Jackknife Instrumental Variable Estimation with Heteroskedasticity

Paul A. Bekker*
Faculty of Economics and Business
University of Groningen

Federico Crudu†
Pontificia Universidad Católica de Valparaíso,
University of Sassari and CRENoS

February 2014

Abstract

We present a new genuine jackknife estimator for instrumental variable inference with unknown heteroskedasticity. It weighs observations such that many-instruments consistency is guaranteed while the signal component in the data is maintained. We show that this results in a smaller signal component in the many-instruments asymptotic variance when compared to estimators that neglect a part of the signal to achieve consistency. Both many-instruments and many-weak-instruments asymptotic distributions are derived using high-level assumptions that allow for the simultaneous presence of weak and strong instruments for different explanatory variables. Standard errors are formulated compactly. We review briefly known estimators and show in particular that our symmetric jackknife estimator performs well when compared to the HLIM and HFUL estimators of Hausman et al. in Monte Carlo experiments.

Key words: Instrumental Variables, Heteroskedasticity, Many Instruments, Jackknife
JEL classification: C12, C13, C23

*p.a.bekker@rug.nl, P.O. Box 800, 9700 AV Groningen, The Netherlands.
†federico.crudu@gmail.com, Avenida Errazuriz 2734, Valparaíso, Chile. Financial support from Marie Curie Excellence Grant MEXT-CT-2006-042471 is gratefully acknowledged.

1 Introduction

The presence of unknown heteroskedasticity is a common setting in microeconometric research. Inference based on many instruments asymptotics, as introduced by Kunitomo (1980), Morimune (1983) and Bekker (1994), shows 2SLS is inconsistent under homoskedasticity. Bekker and Van der Ploeg (2005) show in general LIML is many-instruments inconsistent as well under heteroskedasticity. A number of estimators have been considered, including the two step feasible GMM estimator of Hansen (1982), the continuously updated GMM estimator of Hansen, Heaton and Yaron (1996), the grouping estimators of Bekker and Van der Ploeg (2005), the jackknife estimators of Angrist, Imbens and Krueger (1999), the modified LIML estimators of Kunitomo (2012) and the HLIM and HFUL estimators of Hausman et al. (2012). In particular this last paper has been important for the approach that we present here.

Our starting point is aimed at formulating a consistent estimator for the noise component in the expectation of the sum of squares of disturbances when projected on the space of instruments. That way a method of moments estimator can be formulated similar to the derivation of LIML as a moments estimator as described in Bekker (1994). Surprisingly the estimator can be described as a symmetric jackknife estimator, where 'omit one' fitted values are used not only for the explanatory variables but instead for all endogenous variables including the dependent variable. Influential papers on jackknife estimation include Phillips and Hale (1977), Blomquist and Dahlberg (1999), Angrist, Imbens and Krueger (1999), Donald and Newey (2000), Ackerberg and Deveraux (2003). Our genuine jackknife estimator shares with LIML the property that the endogenous variables are treated symmetrically in the sense that it is invariant to the type of normalization, as discussed by Anderson (2005).

Hausman et al. (2012) and Chao et al. (2012, 2014) use a LIML version of the JIVE2 estimator of Angrist, Imbens and Krueger (1999). The JIVE2 estimator treats endogenous variables symmetrically, but it is not a genuine jackknife estimator. In case of homoskedasticity and many weak instruments, while assuming the number of instruments grows slower than the number of observations, the authors show the HLIM estimator is as efficient as LIML. Thus it seems the efficiency problems of jackknife estimators noted in Davidson and McKinnon (2006) are overcome. Here we show there is room for improvement. The symmetric jackknife estimator is a genuine jackknife estimator and it has a signal component that is larger than that found for HLIM, resulting in a smaller component in the asymptotic covariance matrix. Monte Carlo experiments show it performs better than HLIM and its Fuller modifications in terms of the bias-variance trade-off.

The asymptotic theory allows for both many instruments and many weak instruments asymptotics. Influential papers in this area include Donald and Newey (2001), Hahn, Hausman and Kuersteiner (2004), Hahn (2002), Hahn and Inoue (2002), Chamberlain and Imbens (2004), Chao and Swanson (2005), Stock and Yogo (2005), Han and Phillips (2006), Andrews and Stock (2007) and Van Hasselt (2010). Our results are formulated concisely. They are based on high level assumptions where the concentration parameter need not grow at the same rate as the number of observations and the quality of instruments may vary over explanatory variables.

The plan of the paper is as follows. In Section 2 we present the model and some earlier estimators. Section 3 uses a method of moments reasoning to formulate a heteroskedasticity robust estimator that is subsequently interpreted as a symmetric jackknife estimator. Asymptotic assumptions and results are given in Section 4 and proved in the Appendix. Section 5 compares asymptotic distributions and Section 6 compares exact distributions based on Monte Carlo simulations. Section 7 concludes.

2 The Model and some estimators

Consider observations in the n vector $\boldsymbol{y}$ and the $n \times g$ matrix $\boldsymbol{X}$ that satisfy

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

$$\boldsymbol{X} = \boldsymbol{Z\Pi} + \boldsymbol{V}, \tag{2}$$

where the g vector $\boldsymbol{\beta}$ and the $k \times g$ matrix $\boldsymbol{\Pi}$ contain unknown parameters, and $\boldsymbol{Z}$ is an $n \times k$ observed matrix of instruments. Similar to Hausman et al. (2012) we assume $\boldsymbol{Z}$ to be nonrandom, or we could allow $\boldsymbol{Z}$ to be random, but condition on it, as in Chao et al. (2012). The assumption $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{Z\Pi}$ is made for convenience and could be generalized as in Hausman et al. (2012), or as in Bekker (1994). The disturbances in the $n \times (1+g)$ matrix $(\boldsymbol{\varepsilon}, \boldsymbol{V})$ have rows $(\varepsilon_i, \boldsymbol{V}_i)$, which are assumed to be independent, with zero mean and covariance matrices

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_i^2 & \boldsymbol{\sigma}_{12i} \\ \boldsymbol{\sigma}_{21i} & \boldsymbol{\Sigma}_{22i} \end{pmatrix}.$$

The covariance matrices of the rows $(y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, are given by

$$\boldsymbol{\Omega}_i = \begin{pmatrix} 1 & \boldsymbol{\beta}' \\ \boldsymbol{0} & \boldsymbol{I}_g \end{pmatrix} \boldsymbol{\Sigma}_i \begin{pmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{\beta} & \boldsymbol{I}_g \end{pmatrix}. \tag{3}$$

Throughout we use the notation where $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ has elements $P_{ij} = \boldsymbol{e}_i'\boldsymbol{P}\boldsymbol{e}_j$, and $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are conformable unit vectors.

The estimators that we consider are related to LIML which is found by minimizing the objective function

$$Q_{\mathrm{LIML}}(\boldsymbol{\beta}) = \frac{(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{X\beta})}{(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X\beta})}. \tag{4}$$

The LIML estimator and Fuller (1977) modifications are given by

$$\hat{\boldsymbol{\beta}} = \{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X} - \lambda_f \boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X}\}^{-1} \{\boldsymbol{X}'\boldsymbol{P}\boldsymbol{y} - \lambda_f \boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}\},$$

$$\lambda_f = \lambda - \alpha/(n-k),$$

$$\lambda = 1/\lambda_{\max}[\{(\boldsymbol{y},\boldsymbol{X})'\boldsymbol{P}(\boldsymbol{y},\boldsymbol{X})\}^{-1} \ (\boldsymbol{y},\boldsymbol{X})'(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{y},\boldsymbol{X})],$$

where $\lambda_{\max}$ indicates the largest eigenvalue. For $\alpha = 0$ LIML is found, which has no moments under normality. Under normality and homoskedasticity, where the matrices $\boldsymbol{\Sigma}_i$ do not vary over $i = 1, \ldots, n$, the Fuller estimator that is found for $\alpha = 1$ has moments and is nearly unbiased. If one wishes to minimize the mean square error, $\alpha = 4$ would be appropriate. However, as shown by Bekker and Van der Ploeg (2005), LIML is in general inconsistent under many-instruments asymptotics with heteroskedasticity.[1]

Similarly, the Hansen (1982) two-step GMM estimator is inconsistent under many-instruments asymptotics. It is found by minimizing

$$Q_{\text{GMM}}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{Z}\left\{\sum_{i=1}^{n} \hat{\sigma}_i^2 \boldsymbol{Z}_i'\boldsymbol{Z}_i\right\}^{-1} \boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{5}$$

where $\hat{\sigma}_i^2 = (y_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})^2$ and $\hat{\boldsymbol{\beta}}$ is a first stage IV estimator such as 2SLS or LIML. A many-instruments consistent version is given by the continuously updated GMM estimator of Hansen, Heaton and Yaron (1996), which is found by minimizing the objective function (5) where $\hat{\sigma}_i^2$ is replaced by $\hat{\sigma}_i^2(\boldsymbol{\beta}) = (y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2$. Newey and Windmeijer (2009) showed this estimator and other generalized empirical likehood estimators are asymptotically robust to heteroskedasticity and many weak instruments. Donald and Newey (2000) gave a jackknife interpretation. However, the efficiency depends on using a heteroskedastic consistent weighting matrix that can degrade the finite sample performance with many instruments as was shown by Hausman et al. (2012) in Monte Carlo experiments.

To reduce problems related to the consistent estimation of the weighting matrix Bekker

[1] When dummy instruments indicate groups and group sizes are equal, LIML is many-instruments consistent even under heteroskedasticity.

and Van der Ploeg (2005) use clustering of observations. If this clustering, or grouping, is formulated as a function of $\boldsymbol{Z}$, it is exogenous and continuously updated GMM estimation can be formulated conditional on it. Bekker and Van der Ploeg (2005) give standard errors that are consistent for sequences where the number of groups grows at the same rate as the number of observations. Contrary to LIML, the asymptotic distribution is not affected by deviations from normality. It uses the between group heteroskedasticity to gain efficiency, yet it loses efficiency due to within group sample variance of the instruments.

Another way to avoid problems of heteroskedasticity is to use the jackknife approach. The jackknife estimator, suggested by Phillips and Hale (1977) and later by Angrist, Imbens and Krueger (1999) and Blomquist and Dahlberg (1999) uses the omit-one-observation approach to reduce the bias of 2SLS in a homoskedastic context. The JIVE1 estimator of Angrist, Imbens and Krueger (1999) is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{JIVE1}} &= (\tilde{\boldsymbol{X}}'\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{y}, \\ \boldsymbol{e}_i'\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{X}}_i &= \frac{\boldsymbol{Z}_i(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} - h_i\boldsymbol{X}_i}{1-h_i}, \end{aligned} \tag{6}$$

where $h_i = P_{ii}$, and $i = 1, \ldots, n$. It is robust against heteroskedasticity and many-instruments consistent. The JIVE2 estimator of Angrist, Imbens and Krueger (1999) is not a genuine jackknife estimator but it shares the many-instruments consistency property with JIVE1. It uses $\tilde{\boldsymbol{X}} = (\boldsymbol{P} - \boldsymbol{D})\boldsymbol{X}$ and thus minimizes a 2SLS-like objective function

$$Q_{\text{JIVE2}}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\{\boldsymbol{P} - \boldsymbol{D}\}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{7}$$

where $\boldsymbol{D} = \text{Diag}(\boldsymbol{h})$ is the diagonal matrix formed by the elements of $\boldsymbol{h} = (h_1, \ldots, h_n)'$. JIVE2 is consistent under many instruments asymptotics as has been shown by Ackerberg and Deveraux (2003). However, Davidson and McKinnon (2006) have shown that the jackknife estimators can have low efficiency relative to LIML under homoskedasticity.

Therefore, Hausman et al. (2012) consider jackknife versions of LIML and the Fuller

(1977) estimator by using the objective function

$$Q_{\text{HLIM}}(\boldsymbol{\beta}) = \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\{\boldsymbol{P} - \boldsymbol{D}\}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}. \tag{8}$$

The estimators are given by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \{\boldsymbol{X}'(\boldsymbol{P} - \boldsymbol{D})\boldsymbol{X} - \hat{\alpha}\boldsymbol{X}'\boldsymbol{X}\}^{-1}\{\boldsymbol{X}'(\boldsymbol{P} - \boldsymbol{D})\boldsymbol{y} - \hat{\alpha}\boldsymbol{X}'\boldsymbol{y}\}, \\ \hat{\alpha} &= \frac{(n+c)\widetilde{\alpha} - c}{n + c\widetilde{\alpha} - c}, \\ \widetilde{\alpha} &= \lambda_{\min}[\{(\boldsymbol{y}, \boldsymbol{X})'(\boldsymbol{y}, \boldsymbol{X})\}^{-1}(\boldsymbol{y}, \boldsymbol{X})'\{\boldsymbol{P} - \boldsymbol{D}\}(\boldsymbol{y}, \boldsymbol{X})]. \end{aligned} \tag{9}$$

For $c = 0$, $\hat{\boldsymbol{\beta}}_{\text{HLIM}}$ is found, and in particular $c = 1$ produces $\hat{\boldsymbol{\beta}}_{\text{HFUL}}$. Hausman et al. (2012) consider many-instruments and many-weak-instruments asymptotics and show the asymptotic distributions are not affected by deviations from normality. The estimators perform much better than the original jackknife estimators.[2]

Kunitomo (2012) considered modifications of the LIML objective function (4) resulting in many-instruments consistent estimators under heteroskedasticity. The modification amounts to replacing the diagonal elements of $\boldsymbol{P}$ by values whose difference with k/n vanishes asymptotically. The resulting estimators have the same many-instruments asymptotic distribution as HLIM. By replacing the diagonal elements of $\boldsymbol{P}$ by values that converge to their average value the LIML modifications lose signal just as HLIM. As the asymptotic requirement does not define a unique estimator, there are many possibilities to use for finite-sample comparisons. Kunitomo (2012) claims one of these improves on HLIM in finite samples.

[2]However, notice that the denominator of the objective function (8) weighs the squared errors with equal weights even if the ith row of $\boldsymbol{Z}$ has small squared elements, or even when $\boldsymbol{Z}_i = \mathbf{0}$. If $\boldsymbol{Z} = (\boldsymbol{Z}_1', \ \boldsymbol{Z}_2')'$, where $\boldsymbol{Z}_1$ has n_1 rows and $\boldsymbol{Z}_2 = \mathbf{O}$, then the redundant parameters in $\boldsymbol{\Sigma}_i$, $i = n_1 + 1, \ldots, n$, affect both the finite-sample and the many-instruments asymptotic distribution of HLIM.

3 A method of moments and jackknife estimator

In order to handle heteroskedasticity the grouping estimators use data clustering. In many cases this means information will be lost in the process, although between-group heteroskedasticity is used to improve efficiency. The jackknife approach maintains original instruments to a larger extent, but seems to remove possibly relevant information on $\boldsymbol{\beta}$ contained in the matrix $(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{D}(\boldsymbol{y}, \boldsymbol{X})$. As an alternative to the objective function Q_{HLIM}, we consider a method-of-moments approach that maintains the signal component in the expectation of $(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{P}(\boldsymbol{y}, \boldsymbol{X})$ and aims at estimating the noise component consistently. Thus we try to maintain the information contained in the data to a larger extent without adding much additional noise. The estimator can be interpreted as a symmetric jackknife estimator.

In order to formulate a method-of-moments estimator similar to LIML, we consider minimizing a criterion given by

$$Q(\boldsymbol{\beta}) = \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{B}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}, \tag{10}$$

where, conditional on the instruments $\boldsymbol{Z}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are fixed positive semidefinite matrices. We find the following conditional expectations

$$\mathrm{E}\left\{(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{A}(\boldsymbol{y}, \boldsymbol{X})\right\} = \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_g \end{pmatrix} \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}, \boldsymbol{I}_g) + \sum_{i=1}^{n} A_{ii}\boldsymbol{\Omega}_i, \tag{11}$$

$$\mathrm{E}\left\{(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X})\right\} = \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_g \end{pmatrix} \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{B}\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}, \boldsymbol{I}_g) + \sum_{i=1}^{n} B_{ii}\boldsymbol{\Omega}_i. \tag{12}$$

The method-of-moments approach amounts to solving these equations for $\boldsymbol{\beta}$. Typically this occurs for solving

$$\left[\mathrm{E}\left\{(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{A}(\boldsymbol{y}, \boldsymbol{X})\right\} - l\,\mathrm{E}\left\{(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X})\right\}\right] \begin{pmatrix} 1 \\ -\boldsymbol{\beta} \end{pmatrix} = \boldsymbol{0}, \tag{13}$$

when the signal matrix $\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{A} - l\boldsymbol{B})\boldsymbol{Z}\boldsymbol{\Pi}$ is positive definite, preferably large, and the noise component vanishes: $\sum_{i=1}^n A_{ii}\boldsymbol{\Omega}_i = l\sum_{i=1}^n B_{ii}\boldsymbol{\Omega}_i$.[3]

For example, LIML uses $\boldsymbol{A} = \boldsymbol{P}$ and $\boldsymbol{B} = \boldsymbol{I}_n - \boldsymbol{P}$, which allows for the solution $\boldsymbol{\beta}$ if all matrices $\boldsymbol{\Omega}_i$ are equal.[4] In case of heteroskedasticity, however, the noise term does not drop out which renders LIML many-instruments inconsistent.

Hausman et al. (2012) achieve consistency by using $\boldsymbol{A} = \boldsymbol{P} - \boldsymbol{D}$, so $A_{ii} = 0$, $i = 1, \ldots, n$ and $\boldsymbol{B} = \boldsymbol{I}_n$. That way (11) alone is enough to solve for $\boldsymbol{\beta}$. By removing the diagonal from $\boldsymbol{P}$, however, the signal contained in $(\boldsymbol{\beta}, \boldsymbol{I}_g)'\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{D}\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}, \boldsymbol{I}_g)$ is removed as well.

In our approach we maintain the signal matrix by requiring $\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{Z}$ and $\boldsymbol{Z}'\boldsymbol{B}\boldsymbol{Z} = \mathbf{O}$ and we remove the noise component by requiring only that the diagonal of $\boldsymbol{A}$ is a scalar multiple of the diagonal of $\boldsymbol{B}$, i.e. $l = \operatorname{tr}(\boldsymbol{A})/\operatorname{tr}(\boldsymbol{B})$ and $A_{ii} = lB_{ii}$, $i = 1, \ldots, n$. In that case (13) is satisfied and the many-instruments consistency of the minimizer of $Q(\boldsymbol{\beta})$ would follow easily.

Without loss of generality we use $l = 1$, so $\boldsymbol{A}$ and $\boldsymbol{B}$ should have the same diagonal, and $\boldsymbol{A} = \boldsymbol{P} + \boldsymbol{\Delta}$ where $\boldsymbol{\Delta}$ satisfies $\boldsymbol{Z}'\boldsymbol{\Delta}\boldsymbol{Z} = \mathbf{O}$. To stay close to LIML we would like $\sum_{i=1}^n \Delta_{ii}\boldsymbol{\Omega}_i$ to be small. As $A_{ii}\boldsymbol{\Omega}_i = h_i\boldsymbol{\Omega}_i + \Delta_{ii}\boldsymbol{\Omega}_i$, and $\mathrm{E}\{(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X})\} = \sum_{i=1}^n A_{ii}\boldsymbol{\Omega}_i$, we find $\sum_{i=1}^n \Delta_{ii}\boldsymbol{\Omega}_i$ is small if $(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X})$ is a good estimator, with small bias, of $\sum_{i=1}^n h_i\boldsymbol{\Omega}_i$. We formulate such an estimator for $\boldsymbol{\Omega}_i$ intuitively as

$$\widehat{\boldsymbol{\Omega}}_i = (\boldsymbol{y}, \boldsymbol{X})'\frac{(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{e}_i\boldsymbol{e}_i'(\boldsymbol{I}_n - \boldsymbol{P})}{\boldsymbol{e}_i'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{e}_i}(\boldsymbol{y}, \boldsymbol{X}), \tag{14}$$

which amounts to choosing

$$\boldsymbol{B} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}(\boldsymbol{I}_n - \boldsymbol{P}). \tag{15}$$

This fixes the diagonal elements of $\boldsymbol{\Delta}$ as the difference between the diagonal elements of $\boldsymbol{B}$ and $\boldsymbol{P}$.[5] A choice for $\boldsymbol{\Delta}$ with such diagonal elements and satisfying $\boldsymbol{Z}\boldsymbol{\Delta}\boldsymbol{Z} = \mathbf{O}$ is given

[3] Actually, only $\sum_{i=1}^n A_{ii}\boldsymbol{\Omega}_i(1, \boldsymbol{\beta}')' = l\sum_{i=1}^n B_{ii}\boldsymbol{\Omega}_i(1, -\boldsymbol{\beta}')'$ needs to hold, which amounts to $\sum_{i=1}^n A_{ii}\sigma_i^2 = l\sum_{i=1}^n B_{ii}\sigma_i^2$ and $\sum_{i=1}^n A_{ii}\boldsymbol{\sigma}_{21i} = l\sum_{i=1}^n B_{ii}\boldsymbol{\sigma}_{21i}$.

[4] A less restrictive sufficient requirement is that the covariances between the error term ε_i and all endogenous variables do not vary.

[5] We find $\boldsymbol{\Delta}_{ii} = -h_i + \sum_{j=1}^n \frac{h_j}{1-h_j}\boldsymbol{e}_i'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{e}_j\boldsymbol{e}_j'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{e}_i = -h_i + \frac{h_i}{1-h_i(1-h_i)^2} + \sum_{j\neq i}\frac{h_j}{1-h_j}P_{ij}^2 =$

by

$$\boldsymbol{\Delta} = \boldsymbol{P}\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{P} - \frac{1}{2}\left\{\boldsymbol{P}\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1} + \boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{P}\right\}. \tag{16}$$

To further motivate the choice of $\boldsymbol{\Delta}$ we observe that the minimizer of $Q(\boldsymbol{\beta})$ in (10) can be interpreted as a normalized symmetric jackknife estimator. That is

$$\hat{\boldsymbol{\beta}}_{\text{SJIVE}} = \arg\min_{\boldsymbol{\beta}}\{Q(\boldsymbol{\beta})\} = \arg\min_{\boldsymbol{\beta}}\{Q_{\text{SJIVE}}(\boldsymbol{\beta})\}, \tag{17}$$

$$Q_{\text{SJIVE}}(\boldsymbol{\beta}) = \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{C}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{B}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}, \tag{18}$$

$$\begin{aligned}
\boldsymbol{C} &= \boldsymbol{A} - \boldsymbol{B}, \qquad (19)\\
&= \boldsymbol{P} - \boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1} + \left\{\boldsymbol{P}\boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1} + (\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{D}\boldsymbol{P}\right\}/2\\
&= (\tilde{\boldsymbol{P}} + \tilde{\boldsymbol{P}}')/2,\\
\tilde{\boldsymbol{P}} &= \boldsymbol{P} - \boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1} + \boldsymbol{D}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{P}\\
&= (\boldsymbol{I}_n - \boldsymbol{D})^{-1}(\boldsymbol{P} - \boldsymbol{D}).
\end{aligned}$$

Notice the diagonal elements of $\boldsymbol{C}$ equal zero. The jackknife estimator $\boldsymbol{\beta}_{\text{JIVE1}} = (\tilde{\boldsymbol{X}}'\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{y}$ in (6) is based on $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{P}}\boldsymbol{X}$. So, if we define $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{P}}\boldsymbol{y}$, then we find the numerator of the objective function is given by

$$\begin{aligned}
(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{C}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= \frac{1}{2}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})\\
&= (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).
\end{aligned}$$

That is to say, genuine jackknife prediction is used for all endogenous variables symmetrically, including the dependent variable. As the statistical problem is basically symmetric in the endogenous variables, it seems a good property the symmetry is maintained in the

$-\frac{h_i^2}{1-h_i} + \sum_{j=1}^{n} \frac{h_j}{1-h_j} P_{ij}^2$. As both terms on the right hand side are bounded, so is $|\Delta_{ii}| \leq h_i \max_i(h_i)/\{1 - \max_i(h_i)\}$. If $\max_i(h_i) \to 0$, then $\Delta_{ii} \to 0$, but that will not happen with many instruments if $k/n \to 0$ does not hold. Yet k/n may be small in practice.

jackknifing procedure, so that the estimator is invariant with respect to the particular type of normalization.

We find the signal matrix is larger for SJIVE than for HLIM since

$$\begin{aligned}\mathrm{E}\{(\boldsymbol{y},\boldsymbol{X})'\boldsymbol{C}(\boldsymbol{y},\boldsymbol{X})\} &= \begin{pmatrix}\boldsymbol{\beta}'\\ \boldsymbol{I}_g\end{pmatrix}\boldsymbol{\Pi}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta},\boldsymbol{I}_g) \geq \\ \mathrm{E}\{(\boldsymbol{y},\boldsymbol{X})'(\boldsymbol{P}-\boldsymbol{D})(\boldsymbol{y},\boldsymbol{X})\} &= \begin{pmatrix}\boldsymbol{\beta}'\\ \boldsymbol{I}_g\end{pmatrix}\boldsymbol{\Pi}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta},\boldsymbol{I}_g) - \begin{pmatrix}\boldsymbol{\beta}'\\ \boldsymbol{I}_g\end{pmatrix}\boldsymbol{\Pi}\boldsymbol{Z}'\boldsymbol{D}\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta},\boldsymbol{I}_g).\end{aligned} \tag{20}$$

In Section 5 we find as well that the signal component in the many-instruments asymptotic covariance matrix is smaller for SJIVE than for HLIM for cases where the heteroskedasticity is not extreme.

To compute the symmetric jackknife estimator and its Fuller modifications, let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $\boldsymbol{X}_2 = \boldsymbol{Z}_2$, where $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$, so the explanatory variables in $\boldsymbol{X}_2$ are assumed to be exogenous. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ be partitioned conformably. Let $\boldsymbol{C}^* = \boldsymbol{C} - \boldsymbol{A}\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{A}$, then the SJIVE estimator and its Fuller modifications (SJEF) can be computed by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X} - \hat{\lambda}\boldsymbol{X}\boldsymbol{B}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{C}\boldsymbol{y} - \hat{\lambda}\boldsymbol{X}\boldsymbol{B}\boldsymbol{y}\right), \tag{21}$$

$$\hat{\lambda} = \lambda - \alpha/\operatorname{tr}(\boldsymbol{B}),$$

$$\lambda = \lambda_{\min}\left[\{(\boldsymbol{y},\boldsymbol{X}_1)'\boldsymbol{B}(\boldsymbol{y},\boldsymbol{X}_1)\}^{-1}\ (\boldsymbol{y},\boldsymbol{X}_1)'\boldsymbol{C}^*(\boldsymbol{y},\boldsymbol{X}_1)\right].$$

For $\alpha = 0$ $\hat{\boldsymbol{\beta}}_{\text{SJIVE}}$ is found. Based on the Monte Carlo experiments we would use a Fuller modification $\hat{\boldsymbol{\beta}}_{\text{SJEF}}$ with $\alpha = 2$. Using Theorem 1 below we compute standard errors as the square root of the diagonal elements of the estimated covariance matrix, which is formulated concisely as

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1}\widetilde{\boldsymbol{X}}'\left(\boldsymbol{C}\boldsymbol{D}_{\hat{\boldsymbol{\varepsilon}}}^2\boldsymbol{C} + \boldsymbol{D}_{\hat{\boldsymbol{\varepsilon}}}\boldsymbol{C}^{(2)}\boldsymbol{D}_{\hat{\boldsymbol{\varepsilon}}}\right)\widetilde{\boldsymbol{X}}(\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1}, \tag{22}$$

$$\widetilde{\boldsymbol{X}} = \boldsymbol{X} - \hat{\sigma}^{-2}\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\sigma}}_{12},$$

where $\widehat{\boldsymbol{C}} = \boldsymbol{C} - \hat{\lambda}\boldsymbol{B}$ and $\boldsymbol{C}^{(2)}$ is the elementwise or Hadamard product $\boldsymbol{C} * \boldsymbol{C}$. The diagonal matrix $\boldsymbol{D}_{\hat{\varepsilon}}$ has the residuals $\hat{\boldsymbol{\varepsilon}} = y - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ on the diagonal. Finally, $\hat{\sigma}^2$ and $\hat{\boldsymbol{\sigma}}_{21}$ are found based on $\widehat{\boldsymbol{\Omega}} = (\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X})/k$, which is transformed to $\widehat{\boldsymbol{\Sigma}}$ similar to (3).

4 Asymptotic distributions

We consider many instruments and many weak instruments parameter sequences to describe the asymptotic distributions of the heteroskedasticity robust estimator SJIVE as given in (21). Our formulation allows for the presence of both weak and strong instruments within a single model. The derivation is based on high-level regularity conditions, since primitive regularity conditions could be formulated very similar to earlier ones. For example, the ones used by Hausman et al. (2012) could be used, although our results hold more generally.

Assumption 1. *The diagonal elements of the hat matrix* $\boldsymbol{P}$ *satisfy* $\max_i h_i \leq 1 - 1/c_u$.

Assumption 2. *The covariance matrices of the disturbances are bounded,* $0 \leq \boldsymbol{\Sigma}_i \leq c_u \boldsymbol{I}_{g+1}$ *and satisfy* $k^{-1}\sum_{i=1}^{n} \boldsymbol{e}_i'\boldsymbol{B}\boldsymbol{e}_i\boldsymbol{\Sigma}_i \to \boldsymbol{\Sigma}$.

We partition $\boldsymbol{\Sigma}$ as we partitioned $\boldsymbol{\Sigma}_i$ and use $\boldsymbol{\Omega}$ defined similar to (3), where $\boldsymbol{\Sigma}$ is used instead of $\boldsymbol{\Sigma}_i$. So we find $\lim_{n\to\infty} k^{-1}\,\mathrm{E}\{\boldsymbol{y}, \boldsymbol{X})\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X})\} = \boldsymbol{\Omega}$, and $\mathrm{E}(\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}) = \boldsymbol{H}$, where $\boldsymbol{H}$ is the signal matrix

$$\boldsymbol{H} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi}. \tag{23}$$

Assumption 3. $\operatorname{plim}_{n\to\infty} k^{-1}(\boldsymbol{y}, \boldsymbol{X})\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X}) = \boldsymbol{\Omega}$, *and* $\operatorname{plim}_{n\to\infty} \boldsymbol{H}^{-1}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}_g$.

Let $r_{\min} = \lambda_{\min}(\boldsymbol{H})$ be the smallest eigenvalue of the signal matrix.

Assumption 4. $r_{\min} \to \infty$.

Let $Q^*_{\text{SJIVE}}(\boldsymbol{\beta}) = kQ_{\text{SJIVE}}(\boldsymbol{\beta})$, then the many-instruments asymptotic approximations are based on the following high-level assumption.

Assumption 5. Many instruments: $k/r_{\min} \to \gamma$, *and*

$$\left\{ \frac{\partial^2 Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right\}^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = - \left\{ \frac{\partial^2 Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right\}^{-1/2} \frac{\partial Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + o_p(1) \overset{a}{\sim} \mathcal{N}\left(\mathbf{0},\ \boldsymbol{\Phi}\right),$$

where $\boldsymbol{\Phi} = \boldsymbol{\Xi}^{-1/2} \boldsymbol{\Theta} \boldsymbol{\Xi}^{-1/2}$ *and*

$$\begin{aligned} \boldsymbol{\Xi} &= \operatorname{plim}_{n\to\infty} \boldsymbol{H}^{-1/2} \frac{\partial^2 Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \boldsymbol{H}^{-1/2}, \qquad \boldsymbol{H}^{-1/2} \frac{\partial Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}(\mathbf{0},\ \boldsymbol{\Theta}), \\ \boldsymbol{\Theta} &= \lim_{n\to\infty} \operatorname{Var} \left\{ \boldsymbol{H}^{-1/2} \frac{\partial Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + o_p(1) \right\}. \end{aligned} \tag{24}$$

The $o_p(1)$ term in (24) is defined in (42) in the Appendix. That is to say, as $\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}/\operatorname{tr}\boldsymbol{B} = \sigma^2 + o_p(1)$, we find using (42) that

$$\boldsymbol{H}^{-1/2} \frac{\partial Q^*_{\text{SJIVE}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{2}{\sigma^2} \boldsymbol{H}^{-1/2} \left\{ (\mathbf{0}, \boldsymbol{I}_g) - \sigma^{-2} \boldsymbol{\sigma}_{21} \boldsymbol{\delta}' \right\} (\boldsymbol{M} - \boldsymbol{S}) \boldsymbol{\delta} - o_p(1).$$

Removing the $o_p(1)$ term leaves the vector given by the first term on right-hand side. It has a finite-sample covariance matrix. We make the mild assumption that these finite-sample covariance matrices converge to the covariance matrix of the limit distribution.

Using Assumptions 1-5 the many-instruments asymptotic distribution of SJIVE can be formulated as in Theorem 1 given below. Its derivation is given in the Appendix. In particular (42) and (43) show that Assumption 5 boils down to the asymptotic normality of

$$\boldsymbol{H}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{H}^{-1/2}(\boldsymbol{Z\Pi} + \widetilde{\boldsymbol{V}})' \boldsymbol{C} \boldsymbol{\varepsilon} + o_p(1), \tag{25}$$

$$\widetilde{\boldsymbol{V}} = \boldsymbol{V} - \boldsymbol{\varepsilon} \boldsymbol{\sigma}_{12} / \sigma^2. \tag{26}$$

Instead of the high-level Assumption 5, more primitive assumptions can be considered for specific cases that are sufficient for Assumption 5 to hold true.

For example, consider primitive conditions rather similar to the assumptions of Hausman et al. (2012). Assume, in addition to Assumptions 1, 2, 3 and 4, the disturbances have bounded fourth-order moments and let

$$\boldsymbol{Z\Pi} = n^{-1/2}\boldsymbol{Z}\widetilde{\boldsymbol{\Pi}}\boldsymbol{D}_{\boldsymbol{\mu}_n}\tilde{\boldsymbol{S}}, \tag{27}$$

where $\tilde{\boldsymbol{S}}$ is nonsingular, $\boldsymbol{D}_{\boldsymbol{\mu}_n} = \mathrm{Diag}(\boldsymbol{\mu}_n)$ and $\boldsymbol{\mu}_n$ is a g-vector with elements μ_{jn} that satisfy either $\mu_{jn} = \sqrt{n}$ or $\mu_{jn}/\sqrt{n} \to 0$. Furthermore, assume the existence of $\widetilde{\boldsymbol{H}} = \lim_{n\to\infty} n^{-1}\widetilde{\boldsymbol{\Pi}}'\boldsymbol{Z}'\boldsymbol{Z}\widetilde{\boldsymbol{\Pi}} > 0$, $\widetilde{\boldsymbol{F}} = \lim_{n\to\infty} n^{-1}\sum_i \sigma_i^2\widetilde{\boldsymbol{\Pi}}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{Z}\widetilde{\boldsymbol{\Pi}}$, and

$$\boldsymbol{\Psi}_w = \lim_{n\to\infty} k^{-1}\boldsymbol{G}, \tag{28}$$

$$\boldsymbol{G} = \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2\left(\sigma_j^2\widetilde{\boldsymbol{\Sigma}}_{i22} + \tilde{\boldsymbol{\sigma}}_{i21}\tilde{\boldsymbol{\sigma}}_{j21}'\right) = \sum_{i\neq j} C_{ij}^2\left(\sigma_j^2\widetilde{\boldsymbol{\Sigma}}_{i22} + \tilde{\boldsymbol{\sigma}}_{i21}\tilde{\boldsymbol{\sigma}}_{j21}'\right), \tag{29}$$

$$\widetilde{\boldsymbol{\Sigma}}_i = \mathrm{Var}(\varepsilon_i, \widetilde{\boldsymbol{V}}_i). \tag{30}$$

Together with the assumption $k/r_{\min} \to \gamma$, these conditions imply Assumption 5 as is further discussed in Appendix 8.4. Of course these conditions are not necessary and other less restrictive specifications may be sufficient as well.

The asymptotic distribution of SJIVE can be formulated based the high-level assumption.

Theorem 1. Many instruments *If Assumptions 1-5 are satisfied, then* $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{SJIVE}$ *is consistent and* $(\boldsymbol{X}\widehat{\boldsymbol{C}}\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \overset{a}{\sim} \mathcal{N}(\boldsymbol{0},\, \boldsymbol{\Psi})$, *where* $\widehat{\boldsymbol{C}} = \boldsymbol{C}-\hat{\lambda}\boldsymbol{B}$ *and* $\boldsymbol{\Psi} = \lim_{n\to\infty}\boldsymbol{H}^{-1/2}(\boldsymbol{F}+\boldsymbol{G})\boldsymbol{H}^{-1/2}$, *where* $\boldsymbol{F} = \sum_{i=1}^{n}\sigma_i^2\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}$ *and* $\boldsymbol{G}$ *is defined in* (29).

To formulate a consistent estimator for $\boldsymbol{\Psi}$ we need a condition that guarantees that $\hat{\varepsilon}_i$ converges to ε_i uniformly in probability. This requires uniform convergence of $\|\boldsymbol{Z}_i\boldsymbol{\Pi}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|$ to zero, which is guaranteed by the assumption below. Let d_i denote the ith diagonal element of the projection matrix $\boldsymbol{P}_{Z\Pi} = \boldsymbol{Z\Pi}(\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z\Pi})^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'$.

Assumption 6. $\max_{1\leq i\leq n} d_i \to 0$.

The restricted specification (27), where $\boldsymbol{P}_{Z\Pi} = \boldsymbol{P}_{Z\widetilde{\Pi}}$ and $\widetilde{\boldsymbol{H}} = \lim_{n\to\infty}\widetilde{\boldsymbol{\Pi}}\boldsymbol{Z}'\boldsymbol{Z}\widetilde{\boldsymbol{\Pi}}/n > 0$, is sufficient for Assumption 6 to hold true.

Theorem 2. *If Assumptions 1-6 are satisfied, a consistent estimator for* $\boldsymbol{\Psi}$ *is given by*

$$\widehat{\boldsymbol{\Psi}} = (\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1/2}\widetilde{\boldsymbol{X}}'\left(\boldsymbol{C}\boldsymbol{D}_{\hat{\varepsilon}}^2\boldsymbol{C} + \boldsymbol{D}_{\hat{\varepsilon}}\boldsymbol{C}^{(2)}\boldsymbol{D}_{\hat{\varepsilon}}\right)\widetilde{\boldsymbol{X}}(\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1/2},$$
$$\widetilde{\boldsymbol{X}} = (\boldsymbol{X} - \hat{\sigma}^{-2}\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\sigma}}_{12})$$

where $\widetilde{\boldsymbol{X}}$, $\boldsymbol{D}_{\hat{\varepsilon}}$, $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, $\boldsymbol{C}^{(2)}$, $\hat{\sigma}^2$ *and* $\hat{\boldsymbol{\sigma}}_{21}$ *are defined as in* (22).

The proofs of Theorems 1 and 2 are given in the Appendix.

The asymptotic covariance matrix $\boldsymbol{\Psi}$ has two terms. Under large-sample asymptotics, when $k/r_{\min} \to 0$ the second term vanishes. As the second term may be relevant in the finite sample, the many instruments asymptotic approximation to the finite distribution is usually more accurate than the large-sample approximation as was shown by Bekker (1994). Under homoskedasticity $\tilde{\boldsymbol{\sigma}}_{i21} = \boldsymbol{0}$, $i = 1, \ldots, n$, so the second part of $\boldsymbol{G}$ drops out.

When instruments are weak the second term may be dominant and the first term may even be negligible. Chao and Swanson (2005) used many-weak-instruments asymptotic sequences and showed the first term actually vanishes, while estimators such as LIML under homoskedasticity are still consistent. Hausman et al. (2012) derived the many-weak-instruments asymptotic distribution of HLIM and HFUL as given in (9). We have a similar result, although in our formulation the asymptotic covariance matrix need not be singular as a result of the inclusion of endogenous variables with different signal strengths.

Let $r_{\max} = \lambda_{\max}(\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi})$ be the largest eigenvalue of the signal matrix.

Assumption 7. Many weak instruments*:* $k/r_{\max} \to \infty$, $k^{1/2}/r_{\min} \to 0$ *and*

$$k^{-1/2}\frac{\partial^2 Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = k^{-1/2}\frac{\partial Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} + o_p(1) \overset{a}{\sim} \mathcal{N}\left(\boldsymbol{0},\ \boldsymbol{\Phi}_w\right),$$
$$\boldsymbol{\Phi}_w = \lim_{n\to\infty}\operatorname{Var}\left\{k^{-1/2}\frac{\partial Q^*_{SJIVE}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} + o_p(1)\right\}. \tag{31}$$

The $o_p(1)$ term in (31) is defined explicitly in (44) in the Appendix. Instead of the normality of (25), Assumption 7 boils down to the asymptotic normality of

$$k^{-1/2}\boldsymbol{H}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=k^{-1/2}(\boldsymbol{Z\Pi}+\widetilde{\boldsymbol{V}})'\boldsymbol{C}\boldsymbol{\varepsilon}+o_p(1)=k^{-1/2}\widetilde{\boldsymbol{V}}'\boldsymbol{C}\boldsymbol{\varepsilon}+o_p(1).$$

Theorem 3. Many weak instruments *If Assumptions 1-4 and 7 are satisfied, then* $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}_{SJIVE}$ *is consistent and* $k^{-1/2}\boldsymbol{X}\widehat{\boldsymbol{C}}\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\overset{a}{\sim}\mathcal{N}(\boldsymbol{0},\ \boldsymbol{\Psi}_w)$, *where* $\boldsymbol{\Psi}_w$ *is defined in* (28).

For the actual computation of standard errors the many weak instruments asymptotic distribution is not needed, since the many-instruments standard errors remain consistent.

5 A comparison of asymptotic distributions

Comparing many-instruments asymptotic distributions usually does not give a clear-cut ordering of asymptotic covariance matrices with positive semidefinite differences. Many-instrument asymptotic distributions are good for detecting inconsistency and to formulate standard errors, but to use them for efficiency comparisons would be too restrictive. Only in restrictive classes such efficiency exists. Typically such classes, as in Kunitomo (2012), exclude relevant alternatives, which makes them less useful for making relevant comparisons. For example, contrary to large-sample asymptotic approximations, there is no guarantee that using more instruments increases asymptotic efficiency. Dropping a weak instrument may well increase efficiency. Consequently, it might happen that estimators that treat high-signal instruments poorly perform quite well when instruments have low signal strength. In parts of the high-dimensional parameter space one estimator may perform well, while in other parts other estimators dominate.

Under homoskedasticity Anderson et al. (2010) described a class of estimators for which LIML is many-instruments asymptotically efficient. However, in an interesting example Hausman et al. (2012, p.224) show that under homoskedasticity LIML is not many-instruments asymptotically efficient relative to HLIM. Similar results can be found in Van der Ploeg and Bekker (1995), Hahn (2002), and Chioda and Jansson (2009). With unknown heteroskedasticity things become more complicated, since instruments and variances may covary in ways that affect the asymptotic distributions substantially. Under heteroskedasticity Kunitomo (2012) describes a class for which a modified LIML estimator is asymptotically optimal. However, similar to the case of homoskedasticity, it is rather restrictive as it excludes relevant alternatives. Interestingly under homoskedasticity LIML is not many-instruments asymptotically efficient relative to its heteroskedastic modification. So, the property of asymptotic efficiency as it is commonly required with large-sample asymptotic theory is simply too strong for inferential comparisons based on many-instruments asymptotic theory.

In addition to these difficulties, the accuracy of the asymptotic approximation to the finite-sample distributions is relevant as well. For example, the continuously updated GMM estimator may perform quite well when compared to the other estimators in terms of asymptotic variances, but its actual performance in the finite sample shows a different picture. So comparing the performance of the estimators and the accuracy of the asymptotic approximations in Monte Carlo simulations remains important.

Before turning to the Monte Carlo experiments, however, we can make a comparison of the signal components in the many-instruments asymptotic covariance matrices of SJIVE and HLIM. To make a comparison we assume the heteroskedasticity is not extreme so that a positive definite difference found under homoskedasticity is present as well in a neighborhood of moderate heteroskedasticity. Under homoskedasticity and normality we compare the asymptotic distributions of LIML, SJIVE and HLIM.

For both SJIVE and HLIM the many-instruments asymptotic approximation to the finite distribution of the estimators can be formulated as a normal distribution with mean

$\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{H}^{-1}(\boldsymbol{C}^*)\left\{\boldsymbol{F}(\boldsymbol{C}^*)+\boldsymbol{G}(\boldsymbol{C}^*,\boldsymbol{\Sigma}^*)\right\}\boldsymbol{H}^{-1}(\boldsymbol{C}^*)$ where

$$\boldsymbol{H}(\boldsymbol{C}^*) = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}^*\boldsymbol{Z}\boldsymbol{\Pi}, \tag{32}$$

$$\boldsymbol{F}(\boldsymbol{C}^*) = \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}^*\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}^*\boldsymbol{Z}\boldsymbol{\Pi}, \tag{33}$$

$$\boldsymbol{G}(\boldsymbol{C}^*,\boldsymbol{\Sigma}^*) = \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^{*2}\left(\sigma_j^2\widetilde{\boldsymbol{\Sigma}}_{i22} + \tilde{\boldsymbol{\sigma}}_{i21}\tilde{\boldsymbol{\sigma}}_{j21}'\right), \qquad \widetilde{\boldsymbol{V}}_j = \boldsymbol{V}_j - \varepsilon_j\boldsymbol{\sigma}_{12}^*/\sigma^{2*}, \tag{34}$$

and the submatrices of $\widetilde{\boldsymbol{\Sigma}}_i$ have been computed as in (30). For SJIVE we have $\boldsymbol{C}^* = \boldsymbol{C}_{\text{SJIVE}} = \boldsymbol{C}$, and $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}$ as defined in (19) and Assumption 2, respectively. For HLIM $\boldsymbol{C}^* = \boldsymbol{C}_{\text{HLIM}} = \boldsymbol{P} - \boldsymbol{D}$ and $\boldsymbol{\Sigma}^* = n^{-1}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i$ should be used. Furthermore, when we assume homoskedasticity and normality, the many-instruments asymptotic distribution for LIML is found for $\boldsymbol{C}^* = \boldsymbol{C}_{\text{LIML}} = \boldsymbol{P} - k(n-k)^{-1}(\boldsymbol{I}_n - \boldsymbol{P})$.

The signal matrices $\boldsymbol{H}(\boldsymbol{C}^*)$ for LIML, SJIVE and HLIM can be ordered. Due to $\boldsymbol{P}\boldsymbol{Z} = \boldsymbol{Z}$ and $\boldsymbol{C}\boldsymbol{Z} = \left\{\boldsymbol{I}_n - (1/2)(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\right\}\boldsymbol{Z}$, we find $\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{Z}$ and so

$$\begin{aligned}
&\boldsymbol{H}_{\text{LIML}} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}_{\text{LIML}}\boldsymbol{Z}\boldsymbol{\Pi} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi} = \\
&\boldsymbol{H}_{\text{SJIVE}} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}_{\text{SJIVE}}\boldsymbol{Z}\boldsymbol{\Pi} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi} > \\
&\boldsymbol{H}_{\text{HLIM}} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}_{\text{HLIM}}\boldsymbol{Z}\boldsymbol{\Pi} = \boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{P}-\boldsymbol{D})\boldsymbol{Z}\boldsymbol{\Pi} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi} - \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{D}\boldsymbol{Z}\boldsymbol{\Pi}.
\end{aligned}$$

Under homoskedasticity the signal components $\boldsymbol{H}^{-1}(\boldsymbol{C}^*)\boldsymbol{F}(\boldsymbol{C}^*)\boldsymbol{H}^{-1}(\boldsymbol{C}^*)$ are given by

$$\begin{aligned}
\boldsymbol{H}_{\text{LIML}}^{-1}\boldsymbol{F}_{\text{LIML}}\boldsymbol{H}_{\text{LIML}}^{-1} &= \sigma^2(\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi})^{-1}, \\
\boldsymbol{H}_{\text{SJIVE}}^{-1}\boldsymbol{F}_{\text{SJIVE}}\boldsymbol{H}_{\text{SJIVE}}^{-1} &= \sigma^2(\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}_{\text{SJIVE}}^2\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi})^{-1}, \\
\boldsymbol{H}_{\text{HLIM}}^{-1}\boldsymbol{F}_{\text{HLIM}}\boldsymbol{H}_{\text{HLIM}}^{-1} &= \sigma^2\{\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{Z}\boldsymbol{\Pi}\}^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n-\boldsymbol{P})^2\boldsymbol{Z}\boldsymbol{\Pi}\{\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{Z}\boldsymbol{\Pi}\}^{-1}.
\end{aligned}$$

As

$$\boldsymbol{Z}'\boldsymbol{C}_{\text{SJIVE}}^2\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{Z} + (1/4)\boldsymbol{Z}'(\boldsymbol{I}_n-\boldsymbol{D})^{-1}(\boldsymbol{I}_n-\boldsymbol{P})(\boldsymbol{I}_n-\boldsymbol{D})^{-1}\boldsymbol{Z} \geq \boldsymbol{Z}'\boldsymbol{Z}, \tag{35}$$

and

$$(\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi})^{-1} \leq \{\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z}\boldsymbol{\Pi}\}^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n - \boldsymbol{P})^2\boldsymbol{Z}\boldsymbol{\Pi}\{\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z}\boldsymbol{\Pi}\}^{-1},$$

we find the signal component in the asymptotic covariance matrix of LIML is smaller than the signal components of both SJIVE and HLIM. When the diagonal elements h_i of $\boldsymbol{P}$ are not too big, the signal components of SJIVE and HLIM are ordered as well, as follows from Theorem 4 below, which is proven in Appendix 8.5.

Theorem 4. *Let the difference between the signal components in the many-instruments asymptotic covariance matrices of HLIM and SJIVE under homoskedasticity be given by* $\sigma^2\boldsymbol{\Upsilon}$*, where* $\boldsymbol{\Upsilon} = \boldsymbol{H}^{-1}(\boldsymbol{C}^*)\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}^{*2}\boldsymbol{Z}\boldsymbol{\Pi}\boldsymbol{H}^{-1}(\boldsymbol{C}^*) - \boldsymbol{H}^{-1}(\boldsymbol{C})\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}^2\boldsymbol{Z}\boldsymbol{\Pi}\boldsymbol{H}^{-1}(\boldsymbol{C})$*, and* $\boldsymbol{C}^* = \boldsymbol{P} - \boldsymbol{D}$*. If* $\boldsymbol{D} < \boldsymbol{I}_n/2$ *then* $\boldsymbol{\Upsilon} \geq 0$ *and* $\operatorname{rank}(\boldsymbol{\Upsilon}) = \operatorname{rank}\{(\boldsymbol{I}_n - \boldsymbol{P}_{Z\Pi})\boldsymbol{D}\boldsymbol{Z}\boldsymbol{\Pi}\}$*, where* $\boldsymbol{P}_{Z\Pi} = \boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'$*.*

Under homoskedasticity and normality, when $\boldsymbol{D} < \boldsymbol{I}_n/2$, the signal components in the asymptotic variances of LIML, SJIVE and HLIM range from small to large.[6] Under heteroskedasticity LIML becomes inconsistent in general. SJIVE and LIML remain consistent and if the difference between their signal components is positive definite, the ranking of the signal components also holds when heteroskedasticity is sufficiently moderate.

For the noise components in the asymptotic covariance matrices of LIML, SJIVE and HLIM there does not seem to be a clear ordering. Under homoskedasticity the matrices $\boldsymbol{G}(\boldsymbol{C}^*, \boldsymbol{\Sigma}^*) = \boldsymbol{R}\sum_i \sum_j C_{ij}^{*2}$ are scalar multiples of the covariance matrix $\boldsymbol{R} = \sigma^2\widetilde{\boldsymbol{\Sigma}}_{22} +$

[6] The restriction $\boldsymbol{D} < \boldsymbol{I}_n/2$, which implies $\max_i P_{ii} < 1/2$ is not very restrictive since in most applications the diagonal elements of $\boldsymbol{P}$ are of order k/n. If the instruments are poorly balanced and the restriction is violated there is ground to doubt the accuracy of all asymptotic approximations.

$\tilde{\boldsymbol{\sigma}}_{21}\tilde{\boldsymbol{\sigma}}_{21}'$. We find

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} C_{\mathrm{LIML},ij}^2 &= \frac{nk}{n-k}, \\
\sum_{i=1}^{n}\sum_{j=1}^{n} C_{\mathrm{SJIVE},ij}^2 &= \sum_{i\neq j} C_{ij}^2 = \sum_{i\neq j} P_{ij}^2\left(1+\frac{h_i}{2(1-h_i)}+\frac{h_j}{2(1-h_j)}\right)^2 > \sum_{i\neq j} P_{ij}^2, \\
\sum_{i=1}^{n}\sum_{j=1}^{n} C_{\mathrm{HLIM},ij}^2 &= \sum_{i\neq j} P_{ij}^2 = k - \sum_{i=1}^{n} h_i^2 \leq k - \sum_{i=1}^{n}\frac{k^2}{n^2} = \frac{k(n-k)}{n} < \frac{nk}{n-k},
\end{aligned}$$

so $\sum_i\sum_j C_{\mathrm{HLIM},ij}^2$ is smaller than both $\sum_i\sum_j C_{\mathrm{SJIVE},ij}^2$ and $\sum_i\sum_j C_{\mathrm{LIML},ij}^2$.[7] However, that does not imply that the noise component in the asymptotic covariance matrix of HLIM, $\boldsymbol{H}_{\mathrm{HLIM}}^{-1}\boldsymbol{R}\boldsymbol{H}_{\mathrm{HLIM}}^{-1}\sum_i\sum_j C_{\mathrm{HLIM},ij}^2$, is smaller than the noise components of either SJIVE or LIML.

We thus find, under homoskedasticity and normality, that although the signal components in the asymptotic covariance matrices favor LIML over SJIVE and HLIM, there is no clear ordering in terms of many-instruments asymptotic efficiency. Indeed Hausman et al. (2012) showed LIML is not efficient relative to HLIM. Still, if normality and homoskedasticity are guaranteed, we would prefer to use LIML. Similarly, under mild heteroskedasticity, when LIML is inconsistent, the signal components in the asymptotic covariance matrices favor SJIVE over HLIM.

6 Monte Carlo simulations

We compare the finite sample properties of the HLIM and SJIVE and their Fuller modifications given by (9) and (21), respectively. We use the same Monte Carlo set up as Hausman et al. (2012). The data generating process is given by $\boldsymbol{y} = \boldsymbol{\iota}\gamma + \boldsymbol{x}\beta + \boldsymbol{\varepsilon}$ and $\boldsymbol{x} = \boldsymbol{z}\pi + \boldsymbol{v}$, where $n = 800$, $\gamma = \beta = 0$. The strength of the instruments is varied by using two values $\pi = 0.1$ or $\pi = 0.2$, so that $\mu^2 = n\pi^2 = 8$ and $\mu^2 = 32$, respectively. Furthermore, $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{I}_n)$ and independently $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$. The disturbances $\boldsymbol{\varepsilon}$ are

[7] We thank an anonymous referee for pointing to the fact that $\sum_{i\neq j} P_{ij}^2 < \sum_{i\neq j} C_{ij}^2$.

generated by

$$\boldsymbol{\varepsilon} = \boldsymbol{v}\rho + \sqrt{\frac{1-\rho^2}{\phi^2+\psi^4}}(\phi\boldsymbol{w}_1 + \psi\boldsymbol{w}_2),$$

where $\rho = 0.3$, $\psi = 0.86$ and conditional on $\boldsymbol{z}$, independent of $\boldsymbol{v}$, $\boldsymbol{w}_1 \sim \mathcal{N}(\mathbf{0},\ \mathrm{Diag}(\boldsymbol{z})^2)$ and $\boldsymbol{w}_2 \sim \mathcal{N}(\mathbf{0},\ \psi^2\boldsymbol{I}_n)$. The values $\phi = 0$ and $\phi = 1.38072$ are chosen such that the R-squared between ε_i^2 and the instruments equals 0 and 0.2, respectively.[8] The instruments $\boldsymbol{Z}$ are given for $k = 2$, $k = 5$ and $k = 15$ by matrices with rows $(1, z_i)$, $(1, z_i, z_i^2, z_i^3, z_i^4)$ and $(1, z_i, z_i^2, z_i^3, z_i^4, z_ib_{1i}, \ldots, z_ib_{10i})$, respectively, where independent of other random variables, the elements $b_{1i}, \ldots, b_{10i}$ are i.i.d. Bernoulli distributed with $p = 1/2$. We used 20,000 simulations of estimates of β.

Figure 1 plots the nine decile ranges—between the 5th and 95th percentiles—and the median bias of Fuller modifications HFUL for $c = 0, 1, 2, 3, 4, 5$, and SJEF for $\alpha = 0, 1, 2, 3, 4, 5$, when $R^2_{\varepsilon^2|z} = 0$. As observed by Hausman et al. (2012), LIML is many-instruments consistent for this case and no big differences were found between HLIM and LIML. Here we see that the HFUL and SJEF estimators are very similar as well and the differences are due mainly to the degree of Fullerization.

When $R^2_{\varepsilon^2|z} = 0.2$ this situation changes. Table 1 compares the outcomes for HFUL when $c = 1$ and SJEF when $\alpha = 2$. We see that SJEF dominates HFUL in terms of median bias and nine decile range. The rejection rates of SJEF are smaller than the ones found for HFUL, indicating that confidence sets based on SJEF are more conservative. Figure 2 plots the median bias and nine-decile ranges for all Fullerizations when $R^2_{\varepsilon^2|z} = 0.2$. We find SJEF performs better for this setup than HFUL. That is to say, HFUL with $c = 1$ is dominated by SJEF with $\alpha = 2$. Other choices of these coefficients result in estimators with varying location and spread characteristics. When instruments are weak Fullerization is useful since allowing for some bias reduces the spread considerably. We find that this trade-off is better for SJEF than for HFUL.

[8] $R^2_{\varepsilon^2|z} = \mathrm{var}\{\mathrm{E}(\varepsilon^2|z)\}/[\mathrm{var}\{\mathrm{E}(\varepsilon^2|z)\} + \mathrm{E}\{\mathrm{var}(\varepsilon^2|z)\}]$.

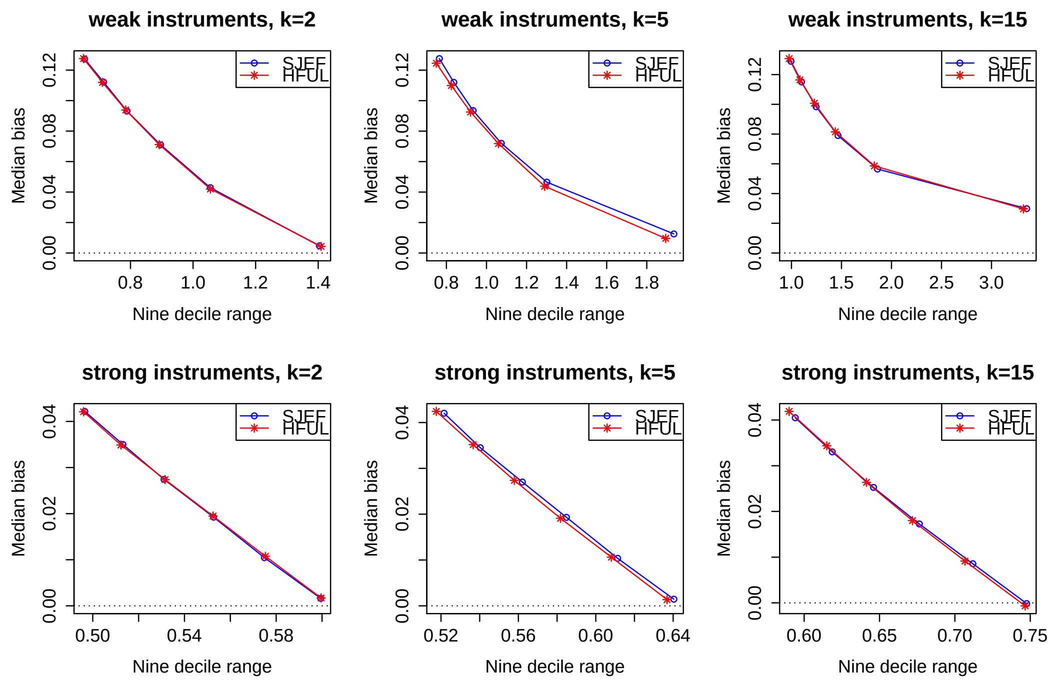

Figure 1: $R^2_{\varepsilon^2|z} = 0$: Median bias against the Nine decile range of HFUL with $c = 0, 1, 2, 3, 4, 5$ from right to left, and SJEF for $\alpha = 0, 1, 2, 3, 4, 5$ from right to left, based on 20,000 replications.

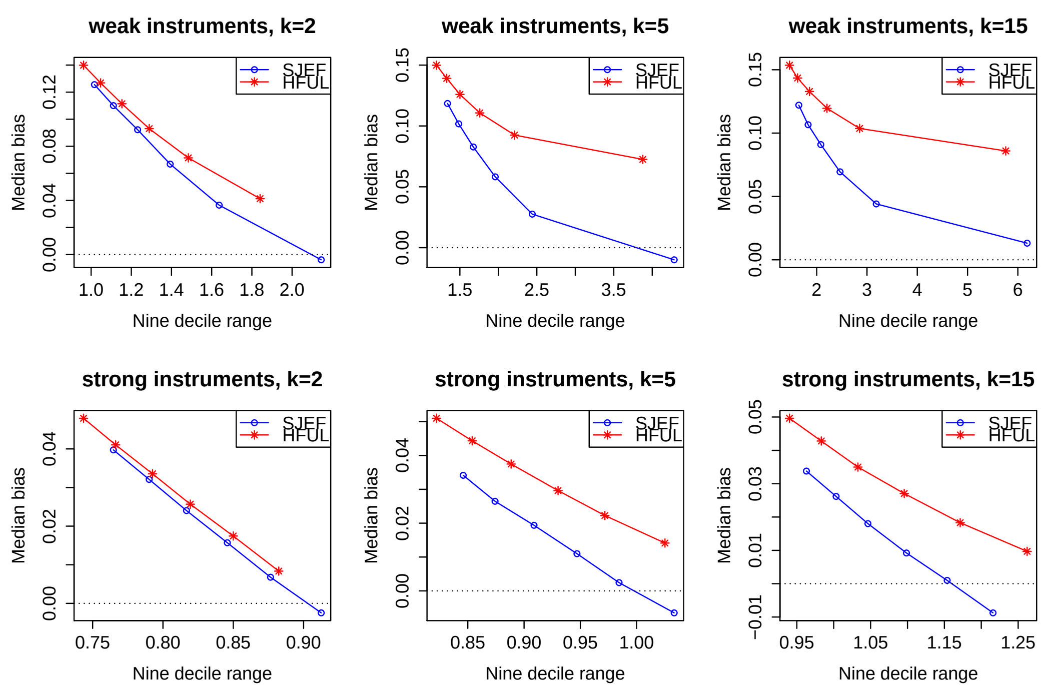

Figure 2: $R^2_{\varepsilon^2|z} = 0.2$: Median bias against the Nine decile range of HFUL with $c = 0, 1, 2, 3, 4, 5$ from right to left, and SJEF for $\alpha = 0, 1, 2, 3, 4, 5$ from right to left, based on 20,000 replications.

μ^2	k	*Median bias* HFUL	SJEF	*Nine decile range* HFUL	SJEF	*Rejection rates* HFUL	SJEF
8	2	0.071	0.067	1.484	1.393	0.033	0.026
8	5	0.093	0.058	2.211	1.959	0.042	0.021
8	15	0.104	0.069	2.853	2.464	0.044	0.030
32	2	0.017	0.016	0.850	0.846	0.048	0.043
32	5	0.022	0.011	0.972	0.947	0.045	0.036
32	15	0.018	0.009	1.171	1.099	0.047	0.039

Table 1: $R^2_{\varepsilon^2|z} = 0.2$: Median bias, Nine decile range and 5% Rejection rates for HFUL ($c = 1$) and SJEF ($\alpha = 2$) based on 20,000 replications.

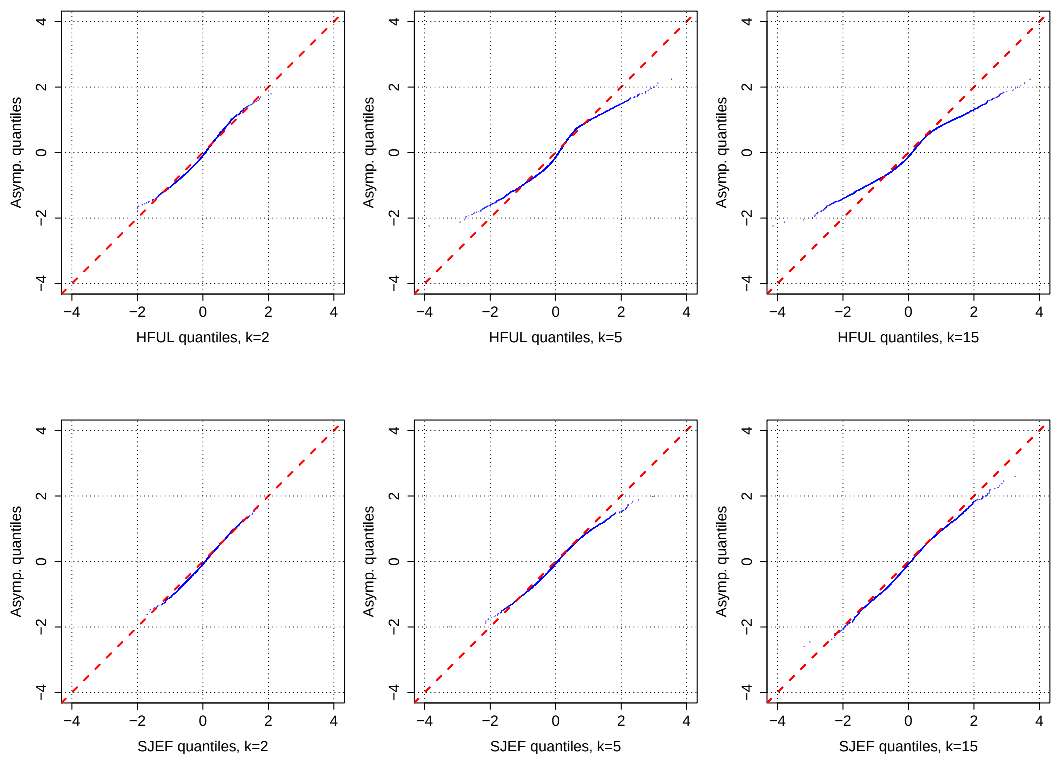

Figure 3: $R^2_{\varepsilon^2|z} = 0.2$, $\mu^2 = 8$: QQ plots for quantiles of HFUL with $c = 1$ and SJEF with $\alpha = 2$ based on 5,000 replications against quantiles of their asymptotic normal approximations.

Concerning the accuracy of the asymptotic approximations we find that in particular HFUL is more spread out than its asymptotic approximation if $k > 2$.[9] This holds for both strong and weak instruments. Figure 3 gives QQ plots for the heteroskedastic case,

[9] Kunitomo (2012) noticed the difference between the finite-sample distribution of HLIM and its asymptotic approximation as well.

$R^2_{\varepsilon^2|z} = 0.2$, with weak instruments, $\mu^2 = 8$. It plots quantiles of the empirical distributions of HFUL with $c = 1$ and SJEF with $\alpha = 2$, based on 5000 replications, against quantiles of the normal distribution with a variance given by element (2,2) of the covariance matrix $\{\boldsymbol{H}(\boldsymbol{C}^*) + f\boldsymbol{\Sigma}^*\}^{-1}\{\boldsymbol{F}(\boldsymbol{C}^*) + \boldsymbol{G}(\boldsymbol{C}^*, \boldsymbol{\Sigma}^*)\}\{\boldsymbol{H}(\boldsymbol{C}^*) + f\boldsymbol{\Sigma}^*\}^{-1}$, where the matrix functions have been defined in (32)-(34), and for SJEF we used $\boldsymbol{C}^* = \boldsymbol{C}$, $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}$ and $f = \alpha = 2$. For HFUL we used $\boldsymbol{C}^* = \boldsymbol{P} - \boldsymbol{D}$, $\boldsymbol{\Sigma}^* = n^{-1}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i$ and $f = nc/(n-c)$, where $c = 1$. For the choice $c = \alpha = 0$ the asymptotic normal approximation has a larger variance but still the empirical distribution of HFUL has fatter tails than the normal. For SJEF we find the asymptotic approximation is more accurate.

7 Conclusion

We considered instrumental variable estimation that is robust against heteroskedasticity. A new estimator has been based on a method-of-moments reasoning and interpreted as a symmetric jackknife estimator. It preserves the signal component in the data. Asymptotic theory based on high level assumptions, which allow for both many instruments and many weak instruments, resulted in a concise formulation of asymptotic distributions and standard errors. We found a smaller signal component in the asymptotic variance when compared to estimators that neglect a part of the signal to achieve consistency. For the noise component such ordering was not found. Sufficient primitive conditions were given as well. A Monte Carlo comparison with the HFUL estimator of Hausman et al. (2012) showed the Fuller modification of symmetric jackknife estimator performs better in terms of trade-off between bias and spread. It showed as well the asymptotic approximation of the finite-sample distribution of the symmetric jackknife estimator is more accurate.

8 Appendix

8.1 Derivation of Theorem 1

To derive Theorem 1 we use the notation $\boldsymbol{\delta} = (1, -\boldsymbol{\beta}')'$, $\boldsymbol{M} = (\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{A}(\boldsymbol{y}, \boldsymbol{X})$ and $\boldsymbol{S} = (\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{B}(\boldsymbol{y}, \boldsymbol{X})$. So $(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{C}\boldsymbol{\varepsilon} = (\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta}$. We find

$$
\begin{aligned}
&\mathrm{E}(\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta} = \boldsymbol{0}, && (36)\\
&\mathrm{Var}\left\{(\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta}\right\} = \mathrm{E}\left\{(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{C}(\boldsymbol{y}, \boldsymbol{X})\boldsymbol{\delta}\boldsymbol{\delta}'(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{C}(\boldsymbol{y}, \boldsymbol{X})\right\}\\
&= \mathrm{E}\sum_{i=1}^{n}\sum_{j=1}^{n}(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{C}\boldsymbol{e}_i\boldsymbol{e}_i'(\boldsymbol{y}, \boldsymbol{X})\boldsymbol{\delta}\boldsymbol{\delta}'(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{e}_j\boldsymbol{e}_j'\boldsymbol{C}(\boldsymbol{y}, \boldsymbol{X})\\
&= \mathrm{E}\sum_{i}\varepsilon_i^2(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{C}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}(\boldsymbol{y}, \boldsymbol{X}) + \mathrm{E}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_i\varepsilon_j C_{ij}^2(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{e}_j\boldsymbol{e}_i'(\boldsymbol{y}, \boldsymbol{X}), && (37)\\
&= \sum_{i=1}^{n}\sigma_i^2\begin{pmatrix}\boldsymbol{\beta}'\\ \boldsymbol{I}_g\end{pmatrix}\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}, \boldsymbol{I}_g) + \sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2\left(\sigma_j^2\boldsymbol{\Omega}_i + \boldsymbol{\Omega}_i\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Omega}_j\right), && (38)
\end{aligned}
$$

where we use the notation $C_{ij}^2 = (C_{ij})^2$. Using Assumption 2 we find $\boldsymbol{\Omega}_i \leq (1 + \boldsymbol{\delta}'\boldsymbol{\delta})c_u^2\boldsymbol{I}_{g+1}$ and

$$
\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2\left(\sigma_j^2\boldsymbol{\Omega}_i + \boldsymbol{\Omega}_i\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Omega}_j\right) &\leq 2c_u^2(1 + \boldsymbol{\delta}'\boldsymbol{\delta})\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2\boldsymbol{I}_{g+1}\\
&= 2c_u^2(1 + \boldsymbol{\delta}'\boldsymbol{\delta})\,\mathrm{tr}(\boldsymbol{C}^2)\boldsymbol{I}_{g+1},
\end{aligned}
$$

which is of order $O(k)$, since $C_{ii} = 0$ and, for $i \neq j$,

$$
C_{ij}^2 = P_{ij}^2\{(1 - h_i)^{-1} + (1 - h_j)^{-1}\}^2 \leq 4c_u^2P_{ij}^2 \tag{39}
$$

for $i, j = 1, \ldots, n$ by Assumption 1 and $\mathrm{tr}(\boldsymbol{P}^2) = k$. Consequently,

$$
\boldsymbol{\delta}'(\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta} = O_p(k^{1/2}). \tag{40}
$$

For the first term we find

$$
\begin{aligned}
\sum_{i=1}^{n} \sigma_i^2 \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_g \end{pmatrix} \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}, \boldsymbol{I}_g) &\le c_u \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_g \end{pmatrix} \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}^2\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}, \boldsymbol{I}_g) \\
&= c_u \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_g \end{pmatrix} \boldsymbol{\Pi}'\boldsymbol{Z}' \left\{ \boldsymbol{I}_n + \frac{1}{4}\boldsymbol{D}^2(\boldsymbol{I}_n - \boldsymbol{D})^{-2} \right\} \boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}, \boldsymbol{I}_g) \\
&\le c_u \left(1 + \frac{c_u^2}{4}\right) \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_g \end{pmatrix} \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\beta}, \boldsymbol{I}_g),
\end{aligned}
$$

so by Assumption 5, where $k/r_{\min} \to \gamma$ we find, using $\boldsymbol{H} = \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Pi}$ as in (23),

$$
\boldsymbol{H}^{-1/2}(\boldsymbol{0}, \boldsymbol{I}_g)(\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta} = O_p(1). \tag{41}
$$

The first derivative of the objective function is given by

$$
\begin{aligned}
\frac{\partial Q^*_{\text{SJIVE}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2 \left\{ \frac{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}}{\operatorname{tr}(\boldsymbol{B})} \right\}^{-1} (\boldsymbol{0}, \boldsymbol{I}_g) \left\{ \boldsymbol{M}\boldsymbol{\delta} - \left(\frac{\boldsymbol{\delta}'\boldsymbol{M}\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} \right) \boldsymbol{S}\boldsymbol{\delta} \right\} \\
&= -2 \left\{ \frac{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}}{\operatorname{tr}(\boldsymbol{B})} \right\}^{-1} (\boldsymbol{0}, \boldsymbol{I}_g) \left(\boldsymbol{I}_{g+1} - \frac{\boldsymbol{S}\boldsymbol{\delta}\boldsymbol{\delta}'}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} \right) (\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta}.
\end{aligned}
$$

Using Assumptions 3 and 5 we find,

$$
\begin{aligned}
-\frac{1}{2} \left\{ \frac{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}}{\operatorname{tr}(\boldsymbol{B})} \right\} \boldsymbol{H}^{-1/2} \frac{\partial Q^*_{\text{SJIVE}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \boldsymbol{H}^{-1/2} \left\{ (\boldsymbol{0}, \boldsymbol{I}_g) - \sigma^{-2}\boldsymbol{\sigma}_{21}\boldsymbol{\delta}' \right\} (\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta} \quad - \\
&\left(\frac{k}{r_{\min}} \right)^{1/2} \left(\frac{\boldsymbol{H}}{r_{\min}} \right)^{-1/2} \left\{ \frac{(\boldsymbol{0}, \boldsymbol{I}_g)\boldsymbol{S}\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} - \frac{\boldsymbol{\sigma}_{21}}{\sigma^2} \right\} k^{-1/2}\boldsymbol{\delta}'(\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta} \\
&= \boldsymbol{H}^{-1/2} \left\{ (\boldsymbol{0}, \boldsymbol{I}_g) - \sigma^{-2}\boldsymbol{\sigma}_{21}\boldsymbol{\delta}' \right\} (\boldsymbol{M} - \boldsymbol{S})\boldsymbol{\delta} + o_p(1).
\end{aligned} \tag{42}
$$

The second derivative of the objective function is given by

$$
\begin{aligned}
\frac{\partial^2 Q^*_{\text{SJIVE}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= 2 \left\{ \frac{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}}{\operatorname{tr}(\boldsymbol{B})} \right\}^{-1} (\boldsymbol{0}, \boldsymbol{I}_g) \left(2\frac{\boldsymbol{S}\boldsymbol{\delta}\boldsymbol{\delta}'}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} - \boldsymbol{I}_{g+1} \right) \left\{ \boldsymbol{M} - \left(\frac{\boldsymbol{\delta}'\boldsymbol{M}\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} \right) \boldsymbol{S} \right\} \quad \times \\
&\qquad \left(2\frac{\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{S}}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} - \boldsymbol{I}_{g+1} \right) \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I}_g \end{pmatrix} \\
&= 2 \left\{ \frac{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}}{\operatorname{tr}(\boldsymbol{B})} \right\}^{-1} (\boldsymbol{0}, \boldsymbol{I}_g)(\boldsymbol{M} - \boldsymbol{S} + \boldsymbol{R}) \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I}_g \end{pmatrix},
\end{aligned}
$$

where

$$\boldsymbol{R} = -2(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta}\frac{\boldsymbol{\delta}'\boldsymbol{S}}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} - 2\frac{\boldsymbol{S}\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}}\boldsymbol{\delta}'(\boldsymbol{M}-\boldsymbol{S}) - \boldsymbol{\delta}'(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta}\left(\frac{\boldsymbol{S}}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} - 4\frac{\boldsymbol{S}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{S}}{(\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta})^2}\right).$$

Due to (40) and (41) we find

$$\boldsymbol{H}^{-1/2}(\boldsymbol{0},\boldsymbol{I}_g)\boldsymbol{R}\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{I}_g\end{pmatrix}\boldsymbol{H}^{-1/2} = o_p(1).$$

Consequently

$$\frac{1}{2}\left\{\frac{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}}{\mathrm{tr}(\boldsymbol{B})}\right\}\boldsymbol{H}^{-1/2}\frac{\partial^2 Q^*_{\mathrm{SJIVE}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\boldsymbol{H}^{-1/2} = \boldsymbol{H}^{-1/2}(\boldsymbol{0},\boldsymbol{I}_g)\left\{\boldsymbol{M}-\boldsymbol{S}\right\}\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{I}_g\end{pmatrix}\boldsymbol{H}^{-1/2} + o_p(1).$$

Based on Assumption 3 we thus find

$$\begin{aligned}\boldsymbol{H}^{-1/2}\frac{\partial^2 Q^*_{\mathrm{SJIVE}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\boldsymbol{H}^{-1/2} &= 2\sigma^{-2}\left\{\boldsymbol{H}^{-1/2}\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\boldsymbol{H}^{-1/2}\right\} + o_p(1)\\ &= 2\sigma^{-2}\left\{\boldsymbol{H}^{-1/2}\,\mathrm{E}(\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X})\boldsymbol{H}^{-1/2}\right\} + o_p(1) = 2\sigma^{-2}\boldsymbol{I}_g + o_p(1). \qquad (43)\end{aligned}$$

So, $\boldsymbol{\Xi} = 2\sigma^{-2}\boldsymbol{I}_g$. Assumption 4 says $r_{\min}\to\infty$, so (43) implies that $\lambda_{\min}\left(\frac{\partial^2 Q^*_{\mathrm{SJIVE}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)\to\infty$, and by Assumption 5 we find $\hat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}$. Finally we find, applying (38), (42) and 43,

$$\begin{aligned}\boldsymbol{\Psi} = \lim_{n\to\infty}\Bigg[\boldsymbol{H}^{-1/2}\Bigg\{&\sum_{i=1}^{n}\sigma_i^2\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi}\\ &+\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^2\left(-\frac{\boldsymbol{\sigma}_{21}}{\sigma^2},\boldsymbol{I}_g\right)\left(\sigma_j^2\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_i\boldsymbol{e}_1\boldsymbol{e}_1'\boldsymbol{\Sigma}_j\right)\left(-\frac{\boldsymbol{\sigma}_{21}}{\sigma^2},\boldsymbol{I}_g\right)'\Bigg\}\boldsymbol{H}^{-1/2}\Bigg].\end{aligned}$$

8.2 Derivation of Theorem 2

To compute standard errors, we estimate $\mathrm{Var}\left\{(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta}\right\}$ using (37). As $\hat{\varepsilon}_i = \varepsilon_i - \boldsymbol{Z}_i\boldsymbol{\Pi}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \boldsymbol{V}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ and $\boldsymbol{H}^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = O_p(1)$, $r_{\min}\to\infty$ and $\boldsymbol{\Omega}_i = O(1)$, we may use $\hat{\varepsilon}_i$ instead of ε_i if $\max_{1\le i\le n}\|\boldsymbol{Z}_i\boldsymbol{\Pi}\boldsymbol{H}^{-1/2}\| = o(1)$, which is guaranteed by Assumption

6. A consistent estimator for $\boldsymbol{\Psi}$ is then given by

$$\widehat{\boldsymbol{\Psi}} = (\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1/2}\left\{(\boldsymbol{0}, \boldsymbol{I}_g) - \hat{\sigma}^{-2}\hat{\boldsymbol{\sigma}}_{21}\hat{\boldsymbol{\delta}}'\right\}\widehat{\mathrm{Var}}\left[(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta}\right] \quad \times$$
$$\left\{(\boldsymbol{0}, \boldsymbol{I}_g)' - \hat{\sigma}^{-2}\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\sigma}}_{12}\right\}(\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1/2},$$

where

$$\widehat{\mathrm{Var}}\left[(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta}\right] = \sum_i \hat{\varepsilon}_i^2(\boldsymbol{y}, \boldsymbol{X})'\widehat{\boldsymbol{C}}\boldsymbol{e}_i\boldsymbol{e}_i'\widehat{\boldsymbol{C}}(\boldsymbol{y}, \boldsymbol{X}) \quad +$$
$$\sum_{i=1}^n\sum_{j=1}^n \hat{\varepsilon}_i\hat{\varepsilon}_j(\widehat{C}_{ij})^2(\boldsymbol{y}, \boldsymbol{X})'\boldsymbol{e}_j\boldsymbol{e}_i'(\boldsymbol{y}, \boldsymbol{X}).$$

The estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1}\left\{(\boldsymbol{0}, \boldsymbol{I}_g) - \hat{\sigma}^{-2}\hat{\boldsymbol{\sigma}}_{21}\hat{\boldsymbol{\delta}}'\right\}\widehat{\mathrm{Var}}\left[(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta}\right] \quad \times \\ &\qquad \left\{(\boldsymbol{0}, \boldsymbol{I}_g)' - \hat{\sigma}^{-2}\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\sigma}}_{12}\right\}(\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1}(\boldsymbol{X} - \hat{\sigma}^{-2}\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\sigma}}_{12})'\left(\widehat{\boldsymbol{C}}\boldsymbol{D}_{\hat{\boldsymbol{\varepsilon}}}^2\widehat{\boldsymbol{C}} + \boldsymbol{D}_{\hat{\boldsymbol{\varepsilon}}}\widehat{\boldsymbol{C}}^{(2)}\boldsymbol{D}_{\hat{\boldsymbol{\varepsilon}}}\right)(\boldsymbol{X} - \hat{\sigma}^{-2}\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\sigma}}_{12})(\boldsymbol{X}'\widehat{\boldsymbol{C}}\boldsymbol{X})^{-1},\end{aligned}$$

which is (22).

8.3 Derivation of Theorem 3

Instead of (42) we now have, using Assumptions 3 and 7,

$$\begin{aligned}-\frac{k^{-1/2}}{2}\left\{\frac{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}}{\mathrm{tr}(\boldsymbol{B})}\right\}\frac{\partial Q^*_{\text{SJIVE}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= \\ & k^{-1/2}\left\{(\boldsymbol{0}, \boldsymbol{I}_g) - \sigma^{-2}\boldsymbol{\sigma}_{21}\boldsymbol{\delta}'\right\}(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta} \quad - \\ & \left\{\frac{(\boldsymbol{0}, \boldsymbol{I}_g)\boldsymbol{S}\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}} - \frac{\boldsymbol{\sigma}_{21}}{\sigma^2}\right\}k^{-1/2}\boldsymbol{\delta}'(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta} \\ &= k^{-1/2}\left\{(\boldsymbol{0}, \boldsymbol{I}_g) - \sigma^{-2}\boldsymbol{\sigma}_{21}\boldsymbol{\delta}'\right\}(\boldsymbol{M}-\boldsymbol{S})\boldsymbol{\delta} + o_p(1). \qquad (44)\end{aligned}$$

Using (38) we thus find

$$\begin{aligned}\boldsymbol{\Psi}_w = \lim_{n\to\infty} k^{-1} \Bigg\{ &\sum_{i=1}^{n} \sigma_i^2 \boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi} \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2 \left(-\frac{\sigma_{21}}{\sigma^2}, \boldsymbol{I}_g\right)\left(\sigma_j^2\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_i\boldsymbol{e}_1\boldsymbol{e}_1'\boldsymbol{\Sigma}_j\right)\left(-\frac{\sigma_{21}}{\sigma^2}, \boldsymbol{I}_g\right)'\Bigg\} \\ = \lim_{n\to\infty} k^{-1} &\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}^2 \left(-\frac{\boldsymbol{\sigma}_{21}}{\sigma^2}, \boldsymbol{I}_g\right)\left(\sigma_j^2\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_i\boldsymbol{e}_1\boldsymbol{e}_1'\boldsymbol{\Sigma}_j\right)\left(-\frac{\boldsymbol{\sigma}_{21}}{\sigma^2}, \boldsymbol{I}_g\right)'.\end{aligned}$$

As (43) remains valid under Assumption 7 we find the result of Theorem 2, where $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ since $r_{\min}/k^{1/2} \to \infty$.

8.4 Primitive assumptions

In order to arrive at asymptotic distributions for their estimators Hausman et al. (2012) and Kunitomo (2012) considered quite primitive assumptions. They applied martingale central limit theorems for quadratic forms as used by Chao et al. (2012) and Anderson et al. (2007), respectively. Here we consider the conditions of specification (27), which are very similar to the ones used by Hausman et al. (2012). The only differences are that their matrix $\boldsymbol{P} - \boldsymbol{D}$ is replaced by our matrix $\boldsymbol{C}$, and their vector $\sum_i \boldsymbol{\sigma}_{21i}/\sum_i \sigma_i^2$ is replaced by our $\boldsymbol{\sigma}_{21}/\sigma^2$.

To verify this, observe that $r_{\min}$ and $\mu_n^2 = \min_{1\leq i\leq g} \mu_{in}^2$ have the same growth rate. Furthermore, if $\boldsymbol{z}_i' = \boldsymbol{e}_i'\boldsymbol{Z}\tilde{\boldsymbol{\Pi}}$, then the existence of the positive definite limit $\widetilde{\boldsymbol{H}}$ implies the existence of c_u such that $\|\sum_i \boldsymbol{z}_i\boldsymbol{z}_i'/n\| \leq c_u$ and $\lambda_{\min}(\sum_i \boldsymbol{z}_i\boldsymbol{z}_i'/n) \geq 1/c_u$ for n sufficiently large, and $\sum_i \|\boldsymbol{z}_i\|^4/n^2 \to 0$, as required.[10]

The asymptotic normality of Assumption 5 can now be verified by following the steps made in Hausman et al. (2012) to derive the asymptotic normality of HLIM. Care should be taken when the martingale central limit theorem is applied. That is to say, Lemma A6

[10]The convergence $\sum_i \|\boldsymbol{z}_i\|^4/n^2 \to 0$ is implied by the convergence $\sum_i\{\boldsymbol{z}_i'(\sum_i \boldsymbol{z}_i\boldsymbol{z}_i'/n)^{-1}\boldsymbol{z}_i\}^2/n^2 \to 0$, since $\sum_i \boldsymbol{z}_i\boldsymbol{z}_i'/n \to \widetilde{\boldsymbol{H}} > 0$. Furthermore, $\sum_i\{\boldsymbol{z}_i'(\sum_i \boldsymbol{z}_i\boldsymbol{z}_i'/n)^{-1}\boldsymbol{z}_i\}^2/n^2 = \sum_i d_i^2$, where d_i is the ith diagonal element of the projection matrix $\boldsymbol{P}_{Z\tilde{\Pi}}$. As $\sum_i d_i^2 \leq g \max_{1\leq i\leq n} d_i$ and $\max_{1\leq i\leq n} d_i \to 0$ since $\sum_i \boldsymbol{z}_i\boldsymbol{z}_i'/n \to \widetilde{\boldsymbol{H}}$, we find $\sum_i \|\boldsymbol{z}_i\|^4/n^2 \to 0$.

in Hausman et al. (2012), which is Lemma A2 of Chao et al. (2012), is formulated explicitly in terms of the off-diagonal elements of $\boldsymbol{P}$. For our purposes the result should remain valid when the off-diagonal elements of $\boldsymbol{C}$ are used instead of $\boldsymbol{P}$. As $C_{ij} = P_{ij}\{1/(1-h_i) + 1/(1-h_j)\}/2$, we find as a result of Assumption 1, $C_{ij}^2 \leq P_{ij}^2 c_u^2$. Using this bound we can check the conditions of the martingale central limit theorem when P_{ij} is replaced by C_{ij} in the same way as is done in the proof of Lemma A2 of Chao et al. (2012). Thus Assumptions 1-4 and $k/r_{\min} \to \gamma$, together with the conditions of specification (27) are sufficient for Assumption 5.

8.5 Derivation of Theorem 4

To prove Theorem 4 let

$$\boldsymbol{H}_1 = (\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}^2\boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{\Pi}'\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{Z}\boldsymbol{\Pi})^{-1},$$
$$\boldsymbol{H}_2 = \{\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{P}-\boldsymbol{D})\boldsymbol{Z}\boldsymbol{\Pi}\}^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{P}-\boldsymbol{D})^2\boldsymbol{Z}\boldsymbol{\Pi}\{\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{P}-\boldsymbol{D})\boldsymbol{Z}\boldsymbol{\Pi}\}^{-1},$$

Using (35) we find

$$\boldsymbol{H}_1 = \boldsymbol{H}^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'\left\{\boldsymbol{I}_n + (1/4)(\boldsymbol{I}_n-\boldsymbol{D})^{-1}(\boldsymbol{I}_n-\boldsymbol{P})(\boldsymbol{I}_n-\boldsymbol{D})^{-1}\right\}\boldsymbol{Z}\boldsymbol{\Pi}\boldsymbol{H}^{-1},$$
$$\boldsymbol{H}_2 = \{\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n-\boldsymbol{D})\boldsymbol{Z}\boldsymbol{\Pi}\}^{-1}\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n-\boldsymbol{D})^2\boldsymbol{Z}\boldsymbol{\Pi}\{\boldsymbol{\Pi}'\boldsymbol{Z}'(\boldsymbol{I}_n-\boldsymbol{D})\boldsymbol{Z}\boldsymbol{\Pi}\}^{-1}.$$

Let $\boldsymbol{L} = \boldsymbol{Z}\boldsymbol{\Pi}\boldsymbol{H}^{-1/2}$, so $\boldsymbol{L}'\boldsymbol{L} = \boldsymbol{I}_g$ and $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{P}_{Z\Pi}$ is a projection matrix and

$$\boldsymbol{H}_1 = \boldsymbol{H}^{-1/2}\boldsymbol{L}'\left\{\boldsymbol{I}_n + (1/4)(\boldsymbol{I}_n-\boldsymbol{D})^{-1}(\boldsymbol{I}_n-\boldsymbol{P})(\boldsymbol{I}_n-\boldsymbol{D})^{-1}\right\}\boldsymbol{L}\boldsymbol{H}^{-1/2},$$
$$\boldsymbol{H}_2 = \boldsymbol{H}^{-1/2}\{\boldsymbol{L}'(\boldsymbol{I}_n-\boldsymbol{D})\boldsymbol{L}\}^{-1}\boldsymbol{L}'(\boldsymbol{I}_n-\boldsymbol{D})^2\boldsymbol{L}\{\boldsymbol{L}'(\boldsymbol{I}_n-\boldsymbol{D})\boldsymbol{L}\}^{-1}\boldsymbol{H}^{-1/2}.$$

After premultiplying by $\boldsymbol{L}'(\boldsymbol{I}_n - \boldsymbol{D})\boldsymbol{L}\boldsymbol{H}^{1/2}$, and postmultiplying by its transpose, we find $\boldsymbol{H}_1 \leq \boldsymbol{H}_2$ if and only if

$$0 \leq \boldsymbol{L}'(\boldsymbol{I}_n - \boldsymbol{D})\left\{\boldsymbol{I}_n - \boldsymbol{P}_{Z\Pi} - (1/4)\boldsymbol{P}_{Z\Pi}(\boldsymbol{I}_n - \boldsymbol{D})^{-1}(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\boldsymbol{P}_{Z\Pi}\right\}(\boldsymbol{I}_n - \boldsymbol{D})\boldsymbol{L} =$$
$$\boldsymbol{L}'(\boldsymbol{I}_n - \boldsymbol{D})(\boldsymbol{I}_n - \boldsymbol{P}_{Z\Pi})\left\{\boldsymbol{I}_n - (1/4)(\boldsymbol{I}_n - \boldsymbol{D})^{-1}(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{I}_n - \boldsymbol{D})^{-1}\right\}(\boldsymbol{I}_n - \boldsymbol{P}_{Z\Pi})(\boldsymbol{I}_n - \boldsymbol{D})\boldsymbol{L},$$

where the equality follows from $(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{L} = \mathbf{O}$. This inequality is implied by $\boldsymbol{D} < (1/2)\boldsymbol{I}_n$ since, if it holds,

$$\boldsymbol{I}_n - (1/4)(\boldsymbol{I}_n - \boldsymbol{D})^{-1}(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{I}_n - \boldsymbol{D})^{-1} \geq \boldsymbol{I}_n - (1/4)(\boldsymbol{I}_n - \boldsymbol{D})^{-2} > 0,$$

which establishes $\boldsymbol{H}_1 \leq \boldsymbol{H}_2$ as required. Furthermore, the rank of the difference is given by $\operatorname{rank}\{(\boldsymbol{I}_n - \boldsymbol{P}_{Z\Pi})(\boldsymbol{I}_n - \boldsymbol{D})\boldsymbol{L}\} = \operatorname{rank}\{(\boldsymbol{I}_n - \boldsymbol{P}_{Z\Pi})\boldsymbol{D}\boldsymbol{Z}\boldsymbol{\Pi}\}$.

References

Ackerberg, D. A. & Devereux, P. J. (2003). Improved JIVE estimators for overidentified linear models with and without heteroskedasticity. *Working paper*.

Anderson, T. W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics*, *127*, 1–16.

Andrews, D. W. K. & Stock, J. H. (2007). Testing with many weak instruments. *Journal of Econometrics*, *138*, 24–46.

Angrist, J. D., Imbens, G. W. & Krueger, A. (1999). Jackknife instrumental variable estimators. *Journal of Applied Econometrics*, *14*, 57–67.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, *54*, 657–682.

Bekker, P. A. & Van der Ploeg, J. (2005). Instrumental variable estimation based on grouped data. *Statistica Neerlandica*, *59*, 239–267.

Blomquist, S. & Dahlberg, M. (1999). Small sample properties of LIML and Jackknife IV estimators: experiments with weak instruments. *Journal of Applied Econometrics*, *14*, 96–88.

Chamberlain, G. & Imbens, G. (2004). Random effects estimators with many instrumental variables. *Econometrica*, *72*, 295–306.

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R. & Woutersen, T. (2012). Combining two consistent estimators. In B. H. Baltagi, R. Carter Hill, W. K. Newey & H. L. White (Eds.), *Advances in econometrics: Essays in honor of Jerry Hausman* (Vol. 29). Emerald Group Publishing, UK.

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R. & Woutersen, T. (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics*, *178*, 15–21.

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

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K. & Woutersen, T. (2012). Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory*, *28*, 42–86.

Chioda, L. & Jansson, M. (2009). Optimal invariant inference when the number of instruments is large. *Econometric Theory*, *25*, 793–805.

Davidson, R. & MacKinnon, J. G. (2006). The case against JIVE. *Journal of Applied Econometrics*, *21*, 827–833.

Donald, S. J. & Newey, W. K. (2000). A Jackknife interpretation of the continuous updating estimator. *Economics Letters*, *67*, 239–243.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica*, *45*, 939–954.

Hahn, J. (2002). Optimal inference with many instruments. *Econometric Theory*, *18*, 140–168.

Hahn, J., Hausman, J. A. & Kuersteiner, G. M. (2004). Estimation with weak instruments:

accuracy of higher-order bias and MSE approximations. *Econometrics Journal*, *7*, 272–306.

Hahn, J. & Inoue, A. (2002). A Monte Carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators. *Econometric Reviews*, *21*, 309–336.

Han, C. & Phillips, P. C. B. (2006). GMM with many moment conditions. *Econometrica*, *74*, 147–192.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.

Hansen, L. P., Heaton, J. & Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics*, *14*, 398–422.

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

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

Kunitomo, N. (2012). An optimal modification of the liml estimation for many instruments and persistent heteroskedasticity. *Annals of the Institute of Statistical Mathematics*, *64*, 881–910.

Morimune, K. (1983). Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica*, *51*, 821–842.

Newey, W. K. & Windmeijer, F. (2009). Generalized method of moments with many weak moment condition. *Econometrica*, *77*, 687–719.

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

Stock, J. & Yogo, M. (2005). Asymptotic distributions of instrumental variables statistics with many instruments. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: essays in honor of Thomas Rothenberg* (chap. 6). Cambridge University Press.

Van der Ploeg, J. & Bekker, P. A. (1995). Efficiency bounds for instrumental variable estimators under group asymptotics. *Working paper, University of Groningen*.

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