. By contrast, all other estimators seem to be dominated by bias regardless of whether the signal

[15]The penalty level recommended for Post-LASSO in BCCH, for example, is proportional to $\sqrt{n \log p}$. We set the penalty at $2.2\sqrt{2n\log(2K)}\sigma_U\sigma_z$, where the standard deviations are set to their true values. We relax the penalty by a factor of $\sqrt{n}$ to $2.2\sqrt{2\log(2K)}\sigma_U\sigma_z$ for selecting instruments for LASSO-JIVE.

[16]Carrasco (2012) requires the input of a tuning (penalty) parameter. Carrasco (2012) provides an expression for the approximate mean squared error of the estimator and suggests choosing the tuning parameter to minimize this criterion. We calculate the optimal tuning parameter for one simulation run assuming the true values for σ_ϵ^2, $\sigma_{\epsilon U}$, and $Z_i'\Pi$ are known. For the remaining simulations, the same tuning parameter is used.

[17]In the Binary simulation with $K = 95$, reading left to right accross the table, LASSO selected 0 instruments in 1257, 910, 54, and 0 runs. With Gaussian instruments and $K = 95$, LASSO selected 0 instruments in 527, 147, 0, and 0 runs. For $K = 190$ in the binary simulation, reading left to right accross the table, LASSO selected 0 instruments in 1343, 1031, 93 and 0 runs. With Gaussian instruments and $K = 190$, LASSO selected 0 instruments in 788, 239, 0, and 0 runs.

is dense or sparse. This demonstrates the robustness of the JIVE and RJIVE in the presence of a weak signal. RJIVE has considerably smaller absolute deviation than its unregularized counterpart, JIVE. Panels C and D show results for a stronger signal ($\mu^2 = 150$). As expected with a strong sparse signal, the Post-LASSO has approximately correct size in the sparse case and smaller median absolute deviation than RJIVE.

The results for $K = 190$ are reported in Table 2. Panels A and B again show results for a weak signal ($\mu^2 = 30$). For the weak signal, RJIVE and JIVE are again the only estimators that produce correct rejection frequencies. They both show small bias relative to median absolute deviation. By contrast, all other estimators seem to be dominated by bias regardless of whether the signal is dense or sparse. Panels C and D show results for a stronger signal ($\mu^2 = 150$). Once again, the Post-LASSO has approximately correct size in the sparse case and has smaller median absolute deviation than RJIVE. This case most clearly shows the relative strength of the Post-LASSO to the RJIVE. Post-LASSO can effectively locate a concentrated strong signal among a set of very many instruments and thus outperforms RJIVE. The RJIVE, on the other hand, dominates all considered procedures across the remainder of the designs and continues to control size and has reasonable risk properties in the strong-sparse-signal case but loses efficiency relative to Post-LASSO. This loss of efficiency in the strong-sparse case appears to be the cost of the additional robustness that RJIVE enjoys relative to Post-LASSO and would likely be apparent for other sparsity-based procedures. Overall, the simulations suggest that RJIVE may usefully complement existing approaches to estimation and inference with many instruments, especially in settings where sparsity is suspect.

5. Empirical Example: Angrist and Krueger (1991)

In this section, we illustrate the use of the RJIVE by revisiting the classic example in the many-instrument literature, Angrist and Krueger (1991). Interest in this example focuses on attempting to estimate the causal effect of schooling on earnings by addressing the potential endogenity of schooling through the use of instrumental variables. The identification strategy and data from Angrist and Krueger (1991) provides many instruments which can be used for schooling, and a substantial body of literature has arisen discussing concerns about the potential biases and inferential problems introduced from using the full set of available instruments. See, for example, Bound, Jaeger, and Baker (1995), Angrist, Imbens, and Krueger (1999), Staiger and Stock (1997) and Hansen, Hausman, and Newey (2008).

As in Angrist and Krueger (1991), we consider the model

$$\begin{aligned}\log(\text{wage}_i) &= \alpha \text{Schooling}_i + w_i'\gamma + \varepsilon_i \\ \text{Schooling}_i &= z_i'\Pi_1 + w_i'\Pi_2 + v_i\end{aligned}$$

where ε_i and v_i are unobservables that satisfy the exclusion restriction $\mathrm{E}[\varepsilon_i|w_i, z_i] = \mathrm{E}[v_i|w_i, z_i] = 0$, $\log(\text{wage}_i)$ is the log(wage) of individual i, Schooling_i is the reported years of completed schooling of individual i, w_i is a vector of control variables, and z_i is a vector of instrumental variables that affect education but do not directly affect the wage. The data were drawn from the 1980 U.S. Census and consist of 329,509 men born between 1930 and 1939. For w_i, we use a set of 510 variables consisting of a constant, 9 year-of-birth dummies, 50 state-of-birth dummies, and 450 state-of-birth $\times$ year-of-birth interactions. As instruments, we use three quarter-of-birth dummies and interactions of these quarter-of-birth dummies with the full set of state-of-birth and year-of-birth controls in w_i giving a total of 1527 potential instruments. Angrist and Krueger (1991) discusses the endogeneity of schooling in the wage equation and provides an argument for the validity of z_i as instruments based on compulsory schooling laws and the shape of the life-cycle earnings profile. We refer the interested reader to Angrist and Krueger (1991) for further details. The coefficient of interest is α, which summarizes the causal impact of education on earnings.

We report results for estimating α from several strategies and for three different instrument sets in Table 3. Each panel in Table 3 gives the results for a different set of instruments. For each set of instruments, we report results from the conventional 2SLS estimator in columns labeled "2SLS", the JIVE estimator in columns labeled "JIVE",[18] the Post-LASSO estimator of BCCH in columns labeled "Post-LASSO",[19] and RJIVE in columns labeled "RJIVE". For RJIVE, we set the ridge penalty matrix as $\sqrt{K}s_{x|w}I_K$ where K is the number of variables in z_i, I_K denotes the $K \times K$ identity matrix, and $s_{x|w}$ is the sample standard deviation of the residuals from the OLS regression of schooling on the controls. We also report heteroskedasticity consistent standard error estimates for each estimator.

Given knowledge of the instruments and the identification argument from Angrist and Krueger (1991), a natural set of instruments is simply the three main quarter-of-birth dummies. Panel A of Table 3 gives results from using this set of instruments. In this case, the four estimators considered give very similar results. These three instruments are relatively powerful and do not seem to result in substantial first-stage overfitting.[20] As such, the 2SLS estimator is fairly well-behaved and lines up well with the other, more robust procedures. Due to the small number and relative strength of the instruments, the ridge regularization using the suggested penalty imposes very little regularization, and the JIVE and RJIVE estimates are nearly identical. Interestingly, LASSO only selects two of the three possible instruments and produces estimates very similar to JIVE.

[18]Specifically, we use the JIV1 estimator of Phillips and Hale (1977). See also Angrist, Imbens, and Krueger (1999) and CSHNW.

[19]We set the penalty parameter in the LASSO according to the refined option in BCCH equation (A.21) using residuals from the regression of schooling on the three main quarter-of-birth effects as $\widehat{v}$.

[20]Hansen, Hausman, and Newey (2008) gives further discussion of this point.

In Panel B of Table 3, we report results using 180 instruments formed by using the three quarter-of-birth main effects and their interactions with the 9 main effects for year-of-birth and 50 main effects for state-of-birth. Angrist and Krueger (1991) reported results from this set of instruments, motivating the use of the additional interactions from the standpoint of increasing efficiency. It is now generally believed that 2SLS estimates using this set of 180 instruments have a substantial bias toward OLS relative to the variability of the estimator due to overfitting of the first-stage which results in potentially misleading inference about the size of the schooling coefficient. As such, procedures that are robust to the presence of many instruments, such as JIVE or selecting instruments via LASSO, have been advocated when this instrument set is used. In our results, we do see that the 2SLS estimator shifts substantively toward the OLS estimate of .0673 when this larger set of instruments is used. On the other hand, all three of the many-instrument-robust point estimates remain near the value estimated using only the three main effects as instruments. We also see that the estimated standard error for each of the many-instrument-robust procedures is smaller than when only three instruments are used, suggesting there is additional signal available in the larger instrument set.[21]

The results reported in Panel C of Table 3 are based on using the full set of 1527 instruments and are the most interesting from the standpoint of the present paper. In this case, we see that both the Post-LASSO and JIVE point estimates have shifted substantively toward the OLS estimate. In contrast, the RJIVE is very stable, remaining around the value estimated by all of the procedures using only three instruments. More interesting is the fact that standard errors from both JIVE and Post-LASSO are now pronouncedly larger than the standard error from the RJIVE. The increase in standard errors for Post-LASSO is due to the fact that LASSO now only selects one variable. This reduction in the number of variables selected is due to the fact that LASSO requires a higher level of signal from each variable before it allows it to enter as a significant predictor with a larger number of variables. Thus, the LASSO estimator remains reasonably stable in this example due to the fact that one of the quarter of birth instruments has a substantial amount of predictive power but may be discarding useful small signals and be inefficient. On the other hand, the leave-one-out fits used as instruments in the JIVE are highly variable due to the large number of essentially uninformative variables used in the first-stage which results in a very high many-instrument-robust standard error. The RJIVE effectively regularizes these uninformative signals out of the problem while apparently capturing more of the signal than LASSO producing an estimator that remains stably at the value obtained when only the strong signals are used while having a smaller estimated variance.

These results demonstrate that RJIVE produces sensible and what appear to be relatively high-quality estimates in this application. As with LASSO, the RJIVE behaves stably without requiring *a priori* information about which are the relevant instruments and produces estimates that are very similar to those obtained from other leading approaches to estimation and inference

[21]In this case, LASSO selects five instruments.

when this information is used. Relative to LASSO, the RJIVE appears to use more signal in this example which may be due to the fact that the signal available in the many state-of-birth $\times$ year-of-birth $\times$ quarter-of-birth interactions is not sparse; i.e. there may be many of these terms that have small effects but provide valuable signal in aggregate. These results suggest that RJIVE may provide a useful complement to currently advocated approaches to dealing with many instruments.

6. Conclusion

To improve efficiency of classical IV techniques, researchers may want to make use of many instruments in order to have stronger signal about exogenous variation in the treatment variable of interest. However, many traditional IV techniques perform poorly when many instruments are used. RJIVE gives a feasible method for using the information present in a large number of instruments. The most important feature of RJIVE is it remains consistent and approximately normal, allowing simple valid inference for treatment effects even without the presence of a strong and sparse first stage signal. The ability to perform well without requiring a sparse first stage is in contrast to high-dimensional IV estimators that rely on variable selection. The dense signal case, where all instruments potentially contribute to variation in the treatment variable, seems like an important setting in practice. The RJIVE is also conceptually straightforward and computationally simple. We show it performs well relative to other many instrument robust procedures through simulations and that it also seems to perform well in the classic Angrist and Krueger example. The results in this paper suggest that the RJIVE may provide researchers with a useful tool when faced with many instruments.

7. Appendix

In the following, we provide proofs of Theorem 1-4. The proofs follow from arguments similar to those of CSHNW modified to account for regularization in performing the jackknife.

7.1. **Lemmas.**

Lemma 1. *Lemma A1 of CSHNW holds with the idempotent matrix P replaced by P^Λ. That is, suppose $P^\Lambda \equiv P^\Lambda(\mathcal{Z}) = Z(Z'Z + \Lambda'\Lambda)Z'$. If conditional on $\mathcal{Z}$ (defined hereafter by $\mathcal{Z} := (\Upsilon, Z, \Lambda)$), the scalar random variables (W_i, Y_i) are independent a.s., then there is a constant C such that*

$$\|\sum_{i\neq j}^{n} P_{ij}^{\Lambda} W_i Y_j - \sum_{i\neq j}^{n} P_{ij}^{\Lambda} \overline{w}_i \overline{y}_j\|_{L_2,\mathcal{Z}}^2 \leqslant C B_n, \ \ a.s.n$$

where $\overline{w}_i = E[W_i|\mathcal{Z}]$, $\overline{y}_i = E[Y_i|\mathcal{Z}]$, $B_n = K\overline{\sigma}_{W_n}^2 \overline{\sigma}_{Y_n}^2 + \overline{\sigma}_{Y_n}^2 \overline{w}_n' \overline{w}_n + \overline{\sigma}_{W_n}^2 \overline{y}_n' \overline{y}_n$ and in the definition of B_n, $\overline{w}_n = E[(W_1, ..., W_n)'|\mathcal{Z}]$, $\overline{y}_n = E[(Y_1, ..., Y_n)'|\mathcal{Z}]$, $\overline{\sigma}_{W_n} = \max_{i\leqslant n} var(W_i|\mathcal{Z})^{1/2}$, $\overline{\sigma}_{Y_n} = \max_{i\leqslant n} var(Y_i|\mathcal{Z})^{1/2}$. Finallly, the norm in the above bound is defined by $\|.\|_{L_2,\mathcal{Z}}^2 = E[(.)^2|\mathcal{Z}]$.

Proof. As in the text, define the $(n+K)\times K$ augmented data matrix $Z^{aug} := \begin{pmatrix} Z' & \Lambda' \end{pmatrix}'$ and the $(n+K)\times(n+K)$ augmented projection matrix $P^{aug} := Z^{aug}(Z^{aug\prime}Z^{aug})^{-1}Z^{aug\prime}$. In addition, define new (degenerate) random variables $W_{n+1},...,W_{n+K}=0$; $Y_{n+1},...,Y_{n+K}=0$. Then note that because P^{Λ} is identical to the principal $n\times n$ submatrix of P^{aug}, the equality of the following sums hold: $\sum_{i\neq j}^{n} P_{ij}^{\Lambda}W_iY_j = \sum_{i\neq j}^{n+K} P^{aug}W_iY_j$ and $\sum_{i\neq j}^{n} P_{ij}^{\Lambda}\overline{w}_i\overline{y}_j = E\left[\sum_{i\neq j}^{n} P_{ij}^{\Lambda}W_iY_j|\mathcal{Z}\right] = E\left[\sum_{i\neq j}^{n+K} P^{aug}W_iY_j|\mathcal{Z}\right] = \sum_{i\neq j}^{n+K} P^{aug}\overline{w}_i\overline{y}_j$.

This implies that

$$\|\sum_{i\neq j}^{n} P_{ij}^{\Lambda}W_iY_j - \sum_{i\neq j}^{n} P_{ij}^{\Lambda}\overline{w}_i\overline{y}_j\|_{L_2,\mathcal{Z}}^2 = \|\sum_{i\neq j}^{n+K} P_{ij}^{aug}W_iY_j - \sum_{i\neq j}^{n+K} P_{ij}^{aug}\overline{w}_i\overline{y}_j\|_{L_2,\mathcal{Z}}^2 \leqslant CB_{n+K} = CB_n$$

The inequality in the above line holds by Lemma A1 of CHSNW et al since P^{aug} is symmetric and idempotent. $CB_n = CB_{n+K}$ holds since $W_{n+1},...,W_{n+K}$ and $Y_{n+1},..,Y_{n+K}$ are degenrate and therefore $\overline{\sigma}^2_{W_n} = \overline{\sigma}^2_{W_{n+K}}$, $\overline{\sigma}^2_{Y_n} = \overline{\sigma}^2_{Y_{n+K}}$, $\overline{y}_n'\overline{y}_n = \overline{y}_{n+K}'\overline{y}_{n+K}$, $\overline{w}_n'\overline{w}_n = \overline{w}_{n+K}'\overline{w}_{n+K}$ and $\text{rank}(P^{\Lambda}) = \text{rank}(P^{aug}) = K$.

□

Lemma 2. *Lemma A2 of CSHNW holds with P^{Λ} replacing P. That is, suppose that the following hold conditional on $\mathcal{Z}$:*

(i) $P^{\Lambda} = P^{\Lambda}(\mathcal{Z}) = Z(Z'Z+\Lambda'\Lambda)^{-1}Z'$;

(ii) $(W_{1n},U_1,\xi_1),...,(W_{nn},U_n,\xi_n)$ are idependent, and $D_{1,n} := \sum_{i=1}^{n} E[W_{in}W_{in}'|\mathcal{Z}]$ satisfies $\|D_{1,n}\| \leqslant C$ a.s.n;

(iii) $E[W_{in}'|\mathcal{Z}]=0, E[U_i|\mathcal{Z}]=0, E[\xi_i|\mathcal{Z}]=0$, and there is a constant C such that $E[\|U_i\|^4|\mathcal{Z}] \leqslant C$ and $E[\|\xi_i\|^4|\mathcal{Z}] \leqslant C$;

(iv) $\sum_{i=1}^{n} E[\|W_{in}\|^4|\mathcal{Z}] \xrightarrow{a.s.} 0$; and

(v) $K\to\infty$ as $n\to\infty$.

Then for

$$D_{2,n} := \sum_{i\neq j}(P_{ij}^{\Lambda})^2\left(E[U_iU_i'|\mathcal{Z}]E[\xi_j^2|\mathcal{Z}] + E[U_i\xi_i|\mathcal{Z}]E[\xi_jU_j'|\mathcal{Z}]\right)/K$$

and any sequences c_{1n} and c_{2n} depending on $\mathcal{Z}$ of conformable vectors with $\|c_{1,n}\| \leqslant C, \|c_{2,n}\| \leqslant C$, and $\Xi_n = c_{1,n}'D_{1,n}c_{1,n} + c_{2,n}'D_{2,n}c_{2,n} > 1/C$ a.s.n. it follows that

$$\bar{Y}_n = \Xi_n^{-1/2}\left(c_{1,n}'\sum_{i=1}^{n}W_{in} + c_{2,n}'\sum_{i\neq j}U_iP_{ij}^{\Lambda}\xi_j/\sqrt{K}\right) \xrightarrow{d} N(0,1),\ \ a.s.$$

Proof. In order to prove the lemma, we would like to apply Lemma A2 of CSHNW using the same augmentation argument that was used to prove Lemma 1. However, the augmentation cannot

be applied immediately, because the augmented projection matrix defined prevously, P^{aug}, need not satisfy $P^{aug}_{ii} \leqslant C < 1$, which is a hypothesis of Lemma A2 in CSHNW. Therefore, we first show that Lemma A2 of CSHNW does indeed hold with any sequence of projection matrices. Once this is accomplished, the augmentation argument will be valid.

We turn to showing that CSHNW Lemma A2 holds without requiring $P_{ii} \leqslant C < 1$ provided P is a sequence of projection matrices. Suppose that $(W_{1n}, U_1, \xi_1), ..., (W_{nn}, U_n, \xi_n)$ satisfy conditions (ii) - (v) of Lemma A2 in CSHNW and that the sequences of matrices $P = P_n$ are projection matrices but don't necessarily satisfy $P_{ii} \leqslant C < 1$. In addition, suppose the vectors c_{1n} and c_{2n} satisfy the hypotheses of the lemma: $\|c_{1,n}\| \leqslant C, \|c_{2,n}\| \leqslant C$, and $\Xi_n := c'_{1,n} D_{1,n} c_{1,n} + c'_{2,n} D_{2,n} c_{2,n} > 1/C$, for $D_{1,n}$ and $D_{2,n}$ defined in the statement of lemma.

We roceed by defining new sequences which are shown to converge in distribution appropriately. The new sequences will then be related to the sequence of interest, $\bar{Y}_n$. This will imply that $\bar{Y}_n$ converges appropriately. The new sequences are given by

$$(\widehat{W}_{i2n}, \widehat{U}_i, \widehat{\xi}_i) := \begin{cases} (W_{in}, U_i, \xi_i) & \text{if } i \leqslant n \\ (0,0,0) & \text{otherwise} \end{cases}$$

Then define new $2n \times 2n$ matrices $\widehat{P}_{2n} := \begin{pmatrix} \frac{1}{2}P_n & \frac{1}{2}P_n \\ \frac{1}{2}P_n & \frac{1}{2}P_n \end{pmatrix}$. The diagonal elements of the newly defined matrices satisfy $[\widehat{P}_{2n}]_{ii} \leqslant \frac{1}{2}$ since $[P_n]_{ii} \leqslant 1$. Note also that $\widehat{P}_{2n}$ are symmetric and idempotent and have rank $K = K_n$. Define $\widehat{K}_{2n} := rank(\widehat{P}_{2n})$.

Finally, let

$$\widehat{c}_{1,2n} := (c'_{1n}\ 0'_{n\times 1})', \quad \widehat{c}_{2,2n} := 2(c'_{2n}\ 0_{n\times 1})',$$

For $\widehat{D}_{2,2n} := \sum_{i\neq j}^{2n} (\widehat{P}_{2n})_{ij}^2 \left(E[\widehat{U}_i \widehat{U}'_i | \mathcal{Z}] E[\widehat{\xi}_j^2 | \mathcal{Z}] + E[\widehat{U}_i \widehat{\xi}_i | \mathcal{Z}] E[\widehat{\xi}_j \widehat{U}'_j | \mathcal{Z}] \right) / \widehat{K}_{2n}$, the equality $\widehat{D}_{2,2n} = \frac{1}{4} D_{2,n}$ holds. In addition, $\widehat{D}_{1,2n} = D_{1,n}$. These equalities, together with the definition $\widehat{\Xi}_{2n} := \widehat{c}'_{1,2n} \widehat{D}_{2n} \widehat{c}_{1,2n} + \widehat{c}'_{2,2n} \widehat{\Sigma}_n \widehat{c}_{2,2n}$, imply that $\widehat{\Xi}_{2n} = \Xi_n$. Therefore, $\widehat{\Xi}_{2n} \geqslant 1/C$ since $\Xi_n \geqslant 1/C$. Then Lemma A2 of CSHNW can be applied to $\widehat{Y}_{2n}$ (if desired, a similar construction can be done that achieves $\widehat{Y}_{2n+1} = \bar{Y}_n$):

$$\widehat{Y}_{2n} := \widehat{\Xi}_{2n}^{-1/2} \left(\widehat{c}'_{1,2n} \sum_{i=1}^{2n} \widehat{W}_{i2n} + \widehat{c}'_{2,2n} \sum_{i\neq j}^{2n} U_i [\widehat{P}_{2n}]_{ij} \xi_j / \sqrt{\widehat{K}_{2n}} \right) \xrightarrow{d} N(0,1), \text{ a.s.;}$$

To relate $\widehat{Y}_{2n}$ back to the original sequence $\bar{Y}_n$, note that $\widehat{c}'_{1,2n} \sum_{i=1}^{2n} \widehat{W}_{i2n} = c_{1,n} \sum_{i=1}^{n} W_{in}$ and that $\widehat{c}'_{2,2n} \sum_{i\neq j}^{2n} \widehat{U}_i [\widehat{P}_{2n}]_{ij} \widehat{\xi}_j / \sqrt{\widehat{K}_{2n}} = 2c'_{2,n} \sum_{i\neq j}^{n} U_i [\frac{1}{2} P_n]_{ij} \xi_j / \sqrt{K} = c'_{2,n} \sum_{i\neq j}^{n} U_i [P_n]_{ij} \xi_j / \sqrt{K}$. Therefore, $\widehat{Y}_{2n} = Y_n$. Then by the convergence of $\widehat{Y}_n$, it follows that

$$\bar{Y}_n \xrightarrow{d} N(0,1) \text{ a.s.}$$

This shows that Lemma A2 of CSHNW still holds without requiring $P_{ii} \leqslant C < 1$ provided P are projection matrices. Therefore, the same augmentation argument as was given for the proof of Lemma 1 can now be used to show Lemma 2. □

Lemma 3. *Lemma A3 of CSHNW holds with P^Λ replacing P. That is, if conditional on $\mathcal{Z}$, (W_i, Y_i),$i = 1, ..., n$ are independent scalars, then there is $C > 0$ such that almost surely,*

$$\| \sum_{i \neq j} {P_{ij}^\Lambda}^2 W_i Y_j - E[\sum_{i \neq j} {P_{ij}^\Lambda}^2 W_i Y_j] \|_{L_2, \mathcal{Z}}^2 \leqslant C B_n'$$

where $B_n' := K \left(\overline{\sigma}_W^2 \overline{\sigma}_Y^2 + \overline{\sigma}_W^2 \overline{\mu}_Y^2 + \overline{\mu}_W^2 \overline{\sigma}_Y^2 \right)$, $\overline{\sigma}_W^2$ and $\overline{\sigma}_Y^2$ use the same notation as Lemma 1 and $\overline{\mu}_W^2 := \max_{i=1,...,n} E[W_i|\mathcal{Z}]$, $\overline{\mu}_Y^2 := \max_{i=1,...,n} E[Y_i|\mathcal{Z}]$.

Proof. The same augmentation argument as used for Lemma 1 holds. □

Lemma 4. *(Modified from CSHNW for RJIVE) Suppose there is a constant $C > 0$ such that, conditional on $\mathcal{Z}$, $(W_1, Y_1, \eta_1), ..., (W_n, Y_n, \eta_n)$ are independent with $E[W_i|\mathcal{Z}] = a_i/\sqrt{n}$, $E[Y_i|\mathcal{Z}] = b_i/\sqrt{n}$, with $|a_i| < C, |b_i| < C$, $E[\eta_i^2|\mathcal{Z}] < C$, $var(W_i|\mathcal{Z}) < C/r_n$, $var(Y_i|\mathcal{Z}) < C/r_n$, and $\max_{i=1,...,n} |[P^\Lambda W]_i| \leqslant C/\sqrt{n}$ where $W = (W_1, ..., W_n)'$, $[P^\Lambda W]_i$ denotes the i^{th} component of $P^\Lambda W$, and r_n is a sequence such that $r_n \to \infty$. Then*

$$A_n := E \left[\sum_{i \neq j \neq k} W_i P_{ik}^\Lambda \eta_k P_{kj}^\Lambda Y_j | \mathcal{Z} \right] = O_P(1), \text{ and}$$

$$\sum_{i \neq j \neq k} W_i P_{ik}^\Lambda \eta_k P_{kj}^\Lambda Y_j - A_n \xrightarrow{p} 0$$

Proof. As was the case in the proof of Lemma 2, we cannot directly apply the augmentation argument. The general augmentation idea will still work, but will require a slight modification. Let $W_{n+1}, ..., W_{n+K} = Y_{n+1}, ..., Y_{n+K} = \eta_{n+1} ... \eta_{n+K} = 0$. Define Z^{aug} and P^{aug} as above. The reason that Lemma A4 as stated in CSHNW does not immediately apply is that we do not assume (nor does it make sense for the augmented variables) that there is π_n such that $\max_{n+1 \leqslant i \leqslant n+K} |a_i - {Z^{aug}}_i' \pi_n| \xrightarrow{a.s.} 0$. Instead, we use the condition $\max_{i=1,...,n} |[P^\Lambda W]_i| \leqslant C/\sqrt{n}$. This is the only alteration needed. Let

$$A_n^{aug} = E \left[\sum_{i \neq j \neq k}^{n+K} W_i P_{ik}^{aug} \eta_k P_{kj}^{aug} Y_j | \mathcal{Z} \right]$$

and note that because of the definitions of W_i, Y_i, η_i for $i > n$, then $A_n = A_n^{aug}$.

Following closely the notation given in CSHNW, let

$$\widehat{\psi}_1 = \sum_{i \neq j \neq k}^{n+K} \tilde{w}_i P_{ik}^{aug} \tilde{\eta}_k P_{kj}^{aug} \tilde{y}_j, \ \widehat{\psi}_2 = \sum_{i \neq j \neq k}^{n+K} \tilde{w}_i P_{ik}^{aug} \tilde{\eta}_k P_{kj}^{aug} \overline{y}_j$$

$$\widehat{\psi}_3 = \sum_{i\neq j\neq k}^{n+K} \tilde{w}_i P_{ik}^{aug} \overline{\eta}_k P_{kj}^{aug} \tilde{y}_j,\ \widehat{\psi}_4 = \sum_{i\neq j\neq k}^{n+K} \tilde{w}_i P_{ik}^{aug} \overline{\eta}_k P_{kj}^{aug} \overline{y}_j$$

$$\widehat{\psi}_5 = \sum_{i\neq j\neq k}^{n+K} \overline{w}_i P_{ik}^{aug} \tilde{\eta}_k P_{kj}^{aug} \tilde{y}_j,\ \widehat{\psi}_6 = \sum_{i\neq j\neq k}^{n+K} \overline{w}_i P_{ik}^{aug} \tilde{\eta}_k P_{kj}^{aug} \overline{y}_j$$

$$\widehat{\psi}_7 = \sum_{i\neq j\neq k}^{n+K} \overline{w}_i P_{ik}^{aug} \overline{\eta}_k P_{kj}^{aug} \tilde{y}_j$$

where conditional means are denoted $\overline{\eta}_i = E[\eta_i|\mathcal{Z}]$, $\overline{w}_i = E[W_i|\mathcal{Z}]$, and $\overline{y}_i = E[Y_i|\mathcal{Z}]$ and deviation from means are denoted $\tilde{\eta}_i = \eta_i - \overline{\eta}_i$, $\tilde{W}_i = W_i - \overline{w}_i$ and $\tilde{Y}_i = Y_i - \overline{y}_i$.

Then, as noted before, $A_n = A_n^{aug}$, and algebraic manipulation gives

$$A_n = A_n^{aug} = \sum_{i\neq j\neq k}^{n+K} W_i P_{ik}^{aug} \eta_k P_{kj}^{aug} Y_j - \sum_{r=1}^{7} \widehat{\psi}_r.$$

The proof Lemma 4 is then completed after showing that $A_n^{aug} = O_P(1)$ and that $\widehat{\psi}_r \xrightarrow{p} 0$ for $r = 1, ..., 7$. A careful verification of the argument in CSHNW reveals all that the arguments showing that $\widehat{\psi}_r \xrightarrow{p} 0$ for $r \in \{1, 2, 3, 4, 5, 7\}$ and that $A_n = O_P(1)$ remain valid in our setting. Therefore, all that is left is showing that $\widehat{\psi}_6 \xrightarrow{p} 0$. First observe that $E[\widehat{\psi}_6|\mathcal{Z}] = 0$. Therefore, we show that $E[\widehat{\psi}_6^2|\mathcal{Z}] \to 0$ and so by Markov inequality, $\widehat{\psi}_6 \xrightarrow{p} 0$.

As before, let

$\overline{\mu}_W^2 = \max_{i\leqslant n} \overline{w}_i^2$, $\overline{\mu}_Y^2 = \max_{i\leqslant n} \overline{y}_i^2$, $\overline{\mu}_\eta = \max_{i\leqslant n} \overline{\eta}_i^2$;

$\overline{\sigma}_W^2 = \max_{i\leqslant n} var(W_i|\mathcal{Z})$, $\overline{\sigma}_Y^2 = \max_{i\leqslant n} var(Y_i|\mathcal{Z})$, $\overline{\sigma}_\eta^2 = \max_{i\leqslant n} var(\eta_i|\mathcal{Z})$

Note that $\overline{\mu}_W \leqslant C/n$, $\overline{\mu}2_Y \leqslant C/n$, $\overline{\mu}_\eta \leqslant C$, and $\overline{\sigma}_W^2 \leqslant C/r_n$, $\overline{\sigma}_Y^2 \leqslant C/r_n$, $\overline{\sigma}_\eta^2 \leqslant C$. Also, let $\breve{w}_i = \sum_{j=1}^{n+K} P_{ij}^{aug} \overline{w}_j$ and $\breve{y}_i = \sum_{j=1}^{n+K} P_{ij}^{aug} \overline{y}_j$.

Then for $i \neq k$, $\sum_{j\notin\{i,k\}}^{n+K} \overline{w}_i P_{ik}^{aug} P_{kj}^{aug} \overline{y}_j = \overline{w}_i P_{ik}^{aug} - \overline{P}_{ik}^{aug} P_{kj}^{aug} \overline{y}_j$. Then for fixed k,

$$\sum_{i\neq k}^{n+K} \sum_{j\notin\{i,k\}}^{n+K} = \sum_{i=1}^{n+K} \left(\overline{w}_i P_{ik}^{aug} \breve{y}_j - \overline{w}_i (P_{ik}^{aug})^2 \overline{y}_i - \overline{w}_i P_{ik}^{aug} P_{kk}^{aug} \overline{y}_k \right)$$

Then, using the the fact that $(A_1 + ... + A_5)^2 \leqslant 5(A_1^2 + ... + A_5^2)$ for any numbers $A_1, ...A_5$, the following sequence of inequalities hold:

$$E[\widehat{\psi}_6^2|\mathcal{Z}] = \sum_{k=1}^{n+K} E[\tilde{\eta}_k^2|\mathcal{Z}] \left(\sum_{i\neq k}^{n+K} \sum_{j\notin\{i,k\}}^{n+K} \overline{w}_i P_{ik}^{aug} P_{kj}^{aug} \overline{y}_j \right)^2$$

$$\leqslant 5\sum_{k=1}^{n+K} E[\tilde{\eta}_k^2|\mathcal{Z}]\left(\breve{w}_k^2\breve{y}_k^2+\left[\sum_{i,j}^{n+K}(P_{kj}^{aug})^2(P_{ki}^{aug})^2\overline{w}_i\overline{y}_i\overline{w}_j\overline{y}_j\right]+\breve{w}_k^2(P_{kk}^{aug})^2\overline{y}_k^2+\overline{w}_k^2(P_{kk}^{aug})^2\breve{y}_k^2+\overline{w}_k^2(P_{kk}^{aug})^4\overline{y}_k^2\right)$$

$$\leqslant 5\overline{\sigma}_\eta^2\left(\sum_{k=1}^n\breve{w}_k^2\breve{y}_k^2+\overline{\mu}_W^2\overline{\mu}_Y^2\sum_{i,j,k}^n(P_{kj}^{aug})^2(P_{ki}^{aug})^2+\overline{\mu}_Y^2\sum_{k=1}^n\breve{w}_k^2+\overline{\mu}_W^2\sum_{k=1}^n\breve{y}_k^2+4n\overline{\mu}_W^2\overline{\mu}_Y^2\right)$$

$$\leqslant 5\overline{\sigma}_\eta^2\left(\sum_{k=1}^n\breve{w}_k^2\breve{y}_k^2+7n\overline{\mu}_W^2\overline{\mu}_Y^2\right)\leqslant C\sum_{k=1}^n\breve{w}_k^2\breve{y}_k^2+Cn/n^2\leqslant C\sum_{k=1}^n\breve{w}_k^2\breve{y}_k^2+o(1)$$

$$\leqslant(\max_{i\leqslant n}|\breve{w}_i|)^2\sum_{k=1}^n\breve{y}_k^2=o(1)\sum_{k=1}^n\breve{y}_k^2\to 0.$$

This completes the proof of the Lemma 4.

□

Lemma 5. *If Assumptions 1-3 are satisfied then*

(i) $S_n^{-1}\tilde{H}S_n^{-1}=\sum_{i\neq j}z_iP_{ij}^\Lambda(1-P_{jj}^\Lambda)^{-1}z_j'/n+o_P(1)$.

(ii). $S_n^{-1}\sum_{i\neq j}X_iP_{ij}^\Lambda(1-P_{jj}^\Lambda)^{-1}\epsilon_j=O_P(1+\sqrt{K/r_n})$.

Proof. The proof is similar to the proof of Lemma A5 in Chao et. al. with our Lemma 1 replacing their Lemma A1. It is included for convenience. The strategy will be to apply Lemma 1 repeatedly to different components of the quantities of interest to obtain the desired bounds. Let e_k be the kth unit vector and apply Lemma 1 with $Y_i=e_k'S_n^{-1}X_i=z_ik/\sqrt{n}+e_k'S_n^{-1}U_i$ and $W_i=e_l'S_n^{-1}X_i(1-P_{ii}^\Lambda)^{-1}$ for some k and l between 1 and G. By assumption 2, $\lambda_{\min}(S_n)\leqslant C\sqrt{r}_n$ which implies that $\|S_n^{-1}\|\geqslant C/\sqrt{r}_n$. Therefore, a.s., all of the following hold:

$$\bar{y}_i:=E[Y_i|\mathcal{Z}]=z_{ik}/\sqrt{n},\quad var(Y_i|\mathcal{Z})\leqslant C/r_n$$

$$\bar{w}_i:=E[W_i|\mathcal{Z}]=z_{ik}(1-P_{ii}^\Lambda)^{-1}/\sqrt{n},\quad var(W_i|\mathcal{Z})\leqslant C/r_n.$$

Then, for $\bar{\sigma}_{Y_n}$, $\bar{\sigma}_{W_n}$, $\bar{y}$ and $\bar{w}$ defined as above , it follows that, a.s.,

$$\sqrt{K}\overline{\sigma}_{W_n}\overline{\sigma}_{Y_n}\leqslant C\sqrt{K}/r_n\to 0$$

$$\overline{\sigma}_{W_n}\sqrt{\overline{y}'\overline{y}}\leqslant Cr_n^{-1/2}\sqrt{\sum_{i=1}^n z_{ik}^2/n}\to 0$$

$$\overline{\sigma}_{Y_n}\sqrt{\overline{w}'\overline{w}}\leqslant Cr_n^{-1/2}\sqrt{\sum_{i=1}^n z_{il}^2(1-P_{ii}^\Lambda)^{-2}/n}\leqslant Cr_n^{-1/2}(1-\max_i P_{ii}^\Lambda)^{-2}\sqrt{\sum_{i=1}^n z_{il}^2/n}\to 0.$$

Then $\sum_{i\neq j} Y_i P_{ij}^{\Lambda} W_j = e_k' S_n^{-1} \sum_{i\neq j} P_{ij}^{\Lambda} X_j' S_n^{-1\prime} e_l/(1-P_{jj}^{\Lambda}) = e_k' S_n^{-1} \tilde{H} S_n^{-1\prime} e_l$ and $P_{ij}^{\Lambda}\overline{w}_i\overline{y}_j = P_{ij}^{\Lambda} z_{ik} z_{jl}/n(1-P_{jj}^{\Lambda})$, and so Lemma 1 is applied to show that

$$E[(e_k' S_n^{-1}\tilde{H} S_n^{-1\prime} e_l - \sum_{i\neq j} e_k' z_i P_{ij}^{\Lambda}(1-P_{jj}^{\Lambda})^{-1} z_j' e_l/n)^2 | \mathcal{Z}] \to 0.$$

Then consider the event $A_{n,v} := \{(|e_k' S_n^{-1}\tilde{H} S_n^{-1\prime} e_l - \sum_{i\neq j} e_k' z_i P_{ij}^{\Lambda}(1-P_{jj}^{\Lambda})^{-1} z_j' e_l/n| > v\}$.

Then by the conditional Markov inequality, for any $v > 0$,

$$\mathrm{P}(A_{n,v}|\mathcal{Z}) \overset{a.s.}{\to} 0.$$

Dominated convergence then gives $\mathrm{P}(A_{n,v}) = E[\mathrm{P}(A_{n,v}|\mathcal{Z})] \to 0$. This can be repeated for all $k, l \leqslant G$, giving the convergence of all components of $S_n^{-1}\tilde{H}S_n^{-1}$ to complete the proof of (i).

To show (ii), apply Lemma 1 with $Y_i = e_k' S_n^{-1} X_i$ and $W_i = \xi_i$. Note that $\overline{w}_i = 0$ and $\overline{\sigma}_{W_n} \leqslant C$. For fixed k and l, Lemma 1 gives

$$E[(e_k' S_n^{-1} \sum_{i\neq j} X_i P_{ij}^{\Lambda} \xi_j e_l)^2 | \mathcal{Z}] \leqslant CK/r_n + C.$$

The conclusion follows by the same argument as for (i).

□

Lemma 6. *If Assumptions 1-4 are satisfied then,* $S_n^{-1}\tilde{H}S_n^{-1} = \sum_{i=1}^n z_i \widehat{z}_i/n + o_P(1)$

Proof. Algebra gives that $\sum_{i=1}^n z_i\widehat{z}_i/n = \sum_{i\neq j} z_i P_{ij}^{\Lambda}(1-P_{ii}^{\Lambda})^{-1} z_j'/n$. Then the result is immediate from Lemma 5. □

Definition 1. The following definitions are convenient for Lemmas 7 and 8 as well as for proving Theorem 4:

(i) $\dot{X}_i = S_n^{-1} X_i$,

(ii) $\widehat{\Sigma}_1 = \sum_{i\neq j\neq k} \dot{X}_i P_{ik}^{\Lambda} \tilde{\xi}_i^2 P_{kj}^{\Lambda} \dot{X}_j'$,

(iii) $\dot{\Sigma}_1 = \sum_{i\neq j\neq k} \dot{X}_i P_{ik}^{\Lambda} \xi_i^2 P_{kj}^{\Lambda} \dot{X}_j'$,

(iv) $\widehat{\Sigma}_2 = \sum_{i\neq j} {P_{ij}^{\Lambda}}^2 (\dot{X}_i \dot{X}_i' \tilde{\xi}_i^2 + \dot{X}_i \tilde{\xi}_j \tilde{\xi}_j \dot{X}_j')$,

(v) $\dot{\Sigma}_2 = \sum_{i\neq j} {P_{ij}^{\Lambda}}^2 (\dot{X}_i \dot{X}_i' \xi_i^2 + \dot{X}_i \xi_j \xi_j \dot{X}_j')$.

Lemma 7. *Under Assumptions 1-6,* $\widehat{\Sigma}_1 - \dot{\Sigma}_1 = o_P(1)$ *and* $\widehat{\Sigma}_1 - \dot{\Sigma}_1 = o_P(K/r_n)$.

Proof. The proof of Lemma 7 is similar to the proof of lemma A7 given in CSHNW.

Let $X_i^{P^{\Lambda}} = X_i/(1-P_{ii}^{\Lambda})$. Then $\tilde{\xi}_i^2 - \xi_i^2 = -2\xi_i {X_i^{P^{\Lambda}}}'(\tilde{\delta}-\delta_0) + [{X_i^{P^{\Lambda}}}'(\tilde{\delta}-\delta_0)]^2$. Let η_i be any component of $-2\xi_i {X_i^{P^{\Lambda}}}'$ or $X_i^{P^{\Lambda}} {X_i^{P^{\Lambda}}}'$. Note that $S_n/\sqrt{n}$ is bounded, so that $\|\Upsilon_i\| \leqslant C$ for some C. Then

$E[\eta_i^2|\mathcal{Z}] \leqslant CE[\xi_i^2|\mathcal{Z}] + CE[\|X_i\|^2|\mathcal{Z}] \leqslant C + C\|\Upsilon_i\|^2 + CE[\|U_i\|^2|\mathcal{Z}] \leqslant C$. By Lemma 4,

$$\sum_{i\neq j\neq k} \dot{X}_i P_{ik}^{\Lambda} \eta_k P_{kj}^{\Lambda} \dot{X}_j' = O_P(1).$$

Considering the expression for $\tilde{\xi}_i^2 - \xi_i^2$, it is clear that the expression for $\widehat{\Sigma}_1 - \dot{\Sigma}_1$ is a sum of terms of the form

$$\Delta \sum_{i\neq j\neq k} \dot{X}_i P_{ik}^{\Lambda} \eta_k P_{kj}^{\Lambda} \dot{X}_j'$$

where $\Delta = o_P(1)$. Therefore, after applying triangle inequality, $\widehat{\Sigma}_1 - \dot{\Sigma}_1 \xrightarrow{p} 0$.

Next consider the second conclusion that $\widehat{\Sigma}_2 - \dot{\Sigma}_2 = o_P(K/r_n)$. Consider the random variables $\widehat{A} = 1 + \|\tilde{\delta}\|$, $\widehat{B} = \|\tilde{\delta} - \delta_0\|$ and $d_i = C + |\epsilon_i| + \|U_i\|$ where C is such that $\|\Upsilon_i\| \leqslant C$. Then the following inequalities all hold:

$\|X_i\| \leqslant C + \|U_i\| \leqslant d_i$,

$\|\dot{X}_i\| \leqslant C r_n^{-1/2} d_i$,

$|\tilde{\xi}_i - \xi_i| \leqslant C|X_i'(\tilde{\delta} - \delta_0)| \leqslant C d_i \widehat{B}$,

$|\xi_i| \leqslant C|X_i'(\tilde{\delta} - \delta_0) + |\tilde{\xi}_i - \xi_i| \leqslant C d_i \widehat{A}$,

$|\tilde{\xi}_i^2 - \xi_i^2| \leqslant (|\xi_i + |\tilde{\xi}_i|)|\tilde{\xi}_i - \xi_i| \leqslant C d_i (1 + \widehat{A}) d_i \widehat{B} \leqslant C d_i^2 \widehat{A}\widehat{B}$,

$\|\dot{X}_i(\tilde{\xi}_i - \xi_i)\| \leqslant C r_n^{-1/2} d_i^2 \widehat{B}$,

$\|\dot{X}_i \tilde{\xi}_i\| \leqslant C r_n^{-1/2} d_i^2 \widehat{A}$,

$\|\dot{X}_i \xi_i\| \leqslant C r_n^{-1/2} d_i^2$.

Because $E[d_i^2|\mathcal{Z}] \leqslant C$, it follows that

$$E[\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 d_i^2 d_j^2 r_n^{-1}|\mathcal{Z}] \leqslant C r_n^{-1} \sum_{i,j} {P_{ij}^{\Lambda}}^2 \leqslant C r_n^{-1} \sum_{i=1}^{n} P_{ii}^{\Lambda} \leqslant CK/r_n.$$

therefore, $\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 d_i^2 d_j^2 r_n^{-1} = O_P(K/r_n)$.

Then bound $\widehat{\Sigma}_2 - \dot{\Sigma}_1 = \sum_{i\neq j} {P_{ij}^{\Lambda}}^2 (\dot{X}_i \dot{X}_i'(\tilde{\xi}_j^2 - \xi_j^2)) + \sum_{i\neq j} {P_{ij}^{\Lambda}}^2 (\dot{X}_i \tilde{\xi}_i \tilde{\xi}_j \dot{X}_j' - \dot{X}_i \xi_i \xi_j \dot{X}_j')$ by considering the two terms on the right hand side seperately as follows:

$$\|\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 (\dot{X}_i \dot{X}_i'(\tilde{\xi}_j^2 - \xi_j^2))\| \leqslant \sum_{i\neq j} {P_{ij}^{\Lambda}}^2 \|\dot{X}_i\|^2 |\tilde{\xi}_j^2 - \xi_j^2| \leqslant C r_n^{-1} \sum_{i\neq j} {P_{ij}^{\Lambda}}^2 d_i^2 d_j^2 \widehat{A}\widehat{B} = o_P(K/r_n)$$

and,

$$\|\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 (\dot{X}_i \tilde{\xi}_i \tilde{\xi}_j \dot{X}_j' - \dot{X}_i \xi_i \xi_j \dot{X}_j')\| \leqslant \sum_{i\neq j} {P_{ij}^{\Lambda}}^2 (\|\dot{X}_i \tilde{\xi}_i\| \|\dot{X}_j(\tilde{\xi}_j^2 - \xi_j^2)\| + \|\dot{X}_j \xi_j\| \|\dot{X}_i(\tilde{\xi}_i - \xi_i)\|)$$

$$\leqslant Cr_n^{-1}\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 d_i^2 d_j^2 \widehat{A}\widehat{B} = o_P(K/r_n).$$

Then triangle inequality implies the second statement of the Lemma.

□

Lemma 8. *Under Assumptions 1 - 6,*

$\dot{\Sigma}_1 = \sum_{i\neq j\neq k} z_i P_{ik}^{\Lambda} E[\xi_k^2|\mathcal{Z}] P_{kj}^{\Lambda} z_j'/n + o_P(1)$ *and,*

$\dot{\Sigma}_2 = \sum_{i\neq j} z_i z_i' E[\xi_i^2|\mathcal{Z}] + S_n^{-1}\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 \left(E[U_iU_i'|\mathcal{Z}]E[\xi_i^2|\mathcal{Z}] + E[U_i\xi_i'|\mathcal{Z}]E[U_j'\xi_j^2|\mathcal{Z}]\right) {S_n^{-1}}' + o_P(K/r_n)$

Proof. The proof of Lemma 8 is similar to the proof given in CSHNW.

Apply Lemma 4 with W_i equal to an element of $\dot{X}_i$ and Y_j equal to an element of $\dot{X}_j$ and $\eta_k = \xi_k^2$ to prove the first conlcusion.

Next, use Lemma 3 to prove the second assertion by the following argument. Note that because $var(\xi^2|\mathcal{Z}) \leqslant C$ and $r_n \leqslant Cn$, then the following inequalities hold with u_{ki} defined by $u_{ki} = e_k' S_n^{-1} U_i$:

$$E[(\dot{X}_{ik}\dot{X}_{il})^2|\mathcal{Z}] \leqslant CE[\dot{X}_{ik}^4 + \dot{X}_{il}^4|\mathcal{Z}]$$
$$\leqslant C(z_{ik}^4/n^2 + E[u_{ki}^4|\mathcal{Z}] + z_{il}^4/n^2 + E[u_{li}^4|\mathcal{Z}]) \leqslant C/r_n^2$$

and

$$E[(\dot{X}_{ik}\xi_i)^2|\mathcal{Z}] \leqslant CE[(z_{ik}^2\xi_i^2/n + u_{ki}^2\xi_i^2|\mathcal{Z}] \leqslant C/n + C/n \leqslant C/r_n.$$

For $\Omega_i := E[U_iU_i'|\mathcal{Z}]$, then $E[\dot{X}_i\dot{X}_i'|\mathcal{Z}] = z_iz_i'/n + S_n^{-1}\Omega_i{S_n^{-1}}'$ and $E[\dot{X}_i\xi_i|\mathcal{Z}] = S_n^{-1}E[U_i\xi_i|\mathcal{Z}]$. Next let W_i be $\dot{X}_{ik}\dot{X}_{il}$ for some k and l, so that

$E[W_i|\mathcal{Z}] = e_k' S_n^{-1}\Omega_i {S_n^{-1}}' e_l + z_{ik}z_{il}/n,$

$|E[W_i|\mathcal{Z}]| \leqslant C/r_n,$

$E[(\dot{X}_{ik}\dot{X}_{il})^2|\mathcal{Z}] \leqslant C/r_n^2.$

Finally, let $Y_i = \xi_i^2$ and note that $|E[Y_i|\mathcal{Z}]| \leqslant C$. Then by Lemma 3,

$$\|\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 W_iY_j - E[\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 W_iY_j]\|_{L_2,\mathcal{Z}}^2 \leqslant CB_n$$

where $B_n = K\left(\overline{\sigma}_W^2\overline{\sigma}_Y^2 + \overline{\sigma}_W^2\overline{\mu}_Y^2 + \overline{\mu}_W^2\overline{\sigma}_Y^2\right) \leqslant \sqrt{K}(C/r_n + C/r_n + C/r_n) \leqslant C\sqrt{K}/r_n$ Therefore, putting in the values for W_i and Y_i,

$$\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 \dot{X}_{ik}\dot{X}_{il}'\xi_j^2 = e_k'\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 (z_iz_i'/n + S_n^{-1}\Omega_i{S_n^{-1}}')e_l E[\xi_j^2|\mathcal{Z}] + O_P(\sqrt{K}/r_n)$$

This bounds the first term in the difference. Similarly, consider an application of Lemma 3 with $W_i = \dot{X}_{ik}\xi_i$ and $Y_i = \dot{X}_{il}\xi_i$. This way $\overline{\sigma}_{W_n}\overline{\sigma}_{Y_n} + \overline{\sigma}_{W_n}\overline{\sigma}_{Y_n} + \overline{\sigma}_{W_n}\overline{\sigma}_{Y_n} \leqslant C/r_n$. Then it follows that

$$\sum_{i\neq j} {P_{ij}^{\Lambda}}^2 \dot{X}_{ik}\xi_i\xi_j\dot{X}_{jl} = e_k' S_n^{-1} \sum_{i\neq j} {P_{ij}^{\Lambda}}^2 E[U_i\xi_i|\mathcal{Z}]E[\xi_j U_j'|\mathcal{Z}]{S_n^{-1}}' e_l + O_P(\sqrt{K}/r_n).$$

Finally, since K grows to infinity with n, $O_P(\sqrt{K}/r_n) = o_P(K/r_n)$. The result follows from application of Triangle inequality.

□

7.2. **Proof of Theorem 1.**

Proof. First we note that $S_n'(\tilde{\delta} - \delta_0)/r_n \xrightarrow{p} 0$ implies $\tilde{\delta} \xrightarrow{p} \delta_0$. This is because

$$\|S_n'(\tilde{\delta} - \delta_0)/r_n\| \geqslant \sqrt{\lambda_{min}(S_n S_n'/r_n)}\|\tilde{\delta} - \delta_0\| \geqslant \sqrt{\lambda_{min}(\tilde{S}_n\tilde{S}_n')}\|\tilde{\delta} - \delta_0\| \geqslant C\|\tilde{\delta} - \delta_0\|.$$

Therefore, it suffices to prove the statement $S_n'(\tilde{\delta} - \delta_0)/r_n \xrightarrow{p} 0$.

First observe that, conditional on $\mathcal{Z}$, $\lambda_{min}(S_n^{-1}\tilde{H}{S_n^{-1}}') = \lambda_{min}(\sum_{i=1}^n z_i\widehat{z}_i'/n + o_P(1))$ by lemma 6. Then $\lambda_{min}(S_n^{-1}\tilde{H}{S_n^{-1}}') \geqslant \lambda_{min}(\sum_{i=1}^n z_i\widehat{z}_i'/n) + o_P(1) \geqslant 1/C + o_P(1)$. Where the last inequality follows from Condition 4 and holds a.s.n. Therefore, $(S_n^{-1}\tilde{H}{S_n^{-1}}')^{-1} = O_p(1)$.

Second, observe that conditional on $\mathcal{Z}$, $S_n^{-1}\sum_{i\neq j} X_i P_{ij}^{\Lambda}\xi_j/\sqrt{r_n} = O_P(1 + \sqrt{K/r_n})/\sqrt{r_n} = o_P(1)$ provided that $\sqrt{K}/r_n \to 0$.

Putting these together shows that conditional on $\mathcal{Z}$,

$$r_n^{-1/2}S_n'(\tilde{\delta} - \delta_0) = (S_n^{-1}\tilde{H}{S_n^{-1}}')^{-1}S_n^{-1}\sum_{i\neq j} X_i P_{ij}^{\Lambda}\xi_j/\sqrt{n} = O_P(1)o_P(1) \xrightarrow{p} 0.$$

To this point, every statement was conditional on $\mathcal{Z}$. The unconditional statement follows by dominated convergence: let $R_n = r_n^{-1/2}S_n'(\tilde{\delta} - \delta_0)$. Then the above argument shows that for any constant $v > 0$, a.s., $\mathrm{P}(\|R_n\| \geqslant v|\mathcal{Z}) \to 0$. Then by dominated convergence, $\mathrm{P}(\|R_n\| \geqslant v) = E[\mathrm{P}(\|R_n\| \geqslant v|\mathcal{Z})] \to 0$. Since v was arbitrary, it follows that $R_n \xrightarrow{p} 0$ proving the theorem.

□

7.3. **Proof of Theorem 2.**

Proof. Let $Y_n = S_n^{-1}\sum_{i\neq j} X_i P_{ij}^{\Lambda}\xi_j$ and notice that it can be decomposed as

$$Y_n = \sum_i (\overline{z}_i - P_{ii}^{\Lambda} z_i)\xi_i/\sqrt{n} + S_n^{-1}\sum_{i\neq j} U_i P_{ij}^{\Lambda}\xi_j.$$

Let $\Gamma_n = var(Y_n|\mathcal{Z})$ so that

$$\Gamma_n = \sum_{i=1}^n (\overline{z}_i - P_{ii}^{\Lambda} z_i)(\overline{z}_i - P_{ii}^{\Lambda} z_i)' E[\xi_i^2|\mathcal{Z}]/n$$

$$+S_n^{-1}\sum_{i\neq j}(P_{ij}^{\Lambda})^2\left(E[U_iU_i'|\mathcal{Z}]E[\xi_i^2|\mathcal{Z}]+E[U_i\xi_i'|\mathcal{Z}]E[U_j'\xi_j^2|\mathcal{Z}]\right)S_n^{-1\prime}.$$

We first show that Γ_n is nondegenerate and bounded. Once that is done, we will show that the central limit theorem given in Lemma 2 holds for Y_n to obtain the desired asymptotic distribution.

Since $P_{ii}^{\Lambda}<C$ a.s.n., $E[\xi_i^2|\mathcal{Z}]=(1-P_{ii}^{\Lambda})^{-2}E[\epsilon_i^2|\mathcal{Z}]\geqslant C$ a.s.n. Then in the positive definite sense

$$\Gamma_n\succeq\sum_{i=1}^{n}(\overline{z}_i-P_{ii}^{\Lambda}z_i)(\overline{z}_i-P_{ii}^{\Lambda}z_i)'E[\xi_i^2|\mathcal{Z}]/n\succeq C\sum_{i=1}^{n}(\overline{z}_i-P_{ii}^{\Lambda}z_i)(\overline{z}_i-P_{ii}^{\Lambda}z_i)'/n$$

Now, let ω be a $G\times 1$ vector with norm $\|\omega\|=1$. Then note that by the Cauchy-Schwartz Inequality and $\lambda_{min}(\sum_i z_i\widehat{z}_i'/n)>C>0$ from assumption 4,

$$0<C<\omega'\sum_{i=1}^{n}z_i\widehat{z}_i'\omega/n=\sum_{i=1}^{n}\omega'z_i(1-P_{ii}^{\Lambda})^{-1}(\overline{z}_i-P_{ii}^{\Lambda}z_i)'\omega/n$$
$$\leqslant\sqrt{\sum_{i=1}^{n}\omega'(\overline{z}_i-P_{ii}^{\Lambda}z_i)(\overline{z}_i-P_{ii}^{\Lambda}z_i)'\omega/n}\sqrt{\sum_{i=1}^{n}\omega'(1-P_{ii}^{\Lambda})^{-2}z_iz_i'\omega/n}.$$

Therefore

$$\sqrt{\sum_{i=1}^{n}\omega'(\overline{z}_i-P_{ii}^{\Lambda}z_i)(\overline{z}_i-P_{ii}^{\Lambda}z_i)'\omega/n}\geqslant C\left(\sqrt{\sum_{i=1}^{n}\omega'(1-P_{ii}^{\Lambda})^{-2}z_iz_i'\omega/n}\right)^{-1}.$$

By $P_{ii}^{\Lambda}\leqslant C<1$, it follows that $\sum_{i=1}^{n}(1-P_{ii}^{\Lambda})^{-2}z_iz_i'/n\leqslant\sum_{i=1}^{n}Cz_iz_i'/n$ in the positive definite sense. This fact along with $\|\sum_{i=1}^{n}z_iz_i'/n\|\leqslant C$ from Assumption 2 give that $\left(\sqrt{\sum_{i=1}^{n}\omega'(1-P_{ii}^{\Lambda})^{-2}z_iz_i'\omega/n}\right)^{-1}\geqslant C$. Therefore, $\sqrt{\sum_{i=1}^{n}\omega'(\overline{z}_i-P_{ii}^{\Lambda}z_i)(\overline{z}_i-P_{ii}^{\Lambda}z_i)'\omega/n}\geqslant C>0$ which in turn implies that $\lambda_{\min}(\Gamma_n)\geqslant C>0$ a.s.n. This regularity in Γ_n will justify using Lemma 2.

Note additionally that, $\|\Gamma_n\|\leqslant C$ a.s.n. because $\sum_{i\neq j}P_{ij}^{\Lambda}/K\leqslant 1$, and $\|\sum_{i=1}^{n}z_zz_i'/n\|<C$ and $E[\xi_i^2|\mathcal{Z}]<C$. Therefore, the eigenvalues of Γ_n^{-1} are bounded away from zero a.s.n.

In anticipation of using the Cramer-Wold device, let α be a nonzero $G\times 1$ vector.

Let $W_{in}=(\overline{z}_i-P_{ii}^{\Lambda}z_i)\xi_i/\sqrt{n}$. Let $c_{1n}=\Gamma_n^{-1/2}\alpha$, and let $c_{2n}=\sqrt{K}S_n^{-1}\Gamma_n^{-1/2}\alpha$. Next, we show that all the conditions of Lemma 2 are satisfied. Condition (i) is satisfied by Assumption 1. Condition (ii) is satisfied since $D_{1,n}\preceq\Gamma_n$. Condition (iii) is satisfied by Assumption 3. Condition (iv) is satisfied by:

$$\begin{aligned}\sum_{i=1}^n E[\|W_{i,n}\|^4|\mathcal{Z}] &= \frac{1}{n^2}\sum_{i=1}^n E[\|(\overline{z}_i - P_{ii}^\Lambda z_i)\xi_i\|^4|\mathcal{Z}] \\ &= \frac{1}{n^2}\sum_i E[\|(\overline{z}_i - P_{ii}^\Lambda z_i)\|^4|\mathcal{Z}]E[\xi_i^4|\mathcal{Z}] \\ &\leqslant \frac{1}{n^2}\sum_{i=1}^n E[\|(\overline{z}_i - P_{ii}^\Lambda z_i)\|^4|\mathcal{Z}]\cdot C \\ &\leqslant \frac{1}{n^2}\sum_{i=1}^n E[2^{4-1}\|\overline{z}_i\|^4 + 2^{4-1}\|P_{ii}^\Lambda z_i\|^4|\mathcal{Z}]\cdot C \\ &\leqslant \frac{1}{n^2}C\sum_{i=1}^n E[\|\overline{z}_i\|^4|\mathcal{Z}] + \frac{1}{n^2}C\sum_{i=1}^n E[\|z_i\|^4|\mathcal{Z}]\end{aligned}$$

By Assumption 5, both terms above are vanishing in the limit as $n \to \infty$. Finally, condition (v) of Lemma 2 is satisfied by Assumption 1.

Note that $c_{1n} = \Gamma_n^{-1/2}\alpha$ and $c_{2n} = (\sqrt{K/r_n})\sqrt{r_n}S_n^{-1}\Gamma_n^{-1/2}$ satisfy $\|c_{1n}\| \leqslant C$ and $\|c_{2n}\| \leqslant C$ a.s.n because of the boundedness of $\sqrt{K/r_n}$, $\sqrt{r_n}S_n^{-1}$, and Γ_n^{-1}. Also, Ξ_n in lemma 2 is given by:

$$\Xi_n = c_{1n}'D_nc_{1n} + c_{2n}'\widehat{\Sigma}_nc_{2n} = \text{var}(\alpha'\Gamma_n^{-1/2\prime}Y_n|\mathcal{Z}) = \alpha'\alpha.$$

An application of lemma 2 yields that

$$(\alpha'\alpha)^{-1/2}\alpha'\Gamma_n^{-1/2}Y_n = \Xi_n^{-1/2}\left(\sum_{i=1}^n c_{1n}'W_{in} + c_{2n}'\sum_{i\neq j}U_iP_{ij}^\Lambda\xi_j/\sqrt{K}\right) := \bar{Y}_n \xrightarrow{d} N(0,1) \text{ a.s.}$$

Therefore, $\alpha'\Gamma_n^{-1/2}Y_n \xrightarrow{d} N(0,\alpha'\alpha)$ a.s., so by the Cramer-Wold device, $\Gamma_n^{-1/2}Y_n \xrightarrow{d} N(0,I_G)$ a.s.

Now, recall V_n is defined by $V_n = H_n^{-1}\Gamma_nH_n^{-1}$ for $H_n = \sum_{i=1}^n z_i\widehat{z}_i'/n$. Let $Q_n = V_n^{-1/2}H_n\Gamma_n^{1/2}$. Q_n is an orthogonal matrix since $Q_nQ_n' = V_n^{-1/2}H_n\Gamma_n^{1/2}\Gamma_n^{1/2\prime}H_nV_n^{-1/2\prime} = V_n^{-1/2}V_nV_n^{-1/2\prime} = I_n$. In addition, Q_n depends only on $\mathcal{Z}$.

Therefore,

$$V_n^{-1/2}(S_n^{-1}\tilde{H}S_n^{-1\prime})^{-1}\Gamma_n^{-1/2} = V_n^{-1/2}(H_n + o_P(1))\Gamma_n^{1/2} = Q_n + o_P(1).$$

Note that because $\Gamma_n^{-1/2}Y_n \xrightarrow{d} N(0,I_G)$ a.s., and Q_n is only a function of $\mathcal{Z}$, we have that $Q_n\Gamma_n^{-1/2}Y_n \xrightarrow{d} N(0,I_G)$. Then by the Slutsky lemma and $\tilde{\delta} = \delta_0 + \tilde{H}^{-1}\sum_{i\neq j}X_iP_{ij}^\Lambda\xi_j$, we have

$$V_n^{-1/2}S_n'(\tilde{\delta} - \delta_0) = V_n^{-1/2}(S_n^{-1}\tilde{H}^{-1}S_n^{-1\prime})^{-1}S_n^{-1}\sum_{i\neq j}X_iP_{ij}^\Lambda\xi_j$$

$$= V_n^{-1/2}(S_n^{-1}\tilde{H}^{-1}S_n^{-1\prime})^{-1}(Y_n + o_P(1))$$

$$= (Q_n + o_P(1))(\Gamma_n^{-1/2}Y_n + o_P(1))$$

$$= Q_n\Gamma_n^{-1/2}Y_n + o_P(1) \xrightarrow{d} N(0, I_G)$$

□

7.4. **Proof of Theorem 3.**

Proof. Because $r_n/K \to 0$, we have that $\sqrt{r_n/K}\sum_i(\overline{z}_i - P_{ii}^{\Lambda}z_i)\xi_i/\sqrt{n} \xrightarrow{p} 0$. Therefore, that term is negligable in the expansion

$$\sqrt{r_n/K}S_n^{-1}\sum_{i\neq j}X_iP_{ij}^{\Lambda}\xi_j = \sqrt{r_n/K}\sum_i(\overline{z}_i - P_{ii}^{\Lambda}z_i)\xi_i/\sqrt{n} + \sqrt{r_n/K}S_n^{-1}\sum_{i\neq j}U_iP_{ij}^{\Lambda}\xi_j.$$

Thus, we consider only the second term on the right hand side and redefine $Y_n := \sqrt{r_n}S_n^{-1}\sum_{i\neq j}U_iP_{ij}^{\Lambda}\xi_j/\sqrt{K}$ for this proof. Let Γ_n be the conditional variance,

$$\Gamma_n = var(Y_n|\mathcal{Z}) = r_n s_n^{-1}\sum_{i\neq j}(P_{ij}^{\Lambda})^2(E[U_iU_i'|\mathcal{Z}]E[\xi_j^2|\mathcal{Z}] + E[U_i\xi_i|\mathcal{Z}]E[U_j'\xi_j|\mathcal{Z}])S_n^{-1\prime}/K.$$

As before, $\|\Gamma_n\| \leqslant C$ a.s.n. Let $\overline{L}_n$ be any sequence of bounded matrices with $\lambda_{\min}(\overline{L}_n\Gamma_n\overline{L}_n') \geqslant C > 0$ a.s.n. Let $Y_n^L = (\overline{L}_n\Gamma_n\overline{L}_n')^{-1/2}\overline{L}_nY_n$. We apply lemma 2 in a similar fashion as for proving theorem 2. Let α be a nonzero vector. Consider $W_{in} = 0$, $C_{1n} = 0$, $c_{2n} = \alpha'(\overline{L}_n\Gamma_n\overline{L}_n')^{-1/2}\overline{L}_n\sqrt{r_n}S_n^{-1}$. Then $var(c_{2n}'\sum_{i\neq j}U_iP_{ij}^{\Lambda}\xi_j/\sqrt{K}|\mathcal{Z}) = \alpha'\alpha > 0$ and Lemma 2 implies that $\alpha'Y_n^L \xrightarrow{d} N(0, \alpha'\alpha)$ a.s. Therefore, $Y_n^L \xrightarrow{d} N(0, I_\ell)$.

Next, for L_n given in the hypothesis of the theorem, let $\overline{L}_n = L_nH_n^{-1}$. This way, $L_nV_n^*L_n' = \overline{L}_n\Gamma_n\overline{L}_n'$. Applying Lemma 6 gives $(S_n^{-1}\tilde{H}S_n^{-1\prime})^{-1} = H_n^{-1} + o_P(1)$, from which it follows that

$$(\overline{L}_n\Gamma_n\overline{L}_n')^{-1/2}L_n(S_n^{-1}\tilde{H}S_n^{-1\prime})^{-1} = (\overline{L}_n\Gamma_n\overline{L}_n')^{-1/2}L_n(H_n + o_P(1)) = (\overline{L}_n\Gamma_n\overline{L}_n')^{-1/2}\overline{L}_n + o_P(1).$$

Finally, $\sqrt{r_n/K}S_n^{-1}\sum_{i\neq j}X_iP_{ij}^{\Lambda}\xi_j = O_P(1)$ by Lemma 5, so

$$(\overline{L}_nV_n^*L_n')^{-1/2}L_n\sqrt{r_n/K}S_n'(\tilde{\delta} - \delta_0)$$

$$= (\overline{L}_n\Gamma_n\overline{L}_n')^{-1/2}L_n(\overline{L}_n\Gamma_n\overline{L}_n')^{-1}\sqrt{r_n/K}S_n^{-1}\sum_{i\neq j}X_iP_{ij}^{\Lambda}\xi_j$$

$$= (\overline{L}_n\Gamma_n\overline{L}_n')^{-1/2}\overline{L}_n + o_P(1))(Y_n + o_P(1)) = Y_n^L + o_P(1) \xrightarrow{d} N(0, I_\ell).$$

□

7.5. **Proof of Theorem 4.**

Proof. We begin by expressing $\Omega_n + \Psi_n$ in terms of $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$.

$$\Omega_n + \Psi_n = \sum_i E[\xi_i^2|\mathcal{Z}](\overline{z}_i - P_{ii}^{\Lambda} z_i)(\overline{z}_i - P_{ii}^{\Lambda} z_i)'/n$$

$$+S_n^{-1} \sum_{i \neq j} {P_{ij}^{\Lambda}}^2 \left(E[U_i U_i'|\mathcal{Z}] E[\xi_i^2|\mathcal{Z}] + E[U_i \xi_i'|\mathcal{Z}] E[U_j' \xi_j^2|\mathcal{Z}] \right) {S_n^{-1}}'$$

$$= \sum_i E[\xi_i^2|\mathcal{Z}](\overline{z}_i - P_{ii}^{\Lambda} z_i)(\overline{z}_i - P_{ii}^{\Lambda} z_i)'/n - \sum_{i \neq j} {P_{ij}^{\Lambda}}^2 z_i z_i' E[\xi_i^2|\mathcal{Z}]$$

$$+ \sum_{i \neq j} {P_{ij}^{\Lambda}}^2 z_i z_i' E[\xi_i^2|\mathcal{Z}] + S_n^{-1} \sum_{i \neq j} {P_{ij}^{\Lambda}}^2 \left(E[U_i U_i'|\mathcal{Z}] E[\xi_i^2|\mathcal{Z}] + E[U_i \xi_i'|\mathcal{Z}] E[U_j' \xi_j^2|\mathcal{Z}] \right) {S_n^{-1}}'$$

$$= \sum_i E[\xi_i^2|\mathcal{Z}](\overline{z}_i - P_{ii}^{\Lambda} z_i)(\overline{z}_i - P_{ii}^{\Lambda} z_i)'/n - \sum_{i \neq j} {P_{ij}^{\Lambda}}^2 z_i z_i' E[\xi_i^2|\mathcal{Z}] + \dot{\Sigma}_2 + o_P(K/r_n)$$

(by Lemma 8)

$$= \sum_i E[\xi_i^2|\mathcal{Z}](\overline{z}_i \overline{z}_i' - P_{ii}^{\Lambda} z_i \overline{z}_i' - P_{ii}^{\Lambda} \overline{z}_i z_i' + {P_{ii}^{\Lambda}}^2 z_i z_i')/n - \sum_{i \neq j} {P_{ij}^{\Lambda}}^2 z_i z_i' E[\xi_i^2|\mathcal{Z}] + \dot{\Sigma}_2 + o_P(K/r_n)$$

$$= \sum_{i \neq j \neq k} z_i P_{ik}^{\Lambda} E[\xi_i^2|\mathcal{Z}] P_{kj}^{\Lambda} z_j'/n + \dot{\Sigma}_2 + o_P(K/r_n)$$

$$= \dot{\Sigma}_1 + o_P(1) + \dot{\Sigma}_2 + o_P(K/r_n) \qquad \text{(by Lemma 8)}$$

$$= \widehat{\Sigma}_1 + o_P(1) + \widehat{\Sigma}_2 + o_P(K/r_n) \qquad \text{(by Lemma 7)}$$

In the case that K/r_n is bounded, Lemma 6 implies that

$$S_n' \tilde{V} S_n = (S_n^{-1} \tilde{H} {S_n^{-1}}')^{-1} (\widehat{\Sigma}_1 + \widehat{\Sigma}_2)(S_n^{-1} \tilde{H} {S_n^{-1}}')^{-1}$$

$$= (H_n^{-1} + o_P(1))(\Omega_n + \Psi_n + o_P(1))(H_n^{-1} + o_P(1)) = V_n + o_P(1)$$

due to H_n^{-1} and $\Omega_n + \Psi_n$ being bounded a.s.n.

In the other case that $K/r_n \to \infty$, then $(r_n/K)(\widehat{\Sigma}_1 + \widehat{\Sigma}_2) = (r_n/K)\Psi_n + (r_n/K)\Omega_n + o_P(1) = (r_n/K)\Psi_n + o_P(1)$. Since $(r_n/K)\Psi_n$ is bounded a.s.n, we have

$$S_n' \tilde{V} S_n = (S_n^{-1} \tilde{H} {S_n^{-1}}')^{-1} (\widehat{\Sigma}_1 + \widehat{\Sigma}_2)(S_n^{-1} \tilde{H} {S_n^{-1}}')^{-1}$$

$$= (H_n^{-1} + o_P(1))((r_n/K)\Psi_n + o_P(1))(H_n^{-1} + o_P(1)) = V_n^* + o_P(1).$$

□

References

Amemiya, T. (1966): "On the Use of Principal Components of Independent Variables in Two-Stage Least-Squares Estimation," *International Economic Review*, 7, 283–303.

Amemiya, T. (1974): "The Non-linear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 2, 105–110.

Andrews, D. W. K., and J. H. Stock (2007): "Inference with Weak Instruments," in *Advances in Economics and Econometrics, Theory and Applications, 9th Congress of the Econometric Society, Volume 3*, ed. by R. Blundell, W. Newey, and T. Persson. Cambridge University Press.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14(1), 57–67.

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

Bai, J., and S. Ng (2009): "Selecting Instrumental Variables in a Data Rich Environment," *Journal of Time Series Econometrics*, 1(1).

——— (2010): "Instrumental Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26, 15771606.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 63, 657–681.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2010): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *forthcoming Econometrica*.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, 37(4), 1705–1732.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90(430), 443–450.

Bühlmann, P. (2006): "Boosting for high-dimensional linear models," *Ann. Statist.*, 34(2), 559–583.

Caner, M. (2009): "LASSO-Type GMM Estimator," *Econometric Theory*, 25, 270–290.

Carrasco, M. (2012): "A Regularization Approach to the Many Instruments Problem," *forthcoming in Journal of Econometrics*.

Carrasco, M., and G. Tchuente Nguembu (2012): "Regularized LIML with Many Instruments," Discussion paper, University of Montreal Working paper.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Chamberlain, G., and G. Imbens (2004): "Random Effects Estimators with Many Instrumental Variables," *Econometrica*, 72, 295–306.

Chao, J., and N. 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*, 28(1), 42–86.

Dicker, L. (2012): "Optimal Estimation and Prediction for Dense Signals in High-Dimensional Linear Models," *ArXiv working paper*.

Donald, S. G., and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–1191.

Fuller, W. A. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–954.

Gautier, E., and A. B. Tsybakov (2011): "High-Dimensional Instrumental Variables Regression and Confidence Sets," *ArXiv working report*.

Hansen, C., J. Hausman, and W. K. Newey (2008): "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26, 398–422.

Kapetanios, G., L. Khalaf, and M. Marcellino (2011): "Factor based identification-robust inference in IV regressions," *working paper*.

Kapetanios, G., and M. Marcellino (2010): "Factor-GMM estimation with large sets of possibly weak instruments," *Computational Statistics & Data Analysis*, 54(11), 26552675.

Kloek, T., and L. Mennes (1960): "Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables," *Econometrica*, 28, 45–61.

Newey, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809–837.

Newey, W. K., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72(1), 219–255.

Okui, R. (2010): "Instrumental Variable Estimation in the Presence of Many Moment Conditions," *forthcoming Journal of Econometrics*.

Phillips, G. D. A., and C. Hale (1977): "The bias of instrumental variable estimators of simultaneous equation systems," *International Economic Review*, 18, 219–228.

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

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20(4), 518–529.

Tibshirani, R. (1996): "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

Table 1. Simulation Results many instruments K = 95

	Dense Signal			Sparse Signal		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
	A. Concentration Parameter = 30. Binary Instruments					
RJIVE	0.014	0.079	0.068	0.037	0.213	0.061
Post-LASSO	0.093	0.093	0.072	0.154	0.157	0.111
LASSO-JIVE	0.092	0.092	0.771	0.259	0.259	0.818
Carrasco	0.086	0.086	0.757	0.239	0.239	0.763
2SLS	0.103	0.103	0.969	0.284	0.284	0.977
JIVE	0.061	0.124	0.069	0.150	0.350	0.072
	B. Concentration Parameter = 30. Gaussian Instruments					
RJIVE	-0.004	0.025	0.061	-0.010	0.087	0.055
Post-LASSO	0.041	0.041	0.275	0.050	0.063	0.125
LASSO-JIVE	0.047	0.047	0.736	0.153	0.153	0.778
Carrasco	0.029	0.029	0.350	0.103	0.103	0.428
2SLS	0.052	0.052	0.958	0.166	0.166	0.966
JIVE	0.011	0.049	0.075	0.047	0.157	0.068
	C. Concentration Parameter = 150. Binary Instruments					
RJIVE	-0.006	0.053	0.051	-0.016	0.144	0.043
Post-LASSO	0.093	0.095	0.320	0.051	0.099	0.088
LASSO-JIVE	0.078	0.078	0.437	0.232	0.232	0.528
Carrasco	0.070	0.071	0.376	0.204	0.204	0.390
2SLS	0.102	0.102	0.700	0.287	0.287	0.715
JIVE	0.001	0.086	0.059	0.035	0.217	0.045
	D. Concentration Parameter = 150. Gaussian Instruments					
RJIVE	-0.001	0.019	0.053	-0.009	0.065	0.043
Post-LASSO	0.036	0.036	0.281	0.027	0.056	0.075
LASSO-JIVE	0.036	0.036	0.365	0.138	0.138	0.493
Carrasco	0.020	0.023	0.167	0.076	0.081	0.191
2SLS	0.049	0.049	0.655	0.154	0.154	0.644
JIVE	-0.001	0.028	0.043	0.004	0.097	0.045

Note: Results are based on 1500 simulation replications. We report Median Bias (Med. Bias), Median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for five different estimators: the RJIVE proposed in this paper (RJIVE); the Post-LASSO IV estimator of Belloni, Chernozhukov, Chen, and Hansen (2012, Post-LASSO), an estimator that uses LASSO model selection with a small penalty level that ensure that instruments are chosen and then uses the selected instruments with the JIVE (LASSO-JIVE); the estimator of Carrasco (2012, Carrasco); 2SLS; and JIVE. In all simulations, the correlation between the first stage error and structural error is set to 0.6.

Table 2. Simulation Results many instruments K = 190

	Dense Signal			Sparse Signal		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
	A. Concentration Parameter = 30. Binary Instruments					
RJIVE	0.028	0.073	0.067	0.086	0.292	0.071
Post-LASSO	0.077	0.077	0.064	0.177	0.179	0.081
LASSO-JIVE	0.078	0.078	0.991	0.306	0.306	0.996
Carrasco	0.069	0.069	0.909	0.268	0.268	0.907
2SLS	0.077	0.077	0.976	0.300	0.300	0.981
JIVE	0.062	0.112	0.054	0.249	0.414	0.063
	B. Concentration Parameter = 30. Gaussian Instruments					
RJIVE	-0.002	0.021	0.053	0.008	0.111	0.063
Post-LASSO	0.035	0.035	0.258	0.053	0.065	0.143
LASSO-JIVE	0.040	0.040	0.980	0.181	0.181	0.986
Carrasco	0.027	0.027	0.613	0.132	0.132	0.699
2SLS	0.039	0.039	0.974	0.176	0.176	0.977
JIVE	0.020	0.046	0.057	0.100	0.201	0.071
	C. Concentration Parameter = 150. Binary Instruments					
RJIVE	-0.002	0.051	0.044	-0.005	0.198	0.049
Post-LASSO	0.077	0.078	0.361	0.068	0.104	0.100
LASSO-JIVE	0.084	0.084	0.861	0.328	0.328	0.905
Carrasco	0.065	0.065	0.559	0.250	0.250	0.557
2SLS	0.081	0.081	0.741	0.318	0.318	0.753
JIVE	0.034	0.112	0.037	0.100	0.413	0.043
	D. Concentration Parameter = 150. Gaussian Instruments					
RJIVE	-0.001	0.015	0.046	-0.007	0.081	0.048
Post-LASSO	0.031	0.032	0.438	0.027	0.056	0.083
LASSO-JIVE	0.040	0.040	0.766	0.195	0.195	0.869
Carrasco	0.019	0.020	0.245	0.107	0.107	0.319
2SLS	0.038	0.038	0.709	0.177	0.177	0.717
JIVE	0.003	0.028	0.047	0.017	0.149	0.046

Note: Results are based on 1500 simulation replications. We report Median Bias (Med. Bias), Median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for five different estimators: the RJIVE proposed in this paper (RJIVE); the Post-LASSO IV estimator of Belloni, Chernozhukov, Chen, and Hansen (2012, Post-LASSO), an estimator that uses LASSO model selection with a small penalty level that ensure that instruments are chosen and then uses the selected instruments with the JIVE (LASSO-JIVE); the estimator of Carrasco (2012, Carrasco); 2SLS; and JIVE. In all simulations, the correlation between the first stage error and structural error is set to 0.6.

Table 3: Estimates of the Return to Schooling in Angrist and Krueger Data

	2SLS	Post-LASSO	JIVE	RJIVE
		A. 3 Instruments		
Schooling Coefficient	0.1079	0.1115	0.1091	0.1091
Estimated Standard Error	0.0196	0.0205	0.0202	0.0202
		B. 180 Instruments		
Schooling Coefficient	0.0928	0.1125	0.1096	0.1062
Estimated Standard Error	0.0097	0.0173	0.0161	0.0157
		C. 1527 Instruments		
Schooling Coefficient	0.0712	0.0862	0.0816	0.1067
Estimated Standard Error	0.0049	0.0254	0.5168	0.0171

Note: This table reports estimates of the returns-to-schooling parameter in the Angrist-Krueger 1991 data using different estimators and different numbers of instruments. In the rows, we give point estimates of the schooling coefficient and heteroskedasticity consistent standard error estimates. We report results for 2SLS, the Post-LASSO estimator of Belloni, Chen, Chernozhukov, and Hansen (2012) (Post-LASSO), JIVE, and our regularized JIVE (RJIVE). Further details are provided in the text. For comparison, the OLS estimate (standard error) of the schooling coefficient is 0.0673 (0.0004).