# Averaged Instrumental Variables Estimators

Yoonseok Lee* Yu Zhou†

April 2015

**Abstract**

We develop averaged instrumental variables estimators as a way to deal with many weak instruments. We propose a weighted average of the preliminary $k$-class estimators, where each estimator is obtained using different subsets of the available instrumental variables. The averaged estimators are shown to be consistent and to satisfy asymptotic normality. Furthermore, its approximate mean squared error reveals that using a small number of instruments for each preliminary $k$-class estimator reduces the finite sample bias, while averaging prevents the variance from inflating. Monte Carlo simulations find that the averaged estimators compare favorably with alternative instrumental-variable-selection approaches when the strength levels of individual IV are similar with each other.

*Keywords*: Averaged estimator, many weak instruments, $k$-class estimator.
*JEL Classifications*: C26, C36.

*Corresponding Author. Department of Economics and Center for Policy Research, Syracuse University, 426 Eggers Hall, Syracuse, NY 13244, U.S.A. *E-mail*: ylee41@maxwell.syr.edu.

†School of Economics, Fudan University, 600 Guoquan Road, Shanghai 200433, P.R.China. *E-mail*: yuzhou@fudan.edu.cn.

# 1 Introduction

Many recent studies on instrumental variables (IV) estimation have considered the cases that the number of instruments grows with the sample size but each individual instrument is only weakly correlated with the endogenous regressors. This paper contributes to this literature by developing averaged IV estimators as a way to deal with many weak instruments.

It has been well understood that using many IV potentially improves asymptotic efficiency but can cause large bias in finite samples (e.g., Morimune (1983)). It also makes the standard inference procedure inaccurate as the standard asymptotic theory may fail to work with many instruments (e.g., Bekker (1994); Lee and Okui (2012)). A natural way of handling many instruments is to choose a subset of them. For example, Donald and Newey (2001), further developed by Donald et al. (2009), choose the number of instruments by minimizing the approximate mean squared error of the IV estimator, with imposing some ordering over instruments. Kuersteiner and Okui (2010) refine this approach by applying model averaging idea of Hansen (2007) to the first stage regression to have a weighted average of the predicted endogenous regressor. Canay (2010) proposes to use a trapezoidal kernel over the moment conditions to select as well as to impose proper weights on them. Carrasco (2012) applies regularization to achieve the mean squared error improvement with many IV. For different approaches with many IV, Belloni et al. (2012) apply the lasso on the sparse first-stage regression and estimate optimal IV.

We propose an alternative but simple solution to deal with large number ($K$) of IV, even when the individual IV strength is unknown and potentially weak. We study a weighted average of the preliminary $k$-class estimators, where each preliminary estimator is obtained using different subsets from the entire IV. The motivation of this averaged IV estimator is straightforward: Since only a small number of instruments are used in each preliminary estimation (namely, $r \ll K$), the finite sample bias is reduced compared with the standard IV estimator using the entire $K$ number of IV; the margin is substantial if $r$ is much smaller than $K$. In the meanwhile, the efficiency loss from using only a subset of IV in each preliminary estimation can be minimized by the weighted average, where we put more weight on the

preliminary estimator that uses stronger subset of instruments. A similar idea on estimator averaging can be found in Sawa (1973), Guggenberger and Sun (2006), and Chen et al. (2015) under different contexts.

Analytically, we show that the averaged IV estimator satisfies the standard first-order many-weak-IV asymptotics (e.g., Chao and Swanson (2005); Hansen et al. (2008); Newey and Windmeijer (2009); Lee and Okui (2012); Chao et al. (2012)) without imposing restrictions on the limit of $K/n$. In order to investigate the averaged IV estimator in depth, we also derive its approximate mean squared error expression. This new estimation complements the similar approaches of Donald and Newey (2001) and Kuersteiner and Okui (2010); but the simulation results show potential benefits from this new estimator, particularly when the levels of individual IV's strength are similar with each other and hence it is somewhat unclear to choose a subset of stronger IV among them.

This new approach has distinct features from the existing ones. First, since it eventually uses all the available IV, it keeps researchers from choosing an arbitrary subset of IV and from imposing an ad hoc ordering over them. Second, we do not need to restrict the limit of $K/n$, and hence the new procedure is computationally feasible even when the number of instruments is larger than the sample size since the size of the subset ($r$) is the effective number of IV in each preliminary estimation. Third, since this approach does not rely on pre-testing or IV-selection procedures, inferences on the averaged IV estimator would not have the potential post-IV-selection inference problem.

The rest of the paper is organized as follows. Section 2 develops the averaged IV estimator and Section 3 lists the technical conditions, based on which the statistical properties of the averaged estimator are derived. Section 4 obtains the first-order asymptotic results of the averaged estimator and develops an overidentifying restriction test. Section 5 further studies the finite sample improvement by deriving the approximate mean squared error and compares the new estimator with other available approaches. Section 6 concludes the paper with some remarks. All the mathematical proofs are in the Appendix.

## 2 Averaged IV Estimators

We consider an instrumental variables (IV) regression model given by

$$y_i = x_i\beta + \varepsilon_i, \tag{1}$$

$$x_i = f_i + u_i = z_i'\pi + u_i \tag{2}$$

for $i = 1, 2, \cdots, n$, where $\beta$ is the main parameter of interest. $y_i$ and $x_i$ are scalar variables and $x_i$ is possibly correlated with an unobserved error $\varepsilon_i$, where the correlation is through the nonzero $cov(\varepsilon_i, u_i)$. For the independently and identically distributed (i.i.d.) vector of unobservables $e_i = (\varepsilon_i, u_i)'$, we define a finite and positive definite matrix $\Sigma = Var\left[e_i|z_i\right]$ as

$$\Sigma = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & \sigma_u^2 \end{pmatrix} \tag{3}$$

conformably as $(\varepsilon_i, u_i)'$, where $\sigma_{\varepsilon u} \neq 0$. We only consider the case with one endogenous regressor but generalization to the vector of endogenous regressors readily follows by properly defining the weight vectors $w_t$ in (5) below. We do not include exogenous regressors in (1) since we could consider all the variables as the fitted residuals of the orthogonal projection on the space spanned by the exogenous regressors.

We assume a $K \times 1$ vector of valid instruments $z_i$ with $\mathbb{E}\left[e_i|z_i\right] = 0$ for all $i$, where $K$ is allowed to increase with the sample size. Similarly as Hahn et al. (2004) and Hansen et al. (2008), the first stage regression (2) assumes $\mathbb{E}\left[x_i|z_i\right] = z_i'\pi$. For $K \to \infty$, it could be understood as an approximation of the unknown $f_i$ by some linear combination of $z_i = (z_{1,i}, \cdots, z_{K,i})'$ as Donald and Newey (2001). In this case, we let $u_i$ include the approximation error that is sufficiently small. Reflecting such idea and for analytical convenience, we further assume that all the instruments are orthogonalized without loss of generality.[1]

[1] This assumption implies that we look at each element of $z_i$ as an orthogonalized linear combination of the raw instrumental variables. A similar idea is used in Donald and Newey (2001) to facilitate the sieve approximation. We could alternatively understand the orthogonalized instruments as the principal components of the raw instrumental variables in data rich environment (e.g., Bai and Ng (2010)). For constructing orthogonal instruments, see Kuersteiner and Okui (2010, p.702).

To describe the averaged IV estimator, we let $z_i^t \in \mathbb{R}^r$ be a sub-vector of $z_i \in \mathbb{R}^K$:

$$z_i^t = (z_{t_1,i}, \cdots, z_{t_r,i})' \text{ for some } \{t_1, \cdots, t_r\} \subset \{1, \cdots, K\},$$

which is an $r$-dimensional subset of IV, where $1 \leq r < \min\{K, n\}$. We impose that $r$ is much smaller than $K$ but we still allow for $r$ to grow with the sample size though it satisfies $r/n \to 0$. More specific conditions are given in the following section. Using each subset of instruments $z_i^t$, we define the preliminary two stage least squares (2SLS) estimator as

$$\widehat{\beta}_{2sls,t} = \left(X'P_tX\right)^{-1} X'P_tY, \tag{4}$$

where $X = (x_1, \cdots, x_n)'$, $Y = (y_1, \cdots, y_n)'$ and $P_t = Z_t \left(Z_t'Z_t\right)^{-1} Z_t'$ for $Z_t = \left(z_1^t, \cdots, z_n^t\right)'$. We then define the *averaged 2SLS estimator* as

$$\widehat{\beta}_{a2sls} = \sum_{t=1}^{T} w_t \widehat{\beta}_{2sls,t} = \sum_{t=1}^{T} w_t \left(X'P_tX\right)^{-1} X'P_tY \tag{5}$$

for some weights $w_t = w_t(n, K, r)$ satisfying $\sum_{t=1}^{T} w_t = 1$ and $0 \leq w_t \leq 1$ for all $t$, where $T = \binom{K}{r}$. In general, for some $\underline{\widehat{\kappa}} = (\widehat{\kappa}_1, \cdots, \widehat{\kappa}_T)'$, we define the averaged $k$-class estimator as

$$\widehat{\beta}_{ak}\left(\underline{\widehat{\kappa}}\right) = \sum_{t=1}^{T} w_t \widehat{\beta}_{k,t}\left(\widehat{\kappa}_t\right), \tag{6}$$

where

$$\widehat{\beta}_{k,t}\left(\widehat{\kappa}_t\right) = \left(X'P_tX - \widehat{\kappa}_t X'X\right)^{-1} \left(X'P_tY - \widehat{\kappa}_t X'Y\right). \tag{7}$$

When $\widehat{\kappa}_t = 0$ for all $t$, we obtain $\widehat{\beta}_{a2sls}$ in (5); when $\widehat{\kappa}_t = \min_\beta (Y - X\beta)'P_t(Y - X\beta)/(Y - X\beta)'(Y - X\beta)$, we have the averaged LIML estimator denoted as $\widehat{\beta}_{aLIML}$; when $\widehat{\kappa}_t = (r-2)/n$, we have the averaged bias-corrected 2SLS estimator denoted as $\widehat{\beta}_{aB2sls}$.

It is worth comparing the averaged IV estimator $\widehat{\beta}_{a2sls}$ with the model-averaging 2SLS estimator by Kuersteiner and Okui (2010). Recall that the model-averaging 2SLS estimator is defined as $\widehat{\beta}_{ma2sls} = (X'(\sum_{t=1}^{K} \widetilde{w}_t P_t)X)^{-1} X'(\sum_{t=1}^{K} \widetilde{w}_t P_t)Y$ for some weights $\widetilde{w}_t$. Though

these two estimators look similar, the difference between them is evident. First, we take the average over the second stage estimators $\widehat{\beta}_{2sls,t}$ in (4), whereas Kuersteiner and Okui (2010) average over the predicted regressors $P_t X$ from the first stage regression (2). Second, we take the weighted average over the possibly non-nested subsets of the (unordered) instruments, whereas Kuersteiner and Okui (2010) take the average over the nested sets of instruments. In particular, they presume that $z_i^t \subset z_i^s$ for any $t < s$ based on the idea of model averaging by Hansen (2007) and the sieve approximation by Donald and Newey (2001). Third, the weights $\widetilde{w}_t$ in Kuersteiner and Okui (2010) can be negative in order to reduce the finite sample bias. In comparison, we restrict the weights $w_t$ to be nonnegative as they are mainly to control for the variance; the bias reduction in our case mostly comes from using a smaller number of IV in the first stage regression.

# 3 Assumptions

In order to derive the statistical properties of the averaged IV estimator, we impose the following conditions.

**Assumption 1** *(i)* $\{y_i, x_i, z_i\}$ *are i.i.d. for* $i = 1, 2, \cdots, n$. *(ii)* $e_i = (\varepsilon_i, u_i)'$ *is i.i.d. with mean zero, positive-definite variance* $\Sigma$ *in (3) and finite fourth moment, conditional on* $z_i$. *(iii) The columns of* $Z = (z_1, \cdots, z_n)'$ *are mutually orthogonal almost surely.*

Assumption 1 is standard in the IV estimation literature. We presume homoskedasticity in $e_i$ (c.f., Hausman et al. (2012)). Assumption 1-(ii) precludes the presence of invalid IV. Assumption 1-(iii) implies that there exists a positive integer $N$ such that for all $n \geq N$, $Z_t$ is of full column rank $r$ almost surely for every $t$. The orthogonality condition is imposed for analytical convenience, though it is stronger than excluding the possibility of multicollinearity in (2).

**Assumption 2** *(i)* $r \to \infty$ *as* $n, K \to \infty$ *satisfying* $r/n \to 0$. *(ii)* $\sup_{1 \leq i \leq n} P_{t,ii} \to_p 0$ *as* $n \to \infty$ *for any* $t = 1, 2, \cdots, T$, *where* $P_{t,ii}$ *is the* $(i,i)$*-th element of* $P_t = Z_t(Z_t'Z_t)^{-1}Z_t'$. *(iii) There exists a nondecreasing sequence of positive real numbers* $\{\lambda = \lambda(n)\}$ *such that* $\lambda/n \to \xi$

*with* $0 \leq \xi \leq 1$, *which satisfies* $\pi' Z' Z \pi / \lambda \to_p \overline{H} < \infty$ *as* $n, K \to \infty$ *for some nonrandom positive definite matrix* $\overline{H}$. *(iv)* $\sup_{1 \leq i \leq n} |\pi' z_i| = O_p(1)$.

We use only a small number ($r$) of instruments when we obtain the preliminary estimator $\widehat{\beta}_{k,t}(\widehat{\kappa}_t)$ in (7). Therefore, the effective number of IV is $r$ here and hence we only need to restrict how fast $r$ can diverge comparing with the sample size $n$ as Assumption 2-(i). The total number of IV ($K$) is no longer important for our asymptotic analysis and we do not need to restrict the limit of $K/n$. Assumption 2-(ii) generally holds under the fixed number of IV or moderately many IV cases. In our case, it holds because $rank(P_t) = r$ and $r/n \to 0$. In Assumption 2-(iii), $\lambda$ can be interpreted as the growth rate of the concentration parameter, where the concentration parameter is defined as $\pi' Z' Z \pi / \sigma_u^2$ in this framework. Therefore, this condition highlights how fast the concentration parameter (or the signal from the instruments) grows as the number of instruments increases with the sample size and hence how weak the instruments are. In particular, as the limit of $\lambda/n$ lies in $[0, 1]$, not every IV can be strong in this case. A similar condition is also found in Chao and Swanson (2005, Assumption 1) and Hansen et al. (2008, Assumption 2). Assumptions 2-(iii) and (iv) rule out lack of identification but they suppose that $Z$ includes weak instruments so that adding more IV does not necessarily enrich information.[2] Further conditions on the ratio between the number of sub-instruments $r$ and $\lambda$ are to be given later.

We decompose $Z\pi = [Z_t \pi_t, Z_{-t} \pi_{-t}]$, where $Z_t$ is the $n \times r$ instrument matrix at the $t$-th choice and $Z_{-t}$ is the $n \times (K - r)$ matrix of the remaining instruments; $\pi_t$ and $\pi_{-t}$ are the corresponding sub-parameter matrices of $\pi$. Since the instruments are mutually orthogonal from Assumption 1-(iii), we can decompose $P = P_t + P_{-t}$, where $P = Z(Z'Z)^{-1}Z'$, $P_t = Z_t(Z_t'Z_t)^{-1}Z_t'$ and $P_{-t} = Z_{-t}(Z_{-t}'Z_{-t})^{-1}Z_{-t}'$. Hence, $P_t Z\pi = Z_t \pi_t$ and $(I - P_t)Z\pi = Z_{-t}\pi_{-t}$ because $(I - P_t)Z = (I - P)Z + P_{-t}Z$ and $(I - P)Z = 0$ by construction.

**Assumption 3** *For each* $t$, *(i) there exists a nondecreasing sequence of positive real numbers* $\{\lambda_t = \lambda_t(n)\}$ *such that* $\pi_t' Z_t' Z_t \pi_t / \lambda_t \to_p \overline{H}$ *as* $n, r \to \infty$, *where* $\overline{H}$ *is defined in Assumption*

[2] Assumption 2-(iv) restricts that $\pi' z_i$ cannot blow up even for large $K$ and hence it could be understood as a counterpart of the (approximate) sparsity condition in Belloni et al. (2012).

*2; (ii) there exist non-negative real numbers* $\delta_t$ *such that* $\sup_{1\leq t\leq T} r^{2\delta_t}(\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda) = O_p(1)$ *and* $\lim_{T\to\infty}\sum_{t=1}^{T} r^{-2\delta_t} < \infty$.

Assumption 3-(i) states that as $r$ increases the signal from each subset of the instruments gets richer and hence $\pi'_t Z'_t Z_t \pi_t/\lambda_t$ has the same probability limit as what $\pi' Z' Z\pi/\lambda$ has, if it is properly normalized by some value $\lambda_t$. Here, $\lambda_t$ can vary across $t$ reflecting the different strength of each subset of instruments $Z_t$. At the same time, Assumption 3-(ii) implies that $\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda = \pi' Z'(I - P_t)Z\pi/\lambda$ is at most $O_p(r^{-2\delta_t})$ for each $t$, where $r^{-2\delta_t}$ measures the ratio of the loss of the signal from using only a subset of the available instruments to the signal from the entire set of instruments, $\pi' Z' Z\pi = O_p(\lambda)$. A similar condition is imposed in Donald and Newey (2001) and Kuersteiner and Okui (2010) in the context of sieve approximation error.

Note that $\delta_t$ carries the information on the strength of the remaining instruments $Z_{-t}$. If the $t$-th subset of instruments $Z_t$ includes most of the relatively stronger IV and hence the rest of the instruments in $Z_{-t}$ are quite weak, then $\delta_t$ is large so that $\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda$ shrinks toward zero fast enough. To the other extreme case, if $Z_t$ only includes very weak IV, then $\delta_t$ is close to zero so that $\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda$ can be $O_p(1)$.[3]

Understanding that the strength of each subset of IV can vary, we need to weaken the effect from the very weak IV subsets in the averaged estimator, so that it has desired statistical properties. One natural way is to impose relatively small weights on the preliminary IV estimates when averaging, if these preliminary estimates are based on relatively weak IV subsets. Without loss of generality, we re-order the index $t$ such that $\delta_t$ in Assumption 3-(ii) satisfies $\delta_t \leq \delta_s$ if $t < s$, which implies that the strength of the subset of the instruments $Z_t$ gets stronger with $t$ and hence $\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda$ shrinks to zero faster. Note that this re-ordering is only for analytical convenience and for formalizing conditions on the weights as in the following assumption; in practice, we do not need to impose such ordering.

[3] If we define $\lambda_{-t}$ such that $\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda_{-t} \to_p \overline{H}$ as $n, r \to \infty$, then we have $\pi' Z' Z\pi/\lambda = (\pi'_t Z'_t Z_t \pi_t/\lambda_t)(\lambda_t/\lambda) + (\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda_{-t})(\lambda_{-t}/\lambda) \to_p \overline{H} \times \lim_{n,r\to\infty}[(\lambda_t/\lambda) + (\lambda_{-t}/\lambda)]$ from Assumptions 2 and 3. It thus requires that $\lim_{n,r\to\infty}[(\lambda_t/\lambda) + (\lambda_{-t}/\lambda)] = 1$. As $\lambda_{-t}/\lambda = O(r^{-2\delta_t})$ from Assumption 3-(ii), therefore, it is either $\lambda_t/\lambda \to 0$ when $\delta_t = 0$; or $\lambda_t/\lambda \to 1$ when $\delta_t > 0$. The interpretation is quite intuitive: as $r, K \to \infty$, one of the partitions of the instruments, $Z_t$ or $Z_{-t}$, should be stronger than the other and hence be dominant in the limit.

**Assumption 4** *(i) For $t, s = 1, 2, \cdots, T$, the weights $w_t = w_t(n, K, r)$ satisfy $0 \leq w_t \leq w_s \leq 1$ for any $t < s$, and $\sum_{t=1}^{T} w_t = 1$. (ii) There exists a sequence of positive integers $\{L = L(T)\}$ such that $L \to \infty$ as $T \to \infty$ and $\sum_{t=1}^{L} w_t \to 0$.*

Assumption 4 imposes sufficiently small weights over the first $L$ weakest subsets of instruments; in this way we can mitigate their effects in constructing the averaged IV estimator. In fact, Assumption 4-(ii) is similar to Kuersteiner and Okui (2010, Assumption 4). It is also closely related to the conditions on the tail behavior of kernel functions in the kernel weighted 2SLS estimators like Canay (2010), Okui (2011), and Kuersteiner (2012), where the kernel weight corresponds to $w_t$ in this case. Note that, as shown in Lemma A.4 in the Appendix, Assumptions 3-(ii) and 4-(ii) together imply that $\sum_{t=1}^{T} w_t r^{-2\delta_t} \to 0$, which yields $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\pi'_{-t} Z'_{-t} Z_{-s} \pi_{-s}/\lambda\right) = o_p(1)$. This condition becomes important to derive that the averaged IV estimator enjoys the standard first-order many-weak-IV asymptotics in the following section.

Lastly, the following condition is required when we consider the statistical properties of the general $k$-class estimators as (6).

**Assumption 5** *For each $t$, $\widehat{\kappa}_t - r/n = o_p(r/n)$.*

Recall that $\widehat{\kappa}_t = \min_\beta (Y - X\beta)' P_t (Y - X\beta)/(Y - X\beta)'(Y - X\beta)$ for $\widehat{\beta}_{aLIML}$ and it can be shown that $\widehat{\kappa}_t - \varepsilon' P_t \varepsilon / n\sigma_\varepsilon^2 = o_p(r/n)$ because $\mathbb{E}(\varepsilon' P_t \varepsilon) = \sigma_\varepsilon^2 tr(P_t) = r\sigma_\varepsilon^2$ as Donald and Newey (2001, Lemma A.7). For $\widehat{\beta}_{aB2sls}$, $\widehat{\kappa}_t = (r-2)/n$ and this condition automatically holds.

# 4 Asymptotic Properties

In this section, we show that the averaged IV estimator has the same first-order asymptotic properties as the standard many-weak-IV estimators by choosing sufficiently small $r$.[4] We also

[4] When $\lim_{n,K\to\infty} K/n \neq 0$, the standard 2SLS estimator using all the IV is known to be inconsistent (e.g., Bekker (1994)). Though the bias-corrected 2SLS or the LIML estimators are consistent in this case, it is still required that $K < n$ to make the estimation procedure computationally feasible. Since $K$ does not have any role in the asymptotic analysis here, the results in this section naturally extend to the case of Bekker (1994) or even for the undersized sample case by choosing sufficiently small $r$.

develop an averaged overidentifying restriction test statistic using the averaged IV estimator. The first theorem obtains consistency of the averaged estimators.

**Theorem 1** *When* $\sqrt{r}/\lambda \to 0$, $\widehat{\beta}_{aLIML} \to_p \beta$ *and* $\widehat{\beta}_{aB2sls} \to_p \beta$ *as* $n, r, \lambda \to \infty$ *under Assumptions 1-5. However,* $\widehat{\beta}_{a2sls} \to_p \beta$ *only when* $r/\lambda \to 0$ *as* $n, r, \lambda \to \infty$ *under Assumptions 1-4.*

Unlike Chao and Swanson (2005), we only need to choose sufficiently small $r$ to achieve consistency instead of imposing conditions on the entire number of IV, $K$. In particular, the order of magnitude of $\lambda$ is obtained relative to $r$, which could allow that the growth rate of the concentration parameter of the entire IV ($\lambda$) can be much slower than the cases of Chao and Swanson (2005). Theorem 1, however, still shows similar findings as Chao and Swanson (2005) in certain aspects: $\widehat{\beta}_{a2sls}$ is less able to withstand instrument weakness than $\widehat{\beta}_{aLIML}$ or $\widehat{\beta}_{aB2sls}$ since it needs $r/\lambda \to 0$ instead of $\sqrt{r}/\lambda \to 0$. But IV cannot be too weak even for the case of $\widehat{\beta}_{aLIML}$ and $\widehat{\beta}_{aB2sls}$ because $r/\lambda$ cannot diverge.

The following theorem shows that the averaged IV estimator asymptotically follows the normal distribution under proper conditions. Recall that we let $P_{t,ij}$ be the $(i,j)$-th element of $P_t = Z_t(Z_t'Z_t)^{-1}Z_t'$. The additional condition given in the following theorem states the absolute summability of the weighted sum of the projection matrices $\sum_{t=1}^{T} w_t P_t$, which gives the Lindeberg condition as van Hasselt (2010), though each projection matrix $P_t$ does not need to be absolutely summable (but its rows are square-summable from the property of the projection matrix).

**Theorem 2** *Let* $\sup_{n\geq 1} \sup_{1\leq i\leq n} \sum_{j=1}^{n} |\sum_{t=1}^{T} w_t P_{t,ij}| = O_p(1)$ *hold. When* $\sqrt{r}/\lambda \to 0$ *but* $r/\lambda \to \varsigma < \infty$, $\sqrt{\lambda}(\widehat{\beta}_{aLIML} - \beta) \to_d \mathcal{N}(0, \Lambda)$ *as* $n, r, \lambda \to \infty$ *and so is* $\widehat{\beta}_{aB2sls}$ *under Assumptions 1-5, where* $\Lambda = \sigma_\varepsilon^2 \overline{H}^{-1} + \varsigma \overline{H}^{-1} \left\{ \mathbb{E}(\varepsilon_i^2 u_i^2) - \sigma_{\varepsilon u}^2 \right\} \overline{H}^{-1}$. *However,* $\widehat{\beta}_{a2sls}$ *satisfies* $\sqrt{\lambda}(\widehat{\beta}_{a2sls} - \beta) \to_d \mathcal{N}(0, \sigma_\varepsilon^2 \overline{H}^{-1})$ *only when* $r/\lambda \to 0$ *as* $n, r, \lambda \to \infty$ *under Assumptions 1-4.*

As we discussed in the previous section, Assumptions 3-(ii) and 4-(ii) together yield that $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\pi'_{-t} Z'_{-t} Z_{-s} \pi_{-s}/\lambda\right) = o_p(1)$. Since

$$\begin{aligned} & \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \frac{\pi'_t Z'_t Z_s \pi_s}{\lambda} \\ = \quad & \frac{\pi' Z' Z \pi}{\lambda} + \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \frac{\pi'_{-t} Z'_{-t} Z_{-s} \pi_{-s}}{\lambda} - 2\sum_{t=1}^{T} w_t \frac{\pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t}}{\lambda} = \frac{\pi' Z' Z \pi}{\lambda} + o_p(1) \end{aligned}$$

from $P_t P_s = I + (I - P_t)(I - P_s) - (I - P_t) - (I - P_s)$ for any $t$ and $s$, this result ensures that the weighted average of all the variances and covariances of the preliminary estimators $\widehat{\beta}_{k,t}(\widehat{\kappa}_t)$ is well described using the probability limit of $\pi' Z' Z \pi/\lambda$. Hence, the asymptotic variance of the averaged IV estimator corresponds to those in the many-weak-IV literature, which are also derived using the probability limit of $\pi' Z' Z \pi/\lambda$. For instance, the results for $\widehat{\beta}_{aLIML}$ and $\widehat{\beta}_{aB2sls}$ resemble those of Hansen et al. (2008) or Chao et al. (2012), where the second term of the asymptotic variance $\Lambda$ depends on the difference $\mathbb{E}(\varepsilon_i^2 u_i^2) - (\mathbb{E}\varepsilon_i u_i)^2$ and the limit of the ratio $r/\lambda$. When $r/\lambda \to 0$, Theorem 2 also shows that the asymptotic variances of $\widehat{\beta}_{aLIML}$, $\widehat{\beta}_{aB2sls}$, and $\widehat{\beta}_{a2sls}$ are the same as $\sigma_\varepsilon^2 \overline{H}^{-1}$, which is the case of the standard IV estimator with $\lambda = n$.

Once we have the averaged $k$-class estimator, one could consider an overidentifying restriction test statistic as $\widetilde{\mathcal{J}} = (\widehat{\varepsilon}'_{ak} P \widehat{\varepsilon}_{ak})/(\widehat{\varepsilon}'_{ak}\widehat{\varepsilon}_{ak}/n)$, where $\widehat{\varepsilon}_{ak}$ is the regression residual from any averaged $k$-class estimator. When the total number of instruments $K$ is small enough (e.g., $K^2/n \to 0$), it can be expected that $\Pr\{\widetilde{\mathcal{J}} \geq q_{\alpha,K-1}\} \to \alpha$ from Theorems 1 and 2 above, where $q_{\alpha,K-1}$ is the $1-\alpha$ quantile of the $\mathcal{X}^2_{K-1}$ distribution (e.g., Newey and Windmeijer (2009)). However, since we do not restrict the limit of $K/n$, we need a different form of the overidentifying restriction test here. In particular, we propose an *averaged overidentifying restriction test statistic* defined as

$$\widehat{\mathcal{J}}_a = \sum_{t=1}^{T} \frac{w_t \widehat{\varepsilon}'_t P_t \widehat{\varepsilon}_t}{\widehat{\varepsilon}'_{ak}\widehat{\varepsilon}_{ak}/n}, \tag{8}$$

where $\widehat{\varepsilon}_t = Y - X\widehat{\beta}_{k,t}(\widehat{\kappa}_t)$ and $\widehat{\varepsilon}_{ak} = Y - X\widehat{\beta}_{ak}(\underline{\widehat{\kappa}}) = \sum_{t=1}^{T} w_t \widehat{\varepsilon}_t$ with $\widehat{\beta}_{k,t}(\widehat{\kappa}_t)$ being either the LIML or the bias-corrected 2SLS estimator for each $t$.[5] In the following theorem, we obtain the many-weak-IV asymptotic result of $\widehat{\mathcal{J}}_a$ similarly as Lee and Okui (2012), which does not require restrictions on $K/n$ as long as the sub-instrument size $r$ is small enough.

**Theorem 3** *Let the conditions of Theorem 2 hold and both $x_i$ and $\varepsilon_i$ have finite eighth moments. For any $r \geq 2$, it holds that* ***(a)*** $\lim_{n,r,\lambda\to\infty} \Pr\{\widehat{\mathcal{J}}_a \geq q_{\alpha,r-1}\} \leq \alpha$ *provided* $\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4 \leq 3$*, where $q_{\alpha,r-1}$ is the $1-\alpha$ quantile of the $\mathcal{X}^2_{r-1}$ distribution and the equality holds when* $\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4 = 3$*;* ***(b)*** $\lim_{n,r,\lambda\to\infty} \Pr\{(\widehat{\mathcal{J}}_a - r)/\sqrt{r(\hat{\mu}_\varepsilon - 1)} \geq q^*_\alpha\} = \alpha$*, where $q^*_\alpha$ is the $1-\alpha$ quantile of the standard normal distribution and* $\hat{\mu}_\varepsilon = (n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_{t,i}^4)/(n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_{t,i}^2)^2$ *for $\widehat{\varepsilon}_{t,i}$ being the $i$-th element of $\widehat{\varepsilon}_t$.*

Since $\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4 = 3$ when $\varepsilon_i$ is normally distributed, Theorem 3-(a) implies that $\widehat{\mathcal{J}}_a$ asymptotically has correct size under normality if the chi-square critical values are used. One limitation of this result is that, however, we could have over-rejection when $\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4 > 3$. Theorem 3-(b), on the other hand, suggests a bias-correction approach to achieve the correct asymptotic size even without normality.

# 5 Discussions

The basic motivation of the averaged IV estimator is in two folds: Since only a small number of instruments are used in each preliminary estimation, the finite sample bias is reduced compared with the standard IV estimator using the entire number of IV; the margin is substantial when $r$ is much smaller than $K$. Meanwhile, the variance of the averaged IV estimator is well controlled by the weighting scheme, where the weights are chosen according to the strength of each subset of instruments.

To be more precise about this idea, we consider the approximate mean squared error (MSE) of the averaged IV estimator based on the Nagar (1959) type asymptotic expansion,

[5] Note that for the many (weak) IV asymptotics, the standard 2SLS estimator is asymptotically biased and so is the Sargan test statistic, unless the ratio of the number of IV to the sample size goes to zero. For this reason, we recommend using either the LIML or the bias-corrected 2SLS estimator when $\lim_{r,\lambda\to\infty} r/\lambda \neq 0$. See Lee and Okui (2012) for further discussions.

similarly as Donald and Newey (2001) and Kuersteiner and Okui (2010). In particular, denoting $\underline{w} = (w_1, \cdots, w_T)'$, we approximate the conditional MSE of the averaged $k$-class estimator, $\mathbb{E}[\lambda(\widehat{\beta}_{ak}(\widehat{\underline{\kappa}}) - \beta)^2|Z]$, by

$$\sigma_\varepsilon^2 H^{-1} + S(r, \underline{w}), \tag{9}$$

where $\lambda(\widehat{\beta}_{ak}(\widehat{\underline{\kappa}}) - \beta)^2 = \widehat{Q}(r, \underline{w}) + \widehat{\gamma}(r, \underline{w})$, $\mathbb{E}[\widehat{Q}(r, \underline{w})|Z] = \sigma_\varepsilon^2 H^{-1} + S(r, \underline{w}) + R(r, \underline{w})$, $H = \pi' Z' Z \pi / \lambda$ and $[\widehat{\gamma}(r, \underline{w}) + R(r, \underline{w})]/tr(S(r, \underline{w})) = o_p(1)$ as $n, \lambda, r \to \infty$. Note that, as we can see from Theorem 2 above, the plim of $\sigma_\varepsilon^2 H^{-1}$ corresponds to the (first-order) variance for all the averaged $k$-class estimators when $r/\lambda \to 0$, and hence finding a explicit expression of $S(r, \underline{w})$ for each case is the main goal here. Based on the assumptions in Section 3, we derive the approximate MSE as follows. We let $v_i = u_i - \varepsilon_i \sigma_{\varepsilon u}/\sigma_\varepsilon^2$, $\sigma_v^2 = Var[v_i|z_i]$, and $\Delta(r, \underline{w}) = \sum_{t=1}^T \sum_{s=1}^T w_t w_s (\pi_{-t}' Z_{-t}' Z_{-s} \pi_{-s}/\lambda) = ||\sum_{t=1}^T w_t (I - P_t) Z\pi/\sqrt{\lambda}||^2$, where $\Delta(r, \underline{w}) = o_p(1)$ from Lemma A.4 in the Appendix.

**Theorem 4** ***(a)*** *If Assumptions 1-4 hold, $\sigma_{\varepsilon u} \neq 0$ and $r^2/\lambda \to 0$, then for $\widehat{\beta}_{a2sls}$ the decomposition (9) holds with $S(r, \underline{w})$ given by $S_{a2sls}(r, \underline{w}) = H^{-1} \left\{ \sigma_{\varepsilon u}^2 \left( r^2/\lambda \right) + \sigma_\varepsilon^2 \Delta(r, \underline{w}) \right\} H^{-1}$.* ***(b)*** *If Assumptions 1-5 hold, $\sigma_v^2 \neq 0$, $\mathbb{E}[\varepsilon_i^2 v_i | z_i] = 0$ and $r/\lambda \to 0$, then for $\widehat{\beta}_{aLIML}$ the decomposition (9) holds with $S_{aLIML}(r, \underline{w}) = H^{-1} \left\{ \sigma_\varepsilon^2 \sigma_u^2 \left( r/\lambda \right) + \sigma_\varepsilon^2 \Delta(r, \underline{w}) \right\} H^{-1}$.* ***(c)*** *If Assumptions 1-4 hold, $\sigma_{\varepsilon u} \neq 0$, $\mathbb{E}[\varepsilon_i^2 u_i | z_i] = 0$ and $r/\lambda \to 0$, then for $\widehat{\beta}_{aB2sls}$ the decomposition (9) holds with $S_{aB2sls}(r, \underline{w}) = H^{-1} \left\{ (\sigma_\varepsilon^2 \sigma_u^2 + \sigma_{\varepsilon u}^2) \left( r/\lambda \right) + \sigma_\varepsilon^2 \Delta(r, \underline{w}) \right\} H^{-1}$.*

The second term in $S_{a2sls}(r, \underline{w})$ corresponds to the second-order asymptotic variance by using only $r$-number of IV for preliminary estimation, whereas the first term corresponds to the squared bias. It is well known that the finite sample bias is proportional to the number of instruments for the 2SLS estimator. Since we only use a subset of the IV instead of using the entire set, the number of (effective) instruments reduces from $K$ to $r$ in each step of preliminary estimation and thus the finite sample bias of the averaged IV estimator is smaller than that of the standard IV estimator for $\sum_{t=1}^T w_t = 1$. In comparison, both terms in $S_{aLIML}(r, \underline{w})$ and $S_{aB2sls}(r, \underline{w})$ are from the second-order asymptotic variance, where these

estimators automatically correct the leading bias.

When $\lambda = n$, Theorem 4 appears to be similar to the results of Donald and Newey (2001) or Kuersteiner and Okui (2010). Since the forms of the estimators are quite different, however, we cannot conduct the formal comparison of Theorem 4 with their results. Instead, we make a partial comparison to illustrate the difference between our estimator with theirs. Recall that for the 2SLS estimator case, Donald and Newey (2001) obtain the second term of (9) as $S_{DN}(K_{DN}) = H^{-1}\{\sigma^2_{\varepsilon u}(K^2_{DN}/n) + D_{DN}(K_{DN}, n)\}H^{-1}$ for some $D_{DN}(K_{DN}, n) = o_p(1)$ and $K^2_{DN}/n \to 0$, where $K_{DN}$ is the increasing number of instruments yielding a proper series approximation to the unknown first stage IV regression function. On the other hand, Kuersteiner and Okui (2010) obtain $S_{KO}(K_{KO}, \underline{\widetilde{w}}) = H^{-1}\{\sigma^2_{\varepsilon u}((\sum_{k=1}^{K_{KO}} k\widetilde{w}_k)^2/n) + D_{KO}(K_{KO}, \underline{\widetilde{w}}, n)\}H^{-1}$ for some weight $\underline{\widetilde{w}} = (\widetilde{w}_1, \cdots, \widetilde{w}_{K_{KO}})'$ satisfying $D_{KO}(K_{KO}, \underline{\widetilde{w}}, n) = o_p(1)$ and $(\sum_{k=1}^{K_{KO}} k\widetilde{w}_k)^2/n \to 0$, in which $K_{KO}$ is some chosen number of instruments that can increase with the sample size. Comparing the first terms in $S_{DN}(K_{DN})$, $S_{KO}(K_{KO}, \underline{\widetilde{w}})$, and $S_{a2sls}(r, \underline{w})$, we can see that the finite sample bias depends on the ratios $K^2_{DN}/n$, $(\sum_{k=1}^{K_{KO}} k\widetilde{w}_k)^2/n$ and $r^2/n$, respectively. This comparison gives a rough idea where the bias reductions come from and how much they can be different with each other.

For further comparisons, we conduct a simulation, whose result is summarized in Tables 1 and 2. We consider the two stage regressions in (1) and (2) with $z_i \sim i.i.d.\mathcal{N}(0, I_K)$ and $(\varepsilon_i, u_i)' \sim i.i.d.\mathcal{N}\left(0, \left(\begin{smallmatrix} 1 & c \\ c & 1 \end{smallmatrix}\right)\right)$, where $\beta = 0.1$ and $(n, K, r) = (200, 16, 2)$, $(200, 32, 2)$. The degree of endogeneity is controlled by $c$, where we consider $c = 0.5$, $0.8$. For the first stage regression (2), we consider the following five specifications of $\pi$:

$$\text{M-1: } \pi_k = \sqrt{R^2_f/K(1-R^2_f)} \text{ for all } k$$

$$\text{M-}m\text{: } \pi_k = \begin{cases} \sqrt{R^2_f/K(1-R^2_f)} & \text{for } k \leq \lfloor \frac{m-1}{4} K \rfloor \\ const_m(K)\left(1 - \frac{k - \lfloor ((m-1)/4)K \rfloor}{K - \lfloor ((m-1)/4)K \rfloor + 1}\right)^4 & \text{for } k > \lfloor \frac{m-1}{4} K \rfloor \end{cases}$$

$$\text{M-5: } \pi_k = const_5(K)\left(1 - \frac{k}{K+1}\right)^4 \text{ for all } k$$

for $m = 2, 3, 4$, where $\lfloor b \rfloor$ denotes the largest integer that does not exceed $b$ and $R_f^2$ is the first-stage $R^2$. The constants $const_m(K)$ and $const_5(K)$ are chosen such that $\pi'\pi = R_f^2/(1-R_f^2)$. We consider the weak IV case by letting $R_f^2 = 0.005$ that yields the concentration parameter as 1, where the concentration parameter is defined as $n\pi'\pi$ in this setup (e.g., Hahn et al. (2004)). M-1 assumes equal strength over the instruments, whereas M-5 assumes the strength of the instruments varies and it decreases gradually; the other models consider convex combinations of these two cases.

Tables 1 and 2 compare the following estimators: the standard 2SLS estimator using all the available $K$ instruments ("`2SLS`"), our averaged 2SLS estimator with $r = 2$ ("`AvgIV`"), the 2SLS estimator using the optimal number of instruments by Donald and Newey (2001) ("`DN`"), and the model-averaged 2SLS estimator using the optimal weights by Kuersteiner and Okui (2010) with $\underline{w} \in \Omega_U = \{\underline{w} : \sum_{t=1}^T w_t = 1\}$ ("`KO`"). For each estimator, we report the median bias ("`Bias`"), the range between the 0.1 and 0.9 quantiles of distribution ("`QR`"), and the median absolute deviation ("`MAD`"). In general, we can observe that `AvgIV` yields better finite sample performance than `2SLS`. In addition, particularly when the levels of individual IV's strength are similar with each other and hence it is difficult to identify strong ones (e.g., M-1 and M-2), we can find that `AvgIV` outperforms `DN` or `KO` especially in improving the precision of the estimator.[6]

**Remark** In practice, we could obtain $(r, \underline{w})$ by minimizing a uniformly consistent estimator of $S(r, \underline{w})$, but we do not pursue it here. In the simulation above, though this procedure would not guarantee the optimal choice, we fix $r$ arbitrarily small (2 in this case) and choose $w_t$ such that it reflects the level of strength of $Z_t$ and hence the conditions in Assumptions 3 and 4 are satisfied. In particular, we let $w_t(r) = R_{f,t}^2 1\{R_{f,t}^2 > \tau(r)\} / \sum_{s=1}^T R_{f,s}^2 1\{R_{f,s}^2 > \tau(r)\}$ for some trimming parameter $0 < \tau(r) < 1$, where $R_{f,t}^2$ is the $R^2$ of the $t$-th first-stage regression. $\tau(r)$ represents a tolerance level of the IV weakness. $w_t(r)$ is small if $Z_t$ is weak; when $Z_t$ is very weak and hence the first-stage $R^2$ is below the threshold $\tau(r)$, then $w_t(r) = 0$ to trim it

[6] `DN` yields the smallest bias in general in this simiulation, which is mainly because it picks one IV as its optimal number most of the cases (more than 60%).

Table 1: Simulation Result with $(n, K, r) = (200, 16, 2)$ and $R_f^2 = 0.005$

| | | 2SLS | | AvgIV | | DN | | KO | |
|---|---|---|---|---|---|---|---|---|---|
| K=16 | $c$ | 0.5 | 0.8 | 0.5 | 0.8 | 0.5 | 0.8 | 0.5 | 0.8 |
| M-1 | Bias | 0.4672 | 0.7626 | 0.4700 | 0.7577 | 0.4815 | 0.7463 | 0.4497 | 0.7595 |
| | QR | 0.5823 | 0.4285 | 0.5771 | 0.4330 | 2.8027 | 2.8589 | 0.6161 | 0.4693 |
| | MAD | 0.4672 | 0.7626 | 0.4700 | 0.7577 | 0.6946 | 0.8198 | 0.4497 | 0.7595 |
| M-2 | Bias | 0.4696 | 0.7457 | 0.4644 | 0.7484 | 0.4021 | 0.7119 | 0.4654 | 0.7413 |
| | QR | 0.5861 | 0.4483 | 0.5913 | 0.4384 | 3.6827 | 2.7784 | 0.6390 | 0.4253 |
| | MAD | 0.4696 | 0.7457 | 0.4644 | 0.7484 | 0.6954 | 0.8055 | 0.4654 | 0.7413 |
| M-3 | Bias | 0.4657 | 0.7464 | 0.4630 | 0.7463 | 0.3998 | 0.7046 | 0.4637 | 0.7381 |
| | QR | 0.5916 | 0.4400 | 0.5912 | 0.4339 | 3.6633 | 2.7112 | 0.6437 | 0.4273 |
| | MAD | 0.4657 | 0.7464 | 0.4630 | 0.7463 | 0.6856 | 0.8027 | 0.4637 | 0.7381 |
| M-4 | Bias | 0.4665 | 0.7510 | 0.4669 | 0.7455 | 0.3981 | 0.7063 | 0.4666 | 0.7429 |
| | QR | 0.5867 | 0.4351 | 0.5856 | 0.4358 | 3.7326 | 2.7486 | 0.6412 | 0.4274 |
| | MAD | 0.4665 | 0.7510 | 0.4669 | 0.7455 | 0.6896 | 0.8044 | 0.4666 | 0.7429 |
| M-5 | Bias | 0.4674 | 0.7508 | 0.4651 | 0.7454 | 0.4024 | 0.7127 | 0.4638 | 0.7436 |
| | QR | 0.5856 | 0.4359 | 0.5840 | 0.4317 | 3.6238 | 2.8085 | 0.6350 | 0.4196 |
| | MAD | 0.4674 | 0.7508 | 0.4651 | 0.7454 | 0.6943 | 0.8106 | 0.4638 | 0.7436 |

out.[7]

# 6 Concluding Remarks

When a large number of instruments are used, the finite sample bias of the IV estimator is always a concern. If we have clear information about the instruments, we could choose a small number of instruments among them. But the finite sample properties of the IV estimator are sensitive to the choice, and a smaller number of instruments could yield efficiency loss. This paper suggests a simple solution how to utilize a large number of potentially-weak instruments when we do not impose any ad hoc information on them. We use subsets of instruments to obtain preliminary IV estimators and average them over all the subset choices; the subsets

[7] Since $F_t = (R_{f,t}^2/r)/((1 - R_{f,t}^2)/(n - r))$, where $F_t$ and $R_{f,t}^2$ is the first-stage $F$-statistic and the $R^2$, respectively, we could obtain $\tau(r)$ as $\tau(r) = r\overline{F}/(n + r(\overline{F} - 1))$ from $F_t > \overline{F}$, where $\overline{F}$ is some preset threshold value of a weak set of instruments in terms of the first-stage $F$ statistic. Furthermore, we can always find $L$ that satisfies Assumption 4 such that $\sum_{t=1}^{L} w_t(r) = \sum_{t=1}^{L} R_{f,t}^2 1\{R_{f,t}^2 < \tau(r)\} / \sum_{s=1}^{T} R_{f,s}^2 1\{R_{f,s}^2 < \tau(r)\} = \sum_{t=1}^{L} X'P_tX \cdot 1\{X'P_tX < \tau(r)X'X\} / \sum_{s=1}^{T} X'P_sX \cdot 1\{X'P_sX < \tau(r)X'X\} \to_p 0$ by adjusting $\tau(r)$ properly.

Table 2: Simulation Result with $(n, K, r) = (200, 32, 2)$ and $R_f^2 = 0.005$

| | | 2SLS | | AvgIV | | DN | | KO | |
|---|---|---|---|---|---|---|---|---|---|
| K=32 | $c$ | 0.5 | 0.8 | 0.5 | 0.8 | 0.5 | 0.8 | 0.5 | 0.8 |
| M-1 | Bias | 0.4965 | 0.7557 | 0.5016 | 0.7539 | 0.4564 | 0.7194 | 0.5084 | 0.7515 |
| | QR | 0.4476 | 0.3045 | 0.4461 | 0.3000 | 3.9679 | 2.2448 | 0.4626 | 0.3258 |
| | MAD | 0.4965 | 0.7557 | 0.5016 | 0.7539 | 0.7432 | 0.8138 | 0.5084 | 0.7515 |
| M-2 | Bias | 0.4792 | 0.7725 | 0.4771 | 0.7699 | 0.4420 | 0.7768 | 0.4704 | 0.7701 |
| | QR | 0.4248 | 0.2905 | 0.4159 | 0.2938 | 2.5570 | 2.0494 | 0.4482 | 0.3079 |
| | MAD | 0.4792 | 0.7725 | 0.4771 | 0.7699 | 0.5837 | 0.8829 | 0.4704 | 0.7701 |
| M-3 | Bias | 0.4791 | 0.7717 | 0.4779 | 0.7703 | 0.4421 | 0.7725 | 0.4694 | 0.7689 |
| | QR | 0.4253 | 0.2904 | 0.4151 | 0.2914 | 2.4868 | 2.0497 | 0.4492 | 0.3081 |
| | MAD | 0.4791 | 0.7717 | 0.4779 | 0.7703 | 0.6109 | 0.8818 | 0.4694 | 0.7689 |
| M-4 | Bias | 0.4796 | 0.7721 | 0.4782 | 0.7703 | 0.4413 | 0.7733 | 0.4707 | 0.7700 |
| | QR | 0.4248 | 0.2903 | 0.4142 | 0.2914 | 2.4356 | 2.0381 | 0.4480 | 0.3069 |
| | MAD | 0.4796 | 0.7721 | 0.4782 | 0.7703 | 0.5722 | 0.8786 | 0.4707 | 0.7700 |
| M-5 | Bias | 0.4799 | 0.7714 | 0.4784 | 0.7703 | 0.4424 | 0.7772 | 0.4713 | 0.7706 |
| | QR | 0.4237 | 0.2903 | 0.4141 | 0.2910 | 2.5439 | 2.0413 | 0.4474 | 0.3086 |
| | MAD | 0.4799 | 0.7714 | 0.4784 | 0.7703 | 0.5830 | 0.8818 | 0.4713 | 0.7706 |

can be overlapped.

This paper, though it provides some empirical guidance, does not fully answer to the following question: how to choose the size of sub-instrument set and the optimal weight for averaging simultaneously. For instance, choice of the optimal $(r, \underline{w})$ by minimizing uniformly consistent estimators of $S(r, \underline{w})$ in Theorem 4 can be considered. However, unlike Donald and Newey (2001) or Kuersteiner and Okui (2010), joint selection of $r$ and $\underline{w}$ is quite challenging in this framework; it also requires high-level assumptions yielding uniform consistency of the estimator for $S(r, \underline{w})$. Even when $r$ is given, any convex envelop of $\underline{w}$ can be a solution and hence we cannot find the unique optimal solution for $\underline{w}$ unless we obtain higher-order approximation of $S(r, \underline{w})$. It also needs to see if these choice are asymptotically optimal (i.e., $\widehat{S}(\widehat{r}, \widehat{\underline{w}}) / \inf_{r, \underline{w}} S(r, \underline{w}) \to_p 1$). We leave this problem for future research.

Unlike recent studies on the moment condition selection (e.g., Andrews (1999); Caner et al. (2015)), which focus on choosing the valid ones among many moment conditions, we assume that all the instruments are valid in this paper. A natural direction for future research includes studying if the idea of averaging can be used to deal with such misspecification problems,

in particular when some instruments violate exogeneity. We expect that our averaging idea could weaken the bias problem from invalid instruments if we formulate the weight as an inversely proportional function to the IV invalidity. For example, we can define $w_t$ using the ratio of the first-stage $F$ statistic (or $R^2$) over the Sargan statistic (or $J$ statistic) for each subset of instruments $Z_t$.

# A Appendix: Mathematical Proofs

## A.1 Useful lemmas

Throughout the Appendix, we let $||A|| = \sqrt{tr(A'A)}$ be the Euclidean norm for a matrix $A$. For notational simplicity, we write $\widehat{\beta}_{ak} = \sum_{t=1}^{T} w_t \widehat{\beta}_{k,t}$ as the general form of the averaged $k$-class estimator in this Appendix instead of (6) and (7).

The first lemma is a modified version of Lemma A.1 in Donald and Newey (2001) so that it can give the Nagar-type asymptotic expansion of the MSE of the averaged IV estimators. For an averaged $k$-class estimator $\widehat{\beta}_{ak}$, we write

$$\sqrt{\lambda}(\widehat{\beta}_{ak} - \beta) = \sum_{t=1}^{T} w_t \widehat{H}_t^{-1} \widehat{h}_t, \tag{A.1}$$

where $\widehat{H}_t = (X'P_tX - \widehat{\kappa}_t X'X)/\lambda$ is nonsingular and $\widehat{h}_t = (X'P_t\varepsilon - \widehat{\kappa}_t X'\varepsilon)/\sqrt{\lambda}$. Though we only consider the scalar $\beta$ in this paper, the first lemma is general enough to consider the vector case of $\beta$.

**Lemma A.1** *For each $t = 1, \cdots, T$, we assume that there exist decompositions $\widehat{h}_t = h + \Phi_t^h + \Gamma_t^h$, $\widehat{H}_t = H + \Phi_t^H + \Gamma_t^H$ and*

$$(h + \Phi_t^h)(h + \Phi_s^h)' - hh'H^{-1}\Phi_s^{H\prime} - \Phi_t^H H^{-1} hh' = \widehat{A}_{ts}(r) + \Gamma_{ts}^A(r)$$

*that satisfy the following conditions: $h = O_p(1)$; $H$ is symmetric with $H = O_p(1)$ and $\det(H)$ is bounded away from zero w.p.a.1;*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \widehat{A}_{ts}(r) \middle| Z\right] = \sigma_\varepsilon^2 H + HS(r, \underline{w})H + o_p(\rho_{r,\lambda}), \tag{A.2}$$

*where $\rho_{r,\lambda} = tr(S(r, \underline{w}))$ with $\rho_{r,\lambda} = o_p(1)$ and the weights $w_t$ satisfy Assumption 4; there exist some sequence of positive real numbers $\{\rho_t^H\}$ and $\{\rho_t^h\}$ such that $||\Phi_t^H|| = O_p(\rho_t^H)$ and $||\Phi_t^h|| = O_p(\rho_t^h)$ satisfying $\sum_{t=1}^{T} w_t \rho_t^h \rho_t^H = o(\rho_{r,\lambda})$, $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \rho_t^h \rho_s^h \rho_t^H = o(\rho_{r,\lambda})$, and $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s (1 + \rho_t^h)(1 + \rho_s^h)\left(\rho_t^H\right)^{\ell_1}\left(\rho_s^H\right)^{\ell_2} = o(\rho_{r,\lambda})$ for $\ell_1, \ell_2 = 1, 2$; $||\Gamma_t^h|| = o_p\left(\rho_{r,\lambda}\right)$; $||\Gamma_t^H|| = o_p\left(\rho_{r,\lambda}\right)$; and $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s ||\Gamma_{ts}^A(r)|| = o_p(\rho_{r,\lambda})$. Then, it holds that $\lambda(\widehat{\beta}_{ak} - \beta)^2 = \widehat{Q}(r, \underline{w}) + \widehat{\gamma}(r, \underline{w})$, $\mathbb{E}[\widehat{Q}(r, \underline{w})|Z] = \sigma_\varepsilon^2 H^{-1} + S(r, \underline{w}) + R(r, \underline{w})$ and $[\widehat{\gamma}(r, \underline{w}) + R(r, \underline{w})]/tr(S(r, \underline{w})) = o_p(1)$ as $n, \lambda, K, r \to \infty$.*

**Proof of Lemma A.1** We first note that

$$\begin{aligned}\widehat{H}_t^{-1}\widehat{h}_t &= H^{-1}\widehat{h}_t - H^{-1}(\widehat{H}_t - H)H^{-1}\widehat{h}_t + H^{-1}(\widehat{H}_t - H)\widehat{H}_t^{-1}(\widehat{H}_t - H)H^{-1}\widehat{h}_t \\ &= H^{-1}\widetilde{h}_t - H^{-1}\Phi_t^H H^{-1}\widetilde{h}_t + H^{-1}\Phi_t^H \widehat{H}_t^{-1}\Phi_t^H H^{-1}\widetilde{h}_t + o_p(\rho_{r,\lambda}) \\ &\equiv \Psi_{A,t} + \Psi_{B,t} + o_p(\rho_{r,\lambda})\end{aligned}$$

for each $t$, where $\widehat{h}_t = \widetilde{h}_t + o_p\left(\rho_{r,\lambda}\right)$ by denoting $\widetilde{h}_t = h + \Phi_t^h$ since $\rho_{r,\lambda} \to 0$. Then the averaged $k$-class estimator (A.1) can be written as

$$\begin{aligned}\lambda(\widehat{\beta}_{ak} - \beta)^2 &= \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \widehat{H}_t^{-1}\widehat{h}_t\widehat{h}_s'\left(\widehat{H}_s^{-1}\right)' \\ &= \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s H^{-1}(B_{ts,1}(r) + B_{ts,2}(r) + B_{ts,3}(r))H^{-1} + o_p\left(\rho_{r,\lambda}\right),\end{aligned}$$

where

$$B_{ts,1}(r) = \widetilde{h}_t\widetilde{h}_s' - hh'H^{-1}\Phi_s^{H\prime} - \Phi_t^H H^{-1}hh' \quad \text{(A.3)}$$

$$B_{ts,2}(r) = -(\Phi_t^h h' + h\Phi_s^{h\prime} + \Phi_t^h\Phi_s^{h\prime})H^{-1}\Phi_s^{H\prime} - \Phi_t^H H^{-1}(\Phi_t^h h' + h\Phi_s^{h\prime} + \Phi_t^h\Phi_s^{h\prime}) \quad \text{(A.4)}$$

which collect terms from $\Psi_{A,t}\Psi_{A,s}'$ and $(\Psi_{A,t}\Psi_{B,s}' + \Psi_{B,t}\Psi_{A,s}')$, respectively. The rest terms are collected in

$$\begin{aligned}B_{ts,3}(r) &= \Phi_t^H H^{-1}\widetilde{h}_t\widetilde{h}_s' H^{-1}\Phi_s^{H\prime} + \Phi_t^H\widehat{H}_t^{-1}\Phi_t^H H^{-1}\widetilde{h}_t\widetilde{h}_s' + \widetilde{h}_t\widetilde{h}_s' H^{-1}\Phi_s^{H\prime}\widehat{H}_t^{-1}\Phi_s^{H\prime} \quad \text{(A.5)} \\ &\quad -\Phi_t^H H^{-1}\widetilde{h}_t\widetilde{h}_s' H^{-1}\Phi_s^{H\prime}\widehat{H}_t^{-1}\Phi_s^{H\prime} - \Phi_t^H\widehat{H}_t^{-1}\Phi_t^H H^{-1}\widetilde{h}_t\widetilde{h}_s' H^{-1}\Phi_s^{H\prime} \\ &\quad +\Phi_t^H\widehat{H}_t^{-1}\Phi_t^H H^{-1}\widetilde{h}_t\widetilde{h}_s' H^{-1}\Phi_s^{H\prime}\widehat{H}_t^{-1}\Phi_s^{H\prime}.\end{aligned}$$

From the assumptions, since $h = O_p(1)$, $H^{-1} = O_p(1)$ and $\widehat{H}_t^{-1} = O_p(1)$, it can be verified that $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s||B_{ts,2}(r)|| = o_p(\rho_{r,\lambda}) \to 0$ and $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s||B_{ts,3}(r)|| = o_p(\rho_{r,\lambda}) \to 0$ as $n, K, r, \lambda \to \infty$. If we let $B_{ts,1}(r) = \widehat{A}_{ts}(r) + \Gamma_{ts}^A(r)$ with $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s||\Gamma_{ts}^A(r)|| = o_p(\rho_{r,\lambda})$ and $\mathbb{E}[\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \widehat{A}_{ts}(r)\,|Z] = \sigma_\varepsilon^2 H + HS(r,\underline{w})H + o_p(\rho_{r,\lambda})$, then

$$\begin{aligned}\lambda(\widehat{\beta}_{ak} - \beta)^2 &= \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s H^{-1}\widehat{A}_{ts}(r)H^{-1} \\ &\quad + \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s H^{-1}(\Gamma_{ts}^A(r) + B_{ts,2}(r) + B_{ts,3}(r))H^{-1} \equiv \widehat{Q}(r,\underline{w}) + \widehat{\gamma}(r,\underline{w}),\end{aligned}$$

where $\widehat{\gamma}(r,\underline{w}) = o_p(\rho_{r,\lambda})$ and $\mathbb{E}[\widehat{Q}(r,\underline{w})|Z] = \sigma_\varepsilon^2 H^{-1} + S(r,\underline{w}) + R(r,\underline{w})$ satisfying $[\widehat{\gamma}(r,\underline{w}) + R(r,\underline{w})]/tr(S(r,\underline{w})) = o_p(1)$ as $n, \lambda, K, r \to \infty$ similarly as Donald and Newey (2001). ■

The following lemmas are useful to prove the main theorems. We denote $P_{t,ij}$ as the $(i,j)$-th element of the projection matrix $P_t$.

**Lemma A.2** *Suppose Assumptions 1-3 hold. For any $t$ and $s$, we have (i) $tr(P_t) = r$; (ii) $\sum_{i=1}^n P_{t,ii}P_{s,ii} = o_p(r)$; (iii) $\sum_{i\neq j} P_{t,ii}P_{s,jj} = r^2 + o_p(r)$; (iv) $\sum_{i\neq j} P_{t,ij}P_{s,ij} = \sum_{i\neq j} P_{t,ij}P_{s,ji} = rank(P_{t\cap s}) + o_p(r) = O_p(r)$, where $P_{t\cap s}$ is the projection matrix of the instruments that are included both in $Z_t$ and $Z_s$; (v) $\sum_{i\neq j} P_{t,ij} = o_p(r)$.*

**Proof of Lemma A.2** (i) and (ii) are straightforward when $t = s$ (e.g., Lemma A.2 in Donald and Newey (2001)). When $t \neq s$, (ii) holds since $\sum_{i=1}^n P_{t,ii}P_{s,ii} \leq \max_{1\leq i\leq n} P_{t,ii} \times \sum_{i=1}^n P_{s,ii} = \max_{1\leq i\leq n} P_{t,ii} \times tr(P_t) = o_p(1)r = o_p(r)$ from Assumption 1-(iv). For (iii), $\sum_{i\neq j} P_{t,ii}P_{s,jj} = (\sum_{i=1}^n P_{t,ii})^2 - \sum_{i=1}^n P_{t,ii}P_{s,ii} = [tr(P_t)]^2 + o_p(r) = r^2 + o_p(r)$ from part (ii). For (iv), if $t = s$, since $P_t$ is symmetric and idempotent, $\sum_{i\neq j} P_{t,ij}P_{t,ij} = tr(P_t'P_t) - \sum_{i=1}^n P_{t,ii}^2 = tr(P_t) + o_p(r) = r + o_p(r)$ from part (ii); if $t \neq s$, $\sum_{i\neq j} P_{t,ij}P_{s,ij} = tr(P_t'P_s) - \sum_{i=1}^n P_{t,ii}P_{s,ii} = tr(P_{t\cap s}) + o_p(r)$ from part (ii) since $tr(P_t'P_s) = tr(P_{t\cap s}) = rank(P_{t\cap s}) < r$. Note that $P_{t\cap t} = P_t$ and the rank of $P_{t\cap s}$ is at most $r-1$ for the case of $t \neq s$. The same results hold for $\sum_{i\neq j} P_{t,ij}P_{s,ji}$ since $\sum_{i\neq j} P_{t,ij}P_{s,ji} = tr(P_tP_s) - \sum_{i=1}^n P_{t,ii}P_{s,ii} = tr(P_t'P_s) - \sum_{i=1}^n P_{t,ii}P_{s,ii}$ as $P_t$ being symmetric. Finally, for part (v), note that $\sum_{i\neq j} P_{t,ij} = \sum_{i=1}^n \sum_{j=1}^n P_{t,ij} - \sum_{i=1}^n P_{t,ii} = \iota_n'P_t\iota_n - tr(P_t) = \iota_n'P_t'P_t\iota_n - r$ since $P_t$ is symmetric and idempotent, where $\iota_n$ is the $n \times 1$ vector of ones. However, $\iota_n'P_t'P_t\iota_n = 2\sum_{i\neq j} P_{t,ii}P_{t,ij} + \sum_{i\neq j} P_{t,ij}^2 + \sum_{i=1}^n P_{t,ii}^2 = 2\sum_{i\neq j} P_{t,ii}P_{t,ij} + r + o_p(r)$ from parts (ii) and (iv), and thus

$$\sum_{i\neq j} P_{t,ij} = 2\sum_{i\neq j} P_{t,ii}P_{t,ij} + o_p(r) \leq \max_{1\leq i\leq n} P_{t,ii} \times 2\sum_{i\neq j} P_{t,ij} + o_p(r) = o_p(1) \times 2\sum_{i\neq j} P_{t,ij} + o_p(r),$$

or $\sum_{i\neq j} P_{t,ij}/(2\sum_{i\neq j} P_{t,ij} + r) \to_p 0$, which implies that $\sum_{i\neq j} P_{s,ij}$ is at most $o_p(r)$. ■

We now let $H = \pi'Z'Z\pi/\lambda$ and $h = \pi'Z'\varepsilon/\sqrt{\lambda}$ for the rest of the proofs, where $H$ is nonsingluar.

**Lemma A.3** *Suppose Assumptions 1-5 hold. For any $t$ and $s$, (i) $H = O_p(1)$ and $h = O_p(1)$; (ii) $\pi'Z'(I-P_t)u/\sqrt{\lambda} = O_p(r^{-\delta_t})$ and $\pi'Z'(I-P_t)\varepsilon/\sqrt{\lambda} = O_p(r^{-\delta_t})$; (iii) $\mathbb{E}[u'P_t\varepsilon|Z] = O_p(r)$ and $\mathbb{E}[\varepsilon'P_t\varepsilon|Z] = O_p(r)$; (iv) $\mathbb{E}[u'P_t\varepsilon\varepsilon'P_su|Z] = \sigma_{\varepsilon u}^2 r^2 + (\sigma_\varepsilon^2\sigma_u^2 + \sigma_{\varepsilon u}^2)rank(P_{t\cap s}) + o_p(r) = \sigma_{\varepsilon u}^2 r^2 + o_p(r^2)$; (v) $\mathbb{E}[\pi'Z'P_t\varepsilon\varepsilon'P_su|Z] = o_p(r)$; (vi) $\mathbb{E}[\pi'Z'\varepsilon\varepsilon'Z\pi H^{-1}u'Z\pi/\lambda^2|Z] = O_p(1/\lambda)$; (vii) $\mathbb{E}[\varepsilon'P_t\varepsilon\varepsilon'Z\pi|Z] = O_p(r)$; (viii) $\mathbb{E}[\varepsilon'P_t\varepsilon\varepsilon'\varepsilon|Z] = nr\sigma_\varepsilon^4 + O_p(r)$, $\mathbb{E}[\varepsilon'P_t\varepsilon u'u|Z] = nr\sigma_\varepsilon^2\sigma_u^2 + O_p(r)$, and $\mathbb{E}[\varepsilon'P_t\varepsilon\varepsilon'v|Z] = O_p(r)$, where $v_i = u_i - \varepsilon_i\sigma_{u\varepsilon}/\sigma_\varepsilon^2$; (ix) $\mathbb{E}[\varepsilon'P_t\varepsilon - \widetilde{\kappa}_t\varepsilon'\varepsilon|Z] = O_p(r/n)$, where $\widetilde{\kappa}_t = (\varepsilon'P_t\varepsilon)/(n\sigma_\varepsilon^2)$.*

**Proof of Lemma A.3** Part (i) follows immediately since $\pi'Z'Z\pi/\lambda = O_p(1)$ is assumed in Assumption 2-(iii) and $\mathbb{E}[||h||^2|Z] = \sigma_\varepsilon^2(\pi'Z'Z\pi/\lambda)$. For part (ii), using $\pi'Z'(I-P_t) = \pi_{-t}'Z_{-t}'$, the result follows since $\mathbb{E}[||\pi'Z'(I-P_t)u/\sqrt{\lambda}||^2|Z] = \mathbb{E}[\pi_{-t}'Z_{-t}'uu'Z_{-t}\pi_{-t}/\lambda|Z] = \sigma_u^2(\pi_{-t}'Z_{-t}'Z_{-t}\pi_{-t}/\lambda) = O_p(r^{-2\delta_t})$ from Assumption 3-(ii) and $\sigma_u^2 < \infty$. The same derivation applies to $\pi'Z'(I-P_t)\varepsilon/\sqrt{\lambda}$ for $\sigma_\varepsilon^2 < \infty$. Part (iii) is from Lemma A.3 in Donald and Newey

(2001) and Lemma A.2 above. For (iv), if we collect non-zero terms, we have

$$\begin{aligned}\mathbb{E}[u'P_t\varepsilon\varepsilon'P_su|Z] &= \sum_{i=1}^n \mathbb{E}[u_i^2\varepsilon_i^2|z_i]P_{t,ii}P_{s,ii} + \sum_{i\neq j}\mathbb{E}[u_i^2|z_i]\mathbb{E}[\varepsilon_j^2|z_j]P_{t,ij}P_{s,ij} \\ &\quad + \sum_{i\neq j}\mathbb{E}[u_i\varepsilon_i|z_i]\mathbb{E}[u_j\varepsilon_j|z_j]\left(P_{t,ii}P_{s,jj} + P_{t,ij}P_{s,ji}\right) \\ &= o_p(r) + \sigma_\varepsilon^2\sigma_u^2 rank(P_{t\cap s}) + \sigma_{\varepsilon u}^2\left(r^2 + rank(P_{t\cap s})\right)\end{aligned}$$

using Lemma A.2 since $(\varepsilon_i, u_i)'$ are i.i.d. and $\mathbb{E}[u_i^2\varepsilon_i^2|z_i] < \infty$ from Assumption 1-(ii). Note that $rank(P_{t\cap s}) \leq r$. Similarly, part (v) is obtained from

$$\begin{aligned}\mathbb{E}\left[\pi'Z'P_t\varepsilon\varepsilon'P_su|Z\right] &= \sum_{i=1}^n \pi'z_iP_{t,ii}P_{s,ii}\mathbb{E}[\varepsilon_i^2u_i|z_i] + \sum_{i\neq j}^n \pi'z_iP_{t,ij}P_{s,jj}\mathbb{E}[\varepsilon_j^2u_j|z_j] \\ &\leq \sup_{1\leq i\leq n}\left|\pi'z_i\right|\left\{\sum_{i=1}^n P_{t,ii}P_{s,ii}\mathbb{E}[\varepsilon_i^2u_i|z_i] + \max_{1\leq j\leq n}P_{s,jj}\cdot\sum_{i\neq j}^n P_{t,ij}\mathbb{E}[\varepsilon_j^2u_j|z_j]\right\}\end{aligned}$$

with probability approaching to one (w.p.a.1, hereafter), where $\mathbb{E}[\varepsilon_i^2u_i|z_i] < \infty$ from Assumption 1-(ii), $\max_{1\leq j\leq n}P_{s,jj} = o_p(1)$ from Assumption 1-(iv), $\sup_{1\leq i\leq n}|\pi'z_i| = O_p(1)$ from Assumption 2-(iv), and by applying Lemma A.2-(ii) and (v). For part (vi), we note that, w.p.a.1,

$$\begin{aligned}\mathbb{E}\left[\left.\frac{\pi'Z'\varepsilon\varepsilon'Z\pi H^{-1}u'Z\pi}{\lambda^2}\right|Z\right] &= \frac{1}{\lambda^2}\sum_{i=1}^n(\pi'z_i)^2H^{-1}\mathbb{E}[\varepsilon_i^2u_i|z_i](\pi'z_i) \\ &\leq \sup_{1\leq i\leq n}\left|\pi'z_i\right|\frac{\pi'Z'Z\pi}{\lambda}H^{-1}\mathbb{E}[\varepsilon_i^2u_i|z_i]\times\frac{1}{\lambda} = O_p\left(\frac{1}{\lambda}\right)\end{aligned}$$

using a similar argument as part (v). For part (vii), if we collect non-zero terms, we have w.p.a.1

$$\mathbb{E}[\varepsilon'P_t\varepsilon\varepsilon'Z\pi|Z] = \sum_{i=1}^n\pi'z_iP_{t,ii}\mathbb{E}[\varepsilon_i^3|z_i] \leq \sup_{1\leq i\leq n}\left|\pi'z_i\right|\sum_{i=1}^n P_{t,ii}\mathbb{E}[\varepsilon_i^3|z_i] = O_p(r)$$

from Lemma A.2-(i) and (v) since $\mathbb{E}[\varepsilon_i^3|z_i]$ is bounded. Part (viii) can be obtained similarly from

$$\begin{aligned}\mathbb{E}[\varepsilon'P_t\varepsilon\varepsilon'\varepsilon|Z] &= \sum_{i=1}^n P_{t,ii}\mathbb{E}[\varepsilon_i^4|z_i] + \sum_{i\neq j}\left(P_{t,ii}+2P_{t,ij}\right)\sigma_\varepsilon^4 = r\mathbb{E}[\varepsilon_i^4|z_i] + nr\sigma_\varepsilon^4 + o_p(r) \\ \mathbb{E}[\varepsilon'P_t\varepsilon u'u|Z] &= \sum_{i=1}^n P_{t,ii}\mathbb{E}[\varepsilon_i^2u_i^2|z_i] + \sum_{i\neq j}\left(P_{t,ii}+2P_{t,ij}\right)\sigma_\varepsilon^2\sigma_u^2 = r\mathbb{E}[\varepsilon_i^2u_i^2|z_i] + nr\sigma_\varepsilon^2\sigma_u^2 + o_p(r)\end{aligned}$$

and

$$\mathbb{E}[\varepsilon' P_t \varepsilon \varepsilon' v | Z] = \sum_{i=1}^{n} P_{t,ii} \mathbb{E}[\varepsilon_i^3 v_i | z_i] = O_p(r)$$

since $\mathbb{E}[v_i \varepsilon_i | z_i] = 0$ by construction and $\mathbb{E}[\varepsilon_i^3 v_i | z_i] = \mathbb{E}[\varepsilon_i^3 u_i | z_i] - \mathbb{E}[\varepsilon_i^4 | z_i] \times (\sigma_{\varepsilon u} / \sigma_\varepsilon^2)$ is bounded from Assumption 1-(ii). Finally, part (ix) can be shown as

$$\mathbb{E}[\varepsilon' P_t \varepsilon - \widetilde{\kappa}_t \varepsilon' \varepsilon | Z] = \mathbb{E}[\varepsilon' P_t \varepsilon - (\varepsilon' P_t \varepsilon \varepsilon' \varepsilon / n \sigma_\varepsilon^2) | Z] = r\sigma_\varepsilon^2 - r\sigma_\varepsilon^2 + O_p(r/n)$$

from part (viii). ■

The following lemma shows that the quadratic form of $\sum_{t=1}^{T} w_t Z_{-t} \pi_{-t}$ is $o_p(\lambda)$. Recall that we define

$$\Delta(r, \underline{w}) = \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \left( \frac{\pi' Z' (I - P_t)(I - P_s) Z \pi}{\lambda} \right) = \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \left( \frac{\pi'_{-t} Z'_{-t} Z_{-s} \pi_{-s}}{\lambda} \right).$$

**Lemma A.4** *If Assumptions 1-4 are satisfied, then (i)* $\sum_{t=1}^{T} w_t r^{-2\delta_t} \to 0$ *and (ii) for given* $r$ *and* $\underline{w} = (w_1, \cdots, w_T)'$*, we have* $\Delta(r, \underline{w}) = O_p(\sum_{t=1}^{T} w_t r^{-2\delta_t}) = o_p(1)$*.*

**Proof of Lemma A.4** For part (i), first note that

$$\sum_{t=1}^{T} w_t r^{-2\delta_t} = \sum_{t=1}^{L} w_t r^{-2\delta_t} + \sum_{t=L+1}^{T} w_t r^{-2\delta_t} \leq \sup_t r^{-2\delta_t} \times \sum_{t=1}^{L} w_t + \sup_t w_t \times \sum_{t=L+1}^{T} r^{-2\delta_t},$$

where $L = L(T)$ is defined as in Assumption 4. Then the first term of the last inequality is $o(1)$ since $\sup_t r^{-2\delta_t} \leq 1$ for $\delta_t \geq 0$ and $r \geq 1$ and $\sum_{t=1}^{L} w_t \to 0$ from Assumption 4. Similarly, the second term is also $o(1)$ since $\sup_t w_t \leq 1$ by construction and $\sum_{t=L+1}^{T} r^{-2\delta_t} \to 0$ since $\sum_{t=1}^{T} r^{-2\delta_t} = O(1)$ from Assumption 3-(ii) and $L \to \infty$ as $T \to \infty$. For part (ii), the Jensen's inequality gives (w.p.a.1)

$$\begin{aligned}
\Delta(r, \underline{w}) &= \frac{1}{\lambda} \left( \sum_{t=1}^{T} w_t Z_{-t} \pi_{-t} \right)' \left( \sum_{s=1}^{T} w_s Z_{-s} \pi_{-s} \right) \leq \frac{1}{\lambda} \sum_{t=1}^{T} w_t \pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t} \\
&= \sum_{t=1}^{T} w_t r^{-2\delta_t} \left( r^{2\delta_t} \frac{\pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t}}{\lambda} \right) \leq \sup_t \left( r^{2\delta_t} \frac{\pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t}}{\lambda} \right) \sum_{t=1}^{T} w_t r^{-2\delta_t} \\
&= O_p \left( \sum_{t=1}^{T} w_t r^{-2\delta_t} \right),
\end{aligned}$$

from Assumption 3-(ii) and Assumption 4, where $\sum_{t=1}^{T} w_t r^{-2\delta_t} \to 0$ as $T \to \infty$ from part (i). ■

**Lemma A.5** *Suppose that Assumptions 1-5 are satisfied,* $\lim_{r,\lambda\to\infty}(r/\lambda) = \varsigma < \infty$ *and* $\sup_{n\geq 1}\sup_{1\leq i\leq n}\sum_{j=1}^{n}|\sum_{t=1}^{T} w_t P_{t,ij}| = O_p(1)$. *Then we have*

$$\lambda^{-1/2}\left(\sum_{t=1}^{T} w_t X'\left(P_t - \left(\frac{r}{n}\right)I\right)\varepsilon\right) \to_d \mathcal{N}\left(0, \sigma_\varepsilon^2\overline{H} + \varsigma\left\{\mathbb{E}(\varepsilon_i^2 u_i^2) - \sigma_{\varepsilon u}^2\right\}\right)$$

*as* $n, r, \lambda \to \infty$.

**Proof of Lemma A.5** The proof goes similarly as Lemma A.1 in Lee and Okui (2012). We let the matrices $U, M, V, C, \Omega$ and $a$ in Theorem 1 of van Hasselt (2010) as $(\varepsilon, X)$, $(0, Z\pi)$, $(\varepsilon, u)$, $\sum_{t=1}^{T} w_t(P_t - (r/n)I)$, $\Sigma$ and $(1,0)'$ in this setup. Then the conditions in Assumption 1-(a) and (b) of van Hasselt (2010) are satisfied from Assumptions 1 and 2. The conditions (c) are also satisfied since we can verify that

$$\begin{aligned}
& \frac{1}{\lambda}\pi' Z'\left[\sum_{t=1}^{T} w_t\left(P_t - \left(\frac{r}{n}\right)I\right)\right]^2 Z\pi \\
= & \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \frac{\pi' Z' P_t P_s Z\pi}{\lambda} - 2\left(\frac{r}{n}\right)\sum_{t=1}^{T} w_t \frac{\pi' Z' P_t Z\pi}{\lambda} + \left(\frac{r}{n}\right)^2 \frac{\pi' Z' Z\pi}{\lambda} \\
= & \left(1 - \frac{r}{n}\right)^2 \frac{\pi' Z' Z\pi}{\lambda} - 2\left(1 - \frac{r}{n}\right)\sum_{t=1}^{T} w_t \frac{\pi'_{-t} Z'_{-t} Z_{-t}\pi_{-t}}{\lambda} + \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \frac{\pi'_{-t} Z'_{-t} Z_{-s}\pi_{-s}}{\lambda} \\
= & \overline{H} + O_p\left(\sum_{t=1}^{T} w_t r^{-2\delta_t}\right) + O_p\left(\Delta(r)\right) \to_p \overline{H} > 0
\end{aligned}$$

as $n, r \to \infty$ (and thus $K, T \to \infty$) from Lemma A.4 and Assumptions 2 and 3. Similarly, letting $d_C$ be the $n \times 1$ vector of the diagonal elements of $\sum_{t=1}^{T} w_t(P_t - (r/n)I)$ and $(z_i^{-t})'$

be the $i$-th row of $Z_{-t}$, we can show that

$$\begin{aligned}
&\frac{1}{\lambda}\pi' Z' \sum_{t=1}^{T} w_t \left(P_t - \left(\frac{r}{n}\right) I\right) d_C \\
= &\sum_{t=1}^{T} w_t \frac{1}{\lambda} \left\{ \left(1 - \frac{r}{n}\right) \pi' Z' - \pi'_{-t} Z'_{-t} \right\} d_C \\
= &\frac{1}{\lambda} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \left\{ \left(1 - \frac{r}{n}\right) \pi' z_i - \pi'_{-t} z_i^{-t} \right\} \left(P_{s,ii} - \frac{r}{n}\right) \\
\leq &\frac{1}{\lambda} \sum_{s=1}^{T} w_s \left(1 - \frac{r}{n}\right) \left| \sum_{i=1}^{n} \pi' z_i \left(P_{s,ii} - \frac{r}{n}\right) \right| + \frac{1}{\lambda} \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \left| \sum_{i=1}^{n} \pi'_{-t} z_i^{-t} \left(P_{s,ii} - \frac{r}{n}\right) \right| \\
\leq &\left(1 - \frac{r}{n}\right) \sum_{s=1}^{T} w_s \left(\frac{\pi' Z' Z \pi}{\lambda}\right)^{1/2} \left(\frac{1}{\lambda} \sum_{i=1}^{n} \left(P_{s,ii} - \frac{r}{n}\right)^2\right)^{1/2} \\
&+ \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \left(\frac{\pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t}}{\lambda}\right)^{1/2} \left(\frac{1}{\lambda} \sum_{i=1}^{n} \left(P_{s,ii} - \frac{r}{n}\right)^2\right)^{1/2} \qquad \text{(A.6)}
\end{aligned}$$

w.p.a.1 from the Cauchy-Schwartz. However, (A.6) is simply $o_p(1)$ since

$$\frac{1}{\lambda} \sum_{i=1}^{n} \left(P_{s,ii} - \frac{r}{n}\right)^2 = \frac{1}{\lambda} \sum_{i=1}^{n} P_{s,ii}^2 - \left(\frac{r}{n}\right)\left(\frac{r}{\lambda}\right) \leq \left(\sup_{1 \leq t \leq T} P_{s,ii}\right) \frac{1}{\lambda} tr\left(P_s\right) + \left(\frac{r}{n}\right)\left(\frac{r}{\lambda}\right)$$

w.p.a.1, where $tr\left(P_s\right) = r$, $r/\lambda \to \varsigma < \infty$, $r/n \to 0$, and $\sup_{1 \leq t \leq T} P_{s,ii} = o_p(1)$. Moreover, $\pi' Z' Z \pi / \lambda \to_p \overline{H} < \infty$ and $\sum_{t=1}^{T} w_t \left[\pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t} / \lambda\right]^{1/2} = O_p(\sum_{t=1}^{T} w_t r^{-\delta_t}) \to 0$ since $0 \leq \sum_{t=1}^{T} w_t r^{-\delta_t} \leq (\sum_{t=1}^{T} w_t r^{-2\delta_t})^{1/2} \to 0$ by the Jensen's inequality and from Lemma A.4-(i). Furthermore,

$$\begin{aligned}
\frac{1}{\lambda} tr\left[\sum_{t=1}^{T} w_t (P_t - \left(\frac{r}{n}\right) I)\right] &= \frac{1}{\lambda} \sum_{t=1}^{T} w_t \left(r - r\right) = 0 \\
\frac{1}{\lambda} tr\left[\left(\sum_{t=1}^{T} w_t (P_t - \left(\frac{r}{n}\right) I)\right)^2\right] &= \frac{1}{\lambda} \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s tr\left[(P_t - \left(\frac{r}{n}\right) I)\left((P_s - \left(\frac{r}{n}\right) I)\right)\right] \\
&= \frac{1}{\lambda} \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \left(tr\left(P_{t \cap s}\right) - 2\frac{r^2}{n} + \frac{r^2}{n}\right) \to \varsigma
\end{aligned}$$

since $tr\left(P_tP_s\right)=tr\left(P_{t\cap s}\right)$ and $\sum_{t=1}^{T}\sum_{s=1}^{T}w_tw_str\left(P_{t\cap s}\right)=O(r)$, and similarly

$$\begin{aligned}\frac{1}{\lambda}d_C'd_C &= \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T}w_tw_s\frac{1}{\lambda}\left(P_{t,ii}-\left(\frac{r}{n}\right)\right)\left(P_{s,ii}-\left(\frac{r}{n}\right)\right)\\ &= \sum_{t=1}^{T}\sum_{s=1}^{T}w_tw_s\frac{1}{\lambda}\left(\sum_{i=1}^{n}P_{t\cap s,ii}-\left(\frac{r}{n}\right)\sum_{i=1}^{n}P_{t,ii}-\left(\frac{r}{n}\right)\sum_{i=1}^{n}P_{s,ii}+\frac{r^2}{n}\right)\\ &= \frac{1}{\lambda}\sum_{t=1}^{T}\sum_{s=1}^{T}w_tw_s\left(tr\left(P_{t\cap s}\right)-2\frac{r^2}{n}+\frac{r^2}{n}\right)\to\varsigma.\end{aligned}$$

Finally, for $c_{ij}$ being the $(i,j)$-th element of $\sum_{t=1}^{T}w_t(P_t-(r/n)I)$, we have

$$\sup_{n\geq1}\sup_{1\leq i\leq n}\sum_{j=1}^{n}|c_{ij}|\leq\sup_{n\geq1}\sup_{1\leq i\leq n}\left(\sum_{j=1}^{n}\left|\sum_{t=1}^{T}w_tP_{t,ij}\right|+\left(\frac{r}{n}\right)\right)=O_p(1)$$

from the condition $\sup_{n\geq1}\sup_{1\leq i\leq n}\sum_{j=1}^{n}\left|\sum_{t=1}^{T}w_tP_{t,ij}\right|=O_p(1)$ since $r/n\to0$. Therefore, the conditions for Theorem 1 of van Hasselt (2010) are all satisfied, which yields

$$\begin{aligned}\frac{1}{\sqrt{\lambda}}\sum_{t=1}^{T}w_tX'\left(P_t-\left(\frac{r}{n}\right)I\right)\varepsilon &= \frac{1}{\sqrt{\lambda}}(0,1)(\varepsilon,X)'\left\{\sum_{t=1}^{T}w_t\left(P_t-\left(\frac{r}{n}\right)I\right)\right\}(\varepsilon,X)(1,0)'\\ &\to_d\mathcal{N}(0,V_{22}),\end{aligned}\tag{A.7}$$

where $V_{22}=\sigma_\varepsilon^2\overline{H}+\varsigma\left\{\mathbb{E}(\varepsilon_i^2u_i^2)-\sigma_{\varepsilon u}^2\right\}$, because it holds that $\mathbb{E}(\sum_{t=1}^{T}w_tX'(P_t-(r/n)I)\varepsilon|Z)=tr[\sum_{t=1}^{T}w_t(P_t-(r/n)I)\mathbb{E}(\varepsilon X'|Z)]=(r-r)\sigma_{\varepsilon u}=0$. ■

**Lemma A.6** *Suppose that Assumptions 1-5 are satisfied and the eighth moments of both $x_i$ and $\varepsilon_i$ exist. Then we have $\widehat{\sigma}_\varepsilon^2\equiv\sum_{i=1}^{n}(y_i-x_i\hat{\beta}_{ak})^2/n\to_p\sigma_\varepsilon^2$ and $\sum_{i=1}^{n}(y_i-x_i\hat{\beta}_{ak})^4/n\to_p\mathbb{E}(\varepsilon_i^4)$ as $n,r,\lambda\to\infty$, where $\hat{\beta}_{ak}$ is either the LIML or bias-corrected 2SLS estimator.*

**Proof of Lemma A.6** The results follow similarly as Lemma A.2 in Lee and Okui (2012) using Lemma A.3 and Theorem 1. Note that we can write

$$\frac{1}{n}(y-X\hat{\beta}_{ak})'(y-X\hat{\beta}_{ak})=\frac{1}{n}(\beta-\hat{\beta}_{ak})'X'X(\beta-\hat{\beta}_{ak})+\frac{2}{n}(\beta-\hat{\beta}_{ak})'X'\varepsilon+\frac{1}{n}\varepsilon'\varepsilon\to_p\sigma_\varepsilon^2.\tag{A.8}$$

■

## A.2 Proofs of the asymptotic results

**Proof of Theorem 1 (Consistency)** We consider the averaged $k$-class estimator defined as

$$\widehat{\beta}_{ak}-\beta=\sum_{t=1}^{T}w_t\left(X'P_tX-\widehat{\kappa}_tX'X\right)^{-1}\left(X'P_t\varepsilon-\widehat{\kappa}_tX'\varepsilon\right).\tag{A.9}$$

Since it is assumed that $\widehat{\kappa}_t - r/n = o_p(r/n)$, similarly as the proof of Theorem 2.1 in Chao and Swanson (2005), we deduce that

$$\left(\widehat{\kappa}_t - \frac{r}{n}\right)\frac{X'X}{\lambda} = o_p\left(\frac{r}{\lambda}\right)\frac{X'X}{n} \to_p 0$$

provided $r/\lambda \to \varsigma < \infty$ and $X'X/n = O_p(1)$ for large $n$. Note that

$$\frac{X'X}{n} = \left(\frac{\lambda}{n}\right)\frac{\pi'Z'Z\pi}{\lambda} + \left(\frac{\lambda}{n}\right)\frac{\pi'Z'u + uZ\pi}{\lambda} + \frac{u'u}{n} \to_p \xi\overline{H} + O_p\left(1/\sqrt{\lambda}\right) + \sigma_u^2,$$

where $\lambda/n \to \xi < \infty$, both $\overline{H}$ and $\sigma_u^2$ are finite, and $\pi'Z'u/\lambda = O_p(1/\sqrt{\lambda})$ from Lemma A1 in Chao and Swanson (2005). We can also show that $\left(\widehat{\kappa}_t - (r/n)\right)(X'\varepsilon/\lambda) \to_p 0$ for each $t$ using a similar argument. It follows that

$$\begin{aligned} & \left(X'P_tX - \widehat{\kappa}_tX'X\right)^{-1}\left(X'P_t\varepsilon - \widehat{\kappa}_tX'\varepsilon\right) \\ = & \left(\frac{X'P_tX}{\lambda} - \left(\frac{r}{n}\right)\frac{X'X}{\lambda}\right)^{-1}\left(\frac{X'P_t\varepsilon}{\lambda} - \left(\frac{r}{n}\right)\frac{X'\varepsilon}{\lambda}\right) + o_p(1) \end{aligned}$$

provided the denominator is $O_p(1)$, which is to be shown below. Since $X'P_tX = X'PX - X'P_{-t}X$, we can show that

$$\begin{aligned} \frac{X'P_tX}{\lambda} - \left(\frac{r}{n}\right)\frac{X'X}{\lambda} &= \left(1-\frac{r}{n}\right)\frac{\pi'Z'Z\pi}{\lambda} - \frac{\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}}{\lambda} \\ &\quad +2\left(1-\frac{r}{n}\right)\frac{\pi'Z'u}{\lambda} - 2\frac{\pi'_{-t}Z'_{-t}u}{\lambda} + \frac{u'P_tu}{\lambda} - \left(\frac{r}{n}\right)\frac{u'u}{\lambda} \\ &= \left(1-\frac{r}{n}\right)\overline{H} + O_p(r^{-2\delta_t}) + \left(1-\frac{r}{n}\right)O_p(1/\sqrt{\lambda}) + O_p(r^{-\delta_t}/\sqrt{\lambda}) \\ &\quad +(r/\lambda)\{\sigma_u^2 + O_p(1/\sqrt{r})\} - (r/n)(n/\lambda)\{\sigma_u^2 + O_p(1/\sqrt{n})\} \\ &= \left(1-\frac{r}{n}\right)\overline{H} + O_p(r^{-2\delta_t}) + \left(1-\frac{r}{n}\right)O_p(1/\sqrt{r}) + O_p(1/r^{\delta_t}\sqrt{\lambda}) \\ &\quad +O_p(\sqrt{r}/\lambda) + O_p(\sqrt{r}/\lambda)O_p(\sqrt{r/n}), \end{aligned}$$

where it is assumed that $\pi'Z'Z\pi/\lambda \to_p \overline{H} < \infty$ and $\sup_{1\le t\le T} r^{2\delta_t}\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda = O_p(1)$. Note that, using Lemma A1 in Chao and Swanson (2005), $\pi'Z'u/\lambda = O_p(1/\sqrt{\lambda})$, $u'P_tu/\lambda = (r/\lambda)\times(u'P_tu/r) = (r/\lambda)\{\sigma_u^2 + O_p(1/\sqrt{r})\}$ and $u'u/\lambda = (n/\lambda)\times(u'u/n) = (n/\lambda)\{\sigma_u^2 + O_p(1/\sqrt{n})\}$; $\pi'_{-t}Z'_{-t}u/\lambda = O_p(r^{-\delta_t}/\sqrt{\lambda})$ from Lemma A.3-(ii) above. Since $\sqrt{r}/\lambda \to 0$ as $n, r, \lambda \to \infty$, it thus follows that

$$\frac{X'P_tX}{\lambda} - \left(\frac{r}{n}\right)\frac{X'X}{\lambda} = \left(1-\frac{r}{n}\right)\overline{H} + o_p(1) \tag{A.10}$$

However, since we consider $r/n \to 0$, it indeed holds that $[X'P_tX/\lambda - (r/n)\,X'X/\lambda]^{-1} \to_p \overline{H}^{-1}$, where $\overline{H}$ is positive definite and bounded. Similarly,

$$\frac{X'P_t\varepsilon}{\lambda} - \left(\frac{r}{n}\right)\frac{X'\varepsilon}{\lambda} = \left(1-\frac{r}{n}\right)\frac{\pi'Z'\varepsilon}{\lambda} - \frac{\pi'_{-t}Z'_{-t}\varepsilon}{\lambda} + \left(\frac{r}{\lambda}\right)\frac{u'P_t\varepsilon}{r} - \left(\frac{r}{n}\times\frac{n}{\lambda}\right)\frac{u'\varepsilon}{n} \to_p 0,$$

where $\pi'Z'\varepsilon/\lambda = O_p(1/\sqrt{\lambda})$, $u'P_t\varepsilon/r = \sigma_{\varepsilon u} + O_p(1/\sqrt{r})$ and $u'\varepsilon/n = \sigma_{\varepsilon u} + O_p(1/\sqrt{n})$ from Lemma A1 in Chao and Swanson (2005); and $\pi'_{-t}Z'_{-t}\varepsilon/\lambda = O_p(r^{-\delta_t}/\sqrt{\lambda})$ from Lemma A.3-(ii) above. It thus follows that $\widehat{\beta}_{ak} - \beta = \sum_{t=1}^{T} w_t(\widehat{\beta}_{ak,t} - \beta) \to_p 0$ since $\sum_{t=1}^{T} w_t = 1 < \infty$.

The consistency of the averaged 2SLS estimator $\widehat{\beta}_{a2sls}$ can be shown similarly, provided $r/\lambda \to 0$, since

$$\begin{aligned}\frac{X'P_tX}{\lambda} &= \frac{\pi'Z'Z\pi}{\lambda} - \frac{\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}}{\lambda} + 2\frac{\pi'Z'u}{\lambda} - 2\frac{\pi'_{-t}Z'_{-t}u}{\lambda} + \frac{u'P_tu}{\lambda} \\ &= \overline{H} + O_p(r^{-2\delta_t}) + O_p(1/\sqrt{\lambda}) + O_p(1/r^{\delta_t}\sqrt{\lambda}) + (r/\lambda)\{\sigma_u^2 + O_p(1/\sqrt{r})\} \to_p \overline{H}\end{aligned}$$

and

$$\frac{X'P_t\varepsilon}{\lambda} = \frac{\pi'Z'\varepsilon}{\lambda} - \frac{\pi'_{-1}Z'_{-t}\varepsilon}{\lambda} + \left(\frac{r}{\lambda}\right)\frac{u'P_t\varepsilon}{r} \to_p 0. \quad \blacksquare$$

**Proof of Theorem 2 (Asymptotic Normality)** We first prove the case of averaged $k$-class estimator given in (A.9). Similarly as the proof of Theorem 1, we have

$$\sqrt{\lambda}\left(\widehat{\beta}_{ak} - \beta\right) = \sum_{t=1}^{T} w_t \left(\frac{X'P_tX}{\lambda} - \left(\frac{r}{n}\right)\frac{X'X}{\lambda}\right)^{-1}\left(\frac{X'P_t\varepsilon}{\sqrt{\lambda}} - \left(\frac{r}{n}\right)\frac{X'\varepsilon}{\sqrt{\lambda}}\right) + o_p(1)$$

since, in the numerator,

$$\left(\widehat{\kappa}_t - \frac{r}{n}\right)\frac{X'\varepsilon}{\sqrt{\lambda}} = o_p\left(\sqrt{\frac{r}{\lambda}}\sqrt{\frac{r}{n}}\right)\frac{X'\varepsilon}{\sqrt{n}} \to_p 0$$

with $X'\varepsilon/\sqrt{n} = \pi'Z'\varepsilon/\sqrt{n} + u'\varepsilon/\sqrt{n} = O_p((\lambda/n)^{-1/2}) + O_p(1)$ from Lemma A1 in Chao and Swanson (2005) and Assumption 2-(iii). From (A.10), $[X'P_tX/\lambda - (r/n)\,X'X/\lambda]^{-1} \to_p \overline{H}^{-1} > 0$ for $r/n \to 0$. Furthermore, from Lemma A.5, $\lambda^{-1/2}(\sum_{t=1}^{T} w_t X'\,(P_t - (r/n)\,I)\,\varepsilon) \to_d \mathcal{N}(0, \sigma_\varepsilon^2\overline{H} + \varsigma\left\{\mathbb{E}(\varepsilon_i^2u_i^2) - \sigma_{\varepsilon u}^2\right\})$ with $r/\lambda \to \varsigma < \infty$. The desired result follows by combining these two results.

For the case of the averaged 2SLS estimator, we have

$$\sqrt{\lambda}\left(\widehat{\beta}_{a2sls} - \beta\right) = \sum_{t=1}^{T} w_t \left(\frac{X'P_tX}{\lambda}\right)^{-1}\frac{X'P_t\varepsilon}{\sqrt{\lambda}} = \overline{H}^{-1}\frac{1}{\sqrt{\lambda}}\sum_{t=1}^{T} w_t X'P_t\varepsilon + o_p(1),$$

where it can be shown that $(1/\sqrt{\lambda})\sum_{t=1}^{T} w_t X'P_t\varepsilon \to_d \mathcal{N}(0, \sigma_\varepsilon^2\overline{H})$ as $n, r, \lambda \to \infty$ by letting $C = \sum_{t=1}^{T} w_t P_t$ in the proof of Lemma A.5 above. $\blacksquare$

**Proof of Theorem 3 (Overidentifying Restriction Test)** Using the same argument as in the proof of Theorem 1, for each $t$, we let

$$\widehat{\varepsilon}_t = y - X\widehat{\beta}_{ak,t} = \left\{ I - X\left( X'\left( P_t - \left(\frac{r}{n}\right) I \right) X \right)^{-1} X'\left( P_t - \left(\frac{r}{n}\right) I \right) \right\} \varepsilon + o_p(1)$$

for either the LIML or bias-corrected 2SLS estimator $\widehat{\beta}_{ak,t}$ since $\widehat{\kappa}_t - r/n = o_p(r/n)$. Then we have

$$\begin{aligned} \widehat{\varepsilon}_t' P_t \widehat{\varepsilon}_t &= \varepsilon' P_t \varepsilon - 2\varepsilon' P_t X \left( X'\left( P_t - \left(\frac{r}{n}\right) I \right) X \right)^{-1} X'\left( P_t - \left(\frac{r}{n}\right) I \right) \varepsilon \\ &\quad + \varepsilon' \left( P_t - \left(\frac{r}{n}\right) I \right) X \left( X'\left( P_t - \left(\frac{r}{n}\right) I \right) X \right)^{-1} X'\left( P_t - \left(\frac{r}{n}\right) I \right) \varepsilon + o_p(1) \\ &= \varepsilon' P_t \varepsilon - \varepsilon' \left( P_t - \left(\frac{r}{n}\right) I \right) X \left( X'\left( P_t - \left(\frac{r}{n}\right) I \right) X \right)^{-1} X'\left( P_t - \left(\frac{r}{n}\right) I \right) \varepsilon + o_p(1) \end{aligned} \tag{A.11}$$

for $r/n \to 0$ as $r, n \to \infty$. We first note that, similarly as Lemma A.1 in Lee and Okui (2012), it can be obtained that

$$\frac{1}{\sqrt{r}} \sum_{t=1}^{T} w_t \varepsilon' \left( P_t - \left(\frac{r}{n}\right) I \right) \varepsilon \to_d \mathcal{N}\left(0, \mathbb{E}(\varepsilon_i^4) - \sigma_\varepsilon^4\right), \tag{A.12}$$

which is from (A.7) in the proof of Lemma A.5 above since

$$\begin{aligned} \frac{1}{\sqrt{\lambda}} \sum_{t=1}^{T} w_t \varepsilon' \left( P_t - \left(\frac{r}{n}\right) I \right) \varepsilon &= \frac{1}{\sqrt{\lambda}} (1,0) (\varepsilon, X)' \left\{ \sum_{t=1}^{T} w_t \left( P_t - \left(\frac{r}{n}\right) I \right) \right\} (\varepsilon, X) (1,0)' \\ &\to_d \mathcal{N}\left(0, \varsigma \left\{ \mathbb{E}(\varepsilon_i^4) - \sigma_\varepsilon^4 \right\}\right), \end{aligned}$$

where $\lim_{r,\lambda\to\infty} (r/\lambda) = \varsigma < \infty$ and we have $\mathbb{E}[\sum_{t=1}^{T} w_t \varepsilon'(P_t - (r/n) I)\varepsilon | Z] = tr[\sum_{t=1}^{T} w_t (P_t - (r/n) I)\mathbb{E}(\varepsilon\varepsilon'|Z)] = (r-r)\sigma_\varepsilon^2 = 0$. By Jensen's inequality, note that $\mathbb{E}(\varepsilon_i^4) - \sigma_\varepsilon^4 > 0$ for $\mathbb{E}(\varepsilon_i^2) = \sigma_\varepsilon^2 > 0$. In addition, we have

$$\frac{1}{\sqrt{\lambda}} X'\left( P_t - \left(\frac{r}{n}\right) I \right) \varepsilon = O_p(1) \quad \text{and} \quad \frac{1}{\lambda} X'\left( P_t - \left(\frac{r}{n}\right) I \right) X \to_p \overline{H} > 0 \tag{A.13}$$

for each $t$, where the first result follows similarly as the proof of Lemma A.5 above by letting $C = P_t - (r/n) I$ for each $t$ (or letting $w_t = 1$ and $w_s = 0$ for $s \neq t$); and the second result is from Lemma A.3 above as $X'(P_t - (r/n) I) X/\lambda = (\pi_t' Z_t' Z_t \pi_t/\lambda) + 2(u' Z_t \pi_t/\lambda) + (u' P_t u/\lambda) - (r/n)(\pi' Z' Z \pi/\lambda) - 2(r/n)(u' Z \pi/\lambda) - (r/n)(u'u/n) \to_p \overline{H}$. By combining these results in (A.11), (A.12) and (A.13), if we define $J_n(r, \underline{w}) \equiv (1/\sqrt{r}) \sum_{t=1}^{T} w_t (\widehat{\varepsilon}_t' P_t \widehat{\varepsilon}_t - (r/n)\varepsilon'\varepsilon)$, we can

derive that

$$J_n(r,\underline{w}) \equiv \frac{1}{\sqrt{r}}\sum_{t=1}^{T} w_t\left(\widehat{\varepsilon}_t' P_t \widehat{\varepsilon}_t - \left(\frac{r}{n}\right)\varepsilon'\varepsilon\right) \tag{A.14}$$
$$= \frac{1}{\sqrt{r}}\sum_{t=1}^{T} w_t \varepsilon'\left(P_t - \left(\frac{r}{n}\right) I\right)\varepsilon + O_p\left(\frac{1}{\sqrt{r}}\right) \to_d \mathcal{N}\left(0, \mathbb{E}(\varepsilon_i^4) - \sigma_\varepsilon^4\right).$$

Now, we let $\widehat{\sigma}_\varepsilon^2 = (\sum_{t=1}^T w_t\widehat{\varepsilon}_t)'(\sum_{t=1}^T w_t\widehat{\varepsilon}_t)/n = (y - X\hat{\beta}_{ak})'(y - X\hat{\beta}_{ak})/n$. Then $\widehat{\sigma}_\varepsilon^2 \to_p \sigma_\varepsilon^2$ from Lemma A.6, and $\widehat{\sigma}_\varepsilon^2 - \varepsilon'\varepsilon/n = O_p(1/\sqrt{\lambda})$ from (A.8) and $\widehat{\beta}_{ak} - \beta = O_p(1/\sqrt{\lambda})$ in Theorem 2. We write

$$\begin{aligned}\frac{J_n(r,\underline{w})}{\sigma_\varepsilon^2} &= \left\{\frac{1}{\sqrt{r}}\sum_{t=1}^{T} w_t\left(\frac{\widehat{\varepsilon}_t' P_t \widehat{\varepsilon}_t}{\widehat{\sigma}_\varepsilon^2} - r\right) + \sqrt{r}\left(\frac{\widehat{\sigma}_\varepsilon^2 - \varepsilon'\varepsilon/n}{\widehat{\sigma}_\varepsilon^2}\right)\right\} \times \left(1 + \frac{\widehat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2}{\sigma_\varepsilon^2}\right) \\ &= \left\{\frac{1}{\sqrt{r}}\left(\widehat{\mathcal{J}}_a - r\right) + O_p\left(\frac{1}{\sqrt{r\lambda}}\right)\right\} \times (1 + o_p(1)) = \frac{1}{\sqrt{r}}\left(\widehat{\mathcal{J}}_a - r\right) + o_p(1),\end{aligned}$$

where $\widehat{\mathcal{J}}_a = \sum_{t=1}^T w_t\widehat{\varepsilon}_t' P_t \widehat{\varepsilon}_t/\widehat{\sigma}_\varepsilon^2$. Since $J_n(r,\underline{w})/\sigma_\varepsilon^2 \to_d \mathcal{N}(0, (\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4) - 1)$ from (A.14), therefore, we can conclude that $(\widehat{\mathcal{J}}_a - r)/\sqrt{r} \to_d \mathcal{N}(0, (\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4) - 1)$. We thus obtain that

$$\frac{\widehat{\mathcal{J}}_a - (r-1)}{\sqrt{(r-1)\left\{(\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4) - 1\right\}}} = \sqrt{\frac{r}{r-1}} \times \frac{\widehat{\mathcal{J}}_a - r}{\sqrt{r\left\{(\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4) - 1\right\}}} + \frac{1}{\sqrt{r-1}} \to_d \mathcal{N}(0,1)$$

as $n, r, \lambda \to \infty$. From this last result, using the normal approximation of the chi-square distribution similarly as Newey and Windmeijer (2009), we derive that

$$\begin{aligned}&\lim_{n,r,\lambda\to\infty} \Pr\left\{\widehat{\mathcal{J}}_a \geq q_{\alpha,r-1}\right\} \\ = &\lim_{n,r,\lambda\to\infty} \Pr\left\{\frac{\widehat{\mathcal{J}}_a - (r-1)}{\sqrt{(r-1)\left\{(\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4) - 1\right\}}} \times \sqrt{\frac{(\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4) - 1}{2}} \geq \frac{q_{\alpha,r-1} - (r-1)}{\sqrt{2(r-1)}}\right\} \leq \alpha\end{aligned}$$

if $\left\{(\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4) - 1\right\}/2 \leq 1$, where $q_{\alpha,r-1}$ being the $1-\alpha$ quantile of the $\mathcal{X}_{r-1}^2$ distribution. This proves the first result.

Note that the inequality above becomes equality when $\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4 = 3$, which is the case of normal $\varepsilon_i$. If we use some consistent estimator $\widehat{\mu}_\varepsilon$ for $\mathbb{E}(\varepsilon_i^4)/\sigma_\varepsilon^4$ using Lemma A.6, then we can use the standard normal critical values $q_\alpha^*$ to obtain the asymptotically correct size as the second result since

$$\lim_{n,r,\lambda\to\infty} \Pr\left\{\widehat{\mathcal{J}}_a \geq q_{\alpha,r}\right\} = \lim_{n,r,\lambda\to\infty} \Pr\left\{\frac{\widehat{\mathcal{J}}_a - r}{\sqrt{r(\widehat{\mu}_\varepsilon - 1)}} \geq q_\alpha^*\right\} = \alpha. \quad \blacksquare$$

## A.3 Proofs of the approximate MSE

We proves Theorem 4 for each estimator separately.

**Proof of Theorem 4 (2SLS Case)** We write $\sqrt{\lambda}(\widehat{\beta}_{a2sls} - \beta) = \sum_{t=1}^{T} w_t \widehat{H}_t^{-1} \widehat{h}_t$, where

$$\widehat{H}_t = X' P_t X/\lambda = H + \Phi_t^H + \Gamma_t^H \quad \text{and} \quad \widehat{h}_t = X' P_t \varepsilon/\sqrt{\lambda} = h + \Phi_t^h + \Gamma_t^h$$

with

$$\begin{aligned}
H &= \pi' Z' Z \pi/\lambda \\
\Phi_t^H &= -\pi' Z'(I - P_t) Z\pi/\lambda + (u' Z\pi + \pi' Z' u)/\lambda \\
\Gamma_t^H &= u' P_t u/\lambda - (u'(I - P_t) Z\pi + \pi' Z'(I - P_t) u)/\lambda \\
h &= \pi' Z' \varepsilon/\sqrt{\lambda} \\
\Phi_t^h &= -\pi' Z'(I - P_t)\varepsilon/\sqrt{\lambda} + u' P_t \varepsilon/\sqrt{\lambda} \\
\Gamma_t^h &= 0.
\end{aligned}$$

In this case, we let $\rho_{r,\lambda} = O_p(r^2/\lambda + \Delta(r, \underline{w})) = O_p(r^2/\lambda + \sum_{t=1}^{T} w_t r^{-2\delta_t}) \to 0$, where the second equality is from Lemma A.4. For each term, we can show that $H = O_p(1)$ and $h = O_p(1)$ from Lemma A.3-(i); $\Phi_t^H = O_p(r^{-2\delta_t} + 1/\sqrt{\lambda})$ since $\pi' Z'(I - P_t) Z\pi/\lambda = \pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t}/\lambda = O_p(r^{-2\delta_t})$ from Assumption 3-(ii) and $u' Z\pi/\lambda = O_p(1/\sqrt{\lambda})$ as in the proof of Theorem 1 above; $\Phi_t^h = O_p(r^{-\delta_t} + r/\sqrt{\lambda})$ since $\pi' Z'(I - P_t)\varepsilon/\sqrt{\lambda} = O_p(r^{-\delta_t})$ and $\mathbb{E}[u P_t \varepsilon | Z]/\sqrt{\lambda} = O_p(r/\sqrt{\lambda})$ as in the proof of Theorem 1 above; and similarly, $\Gamma_t^H = O_p(r/\lambda + r^{-\delta_t}/\sqrt{\lambda})$. Therefore, the conditions in Lemma A.1 are satisfied and we obtain the decomposition

$$\begin{aligned}
\lambda(\widehat{\beta}_{a2sls} - \beta)^2 &= \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \widehat{H}_t^{-1} \widehat{h}_t \widehat{h}_s \widehat{H}_s^{-1} \\
&= \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s H^{-1} (B_{ts,1}(r) + B_{ts,2}(r) + B_{ts,3}(r)) H^{-1} + o_p(\rho_{r,\lambda}),
\end{aligned}$$

where $B_{ts,1}(r)$, $B_{ts,2}(r)$ and $B_{ts,3}(r)$ are given in (A.3), (A.4) and (A.5), respectively. Since $\sum_{t=1}^{T} w_t r^{-2\delta_t} = o(1)$ from Lemma A.4-(i) above, it can be verified that

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s B_{ts,2}(r) &= \sum_{t=1}^{T}\sum_{t=1}^{T} w_t w_s \left[O_p(r^{-\delta_t} + r/\sqrt{\lambda}) O_p(r^{-2\delta_s} + 1/\sqrt{\lambda})\right] = o_p(\rho_{r,\lambda}) \\
\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s B_{ts,3}(r) &= \left[\sum_{t=1}^{T} w_t O_p(r^{-2\delta_t} + 1/\sqrt{\lambda})\right]^2 = o_p(\rho_{r,\lambda})
\end{aligned}$$

as $\lambda \to \infty$, where $\rho_{r,\lambda} \to_p 0$ and $r/\lambda \to 0$.

In order to find the expression as (A.2) in Lemma A.1, we observe that $\Gamma_{ts}^{A}(r) = 0$ in this case. Therefore, the expression (A.2) can be derived by looking at individual terms of $B_{ts,1}(r)$ as follows:

$$\mathbb{E}[hh'|Z] = \mathbb{E}\left[\pi' Z' \varepsilon\varepsilon' Z\pi/\lambda | Z\right] = \sigma_\varepsilon^2 H$$

from Lemma A.3-(i),

$$\begin{aligned}
\mathbb{E}[\Phi_t^h \Phi_s^{h\prime}|Z] &= \mathbb{E}\left[\pi' Z'(I-P_t)\varepsilon\varepsilon'(I-P_s)Z\pi/\lambda|Z\right] + \mathbb{E}\left[uP_t\varepsilon\varepsilon' P_s u/\lambda|Z\right] \\
&\quad -\mathbb{E}\left[\pi' Z'(I-P_t)\varepsilon\varepsilon' P_s u/\lambda|Z\right] - \mathbb{E}\left[u' P_t\varepsilon\varepsilon'(I-P_s)Z\pi/\lambda|Z\right] \\
&= \sigma_\varepsilon^2(\pi'_{-t} Z'_{-t} Z_{-s}\pi_{-s}/\lambda) + \sigma_{\varepsilon u}^2(r^2/\lambda) + o_p(r^2/\lambda) \\
&\quad -\mathbb{E}\left[\pi' Z'(I-P_t)\varepsilon\varepsilon' P_s u/\lambda|Z\right] - \mathbb{E}\left[u' P_t\varepsilon\varepsilon'(I-P_s)Z\pi/\lambda|Z\right],
\end{aligned}$$

from Lemma A.3-(iv), and

$$\begin{aligned}
\mathbb{E}[h\Phi_t^{h\prime}|Z] &= \mathbb{E}\left[-\pi' Z'\varepsilon\varepsilon'(I-P_t)Z\pi/\lambda + \pi' Z'\varepsilon\varepsilon' P_t u/\lambda|Z\right] \\
&= -\mathbb{E}\left[\pi' Z'(I-P_t)\varepsilon\varepsilon'(I-P_t)Z\pi/\lambda|Z\right] - \mathbb{E}\left[\pi' Z' P_t\varepsilon\varepsilon'(I-P_t)Z\pi/\lambda|Z\right] \\
&\quad +\mathbb{E}\left[\pi' Z'\varepsilon\varepsilon' P_t u/\lambda|Z\right] \\
&= -\sigma_\varepsilon^2(\pi'_{-t} Z'_{-t} Z_{-t}\pi_{-t}/\lambda) + 0 + \mathbb{E}\left[\pi' Z'\varepsilon\varepsilon' P_t u/\lambda|Z\right].
\end{aligned}$$

Moreover, from Lemma A.3-(vi), we have

$$\begin{aligned}
\mathbb{E}[hh' H^{-1}\Phi_t^{H\prime}|Z] &= -\sigma_\varepsilon^2(\pi'_{-t} Z'_{-t} Z_{-t}\pi_{-t}/\lambda) + 2\mathbb{E}\left[\pi' Z'\varepsilon\varepsilon' Z\pi H^{-1} u' Z\pi/\lambda^2|Z\right] \\
&= -\sigma_\varepsilon^2(\pi'_{-t} Z'_{-t} Z_{-t}\pi_{-t}/\lambda) + O_p(1/\lambda).
\end{aligned}$$

Therefore, by combining these results, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \mathbb{E}[(h+\Phi_t^h)(h+\Phi_s^h)' - hh'H^{-1}\Phi_s^{H\prime} - \Phi_t^H H^{-1} hh' | Z] \\
= \;& \sigma_\varepsilon^2 H + \sigma_{\varepsilon u}^2 \left(\frac{r^2}{\lambda}\right) + \sigma_\varepsilon^2 \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\frac{\pi'_{-t} Z'_{-t} Z_{-s}\pi_{-s}}{\lambda}\right) \\
& + 2\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \mathbb{E}\left[\left.\frac{\pi' Z' P_t \varepsilon\varepsilon' P_s u}{\lambda}\right| Z\right] + O_p\left(\frac{1}{\lambda}\right) + o_p\left(\frac{r^2}{\lambda}\right) \\
= \;& \sigma_\varepsilon^2 H + \sigma_{\varepsilon u}^2 \left(\frac{r^2}{\lambda}\right) + \sigma_\varepsilon^2 \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\frac{\pi'_{-t} Z'_{-t} Z_{-s}\pi_{-s}}{\lambda}\right) + o_p\left(\rho_{r,\lambda}\right)
\end{aligned}$$

since $O_p(1/\lambda) + o_p(r^2/\lambda) = o_p\left(\rho_{r,\lambda}\right)$ and $\mathbb{E}\left[\pi' Z' P_t \varepsilon\varepsilon' P_s u/\lambda | Z\right] = o_p(r/\lambda) = o_p\left(\rho_{r,\lambda}\right)$ from Lemma A.3-(v). Therefore, $\lambda(\widehat{\beta}_{a2sls} - \beta)^2$ satisfies the decomposition in Lemma A.1 if we define $\rho_{r,\lambda} = tr(S(r,\underline{w}))$ with

$$S(r,\underline{w}) = H^{-1}\left\{\sigma_\varepsilon^2 \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\frac{\pi'_{-t} Z'_{-t} Z_{-s}\pi_{-s}}{\lambda}\right) + \sigma_{\varepsilon u}^2\left(\frac{r^2}{\lambda}\right)\right\} H^{-1}.$$

Note that $\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\pi'_{-t} Z'_{-t} Z_{-s}\pi_{-s}/\lambda\right) = o_p(1)$ from Lemma A.4-(ii). ■

**Proof of Theorem 4 (LIML Case)** We write $\sqrt{\lambda}(\widehat{\beta}_{aLIML} - \beta) = \sum_{t=1}^{T} w_t \widehat{H}_t^{-1}\widehat{h}_t$, where

$$\begin{aligned}
\widehat{H}_t &= X' P_t X/\lambda - \widehat{\kappa}_t X' X/\lambda = H + \Phi_t^H + \Gamma_t^H \;\text{ and} \\
\widehat{h}_t &= X' P_t \varepsilon/\sqrt{\lambda} - \widehat{\kappa}_t X'\varepsilon/\sqrt{\lambda} = h + \Phi_t^h + \Gamma_t^h
\end{aligned}$$

with $\widehat{\kappa}_t = \min_\beta (Y - X\beta)' P_t (Y - X\beta)/(Y - X\beta)'(Y - X\beta)$. For each $t$, we define

$$\begin{aligned}
H &= \pi' Z' Z\pi/\lambda \\
\Phi_t^H &= -\pi' Z'(I - P_t) Z\pi/\lambda + (u' Z\pi + \pi' Z' u)/\lambda - \widetilde{\kappa}_t H \\
\Gamma_t^H &= (u' P_t u - \widetilde{\kappa}_t u' u)/\lambda - (u'(I - P_t) Z\pi + \pi' Z'(I - P_t) u)/\lambda - \widehat{\kappa}_t X' X/\lambda + \widetilde{\kappa}_t (H + u' u/\lambda) \\
h &= \pi' Z'\varepsilon/\sqrt{\lambda} \\
\Phi_t^h &= -\pi' Z'(I - P_t)\varepsilon/\sqrt{\lambda} + v' P_t \varepsilon/\sqrt{\lambda} - \widetilde{\kappa}_t h - \widetilde{\kappa}_t v'\varepsilon/\sqrt{\lambda} + (\sigma_{u\varepsilon}/\sigma_\varepsilon^2)(\varepsilon' P_t \varepsilon - \widetilde{\kappa}_t \varepsilon'\varepsilon)/\sqrt{\lambda} \\
\Gamma_t^h &= -(\widehat{\kappa}_t - \widetilde{\kappa}_t) X'\varepsilon/\sqrt{\lambda}.
\end{aligned}$$

with $v_i = u_i - \varepsilon_i \sigma_{u\varepsilon}/\sigma_\varepsilon^2$ and $\widetilde{\kappa}_t = (\varepsilon' P_t \varepsilon)/(n\sigma_\varepsilon^2)$. Note that from Lemma A.7 of Donald and Newey (2001), we have $\widehat{\kappa}_t - \widetilde{\kappa}_t = o_p(r/n)$, which corresponds to the Assumption 5 (i.e., $\widehat{\kappa}_t - r/n = o_p(r/n)$) since $\mathbb{E}[\varepsilon' P_t \varepsilon/n | Z] = \sigma_\varepsilon^2 tr(P_t)/n = \sigma_\varepsilon^2 (r/n)$. In this case, we let $\rho_{r,\lambda} = O_p(r/\lambda + \Delta(r,\underline{w})) = O_p(r/\lambda + \sum_{t=1}^{T} w_t r^{-2\delta_t}) \to 0$, where the second equality is from Lemma A.4. For each term, similarly as in the proof of the 2SLS case, we

can show that $H = O_p(1)$ and $h = O_p(1)$; $\Phi_t^H = O_p(r^{-2\delta_t} + 1/\sqrt{\lambda} + r/n)$ since $\pi' Z'(I - P_t)Z\pi/\lambda = \pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda = O_p(r^{-2\delta_t})$, $u'Z\pi/\lambda = O_p(1/\sqrt{\lambda})$ and $\mathbb{E}[\widetilde{\kappa}_t H|Z] = (r/n)H$. For $\Phi_t^h$, Lemma A.3 gives $\pi' Z'(I - P_t)\varepsilon/\sqrt{\lambda} = O_p(r^{-\delta_t})$, $\widetilde{\kappa}_t h = (\varepsilon' P_t \varepsilon\varepsilon' Z\pi)/(n\sqrt{\lambda}\sigma_\varepsilon^2) = O_p(r/(n\sqrt{\lambda}))$, and $\widetilde{\kappa}_t v'\varepsilon/\sqrt{\lambda} = (\varepsilon' P_t \varepsilon\varepsilon' v)/(n\sqrt{\lambda}\sigma_\varepsilon^2) = O_p(r/(n\sqrt{\lambda}))$, which yields $\Phi_t^h = v' P_t\varepsilon/\sqrt{\lambda} + (\sigma_{u\varepsilon}/\sigma_\varepsilon^2)(\varepsilon' P_t\varepsilon - \widetilde{\kappa}_t\varepsilon'\varepsilon)/\sqrt{\lambda} + O_p(r^{-\delta_t} + r/(n\sqrt{\lambda}))$. For $h = O_p(1)$, it thus follows that $B_{ts,2}(r)$ in (A.4) satisfies

$$
\begin{aligned}
&\sum_{t=1}^T \sum_{s=1}^T w_t w_s B_{ts,2}(r) \\
= \;& -4 \sum_{t=1}^T w_t \left\{ \frac{v' P_t \varepsilon}{\sqrt{\lambda}} + \frac{\sigma_{u\varepsilon}}{\sigma_\varepsilon^2} \cdot \frac{\varepsilon' P_t \varepsilon - \widetilde{\kappa}_t \varepsilon'\varepsilon}{\sqrt{\lambda}} + O_p\left( r^{-\delta_t} + \frac{r}{n\sqrt{\lambda}} \right) \right\} O_p\left( r^{-2\delta_t} + \frac{1}{\sqrt{\lambda}} + \frac{r}{n} \right) \\
& -2 \left[ \sum_{t=1}^T w_t \left\{ \frac{v' P_t \varepsilon}{\sqrt{\lambda}} + \frac{\sigma_{u\varepsilon}}{\sigma_\varepsilon^2} \cdot \frac{\varepsilon' P_t \varepsilon - \widetilde{\kappa}_t \varepsilon'\varepsilon}{\sqrt{\lambda}} + O_p\left( r^{-\delta_t} + \frac{r}{n\sqrt{\lambda}} \right) \right\} \right]^2 O_p\left( r^{-2\delta_t} + \frac{1}{\sqrt{\lambda}} + \frac{r}{n} \right) \\
= \;& o_p(\rho_{r,\lambda})
\end{aligned}
$$

since $(\varepsilon' P_t\varepsilon - \widetilde{\kappa}_t\varepsilon'\varepsilon)/\sqrt{\lambda} = O_p(r/(n\sqrt{\lambda}))$ from Lemma A.3-(ix) and $\sum_{t=1}^T w_t v' P_t\varepsilon/\sqrt{\lambda} = O_p(1)$ by CLT as in Lemma A.5 above and Lemma A.1 of Lee and Okui (2012). Note that $\mathbb{E}[v' P_t\varepsilon|Z] = r\mathbb{E}[\varepsilon_i v_i|Z] = 0$ by construction. It can be also verified that $B_{ts,3}(r)$ in (A.5) satisfies $\sum_{t=1}^T \sum_{s=1}^T w_t w_s B_{ts,3}(r) = o_p(\rho_{r,\lambda})$ similarly. Furthermore, $\Gamma_t^H = o_p(\rho_{r,\lambda})$ since $\mathbb{E}[(u' P_t u - \widetilde{\kappa}_t u'u)/\lambda|Z] = O_p(r/n\lambda) = o_p(r/\lambda)$, $\pi' Z'(I - P_t)u/\lambda = O_p(r^{-\delta_t}/\sqrt{\lambda})$, $\widehat{\kappa}_t X'X/\lambda - \widetilde{\kappa}_t(H + u'u/\lambda) = (\widehat{\kappa}_t - \widetilde{\kappa}_t)X'X/\lambda + \widetilde{\kappa}_t(X'X/\lambda - H - u'u/\lambda) = o_p(r/n) + O_p(r/n)O_p(1/\sqrt{\lambda}) = o_p(\rho_{r,\lambda})$ since $X'X/\lambda - H - u'u/\lambda = 2\pi' Z'u/\lambda$ with $\pi' Z'u/\sqrt{\lambda} = O_p(1)$ by CLT; and finally $\Gamma_t^h = o_p(r/n)O_p(\sqrt{n/\lambda}) = o_p(\rho_{r,\lambda})$ also by CLT.

In order to find the expression as (A.2) in Lemma A.1, we let $\Phi_t^h = \Phi_{t,1}^h + \Phi_{t,2}^h$, where $\Phi_{t,1}^h = -\pi' Z'(I - P_t)\varepsilon/\sqrt{\lambda} + \nu' P_t\varepsilon/\sqrt{\lambda}$ and $\Phi_{t,2}^h = -\widetilde{\kappa}_t h - \widetilde{\kappa}_t v'\varepsilon/\sqrt{\lambda} + (\sigma_{u\varepsilon}/\sigma_\varepsilon^2)(\varepsilon' P_t\varepsilon - \widetilde{\kappa}_t\varepsilon'\varepsilon)/\sqrt{\lambda}$. We observe that $\Gamma_{ts}^A(r) = \Phi_{t,1}^h \Phi_{s,2}^{h\prime} + \Phi_{t,2}^h \Phi_{s,1}^{h\prime} + \Phi_{t,2}^h \Phi_{s,2}^{h\prime}$ and $\sum_{t=1}^T \sum_{s=1}^T w_t w_s \Gamma_{ts}^A(r) = o_p(\rho_{r,\lambda})$. Then, similarly as the 2SLS case, the expression (A.2) can be derived by looking at the dominating terms in $B_{ts,1}(r)$ as follows:

$$
\begin{aligned}
\mathbb{E}\left[hh'|Z\right] &= \mathbb{E}[\pi' Z'\varepsilon\varepsilon' Z\pi/\lambda|Z] = \sigma_\varepsilon^2 H \\
\mathbb{E}\left[h\Phi_{t,1}^{h\prime}|Z\right] &= -\sigma_\varepsilon^2 \pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda + \mathbb{E}\left[\pi' Z'\varepsilon\varepsilon' P_t v/\lambda|Z\right] \\
\mathbb{E}\left[\Phi_{t,1}^h \Phi_{s,1}^{h\prime}|Z\right] &= \sigma_\varepsilon^2 \pi'_{-t}Z'_{-t}Z_{-s}\pi_{-s}/\lambda + \sigma_\varepsilon^2\sigma_v^2(r/\lambda) + o_p(r/\lambda) \\
&\quad -\mathbb{E}\left[\pi' Z'(I - P_t)\varepsilon\varepsilon' P_s v/\lambda|Z\right] - \mathbb{E}\left[v' P_t\varepsilon\varepsilon'(I - P_s)Z\pi/\lambda|Z\right]
\end{aligned}
$$

from Lemma A.3 for $\sigma_{\varepsilon v} = 0$ by construction. It can be also shown that $\sum_{t=1}^T w_t \mathbb{E}\left[h\Phi_{t,2}^{h\prime}|Z\right] = o_p(\rho_{r,\lambda})$ and $\sum_{t=1}^T w_t \mathbb{E}[hh'H^{-1}\Phi_t^{H\prime}|Z] = -\sigma_\varepsilon^2 \sum_{t=1}^T w_t(\pi'_{-t}Z'_{-t}Z_{-t}\pi_{-t}/\lambda) + o_p(\rho_{r,\lambda})$. There-

fore, by combining these results, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \mathbb{E}[(h+\Phi_t^h)(h+\Phi_s^h)' - hh'H^{-1}\Phi_s^{H\prime} - \Phi_t^H H^{-1}hh'|Z] \\
= \;& \sigma_\varepsilon^2 H + \sigma_\varepsilon^2\sigma_v^2\left(\frac{r}{\lambda}\right) + \sigma_\varepsilon^2 \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\frac{\pi'_{-t}Z'_{-t}Z_{-s}\pi_{-s}}{\lambda}\right) \\
& + 2\sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \mathbb{E}\left[\left.\frac{\pi' Z' P_t \varepsilon\varepsilon' P_s v}{\lambda}\right| Z\right] + o_p\left(\rho_{r,\lambda}\right) \\
= \;& \sigma_\varepsilon^2 H + \sigma_\varepsilon^2\sigma_v^2\left(\frac{r}{\lambda}\right) + \sigma_\varepsilon^2 \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\frac{\pi'_{-t}Z'_{-t}Z_{-s}\pi_{-s}}{\lambda}\right) + o_p\left(\rho_{r,\lambda}\right)
\end{aligned}$$

since $\mathbb{E}\left[\pi' Z' P_t \varepsilon\varepsilon' P_s v/\lambda|Z\right] = 0$ similarly as Lemma A.3-(v), where $\mathbb{E}[\varepsilon_i^2 v_i|z_i] = 0$ is assumed in this case. Therefore, as in the previous proof, $\lambda(\widehat{\beta}_{aLIML} - \beta)^2$ satisfies the decomposition in Lemma A.1 if we define $\rho_{r,\lambda} = tr(S(r,\underline{w}))$ with

$$S(r,\underline{w}) = H^{-1}\left\{\sigma_\varepsilon^2 \sum_{t=1}^{T}\sum_{s=1}^{T} w_t w_s \left(\frac{\pi'_{-t}Z'_{-t}Z_{-s}\pi_{-s}}{\lambda}\right) + \sigma_\varepsilon^2\sigma_v^2\left(\frac{r}{\lambda}\right)\right\} H^{-1}. \quad \blacksquare$$

**Proof of Theorem 4 (B2SLS Case)** We write $\sqrt{\lambda}(\widehat{\beta}_{aB2sls} - \beta) = \sum_{t=1}^{T} w_t \widehat{H}_t^{-1}\widehat{h}_t$, where

$$\begin{aligned}
\widehat{H}_t &= X' P_t X/\lambda - ((r-2)/n)X'X/\lambda = H + \Phi_t^H + \Gamma_t^H \quad \text{and} \\
\widehat{h}_t &= X' P_t \varepsilon/\sqrt{\lambda} - ((r-2)/n)X'\varepsilon/\sqrt{\lambda} = h + \Phi_t^h + \Gamma_t^h
\end{aligned}$$

with

$$\begin{aligned}
H &= \pi' Z' Z\pi/\lambda \\
\Phi_t^H &= -\pi' Z'(I - P_t)Z\pi/\lambda + (u'Z\pi + \pi' Z'u)/\lambda - ((r-2)/n)H \\
\Gamma_t^H &= (u' P_t u - ((r-2)/n)u'u)/\lambda - (u'(I-P_t)Z\pi + \pi' Z'(I-P_t)u)/\lambda \\
&\quad -((r-2)/n)\{X'X/\lambda - (H + u'u/\lambda)\} \\
h &= \pi' Z'\varepsilon/\sqrt{\lambda} \\
\Phi_t^h &= -\pi' Z'(I-P_t)\varepsilon/\sqrt{\lambda} + (u' P_t \varepsilon - (r-2)\sigma_{\varepsilon u})/\sqrt{\lambda} \\
&\quad -((r-2)/n)h - ((r-2)/n)(u'\varepsilon - n\sigma_{\varepsilon u})/\sqrt{\lambda} \\
\Gamma_t^h &= 0.
\end{aligned}$$

In this case, we let $\rho_{r,\lambda} = O_p(r/\lambda + \Delta(r,\underline{w})) = O_p(r/\lambda + \sum_{t=1}^{T} w_t r^{-2\delta_t}) \to 0$, where the second equality is from Lemma A.4. Note that, by letting $\widehat{\kappa}_t = \widetilde{\kappa}_t = (r-2)/n$, each term satisfies the same result as for the LIML case above except $\Phi_t^h$. For $\Phi_t^h$, however, it can be similarly shown that $\Phi_t^h = (u' P_t \varepsilon - r\sigma_{\varepsilon u})/\sqrt{\lambda} + O_p(r^{-\delta_t} + r/n + r/\sqrt{n\lambda})$ since $\pi' Z'(I-P_t)\varepsilon/\sqrt{\lambda} = O_p(r^{-\delta_t})$, $h = O_p(1)$ and $(u'\varepsilon - n\sigma_{\varepsilon u})/\sqrt{n} = O_p(1)$ by CLT so that $((r-$

$2)/n)(u'\varepsilon - n\sigma_{\varepsilon u})/\sqrt{\lambda} = O_p(r/\sqrt{n\lambda})$. Therefore, it also holds that $\sum_{t=1}^T \sum_{s=1}^T w_t w_s B_{ts,2}(r) = o_p(\rho_{r,\lambda})$ and $\sum_{t=1}^T \sum_{s=1}^T w_t w_s B_{ts,3}(r) = o_p(\rho_{r,\lambda})$ since $\sum_{t=1}^T w_t (u' P_t \varepsilon - r\sigma_{\varepsilon u})/\sqrt{\lambda} = O_p(1)$ by CLT as in the previous proof.

In order to find the expression as (A.2) in Lemma A.1, we let $\Phi_t^h = \Phi_{t,1}^h + \Phi_{t,2}^h$, where $\Phi_{t,1}^h = -\pi' Z'(I - P_t)\varepsilon/\sqrt{\lambda} + (u' P_t \varepsilon - (r-2)\sigma_{\varepsilon u})/\sqrt{\lambda}$ and $\Phi_{t,2}^h = -((r-2)/n)h - ((r-2)/n)(u'\varepsilon - n\sigma_{\varepsilon u})/\sqrt{\lambda}$. We observe that $\Gamma_{ts}^A(r) = \Phi_{t,1}^h \Phi_{s,2}^{h\prime} + \Phi_{t,2}^h \Phi_{s,1}^{h\prime} + \Phi_{t,2}^h \Phi_{s,2}^{h\prime}$ and $\sum_{t=1}^T \sum_{s=1}^T w_t w_s \Gamma_{ts}^A(r) = o_p(\rho_{r,\lambda})$ as in the LIML case above. Then, the expression (A.2) can be derived by looking at the dominating terms in $B_{ts,1}(r)$ as follows:

$$\begin{aligned}
\mathbb{E}\left[hh'|Z\right] &= \mathbb{E}[\pi' Z' \varepsilon\varepsilon' Z\pi/\lambda|Z] = \sigma_\varepsilon^2 H \\
\mathbb{E}\left[h\Phi_{t,1}^{h\prime}|Z\right] &= -\sigma_\varepsilon^2 \pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t}/\lambda + \mathbb{E}\left[\pi' Z' \varepsilon\varepsilon' P_t u/\lambda|Z\right] \\
\mathbb{E}\left[\Phi_{t,1}^h \Phi_{s,1}^{h\prime}|Z\right] &= \sigma_\varepsilon^2 \pi'_{-t} Z'_{-t} Z_{-s} \pi_{-s}/\lambda + (\sigma_\varepsilon^2 \sigma_u^2 + \sigma_{\varepsilon u}^2)(r/\lambda) + o_p(r/\lambda) \\
&\quad -\mathbb{E}\left[\pi' Z'(I - P_t)\varepsilon\varepsilon' P_s u/\lambda|Z\right] - \mathbb{E}\left[u' P_t \varepsilon\varepsilon'(I - P_s) Z\pi/\lambda|Z\right],
\end{aligned}$$

from Lemma A.3. Moreover, it can be also shown that $\sum_{t=1}^T w_t \mathbb{E}\left[h\Phi_{t,2}^{h\prime}|Z\right] = o_p(\rho_{r,\lambda})$ and $\sum_{t=1}^T w_t \mathbb{E}[hh' H^{-1} \Phi_t^{H\prime}|Z] = -\sigma_\varepsilon^2 \sum_{t=1}^T w_t (\pi'_{-t} Z'_{-t} Z_{-t} \pi_{-t}/\lambda) + o_p(\rho_{r,\lambda})$. Therefore, by combining these results, we have

$$\begin{aligned}
&\sum_{t=1}^T \sum_{s=1}^T w_t w_s \mathbb{E}[(h + \Phi_t^h)(h + \Phi_s^h)' - hh' H^{-1} \Phi_s^{H\prime} - \Phi_t^H H^{-1} hh'|Z] \\
= \;& \sigma_\varepsilon^2 H + \sigma_\varepsilon^2 \sigma_v^2 \left(\frac{r}{\lambda}\right) + \sigma_\varepsilon^2 \sum_{t=1}^T \sum_{s=1}^T w_t w_s \left(\frac{\pi'_{-t} Z'_{-t} Z_{-s} \pi_{-s}}{\lambda}\right) \\
& + 2 \sum_{t=1}^T \sum_{s=1}^T w_t w_s \mathbb{E}\left[\left.\frac{\pi' Z' P_t \varepsilon\varepsilon' P_s u}{\lambda}\right| Z\right] + o_p\left(\rho_{r,\lambda}\right) \\
= \;& \sigma_\varepsilon^2 H + \sigma_\varepsilon^2 \sigma_v^2 \left(\frac{r}{\lambda}\right) + \sigma_\varepsilon^2 \sum_{t=1}^T \sum_{s=1}^T w_t w_s \left(\frac{\pi'_{-t} Z'_{-t} Z_{-s} \pi_{-s}}{\lambda}\right) + o_p\left(\rho_{r,\lambda}\right)
\end{aligned}$$

since $\mathbb{E}\left[\pi' Z' P_t \varepsilon\varepsilon' P_s u/\lambda|Z\right] = 0$ from Lemma A.3-(v), where $\mathbb{E}[\varepsilon_i^2 u_i|z_i] = 0$ is assumed in this case. Therefore, as in the previous proof, $\lambda(\widehat{\beta}_{aB2sls} - \beta)^2$ satisfies the decomposition in Lemma A.1 if we define $\rho_{r,\lambda} = tr(S(r, \underline{w}))$ with

$$S(r, \underline{w}) = H^{-1} \left\{ \sigma_\varepsilon^2 \sum_{t=1}^T \sum_{s=1}^T w_t w_s \left(\frac{\pi'_{-t} Z'_{-t} Z_{-s} \pi_{-s}}{\lambda}\right) + (\sigma_\varepsilon^2 \sigma_u^2 + \sigma_{\varepsilon u}^2) \left(\frac{r}{\lambda}\right) \right\} H^{-1}.$$

Note that we can rewrite $(\sigma_\varepsilon^2 \sigma_u^2 + \sigma_{\varepsilon u}^2)$ as $(\sigma_\varepsilon^2 \sigma_v^2 + 2\sigma_{\varepsilon u}^2)$ in this expression since $\sigma_v^2 = \sigma_u^2 - (\sigma_{\varepsilon u}^2/\sigma_\varepsilon^2)$ for $v_i = u_i - \varepsilon_i \sigma_{u\varepsilon}/\sigma_\varepsilon^2$. ■

## References


Andrews, D.W.K. (1999). Consistent moment selection procedures for generalized method of moments estimation, *Econometrica*, 67, 543-564.

Bai, J. and S. Ng (2010). Instrumental variable estimation in a data rich environment, *Econometric Theory*, 26(6), 1577-1606.

Bekker, P.A.(1994) Alternative approximations to the distributions of instrumental variable estimators, *Econometrica*, 62(3), 657-681.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain, *Econometrica*, 80(6), 2369-2429.

Canay, I.A. (2010). Simultaneous selection and weighting of moments in GMM using a trapezoidal kernel, *Journal of Econometrics,* 156(2), 284-303.

Caner, M., X. Han, and Y. Lee (2015). Adaptive Elastic Net GMM estimation with many invalid moment conditions: Simultaneous model and moment selection, *CPR Working Papers Series*, No. 177, Syracuse University.

Carrasco, M. (2012). A regularization approach to the many instruments problem, *Journal of Econometrics*, 170(2), 383-398.

Chao, J.C. and N.R. Swanson (2005). Consistent estimation with a large number of weak instruments, *Econometrica*, 73(5), 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.

Chen, X., D.T. Jacho-Chavez, and O. Linton (2015). Averaging of an increasing number of moment condition estimators, *Econometric Theory*, forthcoming.

Donald, S.G., G. Imbens, and W.K. Newey (2009). Choosing instrumental variables in conditional moment restriction models, *Journal of Econometrics*, 152(1), 28-36.

Donald, S.G. and W.K. Newey (2001). Choosing the number of instruments, *Econometrica*, 69(5), 1161-1191.

Guggenberger, P. and Y. Sun (2006). Bias-reduced log-periodogram and Whittle estimation of the long-memory parameter without variance inflation, *Econometric Theory*, 22(5), 863–912.

Hahn, J., J. Hausman, and G. Kuersteiner (2004). Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations, *Econometrics Journal*, 7(1), 272–306.

Hansen, B.E. (2007). Least squares model averaging, *Econometrica*, 75(4), 1175-1189.

Hansen, C. and J. Hausman, and W.K. Newey (2008). Estimation with many instrumental variables, *Journal of Business and Economic Statistics*, 26(4), 398-422.

Hausman, J.A., W.K. Newey, T. Woutersen, J.C. Chao, and N.R. Swanson (2012). Instrumental variable estimation with heteroskedasticity and many instruments, *Quantitative Economics*, 3(2), 211-255.

Kuersteiner, G. (2012). Kernel weighted GMM estimators for linear time series models, *Journal of Econometrics*, 170(2), 399-421.

Kuersteiner, G. and R. Okui (2010). Constructing optimal instruments by first stage prediction averaging, *Econometrica,* 78(2), 697-718.

Lee, Y. and R. Okui (2012). Hahn-Hausman test as a specification test, *Journal of Econometrics*, 167(1), 133-139.

Morimune, K. (1983). Approximate distribution of the k-class estimators when the degree of overidentifiability is large compared with the sample size, *Econometrica*, 51(3), 821–841.

Nagar, A.L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations, *Econometrica*, 27(4), 575-595.

Newey, W.K and F. Windmeijer (2009). GMM with many weak moment conditions, *Econometrica*, 77(3), 687-719.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions, *Journal of Econometrics*, 165(1), 70-86.

Sawa, T. (1973). Almost unbiased estimator in simultaneous equation systems, *International Economic Review*, 14(1), 97–106.

van Hasselt, M. (2010). Many instruments asymptotic approximations under nonnormal error distributions, *Econometric Theory*, 26(2), 633–645.