Many instruments and/or regressors: a friendly guide

STANISLAV ANATOLYEV*
CERGE-EI and New Economic School

September 2018

Abstract

This article surveys the state of the art in the econometrics of regression models with many instruments or many regressors based on alternative – namely, dimension – asymptotics. We list critical results of dimension asymptotics that lead to better approximations of properties of familiar and alternative estimators and tests when the instruments and/or regressors are numerous. Then we consider the problem of estimation and inference in the basic linear instrumental variables regression setup with many strong instruments. We describe the failures of conventional estimation and inference, as well as alternative tools that restore consistency and validity. We then add various other features to the basic model such as heteroskedasticity, instrument weakness, etc., in each case providing a review of the existing tools for proper estimation and inference. Subsequently, we consider a related but different problem of estimation and testing in a linear mean regression with many regressors. We also describe various extensions and connections to other settings, such as panel data models, spatial models, time series models, and so on. Finally, we provide practical guidance regarding which tools are most suitable to use in various situations when many instruments and/or regressors turn out to be an issue.

KEYWORDS: linear regression, instrumental variables, alternative asymptotic theory, dimension asymptotics, many instruments, weak instruments, many regressors.

JEL CODES: C12, C13, C18, C21, C26, C55

*Addresses: CERGE-EI, Politických vězňů 7, 11121 Prague 1, Czech Republic, stanislav.anatolyev@cerge-ei.cz; New Economic School, 45 Skolkovskoe shosse, Moscow, 121353, Russia, sanatoly@nes.ru. This research was supported by the grant 17-26535S from the Czech Science Foundation and Access Industries professorship at the NES. I thank the Associate Editor and two anonymous expert referees for constructive comments and useful suggestions. I also thank seminar audiences at the CERGE-EI, NES, UNSW, University of Sydney, and University of Hong Kong.

1 Introduction

In some instrumental variables (IV) models, there is abundance of instruments to exploit, so their total number is often comparable to a sample size. Because the traditional asymptotic theory presumes their small number, classical IV procedures, which rely on this presumption, may exhibit various distortions when it fails. Consider an illustrative empirical application from Hahn and Hausman (2002, Section 6) where a structural coefficient (price elasticity) in a certain demand equation is estimated, with the number of observations equalling 1452. Five different estimators are used in two different scenarios: one where only 2 excluded instruments are used, and one where 132 excluded instruments are used. These five estimators are asymptotically equivalent from the point of view of the standard asymptotic theory. The figures are reported in Table 1.

	estimate_1	estimate_2	estimate_3	estimate_4	estimate_5
2 excluded instruments	-2.03	-2.31	-2.03	-2.31	-2.05
	(0.465)	(0.500)	(0.465)	(0.502)	(0.469)
132 excluded instruments	-1.24	-8.01	-1.21	-4.64	-1.18
	(0.194)	(0.789)	(0.194)	(0.293)	(0.211)

Table 1. Various estimates of price elasticity in the Hahn and Hausman (2002) application.

While the estimates and their standard errors are indeed numerically similar in the case of a couple of excluded instruments, some of them are way off when many more instruments are used. In fact, the discrepancies are so large that any conclusion from the conventional asymptotics in this example is immediately compromised. Obviously, a different theory is needed that would discriminate properties of these estimators.

The research on many instruments (as well as research on weak instruments and a combination thereof) was catalyzed by the paradigmatic Angrist and Krueger (1991) returns to schooling application, which uses quarter-of-birth effects and their interactions with other variables as instruments. Since the publication of this article, this research area has been recognized as an important and empirically relevant subfield. Paul Bekker pioneered this research direction with a seminal article (1994), where he proposed new standard errors that account for the numerosity of instruments using dimension asymptotics with the number of instruments increasing proportionally with sample size, although the machinery of dimension asymptotics in the present context was used earlier in, e.g., Kunitomo (1980) and Morimune (1983) to compare the properties of familiar estimators.[1] Starting from Chao and Swanson (2005), the many instrument asymptotic framework has blended with the weak instrument asymptotics (Staiger and Stock, 1997), which has also been an important step given the weakness of big instrument sets in practice (including the same Angrist and Krueger (1991) application). Imbens (2010) notes that in empirically relevant situations,

[1] Kunitomo (1980) explores the properties of 2SLS and LIML estimators in a simplified model and finds that as the degree of overidentification becomes comparable with the sample size, 2SLS becomes extremely inferior to LIML. Morimune (1983) derives asymptotic expansions of k-class estimators (Nagar, 1959) in the many instrument framework and finds that the 2SLS estimator is inconsistent while the LIML estimator and its Fuller's (1977) modification are consistent but have higher variance than under the large sample asymptotics.

it is the numerosity of instruments that is more responsible for misleading 2SLS inferences than the instruments' weakness. In the early 2010s, there was a burst of research on many (possibly weak) instruments in conditionally heteroskedastic models, the main contributions being made in Chao, Swanson, Hausman, Newey and Woutersen (2012), Hausman, Newey, Woutersen, Chao and Swanson (2012), and Chao, Hausman, Newey, Swanson and Woutersen (2014). Simultaneously, an interest in many regressors econometrics – a different but related research area – has also gone from inference in a simple homoskedastic regression model (Calhoun, 2011; Anatolyev, 2012) to a general conditionally heteroskedastic model with many covariates (Cattaneo, Jansson and Newey, 2018a, 2018b). Some literature (Anatolyev, 2013; Kolesár, 2018) considers models where both features – many instruments and many covariates – are present at the same time.

More than a quarter of a century has passed since the publication of the Angrist and Krueger (1991) application, and more than two decades since the theoretical work of Bekker (1994), but empirical microeconomists are still largely unaware of how to correctly utilize a host of available instruments or to properly handle large regressor sets. This survey aims to close the gap. The survey is designed to provide empirical researchers and graduate-level econometrics students with an overview of the field and concrete instructions for how to use relevant tools.

2 Structure of survey

Most recent theoretical literature tends to deal with all of the possible complications involved, and thus the degree of sophistication can become a stumbling block for applied researchers. The survey is structured so that complications are gradually scaffolded on an easily comprehensible skeleton model – a linear homoskedastic instrumental variables model. In this section, we describe a roadmap for how the survey is organized.

Section 3 presents the statistical foundations for many instrument/regressor asymptotics; an uninterested reader may skip this section during a first reading and jump to Section 4. First, we reiterate the idea of alternative asymptotic approximations, in particular, dimension asymptotics. Second, the historical evolution of limit theorems for bilinear forms – the main building block of many instrument/regressor asymptotics – is shown, starting with an example of a simplest quadratic form, to show the differences between the approaches – the traditional asymptotics and dimension asymptotics – and how they interlock. Third, a certain important asymptotic property of the diagonal of the projection matrix is discussed.

In Section 4, a linear homoskedastic instrumental variables model is considered in detail. The reasons why 2SLS estimation fails and various remedies for such failures (bias correction, jackknifing, and LIML) are discussed. The alternative asymptotic distributions of consistent estimators are presented, initially for the case of normal errors. Then, we relax error normality to examine the arising complications and discuss how they can be handled. In addition to inferential procedures regarding parameter restrictions, the issues of asymptotic efficiency and specification testing are considered. Finally, we tackle another type of complication that arises when many exogenous regressors (covariates) are present in addition to many instruments.

A major complication in the present context, unlike the ones mentioned above, is conditional heteroskedasticity, which leads to a breakdown of the consistency of LIML, a leading estimator under conditional homoskedasticity. Section 5 reviews the problem and the proposed ways of solving it, the leading method being based on mixing LIML with jackknifing. Estimation and inference procedures robust to conditional heteroskedasticity are described.

In applied microeconometric problems, the numerosity of instruments is often accompanied by their weakness as a group. Section 6 reviews the trade-offs between instrument numerosity and their strength, as well as how they affect the rate of convergence, asymptotic distribution and ultimately asymptotic inference.

Section 7 focuses on a related but different setup. A similar asymptotic analysis is needed in mean regressions with a significant number of regressors that is possibly comparable to the sample size. Such problems call for the many regressor asymptotics, a tool that works similarly to the many instrument machinery. Hypothesis testing and parameter estimation are quite distinct problems in this setting. The relevant literature and methods for solving both problems are reviewed in this section.

Section 8 briefly surveys various further modifications of the basic instrumental variables and regression models and procedures, such as proposed alternative estimators, bootstrap inference, nonlinearity, as well as their extensions to particular settings such as panel data, spatial data, robust regression and time series models. A connection to an adjacent subfield, very many instruments/regressors, is also briefly highlighted.

Section 9 provides discussion and practical guidance regarding which tools are most suitable to use in situations when many instruments and/or regressors turn out to be an issue, as well as in situations outside the scope of this survey. This section is meant to help empirical microeconomists with selecting the most appropriate route in handling their problems. Finally, Section 10 provides a view on perspectives of the field and concludes.

By default, we use the term 'many instruments' to indicate dimension asymptotics, even in cases when one should in fact use the term 'many regressors' or 'many covariates.'

3 Asymptotic tools

This section presents the statistical foundations for many instrument/regressor asymptotics. An uninterested reader may skip it during a first reading and jump directly to Section 4 for the many instrument framework or to Section 7 for the many regressor framework.

3.1 Alternative and dimension asymptotics

Theoretical econometricians realized a while ago that traditional econometric tools often generate poor inferences due to newly discovered peculiarities of data and that they are not able to handle tasks that are rendered more challenging by wider data availability, more sophisticated modeling, and accumulated empirical experience. Alternative asymptotic theories, in particular, have become increasingly popular in recent decades among both theorists and practitioners. The coverage of such theories is broad: regression models and instrumental variable models, unit root and cointegration models, problems of long-run forecasting, post-selection inference, and others; see a review in Anatolyev and Gospodinov (2011b, Chapter 6). The usefulness of alternative asymptotic theories for a careful practitioner has been forcefully emphasized, among others, by Bekker (1994) in the context of models with many instruments and by Leeb and Pötscher (2005) in the context of post-selection inference. In a nutshell, the proceedings of an alternative asymptotic theory allow one to more adequately approximate the actual distributions of estimators and test statistics than the traditional large sample asymptotic theory is able to do.

Broadly speaking, there are two types of alternative asymptotic theories – drifting parameter asymptotics and dimension asymptotics – plus mixtures of the two. Drifting parameters are designed to acknowledge the locality of some parameters to their focal values, e.g., of the autoregressive parameter to its unit root value (near unit roots), of the correlation coefficient between the instrument and the regressor to zero (weak instruments), of the alternative hypothesis to the null (local test power), of a parameter of a competing model to zero (post-selection inference), and so on, and are of high importance in modern econometrics.

Dimension asymptotics serve the same purpose but deal with cases when the dimensionality of some parameters is large and possibly comparable to the sample size (although we exclude cases when dimensionality is not high enough to change the conventional conclusions). The idea of dimension asymptotics is rumored to belong to Kolmogorov who, in the context of discriminant analysis, proposed in 1967 an asymptotic approach in which the sample size increases along with the number of parameters, such that their ratio tends to a constant (Serdobolskii, 2007). Independently, similar asymptotics were being developed in the random matrix theory (Marchenko and Pastur, 1967), leading to a line of work on estimating covariance matrices in large dimensions (Yao, Zheng and Bai, 2015). For microeconometricians, many instruments and/or many regressor asymptotics featured in this survey are arguably the most important developments of dimension asymptotics.

3.2 Illustrative example

To see the differences in how the traditional and many instrument asymptotics work, let us consider the following quadratic form:

$$e'Pe = \sum_{i,j=1}^{n} P_{ij} e_i e_j,$$

where the error series $\{e_i\}_{i=1}^n$ collected in $n \times 1$ vector e is IID with zero mean and finite variance σ^2, and $P = Z\left(Z'Z\right)^{-1} Z'$ is $n \times n$ projection matrix associated with instruments $\{z_i\}_{i=1}^n$ collected in $n \times \ell$ matrix Z treated for simplicity as fixed. Under the traditional large sample asymptotics, as $n \to \infty$,

$$\frac{e'Pe}{n} = \frac{e'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'e}{n} \xrightarrow{p} 0 \cdot \left(\lim \frac{Z'Z}{n}\right)^{-1} \cdot 0 = 0,$$

and, when appropriately normalized,

$$e'Pe = \frac{e'Z}{\sqrt{n}}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'e}{\sqrt{n}} \xrightarrow{d} \sigma^2\chi^2(\ell).$$

Under the many instrument asymptotics, as $\ell \to \infty$ and $\ell/n \to \alpha \neq 0$, the first step above immediately fails (e.g., the limit of $n^{-1}Z'Z$ is not well defined). Instead, under suitable conditions, the normalized quadratic form converges to a non-zero value: as $\ell \to \infty$ and $\ell/n \to \alpha$,

$$\frac{e'Pe}{n} \xrightarrow{p} \alpha\sigma^2 \neq 0$$

as long as $\alpha \neq 0$. Indeed, consider its expectation:

$$E\left[\frac{e'Pe}{n}\right] = \frac{1}{n}E\left[\operatorname{tr}\left(e'Pe\right)\right] = \frac{1}{n}\operatorname{tr}\left(PE\left[ee'\right]\right) = \frac{1}{n}\operatorname{tr}\left(P\sigma^2 I_n\right) = \sigma^2\frac{\ell}{n} \to \alpha\sigma_e^2.$$

Because the limit is non-zero, the quadratic form has to be recentered before being normalized. It turns out that its normalization by $\sqrt{n}$ instead of n is required:

$$\frac{e'(P-\alpha I_n)e}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, 2\alpha\left(1-\alpha\right)\sigma^4 + \kappa\left(\lim\frac{1}{n}\sum_{i=1}^{n}P_{ii}^2 - \alpha^2\right)\sigma^4\right),$$

where

$$\kappa = \frac{E[e_i^4]}{\sigma^4} - 3$$

is the excess kurtosis coefficient of e_i. Apart from normalization, several important things should be noted. First and most intriguing, a quadratic form gives rise to the *normal* asymptotic distribution instead of the habitual chi-square. Second, the recentered and normalized statistic is asymptotically *non-pivotal* in a complex way. Third, there is dependence of the asymptotic variance on the fourth order moment of e_i (unless P_{ii} all converge to α) and on (the asymptotic behavior of) the diagonal elements of P (unless e_i is mesokurtic).

Heuristically, the asymptotic normality of the quadratic form comes from instrument numerosity, i.e. infinity of ℓ in the limit. It is well known that, as the degrees of freedom increase, the chi-square distribution becomes closer to the normal distribution. Indeed, as $\ell \to \infty$, if we represent the chi-square distribution as a sum of independent squared standard normals, recenter the result and renormalize it, we get for $\xi_j \sim$ IID $\mathcal{N}(0,1)$ that

$$\frac{\chi^2(\ell) - \ell}{\sqrt{2\ell}} \overset{d}{=} \frac{1}{\sqrt{2\ell}}\sum_{j=1}^{\ell}(\xi_j^2 - 1) \overset{a}{\sim} \mathcal{N}(0,1),$$

which implies asymptotic normality of $e'(P-\alpha I_n)e$, though with incorrect asymptotic variance:

$$\frac{e'(P-\alpha I_n)e}{\sqrt{n}} \overset{a}{\sim} \mathcal{N}\left(0, 2\alpha\sigma^4\right).$$

The discrepancy in asymptotic variances arises because asymptotically there is a bounded amount n/ℓ of 'information' per each ξ_j rather than its infinite amount. It is this discrepancy that invalidates conventional χ^2 critical values that do not account for it in the presence of many instruments.

Note again that the asymptotic normality is driven by asymptotic infinity of instruments rather than only by an asymptotically infinite sample, i.e. it comes from $\ell \to \infty$ rather than only from $n \to \infty$. Such asymptotic normality can also arise when this condition holds without the asymptotic proportionality of ℓ to n. In fact, it may arise even when $\ell \to \infty$ slower than $n \to \infty$ (see, e.g., Donald, Imbens and Newey, 2003).

Apart from the literature on many instruments/regressors, the asymptotic normality of quadratic forms has been utilized in spatial econometrics (e.g., Kelejian and Prucha, 2001) and vast dimensional covariance matrix estimation (e.g., Ledoit and Wolf, 2002).

3.3 Large sample theorems

In his article, Bekker (1994) used the exact statistical properties of the quadratic form based on the exact normality of its elements. There is a broad statistical literature on asymptotic behavior of U-statistics; see, for example, Götze and Tikhomirov (1999). The conditions for the large sample theorems developed there are typically too tight to be applicable in the present context. However, de Jong (1987) provides a useful result, which was used in Calhoun (2011).

Theorem (de Jong, 1987) Let

$$Q_n = \sum_{i \neq j} a_{ij} \varepsilon_i \varepsilon_j$$

be a quadratic form in independent random variables $\{\varepsilon_i\}_{i=1}^n$ with $E[\varepsilon_i] = 0$ and $E\left[\varepsilon_i^2\right] = 1$, with $\mu_1, ..., \mu_n$ the eigenvalues of the symmetric matrix $\|a_{ij}\|_{i,j=1}^n$ with no diagonal elements, $a_{ii} = 0$ for all i. Suppose there exists a sequence of real numbers K_n such that $K_n^2 \sigma_{Q_n}^{-2} \max_{1 \leq j \leq n} \sum_{i=1}^n a_{ij}^2 \to 0$ and $\max_{1 \leq i \leq n} E\left[\varepsilon_i^2 \mathbb{I}_{\{|\varepsilon_i| > K_n\}}\right] \to 0$ as $n \to \infty$. If the eigenvalues of the matrix $\|a_{ij}\|_{i,j=1}^n$ are negligible, $\sigma_{Q_n}^{-2} \max_{1 \leq i \leq n} \mu_i^2 \to 0$ as $n \to \infty$, then

$$\frac{Q_n}{\sigma_{Q_n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $n \to \infty$, where

$$\sigma_{Q_n}^2 = 2 \sum_{i \neq j} a_{ij,n}^2.$$

Arguably, one of the first sources in the econometric literature where large sample theorems for quadratic/bilinear forms suitable for handling setups with many instruments/regressors was Kelejian and Prucha (2001) in the context of spatial data analysis. This CLT was used, in particular, in Anatolyev (2012) and Anatolyev and Gospodinov (2011a).

Theorem (Kelejian and Prucha, 2001) Assume the elements of the random array $\{\varepsilon_{i,n}\}_{i=1}^n$ are independent and satisfy $E[\varepsilon_{i,n}] = 0$ and $\sup_n \sup_{1 \leq i \leq n} E[|\varepsilon_{i,n}|^{4+\eta}] < \infty$ for some $\eta > 0$. Consider arrays of real numbers $\{a_{ij,n}\}_{i,j=1}^n$ and $\{b_{i,n}\}_{i=1}^n$ such that $a_{ij,n} = a_{ji,n}$, $\sup_n \sup_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij,n}| < \infty$ and $\sup_n n^{-1} \sum_{i=1}^n |b_{i,n}|^{2+\chi} < \infty$ for some $\chi > 0$. Define a bilinear form

$$Q_n = \sum_{i=1}^n \sum_{j=1}^n a_{ij,n} \varepsilon_{i,n} \varepsilon_{j,n} + \sum_{i=1}^n b_{i,n} \varepsilon_{i,n}.$$

Then, as $n \to \infty$,

$$\frac{Q_n - \mu_{Q_n}}{\sigma_{Q_n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

provided that $\sup_n n^{-1}\sigma^2_{Q_n} > 0$, where

$$\mu_{Q_n} = \sum_{i=1}^{n} a_{ii,n} E\left[\varepsilon^2_{i,n}\right]$$

and

$$\sigma^2_{Q_n} = 4\sum_{i=1}^{n}\sum_{j=1}^{i-1} a^2_{ij,n} E\left[\varepsilon^2_{i,n}\right] E\left[\varepsilon^2_{j,n}\right] + \sum_{i=1}^{n} b^2_{i,n} E\left[\varepsilon^2_{i,n}\right] + \sum_{i=1}^{n}\left(a^2_{ii,n}\text{var}\left[\varepsilon^2_{i,n}\right] + 2a_{ii,n}b_{i,n}E\left[\varepsilon^3_{i,n}\right]\right).$$

This CLT, unlike most in the statistical literature, allows the coefficients to depend on the sample size. In addition, it allows the random elements $\varepsilon_{i,n}$ to be heterogeneous across i.

After a while, the many instrument literature was enriched by central limit theorems targeted to the needed structure of the coefficients, in particular, associated with a projection matrix. The CLT of Hansen, Hausman and Newey (2008) for bilinear forms covering the case of fixed regressors and conditional homoskedasticity reads as follows:

Theorem (Hansen, Hausman and Newey, 2008) Suppose $\ell \to \infty$ as $n \to \infty$. Let P be a constant idempotent matrix with rank ℓ, let $(w_{1n}, v_1, u_1), ..., (w_{nn}, v_n, u_n)$ be independent, and let (v_i', u_i) have bounded 4th moments, with $E[v_i] = E[u_i] = 0$, $E[(v_i', u_i)(v_i', u_i)'] =$ const, with $E[||w_{in}||^4] \to 0$ and bounded $\sum_{i=1}^n E[w_{in}w_{in}']$. Then, for any sequence of bounded vectors c_{1n} and c_{2n} such that

$$V_n = c_{1n}'\left(\sum_{i=1}^{n} E[w_{in}w_{in}']\right)c_{1n} + \left(1 - \frac{\sum_{i=1}^n P_{ii}^2}{\ell}\right)c_{2n}'(E[v_iv_i']E[u_i^2] + E[v_iu_i]E[u_iv_i'])c_{2n}$$

is bounded away from zero, it follows that

$$V_n^{-1/2}\left(c_{1n}'\sum_{i=1}^{n} w_{in} + c_{2n}'\sum_{i\neq j}\frac{v_jP_{ij}u_i}{\sqrt{\ell}}\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Note that the linear part is asymptotically mean zero while the bilinear part is of jackknife type (i.e. there are no terms where both random elements have equal indexes), hence the entire form has zero mean. The two parts are uncorrelated, hence no covariance term in the asymptotic variance expression.

One more CLT for symmetric quadratic forms was developed in van Hasselt (2010), where the central matrix does not need to be a projection, the vector elements have non-zero mean, and the linear part is incorporated into the quadratic part. However, the assumptions underlying van Hasselt (2010)'s CLT imposed on the central matrix and vector elements' means seem to be unnecessarily stringent for the given context.

An extension of the CLT of Hansen, Hausman and Newey (2008) given above is provided in Chao, Swanson, Hausman, Newey and Woutersen (2012). The generalization concerns stochastic instruments and heterogeneous elements of the bilinear form thus allowing for conditional heteroskedasticity. The CLT reads as follows:

Theorem (Chao, Swanson, Hausman, Newey and Woutersen, 2012) Suppose $\ell \to \infty$ as $n \to \infty$. Let P be Z-measurable symmetric idempotent matrix with rank ℓ with diagonal elements uniformly smaller than unity almost surely. Let, conditional on Z, $(w_{1n}, v_1, u_1), ..., (w_{nn}, v_n, u_n)$ be independent, (v_i', u_i) have uniformly bounded 4th conditional moments, with $E[w_{in}|Z] = E[v_i|Z] = E[u_i|Z] = 0$, with $\sum_{i=1}^n E[||w_{in}||^4|Z] \overset{AS}{\to} 0$ and uniformly bounded $\|\sum_{i=1}^n E[w_{in}w_{in}'|Z]\|$ almost surely. Then, for any sequence of uniformly bounded vectors c_{1n} and c_{2n} depending on Z such that

$$V_n = c_{1n}' \left(\sum_{i=1}^n E[w_{in}w_{in}'|Z] \right) c_{1n} + c_{2n}' \frac{1}{\ell} \sum_{i \neq j} P_{ij}^2 (E[v_i v_i'|Z] E[u_i^2|Z] + E[v_i u_i|Z] E[u_i v_i'|Z]) c_{2n}$$

is bounded away from zero almost surely, it follows that

$$V_n^{-1/2} \left(c_{1n}' \sum_{i=1}^n w_{in} + c_{2n}' \sum_{i \neq j} \frac{v_i P_{ij} u_j}{\sqrt{\ell}} \right) \overset{d}{\to} \mathcal{N}(0, 1)$$

almost surely.

This CLT may also be applicable when the central matrix is symmetric but not a projection, see a remark in Bekker and Crudu (2015, appendix A.4).

Cattaneo, Jansson and Newey (2018a) presented a unifying framework within which the asymptotic behavior of bilinear forms can be considered. The d-vector V-statistic

$$S_n = \sum_{i=1}^n \sum_{j=1}^n s_{ij},$$

where each s_{ij} contains as a factor an element of a projection matrix of rank ℓ, can be decomposed into three parts:

$$S_n = B_n + \Psi_n + \Phi_n,$$

where $B_n = E[S_n]$ is the bias term, $\Psi_n = \sum_{i=1}^n \psi_i$ is a sum of independent zero mean terms, and $\Phi_n = \sum_{i \neq j} \phi_{ij}$ is a degenerate U-statistic which is a martingale difference sum uncorrelated with Ψ_n. While the component B_n can be made asymptotically vanish, both the terms Ψ_n and Φ_n can contribute to the asymptotic variance as $\ell \to \infty$ and $n \to \infty$:

$$\begin{pmatrix} \text{var}[\Psi_n]^{-1/2} \Psi_n \\ \text{var}[\Phi_n]^{-1/2} \Phi_n \end{pmatrix} \overset{d}{\to} \mathcal{N}(0, I_{2d}).$$

When ℓ grows slower than n, the second component is of smaller order than the first component, and the asymptotic variance of S_n equals that of Ψ_n. When ℓ grows proportionately with n, the rates are the same, and the asymptotic variance of S_n is composed of those of Ψ_n and Φ_n. In the central limit theorems stated above, Ψ_n corresponds to the 'linear' part, and Φ_n to the 'bilinear' part. Correspondingly, the asymptotic variance is composed of two parts, $\text{var}[\Psi_n]$ and $\text{var}[\Phi_n]$, the 'traditional' asymptotic variance and the 'dimensionality', or 'many instrument/regressor', asymptotic variance.

3.4 Diagonal of projection matrix

Some results in the many instrument/regressor literature (see subsections 4.5, 4.6, 4.9, and 7.2) rely on the assumption that the main diagonal of the projection matrix $P = Z(Z'Z)^{-1} Z'$ associated with the instrument/regressor matrix Z is asymptotically homogeneous, i.e. $P_{ii} \to \alpha$ for all

$i = 1, ..., n$. This property is also called an asymptotically balanced design of instruments/regressors. Sometimes it is formulated in a different form, e.g., as $n^{-1} \sum_{i=1}^{n} (P_{ii} - \alpha)^2 \to 0$ (Anderson, Kunitomo and Matsushita, 2010; Kunitomo, 2012) or as $n^{-1} \sum_{i=1}^{n} |P_{ii} - \alpha| \to 0$ (Anatolyev and Gospodinov, 2011) or (which is unnecessarily stronger) $\max_{i=1,...,n} |P_{ii} - \alpha| \to 0$ (Anatolyev, 2012). The convergence is understood in the usual deterministic sense if the instruments/regressors are treated as fixed, and in the sense of convergence in probability if they are treated as random IID variables. The point of concentration of diagonal elements is α because

$$n^{-1} \sum_{i=1}^{n} P_{ii} = n^{-1} \sum_{i=1}^{n} \mathrm{tr} \left(z_i' \left(Z'Z \right)^{-1} z_i \right) = n^{-1} \mathrm{tr} \left(\left(Z'Z \right)^{-1} \sum_{i=1}^{n} z_i z_i' \right) = n^{-1} \mathrm{tr} \left(I_\ell \right) \to \alpha$$

in the fixed instrument case, and similarly in the random IID case,

$$E\left[P_{ii}\right] = n^{-1} \sum_{i=1}^{n} E\left[\mathrm{tr} \left(z_i' \left(Z'Z \right)^{-1} z_i \right) \right] = n^{-1} \mathrm{tr} \left(E\left[\left(Z'Z \right)^{-1} \sum_{i=1}^{n} z_i z_i' \right] \right) = n^{-1} \mathrm{tr} \left(E\left[I_\ell\right] \right) \to \alpha.$$

Asymptotically balanced design essentially means that variability of P_{ii} around α asymptotically vanishes.

The random matrix theory (e.g., Bai and Silverstein, 2010; Yao, Zheng and Bai 2015), a specific direction in mathematical statistics, concentrates on similar questions and is able to provide some answers in some special cases such as under independence of instruments within the instrument vector. More tailored to econometric setups, Anatolyev and Yaskov (2017) present sufficient conditions for the asymptotic homogeneity of the diagonal of an instrument projection matrix in an IID framework and provide examples of when the instrument design is asymptotically balanced and when it is not. Situations where asymptotically balanced design holds include independent instruments (including gaussian), when instruments are drawn from a log-concave distribution (again including gaussian), when instruments follow a factor model, and some other cases. On the other hand, some situations considered in the many instrument literature, such as those of dummy instruments – standalone and interacted with other instruments like in Angrist and Krueger (1991) – are characterized by non-zero asymptotic variation in diagonal elements. Intuitively, the asymptotic homogeneity of the diagonal occurs when there is sufficient mixing within the instrument vector so that a sort of a law of large numbers holds. Anatolyev and Yaskov (2017) also numerically calibrate the asymptotic variations along the diagonal and conclude that ignoring them may have a non-trivial influence on inferences.

4 Basic model with many instruments

Consider the setup of the paradigmatic returns to schooling application of Angrist and Krueger (1991):

$$\log Y_i = \beta_0 edu_i + w_i' \delta_0 + e_i,$$

where Y_i is yearly income, edu_i is endogenous (years of) education, and w_i contains exogenous regressors, including an intercept. The set of instruments Angrist and Krueger (1991) use contains quarter-of-birth dummies (3) and their interactions with year-of-birth dummies ($9 \times 3 = 27$) and with state-of-birth dummies ($50 \times 3 = 150$), totaling to 180 instruments. The rationale behind the use of such variables as instruments is that while they are clearly exogenous, the US schooling laws prescribe that schooling starts depend on birthdates; these laws vary across the US states and change over time. This induces correlation between the education variable and these instruments, making them relevant. The number of these instruments, 180, does not seem to be that small a number that the traditional asymptotics would approve.

The above description does not include all the important facts about the Angrist and Krueger (1991) application but we will wait and reveal others in due time when they become of interest to our current focus.

4.1 Setup

The basic instrumental variables model is linear, conditionally homoskedastic, and is estimated with an IID sample. The structural equation is

$$y_i = x_i'\beta_0 + e_i,$$

where β_0 is $k \times 1$ vector of structural coefficients of interest, or in matrix notation, $Y = X\beta_0 + e$, where $Y = (y_1, ..., y_n)'$ is $n \times 1$, $X = (x_1, ..., x_n)'$ is $n \times k$, and $e = (e_1, ..., e_n)'$ is $n \times 1$. The reduced form equation is

$$x_i = z_i'\Gamma + u_i,$$

where z_i is $\ell \times 1$ vector of instruments, and Γ is $\ell \times k$ matrix of reduced form coefficients, or in matrix notation, $X = Z\Gamma + U$, where $U = (u_1, ..., u_n)'$ is $n \times k$. We assume that the rank of the instrument matrix $Z = (z_1, ..., z_n)'$ equals its column dimension ℓ. The structural and reduced form errors satisfy

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} | z_i \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \Sigma' \\ \Sigma & \Omega \end{pmatrix} \right).$$

Define the projection matrix associated with the instruments,

$$P = Z \left(Z'Z \right)^{-1} Z'.$$

Also, define $M = I_n - P$. Under conditional homoskedasticity, it is technically easier to treat instruments as fixed. Then the matrix P is a constant symmetric idempotent matrix of rank ℓ.

The key assumption of the many instrument asymptotics is

$$\frac{\ell}{n} = \alpha + o\left(\frac{1}{\sqrt{n}}\right)$$

as $n \to \infty$, where $\alpha \geq 0$. We include $\alpha = 0$ for the traditional asymptotics to be a special case. The order of discrepancy allows one to equalize the ratio ℓ/n and α in (first order) asymptotic inferences.

Along with this assumption, we need an assumption of strong but boundedly informative instruments:

$$\frac{\Gamma' Z' Z \Gamma}{n} \to Q,$$

where matrix Q has full rank. This condition means that even asymptotically (within the thought experiment where the instruments arrive continuously and whose number eventually becomes infinite), the instruments explain only a portion of the right hand side variables, and Q being of full rank means that this portion is nontrivial (see the discussion of the notion of concentration parameter in subsection 6.1). Note that this condition implies that some individual instruments may be strong, but most (almost all) must be weak or even irrelevant.

4.2 Inconsistency of 2SLS

Under the many instrument asymptotics, the standard two-stage least-squares estimator is compromised. Recall that the 2SLS estimator is

$$\hat{\beta}_{2SLS} = \left(X'PX\right)^{-1} X'PY.$$

Under the traditional asymptotics with fixed ℓ, the 2SLS estimator is consistent:

$$\hat{\beta}_{2SLS} = \beta_0 + \left(\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'X}{n}\right)^{-1}\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'e}{n} \xrightarrow{p} \beta_0,$$

as $n^{-1}Z'e \xrightarrow{p} 0$, provided that $\lim n^{-1}Z'Z$ exists and is non-singular. However, under the many instrument asymptotics, schematically,

$$\frac{X'PX}{n} = \frac{\Gamma'Z'Z\Gamma}{n} + \frac{\Gamma'Z'U}{n} + \frac{U'Z\Gamma}{n} + \frac{U'PU}{n} \xrightarrow{p} Q + \alpha\Omega$$

and

$$\frac{X'Pe}{n} = \frac{\Gamma'Z'e}{n} + \frac{U'Pe}{n} \xrightarrow{p} \alpha\Sigma.$$

The last terms turn out to be asymptotically non-negligible because of instrument numerosity (recall subsection 3.2). Hence,

$$\hat{\beta}_{2SLS} = \beta_0 + \left(\frac{X'PX}{n}\right)^{-1}\frac{X'Pe}{n} \xrightarrow{p} \beta_0 + (Q+\alpha\Sigma)^{-1}\alpha\Sigma \neq \beta_0,$$

unless $\alpha = 0$ (the instruments are fewer than many) or $\Sigma = 0$ (there is no endogeneity).

Intuitively, the 'incidental parameter problem' in the first stage of the 2SLS procedure – projecting right side variables on the high dimensional instrument space – results in big noise that contaminates the second stage estimation.

4.3 Consistent estimators

There are a few solutions to the problem of inconsistent estimation. One, the bias-corrected 2SLS (B2SLS) estimator (Nagar, 1959; Donald and Newey, 2001), results from the direct removal of the source of the bias in the 2SLS formula. The diagonal elements of P are recentered in as follows:

$$\hat{\beta}_{B2SLS} = \left(X'\left(P - \frac{\ell}{n}I_n\right)X\right)^{-1} X'\left(P - \frac{\ell}{n}I_n\right)Y.$$

Under the traditional large sample asymptotics, the correction is negligible, while under many instrument asymptotics,

$$\begin{aligned}
\hat{\beta}_{B2SLS} - \beta_0 &= \left(\frac{X'PX}{n} - \frac{\ell}{n}\frac{X'X}{n}\right)^{-1}\left(\frac{X'Pe}{n} - \frac{\ell}{n}\frac{X'e}{n}\right) \\
&\xrightarrow{p} \left(Q + \alpha\Omega - \alpha\left(Q + \Omega\right)\right)\left(\alpha\Sigma - \alpha\Sigma\right) \\
&= 0.
\end{aligned}$$

Another solution uses the idea of jackknifing, leading to the following most natural version of the jackknife instrumental variables (JIV) estimator (Angrist, Imbens and Krueger, 1999; Ackerberg and Devereux, 2009). Denote $D = \mathrm{dg}\{P_{ii}\}_{i=1}^{n}$. In the JIV estimator, the diagonal elements of P are completely removed:

$$\begin{aligned}
\hat{\beta}_{JIV} &= \left(X'(P-D)X\right)^{-1} X'(P-D)Y \\
&= \left(\sum_{i\neq j} P_{ij}x_i x_j'\right)^{-1} \sum_{i\neq j} P_{ij}x_i y_j.
\end{aligned}$$

Under the many instrument asymptotics,

$$\hat{\beta}_{JIV} - \beta_0 \xrightarrow{p} \left(\lim \frac{1}{n}\sum_{i\neq j} P_{ij}E\left[x_i x_j'\right]\right)^{-1} \lim \frac{1}{n}\sum_{i\neq j} P_{ij}E\left[x_i e_j\right] = 0,$$

because those terms in the numerator that possess non-zero expectation due to endogeneity, have been removed. Intuitively, the leave-one-out machinery reduces overfitting in the first stage of the 2SLS procedure.[2]

Yet another solution is provided by the limited information maximum likelihood (LIML) estimator (Anderson and Rubin, 1949; Anderson, 2005):

$$\begin{aligned}
\hat{\beta}_{LIML} &= \arg\min_{\beta} \frac{(Y - X\beta)'P(Y - X\beta)}{(Y - X\beta)'M(Y - X\beta)} \\
&= \arg\min_{\beta} \frac{(Y - X\beta)'P(Y - X\beta)}{(Y - X\beta)'(Y - X\beta)}.
\end{aligned}$$

Note that if not for the denominator of the LIML objective function, the minimizer would be the 2SLS estimator. The presence of the estimation of the structural error variance in the denominator is a continuously updated GMM type correction known to reduce the second-order asymptotic bias under traditional asymptotics (see Newey and Smith, 2004).

Computationally, it is more convenient to represent the LIML estimator in the following form:

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

[2] This version of the jackknife estimator appears most elegant. There are other versions of JIV estimation described, for example, in Poi (2006): one that also originates from Angrist, Imbens & Krueger (1999), whose performance in simulations is very close to that of the one we consider; and two versions that originate from Blomquist and Dahlberg (1999), which exhibit poor behavior under many instruments (Poi, 2006).

where $\hat{\alpha}$ is the minimal eigenvalue of the matrix $\left(\bar{X}'\bar{X}\right)^{-1}\bar{X}'P\bar{X}$, where $\bar{X} = [Y\, X]$. Another useful fact is that $\hat{\alpha}$ equals $F(\hat{\beta}_{LIML})$, where

$$F(\beta) \equiv \frac{(Y - X\beta)' P (Y - X\beta)}{(Y - X\beta)' (Y - X\beta)}.$$

The LIML consistency follows from the fact that the true parameter solves the population analog of the LIML optimization problem (Newey, 2004; Hansen, Hausman and Newey, 2008):

$$\begin{aligned}
\hat{\beta}_{LIML} &= \arg\min_{\beta} \frac{n^{-1}(Y - X\beta)' P (Y - X\beta)}{n^{-1}(Y - X\beta)'(Y - X\beta)} \\
&= \arg\min_{\beta} \frac{n^{-1}(e - X(\beta - \beta_0))' P (e - X(\beta - \beta_0))}{n^{-1}(e - X(\beta - \beta_0))'(e - X(\beta - \beta_0))} \\
&\xrightarrow{p} \arg\min_{\beta} \frac{\alpha\sigma^2 - 2\alpha(\beta - \beta_0)'\Sigma + (\beta - \beta_0)'(Q + \alpha\Omega)(\beta - \beta_0)}{\sigma^2 - 2(\beta - \beta_0)'\Sigma + (\beta - \beta_0)'(Q + \Omega)(\beta - \beta_0)} \\
&= \arg\min_{\beta} \left\{\alpha + (1 - \alpha)\frac{(\beta - \beta_0)'Q(\beta - \beta_0)}{\sigma^2 - 2(\beta - \beta_0)'\Sigma + (\beta - \beta_0)'(Q + \Omega)(\beta - \beta_0)}\right\} \\
&= \beta_0
\end{aligned}$$

as long as $\alpha \neq 1$. Even though the 'incidental parameter problem' is still present, it does not affect the consistency of LIML estimation. Kolesár (2018) explains this robustness by the coincidence of the LIML estimator, at least under normality, with the maximum invariance likelihood estimator, one whose number of parameters is asymptotically fixed. The asymptotic variance, however, is inflated by the presence of many instruments; see the next subsection.

A well-known problem with LIML is the non-existence of moments (Fuller, 1977; Chao, Hausman, Newey, Swanson and Woutersen, 2012a). This phenomenon arises because the denominator of $\hat{\beta}_{LIML}$ possesses too much probability mass in the vicinity of zero. An asymptotically negligible perturbation in recentering the projection is able to solve the moment existence problem, see details in Chao, Hausman, Newey, Swanson and Woutersen (2012a). The Fuller (1977) (FULL) estimator that achieves this goal is

$$\hat{\beta}_{FULL} = \left(X'(P - \mathring{\alpha}I_n)X\right)^{-1} X'(P - \mathring{\alpha}I_n)Y$$

where

$$\mathring{\alpha} = \frac{\hat{\alpha} - (1 - \hat{\alpha})C/n}{1 - (1 - \hat{\alpha})C/n} \tag{1}$$

for some C (e.g., Hansen, Hausman and Newey, 2008). The most popular version uses the value $C = 1$, in which case the FULL estimator possesses moments of all orders.

4.4 Asymptotic distributions under normal errors

Assume for the time being that the structural and reduced form errors are jointly gaussian. Under traditional large sample asymptotics, the LIML estimator is consistent and asymptotically normal:

$$\sqrt{n}\left(\hat{\beta}_{LIML} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V_1^{LIML}\right),$$

where

$$V_1^{LIML} = \sigma^2 Q^{-1}.$$

The B2SLS and Fuller estimators are asymptotically equivalent to LIML. The construction of the JIV estimator entails loss of information; as a result, the efficiency of estimation somewhat suffers:

$$\sqrt{n}\left(\hat{\beta}_{JIV} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V_1^{JIV}\right),$$

where

$$V_1^{JIV} = \sigma^2 Q_-^{-1} Q_{-2} Q_-^{-1},$$

$Q_- = \lim n^{-1}\Gamma' Z' (I - D) Z\Gamma$ and $Q_{-2} = \lim n^{-1}\Gamma' Z' (I - D)^2 Z\Gamma$. Large efficiency losses are confirmed by simulations in Davidson and MacKinnon (2006).

Under many instrument asymptotics and error normality, Bekker (1994), Hansen, Hausman and Newey (2008), van Hasselt (2010) and Chao, Swanson, Hausman, Newey and Woutersen (2012) have shown that the B2SLS, JIV, LIML and FULL estimators are asymptotically normal, with the asymptotic variance containing two components:

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V_1 + V_2\right),$$

where V_1 is the conventional asymptotic variance shown above, and instrument numerosity manifests itself in the additional variance component V_2. Note that the rate of convergence is conventional $\sqrt{n}$ because is it proportional to $\sqrt{\ell}$ within the many instrument framework. The additional variance V_2 is necessarily positive definite, and different for different estimators:

$$\begin{aligned}
V_2^{B2SLS} &= \frac{\alpha}{1-\alpha} Q^{-1}\left(\sigma^2\Omega + \Sigma\Sigma'\right) Q^{-1}, \\
V_2^{LIML} &= \frac{\alpha}{1-\alpha} Q^{-1}\left(\sigma^2\Omega - \Sigma\Sigma'\right) Q^{-1}, \\
V_2^{JIV} &= \alpha\left(1-\gamma\right) Q_-^{-1}\left(\sigma^2\Omega + \Sigma\Sigma'\right) Q_-^{-1},
\end{aligned}$$

where $\gamma = \lim \ell^{-1}\sum_{i=1}^{n} P_{ii}^2$.

These components of asymptotic variances are positively related to α, the degree of instrument numerosity. Perhaps even more importantly, they contain indicators of instrument strength Q and Q_- in the denominator, which occurs twice in comparison to appearing once in the V_1 component. This means that when the instruments are not too strong, the 'many instrument' component V_2 may well exceed the 'traditional' component V_1 (Newey, 2004); see Section 6.

4.5 Asymptotic distributions under non-normal errors

Now we depart with the assumption of error normality. Unlike in the traditional asymptotic framework, under many instruments (and some additional assumptions), this results in a change of the form the asymptotic variances take. Namely, two additional terms are responsible for deviations of certain third and fourth error moments from their counterparts under error normality:

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V_1 + V_2 + V_3 + V_4\right),$$

where V_3 is additional variance associated with error 3rd moments, and V_4 is additional variance associated with error 4th moments. Define $\tilde{u}_i = u_i - \sigma^{-2}\Sigma e_i$ to be population residuals in the least squares projection of reduced form errors on the structural errors, and let $\pi_\alpha = \lim n^{-1}\Gamma' Z' (D - \alpha I)\, \iota_n$, where ι_n is an $n \times 1$ vector of ones, exist. The exact formulas are:

$$\begin{aligned} V_3^{B2SLS} &= \frac{1}{1-\alpha} Q^{-1} \left(\pi_\alpha E\left[u_i' e_i^2\right] + E\left[u_i e_i^2\right] \pi_\alpha' \right) Q^{-1}, \\ V_3^{LIML} &= \frac{1}{1-\alpha} Q^{-1} \left(\pi_\alpha E\left[\tilde{u}_i' e_i^2\right] + E\left[\tilde{u}_i e_i^2\right] \pi_\alpha' \right) Q^{-1}, \end{aligned}$$

and

$$\begin{aligned} V_4^{B2SLS} &= \frac{\alpha(\gamma-\alpha)}{(1-\alpha)^2} Q^{-1} \left(E\left[u_i u_i' e_i^2\right] - \sigma^2 \Omega - 2\Sigma\Sigma' \right) Q^{-1}, \\ V_4^{LIML} &= \frac{\alpha(\gamma-\alpha)}{(1-\alpha)^2} Q^{-1} E\left[\tilde{u}_i \tilde{u}_i' \left(e_i^2 - \sigma^2\right)\right] Q^{-1}, \end{aligned}$$

while

$$V_3^{JIV} = V_4^{JIV} = 0$$

thanks to the removal of the terms that may give rise to higher-order moments.

The V_3 and V_4 components are zero under the asymptotically balanced instrument design (see subsection 3.4), in which case $\pi_\alpha = 0$ and $\gamma = \alpha$. Under general instrument design, they may not be zero but the formulas can be simplified under restrictions on the joint error distribution. Consider the special case of error moment independence: suppose that $E[e_i|\tilde{u}_i] = 0$, $E\left[e_i^2|\tilde{u}_i\right] = \sigma^2$ and $E\left[e_i^3|\tilde{u}_i\right] = v_3$. Then

$$\begin{aligned} V_3^{B2SLS} &= \frac{1}{1-\alpha} \frac{v_3}{\sigma^2} Q^{-1} \left(\pi_\alpha \Sigma' + \Sigma \pi_\alpha' \right) Q^{-1}, \\ V_4^{B2SLS} &= \frac{\alpha(\gamma-\alpha)}{(1-\alpha)^2} \left(\frac{v_4}{\sigma^4} - 3 \right) Q^{-1} \Sigma\Sigma' Q^{-1}, \end{aligned}$$

while

$$V_3^{LIML} = V_4^{LIML} = 0.$$

If we additionally assume error normality, we also get

$$V_3^{B2SLS} = V_4^{B2SLS} = 0.$$

Abutaliev and Anatolyev (2013) compute the numerical values of various components for the LIML estimator for a model with Skewed Student errors with gaussian copula superimposed and group instruments as in Bekker and van der Ploeg (2005). It turns out that indeed, for small first-stage R^2 (and thus small Q relative to Ω), the 'many instrument' correction V_2^{LIML} may be very large relative to the 'traditional' component V_1^{LIML}. For a large degree of endogeneity, the 'excess kurtosis' correction V_4^{LIML} may be quite perceptible relative to V_1^{LIML}, while most often the 'skewness' correction V_3^{LIML} is much smaller in absolute value relative to V_1^{LIML}. The latter evidence is consistent with the many weak instrument theory (see subsection 6.2).

4.6 Asymptotic efficiency

When possessing the formulas for asymptotic variance components, one can rank the 'traditional' asymptotic variance components easily: $V_1^{B2SLS} = V_1^{LIML} \leq V_1^{JIV}$, but the 'many instrument' components are more ambiguous. It is clear that $V_2^{B2SLS} \geq V_2^{LIML}$, strictly so if $\alpha \neq 0$ and $\Sigma \neq 0$, but it is not clear how these are related to V_2^{JIV}. The presence of the V_3 and V_4 components complicates the picture even further. Under the asymptotically balanced instrument design (see subsection 3.4) when the last two components are absent, one can additionally obtain that $V_1^{B2SLS} = V_1^{JIV}$ and $V_2^{B2SLS} = V_2^{JIV}$, and hence $LIML \succeq B2SLS \sim JIV$ in terms of asymptotic efficiency. Under error moment independence, the signs of V_3^{B2SLS} and V_4^{B2SLS} depend on the interaction of the sign of $\pi_\alpha \Sigma'$ and the direction of skewness, while the sign of V_4^{B2SLS} is determined by whether the structural error is lepto- or platykurtic, and both components may have a positive or negative effect on the total B2SLS asymptotic variance.

More conceptually, with many instruments, classical results on parametric efficiency are not applicable because of the presence of many incidental parameters. The ranking of available estimators in terms of asymptotic efficiency is more problematic than in the traditional large sample framework due to the presence of several components in the asymptotic variances. More instruments, which implies higher Q and higher α, may result in a higher or lower asymptotic variance of a given estimator.

In setups with either asymptotically balanced instrument design or elliptically contoured error distribution, Anderson, Kunitomo and Matsushita (2010) show the asymptotic efficiency of LIML in a set of consistent estimators taking the form $\phi(S, S^\perp)$, where $S = \bar{X}'P\bar{X}$ and $S^\perp = \bar{X}'M\bar{X}$ are a pair of sufficient statistics, for a smooth function ϕ with bounded derivatives. Chioda and Jansson (2009) expand the class of estimators to include all regular rotation-invariant estimators and show that the asymptotic efficiency bound coincides with the LIML asymptotic variance, although they prove their results under error normality.

Hahn (2002) finds out that when rotation invariance is not imposed, the asymptotic efficiency bound is strictly lower than that attained by LIML. He argues that it has to lie between the classical few instrument efficiency bound and the semiparametric efficiency bound of Chamberlain (1987), which however coincide and are equal to $\sigma^2 Q^{-1}$. Heuristically, if ℓ increases slower than n, then formally $\alpha = 0$ and the LIML asymptotic variance achieves the few instrument efficiency bound. Thus, LIML and asymptotically equivalent estimators like FULL are not globally efficient, and neither are any estimators whose asymptotic variance depends on α. It is not clear though that such an efficiency bound is possible to achieve in practice, as this requires the use of a restricted subset of available instruments, and it is not clear how to choose this restricted set. Hahn (2002) conjectures that selecting instruments using tools in Donald and Newey (2001) might help attain the efficiency bound.

In a model with one endogenous regressor and non-normal errors, Kolesár (2018) proposes a *minimum distance estimator* that efficiently employs all information in S and $S^\perp$, which is asymptotically more efficient that the LIML estimator, and describes its asymptotic variance.

4.7 Variance estimation

For any consistent estimator considered above, one can consistently estimate all components of the asymptotic variance in a straightforward way. It is important, as in the rest of variance estimation business, to make an emerging estimator positive definite by construction. For example, Bekker (1994) and Hansen, Hausman and Newey (2008) suggest the following variance estimate for LIML under error normality:

$$\hat{V}_{1+2}^{LIML} = \hat{H}^{-1}\hat{\Sigma}_{1+2}^{LIML}\hat{H}^{-1},$$

where $\hat{H} = X'(P - \tilde{\alpha}I_n)X$ and

$$\hat{\Sigma}_{1+2}^{LIML} = \hat{\sigma}^2\left((1-\tilde{\alpha})^2\tilde{X}'P\tilde{X} + \tilde{\alpha}^2\tilde{X}'M\tilde{X}\right),$$

where $\tilde{\alpha} = \hat{e}'P\hat{e}/\hat{e}'\hat{e} \xrightarrow{p} \alpha$, $\hat{\sigma}^2 = \hat{e}'\hat{e}/(n-k) \xrightarrow{p} \sigma^2$, and $\tilde{X} = X - (\hat{e}'X/\hat{e}'\hat{e})\hat{e}$, and where $\hat{e}$ are LIML residuals. Unfortunately, the V_3 and V_4 components have a complex structure, and it does not seem possible to ensure positive definiteness of the whole variance estimate by construction. Hansen, Hausman and Newey (2008) propose the following consistent estimator:

$$\hat{V}^{LIML} = \hat{H}^{-1}\left(\hat{\Sigma}_{1+2}^{LIML} + \hat{\Sigma}_{3}^{LIML} + \hat{\Sigma}_{4}^{LIML}\right)\hat{H}^{-1},$$

where, in addition to above,

$$\hat{\Sigma}_{3}^{LIML} = \frac{1}{n}\sum_{i=1}^{n}\left(P_{ii} - \frac{\ell}{n}\right)(PX)_i\sum_{i=1}^{n}\hat{e}_i^2(M\tilde{X})_i'$$

and

$$\hat{\Sigma}_{4}^{LIML} = \frac{\sum_{i=1}^{n}P_{ii}^2 - n^{-1}\ell^2}{n - 2\ell + \sum_{i=1}^{n}P_{ii}^2}\sum_{i=1}^{n}(\hat{e}_i^2 - \hat{\sigma}^2)(M\tilde{X})_i(M\tilde{X})_i'.$$

4.8 Parameter inference

Once consistent estimates of asymptotic variance are available, testing structural parameter restrictions is straightforward. In particular, for the null hypothesis

$$H_0 : c'\beta = c'\beta_0,$$

where c is $k \times 1$ vector of constants, Hansen, Hausman and Newey (2008) establish that the LIML t-statistic is asymptotically standard normal

$$\frac{c'(\hat{\beta}_{LIML} - \beta_0)}{\sqrt{c'\hat{V}^{LIML}c}} \xrightarrow{d} \mathcal{N}(0,1),$$

and the decision rule is usual based on the quantiles of the standard normal distribution. The FULL estimator can be used in place of LIML.

4.9 Specification testing

Specification testing is related to the hypothesis of valid instruments

$$H_0: \; E[z_i e_i] = 0$$

(or, rather, $H_0: \; n^{-1}Z'e \xrightarrow{p} 0$ for fixed Z). The Hansen (1992) overidentification J test statistic and its asymptotic distribution under large sample asymptotics are:

$$J = \frac{\hat{e}'P\hat{e}}{\hat{\sigma}^2} \xrightarrow{d} \chi^2(\ell - k).$$

Now, under many instrument asymptotics, $J \xrightarrow{p} \infty$, so the statistic should be recentered and renormalized (recall subsection 3.2). Lee and Okui (2012) derive that the modified J test statistic

$$J_m = \frac{\hat{e}'\left(P - (\ell/n)I_n\right)\hat{e}}{\hat{\sigma}^2}$$

has the following asymptotic distribution:

$$\frac{J_m}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, 2\alpha(1-\alpha) + \kappa\left(\lim \frac{1}{n}\sum_{i=1}^{n} P_{ii}^2 - \alpha^2\right)\right),$$

where

$$\kappa = \frac{E[e_i^4]}{\sigma^4} - 3$$

is excess kurtosis of e_i (cf. subsection 3.2). Pivotization of the statistic is straightforward, and the testing is one-sided. An analogous test can be constructed on the basis of the J statistic modified in the jackknife fashion; see subsection 5.3.

In the situation of asymptotically balanced instrument design (see subsection 3.4) implying $\lim n^{-1}\sum_{i=1}^{n} P_{ii}^2 = \alpha^2$, Anatolyev and Gospodinov (2011a) work with the conventional J statistic

$$J = \frac{\hat{e}'P\hat{e}}{\hat{\sigma}^2},$$

and propose the *robust J test* that prescribes rejecting H_0 at significance level ϕ when

$$J > q_{\phi^*}^{\chi^2(\ell-k)}$$

where

$$\phi^* = \Phi\left(\sqrt{1 - \frac{\ell}{n}} \cdot \Phi^{-1}(\phi)\right).$$

That is, testing is carried out in the conventional way using χ^2 critical values but with an adjusted significance level $\phi^* < \phi$ in place of the target level ϕ. This test is robust to the type of asymptotic framework as it is valid both when there are few instruments (in which case $\phi^* = \phi$ asymptotically) and when instruments are many (and the instrument design is asymptotically balanced). Kaffo and Wang (2017) note that the test also works under many weak instrument asymptotics (see Section 6).

4.10 Approximate reduced form

Recall that in the basic setup, the reduced form equation is linear:

$$x_i = z_i'\Gamma + u_i.$$

Hansen, Hausman and Newey (2008) (as well as Hausman, Newey, Woutersen, Chao and Swanson, 2012; Chao, Swanson, Hausman, Newey and Woutersen, 2012) consider a more realistic situation. They consider a general reduced form

$$x_i = \Upsilon_i + u_i,$$

where *unknown* Υ_i is approximated by a linear function of z_i. This linear combination may primarily originate from expanding the unknown non-linear function of few basic instruments using approximating functions such as powers, B-splines, etc. as in Newey (1990).

It turns out that the previous results go through if the approximation is good enough, in the sense that for each n there exists a linear combination Γ_n of instruments z_i such that

$$\frac{1}{n}\sum_{i=1}^{n} ||\Upsilon_i - z_i'\Gamma_n||^2 \overset{AS}{\to} 0.$$

4.11 Many exogenous regressors

Consider again the setup of Angrist and Krueger (1991):

$$\log Y_i = \beta_0 edu_i + w_i'\delta_0 + e_i,$$

where Y_i is yearly income, and edu_i is endogenous (years of) education. The set of exogenous regressors (also known as 'included instruments' or 'covariates') in w_i contains the intercept (1), year-of-birth dummies (9) and state-of-birth dummies (50), so in total there are 60 exogenous regressors; this number is comparable to a number of excluded instruments (180). Such inclusions of many exogenous variables into the right hand side are typically done as a guard against model misspecification or in order to approximate some important but unobservable factor.

Anatolyev (2013) extends the basic setup discussed previously to

$$y_i = x_i'\beta_0 + w_i'\delta_0 + e_i,$$

where w is an additional m-vector of exogenous regressors (or covariates, or included instruments), and δ_0 is m-vector of coefficients, which are nuisance parameters in this context. An additional Bekker (1994) type assumption imposed on the dimensionality of w characterizes the *many exogenous regressor asymptotics*:

$$\frac{m}{n} \to \mu > 0$$

as $n \to \infty$.

Collect w_i for all $i = 1, ..., n$ into $n \times m$ matrix W with the associated projection and residual matrices $P_W = W\left(W'W\right)^{-1} W'$ and $M_W = I_n - P_W$. Also, similarly denote by $P_{Z^\perp}$ and $M_{Z^\perp}$ the projection and residual matrices associated with $Z^\perp = M_W Z$.

The additional assumption has certain effects on the consistency of some estimators. The usual bias-corrected 2SLS estimator ceases to be consistent, but appropriate bias correction as in

$$\hat{\beta}_{B2SLS} = \left(X' \left(P_{Z^\perp} - \frac{\ell}{n-m} M_W \right) X \right)^{-1} X' \left(P_{Z^\perp} - \frac{\ell}{n-m} M_W \right) Y$$

restores consistency. The LIML estimator is still consistent, but numerosity of exogenous regressors has certain effects on the form of asymptotic variance, for example,

$$V_1^{LIML} + V_2^{LIML} = \sigma^2 Q_W^{-1} + \frac{\alpha}{1-(1-\mu)^{-1}\alpha} Q_W^{-1} \left(\sigma^2 \Omega - \Sigma\Sigma' \right) Q_W^{-1}$$

where now $Q_W = \lim n^{-1} \Gamma' Z' M_W Z \Gamma$, all other notation being the same as before. The other asymptotic variance components also change in mysterious ways. Anatolyev (2013) also proposes an adaptation of the Lee and Okui (2012) modified J specification test and extends the notion of a k-class estimator (Nagar, 1959) to the situation with numerous exogenous regressors.

Asymptotic variance estimation is further complicated by the fact that under the many instrument and exogenous regressor asymptotics, the habitual estimates of error moments based on sample moments of residuals are no longer consistent. For example, the average squared residual $n^{-1} \sum_{i=1}^n \hat{e}_i^2$ is consistent for $(1-\mu)\sigma^2$ rather than for error variance σ^2, which makes a difference when $\mu \neq 0$. Likewise, the average cubed residual $n^{-1} \sum_{i=1}^n \hat{e}_i^3$ does not consistently estimate $E\left[e_i^3\right]$, but the product $\left(\sum_{j=1}^n \sum_{f=1}^n (M_W)_{jf}^3 \right)^{-1} \sum_{i=1}^n \hat{e}_i^3$ does. For details, see Anatolyev (2013); see also Calhoun (2011) on the consistent estimation of variance and kurtosis in the context of linear regression with many regressors.

In some recent work on many instruments, the numerosity of exogenous regressors is explicitly embedded in the setup; see Kolesár (2018) and Evdokimov and Kolesár (2018).

5 Heteroskedastic model with many instruments

5.1 Setup and estimation

The model equations are the same:

$$y_i = x_i' \beta_0 + e_i$$

and

$$x_i = z_i' \Gamma + u_i,$$

but now the structural and reduced form errors are heteroskedastic:

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} | z_i \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \Sigma_i' \\ \Sigma_i & \Omega_i \end{pmatrix} \right),$$

where the dependence of variances and covariances on i comes from the dependence on z_i if those are treated as random variables.

The bad news is that, generally, the LIML estimator ceases to be consistent. This was first shown in Bekker and van der Ploeg (2005). Chao, Swanson, Hausman, Newey and Woutersen (2012) and Hausman, Newey, Woutersen, Chao and Swanson (2012) provide a systematic treatment of the many instrument problem under heteroskedasticity.

The good news is that consistent estimation is possible under heteroskedasticity using the JIV estimator as it does not contain the terms where endogeneity kicks in and contaminates expectations (Ackerberg and Devereux, 2009; Chao, Swanson, Hausman, Newey and Woutersen, 2012). Indeed, the 2SLS estimator solves

$$\hat{\beta}_{2SLS} \xrightarrow{p} \arg\min_{\beta} \lim E\left[n^{-1}(Y-X\beta)' P (Y-X\beta)\right]$$
$$= \arg\min_{\beta} \left\{ (\beta-\beta_0)'Q_-(\beta-\beta_0) + \lim \frac{1}{n}\sum_{i=1}^{n} P_{ii} E\left[\left(y_i - x_i'\beta\right)^2\right] \right\},$$

where $Q_- = \lim n^{-1}\Gamma' Z'(P-D)Z\Gamma$, which is a positive definite matrix (Hausman, Newey, Woutersen, Chao and Swanson, 2012). The minimizer will differ from β_0 as the second term in the limiting objective function is not minimized at β_0 because of endogeneity, and is of the same order of magnitude as the first term. Removal of the problematic term in the sample objective function leads to the JIV estimator

$$\hat{\beta}_{JIV} = \arg\min_{\beta} n^{-1}(Y-X\beta)'(P-D)(Y-X\beta) \xrightarrow{p} \arg\min_{\beta} (\beta-\beta_0)'Q_-(\beta-\beta_0) = \beta_0.$$

However, as noted previously, since JIV has low efficiency under homoskedasticity (Davidson and MacKinnon, 2006), it is hardly highly efficient in heteroskedastic situations.

Hausman, Newey, Woutersen, Chao and Swanson (2012) propose how to combine consistency of JIV and relative efficiency of LIML. Following Hausman, Newey, Woutersen, Chao and Swanson (2012), the LIML estimator is inconsistent for a similar reason as the 2SLS estimator:

$$\hat{\beta}_{LIML} \xrightarrow{p} \arg\min_{\beta} \frac{\lim E\left[n^{-1}(Y-X\beta)' P (Y-X\beta)\right]}{\lim E\left[n^{-1}(Y-X\beta)'(Y-X\beta)\right]}$$

$$\begin{aligned} &= \arg\min_{\beta} \frac{(\beta-\beta_0)'Q_-(\beta-\beta_0) + \lim n^{-1}\sum_{i=1}^{n} P_{ii} E\left[\left(y_i - x_i'\beta\right)^2\right]}{\lim E\left[n^{-1}(Y-X\beta)'(Y-X\beta)\right]} \\ &\neq \beta_0 \end{aligned}$$

because of the presence of the second term in the numerator of the limiting objective function. Hausman, Newey, Woutersen, Chao and Swanson (2012) give an exact condition when the LIML estimator is inconsistent in terms of non-zero correlation between the diagonal elements of P and certain endogeneity indicators. When LIML does not work, the remedy is familiar – the removal of this source of inconsistency. This gives rise to the *HLIM* estimator

$$\begin{aligned} \hat{\beta}_{HLIM} &= \arg\min_{\beta} \frac{(Y-X\beta)'(P-D)(Y-X\beta)}{(Y-X\beta)'(Y-X\beta)} \\ &\xrightarrow{p} \arg\min_{\beta} \frac{(\beta-\beta_0)'Q_-(\beta-\beta_0)}{\lim E\left[n^{-1}(Y-X\beta)'(Y-X\beta)\right]} = \beta_0. \end{aligned}$$

Like with LIML, there is no need to solve an explicit optimization problem. Hausman, Newey, Woutersen, Chao and Swanson (2012) note that the HLIM estimator can be computed as

$$\hat{\beta}_{HLIM} = \left(X'(P - D - \hat{\alpha} I_n)X\right)^{-1} X'(P - D - \hat{\alpha} I_n)Y$$

where $\hat{\alpha}$ is the minimal eigenvalue of $\left(\bar{X}'\bar{X}\right)^{-1} \bar{X}'(P - D)\bar{X}$.

A Fuller-type analog of the HLIM called *HFUL*, which no longer has the moment problem can be constructed in a way similar to how the FULL estimator is constructed based on LIML (Hausman, Newey, Woutersen, Chao and Swanson, 2012; Chao, Hausman, Newey, Swanson and Woutersen, 2012a); see equation (1). Hausman, Newey, Woutersen, Chao and Swanson (2012) show that the HFUL estimator does possess low order moments under certain additional assumptions.

Interestingly, Chao, Hausman, Newey, Swanson and Woutersen (2012b) show that the HLIM estimator can be viewed as a certain convex linear combination of the JIV estimator and the 'reverse' JIV estimator (that is, one where z_i's are used as instruments in a reverse regression of x_i's on y_i's). They also suggest a way to construct other estimators based on this combination that are asymptotically equivalent to HLIM and HFUL. An asymptotically optimal modification of LIML of Kunitomo (2012) modifies the projectors in the LIML objective function and obtains the *AOM-LIML* estimator that is asymptotically equivalent to HLIM under many instruments and heteroskedasticity and has some advantages in finite samples.

5.2 Asymptotic distributions

Hausman, Newey, Woutersen, Chao and Swanson (2012) derive the asymptotic distribution of the HLIM estimator under many instrument asymptotics:

$$\sqrt{n}\left(\hat{\beta}_{HLIM} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V^{HLIM}\right),$$

where

$$V^{HLIM} = V_1^{HLIM} + V_2^{HLIM} = Q_-^{-1}\Sigma_- Q_-^{-1} + \alpha Q_-^{-1}\Psi_- Q_-^{-1},$$

$\Sigma_- = \lim n^{-1}\Gamma' Z'(I - D)\mathrm{dg}\{\sigma_{e,i}^2\}_{i=1}^n (I - D)Z\Gamma$, $\Psi_- = \lim \ell^{-1}\sum_{i\neq j} P_{ij}^2 \left(\sigma_i^2 \tilde{\Omega}_j + \Sigma_{\tilde{u}e,i}\Sigma_{\tilde{u}e,j}'\right)$, with

$$\begin{pmatrix} e_i \\ \tilde{u}_i \end{pmatrix} | z_i \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \tilde{\Sigma}_i' \\ \tilde{\Sigma}_i & \tilde{\Omega}_i \end{pmatrix} \right),$$

where the variances and covariances labeled by tilde are computed for rows of

$$\tilde{U} = U - e\frac{\sum_{i=1}^n \Sigma_i'}{\sum_{i=1}^n \sigma_i^2}.$$

Note that the HLIM asymptotic variance, like that of the JIV estimator under homoskedasticity, does not contain the terms that originate from deviations of third and fourth order moments from their normal counterparts, for the same reason. In Chao, Swanson, Hausman, Newey and Woutersen (2012), the asymptotic variance of the JIV under heteroskedasticity is presented too.

The sandwich form of the first, 'traditional,' asymptotic variance component indicates the asymptotic inefficiency of HLIM. Indeed, Hausman, Newey, Woutersen, Chao and Swanson (2012) establish that under homoskedasticity, the HLIM estimator is asymptotically more efficient than the JIV estimator, but neither of the HLIM and LIML estimators dominate each other. The reason for inefficiency of HLIM is that a part of the 'signal' is removed because of jackknifing in its construction. Bekker and Crudu (2015) find a way to preserve the signal neglected in HLIM. Their *symmetric jackknife estimator* has a smaller 'traditional' component in the asymptotic variance than the HLIM estimator if conditional heteroskedasticity is not extreme, though there is no clear ordering in the 'many instrument' component.

Hausman, Newey, Woutersen, Chao and Swanson (2012) provide a valid and robust variance estimator for the HLIM estimator:

$$\hat{V}^{HLIM} = \hat{Q}_{-}^{-1}\hat{\Sigma}^{HLIM}\hat{Q}_{-}^{-1},$$

where $\hat{Q}_{-} = X'(P - D - \hat{\alpha}I_n)X$, and

$$\hat{\Sigma}^{HLIM} = \sum_{i=1}^{n}((P\hat{X})_i(P\hat{X})_i' - P_{ii}\hat{X}_i(P\hat{X})_i' - P_{ii}(P\hat{X})_i\hat{X}_i')\hat{e}_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}^2\hat{X}_i\hat{X}_j'\hat{e}_i\hat{e}_j,$$

where $\hat{X} = X - \hat{e}(\hat{e}'X/\hat{e}'\hat{e})$ and $\hat{e}$ are HLIM residuals.[3] The same construction can be applied for the asymptotically equivalent HFUL estimator. The variance estimator is robust to many instruments, i.e. it still works correctly when instruments are few. Chao, Swanson, Hausman, Newey and Woutersen (2012) describe asymptotic variance estimation for the JIV estimator too.

5.3 Asymptotic inference

Once consistent estimates of asymptotic variance are available, testing structural parameter restrictions is straightforward. For the null hypothesis

$$H_0 : c'\beta = c'\beta_0,$$

where c is $k \times 1$ vector of constants, an appropriate t-statistic and its asymptotic distribution are

$$\frac{c'(\hat{\beta}_{HLIM} - \beta_0)}{\sqrt{c'\hat{V}^{HLIM}c}} \xrightarrow{d} \mathcal{N}(0, 1),$$

and the decision rule is the usual one based on quantiles of the standard normal distribution (Hausman, Newey, Woutersen, Chao and Swanson, 2012).

Chao, Hausman, Newey, Swanson and Woutersen (2014) generalize the specification J test for the heteroskedastic case. Their statistic is based on the jackknife modification of the familiar (see subsection 4.9) quadratic form:

$$\hat{T} = \frac{\hat{e}'(P - D)\hat{e}}{\sqrt{\hat{V}}} + \ell,$$

where

$$\hat{V} = \frac{1}{\ell}\sum_{i \neq j} \hat{e}_i^2 P_{ij}^2 \hat{e}_j^2$$

[3] Hausman, Newey, Woutersen, Chao and Swanson (2012) also suggest an alternative representation of the second term that avoids double summation from 1 to n, to simplify computations.

is an estimate of the variance of the modified quadratic form. The test is one-sided, and the decision rule is to reject the null of instrument validity when

$$\hat{T} > q_{\phi}^{\chi^2(\ell-k)}.$$

Note that here the chi-squared distribution is used as an asymptotic equivalent of the normal distribution when instruments are many, as in the robust test of Anatolyev and Gospodinov (2011a) (see subsection 4.9): while the modified quadratic term is asymptotically centered normal, the additional 'ℓ' term makes it asymptotically chi-squared when ℓ is large (cf. subsection 3.2). The test described works with many instruments whether there is homo- or heteroskedasticity. It also works with few instruments under homoskedasticity, but it does not work under heteroskedasticity when instruments are few.

Crudu, Mellace, and Sandor (2017) design a hybrid parameter and specification test based on the Anderson–Rubin statistic and Bekker and Crudu (2015) symmetric jackknife objective function. In a way, this is a generalization of the Anatolyev and Gospodinov (2011a) test robust to conditional heteroskedasticity and asymptotically unbalanced instrument design.

6 Many weak instruments

6.1 Concentration parameter

Let us return to the homoskedastic version of the instrumental variables model with many instruments. Recall that another critical assumption of many instrument asymptotics, apart from $\ell/n \to \alpha > 0$ as $n \to \infty$, is the asymptotic non-triviality of information embedded in the set of instruments: $n^{-1}\Gamma' Z' Z \Gamma \to Q$, and Q is positive definite. In fact, this is a situation of *many strong instruments*: instruments are many and they are strong as a group, though most are individually weak or even irrelevant.

However, in microeconometric practice, a more typical situation is when the number of instruments is indeed large but still tiny relative to the sample size, so that the alternative assumption that $\ell \to \infty$ but $\ell/n \to 0$ (in previous notation, $\alpha = 0$) is more to the point. At the same time, the instruments are little informative and weak as a group, so that a more suitable assumption would be $n^{-1}\Gamma' Z' Z \Gamma \to 0$ (in previous notation, $Q = 0$). This is a situation of *many weak instruments* (e.g., Chao and Swanson, 2005; Stock and Yogo, 2005b).

In this and related contexts, a sensible measure of instrument informativeness is the *concentration parameter* (assuming $k = 1$)

$$\mu_n^2 = \frac{\Gamma' Z' Z \Gamma}{\Omega},$$

which measures the balance between the explained and unexplained portions in the reduced form equation. In the traditional large sample asymptotics or in the many instrument asymptotics, $\mu_n^2 = O(n)$. In the many weak instrument asymptotics, $\mu_n^2 = o(n)$.

Let us get back to the Angrist and Krueger (1991) application. We know that the instruments are numerous, as $\ell = 180$. However, this number is still tiny compared to the sample size: $n \approx 330,000$. That is, ℓ/n is essentially zero. If we look at the concentration parameter though, its value is estimated at $\mu_n^2 \approx 257$ (Hansen, Hausman and Newey, 2008). This figure is also of a much smaller order than the sample size, and it is of the same order as ℓ, the number of instruments.

6.2 Asymptotic distributions

A relatively simple asymptotic conclusion results under the following assumptions (Stock and Yogo, 2005b; Hansen, Hausman and Newey, 2008). The number of instruments is large but tiny related to the sample size:

$$\ell \to \infty, \quad \frac{\ell}{n} \to 0.$$

A proper terminology for such a situation is *moderately many* instruments as opposed to many. The set of instruments is now little informative relative to the sample size, but well informative relative to the their quantity:

$$\frac{\Gamma' Z' Z \Gamma}{\ell} \to \tilde{Q},$$

where $\tilde{Q}$ is positive definite. It follows that in terms of the concentration parameter, $\mu_n^2 = O(\ell) = o(n)$. Then (for example), the LIML estimator under error normality is asymptotically normal:

$$\sqrt{\ell}\left(\hat{\beta}_{LIML} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\tilde{Q}^{-1} + \tilde{Q}^{-1}\left(\sigma^2\Omega - \Sigma\Sigma'\right)\tilde{Q}^{-1}\right).$$

Note that the rate of convergence is $\sqrt{\ell} = o\left(\sqrt{n}\right)$. Recall that it is numerosity of instruments that gives rise to the asymptotic normality of bilinear forms contained in the LIML estimator. Note also the similarity of the form of asymptotic variance to that in the many instrument case (see subsection 4.4). Hansen, Hausman and Newey (2008) show via simulations that these asymptotics provide a good approximation when ℓ is sufficiently large even when instruments are little informative as in Angrist and Krueger (1991).

More general cases leading to more sophisticated asymptotic statements are discussed in Chao and Swanson (2005), Hansen, Hausman and Newey (2008), Hausman, Newey, Woutersen, Chao and Swanson (2012) and Chao, Swanson, Hausman, Newey and Woutersen (2012). In particular, they cover the possibilities that the number of instruments grows (i.e. $\ell \to \infty$) as fast as the sample size or slower (i.e. $\ell/n \to \alpha \geq 0$) and as fast as their informativeness (i.e., $\mu_n^2 = O(\ell)$) or even faster (i.e., $\mu_n^2 = o(\ell)$ if $\alpha = 0$). Moreover, different regressors may correspond to instruments of different degrees of informativeness: this degree is $\sqrt{n}$ for the regressors which are included instruments, and is $o(\sqrt{n})$ for the truly endogenous regressors. This leads to different rates of convergence of estimates of different coefficients in the structural form. The parameter estimators, asymptotic variance estimators and inference procedures proposed in Hansen, Hausman and Newey (2008), Hausman, Newey, Woutersen, Chao and Swanson (2012) and Chao, Swanson, Hausman, Newey and Woutersen (2012) are all robust to many weak instruments.

One of the interesting and peculiar results within this general framework is the following (see Chao and Swanson, 2005; Hansen, Hausman and Newey, 2008). Let $\ell \to \infty$ and $\ell/n \to 0$, i.e. instruments are moderately many; let $\ell/\mu_n^2 \to \infty$, i.e. instrument numerosity exceeds their informativeness; yet let $\sqrt{\ell}/\mu_n^2 \to 0$, i.e. they are not too many so that the parameter identification obtains. Relax the assumption of error normality. Then the LIML estimator has the following asymptotics:

$$\frac{\mu_n^2}{\sqrt{\ell}}\left(\hat{\beta}_{LIML}-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,V_2+V_4\right),$$

where the variance components V_2 and V_4 have the nature described in subsection 4.4. Note that the rate of convergence is dictated by the exceedance of the degree of instrument informativeness over their numerosity, and may be quite small if the 'signal' in the instrument set only slightly dominates the 'noise'. Another peculiarity is that the 'traditional' LIML variance component V_1 and 'skewness' component V_3 are dominated by the 'many instruments' component V_2 and 'excess kurtosis' component V_4 and do not survive in the limit.

6.3 Asymptotic inference

For the null hypothesis of structural parameter restrictions

$$H_0 : c'\beta = c'\beta_0,$$

where c is $k \times 1$ vector of constants, an appropriate t-statistic and its asymptotic distribution are

$$\frac{c'(\hat{\beta}_{EST}-\beta_0)}{\sqrt{c'\hat{V}^{EST}c}}\xrightarrow{d}\mathcal{N}\left(0,1\right),$$

where EST is $LIML$ or $FULL$ in a homoskedastic model and $HLIM$ or $HFUL$ in a heteroskedastic model. The decision rule is the usual one based on quantiles of the standard normal distribution. Importantly, it is shown in Hansen, Hausman and Newey (2008) and Hausman, Newey, Woutersen, Chao and Swanson (2012) that the above pivotization takes care of correct rates of convergence even if they are different across different parameters. Hansen, Hausman and Newey (2008) compute that the returns to schooling coefficient in the Angrist and Krueger (1991) application, is estimated by FULL at 0.106 with the confidence interval $(0.078, 0.134)$.

Hansen, Hausman and Newey (2008) propose a modified version of the Lagrange multiplier statistic of Kleibergen (2002) and Moreira (2009), which is asymptotically $\chi^2(k)$ distributed under the null of a true parameter vector under many weak instruments. This test can be used to construct confidence regions. In addition, Hansen, Hausman and Newey (2008) conjecture that the Stock and Yogo (2005a) test for weak instruments can be used in the context of many (possibly) weak instruments if one uses the above many weak instrument asymptotics in place of one based on few weak instruments.

7 Many regressors

Sometimes in applied practice, the number of included regressors in a mean regression is large and comparable with the number of observations. In this section, we change our focus and consider mean regression models when the number of regressors is sizable.

Apart from different empirical contexts, the many instrument setup and many regressor setup are sharply different: in their basic formulations, the former assumes an asymptotically fixed number of parameters in the estimated equation (often, just one parameter of a single endogenous variable),[4] while the latter contains many of them, an asymptotically infinite number.

7.1 Literature on many parameter models

The literature on econometrics of parametric regression models with many regressors is much thinner than that on instrumental variables models with many instruments, for several reasons. One is that setups with endogeneity and hence instrumental variables arise in microeconometric practice more often and more naturally than mean regressions with full exogeneity of all regressors. Second, the abundance of instruments may well be due to the formulation of the model in the conditional restriction form, which gives rise to a possibility of creating nonlinear combinations and interactions of a given small set of basic instruments, while a number of regressors/covariates is a built-in feature of a given conditional mean formulation. Third, big instrument sets may be driven by a desire to utilize more information in the instruments in the face of their weakness, while in the mean regression context there is no room for such a phenomenon. As a result, the many regressor econometrics has received less attention than many instruments. Nevertheless, the literature on many regressors, however thin, does exist and contributes to methods that an applied microeconometrician needs to possess. The tools of dimension asymptotics required in the many regressors framework often utilize the same or similar asymptotic theorems as those described in Section 3.

The problem with many regressors is a special case of a problem with many parameters, i.e. when the number of parameters increases to infinity with the sample size. The statistical literature was interested in estimation of parameters in a linear regression under dimension asymptotics back in 70-80's; see, e.g., Huber (1973) and Portnoy (1984, 1985). The interest was focused mainly on establishing maximal growth rates of a number of regressors relative to the sample size when the usual asymptotic properties of estimators (consistency and asymptotic normality) do not change; for example, Portnoy (1985) establishes that $(m \log m)^{3/2}/n \to 0$ is sufficient for asymptotic normality to hold. Of course, the ratio m/n is asymptotically negligible ($\mu = 0$). Koenker and Machado (1999) show that the traditional inference in general GMM problems is valid when the dimensionality of the problem grows no faster than $O\left(n^{1/3}\right)$.

The interest in the framework with $m = O(n)$ in the regression context arose early. For example, Huber (1973) pays special attention to the case when m/n is non-negligible. However, appropriate asymptotic tools were not available at the time; Huber (1973, p.82) conjectures that such a setup is "unlikely to yield to a reasonable simple asymptotic theory." Much later, when suitable large sample theorems became invented and documented, corresponding asymptotics for OLS estimation and inference were developed, in particular in Calhoun (2011) and Anatolyev (2012), who construct tests for multiple linear hypothesis about coefficients in a homoskedastic model but do not consider their estimation (see subsection 7.2). El Karoui, Bean, Bickel, Lim,

[4]... of course, not counting nuisance parameters, for example, implicitly estimated in the first stage of 2SLS.

and Yu (2013) derive the asymptotic normality of least squared estimates of parameters in a model where errors are independent of gaussian regressors. Li and Müller (2016) further propose a test for parameters of interest in a heteroskedastic model when the explanatory power of nuisance covariates have a bounded asymptotic growth. In Cattaneo, Jansson and Newey (2018b), the estimation of an asymptotically finite set of parameters of interest is considered, and consistent variance estimators in a heteroskedastic environment are developed (see subsection 7.3).

How empirically justified are these developments? On the one hand, Koenker (1988) discovers in his metastudy that the parameter dimensionality in empirical studies of cross-sectional wage equations is roughly $O\left(n^{1/4}\right)$. However, the practical importance of the case $m = O\left(n\right)$ in crystallography was mentioned as early as in Huber (1973). In economics, a perceptible ratio of m to n may occur, for example, in growth regressions, regressions run for few transition or third world countries, or predictive regressions with many predictors. Calhoun (2011) illustrates his proposed test with a macroeconomic application to monetary policy shocks and with an application based on previous studies of economic growth. Last, but not least, a big number of regressors can also result from approximating the unknown regression function by expanding to a series of approximating functions (see subsection 7.3), as in Cattaneo, Jansson and Newey (2018a).

In the remainder of this section, we describe from a practical viewpoint the methods of hypothesis testing and parameter estimation in linear regression models with many regressors.

7.2 Hypothesis testing

Earlier econometric literature noticed problems that the classical tests experience when there are many regressors and many restrictions. Berndt and Savin (1977) document huge conflicts between the classical Wald, likelihood ratio and Lagrange multiplier tests. Rothenberg (1984) discovers a big error in approximating the Wald statistic by chi-squared distribution in these circumstances. Burnside and Eichenbaum (1996) find out that the actual size of the Wald test increases sharply with a number of restrictions imposed under the null. The conventional testing tools developed under the traditional asymptotics fail to account for these phenomena. The analysis of inference in the regression context using the dimension asymptotics is undertaken in Anatolyev (2012) and Calhoun (2011).

Anatolyev (2012) considers a linear homoskedastic regression with fixed regressors

$$y_i = x_i'\gamma + e_i,$$

where $E\left[e_i\right] = 0$, $E[e_i^2] = \sigma_e^2$, and x_i and γ are $m \times 1$ vectors. A standard hypothesis containing $r \leq m$ linear restrictions

$$H_0 : R\gamma = q,$$

where q is $r \times 1$ and matrix R has full row rank r, is tested. Denote

$$x_i^* = R\left(X'X\right)^{-1} x_i,$$

and let P and P^* be projection matrices associated with x_i and x_i^*, respectively.

The Bekker-type *many regressor and* (possibly) *restriction asymptotics* assumes that as $n \to \infty$,

$$\frac{m}{n} \to \mu > 0,$$

and either r is fixed, or

$$\frac{r}{n} \to \rho > 0.$$

Anatolyev (2012) additionally assumes an asymptotically balanced design of regressors x_i and vectors x_i^* (see subsection 3.4).

The results are qualitatively different depending on whether the restrictions tested are few or many, i.e. whether r is asymptotically fixed or $\rho > 0$. The former case includes a null that an individual regression coefficient takes a particular value, most often zero, and a null of joint insignificance of several regression coefficients. The latter setup includes a test for many exclusion restrictions, e.g., joint significance test for all slope coefficients except, possibly, their small number.

If the restrictions are few, then under H_0, the standard trio of statistics, some adjusted for degrees of freedom, are asymptotically chi-squared:

$$\begin{aligned} &W \xrightarrow{d} \chi^2(r), \\ &\left(1 - \frac{m}{n}\right) LR \xrightarrow{d} \chi^2(r), \\ &\left(1 - \frac{m}{n}\right) LM \xrightarrow{d} \chi^2(r). \end{aligned}$$

If $r = 1$, the conventional t-statistic is asymptotically standard normal. That is, apart from the degrees-of-freedom adjustment for the likelihood ratio and Lagrange multiplier tests, the conventional tests work, and they are (after the adjustment) robust to the numerosity of regressors.

When the restrictions are many, the results are strikingly different. The W, LR and LM tests are asymptotically incorrect for a similar reason as the chi-square tests are distorted when there are many instruments (see subsection 3.2). Anatolyev (2012) shows that the suitably recentered and normalized test statistics are asymptotically normal in the framework of many regressors and restrictions,[5] and the asymptotic normality can be converted to asymptotically equivalent chi-squaredness. In particular,

$$\sqrt{r}\left(\frac{W}{r} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, 2\left(1 + \lambda\right)\right),$$

where

$$\lambda = \frac{\rho}{1 - \mu}$$

(cf. subsection 3.2). The emerging corrected Wald test rejects H_0 if

$$W > q_{\phi^C}^{\chi^2(r)},$$

where $\phi^C = \Phi(\sqrt{1 + r/(n - m)}\Phi^{-1}(\phi))$, and ϕ is a target significance level. That is, the corrected Wald test uses a corrected significance level so that the target significance level is asymptotically

[5] Normal asymptotic approximations for traditional asymptotic chi-squared tests can be found in earlier literature where the dimensionality grows slower than the sample size. See, for example, Hong and White (1995) for regression specification testing and Donald, Imbens and Newey (2003) for empirical likelihood based conditional moment testing.

obtained. Note that the corrected test is robust to the numerosity of regressors and restrictions as $\phi^C \to \phi$ when $\rho \to 0$, and the conventional Wald test results. The LR and LM tests can be corrected similarly, though require additional normalization.

Anatolyev (2012) also considers the 'exact F' test, which compares the F-statistic against the critical values of the Fisher-Snedecor distribution and controls the size exactly in a normal linear regression. It is shown that under an asymptotically balanced regressor design, the 'Exact F' test is asymptotically valid without error normality, irrespective of whether the regressors are many or few and whether the restrictions are many or few. Calhoun (2011) shows that when neither error normality nor asymptotically balanced design hold, there is an extra term in the asymptotic variance of the rescaled F statistic (cf. subsection 3.2). Assume, for simplicity, the conditional homokurticity of errors[6] and denote the excess kurtosis by κ. Then

$$\sqrt{r}\,(F-1) \xrightarrow{d} \mathcal{N}\left(0, 2\,(1+\lambda) + \frac{\kappa}{\rho}\lim \frac{1}{n}\sum_{i=1}^{n}\left((P_{ii}^{*}-\rho)+\lambda\,(P_{ii}^{*}-\mu)\right)^{2}\right).$$

Based on this result, Calhoun (2011) also proposes a corrected F statistic that takes into account this increase in asymptotic variance. The corrected F statistic also adapts to a situation of few regressors in which case it mimics the properties of the regular F statistic. It turns out to be very tricky to consistently estimate the excess kurtosis parameter κ using regression residuals (cf. subsection 4.11).

7.3 Parameter estimation and inference

When m is an asymptotically nonnegligible fraction of n, one has to be careful with the asymptotic properties of parameter estimates whose number is asymptotically increasing. It makes sense to focus on an asymptotically fixed number of regression parameters of interest, the others being nuisance parameters:

$$y_i = x_i'\beta_0 + z_i'\gamma + e_i, \quad E\left[e_i|x_i, z_i\right] = 0,$$

where the dimension of β_0 is fixed, while the dimension of γ grows proportionately with the sample size. The conditionally homoskedastic case falls into the instrumental variables setup with few instruments of Anatolyev (2013, Section 3) and the partial linear regression setup of Cattaneo, Jansson and Newey (2018a, subsection 3.2). It turns out that in a linear homoskedastic regression, the traditional OLS estimation and inference are valid once a degrees-of-freedom adjustment is made, despite the presence of many nuisance covariates.

Cattaneo, Jansson and Newey (2018b) consider the case of conditional heteroskedasticity. They point out that the Eicker–White estimator asymptotic variance of the OLS estimator is asymptotically invalid, and propose another estimator whose 'middle of the sandwich' matrix is constructed as

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_{ij}(MX)_i(MX)_i'\hat{e}_j^2,$$

[6]That is, almost sure constancy of their conditional fourth moments.

where M is the residual projection matrix associated with the nuisance regressors, $\hat{e}_i$'s are OLS residuals, and the matrix $K = \|\kappa_{ij}\|_{i,j=1}^{n}$ is $K = (M \odot M)^{-1}$. Such a choice of K ensures that the target conditional expectation

$$\frac{1}{n}\sum_{i=1}^{n}(MX)_i(MX)_i' E\left[e_i^2|X,Z\right]$$

matches the leading term in the conditional expectation of the variance estimator:

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_{ij}(MX)_i(MX)_i'\hat{e}_j^2\middle| X,Z\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\kappa_{ik}M_{kj}^2(MX)_i(MX)_i'E\left[e_j^2|X,Z\right] + o(1).$$

The validity of this variance estimator requires that the number of nuisance regression parameters is less numerous than half of the sample size. Interestingly, the above estimator can be viewed as an Eicker–White one with bias-corrected squared residuals. Anatolyev (2018) proposes to set $K = (M \odot M - P_M \odot P_M)^{-1}$, where P_M is a projection matrix associated with MX. This correction is aimed to reduce the bias of the variance estimator to improve the control over the test size.

Cattaneo, Jansson and Newey (2018a, 2018b) also consider a partial linear regression

$$y_i = x_i'\beta_0 + g(z_i) + e_i, \quad E\left[e_i|x_i,z_i\right] = 0,$$

where a series expansion of the nonlinear part $g(z_i)$ generates many nuisance regressors. Cattaneo, Jansson and Newey (2018a) show that under conditional homoskedasticity, the OLS estimator and usual variance estimator are consistent if $g(z_i)$ and $E\left[x_i|z_i\right]$ are approximated by linear forms in nuisance regressors well enough. Cattaneo, Jansson and Newey (2018b) derive conditions, under which their asymptotic variance estimate is valid.

Kline, Saggio and Sølvsten (2018) propose a general consistent estimator for quadratic forms in regression parameters in a many-regressor heteroskedastic linear regression. They use jackknife-type correction for the bias arising from 'diagonal' elements in the quadratic form and the leave-one-out principle for estimation of individual variances. This theory is helpful, in particular, in constructing variance estimates for and doing inference about linear combinations of regression parameters.

Numerosity of regressors may manifest itself in circumstances other than the inclusion of them as regressors in an estimated equation. Cattaneo, Jansson and Ma (2018) investigate the consequences of including a (moderately) large number of regressors in the first step regression on parameter estimates in the second step, in a general GMM framework with generated regressors. The critical rate of numerosity of regressors turns out to be $m = O\left(\sqrt{n}\right)$, in which case the numerosity of first step regressors induces an inconsistency bias in the distribution of the second step estimates, in addition to higher-order effects on the asymptotic variance. Cattaneo, Jansson and Ma (2018b) propose a hybrid procedure containing jackknifing to eliminate the asymptotic bias together with bootstrapping to account for additional variability caused by the generated regressors and jackknife bias correction.

8 Modifications and extensions

8.1 Alternative estimators

In addition to the main procedures described in the previous sections, some other estimators of structural coefficients in the (possibly weak) many instrument environments have been proposed in the literature.

In a model with only one endogenous regressor and normal errors, Chamberlain and Imbens (2004) develop a *random effects* (RE) estimator whose likelihood function puts a random coefficients structure on the coefficients in the reduced form. One of the restricted versions of the RE estimator is equivalent to LIML. The unrestricted RE estimator and its standard error bring improvements in inference in finite samples when there are many weak instruments. Chamberlain and Imbens (2004) find that the returns to schooling coefficient in the Angrist and Krueger (1991) application[7] is estimated by RE at 0.096 with the confidence interval $(0.056, 0.139)$ as opposed to 0.094 and $(0.061, 0.129)$ when estimated by LIML. The difference is small, especially compared to the 2SLS result with a (biased) point estimate of 0.073 and a (invalid) confidence interval of $(0.057, 0.089)$.

In a similar model with non-normal errors, Kolesár (2018) investigates the class of *minimum distance* (MD) estimators based on moment conditions for the two sufficient statistics. He shows that under error normality, the efficient MD estimator is identical to the RE estimator. He also proposes an efficient minimum distance estimator under error non-normality, which is asymptotically superior to the LIML estimator.

Bekker and Wansbeek (2016) construct a *concentrated instrumental variables* estimator that uses 2SLS in two (or more) steps thus avoiding LIML. In the first step, a 'concentrated' few-dimensional instrument is constructed, which is used via 2SLS in the second step (which may be iterated to achieve consistency). Under conditional homoskedasticity and error normality, this estimator is asymptotically equivalent to LIML.

Sølvsten (2017) adapts LIML estimation to non-normal, possibly thick-tailed structural errors. His *optimal robust estimator* is minimax optimal in a class of estimators that includes LIML. Sølvsten (2017) establishes consistency and asymptotic normality under many instrument asymptotics, and also provides a consistent variance estimator. In simulations, it is found that efficiency gains from the optimal robust estimator when may be substantial relative to that of LIML when the structural errors have thick tails.

8.2 Bootstrapping

It is well known that the presence of weak instruments (recall that many instruments imply that almost all instruments are individually weak) invalidates conventional bootstraps. Wang and Kaffo (2016) consider a variety of bootstrap procedures for the LIML and FULL estimators in the context of a conditionally homoskedastic regression with many or many weak instruments. One is the standard residual bootstrap, with instruments staying fixed in bootstrap samples. Another is

[7] Chamberlain and Imbens (2004) use a cut sample (only from the 1st and 4th quarters) and an inflated set of instruments (in addition to interactions with states and years, interactions with state/year combinations).

Davidson and MacKinnon's (2008) restricted efficient bootstrap, where the null is imposed in bootstrap samples and reduced form estimates are more efficient. Wang and Kaffo (2016) show that these bootstraps are not asymptotically valid because the instrument strength is exaggerated in bootstrap samples, and, as a result, the asymptotic variance and even the rate of convergence may be distorted in them. Wang and Kaffo (2016) propose modifications of the restricted efficient bootstrap that allow one to reach asymptotic validity, at least under asymptotically balanced instrument design (see subsection 3.4) or when instruments are few compared to sample size but whose number is comparable to the concentration parameter. This is achieved by appropriately rescaling the residuals and using alternative reduced form estimates. Wang and Kaffo (2016) verify via simulations that their bootstrap inferences, both percentile and percentile-t, yield a much more correct size control than the baseline alternatives, and have good power properties. Kaffo and Wang (2017) propose bootstrap alternatives to Anatolyev and Gospodinov's (2011) corrected J and AR tests under asymptotically balanced instrument design.

In the regression context, El Karoui and Purdom (2016) explain why the residual bootstrap does not work and needs modification when regressors are many; the pairs bootstrap has problems as well. Richard (2017) studies the properties of bootstrap tests under many regressors and many restrictions. In the framework of Anatolyev (2012), where the regressors have an asymptotically balanced design, he shows the asymptotic validity of the (fixed regressor) bootstrap F, LR and LM tests, which are also robust to the numerosity of regressors and restrictions. Under an asymptotically unbalanced design, Richard (2017) proposes bootstrapping Calhoun's (2011) modified F statistic instead. Richard (2017) notices that because of a mismatch between the bootstrap and sample third cumulants, such a bootstrap is not likely to provide asymptotic refinements in general.

8.3 Panel data models

Dynamic panel data models provide an environment where many instruments arise naturally after first differencing. If T is the panel's time dimension, the number of instruments implied by the difference or system GMM procedures is of order $O(T^2)$, and if T is even moderately large, the number of instruments can be comparable to the cross-sectional dimension. Some of the related literature focuses on deriving biases of various GMM estimators (e.g., Bun and Kiviet, 2006) or simply reducing the instrument count by limiting lag depth (see Roodman, 2009). Other literature tries to invent ways to eliminate biases or constructing alternative estimators. For example, Okui (2009) proposes algorithms to select moderately many instruments when $T = o(n)$. Arellano (2016), using an auxiliary parametric model for the instruments, constructs an optimal instrument such that the corresponding estimator lacks the biases in the $T = O(n)$ asymptotic setup and has optimality properties in the fixed T setup.

Alvarez and Arellano (2003) analyze a number of estimators, including GMM and LIML, in an autoregressive dynamic panel model under the joint asymptotics $n, T \to \infty$ when $T/n \to \tau > 0$. They utilize all available instruments in differences, whose number is of order $\ell = O(T^2) = O(n^2)$ so the ratio of the number of orthogonality conditions to the total sample size $\ell/(nT)$ is asymptotically a nonzero constant. Interestingly, despite such abundant instrument numerosity, both GMM and LIML estimators of the autoregressive coefficient turn out to be consistent, asymptotically normal, and asymptotically equally efficient, though having different higher-order biases. Alvarez and Arellano (2003) explain the consistency of GMM by the fact that the endogeneity bias asymptotically goes down. Akashi and Kunitomo (2012) consider a dynamic panel model with generic endogenous regressors. They show that when $T/n \to \tau > 0$, the differences-GMM estimator is inconsistent, while the LIML estimator is consistent and asymptotically normal with asymptotic variance containing four components along the lines of subsection 4.5.

8.4 Spatial data models

In spatial autoregressions, the spatial autoregressive term induces endogeneity, hence causing a problem of instrumental variables estimation. Liu (2012) considers the genuine LIML estimator and shows that if instruments are moderately many, i.e. $\ell = o(n)$, then the LIML estimator is consistent, but it is not if instrument are many, i.e. $\ell = O(n)$, because of spatial correlation. Liu and Lee (2013) consider instead 2SLS estimation and aim at the selection of (moderately many) instruments using Donald and Newey's (2001) criterion of minimization of the approximate mean squared error.

8.5 Robust regression

El Karoui, Bean, Bickel, Lim and Yu (2013) consider a linear homoskedastic many-regressor regression model whose parameters are estimated via loss functions other than quadratic. It turns out that non-quadraticity of the loss leads to additional complications. In particular, the norm of the difference between the vectors of estimates and true parameters asymptotically depends on the error distribution. Another peculiarity is that the residuals cease to be a linear transformation of the true errors; this, in particular, limits possibilities of residual bootstrapping (El Karoui and Purdom, 2016). These issues are an active research area in high-dimensional statistics.

8.6 Time series models

Unfortunately, when there is dependence across equations for different i, the tools described in Section 3 do not work. Consider, for example, the same quadratic form. Recall that if there is no dependence, under many instrument asymptotics

$$\frac{e'Pe}{n} \xrightarrow{p} \alpha\sigma^2.$$

Under dependence, the same result is not likely to hold. In an attempt to compute an expectation of the quadratic form, if one conditions on Z as before,

$$E\left[\frac{e'Pe}{n}\right] = E\left[\frac{E\left[\mathrm{tr}(e'Pe)|Z\right]}{n}\right] = E\left[\frac{E\left[\mathrm{tr}(PE[ee'|Z])\right]}{n}\right],$$

this leads to nowhere as $E[ee'|Z]$ is likely to depend on Z in a complex way even under conditional homoskedasticity because there is cross-equational dependence, in particular, between e_ie_j and z_k for $k > \max\{i, j\}$ if the instruments are predetermined but not strictly exogenous, as is often the case in time series models.

8.7 Nonlinear models

The methods related to many instruments and/or regressors are derived for linear models and are hard to generalize to nonlinear equations. Han and Phillips (2006) develop a general theory for GMM with many (possibly weak) moment conditions, including the case of many moments, i.e. $\ell = O(n)$ where now ℓ stands for a number of moment conditions. Newey and Windmeijer (2009) present an extension to continuously updated GMM (an analog of LIML), jackknife GMM (an analog of JIV) and generalized empirical likelihood (GEL) estimators with many weak moment conditions when the moments are moderately many, i.e. $\ell = o(n)$.

8.8 Very many instruments/regressors

There is a broad econometric literature about dealing with big sets of instruments, with cardinality possibly exceeding a sample (i.e. $\ell > n$) or even being a continuum. This literature is inspired by existing statistical techniques related to big data processing. For example, a lasso-type dimension reduction is proposed in Belloni, Chen, Chernozhukov and Hansen (2012); some regularization procedures for the LIML estimator are also proposed in Carrasco and Tchuente (2015). These methods aim at extracting the most useful and relevant information from a host of instruments, and often rely on strong assumptions about the data structure such as sparsity. Belloni, Chernozhukov and Hansen (2014) analogously propose handling a very many regressor problem that may result, in particular, from the approximation of unknown nonlinear forms in partially linear regressions.

Methods of dealing with very many instruments/regressors are useful, should a researcher be endowed with such big instrument sets. Afterwards, the many (weak) instrument/regressor techniques described here may be applied as the resulting instrument sets may still be sizable and call for improved estimation and inference.

9 Empirical strategies

Previous sections describe a lot of econometric tools pertaining to estimation and inference in linear models with many instruments and/or regressors. In this section we give some guidance on which tools should be applied when an empirical researcher is confronted with one or both phenomena depending on the particularities of a given empirical problem.

9.1 Many instruments

This survey focuses on those cases when the set of instruments is large in the sense that the number of instruments is comparable to the sample size, provided that the instruments are strong as a group (the case of many instruments) or is moderately large in the sense that their number is large but much smaller compared to the sample size, provided that the instruments are weak as a group (the case of many weak instruments). Suppose that a researcher instead faces a linear IV setup in which the number of instruments is moderately large and that there are no doubts about their strength, a situation falling outside the focus of this survey.[8] A researcher may then apply methods for selecting the most informative and least noisy instruments from the given set by exploiting the trade-off between first order asymptotic variance and higher order asymptotic bias and/or variance to reach minimal asymptotic MSE. See Donald and Newey (2001) on selecting instruments in a linear homoskedastic model, and Donald, Imbens, and Newey (2009) on selecting instruments in a conditional nonlinear heteroskedastic model. Another, possibly better, option, is to employ shrinkage or model averaging methods for the reduced form to construct the optimal instrument, as these methods, unlike instrument selection, utilize information from the whole set of instruments and do not require the pre-ordering of instruments. Okui (2011) suggests shrinking some OLS coefficient estimates in the first stage of 2SLS, the shrinkage intensity being guided by the asymptotic higher order MSE criterion, and then using the predicted values as instruments; Canay (2010) exploits the same idea but shrinks coefficients in a smoother manner and uses heteroskedasticity-robust GMM estimation. Kuersteiner and Okui (2010) propose a model averaging procedure that generalizes instrument selection and dominates it in terms of asymptotic higher order MSE.[9]

In the opposite situation, when the number of instruments is huge and even exceeds the sample size, dimension reduction methods are called for; one can recommend, in particular, lasso-type shrinkage tools developed in Belloni, Chen, Chernozhukov and Hansen (2012) for conditionally heteroskedastic, non-gaussian models. One has to be aware that their application requires a sparse or approximately sparse structure of the reduced form, which essentially means that all information is concentrated in at most a moderately large subset of instruments. When the information instead is scattered across many instruments, this method is not meant to achieve that goal.[10] An alternative option is to use various regularization methods as in Hansen and Kozbur (2014), Carrasco and Tchuente (2015) and Carrasco and Doukali (2017).[11] These methods do not require sparsity, ordering or strength of instruments; all instruments are used even if they are weak.

If the setup does fall into the category this survey is focused on – the instruments are many and strong (Sections 4–5) or the instruments are moderately many and weak as a group but their strength is comparable or smaller than their numerosity (Section 6) – the decision of which exact tools to apply depends on whether there is conditional homoskedasticity and, if yes, whether the design of the instrument projection matrix is asymptotically balanced and/or the errors have gaussian third and fourth moments. Most recent developments provide most robust tools that will work irrespective of homo/heteroskedasticity and the instrument design; they are also robust to whether the instruments are strong or weak as a group as well as to differences in rates of

[8]This situation formally may be considered as a trivial case of the many instrument theory, when the noise coming from instrument numerosity is negligible. The described tools can then still be applied, though without a point.

[9]These methods do require $\ell = o(n)$; Donald and Newey (2001), for example, impose the conditions $\ell/n \to 0$ and $\max_{1\leq i\leq n} P_{ii} \xrightarrow{p} 0$, and so do Okui (2011) and Kuersteiner and Okui (2010), while Canay (2010) requires $\ell = o(n^{1/2})$.

[10]What is critical in the simulation design in Belloni, Chen, Chernozhukov and Hansen (2012) is that the information is distributed exponentially across the instruments, which essentially makes only a few of them strong. Setting $\ell = 100$ with $n = 100$ or 250, i.e. ℓ not exceeding n, the authors can apply LIML/FULL as well. They report that "the simulation results verify that FULL becomes more appealing as the sparsity assumption breaks down" (p.2403).

[11]In general, the method of Belloni, Chen, Chernozhukov and Hansen (2012) allows for huge ℓ restricted above by $\log \ell = O(n^{1/3})$. Carrasco and Tchuente (2015) and Carrasco and Doukali (2017) do not at all restrict the number of instruments which may be even infinite.

convergence across different parameters. Naturally, these tools are robust to few instruments and the type of asymptotic framework. Apart from and in addition to that, the literature seems to have converged to the dominance of Fuller-type modification over LIML-based estimators, because of the existence of their moments.

Thus, in the general case, the most preferable estimator is HFUL (with $C = 1$) described in subsection 5.1, accompanied with its variance estimator described in subsection 5.2, and the most preferable specification test is described in subsection 5.3.[12] If the model is homoskedastic (or sufficiently close to homoskedastic), however, there is a choice. There may be benefits for switching to estimators and tests targeted to the conditionally homoskedastic model, as the (LIML and FULL) parameter estimates may be more efficient (see subsection 5.2), the variance estimators and, consequently, the test statistics have a simpler form (see subsection 4.7 vs subsection 5.2) and thus contain less noise in finite samples. If, in addition, the instrument design seems to be sufficiently balanced (which can be verified directly from the data on instruments) or the error distribution has zero skewness and zero excess kurtosis (which is testable), further simplifications and thus finite sample improvements are possible (see subsections 4.5 and 4.9). An alternative is to sacrifice some efficiency and keep using HFUL under conditional homoskedasticity. Another reason for doing that is, apart from the universal robustness of HFUL, the absence of third and forth moment terms in its asymptotic variance (unlike that of LIML or FULL) – the terms that are most tedious to estimate (see subsection 4.7 vs subsection 5.2). Yet another phenomenon an empirical researcher has to watch for is the possible numerosity of covariates in the structural equation (see subsection 4.11).[13]

Often, the instrument set is formed by the researcher, for example, by adding interactions among those already in use or by adding their nonlinear functions. Thus, there is a choice of which asymptotic framework the problem eventually falls into. While more instruments means more asymptotic efficiency, newly generated instruments taken together may be weak, which may not justify their inclusion given the increasing dimensionality of the asymptotic framework.[14] On the other hand, by including them up to a reasonable limit and using the robust tools described above does not really do any harm.

Which relationships between instrument numerosity ℓ and sample size n may be characterized as 'many instruments' or 'moderately many instruments'? Bruce Hansen notes in his online textbook (Hansen, 2018): "*In an application, users should calculate the 'many instrument ratio'* $\alpha = \ell/n$*. Unfortunately, there is no known rule-of-thumb for* α *which should lead to acceptable inference,*

[12]An exception is that the specification test is not robust to heteroskedasticity coupled with few instruments; see subsection 5.3.

[13]To the best of our knowledge, numerosity of exogenous covariates has not yet been incorporated in the most general framework.

[14]To date, there are no available methods on instrument selection/shrinkage/model averaging applicable to the 'many instrument' case $\ell = O(n)$ in contrast to those cited above for the case $\ell = o(n)$. Likewise, there are no methods for testing for instrument weakness in this framework, analogous to the Stock and Yogo (2005a) test, although Hansen, Hausman and Newey (2008) conjecture that this test can still be used in the context of many (possibly weak) instruments if one uses the many weak instrument asymptotics in place of the few weak instruments asymptotics.

but a minimum criterion is that if $\alpha \geq 0.05$ you should be seriously concerned about the many-instrument problem." To try to answer the question more precisely, one can trace simulation sections of theoretical studies. For example, the simulations in Donald and Newey (2001) and Kuersteiner and Okui (2010) are tuned to $n = 100$ with $\ell = 20$ and $n = 1000$ with $\ell = 30$; those in Donald, Imbens, and Newey (2009) are tuned to $n = 200$ with $\ell = 10$ and $n = 800$ with $\ell = 20$. All three papers consider designs with the first stage $R^2 = 0.10$, although sometimes also include the weak instrument design with $R^2 = 0.01$. These are situations of 'moderately many instruments,' in the latter case coupled with weak instruments. Anatolyev and Gospodinov (2011a) set $\ell/n = 0.04, 0.2, 0.5, 0.8$. While the first and possibly second value also correspond to the same situation, the other two clearly belong to the 'many instruments' case. When instruments are weak, the relationship between instrument numerosity ℓ and concentration parameter μ_n^2 becomes important as well. For example, Hausman, Newey, Woutersen, Chao and Swanson (2012) set $n = 800$ with $\ell = 30$ and $\mu_n^2 = 8, 32$, as well as with $\ell = 10$ and $\mu_n^2 = 16$. Chao, Hausman, Newey, Swanson, and Woutersen (2014) choose $n = 800$ with $\ell = 10, 30, 50$ and $\mu_n^2 = 8, 32$. Bekker and Crudu (2015) use the same design, also setting $n = 800$ with $\mu_n^2 = 8, 32$ but $\ell = 2, 5, 15$. These figures are able to provide an idea about what kind of practical situations are meant to fit the theory.

9.2 Many regressors

In models with many regressors, the situation is more straightforward, as there are no weakness issues. If the regressor set is numerous but still very small compared to the sample size ('moderately many regressors'), one may still apply traditional asymptotic tools for estimation and inference. If, on the contrary, the regressors are abundant ('very many regressors') because of, for example, lots of potential covariates may be determinants of the outcome variable, then dimension reduction tools, such as lasso-based methods of Belloni, Chernozhukov and Hansen (2014), are in order. However, if the number of regressors is smaller but comparable to the number of observations ('many regressors'), one should apply the methods described in Section 7 of this survey.

The estimator of choice is OLS, which is consistent for coefficients of interest (of which there should be few). If the target is to test (few or many) restrictions about regression parameters in a homoskedastic regression (for example, in the course of model building), the relevant tools are described in subsection 7.2.[15] The preferred test is the F test, possibly robustified against an asymptotically unbalanced regressor design. Under heteroskedasticity, the appropriate variance estimator for OLS estimates of coefficients of interest is given in subsection 7.3. The 'many regressors/restrictions' situation is distinguished from the 'moderately many regressors/restrictions' situation in the same way as 'many instruments' are distinguished from 'moderately many instruments.'

[15]To the best of our knowledge, there are no available methods to date on testing many restrictions in the conditionally heteroskedastic regression with many regressors.

10 Concluding remarks

We have surveyed the state of the art in the econometrics of regression models with many instruments and/or many regressors. While the ideas behind the modern analysis of such models were put forth several decades ago in 1980-90s, a burst of developments has occurred quite recently within the last decade, and some are still occurring. There are some open questions regarding, in particular, estimation efficiency, the selection of instruments, and the applicability to nonlinear and time series models. There is also, unfortunately, a significant gap between the existing theory and prevailing practice, as often happens when alternative approximation ideas and advanced econometric methods are involved. This survey will hopefully narrow that gap.

References

Abutaliev, A. and S. Anatolyev (2013): "Asymptotic variance under many instruments: numerical computations," *Economics Letters*, 118(2), 272–274.

Ackerberg, D.A. and P.J. Devereux (2009): "Improved JIVE estimators for overidentified linear models with and without heteroskedasticity," *Review of Economics and Statistics*, 91(2), 351–362.

Akashi, K. and N. Kunitomo (2012): "Some properties of the LIML estimator in a dynamic panel structural equation," *Journal of Econometrics*, 166(2), 167–183.

Alvarez, J. and M. Arellano (2003): "The time series and cross-section asymptotics of dynamic panel data estimators," *Econometrica*, 71(4), 1121–1159.

Anatolyev, S. (2012): "Inference in regression models with many regressors," *Journal of Econometrics*, 170(2), 368–382.

Anatolyev, S. (2013): "Instrumental variables estimation and inference in the presence of many exogenous regressors," *Econometrics Journal*, 16(1), 27–72.

Anatolyev, S. (2018): "Almost unbiased variance estimation in linear regressions with many covariates," *Economics Letters*, 169, 20–23.

Anatolyev, S. and N. Gospodinov (2011a): "Specification testing in models with many instruments," *Econometric Theory*, 27(2), 427–441.

Anatolyev, S. and N. Gospodinov (2011b): *Methods for Estimation and Inference in Modern Econometrics*. Boca Raton, CRC Press–Taylor & Francis.

Anatolyev, S. and P. Yaskov (2017): "Asymptotics of diagonal elements of projection matrices under many instruments/regressors," *Econometric Theory*, 33(3), 717–738.

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

Anderson, T.W., N. Kunimoto and Y. Matsushita (2010): "On the asymptotic optimality of the LIML estimator with possibly many instruments," *Journal of Econometrics*, 157(2), 191–204.

Anderson, T.W., N. Kunimoto and Y. Matsushita (2011): "On finite sample properties of alternative estimators of coefficients in a structural equation with many instruments," *Journal of Econometrics*, 165(1), 58–69.

Anderson, T.W. and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics*, 20(1), 46–63.

Angrist, J.D., G.W. Imbens and A. Krueger (1999): "Jackknife instrumental variables estimation," *Journal of Applied Econometrics*, 14(1), 57–67.

Angrist, J. and A. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?" *Quarterly Journal of Economics*, 106(4), 979–1014.

Arellano, M. (2016): "Modelling optimal instrumental variables for dynamic panel data models," *Research in Economics*, 70(2), 238–261.

Bai, Z. and J. Silverstein (2010): *Spectral analysis of large dimensional random matrices*, 2$^{\text{nd}}$ edition, New York, Springer.

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

Bekker, P.A. and F. Crudu (2015): "Jackknife instrumental variable estimation with heteroskedasticity," *Journal of Econometrics*, 185(2), 332–342.

Bekker, P.A. and J. van der Ploeg (2005): "Instrumental variable estimation based on grouped data," *Statistica Neerlandica*, 59(3), 239–267.

Bekker, P.A. and T. Wansbeek (2016): "Simple many-instruments robust standard errors through concentrated instrumental variables," *Economics Letters*, 149, 52–55.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 81(2), 608–650.

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.

Berndt, E.R. and N. E. Savin (1977): "Conflict among criteria for testing hypotheses in the multivariate linear regression model," *Econometrica*, 45(5), 1263–1277.

Blomquist, S. and M. Dahlberg (1999): "Small sample properties of LIML and jackknife IV estimators: experiments with weak instruments," *Journal of Applied Econometrics*, 14(1), 69–88.

Bun, M. and J. Kiviet (2006): "The effects of dynamic feedbacks on LS and MM estimator accuracy in panel data models," *Journal of Econometrics*, 132(2), 409–444.

Burnside, C. and M. Eichenbaum (1996): "Small sample properties of GMM-based Wald tests," *Journal of Economic & Business Statistics*, 14(3), 294–308.

Calhoun, G. (2011): "Hypothesis testing in linear regression when k/n is large," *Journal of Econometrics*, 165(2), 163–174.

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

Carrasco, M.and M. Doukali (2017): “Efficient estimation using regularized jackknife IV estimator,” *Annals of Economics and Statistics*, 128, 109–149.

Carrasco, M.and G. Tchuente (2015): “Regularized LIML for many instruments,” *Journal of Econometrics*, 186(2), 427–442.

Cattaneo, M. D., M. Jansson, and X. Ma (2018): “Two-step estimation and inference with possibly many included covariates,” *Review of Economic Studies*, accepted for publication.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018a): “Alternative asymptotics and the partially linear model with many regressors,” *Econometric Theory*, 34(2), 277–301.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018b): “Inference in linear regression models with many covariates and heteroskedasticity,” *Journal of American Statistical Association*, `DOI: 10.1080/01621459.2017.1328360`.

Chamberlain, G. (1987): “Asymptotic efficiency in estimation with conditional moment restrictions,” *Journal of Econometrics*, 34(3), 305–334.

Chamberlain, G. and G. Imbens (2004): “Random effects estimators with many instrumental variables,” *Econometrica*, 72(1), 295–306.

Chao, J.C., J.A. Hausman, W.K. Newey, N.R. Swanson, and T. Woutersen (2012a): “An expository note on the existence of moments of Fuller and HFUL estimators,” in B.H. Baltagi, R.C. Hill, W.K. Newey, H.L. White (ed.) Essays in Honor of Jerry Hausman (*Advances in Econometrics*, 29), New York, Emerald Group Publishing Limited, 87–106.

Chao, J.C., J.A. Hausman, W.K. Newey, N.R. Swanson, and T. Woutersen (2012b): “Combining two consistent estimators,” in B.H. Baltagi, R.C. Hill, W.K. Newey, H.L. White (ed.) Essays in Honor of Jerry Hausman (*Advances in Econometrics*, 29), New York, Emerald Group Publishing Limited, 33–53.

Chao, J.C., J.A. Hausman, W.K. Newey, N.R. Swanson, and T. Woutersen (2014): “Testing overidentifying restrictions with many instruments and heteroskedasticity,” *Journal of Econometrics*, 178(1), 15–21.

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.

Chioda, L. and M. Jansson (2009): “Optimal invariant inference when the number of instruments is large,” *Econometric Theory*, 25(3), 793–805.

Crudu, F., G. Mellace, and Z. Sandor (2017): "Inference in instrumental variables models with heteroskedasticity and many instruments," manuscript, University of Siena.

Davidson, R. and J.G. MacKinnon (2006): "The case against JIVE," *Journal of Applied Econometrics*, 21(6), 827–833.

Davidson, R. and J.G. MacKinnon (2008): "Bootstrap inference in a linear equation estimated by instrumental variables," *Econometrics Journal*, 11(3), 443–477.

de Jong, P. (1987): "A central limit theorem for generalized quadratic forms," *Probability Theory and Related Filds*, 75(2), 261–277.

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

Donald, S.G., G.W. Imbens, and W.K. Newey (2003): "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics*, 117(1), 55–93.

Donald, S.G., G.W. Imbens, and W.K. Newey (2009): "Choosing instrumental variables in conditional moment restriction models," *Journal of Econometrics*, 152(1), 26–38.

El Karoui, N., D. Bean, P.J. Bickel, C. Lim, and B. Yu (2013): "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, 110(36), 14557–14562.

El Karoui, N. and E. Purdom (2016): "Can we trust the bootstrap in high-dimension?" manuscript, `arXiv:1608.00696`.

Evdokimov, K.S. and M. Kolesár (2018): "Inference in instrumental variables analysis with heterogeneous treatment effects," manuscript, Princeton University.

Fuller, W.A. (1977): "Some properties of a modification of the limited information estimator," *Econometrica*, 45(4), 939–954.

Götze F. and Tikhomirov A.N. (1999): "Asymptotic distribution of quadratic forms," *Annals of Probability*, 27(2), 1072–1098.

Hahn, J. (2002): "Optimal inference with many instruments," *Econometric Theory*, 18(1), 140–168.

Hahn, J. and J. Hausman (2002): "A new specification test for the validity of instrumental variables," *Econometrica*, 70(1), 163–189.

Hansen, B.E. (2018): *Econometrics*, online textbook, University of Wisconsin-Madison, available at `www.ssc.wisc.edu/~bhansen/econometrics`.

Hansen, L.P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50(4), 1029–1054.

Hansen, C., J. Hausman and W.K. Newey (2008): "Estimation with many instrumental variables," *Journal of Business & Economics Statistics*, 26(4), 398–422.

Hansen, C. and D. Kozbur (2014): "Instrumental variables estimation with many weak instruments using regularized Jive", *Journal of Econometrics*, 182(2), 290–308.

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.

Hong, Y. and H. White (1995): "Consistent specification testing via nonparametric series regression," *Econometrica*, 63(5), 1133–1159.

Huber, P.J. (1973): "Robust regression: Asymptotics, conjectures and Monte Carlo," *Annals of Statistics*, 1(5), 799–821.

Imbens, G. (2010): "Lectures on evaluation methods: weak instruments," Lecture notes 4, Impact Evaluation Network, Miami.

Kaffo, M. and W. Wang (2017): "On bootstrap validity for specification testing with many weak instruments," *Economics Letters*, 157, 107–111.

Koenker, R. (1988): "Asymptotic theory and econometric practice," *Journal of Applied Econometrics*, 3(2), 139–147

Koenker, R. and J.A.F. Machado (1999): "GMM Inference when the number of moment conditions is large," *Journal of Econometrics*, 93(2), 327–344.

Kolesár, M. (2018): "Minimum distance approach to inference with many instruments," *Journal of Econometrics*, 204(1), 86–100.

Kelejian, H.H. and I.R. Prucha (2001): "On the asymptotic distribution of the Moran I test statistic with applications," *Journal of Econometrics* 104(2), 219–257.

Kleibergen, F. (2002): "Pivotal statistics for testing structural parameters in instrumental variables regression," *Econometrica*, 70(5), 1781–1803.

Kline, P., R. Saggio, and M. Sølvsten (2018): "Leave-out estimation of variance components," manuscript, `arXiv:1806.01494`.

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

Kunitomo, N. (1980): "Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations," *Journal of American Statistical Association*, 75(371), 693–700.

Kunimoto, N. (2012): "An optimal modification of the LIML estimation for many instruments and persistent heteroscedasticity," *Annals of the Institute of Statistical Mathematics*, 64(5), 881–910.

Ledoit, O. and M. Wolf (2002): "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size," *Annals of Statistics*, 30(4), 1081–1102.

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

Leeb, H. and B.M. Pötscher (2005): "Model selection and inference: facts and fiction," *Econometric Theory*, 21(1), 21–59.

Li, C. and U.K. Müller (2016): "Linear regression with many controls of limited explanatory power," manuscript, Princeton University.

Liu, X. (2012): "On the consistency of the LIML estimator of a spatial autoregressive model with many instruments," *Economics Letters*, 116(3), 472–475.

Liu, X. and L.-f. Lee (2013): "Two stage least squares estimation of spatial autoregressive models with endogenous regressors and many instruments," *Econometric Reviews*, 32(5-6), 734–753.

Marchenko, V.A. and L.A. Pastur (1967): "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, 72(114), 507–536.

Moreira, M.J. (2009): "Tests with correct size when instruments can be arbitrarily weak," *Journal of Econometrics*, 152(2), 131–140.

Morimune, K. (1983): "Approximate distributions of 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), 573–595.

Newey, W.K. (1990): "Efficient instrumental variables estimation of nonlinear models," *Econometrica*, 58(4), 809–837.

Newey, W.K. (2004): "Many instrument asymptotics," manuscript, MIT.

Newey, W.K. and R.J. Smith (2004): "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72(1), 219–255.

Okui, R. (2009): "The optimal choice of moments in dynamic panel data models", *Journal of Econometrics*, 151(1), 1–16.

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

Poi, B. (2006): "Jackknife instrumental variables estimation in Stata," *Stata Journal*, 6(3), 364–376.

Portnoy, S. (1984): "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency," *Annals of Statistics*, 12(4), 1298–1309.

Portnoy, S. (1985): "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation," *Annals of Statistics*, 13(4), 1403–1417.

Richard, P. (2017): "Bootstrap tests in linear models with many regressors," manuscript, Université de Sherbrooke.

Roodman, D. (2009): "A note on the theme of too many instruments," *Oxford Bulletin of Economics and Statistics*, 71(1), 135–158.

Rothenberg, T.J. (1984): "Approximating the distributions of econometric estimators and test statistics." In: Griliches, Z. and M.D. Intriligator, eds., *Handbook of Econometrics*, vol. 2. New York, North-Holland.

Serdobolskii, V. (2007): *Multiparametric Statistics*, Elsevier Science.

Sølvsten, M. (2017): "Robust estimation with many instruments," manuscipt, University of Wisconsin.

Staiger, D. and J.H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65(3), 557–586.

Stock, J.H. and M. Yogo (2005a): "Testing for weak instruments in linear IV regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by D.W.K. Andrews and J.H. Stock, Cambridge: Cambridge University Press, chapter 5, 80–108.

Stock, J.H. and M. Yogo (2005b): "Asymptotic distributions of instrumental variables statistics with many instruments," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by D.W.K. Andrews and J.H. Stock, Cambridge: Cambridge University Press, chapter 6, 109–120.

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

Wang, W. and M. Kaffo (2016): "Bootstrap inference for instrumental variable models with many weak instruments," *Journal of Econometrics*, 192(1), 231–268.

Yao, J., S. Zheng, and Z. Bai (2015): *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge, Cambridge University Press.