Efficient estimation with many weak instruments using regularization techniques*

Marine Carrasco
University of Montreal, CIREQ, CIRANO

Guy Tchuente
University of Kent

January 2015

Abstract

The problem of weak instruments is due to a very small concentration parameter. To boost the concentration parameter, we propose to increase the number of instruments to a large number or even up to a continuum. However, in finite samples, the inclusion of an excessive number of moments may be harmful. To address this issue, we use regularization techniques as in Carrasco (2012) and Carrasco and Tchuente (2014). We show that normalized regularized 2SLS and LIML are consistent and asymptotically normally distributed. Moreover, our estimators are asymptotically more efficient than most competing estimators. Our simulations show that the leading regularized estimators (LF and T of LIML) work very well (are nearly median unbiased) even in the case of relatively weak instruments. An application to the effect of institutions on output growth completes the paper.

Key Words: Many weak instruments, LIML, 2SLS, regularization methods, semiparametric efficiency bound.

*The authors wish to thank the editors and two referees for helpful comments. Carrasco gratefully acknowledges financial support from SSHRC.

1 Introduction

This paper considers the estimation of the regression coefficient in a linear model using many weak instruments. The literature on weak instruments (see e.g. Staiger and Stock (1997)) has initially focused on the case where the number of instruments is fixed and the correlation between the endogenous regressors and the instruments goes to zero at the $\sqrt{n}$ rate. In this setting, the parameter is not identified and two-stage least squares (2SLS) and limited maximum likelihood (LIML) estimators are not consistent and converge to nonstandard distributions. Subsequent work (see, e.g. Chao and Swanson (2005), Hansen, Hausman, and Newey (2008), and Newey and Windmeijer (2009)) focused on situations where the number of instruments grows with the sample size and the correlation between the endogenous regressors and the instruments goes to zero at a rate greater than $\sqrt{n}$. Consequently, the parameter is identified. However, the best rate you can get for any estimator is slower than the usual $\sqrt{n}$ and the Gaussian asymptotic approximation may be poor. Our paper fits in the second strand of literature. The problem of many weak instruments has recently received considerable attention in both theoretical and applied econometrics. Empirical examples include Angrist and Krueger (1991) who estimate the return to schooling, Eichenbaum, Hansen, and Singleton (1988) who consider consumption asset pricing models.

Building on the early work by Carrasco and Florens (2000), Carrasco (2012) and Carrasco and Tchuente (2014) proposed respectively regularized versions of 2SLS and LIML estimators for many strong instruments. The regularization permits to address the singularity of the covariance matrix resulting from many instruments. These papers use three regularization methods borrowed from inverse problem literature. The first estimator is based on Tikhonov (ridge) regularization, the second estimator is based on an iterative method called Landweber-Fridman (LF), the third estimator is based on the principal components associated with the largest eigenvalues. We extend these previous works to allow for the presence of a large number of weak instruments or weak identification. We consider a linear model with homoskedastic error and allow for weak identification as in Hansen, Hausman, and Newey (2008) and Newey and Wind-

meijer (2009). This specification helps us to have different types of weak instruments sequences, including the many instruments sequence of Bekker (1994) and the many weak instruments sequence of Chao and Swanson (2005). We impose no condition on the number of moment conditions since our framework allows for an infinite countable or even a continuum of instruments. The advantage of regularization is that all available moments can be used without discarding any a priori. We show that regularized 2SLS and LIML are consistent in the presence of many weak instruments. If properly normalized, the regularized 2SLS and LIML are asymptotically normal and reach the semiparametric efficiency bounds. Therefore, their asymptotic variance is smaller than that of Hansen, Hausman, and Newey (2008) and Newey and Windmeijer (2009). All these methods involve a regularization parameter, which is the counterpart of the smoothing parameter that appears in the nonparametric literature. A data driven method was developed in Carrasco (2012) and Carrasco and Tchuente (2014) to select the best regularization parameter when the instruments are strong. We use these methods in our simulations for selecting the regularization parameter when the instruments are weak but we do not prove that these methods are valid in this case. A related paper is that of Hansen and Kozbur (2014) who propose a regularized jackknife instrumental variables estimator in a strong instruments setting where the design is not sparse.

Our Monte Carlo experiment shows that the leading regularized estimators (LF and T LIML) perform very well (are nearly median unbiased) even in the case of weak instruments.

The paper is organized as follows. Section 2 introduces the three regularization methods we consider and the associated estimators. Section 3 derives the asymptotic properties of the estimators. Section 4 discusses efficiency and related results. Section 5 presents Monte Carlo experiments. Section 6 considers an application to the effect of social infrastructure on per capita income. Section 7 concludes. The proofs are collected in Appendix.

2 Presentation of the regularized 2SLS and LIML estimators

This section presents the weak instruments setup and the estimators used in this paper. Estimators studied here are the regularized 2SLS and LIML estimators introduced in Carrasco (2012) and Carrasco and Tchuente (2014). They can be used with many or even a continuum of instruments. This work extends previous works by allowing for weak instruments as in Hansen, Hausman, and Newey (2008).

Our model is inspired by Hausman, Newey, Woutersen, Chao, and Swanson (2012). The model is

$$\begin{cases} y_i = W_i'\delta_0 + \varepsilon_i, \\ W_i = \gamma_i + u_i, \end{cases}$$

$i = 1, 2, ..., n$. The parameter of interest δ_0 is a finite dimensional $p \times 1$ vector.

$E(u_i|x_i) = E(\varepsilon_i|x_i) = 0$; $E(\varepsilon_i^2|x_i) = \sigma_\varepsilon^2$. y_i is a scalar and x_i is a vector of exogenous variables. Some rows of W_i may be exogenous, with the corresponding rows of u_i being zero. $\gamma_i = E(W_i|x_i)$ is a $p \times 1$ vector of reduced form values with $E(\gamma_i\varepsilon_i) = 0$. γ_i is the optimal instrument which is typically unknown. The estimation is based on a set of instruments $Z_i = Z(\tau; x_i)$, indexed by $\tau \in S$. The index τ may be an integer or may take its values in an interval. Examples of Z_i are the following.

- when x_i is a large $L \times 1$ vector, then one can select $Z_i = x_i$. In this case, $S = \{1, 2, ...L\}$ thus we have L instruments.
- assume that x_i is a scalar and $Z(\tau; x_i) = (x_i)^{\tau-1}$ with $\tau \in S = \mathbb{N}$, we obtain an infinite countable sequence of instruments.
- assume that x_i is a vector and $Z(\tau; x_i) = exp(i\tau' x_i)$ where $\tau \in S = \mathbb{R}^{dim(x_i)}$, we obtain a continuum of moment.

To simplify the presentation, we will present the estimators in the case where Z_i is a $L \times 1$ vector of instruments where L is some large integer. The theoretical results of Section 3 are proved for an arbitrary L which may be finite or infinite (case with a countable sequence or a continuum of instruments). In all cases, L does not depend on

n. The presentation of the estimators in the case with an infinite number of instruments is left in Appendix A.

This model allows for γ_i to be a linear or a non linear combination of Z_i. The model also allows for γ_i to approximate the reduced form. For example, we could let γ_i be a vector of unknown functions of x_i and Z_i could be power functions of x_i or interactions between elements of x_i. Adding extra instruments is a way to boost the concentration parameter as illustrated in the application in Section 6.

The estimate δ is based on the orthogonality condition.

$$E[(y_i - W_i'\delta)Z_i] = 0$$

where the vector of instruments Z_i has dimension L.

Let $W = \begin{pmatrix} W_1' \\ W_2' \\ . \\ . \\ W_n' \end{pmatrix}$ $n \times p$ and $u = \begin{pmatrix} u_1' \\ u_2' \\ . \\ . \\ u_n' \end{pmatrix}$ $n \times p$.

Let $\mathbf{Z}$ denote the $n \times L$ matrix having rows corresponding to Z_i'. Denote ψ_j the eigenvectors of the $n \times n$ matrix $\mathbf{Z}\mathbf{Z}'/n$ associated with eigenvalues λ_j. Recall that two-stage least squares (2SLS) and LIML estimators involve a projection matrix

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

The matrix $\mathbf{Z}'\mathbf{Z}$ may become nearly singular when L gets large. Moreover, $\mathbf{Z}'\mathbf{Z}$ is singular whenever $L \geq n$. To cover these cases, we will consider a regularized version of the inverse of the matrix $\mathbf{Z}'\mathbf{Z}$. For an arbitrary $n \times 1$ vector v, we define the $n \times n$ matrix P^α as

$$P^\alpha v = \frac{1}{n}\sum_{j=1}^{n} q\left(\alpha, \lambda_j^2\right)\left(v'\psi_j\right)\psi_j$$

where $q\left(\alpha, \lambda_j^2\right)$ is a weight that takes different forms depending on the regularization schemes. We consider three types of regularization:

- The Tikhonov (T) regularization: $q\left(\alpha, \lambda_j^2\right) = \dfrac{\lambda_j^2}{\lambda_j^2 + \alpha}$.

- The Landweber-Fridman (LF) regularization: $q(\alpha, \lambda_j^2) = [1-(1-c\lambda_j^2)^{1/\alpha}]$, where c is a constant such that $0 < c < 1/\left\|Z'Z/n\right\|^2$ and $\left\|Z'Z/n\right\|$ denotes the largest eigenvalue of $Z'Z/n$.
- The Spectral Cut-off (SC): $q(\alpha, \lambda_j^2) = I(\lambda_j^2 \geq \alpha)$.

Note that all these regularization techniques involve a tuning parameter α. The case $\alpha = 0$ corresponds to the case without regularization, $q\left(\alpha, \lambda_j^2\right) = 1$. Then, we obtain

$$P^0 = P = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'.$$

Consider regularized k-class estimators defined as follows:

$$\hat{\delta}_\nu = (W'(P^\alpha - \nu I_n)W)^{-1}W'(P^\alpha - \nu I_n)y.$$

where ν is either a constant term or a random variable. The case where $\nu = 0$ corresponds to regularized 2SLS estimator studied in Carrasco (2012):

$$\hat{\delta} = (W'P^\alpha W)^{-1}W'P^\alpha y$$

and the case $\nu = \nu_\alpha = \min_\delta \dfrac{(y-W\delta)'P^\alpha(y-W\delta)}{(y-W\delta)'(y-W\delta)}$ corresponds to the regularized LIML studied in Carrasco and Tchuente (2014). We denote $\hat{\delta}$ the regularized 2SLS estimator and $\hat{\delta}_L$ the regularized LIML estimators.

We study both 2SLS and LIML because LIML may have some advantages over 2SLS. For example when the number of instruments, L, increases with the sample size, n, so that $L/n \to c$ (with c constant), the standard 2SLS estimator is not consistent whereas standard LIML estimator is consistent.

3 Asymptotic properties

Carrasco (2012) and Carrasco and Tchuente (2014) focused on strong instruments. They found that regularized 2SLS and LIML estimators are asymptotically normal and attain the semiparametric efficiency bound. Here, we extend Carrasco (2012) and Carrasco and Tchuente (2014) results to the case of many weak instruments.

The weakness of the instruments is measured by the concentration parameter. For $p = 1$, the concentration parameter is equal to

$$CP = \frac{\sum_{i=1}^{n} \gamma_i^2}{E\left(u_i^2\right)}.$$

When the instruments are weak in the sense of Staiger and Stock (1997), CP converges to a constant and the parameter δ is not identified. This case is not considered here. We will maintain the assumption that CP diverges. It may diverge at the n rate (strong instruments) or at a slower rate (many weak IV asymptotics). By adding more instruments in the first stage equation:

$$W = Z\Pi + U,$$

the concentration parameter

$$CP = \frac{\Pi' Z' Z \Pi}{E\left(u_i^2\right)}$$

does not decrease and actually increases if these instruments contain non trivial information. Hence, adding more instruments is a way to boost the concentration parameter. Where do you get these new instruments? If you already have exogenous instruments, it is possible to interact them as it has been done for the estimation of the return to schooling (Angrist and Krueger (1991)) or take higher order powers of the same instruments as in Dagenais and Dagenais (1997). In the case of panel data, the use of lag variables is usually a source of many instruments. We provide an empirical application of the use of many weak instruments in Section 6.

Assumption 1:

(i) There exists a $p \times p$ matrix $S_n = \tilde{S}_n diag(\mu_{1n}, ..., \mu_{pn})$ such that $\tilde{S}_n$ is bounded, the smallest eigenvalue of $\tilde{S}_n \tilde{S}_n'$ is bounded away from zero; for each j, either $\mu_{jn} = \sqrt{n}$ (strong identification) or $\mu_{jn}/\sqrt{n} \to 0$ (weak identification), $\mu_n = \min_{1 \leq j \leq p} \mu_{jn} \to \infty$ and $\alpha \to 0$.

(ii) There exists a function $f_i = f(x_i)$ such that $\gamma_i = S_n f_i / \sqrt{n}$ and $\mu_n S_n^{-1} \to S_0$. $\sum_{i=1}^{n} \|f_i\|^4 / n^2 \to 0$, $\sum_{i=1}^{n} f_i f_i' / n$ is bounded and uniformly nonsingular.

These conditions allow for both many strong and many weak instruments. If $\mu_{jn} = \sqrt{n}$ this leads to asymptotic theory like in Kunitomo (1980), Morimune (1983), and Bekker (1994), but here we use regularization parameter instead of having an increasing sequence of instruments. For μ_n^2 growing slower than n, the convergence rate will be slower that $\sqrt{n}$, leading to an asymptotic approximation as that of Chao and Swanson (2007), Chao, Swanson, Hausman, Newey, and Woutersen (2012a), Chao, Swanson, Hausman, Newey, and Woutersen (2012b), and Chao, Swanson, Hausman, Newey, and Woutersen (2014). This is the case where we have many instruments without strong identification. Assumption 1 also allows for some components of the reduced form to give only weak identification (corresponding to $\mu_{jn}/\sqrt{n} \to 0$ which allows the concentration parameter to grow slower than $\sqrt{n}$), and other components (corresponding to $\mu_{jn} = \sqrt{n}$) to give strong identification for some coefficients of the reduced form. In particular, this condition allows for fixed constant coefficients in the reduced form. This specification of weak instruments can also be viewed as a generalization of Chao and Swanson (2007) but differs from that of Antoine and Lavergne (2012) who define the identification strength through the conditional moments that flatten as the sample size increases. To illustrate Assumption 1, let us consider the following example.

Example 1: Assume that $p = 2$, $\tilde{S}_n = \begin{pmatrix} 1 & 0 \\ \pi_{21} & 1 \end{pmatrix}$, and $\mu_{jn} = \begin{cases} \sqrt{n}, & j = 1 \\ \mu_n, & j = 2 \end{cases}$ with $\mu_n/\sqrt{n} \to 0$.

Then for $f(x_i) = (f_{1i}', f_{2i}')'$ the reduced form is

$$\gamma_i = \begin{pmatrix} f_{1i} \\ \pi_{21} f_{1i} + \dfrac{\mu_n}{\sqrt{n}} f_{2i} \end{pmatrix}.$$

We also have

$$\mu_n S_n^{-1} \to S_0 = \begin{pmatrix} 0 & 0 \\ -\pi_{21} & 1 \end{pmatrix}.$$

Assumption 2:

(i) The operator K is nuclear.

(ii) The ath row of γ, denoted γ_a, belongs to the closure of the linear span of $\{Z(.; x)\}$ for $a = 1, ..., p$.

(iii) $E(Z(.,x_i)f_{ia})$ belong to the range of K.

Condition (i) refers to the covariance operator K defined in Appendix A. K is nuclear provided its trace is finite, see for instance Carrasco and Florens (2014). This assumption is trivially satisfied if L is finite but may or may not be satisfied when L is infinite. This assumption implies in particular that the smallest eigenvalues decrease to zero sufficiently fast. For this to be true, the Z_i have to be correlated with each other. If $E\left(Z_i Z_i'\right) = I_L$ as in Assumption 5 of Newey and Windmeijer (2009), all the eigenvalues of the operator K equal 1 and hence K is not nuclear when L goes to infinity. To see whether Condition (i) is realistic, we examine the properties of the sample counterpart of K, namely $K_n = Z'Z/n$, in three applications: the return to schooling using 240 instruments from Angrist and Krueger (1991) (see also Carrasco and Tchuente (2014)), the elasticity of intertemporal substitution (see Carrasco and Tchuente (2014)), and the application on the effect of institutions on growth (see Section 6 of this paper). In the table below, we report the smallest eigenvalue, the largest eigenvalue, the condition number (which is the ratio of the largest eigenvalue on the smallest eigenvalue) and the trace of $Z'Z/n$ in two cases: raw data and standardized instruments. In the standardized case, the instruments are divided with their standard deviation. This standardization has no impact on 2SLS and LIML estimators which are scale invariant. However, our estimators are not scale invariant and standardization may improve the results. Such standardizations are customary whenever regularizations are used, see for instance De Mol, Giannone, and Reichlin (2008) and Stock and Watson (2012). We observe that in all applications, the smallest eigenvalue is close to zero so the instruments are strongly correlated[1]. The condition number - which is scale invariant - is an indicator on how ill-posed the matrix K_n is. The higher the condition number, the more imprecise the inverse of K_n will be. The smallest possible condition number is 1 (which corresponds to the identity matrix). Here, the condition numbers are all very large which suggests that regularization will be helpful to improve the reliability of the estimate of K^{-1}. The trace of K_n appears to be finite throughout the applications.

Condition (ii) guarantees that the optimal instrument f can be approached by a

[1] A word of caution: when the number of instruments is large enough relative to the sample size, the sample covariance matrix $Z'Z/n$ will be near singular or singular which does not mean that the smallest eigenvalue of K is not bounded away from 0 in the population. Moreover, eigenvalues are not scale invariant.

Table 1: Properties of $Z'Z/n$

	Largest eigenvalue	Smallest eigenvalue	Condition number	Trace
Angrist and Krueger	1.35	0.0000107	126168.22	5.05
Angrist and Krueger standardized	5.93	0.0012	4941.66	244.47
EIS	1550	1.41×10^{-13}	1.09929×10^{16}	1550
EIS standardized	11.8	2.35×10^{-5}	5.06×10^{5}	11.89
Institutions	474×10^{7}	9.47×10^{-6}	5.00528×10^{14}	4.78×10^{9}
Institutions standardized	28.9	0.000116	249137.93	43.58

sequence of instruments. It is similar to Assumption 4 in Hansen, Hausman, and Newey (2008). Condition (iii) is a technical assumption which can also be found in Carrasco (2012). Assumptions 2(ii) and (iii) are needed only for efficiency.

Proposition 1. *(Asymptotic properties of regularized 2SLS with many weak instruments)*

Assume $\{y_i; W_i; x_i\}$ *are iid,* $E(\varepsilon_i^2|X) = \sigma_\varepsilon^2$, α *goes to zero as* n *goes to infinity. Moreover, Assumptions 1 and 2 are satisfied. Then, the T, LF, and SC estimators of 2SLS satisfy:*

1. *Consistency:* $S_n'(\hat{\delta} - \delta_0)/\mu_n \to 0$ *in probability as* n, $n\alpha^{\frac{1}{2}}$ *and* μ_n *go to infinity.*
2. *Asymptotic normality:*

   $S_n'(\hat{\delta} - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 \left[E(f_i f_i')\right]^{-1}\right)$

   as n, $n\alpha$ *and* μ_n *go to infinity, where* $E(f_i f_i')$ *is a nonsingular* $p \times p$ *matrix.*

Proof In Appendix.

The first point of Proposition 1 implies the consistency of the estimator, namely $(\hat{\delta} - \delta_0) \to 0$ (see the proof of Theorem 1 in Hansen and Kozbur (2014)). Moreover, Proposition 1 shows that the three estimators have the same asymptotic distribution. Instead of restricting the number of instruments (which may be very large or infinite), we impose restrictions on the regularization parameter which goes to zero. This insures us that all available and valid instruments are used in an efficient way even if they are weak. To obtain consistency, the condition on α is $n\alpha^{\frac{1}{2}}$ go infinity, whereas for the asymptotic normality, we need $n\alpha$ go to infinity. This means that α is allowed to go

to zero at a slower rate. However, this rate does not depend on the weakness of the instruments.

Interestingly, our regularized 2SLS estimators reach the semiparametric efficiency bound. This result will be further discussed in Section 4.

We are now deriving the asymptotic properties of the regularized LIML with many weak instruments.

Proposition 2. *(Asymptotic properties of regularized LIML with many weak instruments)*

Assume $\{y_i; W_i; x_i\}$ *are iid,* $E(\varepsilon_i^2|X) = \sigma_\varepsilon^2$, $E\left(\varepsilon_i^4|X\right) < \infty$, $E\left(u_{bi}^4|X\right) < \infty$, α *goes to zero as n goes to infinity. Moreover, Assumptions 1 and 2 are satisfied. Then, the T, LF, and SC estimators of LIML with weak instruments satisfy:*

1. *Consistency:* $S_n'(\hat{\delta}_L - \delta_0)/\mu_n \to 0$ *in probability as n,* μ_n *and* $\mu_n^2\alpha$ *go to infinity.*
2. *Asymptotic normality:*

   $S_n'(\hat{\delta}_L - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 \left[E(f_i f_i')\right]^{-1}\right)$

   as n, μ_n *and* $\mu_n^2\alpha$ *go to infinity where* $E(f_i f_i')$ *is a nonsingular* $p \times p$ *matrix.*

Proof In Appendix.

Again, Proposition 1 implies the consistency of the estimator, namely $(\hat{\delta} - \delta_0) \to 0$. Interestingly we obtain the same asymptotic distribution as in the many strong instruments case (with a slower rate of convergence). We also find the same speed of convergence as in Hansen, Hausman, and Newey (2008) and Newey and Windmeijer (2009). For the consistency and asymptotic normality, $\mu_n^2\alpha$ needs to go to infinity, which means that the regularization parameter should go to zero at a slower rate than the concentration parameter. The asymptotic variance of regularized LIML corresponds to the lower bound and is smaller than that obtained in Hansen, Hausman, and Newey (2008). We believe that the reason, why Hansen, Hausman, and Newey (2008) obtain a larger asymptotic variance than us, is that they use the number of instruments as regularization parameter. As a result, they can not let L grow fast enough to reach efficiency. Our estimator involves the extra tuning parameter α which is selected so that extra terms in the variance vanish asymptotically. Moreover, we

assume that the set of instruments is sufficiently rich to span the optimal instrument (Assumption 2(ii)).

Example 1:(cont)
$S_n'(\hat{\delta} - \delta_0) = \begin{pmatrix} \sqrt{n}[(\hat{\delta}_1 - \delta_{01}) + \pi_{21}(\hat{\delta}_2 - \delta_{02})] \\ \mu_n(\hat{\delta}_2 - \delta_{02}) \end{pmatrix}$ is jointly asymptotically normal. The linear combination $(\hat{\delta}_1 - \delta_{01}) + \pi_{21}(\hat{\delta}_2 - \delta_{02})$ converges at rate $\sqrt{n}$. This is the coefficient of f_{i1} in the reduced form equation for y_i. And the estimator of the coefficient δ_2 of W_{i2} converges at rate $\dfrac{1}{\mu_n}$.

Now, as in Newey and Windmeijer (2009), we consider a t-ratio for a linear combination $c'\delta$ of the parameter of interest. The following proposition is a corollary of Proposition 2.

Proposition 3. *Under the assumptions of Proposition 2 and if we assume that there exist r_n, c and $c^* \neq 0$ such that $r_n S_n^{-1} c \to c^*$ and $S_n' \hat{\Phi} S_n / n \to \Phi$ in probability with $\Phi = \sigma_\varepsilon^2 \left[E(f_i f_i')\right]^{-1}$.*
Then,

$$\frac{c'(\hat{\delta}_L - \delta_0)}{\sqrt{c'\hat{\Phi}c}} \xrightarrow{d} \mathcal{N}(0,1)$$

as n and $\mu_n^2 \alpha$ go to infinity.

This result allows us to form confidence intervals and test statistics for a single linear combination of parameters in the usual way.

4 Efficiency and Related Literature

4.1 Efficiency

If the optimal instrument γ_i were known, the estimator would be solution of

$$\frac{1}{n}\sum_{i=1}^{n} \gamma_i \left(y_i - W_i'\delta\right) = 0.$$

Hence,

$$\begin{aligned}
\hat{\delta} &= \left(\sum_{i=1}^{n}\gamma_i W_i'\right)^{-1}\sum_{i=1}^{n}\gamma_i y_i, \\
\hat{\delta}-\delta_0 &= \left(\sum_{i=1}^{n}\gamma_i W_i'\right)^{-1}\sum_{i=1}^{n}\gamma_i \varepsilon_i \\
&= \left(S_n\frac{\sum_{i=1}^{n} f_i f_i'}{n}S_n' + S_n\frac{\sum_{i=1}^{n} f_i u_i}{\sqrt{n}}\right)^{-1} S_n\frac{\sum_{i=1}^{n} f_i\varepsilon_i}{\sqrt{n}},
\end{aligned}$$

$$\begin{aligned}
S_n\left(\hat{\delta}-\delta_0\right) &= \left(\frac{\sum_{i=1}^{n} f_i f_i'}{n} + \frac{\sum_{i=1}^{n} f_i u_i}{\sqrt{n}}S_n'^{-1}\right)^{-1}\frac{\sum_{i=1}^{n} f_i\varepsilon_i}{\sqrt{n}} \\
&\xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2\left[E\left(f_i f_i'\right)\right]^{-1}\right).
\end{aligned}$$

So the lowest asymptotic variance that can be obtained is $\sigma_\varepsilon^2\left[E\left(f_i f_i'\right)\right]^{-1}$. We will refer to this as the semiparametric efficiency bound[2].

In Carrasco (2012, Section 2.4), it was shown that the regularized 2SLS estimator coincides with a 2SLS estimator that uses a specific nonparametric estimator, $\hat{\gamma}_i$, of γ_i :

$$\hat{\delta} = \left(\sum_{i=1}^{n}\hat{\gamma}_i' W_i\right)^{-1}\sum_{i=1}^{n}\hat{\gamma}_i' y_i.$$

This may explain why for the regularized 2SLS estimator, the conditions on α are not related to μ_n whereas, in the case of LIML, the rate of convergence of α depends on how weak the instruments are.

4.2 Related Literature

In the literature on many weak instruments, the asymptotic behavior of estimators depends on the relation between the number of moment conditions L and sample size n. For the CUE, L and n need to satisfy $L^2/n \to 0$ for consistency and $L^3/n \to 0$ for asymptotic normality. Under homoskedasticity, Stock and Yogo (2005) require $L^2/n \to 0$. Hansen, Hausman, and Newey (2008) allowed L to grow at the same rate

[2]We do not provide a formal proof that this bound is the semiparametric bound. This proof is beyond the scope of the present paper. We refer the interested readers to Newey (1990), Newey (1993), and Chamberlain (1992).

as n, but restricted L to grow slower than the square of the concentration parameter, for the consistency of LIML and FULL. Andrews and Stock (2006) require $L^3/n \to 0$ when normality is not imposed.

Just as 2SLS is not consistent if L is too large relative to n, LIML estimator is not feasible if $L > N$ because the matrix $Z'Z$ is not invertible. Therefore, some form of regularization needs to be implemented to obtain consistent estimators when the number of instruments is really large. The introduction of a regularization parameter, α, permits to free L from any constraint (conditions are on α and not on L). Consequently, the regularized estimator can reach the semiparametric efficiency bound. There is another major difference between our work and the many instruments literature. In the many instruments literature (see e.g. Newey and Windmeijer, 2009), the smallest and largest eigenvalues of the $Z'Z/n$ are bounded away from 0 and from above, respectively. In our case, we suppose that the eigenvalues of the covariance operator K (which corresponds to the limit of $Z'Z/n$ or a rescaled version of it) are summable, which means that the smallest eigenvalues must decrease to zero sufficiently fast.

Caner and Yildiz (2012) in a recent work consider a Continuous Updating Estimator (CUE) with many weak moments under nearly singular design. They show that the nearly singular design affects the form of asymptotic covariance matrix of the estimator compared to that of Newey and Windmeijer (2009). Our work is also related to Hausman, Lewis, Menzel, and Newey (2011) who modify the continuous updating estimator (CUE) by introducing two tuning parameters which perform a Tykhonov-type regularization. They show that their estimator has finite moments when the regularization parameters are positive. On the other hand, their estimator is shown to be asymptotically equivalent to the conventional CUE under many weak asymptotics when the regularization parameters go to zero. There are two main differences with our approach. First, they introduce two tuning parameters instead of one. Second, they restrict the number of moments as in Newey and Windmeijer (2009), whereas we allow for the number of instruments to exceed the sample size.

Belloni, Chen, Chernozhukov, and Hansen (2012) propose to use an alternative regularization named lasso in the IV context. This regularization imposes a l_1 type penalty on the first stage coefficient. Assuming that the first stage equation is approxi-

mately sparse, they show that the postlasso estimator reaches the asymptotic efficiency bound.

Chao, Swanson, Hausman, Newey, and Woutersen (2012c) explain why Fuller estimator has moments and argue that "the Fuller modification amounts to a ridge-regression-type perturbation of the denominator of the 2SLS". The existence of moments of regularized LIML estimator is shown in Carrasco and Tchuente (2014). Note that, in the regularized LIML/2SLS, the regularization is applied to the covariance matrix, whereas, in Fuller estimator, a penalty term is added to the denominator.

Table 2 gives an overview of the assumptions used in the main papers on many weak instruments.

Table 2: Comparison of different IV asymptotics

	Number of instruments	Extra assumptions
Conventional	Fixed L	
Phillips (1989)	Fixed L, $Cov(W,x)=0$	
Staiger and Stock (1997)	Fixed L, $Cov(W,x)=O(n^{-1/2})$	
Bekker (1994)	$L/n \to c<1$, $\mu_n^2=O(n)$	
Han and Phillips (2006)	$L\to\infty$ and $\frac{L}{nc_n}\to c$ c_n μ_n constant or zero	
Chao and Swanson (2005)	$\frac{L}{\mu_n^2}\to 0$ or $\frac{L^{1/2}}{\mu_n^2}\to 0$	
Hansen et al. (2008)	(I) $\frac{L}{\mu_n^2}$ bounded or (II) $\frac{L}{\mu_n^2}\to\infty$	$\sum z_i z_i'/n$ nonsingular
Newey and Windmeijer (2009)	$L\to\infty$, $\frac{L}{\mu_n^2}$ bounded, $\frac{L^3}{n}\to 0$	
Carrasco (2012)	No condition on L, possibly continuum strong instruments	Compactness of covariance matrix
Belloni et al. (2012)	$log(L)=o(n^{1/3})$, strong instruments	Approximately sparse first stage equation

5 Monte Carlo study

We now carry out a Monte Carlo simulation for the simple linear IV model where the disturbances and instruments have a Gaussian distribution and the instruments are independent from each other as in Newey and Windmeijer (2009). The design of this experiment involves the correlation coefficient ρ between the structural and reduced

form errors, the concentration parameter (CP), and the number of instruments L.

The data generating process is given by:

$$y_i = W_i'\delta_0 + \varepsilon_i,$$

$$W_i = x_i'\pi + u_i,$$

$$\varepsilon_i = \rho u_i + \sqrt{1-\rho^2}v_i,$$

$$u_i \sim \mathcal{N}(0,1),\ v_i \sim \mathcal{N}(0,1),\ x_i \sim \mathcal{N}(0, I_L)$$

$$\pi = \sqrt{\frac{CP}{Ln}}\iota_L$$

where ι_L is an L-vector of ones. The sample size is $n = 500$. The instruments are $Z_i = x_i$ and the number of instruments L equals 30 and 50. Note that this setting is not favorable for us because the eigenvalues of the matrix $Z'Z/n$ are all equal to 1. If L were infinite, the matrix $Z'Z/n$ would become an infinite dimensional identity matrix which is not nuclear and hence Assumption 2 would not hold. However, here L being no larger than 50, K is nuclear.

In the simulations, $\rho = 0.5$ and $\delta_0 = 0.1$. The values of CP equal 8, 35, and 65.

The estimators we proposed in this paper depend on a regularization (smoothing) parameter α that needs to be selected. In the simulations, we use a data-driven method[3] for selecting α based on an expansion of the MSE and proposed in Carrasco (2012) and Carrasco and Tchuente (2014). These selection criteria were derived assuming strong instruments and may not be valid in the presence of weak instruments. Providing a robust to weak instruments selection procedure is beyond the scope of this paper.

We report the median bias (Med.bias), the median of the absolute deviation of the estimator from the true value (Med.abs), the difference between the 0.1 and 0.9 quantiles (dis) of the distribution of each estimator, and the coverage rate (Cov.) of a nominal 95% confidence interval for unfeasible instrumental variable estimator (IV) using the true optimal instrument, regularized two-stage least squares (T2SLS (Tikhonov),

[3]The optimal α for Tikhonov is searched over the interval [0.01,0.5] with 0.01 increment. The range of values for the number of iterations for LF is from 1 to 300, and the number of principal components ranges from 1 to the number of instruments.

L2SLS (Landweber Fridman), P2SLS (Principal component)), LIML and regularized LIML (TLIML (Tikhonov), LLIML (Landweber Fridman), PLIML (Principal component)) and Donald and Newey's (2001) 2SLS and LIML (D2SLS and DLIML). For all the regularized LIML, DLIML, and standard LIML estimators, the starting values for the minimization needed in the estimation of ν (see Section 2) are the 2SLS using all the instruments. For confidence intervals, we compute the coverage probabilities using the following estimate of asymptotic variance as in Donald and Newey (2001) and Carrasco (2012):

$$\hat{V}(\hat{\delta}) = \frac{(y - W\hat{\delta})'(y - W\hat{\delta})}{n} \left(\hat{W}'W^{-1}\right)^{-1} \hat{W}'\hat{W} \left(W'\hat{W}\right)^{-1}$$

where $\hat{W} = P^{\alpha}W$ for 2SLS and $\hat{W} = (P^{\alpha} - \nu I_n)\, W$ for LIML. Note that the formulaes for the confidence intervals is the same as for strong instruments (see Carrasco and Tchuente (2014)).

Table 3 reports simulation results. We use different strength (measured by the concentration parameter) of instruments and number of instruments. We investigate the case of very weak instruments for example, when $CP = 8$ and $L = 50$, the first stage F-statistic equals $\dfrac{CP}{L} + 1 = 1.16$.

We observe that

(a) The performances of the regularized estimators increase with the strength of instruments but decrease with the number of instruments. Providing regularization parameter selection procedure robust to weak instruments would certainly improve these results.

(b) The bias of regularized LIML is quite a bit smaller than that of regularized 2SLS.

(c) The bias of our regularized estimators are smaller than those of the corresponding Donald and Newey's estimators. On the other hand, DN estimator has often better coverage.

(d) LF LIML and T-LIML estimators have very low median bias even in the case of relatively weak instruments ($CP = 8$).

(e) The coverage of our estimators deteriorates when the instruments are weak.

Table 3: Simulations results for regularized 2SLS and LIML with $CP = 8$, 35 and 65; $L = 30$ and 50; $n = 500$; 1000 replications.

		T2SLS	**L2SLS**	**P2SLS**	**D2SLS**	**TLIML**	**LLIML**	**PCLIML**	**DLIML**	**IV**
L=30										
CP=8	Med.bias	0.4012	0.3875	0.3271	0.3802	0.0889	0.0800	0.3228	0.3628	-0.0096
	Med.abs	0.4012	0.3875	0.4947	0.5419	0.4472	0.4568	0.4511	0.4513	0.2624
	Disp	0.3884	0.4271	2.5366	2.6333	2.4053	2.2558	1.4859	1.5308	1.0184
	Cov	0.2260	0.3270	0.7570	0.7180	0.9180	0.9210	0.8030	0.7880	0.9540
CP=35	Med.bias	0.2195	0.2070	0.2245	0.2476	-0.0172	-0.0097	0.0858	0.0997	-0.0157
	Med.abs	0.2195	0.2070	0.2623	0.2914	0.1465	0.1488	0.1652	0.1680	0.1100
	Disp	0.2814	0.3115	0.6175	0.8039	0.6028	0.6232	0.6382	0.6008	0.4424
	Cov	0.5050	0.5900	0.7100	0.682	0.954	0.9500	0.8590	0.8700	0.9600
CP=65	Med.bias	0.1567	0.1456	0.1548	0.1992	-0.0083	-0.0046	0.0194	0.0184	-0.0072
	Med.abs	0.1568	0.1476	0.1826	0.2154	0.0999	0.0919	0.0972	0.1012	0.0813
	Disp	0.2503	0.2703	0.4139	0.4582	0.381	0.3849	0.3825	0.3747	0.3034
	Cov	0.6270	0.7070	0.7560	0.6960	0.960	0.9630	0.9130	0.913	0.9630
L=50										
CP=8	Med.bias	0.4273	0.4189	0.3969	0.4166	0.1178	0.1327	0.3860	0.4089	-0.0042
	Med.abs	0.4273	0.4189	0.5418	0.6089	0.5225	0.5163	0.4730	0.5253	0.2548
	Disp	0.3112	0.3571	2.4463	3.1653	3.1549	3.4005	1.5870	1.7621	1.0971
	Cov	0.0710	0.1480	0.7450	0.7730	0.9190	0.9200	0.7960	0.7930	0.9510
CP=35	Med.bias	0.2906	0.2740	0.2654	0.2875	-0.0188	0.0072	0.1475	0.1998	0.0016
	Med.abs	0.2906	0.2740	0.2961	0.3312	0.1766	0.1810	0.2103	0.2499	0.1178
	Disp	0.2527	0.2811	0.8402	1.2808	0.7100	0.7584	0.7554	0.8201	0.4587
	Cov	0.1860	0.3110	0.6730	0.6640	0.9420	0.9500	0.8420	0.7830	0.9580
CP=65	Med.bias	0.2137	0.1953	0.2020	0.2402	0.0082	0.0078	0.0695	0.0739	0.0041
	Med.abs	0.2137	0.1953	0.2214	0.2614	0.1127	0.1128	0.1212	0.1245	0.0812
	Disp	0.2075	0.2474	0.4840	0.5215	0.4333	0.4352	0.4261	0.4587	0.3231
	Cov	0.3010	0.4710	0.6840	0.6500	0.9610	0.9580	0.8520	0.8660	0.9530

6 Empirical application: Institutions and Growth

This section revisits Hall and Jones (1999) empirical work. Hall and Jones (1999) argue that the difference between output per worker across countries is mainly due to the differences in institution and government policies - the so-called social infrastructure. They write "Countries with corrupt government officials, severe impediments to trade, poor contract enforcement, and government interference in production will be unable to achieve levels of output per worker anywhere near the norms of western Europe, northern America, and eastern Asia." To quantify the effect of social infrastructure on per capita income, they use two-stage least squares (2SLS) with four instruments: the fraction of population speaking English at birth (EnL), the fraction of population speaking one of the five major European languages at birth (EuL), the distance from the equator[4] (latitude , Lt) and Romer and Frankel (1999) geography-predicted trade intensity (FR). The linear IV regression model is given by:

$$y = c + \delta S + \varepsilon,$$

where y is an $n \times 1$ vector of log income per capita, S is an $n \times 1$ vector which is the proxy for social infrastructure, θ is an $L \times 1$ vector, c and δ are scalars. Dmitriev (2013) points out the fact that the instruments[5] $X = [EnL, EuL, Lt, FR]$ are weak. To address this issue, we increased the number of instruments from 4 to 18. The 18 instruments in our regression are derived from X and are given by[6]
$Z = [X, X.^2, X.^3, X(:,1) * X(:,2), X(:,1) * X(:,3), X(:,1) * X(:,4), X(:,2) * X(:,3), X(:,2) * X(:,4), X(:,3) * X(:,4)]$ where all instruments are divided by their standard deviation.

The use of many instruments increased the concentration parameter from $\hat{\mu}_n^2 = 28.6$ to $\hat{\mu}_n^2 = 51.48$. However, it also increased the condition number of the $Z'Z$ matrix from $1.08e+04$ for 4 instruments to $2.48e+05$ for 18. As the regularized 2SLS and LIML

[4] The distance from the equator is measured as the absolute value of latitude in degrees divided by 90 to place it on a 0 to 1 scale.

[5] This corresponds to the specification (iv) of Dmitriev (2013).

[6] $X.^k = [X_{ij}^k]$, $X(:,k)$ is the k^{th} column of X and $X(:,k) * X(:,l)$ is a vector of interactions between columns k and l.

correct the bias due to the use of many instruments, they should provide better point estimates.

We use a sample of 79 countries for which no data were imputed[7]. The results are reported in Table 4 below. Robust to heteroskedasticity standard errors are given in parentheses. They are computed using the formula of Carrasco and Tchuente (2014):

$$\hat{V}(\hat{\delta}) = \left(\hat{W}'W\right)^{-1} \hat{W}'\widehat{\Omega}\hat{W} \left(W'\hat{W}\right)^{-1}$$

where $\hat{W} = P^{\alpha}W$ for 2SLS, $\hat{W} = (P^{\alpha} - \nu I_n)\, W$ for LIML, and $\widehat{\Omega}$ is the diagonal matrix with ith diagonal element equal to $\hat{\varepsilon}_i^2 = \left(y_i - W_i'\hat{\delta}\right)^2$.

Table 4: Institutions and growth

2SLS (4)	**2SLS (18)**	**T2SLS**	**L2SLS**	**P2SLS**
4.6612 (0.7027)	4.0124 (0.5041)	4.2916 (0.338)	4.27 (0.431)	4.03 (0.327)
		α=0.01	Number of iterations 1000	Number of eigenvalues 15
LIML (4)	**LIML (18)**	**TLIML**	**LLIML**	**PLIML**
5.2683 (0.7602)	5.7090 (0.899)	5.3062 (0.631)	4.73 (0.687)	5.57 (0.846)
		α=0.01	Number of iterations 1000	Number of eigenvalues 15
$\hat{\mu}_n^2$=28.6	$\hat{\mu}_n^2$=51.48			

NB: We report 2SLS and LIML for 4 and 18 instruments. For LIML with 18 instruments, we report the many instrument robust standard error of Hansen, Hausman, and Newey (2008) in parentheses. The regularized estimators are computed for 18 instruments. For the regularized estimators, the heteroskedasticity robust standard errors are given in parentheses.

Our findings suggest that "social infrastructure" has a significant causal effect on long-run economic performance throughout the world. The use of many instruments first increase the bias as illustrated by the fact that the distance between 2SLS and LIML is larger when 18 instruments are used. When the regularization is introduced, this gap shrinks. For instance, for LF regularization, LIML and 2SLS are very close, this may be due to bias correction. But, for PC, the gap remains wide. The reason may be due to the lack of factor structure in the instruments.

[7]The data were downloaded from Charles Jones' webpage: http://www.stanford.edu/~chadj/HallJones400.asc

7 Conclusion

This paper illustrates the usefulness of regularization techniques for estimation in the many weak instruments framework. We derived the properties of the regularized 2SLS and LIML estimators in the presence of many or a continuum of moments that may be weak. We show that if well normalized the regularized 2SLS and LIML are consistent and reach the semiparametric efficiency bound. Our simulations show that the leading regularized estimators (LF and T of LIML) perform well.

In this work, we restricted our investigation to 2SLS and LIML with weak instruments. It would be interesting, for future research, to study the behavior of regularized version of other k-class estimators, such as FULL (Fuller (1977)) and bias adjusted 2SLS or other estimators as generalized method of moments or generalized empirical likelihood, in presence of many weak instruments. This will help us to have results that can be compared with those of Newey and Windmeijer (2009) and Hansen, Hausman, and Newey (2008). Another topic of interest is the use of our regularization tools to provide version of robust tests for weak instruments as Anderson-Rubin tests, that can be used with a large number or a continuum of moment conditions.

References

Andrews, D. W., and J. H. Stock (2006): "Inference with Weak Instruments," in *Advances in Economics and Econometrics*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 3. Cambridge University Press.

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

Antoine, B., and P. Lavergne (2012): "Conditional Moment Models under Semi-Strong Identification," Discussion Papers dp11-04, Department of Economics, Simon Fraser University.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657–81.

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.

Caner, M., and N. Yildiz (2012): "CUE with many weak instruments and nearly singular design," *Journal of Econometrics*, 170(2), 422–441.

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

Carrasco, M., and J.-P. Florens (2000): "Generalization Of Gmm To A Continuum Of Moment Conditions," *Econometric Theory*, 16(06), 797–834.

——— (2014): "On the Asymptotic Efficiency of GMM," *Econometric Theory*, 30(2), 372–406.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 77. Elsevier.

Carrasco, M., and G. Tchuente (2014): "Regularized LIML for many instruments," Discussion paper, forthcoming in Journal of Econometrics.

Chamberlain, G. (1992): "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60(3), 567–596.

Chao, J. C., and N. R. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73(5), 1673–1692.

——— (2007): "Alternative approximations of the bias and MSE of the IV estimator under weak identification with an application to bias correction," *Journal of Econometrics*, 137(2), 515–555.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012a): "Asymptotic Distribution of JIVE in a Heteroskedastic Regression with Many Instruments," *Econometric Theory*, 28, 42–86.

——— (2012b): "Combining Two Consistent Estimators," in *Advances in Econometrics: Essays in Honor of Jerry Hausman*, ed. by W. N. Badi Baltagi, Carter Hill, and H. White. Emerald Group Publishing.

——— (2012c): "An Expository Note on the Existence of Moments of Fuller and HFUL Estimators," in *Advances in Econometrics: Essays in Honor of Jerry Hausman*, ed. by W. N. Badi Baltagi, Carter Hill, and H. White. Emerald Group Publishing.

——— (2014): "Testing Overidentifying Restrictions with Many Instruments and Heteroskedasticity," *Journal of Econometrics*, 178, 15–21.

Dagenais, M. G., and D. L. Dagenais (1997): "Higher moment estimators for linear regression models with errors in the variables," *Journal of Econometrics*, 76(1-2), 193–221.

De Mol, C., D. Giannone, and L. Reichlin (2008): "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?," *Journal of Econometrics*, 146(2), 318–328.

Dmitriev, A. (2013): "Institutions and growth: evidence from estimation methods robust to weak instruments," *Applied Economics*, 45(13), 1625–1635.

Eichenbaum, M. S., L. P. Hansen, and K. J. Singleton (1988): "A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice under Uncertainty," *The Quarterly Journal of Economics*, 103(1), 51–78.

Fuller, W. A. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45(4), 939–953.

Hall, R. E., and C. I. Jones (1999): "Why Do Some Countries Produce So Much More Output Per Worker Than Others?," *The Quarterly Journal of Economics*, 114(1), 83–116.

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

Hansen, C., and D. Kozbur (2014): "Instrumental Variables Estimation with Many Weak Instruments Using Regularized JIVE," *working paper*.

Hausman, J., R. Lewis, K. Menzel, and W. Newey (2011): "Properties of the CUE estimator and a modification with moments," *Journal of Econometrics*, 165(1), 45 – 57.

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.

Kunitomo, N. (1980): "Asymptotic Expansions of the Distributions of Estimators in a Linear Functional Relationship and Simultaneous Equations," *Journal of the American Statistical Association*, 75(371), 693–700.

Morimune, K. (1983): "Approximate Distributions of k-Class Estimators When the Degree of Overidentifiability Is Large Compared with the Sample Size," *Econometrica*, 51(3), 821–41.

Newey, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58(4), 809–837.

——— (1993): "Efficient Estimation of Models with Conditional Moment Restrictions," in *Handbook of Statistics*, ed. by G. Maddala, C. Rao, and H. Vinod, vol. 11, pp. 419–454. Elsevier.

Newey, W. K., and F. Windmeijer (2009): "Generalized Method of Moments With Many Weak Moment Conditions," *Econometrica*, 77(3), 687–719.

Romer, D. H., and J. A. Frankel (1999): "Does Trade Cause Growth?," *American Economic Review*, 89(3), 379–399.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

Stock, J., and M. Watson (2012): "Generalised Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business and Economic Statistics*, 30(4), 481–493.

Stock, J. H., and M. W. Watson (2002): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97(460), pp. 1167–1179.

Stock, J. H., and M. Yogo (2005): "Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments," in *Identification and Inference for Econometric Models*, ed. by D. W. K. Andrews, and J. H. Stock, pp. 109–120. Cambridge University Press.

A General notation

Here we consider the general case where the estimation is based on a sequence of instruments $Z_i = Z(\tau; x_i), \tau \in S$. Let π be a positive measure on S. We denote $L^2(\pi)$ the Hilbert space of square integrable functions with respect to π.

We define the covariance operator K of the instruments as

$$K : L^2(\pi) \to L^2(\pi)$$

$$(Kg)(\tau_1) = \int E(Z(\tau_1; x_i)\overline{Z(\tau_2; x_i)})g(\tau_2)\pi(\tau_2)d\tau_2$$

where $\overline{Z(\tau_2; x_i)}$ denotes the complex conjugate of $Z(\tau_2; x_i)$.

K is assumed to be a nuclear operator. Let λ_j and ϕ_j, $j = 1, 2, ...$ be respectively, the eigenvalues (ranked in decreasing order) and orthonormal eigenfunctions of K. K can be estimated by K_n defined as:

$$K_n : L^2(\pi) \to L^2(\pi)$$

$$(K_n g)(\tau_1) = \int \frac{1}{n}\sum_{i=1}^{n} Z(\tau_1; x_i)\overline{Z(\tau_2; x_i)}g(\tau_2)\pi(\tau_2)d\tau_2.$$

If the number of moment conditions is infinite then the inverse of K_n needs to be regularized because it is not continuous. By definition (see Kress, 1999, page 269), a regularized inverse of an operator K is

$$R_\alpha : L^2(\pi) \to L^2(\pi)$$

such that $\lim_{\alpha \to 0} R_\alpha K\varphi = \varphi, \forall \varphi \in L^2(\pi)$.

Three different types of regularization schemes are considered: Tikhonov (T), Landwerber Fridman (LF), Spectral cut-off (SC) or Principal Components (PC). They are defined as follows:

1. **Tikhonov(T)**

   This regularization scheme is related to the ridge regression.

$$(K^\alpha)^{-1} = (K^2 + \alpha I)^{-1} K$$

$$(K^\alpha)^{-1} r = \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j^2 + \alpha} \langle r, \phi_j \rangle \phi_j$$

   where $\alpha > 0$ is the regularization parameter. A fixed α would result in a loss of efficiency. For the estimator to be asymptotically efficient, α has to go to zero at a certain rate which will be determined later on. This regularization is closely related to ridge regularization. Ridge regularization was first used in regression in a context where there were too many regressors. The aim was then to stabilize the inverse of $X'X$ by replacing $X'X$ by $X'X + \alpha I$. However, this was done at the expense of a bias relative to OLS estimator. In the IV regression, the 2SLS estimator has already a bias and the use of many instruments usually increases its bias. The selection of an appropriate ridge parameter for the first step regression helps to reduce this bias. This explains why, in the IV case, ridge regularization is useful.

2. **Landweber Fridman (LF)**

   This method of regularization is iterative. Let $0 < c < 1/\|K\|^2$ where $\|K\|$ is the largest eigenvalue of K (which can be estimated by the largest eigenvalue of K_n). $\hat{\varphi} = (K^\alpha)^{-1} r$ is computed using the following algorithm:

$$\begin{cases} \hat{\varphi}_l = (1 - cK^2)\hat{\varphi}_{l-1} + cKr, & \text{l=1,2,...,} \dfrac{1}{\alpha} - 1; \\ \hat{\varphi}_0 = cKr, \end{cases}$$

   where $\dfrac{1}{\alpha} - 1$ is some positive integer. We also have

$$(K^\alpha)^{-1} r = \sum_{j=1}^{\infty} \frac{[1 - (1 - c\lambda_j^2)^{\frac{1}{\alpha}}]}{\lambda_j} \langle r, \phi_j \rangle \phi_j.$$

3. **Spectral cut-off (SC)**

   This method consists in selecting the eigenfunctions associated with the eigenvalues greater than some threshold. The aim is to select those who have greater contribution.

$$(K^\alpha)^{-1}r = \sum_{\lambda_j^2 \geq \alpha} \frac{1}{\lambda_j}\langle r, \phi_j\rangle \phi_j$$

   for $\alpha > 0$.

   This method is similar to principal components (PC) which consists in using the first eigenfunctions:

$$(K^\alpha)^{-1}r = \sum_{j=1}^{1/\alpha} \frac{1}{\lambda_j}\langle r, \phi_j\rangle \phi_j$$

   where $\dfrac{1}{\alpha}$ is some positive integer. It is equivalent to projecting on the first principal components of W. Interestingly, this approach is used in factor models where W_i is assumed to depend on a finite number of factors (see Bai and Ng (2002), Stock and Watson (2002)) As the estimators based on PC and SC are identical, we will use PC and SC interchangeably.

These regularized inverses can be rewritten in common notation as:

$$(K^\alpha)^{-1}r = \sum_{j=1}^{\infty} \frac{q(\alpha, \lambda_j^2)}{\lambda_j}\langle r, \phi_j\rangle \phi_j$$

where for T: $q(\alpha, \lambda_j^2) = \dfrac{\lambda_j^2}{\lambda_j^2 + \alpha}$,
for LF: $q(\alpha, \lambda_j^2) = [1 - (1 - c\lambda_j^2)^{1/\alpha}]$,
for SC: $q(\alpha, \lambda_j^2) = I(\lambda_j^2 \geq \alpha)$, for PC $q(\alpha, \lambda_j^2) = I(j \leq 1/\alpha)$.

In order to compute the inverse of K_n we have to choose the regularization parameter α. Let $(K_n^\alpha)^{-1}$ be the regularized inverse of K_n and P^α a $n \times n$ matrix defined as in Carrasco (2012) by $P^\alpha = T(K_n^\alpha)^{-1}T^*$ where

$$T : L^2(\pi) \rightarrow \mathbb{R}^n$$

$$Tg = \begin{pmatrix} \langle Z_1, g \rangle \\ \langle Z_2, g \rangle \\ . \\ . \\ \langle Z_n, g \rangle \end{pmatrix}$$

and

$$T^* : \mathbb{R}^n \rightarrow L^2(\pi)$$

$$T^* v = \frac{1}{n} \sum_{i=1}^{n} Z_i v_i$$

such that $K_n = T^*T$ and TT^* is an $n \times n$ matrix with typical element $\frac{\langle Z_i, Z_j \rangle}{n}$. Let $\hat{\phi}_j$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... > 0$, $j = 1, 2, ...$ be the orthonormalized eigenfunctions and eigenvalues of K_n. $\hat{\lambda}_j$ are consistent estimators of λ_j the eigenvalues of TT^*. We then have $T\hat{\phi}_j = \sqrt{\lambda_j}\psi_j$ and $T^*\psi_j = \sqrt{\lambda_j}\hat{\phi}_j$.
For $v \in \mathbf{R}^n$, $P^\alpha v = \sum_{j=1}^{\infty} q(\alpha, \lambda_j^2) \langle v, \psi_j \rangle \psi_j$. It follows that for any vectors v and w of $\mathbf{R}^n$:

$$\begin{aligned} v' P^\alpha w &= v' T (K_n^\alpha)^{-1} T^* w \\ &= \left\langle (K_n^\alpha)^{-1/2} \sum_{i=1}^{n} Z_i(.)\, v_i, (K_n^\alpha)^{-1/2} \frac{1}{n} \sum_{i=1}^{n} Z_i(.)\, w_i \right\rangle. \end{aligned} \tag{1}$$

B Proofs

Proof of Proposition 1:

We first prove the consistency of our estimator.
Let $g_n = \frac{1}{n}\sum_{i=1}^{n} Z_i W_i = S_n \left[\frac{1}{n} \sum_{i=1}^{n} Z_i f_i \right] / \sqrt{n} + \frac{1}{n} \sum_{i=1}^{n} Z_i u_i = S_n g_{n1} / \sqrt{n} + g_{n2}$ (remember that g_n is a function indexed by τ and Z_i is also a function of τ, such a representation can handle both countable and continuum of instruments). Note that $g_{n2} = \frac{1}{n} \sum_{i=1}^{n} Z_i u_i = o_p(1)$, $\sqrt{n} g_{n2} = O_p(1)$ and $S_n/\sqrt{n}$ is bounded by Assumption 1(i).

$$\hat{\delta} - \delta_0 = (W'P^{\alpha}W)^{-1}W'P^{\alpha}\varepsilon$$

We have $S_n'(\hat{\delta} - \delta_0)/\mu_n = [S_n^{-1}W'P^{\alpha}WS_n^{-1'}]^{-1}[S_n^{-1}W'P^{\alpha}\varepsilon/\mu_n]$ and by construction[8] of P^{α} :

$$\begin{aligned} W'P^{\alpha}W &= n\left\langle (K_n^{\alpha})^{-1/2}g_n, (K_n^{\alpha})^{-1/2}g_n' \right\rangle \\ &= S_n\left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}g_{n1}' \right\rangle S_n' \\ &\quad + S_n\left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}g_{n2}' \right\rangle \sqrt{n} \\ &\quad + \left\langle (K_n^{\alpha})^{-1/2}g_{n2}, (K_n^{\alpha})^{-1/2}g_{n1}' \right\rangle S_n'\sqrt{n} \\ &\quad + \left\langle (K_n^{\alpha})^{-1/2}g_{n2}, (K_n^{\alpha})^{-1/2}g_{n2}' \right\rangle n. \end{aligned}$$

$$\begin{aligned} S_n^{-1}W'P^{\alpha}WS_n^{-1'} &= \left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}g_{n1}' \right\rangle \\ &\quad + \left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}\sqrt{n}g_{n2}' \right\rangle S_n^{-1'} \\ &\quad + S_n^{-1}\left\langle (K_n^{\alpha})^{-1/2}\sqrt{n}g_{n2}, (K_n^{\alpha})^{-1/2}g_{n1}' \right\rangle \\ &\quad + S_n^{-1}\left\langle (K_n^{\alpha})^{-1/2}\sqrt{n}g_{n2}, (K_n^{\alpha})^{-1/2}\sqrt{n}g_{n2}' \right\rangle S_n^{-1'}. \end{aligned}$$

Hence,

$$S_n^{-1}[W'P^{\alpha}W]S_n^{-1'} = \left\langle (K_n^{\alpha})^{-\frac{1}{2}}g_{n1}, (K_n^{\alpha})^{-\frac{1}{2}}g_{n1}' \right\rangle + o_p(1).$$

At this stage, we can apply the same proof as that of Proposition 1 of Carrasco (2012) which shows that

$$\left\langle (K_n^{\alpha})^{-\frac{1}{2}}g_{n1}, (K_n^{\alpha})^{-\frac{1}{2}}g_{n1}' \right\rangle \to \left\langle g_1, g_1' \right\rangle_K$$

in probability as n and $n\alpha^{\frac{1}{2}}$ go to infinity, with $\left\langle g_1, g_1' \right\rangle_K$ a $p \times p$ matrix with (a, b) element $\left\langle K^{-\frac{1}{2}}E(Z(.,x_i)f_{ia}), K^{-\frac{1}{2}}E(Z(.,x_i)f_{ib}) \right\rangle$ which is assumed to be nonsingular.

[8]Let g and h be two p vectors of functions of $L^2(\pi)$. By a slight abuse of notation, $\left\langle g, h' \right\rangle$; denotes the matrix with elements $\left\langle g_a, h_b \right\rangle$ $a, b = 1, ..., p$

$$\begin{aligned}
\frac{S_n^{-1}W'P^{\alpha}\varepsilon}{\mu_n} &= \frac{nS_n^{-1}}{\mu_n}\left\langle (K_n^{\alpha})^{-1/2}g_n, (K_n^{\alpha})^{-1/2}\frac{1}{n}\sum_{i=1}^{n}Z_i\varepsilon_i\right\rangle \\
&= \frac{1}{\mu_n}\left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i\right\rangle \\
&\quad +\frac{\mu_n S_n^{-1}}{\mu_n^2}\left\langle (K_n^{\alpha})^{-1/2}\sqrt{n}g_{n2}, (K_n^{\alpha})^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i\right\rangle \\
&= o_p(1)
\end{aligned}$$

because $\mu_n S_n^{-1} \to S_0$ by Assumption 1(ii) and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i = O_p(1)$. This proves the consistency of the regularized 2SLS.

For the asymptotic normality we write

$$S_n'(\hat{\delta}-\delta_0) = [S_n^{-1}W'P^{\alpha}WS_n'^{-1}]^{-1}[S_n^{-1}W'P^{\alpha}\varepsilon]$$

We then have

$$\begin{aligned}
S_n^{-1}W'P^{\alpha}\varepsilon &= nS_n^{-1}\big\langle (K_n^{\alpha})^{-1}g_n, \frac{1}{n}\sum_{i=1}^{n}Z_i\varepsilon_i\big\rangle \\
&= \left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i\right\rangle \\
&\quad +S_n^{-1}\left\langle (K_n^{\alpha})^{-1/2}\sqrt{n}g_{n2}, (K_n^{\alpha})^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i\right\rangle \\
&= \left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i\right\rangle \\
&\quad +o_p(1).
\end{aligned}$$

Moreover,

$$\left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i\right\rangle \tag{2}$$

$$= \big\langle (K_n^{\alpha})^{-1}g_{n1} - K^{-1}g_1, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i\big\rangle \tag{3}$$

$$+ \big\langle K^{-1}g_1, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i)\big\rangle.$$

The first term is negligible since
$\left\langle (K_n^\alpha)^{-1} g_{n1} - K^{-1} g_1, \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \varepsilon_i) \right\rangle \leq \|(K_n^\alpha)^{-1} g_{n1} - K^{-1} g_1\| \| \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \varepsilon_i \| = o_p(1) O_p(1).$

By the functional central limit theorem, we obtain the following result
$\left\langle K^{-1} g_1, \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \varepsilon_i \right\rangle \to \mathcal{N}(0, \sigma_\varepsilon^2 \left\langle g_1, g_1' \right\rangle_K)$ as n and $n\alpha$ go to infinity.

We then apply the continuous mapping theorem and Slutzky's theorem to show that $S_n'(\hat{\delta} - \delta_0) \xrightarrow{d} \mathcal{N}(0, \sigma_\varepsilon^2 \left\langle g_1, g_1' \right\rangle_K^{-1})$.

By assumption, $g_{1a} = E(Z(., x_i) f_{ia})$ belong to the range of K. Let $L^2(Z)$ be the closure of the space spanned by $\{Z(x, \tau), \tau \in I\}$ and g_1 is an element of this space. If $f_i \in L^2(Z)$ we can compute the inner product in the RKHS and show that

$$\left\langle g_{1a}, g_{1b} \right\rangle_K = E(f_{ia} f_{ib}).$$

This can be seen by applying Theorem 6.4 of Carrasco, Florens, and Renault (2007). It follows that

$$S_n'(\hat{\delta} - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 \left[E\left(f_i f_i'\right)\right]^{-1}\right)$$

This completes the proof of Proposition 1.

Proof of Proposition 2:

To prove this proposition, we need three lemmas. The first lemma corresponds to lemma A0 of Hansen, Hausman, and Newey (2008).

Lemma 1: Under assumption 1 if $\|S_n'(\hat{\delta}_L - \delta_0)/\mu_n\|^2 / (1 + \|\hat{\delta}_L\|^2) \xrightarrow{P} 0$ then $\|S_n'(\hat{\delta}_L - \delta_0)/\mu_n\| \xrightarrow{P} 0$.

Proof: The proof of this lemma is the same as that of lemma A0 in Hansen, Hausman, and Newey (2008).

Lemma 2: Suppose that the assumptions of Proposition 2 hold. Then,

$$Var(\varepsilon' P^\alpha u_a) \leq C(\sum_j q_j^2),$$

$$\begin{aligned} \varepsilon' P^\alpha u_a - E(\varepsilon' P^\alpha u_a | X) &= O((\sum_j q_j^2)^{\frac{1}{2}}), \\ \frac{\varepsilon' P^\alpha \varepsilon}{\mu_n^2} &= O_p\left(\frac{1}{\alpha \mu_n^2}\right) = o_p(1). \end{aligned}$$

Proof:

For notational simplicity, we suppress the conditioning on X. Let $E(\varepsilon_i^2) = \sigma_\varepsilon^2$, $E(\varepsilon_i u_{ai}) = \sigma_{\varepsilon u_a}$ and $E(u_{ai}' u_{ai}) = \sigma_{u_a}^2$,

$$Var(\varepsilon' P^\alpha u_a) = E(\varepsilon' P^\alpha u_a u_a' P^\alpha \varepsilon) - E(\varepsilon' P^\alpha u_a) E(u_a' P^\alpha \varepsilon).$$

Using the spectral decomposition of P^α, we have

$$\begin{aligned} E(\varepsilon' P^\alpha u_a u_a' P^\alpha \varepsilon) &= \frac{1}{n^2} \sum_{j,l} q_j q_l E\Big\{ (\varepsilon' \psi_l)(u_a' \psi_l)' (\varepsilon' \psi_j)(u_a' \psi_j) \Big\} \\ &= \frac{1}{n^2} \sum_{j,l} q_j q_l E\Big\{ \sum_i \varepsilon_i u_{ai}' \psi_{li}^2 \sum_b \varepsilon_b u_{ab} \psi_{jb}^2 \\ &+ \sum_c \varepsilon_c u_{ac}' \psi_{lc} \psi_{jc} \sum_d \varepsilon_d u_{ad} \psi_{jd} \psi_{ld} \\ &+ \sum_c \varepsilon_c^2 \psi_{lc} \psi_{jc} \sum_d u_{ad}' u_{ad} \psi_{jd} \psi_{ld} \Big\} \\ &= (\sum_j q_j)^2 \sigma_{\varepsilon u_a}' \sigma_{\varepsilon u_a} + (\sigma_{\varepsilon u_a}' \sigma_{\varepsilon u_a} + \sigma_\varepsilon^2 \sigma_{u_a}^2) \sum_j q_j^2 \end{aligned}$$

by the fact that (u_{ai}, ε_i) are independent across i and the eigenvectors are orthonormal.

$$\begin{aligned} E(\varepsilon' P^\alpha u_a) &= \frac{1}{n} \sum_l q_l E\{ (\sum_k u_{ak}' \psi_{lk}) (\sum_i \varepsilon_i \psi_{li}) \} \\ &= \frac{1}{n} \sum_l q_l n \sigma_{\varepsilon u_a}' \\ &= \sigma_{\varepsilon u_a}' (\sum_j q_j). \end{aligned}$$

Thus

$$Var(\varepsilon' P^\alpha u_a) = (\sigma_{\varepsilon u_a}' \sigma_{\varepsilon u_a} + \sigma_\varepsilon^2 \sigma_{u_a}^2) \sum_j q_j^2 \leq C (\sum_j q_j^2).$$

The second conclusion follows by Markov inequality.

$$\begin{aligned} E\left(\varepsilon' P^\alpha \varepsilon\right) &= tr\left(P^\alpha E\left(\varepsilon \varepsilon'\right)\right) \\ &= \sigma_\varepsilon^2 (\sum_j q_j) = O_p\left(1/\alpha\right). \end{aligned}$$

Using the result for $\varepsilon' P^\alpha u_a$ with ε in place of u_a, we obtain

$$Var(\varepsilon' P^\alpha \varepsilon) \leq C(\sum_j q_j^2).$$

It follows that $\left(\varepsilon' P^\alpha \varepsilon - E\left(\varepsilon' P^\alpha \varepsilon\right)\right)/\mu_n^2 = O_p\left(\left(\sum_j q_j^2\right)^{1/2}/\mu_n^2\right) = o_p\left(\sum_j q_j/\mu_n^2\right)$. Hence, the third equality holds.

Lemma 3: Suppose that the assumptions of Proposition 2 hold. Let $\hat{A} = \dfrac{f' P^\alpha f}{n}$ and $\hat{B} = \dfrac{\bar{W}' \bar{W}}{n}$ with $\bar{W} = [y, W]$, there exist two constants C and C' such that $\hat{A} \geq C I_p$ and $\|\hat{B}\| \leq C'$.

Proof: By the definition of P^α, we have (see Equation (1)):

$$\hat{A} = \frac{f' P^\alpha f}{n} = \left\langle (K_n^\alpha)^{-\frac{1}{2}} f_n, (K_n^\alpha)^{-\frac{1}{2}} f_n' \right\rangle$$

with

$$f_n = \frac{1}{n} \sum_i Z_i f_i.$$

By Lemma 5(i) of Carrasco (2012) and the law of large numbers,

$$\frac{f' P^\alpha f}{n} = \frac{f' f}{n} + o_p(1) = E(f_i' f_i) + o_p(1)$$

as α goes to zero. Because $E(f_i' f_i)$ is positive definite, there exists a constant C such that

$$\hat{A} \geq C I_p$$

with probability one.

We have $\bar{W} = [y, W] = W D_0 + \varepsilon e$ where $D_0 = [\delta_0, I]$, δ_0 is the true value of the

parameter and e is the first unit vector.

$$\begin{aligned}\hat{B} &= \frac{\bar{W}'\bar{W}}{n} \\ &= D_0'S_n\frac{f'f}{n}S_n'D_0/n + D_0'S_n\frac{f'u}{n}D_0/\sqrt{n} + D_0'S_n\frac{f'\varepsilon}{n}e/\sqrt{n} \\ &+ D_0'\frac{u'f}{n}S_n'D_0/\sqrt{n} + D_0'\frac{u'u}{n}D_0 + D_0'\frac{u'\varepsilon}{n}e \\ &+ e'\frac{\varepsilon'f}{n}S_n'D_0/\sqrt{n} + e'\frac{\varepsilon'u}{n}D_0 + e'\frac{\varepsilon'\varepsilon}{n}e.\end{aligned}$$

Using the law of large numbers, we can conclude that $\|\hat{B}\| \leq C'$, where C' is a constant, with probability one.

Proof of consistency

Let us consider

$$\hat{Q}(\delta) = \frac{(y - W\delta)'P^{\alpha}(y - W\delta)/\mu_n^2}{(y - W\delta)'(y - W\delta)/n}.$$

$$\hat{\delta}_L = argminQ(\delta).$$

For $\delta = \delta_0$, $\hat{Q}(\delta_0) = \dfrac{\varepsilon'P^{\alpha}\varepsilon/\mu_n^2}{\varepsilon'\varepsilon/n}$. With probability one $\varepsilon'\varepsilon/n > C$ and by lemma 2

$$\varepsilon'P^{\alpha}\varepsilon/\mu_n^2 = o_p(1).$$

Hence $\hat{Q}(\delta_0) = o_p(1)$.

Since $0 \leq \hat{Q}(\hat{\delta}_L) \leq \hat{Q}(\delta_0)$ it is easy to see that $\hat{Q}(\hat{\delta}_L) = o_p(1)$.

Let us show that

$$\mu_n^{-2}(y - W\delta)'P^{\alpha}(y - W\delta) \geq C\|S_n'(\delta - \delta_0)/\mu_n\|^2.$$

Let $D(\delta) = \mu_n^{-2}(y - W\delta)'P^{\alpha}(y - W\delta) = \mu_n^{-2}(1, -\delta')\bar{W}'P^{\alpha}\bar{W}(1, -\delta')'$. Moreover, $D(\delta) = \mu_n^{-2}(1, -\delta')D_0'S_n\dfrac{f'P^{\alpha}f}{n}S_n'D_0(1, -\delta')' + o_p(1) = \mu_n^{-2}(1, -\delta')D_0'S_nE\left(ff'\right)S_n'D_0(1, -\delta')' + o_p(1)$. It follows from lemma 3 that

$$D(\delta) \geq C\|S_n'(\delta - \delta_0)/\mu_n\|^2.$$

We also have that $(y - W\delta)'(y - W\delta)/n = (1, -\delta')\hat{B}(1, -\delta')'$. Hence,

$$\frac{\|S_n'(\hat{\delta}_L - \delta_0)/\mu_n\|^2}{(1 + \|\hat{\delta}_L\|^2)} \leq C\hat{Q}(\hat{\delta}_L).$$

Then by Lemma 1 we have $S_n'(\hat{\delta}_L - \delta_0)/\mu_n \to 0$ in probability as n and $\mu_n^2\alpha$ go to infinity. This proves the consistency of LIML with many weak instruments.

Now let us prove the asymptotic normality.

Proof of asymptotic normality

Denote $A(\delta) = (y - W\delta)'P^\alpha(y - W\delta)/2$, $B(\delta) = (y - W\delta)'(y - W\delta)$ and

$$\Lambda(\delta) = \frac{A(\delta)}{B(\delta)}.$$

We know that the LIML is $\hat{\delta}_L = argmin\Lambda(\delta)$.

We calculate the gradient and Hessian $\Lambda_\delta(\delta) = B(\delta)^{-1}[A_\delta(\delta) - \Lambda(\delta)B_\delta(\delta)]$,

$\Lambda_{\delta\delta}(\delta) = B(\delta)^{-1}[A_{\delta\delta}(\delta) - \Lambda(\delta)B_{\delta\delta}(\delta)] - B(\delta)^{-1}[B_\delta(\delta)\Lambda_\delta'(\delta) - \Lambda_\delta(\delta)B_\delta'(\delta)]$.

Then by the mean-value theorem applied to the first-order condition $\Lambda_\delta(\hat{\delta}) = 0$, we have:

$$S_n'(\hat{\delta}_L - \delta_0) = -[S_n^{-1}\Lambda_{\delta\delta}(\tilde{\delta})S_n^{-1'}]^{-1}[S_n'\Lambda_\delta(\delta_0)]$$

where $\tilde{\delta}$ is the mean-value. By the consistency of $\hat{\delta}_L$, $\tilde{\delta} \to \delta_0$.

It then follows that

$$\begin{aligned} B_\delta(\tilde{\delta})/n &= -2\sum_i W_i\tilde{\varepsilon}_i/n, \\ &= -2\sum_i(\gamma_i + u_i)\tilde{\varepsilon}_i/n \\ &= -2S_n/\sqrt{n}(\sum_i f_i\tilde{\varepsilon}_i/n) - 2(\sum_i u_i\tilde{\varepsilon}_i/n) \\ &= -2\sigma_{u\varepsilon} + o_p(1) \end{aligned}$$

under the assumption that $S_n/\sqrt{n}$ is bounded, with $\tilde{\varepsilon}_i = (y_i - W_i'\tilde{\delta})$ and $\sigma_{u\varepsilon} = E(u_i\varepsilon_i)$.

$$B(\tilde{\delta})/n \xrightarrow{P} \sigma_\varepsilon^2,\ B_\delta(\tilde{\delta})/n \xrightarrow{P} -2\sigma_{u\varepsilon}$$

$$\Lambda(\delta) = \frac{(y - W\delta)' P^{\alpha} (y - W\delta)/2n}{(y - W\delta)'(y - W\delta)/n}$$

For $\delta = \delta_0$, $\Lambda(\delta_0) = \dfrac{\varepsilon' P^{\alpha} \varepsilon/2n}{\varepsilon'\varepsilon/n}$. With probability one, $\varepsilon'\varepsilon/n > C$, and by Lemma 2 and $\mu_n^2 \leq n$,

$$\varepsilon' P^{\alpha} \varepsilon/n = o_p(1).$$

We have $\Lambda(\delta_0) = o_p(1)$. Therefore, $\Lambda(\tilde{\delta}) \xrightarrow{P} 0$. By the first order condition, we also have

$$\Lambda_{\delta}(\tilde{\delta}) \xrightarrow{P} 0.$$

$B_{\delta\delta}(\tilde{\delta}) = 2W'W/n \xrightarrow{P} 2E(W_i W_i')$, $A_{\delta\delta}(\tilde{\delta})/n = W' P^{\alpha} W/n$.

We can then conclude that $\Lambda_{\delta\delta}(\tilde{\delta}) = nB^{-1}(\tilde{\delta})[A_{\delta\delta}(\tilde{\delta})/n] + o_p(1)$. Hence

$$\begin{aligned} n\tilde{\sigma}_{\varepsilon}^2 \Lambda_{\delta\delta}(\tilde{\delta}) &= W' P^{\alpha} W \\ &= S_n \left\langle (K_n^{\alpha})^{-\frac{1}{2}} g_{n1}, (K_n^{\alpha})^{-\frac{1}{2}} g_{n1}' \right\rangle S_n' + o_p(1) \\ &= S_n H S_n' + o_p(1) \end{aligned}$$

with $H = E(f(x_i) f(x_i)')$ and $\tilde{\sigma}_{\varepsilon}^2 = (y - W\tilde{\delta})'(y - W\tilde{\delta})/n$.

Hence

$$n\tilde{\sigma}_{\varepsilon}^2 S_n^{-1} \Lambda_{\delta\delta}(\tilde{\delta}) S_n^{-1'} = H + o_p(1).$$

Let $\hat{\phi} = \dfrac{W'\varepsilon}{\varepsilon'\varepsilon}$, $\phi = \dfrac{\sigma_{u\varepsilon}}{\sigma_{\varepsilon}^2}$ and $v = u - \varepsilon\phi'$. It is useful to remark that $v' P^{\alpha} \varepsilon = O_p(1/\sqrt{\alpha})$ using Lemma 2 with v in place of u and $E(u_i v_i) = 0$. Moreover, $\hat{\phi} - \phi = O_p(1/\sqrt{n})$ by the central limit theorem and the delta method. Hence, $(\hat{\phi} - \phi)\varepsilon' P^{\alpha} \varepsilon = O_p(1/\alpha\sqrt{n})$.

Furthermore, $f'(I - P^{\alpha})\varepsilon/\sqrt{n} = O_p(\Delta_{\alpha}^2) = o_p(1)$ by Lemma 5(ii) Carrasco (2012) with $\Delta_{\alpha} = tr(f'(I - P^{\alpha})^2 f/n) = O_p\left(\alpha^{\min(\beta,2)}\right) = o_p(1)$. We have

$$\begin{aligned} -n\tilde{\sigma}^2 S_n^{-1} \Lambda_{\delta}(\delta_0) &= S_n^{-1}(W' P^{\alpha} \varepsilon - \varepsilon' P^{\alpha} \varepsilon \frac{W'\varepsilon}{\varepsilon'\varepsilon}) \\ &= f'\varepsilon/\sqrt{n} - f'(I - P^{\alpha})\varepsilon/\sqrt{n} + S_n^{-1} v' P^{\alpha} \varepsilon - S_n^{-1}(\hat{\phi} - \phi)\varepsilon' P^{\alpha} \varepsilon \\ &= f'\varepsilon/\sqrt{n} + o_p(1) + S_n^{-1} O_p(1/\sqrt{\alpha}) + S_n^{-1} O_p(1/\alpha\sqrt{n}) \\ &= f'\varepsilon/\sqrt{n} + o_p(1) \xrightarrow{d} \mathcal{N}(0, \sigma_{\varepsilon}^2 H) \end{aligned}$$

as n, $\alpha\mu_n^2$ go to infinity under the assumption $\mu_n S_n^{-1} \to S_0$.

The conclusion follows from Slutzky's theorem.