Université de Montréal

Regularized Jackknife estimation with many instruments

par

Mohamed Doukali

Département de sciences économiques

Faculté des arts et sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en sciences économiques

Septembre, 2016

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Regularized Jackknife estimation with many instruments

présentée par :
Mohamed Doukali

a été évaluée par un jury composé des personnes suivantes :

René Garcia,	président-rapporteur
Marine Carrasco,	directrice de recherche
Benoit Perron,	membre du jury
Prosper Dovonon,	examinateur externe
Pierre Duchesne,	représentant du doyen de la FAS

Thèse acceptée le 25 octobre 2016

à mes parents, Abdelaziz Doukali et Hafida Moutaouakil Alaoui

Remerciements

J'ai réalisé ce travail grâce à l'appui de nombreuses personnes qui m'ont aidé directement ou indirectement. La liste de ces personnes ne sera jamais exhaustive.

Je tiens tout d'abord à remercier ma directrice de recherche Marine Carrasco pour sa présence continue, ses précieux conseils, pour m'avoir accompagné dans toute cette thèse et surtout d'avoir cru en moi. Elle est la personne clé dans la réalisation de cette thèse. Merci infiniment Marine. Mes remerciements vont également à Benoit Perron, qui m'a suivi depuis la première présentation orale de cette thèse, pour ses conseils, ses suggestions et critiques constructives.

Je tiens aussi à remercier l'ensemble du corps professoral du département de sciences économiques de l'Université de Montréal, particulièrement Ilze Kalnina, William Mccausland, Silvia Gonçalves, Lars Ehlers, Yves Sprumont, Emanuela Cardia et Michel Poitevin qui m'ont dispensé les enseignements tout au long du programme.

Je n'oublierai pas de remercier mes amis et collègues du programme de Doctorat avec qui j'ai passé des moments inoubliables, particulièrement ces discussions enrichissantes dans les couloirs du département qui m'ont permis de raffermir mes connaissances en économie. Je remercie en particulier Alex, Ada, Jonathan, David, Guy, Idrissa, Ismail, Désiré, Juste, Laurent, Louphou, Vincent et Tewou. Ce parcours fut difficile d'un point de vue intellectuel, mais aussi moral, à des milliers de Kilomètres du beau soleil du Maroc; je devais faire face à ce rude hiver de Montréal. J'ai trouvé sur place une grande "famille" qui m'a sans cesse soutenu dans toutes les épreuves. Je ne pourrai terminer sans remercier mes chers amis Nabil et Omar qui ont été d'un soutien permanent.

Je remercie le Département de sciences économiques de l'Université de Montréal, le Centre Inter-universitaire de Recherche en économie Quantitative (CIREQ) et la Faculté des études supérieures et postdoctorales pour le soutien financier et logistique. Mes remerciements vont également à tout le personnel du département de sciences économiques de l'université de Montréal et à celui du CIREQ.

Enfin, je tiens à remercier ma famille: mes parents Abdelaziz et Hafida pour leur soutien inconditionnel, pour m'avoir encouragé à dépasser mes limites, et surtout cette éducation qu'ils m'ont inculquée. Mon merveilleux petit frère Saad qui m'a toujours gratifié de ses sages conseils et de sa bonne humeur. Mes deux sœurs Sara et Sanae pour leur soutien constant. Je ne peux que remercier Dieu de m'avoir gratifié d'une telle famille.

Résumé

Dans cette thèse, je me suis intéressé aux modèles à variables instrumentales (VI) où les instruments sont nombreux et potentiellement faibles. La théorie asymptotique n'étant pas toujours une bonne approximation de la distribution d'échantillonnage des estimateurs et statistiques de tests, je considère la régularisation et la méthode Jackknife pour améliorer la précision des estimateurs.

Dans le premier chapitre (co-écrit avec Marine Carrasco), nous étudions l'estimation efficace d'un paramètre de dimension finie dans un modèle linéaire où le nombre d'instruments peut être très grand. Cependant, en échantillons finis, l'utilisation d'un grand nombre de conditions de moments accroit le biais des estimateurs VI. Cette situation pourrait s'aggraver en présence d'instruments faibles. Nous proposons une version régularisée de l'estimateur Jackknife (RJIVE) basée sur trois méthodes de régularisations différentes, Tikhonov, Landweber Fridman et composantes principales, qui réduisent le biais. Nous montrons par la suite que les estimateurs RJIVE sont convergents et asymptotiquement normaux. Ces méthodes font chacune intervenir un paramètre d'ajustement, qui doit être sélectionné. Nous dérivons une méthode basée uniquement sur les données pour sélectionner le paramètre de régularisation, i.e. minimiser la perte espérée d'utilité. Des simulations Monte Carlo montrent que nos estimateurs proposés se comportent mieux en comparaison avec l'estimateur Jackknife sans régularisation.

Dans le deuxième chapitre (co-écrit avec Marine Carrasco), nous proposons une version modifiée du test de suridentification dans un contexte où le nombre d'instruments peut être très grand. Notre test d'hypothèse combine deux techniques: la méthode de Jackknife et la technique de Tikhonov. Nous montrons théoriquement que ledit test atteint asymptotiquement le seuil de probabilité en dessous duquel on est prêt à rejeter l'hypothèse nulle. Les simulations montrent la dominance de notre test par rapport à d'autres J tests existants dans la littérature en terme de niveau et de puissance du test.

Dans le dernier chapitre, je propose un nouveau estimateur basé sur la version Jackknife de l'estimateur du maximum de vraisemblance à information limitée régularisé (JLIML) dans un environnement riche en données où le nombre d'instruments (possiblement faibles) peut être aussi très grand. Je montre que l'estimateur JLIML régularisé est convergent et asymptotiquement normal. Les propriétés des estimateurs proposés sont évaluées à travers une étude Monte-Carlo, et une illustration empirique portant sur l'élasticité de substitution inter-temporelle.

Mots-clés: Modèles de grande dimension, Jackknife, Régularisation, Variable instrumentale faibles, Test de suridentification, Erreur quadratique moyenne, Hétéroscédasticité.

Abstract

In this thesis, I have been interested in the instrumental variables (IV) models with many instruments and possibly, many weak instruments. Since the asymptotic theory is often not a good approximation to the sampling distribution of estimators and test statistics, I consider the Jackknife and regularization methods to improve the precision of IV models.

In the first chapter (co-authored with Marine Carrasco), we consider instrumental variables (IV) regression in a setting where the number of instruments is large. However, in finite samples, the inclusion of an excessive number of moments may increase the bias of IV estimators. Such a situation can arise in presence of many possibly weak instruments. We propose a Jackknife instrumental variables estimator (RJIVE) combined with regularization techniques based on Tikhonov, Principal Components and Landweber Fridman methods to stabilize the projection matrix. We prove that the RJIVE is consistent and asymptotically normally distributed. We derive the rate of the mean square error and propose a data-driven method for selecting the tuning parameter. Simulation results demonstrate that our proposed estimators perform well relative to the Jackknife estimator with no regularization.

In the second chapter (co-authored with Marine Carrasco), we propose a new overidentifying restrictions test in a linear model when the number of instruments (possibly weak) may be smaller or larger than the sample size or even infinite in a heteroskedastic framework. The proposed J test combines

two techniques: the Jackknife method and the Tikhonov technique. We theoretically show that our new test achieves the asymptotically correct size in the presence of many instruments. The simulations show that our modified J statistic test has better empirical properties in small samples than existing J tests in terms of the empirical size and the power of the test.

In the last chapter, I consider instrumental variables regression in a setting where the number of instruments is large. However, in finite samples, the inclusion of an excessive number of moments may be harmful. We propose a Jackknife Limited Information Maximum Likelihood (JLIML) based on three different regularizations methods: Tikhonov, Landweber-Fridman, and Principal Components. We show that our proposed regularized Jackknife estimators JLIML are consistent and asymptotically normally distributed under heteroskedastic error. Finally, the proposed estimators are assessed through Monte Carlo study and illustrated using the elasticity of intertemporal substitution empirical example.

Keywords: High-dimensional models, Jackknife, Regularization methods, Overidentification test, Many weak instruments, MSE, Heteroskedasticity.

Contents

List of Figures

List of Tables

Chapter 1

Efficient estimation using regularized Jackknife IV estimator.

1.1 Introduction

Instrumental variables (IV) regression is largely used in economic research to calculate treatment effects for endogenous regressors[1]. However, IV estimates of structural effects are often imprecise in practice. One solution to increase the precision of IV estimators is to use all the moment conditions available. Empirical examples such as Eichenbaum, Hansen, and Singleton (1988) who consider consumption asset pricing models and Angrist and Krueger (1991) who measure return to schooling, were the first to show the problem of the presence of many instruments. Theoretical literature on weak/strong instruments showed also that the inclusion of many moments can improve the precision of IV estimators but the usual Gaussian asymptotic approximation can be poor and IV estimators may be biased (see among others Staiger and Stock (1997) and Chao and Swanson (2005), Hansen, Hausman, and Newey (2008), Chao, Swanson, Hausman, Newey, and Woutersen (2012a) (CSHNW), Bekker (1994) and Newey and Smith (2004)). To deal with the problem of many instruments, this paper proposes new estimators for instrumental variables models when the number of instruments is not restricted and may be smaller or larger than the sample size n or even infinite. We propose a regularized version of the Jackknife estimator based on three regularization techniques. The first estimator is based on the Tikhonov method also called the ridge regularization, the second estimator is based on an iterative method called Landweber-Fridman (LF) and the third estimator is based on the principal components associated with the largest eigenvalues (see Kress (1999) and Carrasco, Florens, and Renault (2007) for a review of regularization schemes). We consider a linear model with heteroskedastic errors and allow for weak identification as in Hansen, Hausman, and

[1]This chapter is a joint work with Marine Carrasco. The authors thank the participants of the CEA 2015, of the 54th Société Canadienne de Science économique, of the 11th CIREQ Conference, and of the seminar at the Queen's University for helpful comments.

Newey (2008) and Newey and Windmeijer (2009). This specification helps us to have different types of weak instruments sequences, including sequence of Bekker (1994) and have many weak instruments of Chao and Swanson (2005). We show that the regularized Jackknife estimators are consistent and asymptotically normal under heteroskedastic error. In the homoskedastic case with strong instruments, our estimators reach the semiparametric efficiency bound.

All regularization procedures involve a regularization parameter α, which is the counterpart of the smoothing parameter in the nonparametric literature. In this paper, we develop a data driven method to select the regularization parameter by minimizing the higher-order expansion of the mean square error (MSE) of our estimators when the instruments are strong.

The simulations show that the leading regularized Jackknife estimators based on the Tikhonov and Landweber-Fridman techniques perform very well (are nearly median unbiased) in comparison with other existing estimators even in the case of weak instruments.

Our paper is related to some other papers in the literature about many instruments. Carrasco (2012) and Carrasco and Doukali (2016) propose regularized versions of the two-stage least squares (2SLS) estimator and the limited information maximum likelihood (LIML) estimator for many instruments. Their proposed estimators are consistent and asymptotically normally distributed in presence of a very large number of instruments. Chao, Swanson, Hausman, Newey, and Woutersen (2012a) (CSHNW) derive the limiting distribution of the Jackknife instrumental variables estimator and give formulas for consistent standard errors in the presence of heteroskedasticity and many instruments, but their estimator perform poorly when the number of instruments L is larger than the sample size n. Regularization has also been introduced in the context of times series and forecasting macroeconomics series using a large number of predictors. In this context, it is assumed that

there is a fixed number of factors that provide a good estimation (see Stock and Watson (2002), Bai and Ng (2010) and De Mol, Giannone, and Reichlin (2008)). Hansen and Kozbur (2014) propose an estimation and inference procedure in the presence of very many instruments, they use a Jackknife estimator combined with a ridge regularization. The condition they impose allow for the number of instruments L to be larger than the sample size n. Also, their estimation procedure do not assume sparsity, in other words, they do not require any prior information about the ordering or the strength of instruments, all instruments are used even if they are weak.
The rest of the paper is organized as follows. Section 2 sets up the model and introduces the regularization methods. Section 3 derives the asymptotic properties of the estimators. Section 4 derives the rate of convergence of the mean square error. Section 5 proposes a data-driven selection of the regularization parameter. Section 6 presents Monte Carlo experiments. An empirical application to measuring the return to education is illustrated in Section 7. Section 8 concludes. The proofs are collected in Appendix.

1.2 Presentation of the regularized Jackknife model

This section presents the model and the regularized Jackknife estimators. It is important to note that throughout the paper, the number of instruments L is not restricted and may be smaller or larger than the sample size n.
We consider the same model as in Chao et al. (2012a):

$$y_i = X_i'\delta_0 + \epsilon_i \tag{1.2.1}$$

$$X_i = \Upsilon_i + u_i \tag{1.2.2}$$

$i = 1, \ldots, n$. The vector of interest is δ_0 which is a $p \times 1$ vector. y_i is a scalar. The vector Υ_i is the optimal instrument which is typically unknown. We assume that y_i and X_i are observed but the Υ_i is not. $E(X_i\epsilon_i) \neq 0$, as a result, X_i is endogenous and the OLS estimator of δ_0 is not consistent. The estimation will be based on a sequence of instruments $Z_i = Z(\tau; \nu_i)$ where ν_i is a vector of exogenous variables and τ is an index taking countable values. Such a situation can arise by taking interactions between some exogenous variables as in Angrist and Krueger (1991), or by non-linear transformations of an exogenous variable as in Dagenais and Dagenais (1997) , or also by allowing lagged dependent variables as in Arellano and Bond (1991).

Assumption 1. *y_i, X_i and ν_i are iid, $E(u_i|\nu_i) = E(\epsilon_i|\nu_i) = 0$; $\Upsilon_i = E(X_i|\nu_i)$ denote the $p \times 1$ reduced form vector with $E(\Upsilon_i\epsilon_i) = 0$.*

It is well known that the two-stage least squares (2SLS) estimator suffers from a small-sample bias in presence of endogeneity that is increased dramatically when many instruments are used and/or the instruments are only weakly correlated with the endogenous variables, see Yogo (2004). To solve the problem of the correlation between estimated instruments and first-stage errors, researchers have proposed the Jackknife method. The Jackknife estimator was first suggested by Phillips and Hale (1977) and popularized by Angrist, Imbens, and Krueger (1999) by using the leave-one-out observation approach to reduce the bias of the 2SLS estimator. CSHNW study the properties of the instrumental variable estimator in the case of many possibly weak instruments, but they assume that the number of instruments L grows slower than the sample size n, which is not the case in our work. Recently, Chao et al. (2012b) propose a Jackknife version of the LIML estimator where again L is smaller than n.

First we recall the expression of the usual Jackknife instrumental variable estimator (JIVE) when the number of instruments is finite.

$$\hat{\delta} = (\hat{\Upsilon}'X)^{-1}(\hat{\Upsilon}'Y) \tag{1.2.3}$$

$$\hat{\delta} = (\sum_{i=1}^{n} \hat{\Upsilon}_i X_i')^{-1} \sum_{i=1}^{n} \hat{\Upsilon}_i y_i \tag{1.2.4}$$

The leave-one-out estimator $\hat{\Upsilon}_i$ is defined as $\hat{\Upsilon}_i = Z_i'\hat{\pi}_{-i}$, where $\hat{\pi}_{-i} = (Z'Z - Z_i Z_i')^{-1}(Z'X - Z_i X_i')$ is the OLS coefficient from running a regression of X on Z using all but the i^{th} observation.
Using the formulation from CSHNW:

$$\hat{\delta} = (\sum_{i=1}^{n} \hat{\pi}_{-i}' Z_i X_i')^{-1} \sum_{i=1}^{n} \hat{\pi}_{-i} Z_i y_i \tag{1.2.5}$$

$$\hat{\pi}_{-i}' Z_i = (X'Z(Z'Z)^{-1}Z_i - P_{ii}X_i)/(1 - P_{ii}) = \textstyle\sum_{i\neq j}^{n} P_{ij}X_j/(1 - P_{ii})$$

where $\sum_{i\neq j}$ denotes the double sum $\sum_i \sum_{j\neq i}$ and P is a $n \times n$ matrix defined as $P = Z(Z'Z)^{-1}Z'$ and P_{ij} denotes the *(i,j)*th element of P.
Then, the JIVE is given by:
$\hat{\delta} = \hat{H}^{-1} \sum_{i\neq j}^{n} X_i P_{ij}(1 - P_{jj})^{-1} y_j$, where $\hat{H} = \sum_{i\neq j}^{n} X_i P_{ij}(1 - P_{jj})^{-1} X_j'$.
When the number of the instruments is large, the inverse of Z'Z needs to be stabilized because it is singular or nearly singular. There are many influential papers on how to deal with many instruments. Bai and Ng (2010) assume that the endogenous variables depend on a small number of factors which are exogenous, they use estimated factors as instruments, but their variable selection procedure is based on the assumption that a small number of selected factors may be a good-approximation of the endogenous variables, and so this variable selection scheme requires a prior information about the model. Belloni, Chen, Chernozhukov, and Hansen (2012) apply an instru-

mental selection method based on Lasso when there is a low dimensional set of instruments that leads to a good approximation of the relationship between instruments and the endogenous variables. Donald and Newey (2001) reduce the dimension of the instruments set by selecting the number of instruments which minimizes an approximate mean square error. However, an ad hoc selection of instruments leads to a loss of efficiency because some instruments are discarded a priori. In this paper, we keep all available instruments by applying regularization on the inverse of $Z'Z$.

Now let us suppose that the number of instruments is finite or countable infinite as in Carrasco (2012). Here are some examples of Z_i.

- If $Z_i = \nu_i$ where ν_i is a L-vector of exogenous variables with a fixed L, then $Z(\tau;\nu_i)$ denotes the τth element of ν_i.

- $Z(\tau;\nu_i) = (\nu_i)^{\tau-1}$ with $\tau \in N$, thus we have an infinite countable sequence of instruments.

Assume τ lies in a space Ξ ($\Xi = \{1,..,L\}$ or $\Xi = \mathbb{N}$) and let π be a positive measure on Ξ. Let K be the covariance operator for instruments from $L^2(\pi)$ to $L^2(\pi)$ such that:

$$(Kg)(\tau) = \sum_{l=1}^{L} E(Z(\tau,\nu_i)Z(\tau_l,\nu_i))g(\tau_l)\pi(\tau_l).$$

where $L^2(\pi)$ denotes the Hilbert space of square integrable functions with respect to π. K is supposed to be a nuclear operator which means that its trace is finite. Let λ_j and ψ_j, $j = 1...$ be respectively the eigenvalues (ordered in decreasing order) and the orthogonal eigenfunctions of K. The operator can be estimated by K_n defined as:

$$K_n : L^2(\pi) \to L^2(\pi)$$
$$(K_n g)(\tau) = \sum_{l=1}^{L} \frac{1}{n} \sum_{i=1}^{n} (Z(\tau,\nu_i)Z(\tau_l,\nu_i))g(\tau_l)\pi(\tau_l).$$

If the number of instruments L is large relatively to n, inverting the operator K is considered as an ill-posed problem which means that the inverse is not continuous and its sample counterpart, K_n, is singular or nearly singular. To solve this problem we need to stabilize the inverse of K_n using regularization. A regularized inverse of an operator K is defined as: $R_\alpha : L^2(\pi) \to L^2(\pi)$ such that $lim_{\alpha\to 0} R_\alpha K\rho = \rho, \forall \rho \in L^2(\pi)$, where α is the regularization parameter (see Kress (1999) and Carrasco, Florens, and Renault (2007)).

Three types of regularization:

We consider three regularization schemes.

[1] Tikhonov (ridge) regularization:

$$\begin{aligned} (K^\alpha)^{-1} &= \left(K^2 + \alpha I\right)^{-1} K. \\ (K^\alpha)^{-1} r &= \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j^2 + \alpha} \langle r, \psi_j \rangle \psi_j. \end{aligned}$$

where $\alpha > 0$, $r \in R^n$ and I is the identity operator.

[2] Spectral cut-off or principal components:

It consists in selecting the eigenfunctions associated with the eigenvalues greater than some threshold.

$$(K^\alpha)^{-1} r = \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j} \langle r, \psi_j \rangle \psi_j.$$

for some $\alpha > 0$ and $r \in R^n$.

[3] Landweber-Fridman iterative method. Let $0 < c < \frac{1}{\lambda_1^2(K)}$ where $\lambda_1(K)$

is the largest eigenvalues of K. define:

$$\begin{aligned} \psi_k &= (1 - cK^2)\psi_{k-1} + cKr,\ k = 1, 2, ...1/\alpha - 1, \\ \psi_0 &= cKr. \end{aligned}$$

where $1/\alpha - 1$ is some positive integer. ψ_k converges to $K^{-1}r$ when the number of iterations k goes to infinity. The earlier we stop the iterations, the more stable is ψ_k. Alternatively, we have:

$$(K^\alpha)^{-1}r = \sum_j^\infty \frac{1 - \left(1 - c\lambda_j^2\right)^{1/\alpha}}{\lambda_j} \langle r, \psi_j \rangle \psi_j.$$

These three regularized inverses of K can be rewritten using a common notation as:

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

where:

- $q\left(\alpha, \lambda_j^2\right) = \lambda_j^2 / \left(\alpha + \lambda_j^2\right)$ for Tikhonov,
- $q\left(\alpha, \lambda_j^2\right) = I\left(\lambda_j^2 \geq \alpha\right)$ for spectral cut-off,
- $q\left(\alpha, \lambda_j^2\right) = 1 - \left(1 - c\lambda_j^2\right)^{1/\alpha}$ for Landweber-Fridman.

Let $(K_n^\alpha)^{-1}$ be the regularized inverse of K_n and P^α a $n \times n$ matrix as defined in Carrasco (2012) by $P^\alpha = T(K_n^\alpha)^{-1}T^*$ where $T : L^2(\pi) \to R^n$ with $Tg = (\langle Z_1, g\rangle, \langle Z_2, g\rangle', ..., \langle Z_n, g\rangle')'$ and $T^* : R^n \to L^2(\pi)$ with $T^*v = \frac{1}{n}\sum_j^n Z_j v_j$.
such that $K_n = T^*T$ and TT^* is a $n \times n$ matrix with typical element $\frac{<Z_i, Z_j>}{n}$. Let $\hat{\phi}_j$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq .... \geq 0$, $j = 1, 2, ...$ be the orthonormalized eigenfunctions and eigenvalues of K_n and ψ_j the eigenfunctions of TT^*. We then have $T\hat{\phi}_j = \sqrt{\lambda}\psi_j$ and $T^*\psi_j = \sqrt{\lambda}_j\hat{\phi}_j$. For $v \in R^n$, $P^\alpha v = \sum_j^\infty q(\alpha, \lambda_j^2) < v, \psi_j > \psi_j$.

In the finite dimensional case, and for an arbitrary $n \times 1$ vector, r, we define the $n \times n$ matrix, P^α, as

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

where α is the regularization parameter. The case $\alpha = 0$ corresponds to the case without regularization, $q\left(\alpha, \lambda_j^2\right) = 1$. Then, we obtain $P^0 = P = Z(Z'Z)^{-1}Z'$.
The regularized version of JIVE is given by:

$$\hat{\delta} = \hat{H}^{-1} \sum_{i \neq j}^{n} X_i P_{ij}^\alpha (1 - P_{jj}^\alpha)^{-1} y_j, \tag{1.2.6}$$

where

$$\hat{H} = \sum_{i \neq j}^{n} X_i P_{ij}^\alpha (1 - P_{jj}^\alpha)^{-1} X_j' \tag{1.2.7}$$

and P_{ij}^α denotes the *(i,j)th* element of P^α. In the special case of ridge regularization, $\hat{\delta}$ has been introduced by Hansen and Kozbur (2014).
Let $\xi_i = (1 - P_{ii}^\alpha)^{-1} \epsilon_i$ and substituting $y_i = X_i' \delta_0 + \epsilon_i$, we have :

$$\hat{\delta} = \delta_0 + \hat{H}^{-1} \sum_{i \neq j}^{n} X_i P_{ij}^\alpha \xi_j. \tag{1.2.8}$$

1.3 Asymptotic Properties of RJIVE

In this section, we establish the asymptotic properties of the Jackknife regularized IV estimator when the errors are heteroskedastic. We also allow for the presence of many weak instruments as in Staiger and Stock (1997). A measure of the strength of the instruments is the concentration parameter, which can be seen as a measure of the information contained in the instruments. If one could approximate the reduced form Υ by a sequence of instruments Z, so that $X = Z'\pi + u$ where $E[u^2|Z] = \sigma_u^2$, the concentration parameter would be given by:

$$CP = \tfrac{\pi' Z' Z \pi}{\sigma_u^2}.$$

Assumption 2. *$\Upsilon_i = S_n f_i/\sqrt{n}$ where $S_n = \hat{S}_n diag(\mu_{1n}, \ldots, \mu_{pn})$ such that $\hat{S}_n$ is $p \times p$ bounded matrix, the smallest eigenvalue of $\hat{S}_n \hat{S}_n'$ is bounded away from zero; for each j, either $\mu_{jn} = \sqrt{n}$ (strong identification) or $\frac{\mu_{jn}}{\sqrt{n}} \to 0$ (weak identification). Moreover $\mu_n = \min_{1<j<p} \mu_{jn} \to \infty$ and $1/(\sqrt{\alpha}\mu_n^2) \to 0$, $\alpha \to 0$. Also there is a constant $\bar{C}$ such that $||\sum_{i=1}^n f_i f_i'/n|| \leq \bar{C}$ and $\lambda_{min}(\sum_{i=1}^n f_i f_i'/n) \geq 1/\bar{C}$, a.s.*

Assumption 2 allows for both strong and weak instruments. If $\mu_{jn} = \sqrt{n}$, the instrument is strong. If μ_{jn}^2 is growing slower than n, this leads to a weak identification as that of Chao and Swanson (2005) and CSHNW. f_i defined in Assumption 2 is unobserved and has the same dimension as the infeasible optimal instrument, Υ_i. Then f_i can be seen as a rescaled version of this optimal instrument.

An illustration of assumption 2 is as follows. Let us consider the simple linear model $y_i = z_{i1}\delta_1 + \delta_{0p} x_{i2} + \epsilon_i$, where z_{i1} is an included instruments and x_{i2} is an endogenous variable. Suppose that x_{i2} is a linear combination of the included instrumental z_{i1} and an unknown excluded instruments z_{ip}, i.e $x_{i2} = \pi_1 z_{i1} + (\frac{\mu_n}{\sqrt{n}}) z_{ip}$. The reduced form is:

$$\Upsilon_i = \begin{pmatrix} z_{i1} \\ x_{i2} \end{pmatrix} = \begin{pmatrix} z_{i1} \\ \pi_1 z_{i1} + (\frac{\mu_n}{\sqrt{n}}) z_{ip} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \pi_1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{\mu_n}{\sqrt{n}} \end{pmatrix} \begin{pmatrix} z_{i1} \\ z_{ip} \end{pmatrix}$$

with

$$\hat{S}_n = \begin{pmatrix} 1 & 0 \\ \pi_1 & 1 \end{pmatrix}, \ \mu_{jn} = \begin{cases} \sqrt{n} & , \quad j = 1 \\ \mu_n & , \quad j = 2 \end{cases}, \text{ with } \frac{\mu_n}{\sqrt{n}} \to 0, \text{ and } f_i = \nu_i = \begin{pmatrix} z_{i1} \\ z_{ip} \end{pmatrix}.$$

Assumption 3. *There is a constant $\bar{C}$, such that conditional on $\mathcal{Z} = (\Upsilon, Z)$,* *the observations* $(\epsilon_1, u_1), ..., (\epsilon_n, u_n)$ *are independent, with* $E[\epsilon_i|\mathcal{Z}] = 0$ *for all* i*,* $E[u_i|\mathcal{Z}] = 0$ *for all* i*,* $sup_i E[\epsilon_i^2|\mathcal{Z}] \leq \bar{C}$*, and* $sup_i E[||u_i||^2|\mathcal{Z}] \leq \bar{C}, a.s..$

This assumption requires the second conditional moments of the disturbances to be bounded.

Assumption 4. *(i) The operator* K *is nuclear. (ii)* Υ_a *(the ath row of* Υ*) belongs to the closure of the linear span of* $\{Z(.;\nu)\}$*for* $a = 1, .., p$*. (iii) There exist a constant* $\bar{C}$ *such that* $P_{ii}^{\alpha} \leq \bar{C} < 1$*,* $i = 1, ..., n$*.*

Assumption 4 is the same as in Carrasco (2012). Condition (i) provides that the smallest eigenvalues of the covariance operator K decreases to zero sufficiently fast. Condition ii) implies that the optimal instrument f can be approached by a sequence of instruments. In finite case, this condition is equivalent to say that f_i can be approached by a linear combination of the instruments where $Z(\nu)^L$ is a subset of the instruments. So, there exists a π_L such that $\sum_{i=1}^n ||f_i - \pi_L Z(\nu)^L||^2/n \to 0$. Condition (iii) is reminiscent of Assumption 1 in CSHNW: "for some $\bar{C} < 1$, $P_{ii} < \bar{C}$, $i = 1, ..., n$". However it is much less restrictive. Indeed, $P_{ii} < \bar{C} < 1$ implies that $\sum_i \frac{P_{ii}}{n} = \frac{L}{n} < 1$, $L =$rank(Z), which restricts the number of instruments. Our condition $P_{ii}^{\alpha} \leq \bar{C} < 1$ implies that trace$(P^{\alpha}) = \sum_i q_i < n$, which implies a condition on α. Recall from Carrasco (2012) that $\sum_i q_i = O(\frac{1}{\alpha})$. So Assumption (iii) implies $\frac{1}{\alpha n} < 1$.

Assumption 5. *There exist a constant* $\bar{C}$*,* $\bar{C} > 0$ *such that* $\sum_{i=1}^n ||f_i||^4/n \to 0$*,* $sup_i E[\epsilon_i^4|\mathcal{Z}] \leq \bar{C}$*, and* $sup_i E[||u_i||^4|\mathcal{Z}] \leq \bar{C}$*.*

Assumption 5 is a standard condition which assumes that fourth moments are bounded.

Theorem 1. *Suppose that Assumptions 1-4 are satisfied. The Tikhonov, Landweber-Fridman, and the Spectral cut-off regularized Jackknife estimators satisfy* $S_n'(\hat{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0$ *as* n*,* μ_n *go to infinity,* α *goes to 0.*

Remark 1. *Here, we remark that the formulation of the JIVE from CSHNW is quite similar, we can use this similarity to show the consistency of our estimators.*

Remark 2. *Theorem 1 implies* $(\hat{\delta} - \delta_0) \xrightarrow{p} 0$.

See proof of remark 2 in Hansen and Kozbur (2014).
The following theorem will state the asymptotic normality of our proposed estimators.
First, let: $\sigma_i^2 = E[\epsilon_i^2|\mathcal{Z}]$, $H_n = \sum_i f_i f_i'/n$, $\Omega_n = \sum_i f_i f_i' \sigma_i^2/n$,
$\Psi_n = S_n^{-1} \sum_{i \neq j}^n (P_{ij}^\alpha)^2 (E[U_i U_i'|\mathcal{Z}] \sigma_j^2 (1 - P_{jj})^{-2} + E[U_i \epsilon_i|\mathcal{Z}](1 - P_{ii})^{-1}$
$E[U_j \xi_j|\mathcal{Z}](1 - P_{jj})^{-1}) S_n'^{-1}$.
Similarly to CSHNW, we give the conditional asymptotic variance of $S_n'(\hat{\delta} - \delta_0)$:

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

Theorem 2. *Suppose that assumptions 1-5 are satisfied and* $\frac{1}{\alpha \mu_n^2}$ *is bounded. Then:*

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

Remark 3. *As in CSHNW, the term* Ψ_n *in the conditional asymptotic variance of* $\hat{\delta}$ *accounts for the presence of many instruments. The order of this term is* $\frac{1}{\alpha \mu_n^2}$. *So the term* Ψ_n *vanishes asymptotically if* $\frac{1}{\alpha \mu_n^2} \to 0$, *then the conditional asymptotic variance is:* $V_n = H_n^{-1} \Omega_n H_n^{-1}$.

Homoskedastic case.

If the errors are homoskedastic, $E[\epsilon_i^2|\mathcal{Z}] = \sigma_\epsilon^2$, and if $\frac{1}{\alpha \mu_n^2} \to 0$, the asymptotic variance of the regularized Jackknife estimator is equal to $\sigma_\epsilon^2 \left[E(f_i f_i')\right]^{-1}$, which corresponds to the semiparametric efficiency bound (Chamberlain (1992)) and is smaller than that obtained in CSHNW. $\alpha \mu_n^2$ needs to go to infinity, which means that the regularization parameter α should go to zero at a slower rate than the concentration parameter μ_n^2 goes to infinity. We believe that the reason, why CSHNW obtain a larger asymptotic variance than us, is that they use the number of instruments as regularization parameter. As a result, they cannot let L grow fast enough to reach efficiency. Our estimator involves an extra tuning parameter which is selected so that the extra term Ψ_n in the variance vanishes asymptotically. Moreover, we assume that the set of instruments is sufficiently rich to span the optimal instrument (Assumption 4(ii)).

It is useful to write the RJIVE as:

$$\hat{\delta} = \hat{H}^{-1} \sum_{i,j=1}^{n} X_i C_{ij}^{\alpha} y_j, \tag{1.3.1}$$

where $\hat{H} = \sum_{i,j=1}^{n} X_i C_{ij}^{\alpha} X_j'$, and $C^{\alpha} = (C_{ij}^{\alpha}) = \begin{cases} \frac{P_{ij}^{\alpha}}{1-P_{ii}^{\alpha}} & if \quad i \neq j \\ C_{ii}^{\alpha} = 0 & if \quad i = j \end{cases}$. Then the RJIVE estimator could be written as:

$$\sqrt{n}(\hat{\delta} - \delta_0) = \frac{(X'C^{\alpha\prime}X)^{-1}}{n} \frac{(X'C^{\alpha\prime}\epsilon)}{\sqrt{n}}. \tag{1.3.2}$$

The asymptotic variance is given by:

$$(X'C^{\alpha\prime}X)^{-1} E[(X'C^{\alpha\prime}\epsilon)(\epsilon' C^{\alpha} X)](X'C^{\alpha\prime}X)^{-1}$$

Now we give an estimator of the asymptotic variance in the homoskedastic case:

$\tilde{\sigma}_{\epsilon}^2 (X'C^{\alpha}X)^{-1}(X'C^{\alpha\prime}C^{\alpha}X)(X'C^{\alpha\prime}X)^{-1}$, where $\tilde{\sigma}_{\epsilon}^2 = \frac{1}{n}\sum_i^n (y_i - X_i\tilde{\delta})^2$ and $\tilde{\delta}$ is the consistent RJIVE estimator.

1.4 Mean square error

The three regularization schemes involve a regularization parameter α. We will choose α that minimizes the mean square error (MSE). To do this, we follow the same approach as in Carrasco (2012), and Donald and Newey (2001) by analyzing the higher-order expansion of the MSE of the regularized JIVE.

In this section, we assume that we deal only with many strong instruments. Let $\Upsilon = f = (f(\nu_1), ..., f(\nu_n))'$. Let $\bar{H}$ be the $p \times p$ matrix $\bar{H} = f'f/n$, $\Sigma_u = E(u_i u_i'|\mathcal{Z})$, $\sigma_{u\epsilon} = E(\epsilon_i u_i|\mathcal{Z})$ and $X = (X_1, ..., X_n)$ and finally let $||A||$ be the Euclidean norm of a matrix A.

Assumption 6. *(i) $H = E(f_i f_i')$ exists and is non singular.*
ii) there is a $\beta \geq 1/2$ such that

$\sum_{j=1}^{\infty} \frac{<E(Z(.,\nu_i)f_a(\nu_i)),\phi_j>^2}{\lambda_j^{2\beta+1}} < \infty$, *where f_a is the a^{th} element of f for $a = 1, 2, ..., p$.*

Assumption 6(i) and 6(ii) are similar to those of Carrasco (2012). Assumption 6(ii) is used to derive the rate of convergence of the MSE. It guarantees that $||f - P^{\alpha}|| = O_P(\alpha^{\beta})$ for LF and SC and $||f - P^{\alpha}|| = O_P(\alpha^{min(2,\beta)})$ for T. The value of β measures how well the instruments approximate the reduced form. The larger β, the better is the approximation.

Assumption 7. *X_i, y_i, ν_i iid, $E[\epsilon_i^2|\mathcal{Z}] = \sigma_\epsilon^2$, and $E[\epsilon_i^4|\mathcal{Z}]$, $E[u_i^4|\mathcal{Z}]$ are bounded. We assume that the instruments are strong.*

Assumption 8. *(i) $E[(\epsilon_i, u_i)'(\epsilon_i, u_i)]$ is bounded, (ii) K is a compact operator with non zero eigenvalues, (iii) $f(\nu_i)$ is bounded.*

A sufficient condition for Assumption 8 (ii) is that the eigenvalues of the operator K are square summable $\sum_{j=1}^{\infty} \lambda_j^2 < 0$.

Theorem 3. *Suppose that Assumptions 6-8 are satisfied. For RJIVE $\hat{\delta}$, the approximate MSE for $\sqrt{n}(\hat{\delta} - \delta_0)$ is given by:*

$$S(\alpha) = H^{-1}\left[\Sigma_u^2 \sigma_\varepsilon^2 \frac{tr\left(C^{\alpha} C^{\alpha\prime}\right)}{n} + \sigma_{u\varepsilon}\sigma_{u\varepsilon}' \frac{tr(C^{\alpha 2})}{n} + \sigma_\varepsilon^2 \frac{f'\left(I - C^{\alpha\prime}\right)\left(I - C^{\alpha}\right)f}{n}\right] H^{-1}$$

Moreover, for LF, SC, $S(\alpha) = O_p(1/\alpha n + \alpha^{\beta})$ *and for T,* $S(\alpha) = O_p(1/\alpha n + \alpha^{min(\beta,2)})$.

We could compare $S(\alpha)$ with the expression of the approximate MSE given by Carrasco (2012) and Carrasco and Doukali (2016).

$$\begin{aligned}
S_{RJIVE}(\alpha) &= H^{-1}\left[\Sigma_u^2\sigma_\varepsilon^2\frac{tr\left(C^\alpha C^{\alpha\prime}\right)}{n} + \sigma_{u\varepsilon}\sigma_{u\varepsilon}'\frac{tr(C^{\alpha 2})}{n} + \sigma_\varepsilon^2\frac{f'\left(I - C^{\alpha\prime}\right)\left(I - C^\alpha\right)f}{n}\right]H^{-1} \\
S_{2SLS}(\alpha) &= H^{-1}[(\sigma_{u\epsilon}\sigma_{u\epsilon}')\frac{[tr(P^\alpha)]^2}{n} + \sigma_\epsilon^2\frac{f'(I-P^\alpha)^2 f}{n}]H^{-1}, \\
S_{LIML}(\alpha) &= \sigma_\epsilon^2 H^{-1}[\Sigma_v\frac{[tr((P^\alpha)^2)]}{n} + \frac{f'(I-P^\alpha)^2 f}{n}]H^{-1}.
\end{aligned}$$

We know that
$S_{RJIVE}(\alpha) \sim \frac{1}{n\alpha} + \alpha^\beta$,
$S_{2SLS}(\alpha) \sim \frac{1}{n\alpha^2} + \alpha^\beta$,
$S_{LIML}(\alpha) \sim \frac{1}{n\alpha} + \alpha^\beta$,
for LF, PC and if $\beta < 2$ in the Tikhonov regularization. For $\beta \geq 2$ the leading term of the Tikhonov regularization is
$S_{RJIVE}(\alpha) \sim \frac{1}{n\alpha} + \alpha^2$,
$S_{2SLS}(\alpha) \sim \frac{1}{n\alpha^2} + \alpha^2$,
$S_{LIML}(\alpha) \sim \frac{1}{n\alpha} + \alpha^2$,
The approximate MSE of regularized Jackknife is of a smaller order in α than of the regularized 2SLS and has the same order as the regularized LIML. The simulation study in Section 1.6 shows that almost everywhere regularized Jackknife performs better than regularized 2SLS and performs as well as the regularized LIML.
We note that the approximate MSE of the 2SLS and RJIVE estimators is composed of two terms, the first one corresponds to the bias term which grows when α goes to 0, and the second term corresponds to the variance term. For the LIML estimator, the leading terms in its approximate MSE come only from the variance terms.

1.5 Selection of the regularization parameter

We want to find α that minimizes the conditional MSE of $v'\delta$ for some arbitrary $p \times 1$ vector v. The conditional MSE is:
$MSE = E[v'(\hat{\delta} - \delta_0)(\hat{\delta} - \delta_0)'v|Z] = v'S(\alpha)v = S_v(\alpha)$
We will replace $S_v(\alpha)$ by an estimate obtained by cross-validation. First, we need to reduce the dimension. If $\delta \in R^p$ for $p > 1$, the regression $X = f + u$ involves $n \times p$ matrices. We can reduce the dimension by post-multiplying by $H^{-1}v$, we have:

$$X_v = f_v + u_v \tag{1.5.1}$$

where X_v, f_v, and u_v are $n \times 1$ vectors such that $X_v = XH^{-1}v$, $f_v = fH^{-1}v$ and $u_v = uH^{-1}v$. We have:

$$\begin{aligned} S(\alpha) &= H^{-1}\left[\Sigma_u^2\sigma_\varepsilon^2 \frac{tr\left(C^\alpha C^{\alpha\prime}\right)}{n} + \sigma_{u\varepsilon}\sigma_{u\varepsilon}' \frac{tr(C^{\alpha 2})}{n} + \sigma_\varepsilon^2 \frac{f'\left(I - C^{\alpha\prime}\right)\left(I - C^\alpha\right) f}{n}\right] H^{-1} \\ &= \sigma_{u_\nu}^2\sigma_\varepsilon^2 \frac{tr\left(C^\alpha C^{\alpha\prime}\right)}{n} + \sigma_{u_\nu\varepsilon}^2 \frac{tr(C^{\alpha 2})}{n} + \sigma_\varepsilon^2 \frac{f_\nu'\left(I - C^{\alpha\prime}\right)\left(I - C^\alpha\right) f_\nu}{n}. \end{aligned}$$

The term $\frac{f_\nu'(I - C^{\alpha\prime})(I - C^\alpha) f_\nu}{n}$ corresponds to the prediction error in the regression (1.5.1). It can be approximated by one of the usual cross-validation techniques (see for instance Li (1987)). Because the trace of C^α is equal to zero, C_p cross-validation, generalized cross-validation, and leave-one-out cross validation coincide. Let

$$\begin{aligned} \hat{R}(\alpha) &= \frac{1}{n}\left\|X_\nu - \hat{f}_\nu(\alpha)\right\|^2 \\ R(\alpha) &= \frac{1}{n}E\left[\left\|f_\nu - \hat{f}_\nu(\alpha)\right\|^2 |\mathcal{Z}\right] \end{aligned}$$

where $\hat{f}_\nu(\alpha) = C^\alpha X_\nu$.

First we show that $\hat{R}(\alpha)$ is a conditionally unbiased estimator of $R(\alpha)$

up to an additive constant.

$$\begin{aligned}\hat{R}(\alpha) &= \frac{1}{n}\left(f_\nu + u_\nu - \hat{f}_\nu(\alpha)\right)'\left(f_\nu + u_\nu - \hat{f}_\nu(\alpha)\right) \\ &= \frac{1}{n}\left(f_\nu - \hat{f}_\nu(\alpha)\right)'\left(f_\nu - \hat{f}_\nu(\alpha)\right) \\ &\quad + \frac{1}{n}u_\nu'\left(f_\nu - \hat{f}_\nu(\alpha)\right) + \frac{1}{n}\left(f_\nu - \hat{f}_\nu(\alpha)\right)' u_\nu \\ &\quad + \frac{1}{n}u_\nu' u_\nu.\end{aligned}$$

Then,

$$E\left(\hat{R}(\alpha)\,|\mathcal{Z}\right) = R(\alpha) + \sigma_{u_\nu}^2.$$

Moreover,

$$\begin{aligned}R(\alpha) &= \frac{1}{n}E\left[\left((1-C^{\alpha\prime})f_\nu + C^{\alpha\prime}u_\nu\right)'\left((1-C^{\alpha\prime})f_\nu + C^{\alpha\prime}u_\nu\right)|Z\right] \\ &= \frac{1}{n}f_\nu'\left(I - C^{\alpha\prime}\right)\left(I - C^{\alpha}\right)f_\nu + \frac{1}{n}\sigma_{u_\nu}^2 tr\left(C^\alpha C^{\alpha\prime}\right).\end{aligned}$$

In the expression of $S(\alpha)$, the term $\sigma_{u_\nu}^2\sigma_\varepsilon^2\frac{tr(C^\alpha C^{\alpha\prime})}{n} + \sigma_\varepsilon^2\frac{f_\nu'(I-C^{\alpha\prime})(I-C^\alpha)f_\nu}{n}$ can be replaced by $\sigma_\varepsilon^2 R(\alpha)$. So $S(\alpha)$ can be estimated[2] by

$$\begin{aligned}\hat{S}(\alpha) &= \hat{\sigma}_\varepsilon^2\hat{R}(\alpha) + \hat{\sigma}_{u_\nu\varepsilon}^2\frac{tr(C^{\alpha 2})}{n} \\ &= \hat{\sigma}_\varepsilon^2\frac{1}{n}\left\|X_\nu - C^\alpha X_\nu\right\|^2 + \hat{\sigma}_{u_\nu\varepsilon}^2\frac{tr(C^{\alpha 2})}{n} \qquad (1.5.2)\end{aligned}$$

where $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_{u_\nu\varepsilon}^2$ are consistent estimators of σ_ε^2 and $\sigma_{u_\nu\varepsilon}^2$. To estimate σ_ε^2, let $\tilde{\varepsilon} = y - X\widetilde{\delta}$ where $\widetilde{\delta}$ is a preliminary estimator (obtained for instance from a finite number of moments), then $\hat{\sigma}_\varepsilon^2 = \tilde{\varepsilon}'\tilde{\varepsilon}/n$. To estimate $\sigma_{u_\nu\varepsilon}^2$, one difficulty arises which comes from the fact that $X_\nu = XH^{-1}\nu$ is not observed. There are two ways to proceed. One ad-hoc solution is to set $H^{-1}\nu = e$ where e is some arbitrary vector chosen a priori for instance $e = (1, ..., 1)'$. Another

[2] We dropped the extra term $\sigma_{u_\nu}^2$ because it does not depend on α.

solution consists in setting ν a priori (as a vector of ones for instance) and estimating H by

$$\tilde{H} = X'C^{\tilde{\alpha}\prime}X/n$$

where $\tilde{\alpha}$ is obtained from a first stage cross-validation criterion based on a single endogenous variable (for instance the first one). Then X_ν in (1.5.2) is replaced by $\tilde{X}_\nu = X\widetilde{H}^{-1}\nu$. Let $\tilde{u}_\nu = (I - C^{\tilde{\alpha}\prime})X\tilde{H}^{-1}\nu$, then, $\hat{\sigma}_{u_\nu\varepsilon} = \tilde{u}_\nu'\tilde{\varepsilon}/n$.

So a feasible estimator of $S(\alpha)$ is given by

$$\tilde{S}(\alpha) = \hat{\sigma}_\varepsilon^2 \frac{1}{n} \left\| \tilde{X}_\nu - C^\alpha X_\nu \right\|^2 + \hat{\sigma}_{u_\nu\varepsilon}^2 \frac{tr(C^{\alpha 2})}{n}.$$

1.6 Simulation study

In this section we present a Monte Carlo study. Our goal is to demonstrate the performance of our estimators and provide a comparison with other standard estimators using a simulation study on a simple DGP.
The data generating process (DGP) is given by:

$$y_i = X_i'\delta_0 + \epsilon_i \tag{1.6.1}$$

$$X_i = f(\nu_i) + u_i \tag{1.6.2}$$

$i = 1, \ldots, n$. $\delta_0 = 0.1$ and $(\epsilon_i, u_i) \sim N(0, \sum)$ with

$$\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

For the purpose of comparison, we consider various setting for the remaining parameters of the model.

Model 1.

In this model, $f(\nu_i)$ is linear as in DN.

$$f(\nu_i) = \nu_i'\pi$$

with $\nu_i \overset{iid}{\sim} N(0, I_L)$, L= 15 and 30.
The ν_i are used as instruments so that $Z_i = \nu_i$. The instruments are independent from each other, this example corresponds to the worse case scenerio for our regularized estimators. Indeed, here all the eigenvalues of the operator K are equal to 1, so there is no information contained in the spectral decomposition of K. Moreover, if L were infinite, K would not be compact, hence our method would not apply. However, in practical applications, it is not plausible that a very large number of instruments would be uncorrelated with each other.

Model 1a.

$\pi_l = d(1 - l/(L+1))^4$, l=1,2,...L where d is chosen so that $\pi'\pi = \frac{R_f^2}{1-R_f^2}$.
Here, the instruments are ordered in decreasing order of importance. This model represents the case where there is some prior information about what instruments are important.

Model1b.

$\pi_l = \sqrt{\frac{R_f^2}{L(1-R_f^2)}}$, l=1,2,3,... L and $R_f^2 = 0.1$. In this case, as π_l is the same for all l, there is no reason to prefer one instrument over another.

Model 2 (Factor Model).

$$X_i = f_{i1} + f_{i2} + f_{i3} + u_i$$

where $f_i = (f_{i1}, f_{i3}, f_{i3})' \sim iidN(0, I_3)$, x_i is a $L \times 1$ vector of instruments constructed f_i through

$$\nu_i = Mf_i + \theta_i$$

where $\theta_i \sim N(0, \sigma_\theta^2)$ with $\sigma_\theta = 0.3$, and M is $L \times 3$ matrix which elements are independently drawn in a U [-1,1].

Model 3 (weak instruments)

Following Chao and Swanson (2005), the concentration parameter is proportional to $\pi'\mathbf{Z_n'Z_n}\pi$, where $\mathbf{Z_n}$ is the $n \times L$ matrix: $\mathbf{Z_n} = (Z_1', ..., Z_n')'$. As here $Z_i \overset{iid}{\sim} N(0, I_L)$, we have $E(\mathbf{Z_n'Z_n}) = nI_L$. Therefore, the concentration parameter can be approximated by:

$$CP = n\pi'\pi = n\sum_{l=1}^{L}\pi_l^2$$

When $\pi_l = \pi_1$ for all l, $CP = nL\pi_1^2$. In model 1b, $CP = n\frac{R_f^2}{1-R_f^2} = 55.5$
Now, we consider $CP = 70$ and a smaller value of CP, namely $CP = 35$.[3]
The simulations are performed using 1000 replications of samples of size $n = 500$. Our proposed estimators depend on a regularization parameter α that needs to be chosen. We use a data-driven method to select α based on an expansion of the MSE [4]. We compute the estimators corresponding to Carrasco and Doukali (2016) regularized two-stage least squares and limited information maximum likelihood: T2SLS and TLIML (Tikhonov), L2SLS and LLIML (Landweber-Fridman), P2SLS and PLIML (principal component). In addition, we consider for each setting the regularized Jackknife estimators Tjack (Tikhonov), Ljack (Landweber Fridman) and Pjack (principal component). The optimal regularization parameter is selected using Mallows C_p.
We report the median bias (M.bias), the median of the absolute deviations of the estimator from the true value (M.abs), the difference between the 0.1 and 0.9 quantiles (dis) of the distribution of each estimator, the mean square error (MSE) and the coverage rate (Cov.) of a nominal 95% confidence interval. To construct the confidence intervals to compute the coverage probabilities, we used the following estimate of asymptotic variance:

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

[3]Note that the values considered by Hausman et al. (2012) correspond to $n\pi_1^2 = \mu^2 = 8$ and 32, so that we consider here smaller value of π_1^2, therefore weaker instruments.

[4]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 constant c in the LF algorithm is set equal to 0.1. For the number of principal components, we search between 1 and the number of instruments.

where $\hat{X} = C^{\alpha} X$.

Table 1.1 shows results for model 1a. We remark that the regularized Jackknife is better than the regularized 2SLS and performs as well as the regularized LIML in almost every case. We observe that the coverage of the regularized 2SLS is very poor while that for the regularized Jackknife is much better. Within the regularized Jackknife, T and LF perform better than the others estimators especially when the number of instruments are very high.

In Model 1b where all the instruments have equal weights and there is no reason to prefer one over another, the regularized Jackknife dominates the regularized 2SLS in terms of bias (see Table 1.2). We can also notice that, when the number of instruments increases, the MSE of the regularized Jackknife becomes greater than those of regularized 2SLS.

The poor performance of SC 2SLS estimator and SC Jackknife estimator in model 1a can be explained by the absence of factor structure.

In model 2 which is a factor model, Table 1.3 shows that the regularized Jackknife estimator has similar performance as the regularized 2SLS for all schemes. The T and LF Jackknife dominate the P Jackknife with respect to all criteria.

Now we consider the model 3 which allows the presence of weak instruments. From Table 1.4 we remark that:

(a) The performance of the regularized estimators increases with the strength of instruments but decreases with the number of instruments.

(b) The bias of regularized Jackknife regularized estimator is quite a bit smaller than that of regularized 2SLS. However, it is larger than that of the regularized LIML especially for weak instruments but the coverage of regularized Jackknife is better than that of regularized LIML.

(c) Tikhonov Jackknife estimator has the smallest bias while LF Jackknife has the smaller MSE. The SC is not recommended in presence of weak in-

struments as both its bias and variance are large.
As expected, regularized Jackknife estimators have smaller bias but larger variance than the regularized 2SLS estimators. This loss of efficiency can be regarded as the cost of improvement of the bias relative to the regularized 2SLS estimators.

Table 1.1: Simulations results of model 1a

		T2SLS	L2SLS	P2SLS	TLIML	LLIML	PLIML	Tjack	Ljac	Pjack
								-	-	
$L = 15$	**M.bias**	0.101	0.114	0.088	0.006	0.007	0.017	0.014	0.006	0.043
	M.abs	0.114	0.100	0.174	0.103	0.102	0.102	0.120	0.109	0.192
	Disp	0.302	0.312	0.843	0.391	0.391	0.390	0.466	0.429	0.993
	MSE	0.023	0.02	12.943	0.025	0.025	4.5e+28	0.040	0.032	476.1
	Cov	0.812	0.819	0.927	0.898	0.902	0.895	0.954	0.958	0.934
								-		
$L = 30$	**M.bias**	0.172	0.167	0.180	0.007	0.009	0.046	0.004	0.003	0.141
	M.abs	0.172	0.169	0.204	0.106	0.108	0.113	0.129	0.123	0.343
	Disp	0.263	0.291	0.457	0.447	0.441	0.433	0.505	0.490	1.760
	MSE	0.039	0.039	11.038	0.032	0.032	Inf	0.087	0.064	35.23
	Cov	0.586	0.632	0.712	0.820	0.812	0.824	0.956	0.952	0.899

Table 1.2: Simulations results of model 1b

		T2SLS	L2SLS	P2SLS	TLIML	LLIML	PLIML	Tjack	Ljac	Pjack
					-	-		-	-	
$L = 15$	**M.bias**	0.098	0.096	0.071	0.000	0.001	0.015	0.011	0.009	0.015
	M.abs	0.110	0.114	0.185	0.103	0.102	0.103	0.108	0.107	0.108
	Disp	0.288	0.296	0.760	0.390	0.386	0.380	0.423	0.417	0.408
	MSE	0.022	0.023	3.319	0.024	0.025	1.4e+28	0.029	0.029	0.153
	Cov	0.838	0.843	0.925	0.897	0.899	0.891	0.924	0.925	0.946
								-	-	
$L = 30$	**M.bias**	0.176	0.165	0.129	0.010	0.011	0.042	0.002	0.004	0.0120
	M.abs	0.176	0.165	0.275	0.107	0.110	0.111	0.126	0.126	0.355
	Disp	0.258	0.276	1.500	0.412	0.421	0.413	0.504	0.485	1.610
	MSE	0.039	0.038	24.344	0.030	0.033	1.3e+29	2.510	0.047	11.17
	Cov	0.576	0.643	0.93	0.829	0.828	0.806	0.962	0.961	0.891

Table 1.3: Simulations results of model 2

		T2SLS	L2SLS	P2SLS	TLIML	LLIML	PLIML	Tjack	Ljac	Pjack
					-	-	-			
$L = 15$	**M.bias**	0.001	0.001	0.001	0.0003	0.0002	0.0002	0.0002	0.0002	0.001
	M.abs	0.017	0.017	0.017	0.0177	0.0177	0.017	0.017	0.017	0.017
	Disp	0.066	0.066	0.066	0.067	0.067	0.067	0.065	0.066	0.065
	MSE	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
	Cov	0.944	0.944	0.942	0.948	0.948	0.948	0.949	0.946	0.946
								-	-	-
$L = 30$	**M.bias**	0.0008	0.0004	0.0004	0.001	0.0009	0.0009	0.001	0.001	0.0004
	M.abs	0.016	0.016	0.016	0.017	0.017	0.0172	0.017	0.017	0.017
	Disp	0.065	0.065	0.065	0.067	0.067	0.067	0.065	0.065	0.065
	MSE	0.0007	0.0007	0.0007	0.0007	0.0007	0.0035	0.0007	0.0007	0.0007
	Cov	0.935	0.933	0.932	0.953	0.955	0.955	0.937	0.937	0.936

Table 1.4: Simulations results of model 3

			T2SLS	L2SLS	P2SLS	TLIML	LLIML	PLIML	Tjack	Ljac	Pjack
									-	-	
L=15	Cp=35	**M.bias**	0.142	0.139	0.100	0.0003	0.001	0.066	0.006	0.013	0.033
		M.abs	0.151	0.151	0.265	0.133	0.135	0.273	0.210	0.215	0.123
		Disp	0.342	0.368	1.389	0.534	0.544	1.426	0.722	0.643	0.614
		MSE	0.039	0.039	66.381	0.036	0.037	Inf	1.528	0.451	5.876
		Cov	0.768	0.792	0.93	0.858	0.868	0.934	0.927	0.937	0.915
						-			-	-	
	Cp=70	**M.bias**	0.081	0.079	0.065	0.001	0.002	0.029	0.027	0.007	0.006
		M.abs	0.097	0.098	0.143	0.092	0.090	0.153	0.114	0.096	0.090
		Disp	0.271	0.286	0.617	0.348	0.350	0.747	0.489	0.385	0.341
		MSE	0.018	0.018	3.345	0.015	0.015	1.5e+18	4.076	0.024	0.020
		Cov	0.838	0.848	0.926	0.902	0.896	0.933	0.937	0.939	0.938
									-	-	
L=30	Cp=35	**M.bias**	0.231	0.216	0.143	0.002	0.007	0.105	0.020	0.013	0.128
		M.abs	0.231	0.217	0.384	0.153	0.155	0.399	0.216	0.181	0.198
		Disp	0.293	0.323	1.999	0.611	0.627	2.031	0.950	0.822	0.753
		MSE	0.064	0.062	104.3	0.050	0.051	Inf	13.64	1.866	5.937
		Cov	0.481	0.559	0.947	0.719	0.734	0.95	0.927	0.956	0.893
									-	-	
	Cp=70	**M.bias**	0.150	0.139	0.082	0.003	0.003	0.055	0.003	0.007	0.047
		Med.	0.151	0.139	0.230	0.102	0.102	0.244	0.105	0.114	0.116
		Disp	0.249	0.262	1.085	0.391	0.398	1.239	0.417	0.442	0.448
		MSE	0.030	0.029	7.716	0.031	0.027	2.5e+26	0.031	0.040	0.364
		Cov	0.625	0.680	0.937	0.833	0.832	0.941	0.873	0.998	0.845

1.7 Empirical application: Returns to schooling

To show that our proposed regularized estimators work well, we use the classic example of Angrist and Krueger (1991). The authors estimate the causal effect of the number of years of schooling on the log of the weekly wage. Because of the problem of the endogeneity of the explanatory variable, the OLS estimate is biased. Angrist and Krueger (1991) proposed to use the quarters of birth as instruments. We use the same model and instruments and we compute different versions of the regularized Jackknife estimators and compare them with other competing IV estimators.
We consider the Angrist and Krueger (1991)'s model:

$$logw = \alpha + \delta education + \beta_1' Y + \beta_2' S + \epsilon$$

where log w=log of weekly wage, education= year of education, Y= year of birth dummy (9), S= state of birth dummy (50). The vector of instruments $Z = (1, Y, S, Q, Q * Y, Q * S)$ includes 240 variables, where Q is quarter-of-birth dummy. The parameter of interest is δ which represents the impact of education on earnings. The sample drawn from the 1980 US Census consists of 325,509 men born between 1930 and 1939.
Table 1.5 reports schooling coefficients generated by different estimators applied to the Angrist and Krueger (1991) data with their standard errors. Table 1.5 shows that all regularized 2SLS and Jackknife give close results. This results confirm what simulations have shown. Table 1.5 shows also that Tjack and Ljack are more reliable than their 2SLS counterpart and Pjack. Pjack gives estimators which are quite bigger than Tjack and Ljack but we do not trust Pjack because there is no factor structure here. The results suggest that the return to education is between 0.0901 and 0.1077.

Table 1.5: Estimates of the returns to education

OLS	**2SLS**	**TSLS**	**L2SLS**	**P2SLS**
0.068 (0.0003)	0.081 (0.010)	0.123 (0.048)	0.129 (0.030)	0.978 (0.041)
	.	$\alpha = 0.00001$	700 iterations	Nb of eignf=81
.	**Jackknife**	**Tjack**	**Ljack**	**Pjack**
. .	0.095 (0.026)	0.107 (0.030)	0.090 (0.059)	0.140 (0.074)
	.	$\alpha = 0.00001$	700 iterations	Nb of eignf=219

1.8 Conclusion

In this paper, we propose a new method for estimation with many weak instruments. We derived the limiting distribution for the regularized Jackknife estimator. This estimator is consistent and asymptotically normal. We show that, thanks to the regularization, our estimators are more efficient than the standard Jackknife IV estimator. All regularization methods involve a tuning parameter which needs to be selected. We propose a data-driven method for selecting this parameter. Simulations show that the regularized estimators (LF and T of Jackknife) perform well (are nearly median unbiased) relative to other competing estimators. These proposed estimators are attractive alternatives to existing methods for researchers working with many weak instruments.

1.9 Appendix

Lemma A0.: *If Assumptions 1-3 are satisfied Then :*

i) $P_{ii}^{\alpha} < 1$ *for* $\alpha > 0$*,*

ii) $\sum_{i\neq j} (P_{ij}^{\alpha})^2 = O(1/\alpha)$,

iii) $\sum_{i\neq j} P_{ij}^{\alpha} = O(1/\alpha)$.

iv) $\sum_{i,l,k,r} P_{ik}^{\alpha} P_{kl}^{\alpha} P_{lr}^{\alpha} P_{ri}^{\alpha} = O(1/\alpha)$.

v) $\sum_{i,j} (P_{ij}^{\alpha})^4 = O(1/\alpha)$.

vi) $\sum_{i,j,k} (P_{ij}^{\alpha})^2 (P_{j,k}^{\alpha})^2 = O(1/\alpha)$.

Proof of Lemma A0:

i) Let $P = Z(Z'Z)^{+}Z'$ a $n \times n$ matrix where $(Z'Z)^{+}$ is the generalized inverse of Z'Z.

P is a projection matrix of rank, say r, with r eigenvalues=1 and $n - r = 0$. Let $P^{\alpha}v = \sum_j q(\alpha, \lambda_j^2) < v, \psi_j > \psi_j$. The eigenvalues of P^{α} are $q(\alpha, \lambda_j^2)$ where $q(\alpha, \lambda_j^2) = 0$ if $\lambda_j = 0$ and $q(\alpha, \lambda_j^2) < 1$ if $\lambda_j \neq 0$ and $\alpha > 0$.

$(P - P^{\alpha})v = \sum_j (q(\alpha, \lambda_j^2) - 1) < v, \psi_j > \psi_j$

We know that $q(\alpha, \lambda_j^2) < 1$, thus, $P^{\alpha} - P$ is negative definite.

Therefore for all vector v: $v'(P^{\alpha} - P)v < 0$. Let $v = (0, ..., 1, ...0) = e_j$, we have:

$e_j'(P^{\alpha} - P)e_j = P_{jj}^{\alpha} - P_{jj} < 0$. Recall that the elements P_{jj} of a projection matrix are such that $0 < P_{jj} < 1$.

Conclusion: $P_{ii}^{\alpha} < 1$ for $\alpha > 0$.

ii) $\sum_{i\neq j} (P_{ij}^{\alpha})^2 \leq \sum_i^n \sum_{i\neq j}^n (P_{ij}^{\alpha})^2 \leq \sum_{i,j}^n (P_{ij}^{\alpha})^2$.

We know that $tr((P^{\alpha})^2) = \sum_{i,j} (P_{ij}^{\alpha})^2 = \sum_j^n q_j^2 \leq \sum_j^n q_j = tr(P^{\alpha}) = O(1/\alpha)$ because $0 \leq q_j \leq 1$. For the last equality see Lemma 4 of Carrasco 2012.

iii) $\sum_{i\neq j} P_{ij}^{\alpha} \leq \sum_{i,j} P_{ij}^{\alpha} \leq tr(P^{\alpha}) = \sum_j q_j = O(1/\alpha)$.

iv) Let $P^{\alpha} = (P_{ij}^{\alpha})_{i,j}$ for $i = 1...$ and $j = 1...$

$(P_{ij}^{\alpha})^2 = (P^{\alpha} \times P^{\alpha})_{ij} = \sum_k P_{ik}^{\alpha} P_{kj}^{\alpha} = a_{ij}$ for $i = 1...$ and $j = 1...$

$(P_{ij}^{\alpha})^4 = \sum_l a_{il} a_{lj} = \sum_l [\sum_k P_{ik}^{\alpha} P_{kl}^{\alpha}][\sum_r P_{lr}^{\alpha} P_{rj}^{\alpha}] = \sum_{l,k,r} P_{ik}^{\alpha} P_{kl}^{\alpha} P_{lr}^{\alpha} P_{rj}^{\alpha}$

Then $tr((P^\alpha)^4) = \sum_{i,l,k,r} P^\alpha_{ik} P^\alpha_{kl} P^\alpha_{lr} P^\alpha_{ri}$.
We know that $tr((P^\alpha)^4) = \sum_j q(\alpha, \lambda_j^2)^4 \leq \sum_j q(\alpha, \lambda_j^2) = tr((P^\alpha)) = O(1/\alpha)$, because
$0 \leq q(\alpha, \lambda_j^2) \leq 1$. Then $tr((P^\alpha)^4) = \sum_{i,l,k,r} P^\alpha_{ik} P^\alpha_{kl} P^\alpha_{lr} P^\alpha_{ri} = O(1/\alpha)$.
v) $\sum_{i,j} (P^\alpha_{ij})^4 \leq \sum_{i,l,k,r} P^\alpha_{ik} P^\alpha_{kl} P^\alpha_{lr} P^\alpha_{ri} = O(1/\alpha)$.
vi) $\sum_{i,j,k} (P^\alpha_{ij})^2 (P^\alpha_{j,k})^2 \leq \sum_{i,l,k,r} P^\alpha_{ik} P^\alpha_{kl} P^\alpha_{lr} P^\alpha_{ri} = tr((P^\alpha)^4) = O(1/\alpha)$.
Let us define some notations that will be used in the following Lemmas. For random variables[5] W_i, Y_i, η_i and $\mathcal{Z} = (Y, Z)$. Let $\bar{w}_i = E[W_i|\mathcal{Z}_i]$, $\bar{y}_i = E[Y_i|\mathcal{Z}_i]$, $\bar{\eta}_i = E[\eta_i|\mathcal{Z}_i]$, $\tilde{W}_i = W_i - \bar{w}_i$ and $\tilde{Y}_i = Y_i - \bar{y}_i$, $\tilde{\eta}_i = \eta_i - \bar{\eta}_i$, $\bar{w}_n = E[(W_1, ..., W_n)'|\mathcal{Z}]$, $\bar{y}_n = E[(Y_1, ..., Y_n)'|\mathcal{Z}]$, $\bar{\mu}_W = max_{i\leq n}|\bar{w}_i|$, $\bar{\mu}_Y = max_{i\leq n}|\bar{y}_i|$, $\bar{\mu}_\eta = max_{i\leq n}|\bar{\eta}_i|$, $\bar{\sigma}^2_{W_n} = max_{i\leq n} var(W_i|\mathcal{Z}_i)^{1/2}$, $\bar{\sigma}^2_{Y_n} = max_{i\leq n} var(Y_i|\mathcal{Z}_i)^{1/2}$.
Define the norm: $||W||^2_{L_2,\mathcal{Z}} = \sqrt{E[W^2|\mathcal{Z}]}$, and let M, CS, T denote the Markov inequality, the Cauchy-Schwarz inequality, and the triangle inequality, respectively.

Lemma A1.

If conditional on $\mathcal{Z}$, the pairs of scalar random variables (W_i, Y_i) are independent across i, P^α is the regularized projection matrix. Then there is a constant C such that:
$|| \sum_{i\neq j}^n P^\alpha_{ij} W_i Y_j - \sum_{i\neq j}^n P^\alpha_{ij} \bar{w}_i \bar{y}_j ||^2_{L_2,\mathcal{Z}} < CB_n$
where $B_n = (1/\alpha)\bar{\sigma}^2_{W_n}\bar{\sigma}^2_{Y_n} + \bar{\sigma}^2_{Y_n}\bar{w}_n'\bar{w}_n + \bar{\sigma}^2_{W_n}\bar{y}_n'\bar{y}_n$ and $\bar{w}_n$ is defined as $\bar{w}_n = E[(W_1, ..., W_n)'|\mathcal{Z}]$, $\bar{y}_n = E[(Y_1, ..., Y_n)'|\mathcal{Z}]$, $\bar{\sigma}^2_{W_n} = max_{i\leq n} var(W_i|\mathcal{Z}_i)^{1/2}$, $\bar{\sigma}^2_{Y_n} = max_{i\leq n} var(Y_i|\mathcal{Z}_i)^{1/2}$.

Proof of Lemma A1.

$\tilde{W}_i = W_i - \bar{w}_i$ and $\tilde{Y}_i = Y_i - \bar{y}_i$. We have:
$\sum_{i\neq j}^n P^\alpha_{ij} W_i Y_j - \sum_{i\neq j}^n P^\alpha_{ij} \bar{w}_i \bar{y}_j = \sum_{i\neq j}^n P^\alpha_{ij} \tilde{W}_i \tilde{Y}_j + \sum_{i\neq j}^n P^\alpha_{ij} \tilde{W}_i \bar{y}_j + \sum_{i\neq j}^n P^\alpha_{ij} \bar{w}_i \tilde{Y}_j$.
Let $D_{1n} = \bar{\sigma}^2_{W_n}\bar{\sigma}^2_{Y_n}$. For $i \neq j$ and $k \neq l$, $E[\tilde{W}_i \tilde{Y}_j \tilde{W}_k \tilde{Y}_l|\mathcal{Z}]$ is zero unless $i = k$

[5] Note that here W_i and η_i are arbitrary scalar variables that will take various forms in the sequel.

and $j = l$ or $i = l$ and $j = k$. Then by Cauchy-schwarz inequality and Lemma 1, we have:

$$\begin{aligned}
E[(\sum_{i\neq j}^{n} P_{ij}^{\alpha}\tilde{W}_i\tilde{Y}_j)^2|\mathcal{Z}] &= \sum_{i\neq j}^{n}\sum_{k\neq l}^{n} P_{ij}^{\alpha}P_{kl}^{\alpha}E[\tilde{W}_i\tilde{Y}_j\tilde{W}_k\tilde{Y}_l|\mathcal{Z}] \\
&= \sum_{i\neq j}^{n}(P_{ij}^{\alpha})^2(E[\tilde{W}_i^2|\mathcal{Z}]E[\tilde{Y}_j^2|\mathcal{Z}] + E[\tilde{W}_i\tilde{Y}_i|\mathcal{Z}]E[\tilde{W}_j\tilde{Y}_j|\mathcal{Z}]) \\
&\leq 2D_{1n}\sum_{i\neq j}^{n}(P_{ij}^{\alpha})^2 \\
&\leq 2D_{1n}\sum_{i}^{n}P_{ii}^{\alpha} \\
&\leq 2D_{1n}(1/\alpha).
\end{aligned}$$

For $\tilde{W} = (\tilde{W}_1, ..., \tilde{W}_n)'$, we have $\sum_{i\neq j} P_{ij}^{\alpha}\tilde{W}_i\bar{y}_j = \tilde{W}P\bar{y} - \sum_i P_{ii}^{\alpha}\bar{y}_i\tilde{W}_i$. By independence across i conditional on Z, we have $E[\tilde{W}_i\tilde{W}_i|\mathcal{Z}] \leq \bar{\sigma}_{W_n}^2 I_n$, so:
$E[((\bar{y}P^{\alpha}\tilde{W})^2|\mathcal{Z}] = \bar{y}'P^{\alpha}E[\tilde{W}_i\tilde{W}_i|\mathcal{Z}]P^{\alpha}\bar{y} \leq \bar{\sigma}_{W_n}^2\bar{y}'P^{\alpha}\bar{y} \leq \bar{\sigma}_{W_n}^2\bar{y}'\bar{y}$,
$E[(\sum_i P_{ii}^{\alpha}\bar{y}_i\tilde{W}_i)^2|\mathcal{Z}] = \sum_i (P_{ii}^{\alpha})^2 E[\tilde{W}_i^2|\mathcal{Z}] \leq \bar{\sigma}_{W_n}^2\bar{y}'\bar{y}$.
Then by triangular inequality we have:
$||\sum_{i\neq j}^{n} P_{ij}^{\alpha}\tilde{W}_i\bar{y}_j|| \leq ||\bar{y}P^{\alpha}\tilde{W}|| + ||\sum_i^n P_{ii}^{\alpha}\tilde{W}_i\bar{y}_i|| \leq C\bar{\sigma}_{W_n}^2\bar{y}'\bar{y}$.
Interchanging the roles of Y_i and W_i we have $\sum_{i\neq j}^{n} P_{ij}^{\alpha}\bar{w}_i\tilde{Y}_j \leq C\sigma_{Y_n}^2\bar{w}'\bar{w}$.

Lemma A2. *(adaptation of Lemma A2 of CSHNW)*
Suppose the following hold conditional on $\mathcal{Z}$:
(i) $P^{\alpha}v = \sum_j q(\alpha, \lambda_j^2) < v, \psi_j > \psi_j$.
(ii)$(W_{1n}, U_1, \epsilon_1), ..., (W_{nn}, U_n, \epsilon_n)$ *are independent, and* $D_{1,n} := \sum_{i=1}^{n} E[W_{in}W_{in}'|\mathcal{Z}]$ *satisfies* $||D_{1,n}|| < C$
(iii) $E[W_{in}'|\mathcal{Z}] = 0$, $E[U_i|\mathcal{Z}] = 0$, $E[\epsilon_i|\mathcal{Z}] = 0$, *and there is a constant* C *such that* $E[||U_i||^4|\mathcal{Z}] \leq C$ *and* $E[\epsilon_i^4|\mathcal{Z}] \leq C$
(iv) $\sum_{i=1}^{n} E[||W_{in}||^4|\mathcal{Z}] \to 0$ *a.s.*
(v) $\alpha \to 0$ *as* $n \to \propto$.

Then for:

$D_{2,n} := \alpha \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 (E[U_i U_i'|\mathcal{Z}]E[\epsilon_j^2|\mathcal{Z}] + E[U_i\epsilon_i|\mathcal{Z}]E[U_j'\epsilon_j|\mathcal{Z}])$

and any sequences c_{1n} *and* c_{2n} *depending on* $\mathcal{Z}$ *with* $||c_{1n}|| \leq C$, $||c_{2n}|| \leq C$, *and* $\sum_n{}^{-1/2} = c_{1n}' D_{1,n} c_{1n} + c_{2n}' D_{2,n} c_{2n} > 1/C$, *it follows that:*

$\bar{Y}_n = \sum_n{}^{-1/2} \sqrt{\alpha}(c_{1n}' \sum_{i=1}^{n} W_{i,n} + c_{2n}' \sum_{i\neq j}^{n} U_i (P_{ij}^{\alpha})^2 \epsilon_j) \to N(0,1).$

Proof Lemma A2.

The proof is similar to that of Lemma A2 CHSNW replacing P by the regularized P^{α}, and using Lemma A1, and Lemma B1, Lemma B2, Lemma B3, and Lemma B4 of CHSNW and replacing the number of instruments L by $1/\alpha$.

Lemma A3. *(adaptation of Lemma A3 of CSHNW)*

If conditional on $\mathcal{Z}$ *and* $(W_i, Y_i) i = 1, ..., n$ *are independent scalars, then there is* $C > 0$ *such that:*

$|| \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 W_i Y_j - E[\sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 W_i Y_j]||_{L_2,Z}^2 < CB_n'$

where $B_n' = (1/\alpha)(\bar{\sigma}_{W_n}^2 \bar{\sigma}_{Y_n}^2 + \bar{\sigma}_{W_n}^2 \bar{\mu}_Y^2 + \bar{\mu}_W^2 \bar{\sigma}_{Y_n}^2)$.

Proof of Lemma A3.

We have:

$\sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 W_i Y_j - \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 \bar{w}_i \bar{y}_j = \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 \tilde{W}_i \tilde{Y}_j + \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 \tilde{W}_i \bar{y}_j + \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 \bar{w}_i \tilde{Y}_j.$

For $i \neq j$ and $k \neq l$, $E[\tilde{W}_i \tilde{Y}_j \tilde{W}_k \tilde{Y}_l|\mathcal{Z}]$ is zero unless $i = k$ and $j = l$ or $i = l$ and $j = k$. Also $|P_{ij}^{\alpha}| \leq P_{ii}^{\alpha} < 1$ by Lemma A0 (i). Also, we have

$(P^\alpha_{ij})^4 \leq (P^\alpha_{ij})^2$ and $\sum_j (P^\alpha_{ij})^2 = O(\frac{1}{\alpha})$ by Lemma A0, so:

$$\begin{aligned}
E[(\sum_{i\neq j}^{n} (P^\alpha_{ij})^2 \tilde{W}_i \tilde{Y}_j)^2 |\mathcal{Z}] &= \sum_{i\neq j}^{n} \sum_{k\neq l}^{n} (P^\alpha_{ij})^2 (P^\alpha_{kl})^2 E[\tilde{W}_i \tilde{Y}_j \tilde{W}_k \tilde{Y}_l |\mathcal{Z}] \\
&= \sum_{i\neq j}^{n} (P^\alpha_{ij})^4 (E[\tilde{W}_i^2|\mathcal{Z}] E[\tilde{Y}_j^2|\mathcal{Z}] + E[\tilde{W}_i \tilde{Y}_i|\mathcal{Z}] E[\tilde{W}_j \tilde{Y}_j|\mathcal{Z}]) \\
&\leq 2\bar{\sigma}^2_{W_n} \bar{\sigma}^2_{Y_n} \sum_{i\neq j}^{n} (P^\alpha_{ij})^4 \\
&\leq 2\bar{\sigma}^2_{W_n} \bar{\sigma}^2_{Y_n} (1/\alpha).
\end{aligned}$$

We have $\sum_{i\neq j}^n (P^\alpha_{ij})^2 \tilde{W}_i \bar{y}_j = \tilde{W} \tilde{P} \bar{y} - \sum_i (P^\alpha_{ii})^2 \bar{y}_i \tilde{W}_i$ where $\tilde{P}^\alpha_{ij} = P^{\alpha 2}_{ij}$. By independence across i conditional on $\mathcal{Z}$, we have $E[\tilde{W}_i \tilde{W}_i|Z] \leq \bar{\sigma}^2_{W_n} I_n$, so:

$$\begin{aligned}
E[((\bar{y}\tilde{P}^\alpha \tilde{W})^2|\mathcal{Z}] &= \bar{y}' \tilde{P}^\alpha E[\tilde{W}_i \tilde{W}_i|Z] \tilde{P}^\alpha \bar{y} \leq \sigma^2_{W_n} \bar{y}' (\tilde{P}^\alpha)^2 \bar{y} \\
&= \sigma^2_{W_n} \sum_{i,j,k} \bar{y}_i (P^\alpha_{ik})^2 (P^\alpha_{kj})^2 \bar{y}_j \leq \sigma^2_{W_n} \bar{\mu}^2_Y \sum_{i,j,k} (P^\alpha_{ik})^2 (P^\alpha_{kj})^2 \\
&= \sigma^2_{W_n} \bar{\mu}^2_Y \sum_k (\sum_i (P^\alpha_{ik})^2)(\sum_j (P^\alpha_{kj})^2) \\
&= \sigma^2_{W_n} \bar{\mu}^2_Y \sum_k (P^\alpha_{kk})^2 \\
&= \sigma^2_{W_n} \bar{\mu}^2_Y (1/\alpha)
\end{aligned}$$

$E[(\sum_i (P^\alpha_{ii})^2 \bar{y}_i \tilde{W}_i)^2|\mathcal{Z}] = \sum_i (P^\alpha_{ii})^4 E[\tilde{W}_i^2|\mathcal{Z}] \bar{y}_i^2 \leq \sigma^2_{W_n} \bar{\mu}^2_Y (1/\alpha)$.
Then by the triangle inequality, we have:
$|| \sum_{i\neq j}^n (P^\alpha_{ij})^2 \tilde{W}_i \bar{y}_j ||^2 \leq ||\bar{y} \tilde{P}^\alpha \tilde{W}||^2 + || \sum_i^n (P^\alpha_{ii})^2 \tilde{W}_i \bar{y}_i ||^2 \leq C(1/\alpha) \sigma^2_{W_n} \bar{\mu}^2_Y$.
Interchanging the roles of Y_i and W_i we have:
$|| \sum_{i\neq j}^n (P^\alpha_{ij})^2 \bar{w}_i \tilde{Y}_j ||^2 \leq C(1/\alpha) \bar{\sigma}^2_{Y_n} \bar{\mu}^2_W$.
Lemma A4. *(adaptation of Lemma A4 of CSHNW)*
Suppose that there is a constant $C > 0$ *such that, conditional on* $\mathcal{Z}$, (W_1, Y_1, η_1), ..., (W_n, Y_n, η_n) *are independent with* $E[W_i|\mathcal{Z}] = a_i/\sqrt{n}$, $E[Y_i|\mathcal{Z}] = b_i/\sqrt{n}$,

$|a_i| \leq C$, $|b_i| \leq C$, $E[\eta_i^2|\mathcal{Z}] \leq C$, $Var(W_i|\mathcal{Z}) \leq C/\mu_n^2$, and $Var(Y_i|\mathcal{Z}) \leq C/\mu_n^2$, and there exists π_n such that $max_{i\leq n}|a_i - Z_i'\pi_n| \to 0$ a.s. and $\frac{1}{\alpha\mu_n^2} \to 0$. Then

$A_n = E[\sum_{i\neq j\neq k} W_i P_{ik}^{\alpha}\eta_k P_{kj}^{\alpha} Y_j|\mathcal{Z}] = O_p(1)$, $\sum_{i\neq j\neq k} W_i P_{ik}^{\alpha}\eta_k P_{kj}^{\alpha} Y_j - A_n \xrightarrow{P} 0$.

Proof of Lemma A4.

Lemma A4 of CSHNW holds with P^{α} replacing P.

Lemma A5.

If Assumptions 1-3 are satisfied, then

i) $S_n^{-1}\hat{H}S_n^{-1} = \sum_{i\neq j}^n f_i P_{ij}^{\alpha}(1 - P_{jj}^{\alpha})^{-1} f_j'/n + o_p(1)$

ii) $S_n^{-1}\sum_{i\neq j}^n X_i P_{ij}^{\alpha}(1 - P_{jj}^{\alpha})^{-1}\epsilon_j' = O_p(1 + (1/(\sqrt{\alpha}\mu_n)))$.

Proof of Lemma A5 .

Let e_k denote the kth unit vector and apply Lemma A1 with $Y_i = e_k' S_n^{-1} X_i = f_{ik}/\sqrt{n} + e_k' S_n^{-1} U_i$ and $W_i = e_l' S_n^{-1} X_i (1 - P_{ii}^{\alpha})^{-1}$ for some k and l. By assumption 2, $\lambda_{min} \geq C/\mu_n$, implying $||S_n^{-1}|| \leq C/\mu_n$.

$E[Y_i|\mathcal{Z}] = f_{ik}/\sqrt{n}$, $Var([Y_i|\mathcal{Z}]) \leq C/\mu_n^2$,

$E[W_i|\mathcal{Z}] = f_{il}/\sqrt{n}(1 - P_{ii}^{\alpha})$, $Var([W_i|\mathcal{Z}]) \leq C/\mu_n^2$,

We have a.s.

$\frac{1}{\sqrt{\alpha}}\bar{\sigma}_{W_n}\bar{\sigma}_{Y_n} \leq C/(\sqrt{\alpha}\mu_n^2) \to 0$, $\bar{\sigma}_{W_n}\sqrt{\bar{y}\bar{y}'} \leq C/\mu_n\sqrt{\sum_i f_{ik}^2/n} \to 0$.

$\bar{\sigma}_{Y_n}\sqrt{\bar{w}\bar{w}'} \leq C/\mu_n\sqrt{\sum_i f_{il}^2(1 - P_{ii}^{\alpha})^{-2}/n} \leq C/\mu_n(1 - max_i P_{ii}^{\alpha})^{-2}\sqrt{\sum_i f_{il}^2/n} \to 0$.

We have $e_k' S_n^{-1}\hat{H}S_n'^{-1} e_l = e_k' S_n^{-1}\sum_{i\neq j}^n X_i P_{ij}^{\alpha} X_j S_n'^{-1} e_l/(1 - P_{jj}^{\alpha}) = \sum_{i\neq j}^n Y_i P_{ij}^{\alpha} W_j$ and $P_{ij}^{\alpha}\bar{w}_i\bar{y}_j = P_{ij}^{\alpha} f_{ik} f_{jl}/n(1 - P_{jj}^{\alpha})$, applying Lemma A1 and the conditional version of Markov, we have that for any $v > 0$ and

$A_n = \{|e_k' S_n'^{-1}\hat{H}S_n'^{-1} e_l - \sum_{i\neq j}^n e_k' f_i P_{ij}^{\alpha}(1 - P_{ii}^{\alpha})^{-1} f_j' e_l/n| \geq v\}$, $P(A_n|\mathcal{Z}) \to 0$.

By the dominated convergence theorem, $P(A_n) = E[P(A_n|\mathcal{Z})]$. The preceding argument establishes the first conclusion for the (k,l)th element. Doing this for every element completes the proof of (i).

For (ii), we apply Lemma A1 with $Y_i = e_k' S_n'^{-1} X_i$ as before and $W_i = \epsilon_i/(1 - P_{ii}^{\alpha})$. Note that $\bar{w}_i = 0$ and $\bar{\sigma}_{W_n} \leq C$. Then by Lemma A1,

$E[\{e_k' S_n'^{-1} \sum_{i\neq j}^n X_i P_{ij}^\alpha (1-P_{ii}^\alpha)^{-1}\epsilon_j\}^2|\mathcal{Z}] \leq c/(\alpha\mu_n^2) + C$.

The conclusion then follows from the fact that $E[A_n|\mathcal{Z}] \leq C$ implies $A_n = O_p(1)$.

Lemma A6.

If assumptions 1-4 are satisfied, then :

$S_n^{-1}\hat{H}S_n^{-1} = \sum_i^n f_i f_i'/n + o_p(1)$.

Proof of Lemma A6.

Lemma A6 of CSHNW holds with P^α replacing P.

Proof of theorem 1.

First we note that $S_n'(\hat{\delta}-\delta_0)/\mu_n \to 0$ implies $\hat{\delta} \xrightarrow{\text{P}} \delta_0$. Because

$||S_n'(\hat{\delta}-\delta_0)/\mu_n| \geq \sqrt{\lambda_{min}(S_n S_n'/\mu_n)}||\hat{\delta}-\delta_0|| \geq C||\hat{\delta}-\delta_0||$ by Assumption 2.

Therefore, it suffices to prove the statement $S_n'(\hat{\delta}-\delta_0)/\mu_n \to 0$.

The steps of the proof of theorem 1 of CSHNW holds with P_{ii}^α replacing P_{ii}.

So we have :

$S_n'(\hat{\delta}-\delta_0)/\mu_n = (S_n^{-1}\hat{H}S_n^{-1})^{-1}S_n^{-1}\sum_{i\neq j}^n X_i P_{ij}^\alpha(1-P_{jj}^\alpha)^{-1}\epsilon_j'/\sqrt{n} = O_p(1)o_p(1) \to 0$ (by lemma A5 and lemma A6)

Proof of theorem 2.

The proof is similar to that of CHSNW replacing P by the regularized P^α, and using Lemma A0 and Lemma A2. Define: $\xi_i = \frac{\epsilon_i}{1-P_{ii}}$ and $Y_n = \sum_i^n f_i(1-P_{ii}^\alpha)\xi_i/\sqrt{n} + S_n^{-1}\sum_{i\neq j}^n U_i P_{ij}^\alpha \xi_j$

We now apply Lemma A2 to establish the asymptotic normality of Y_n conditional on $\mathcal{Z}$, and we conclude after that the asymptotic normality of

$S_n^{-1}\sum_{i\neq j}^n W_i P_{ij}^\alpha \xi_j$.

Let $\Gamma_n = Var(Y_n|\mathcal{Z})$, so:

$\Gamma_n = \sum_i f_i f_i'(1-P_{ii}^\alpha)^2 E[{\xi_i}^2|\mathcal{Z}]/n + S_n^{-1}\sum_{i\neq j}^n (P_{ij}^\alpha)^2 (E[U_i U_i'|\mathcal{Z}]E[\xi_j^2|\mathcal{Z}] + E[U_i\xi_i|\mathcal{Z}]E[U_j'\xi_j|Z])S_n^{-1}$,

Note that $||\Gamma_n|| \leq C$ see Hansen and Kozbur (2014).

Now, let β be a $p\times 1$ nonzero vector and $W_i = f_i(1-P_{ii}^\alpha)\xi_i/\sqrt{n}$, $c_{1n} = {\Gamma_n}^{-1/2}\beta$ and $c_{2n} = \sqrt{1/\alpha}S_n^{-1}{\Gamma_n}^{-1/2}\beta$.

We show that all conditions of Lemma A2 are satisfied:

(i) Condition (i) is satisfied.

(ii) $D_{1,n} = \sum_i E[W_i W_i'|\mathcal{Z}] = \sum_i E[(f_i(1-P_{ii}^{\alpha})\xi_i/\sqrt{n})(f_i(1-P_{ii}^{\alpha})\xi_i/\sqrt{n})'|\mathcal{Z}]$.
$= E[||\sum_i^n f_i f_i' \epsilon_i^2/\sqrt{n}||^2|\mathcal{Z}] = \sum_i^n ||f_i f_i'||^2 E[\xi_i^2|\mathcal{Z}] \leq C\sum_i^n ||f_i f_i'||^2/n \xrightarrow{\text{a.s}} 0$.

Then, condition (ii) is satisfied.

(iii) Condition (iii) is satisfied by Assumptions 3 and 5.

(iv) Condition (iv) is satisfied. It is the same condition as in Theorem 2 of CSHNW.

(v) Condition (v) is satisfied by Assumption 2.

Now, Let us apply Lemma A2. Note that c_{1n} and c_{2n} are bounded. Moreover, the Σ_n of lemma A2 is:

$\Sigma_n = Var(c_{1n}'\sum_i W_{in} + c_{2n}\sum_{i\neq j}^n U_i P_{ij}^{\alpha}\xi_j)\sqrt{\alpha}|\mathcal{Z}) = Var(\beta'\Gamma_n^{-1/2'} Y_n|\mathcal{Z}) = \beta'\beta$.

We have:

$(\beta'\beta)^{-1/2}\beta'\Gamma_n^{-1/2}Y_n = \Sigma_n^{-1/2}(c_{1n}'\sum_i W_{in} + c_{2n}\sum_{i\neq j}^n U_i P_{ij}^{\alpha}\xi_j)\sqrt{\alpha} \to N(0,1)$

It follows that $\beta'\Gamma_n^{-1/2}Y_n \to N(0,\beta'\beta)$, so by the Cramer-Wold device, $\Gamma_n^{-1/2}Y_n \to N(0, I_p)$.

Recall that $V_n = H_n^{-1}\Gamma_n H_n^{-1}$ for $H_n = \sum_i f_i f_i'$. Let $B_n = V_n^{-1/2} H_n \Gamma_n^{1/2}$. B_n is an orthogonal matrix since $B_n B_n' = V_n^{-1/2} H_n \Gamma_n^{1/2}\Gamma_n^{1/2'} H_n V_n^{-1/2'} = I_n$, note also that B_n depends only on $\mathcal{Z}$.

Therefore,

$V_n^{-1/2}(S_n^{-1}\hat{H}S_n^{-1'})^{-1}\Gamma_n^{-1/2} = V_n^{-1/2}(H_n + o_p(1))\Gamma_n^{1/2} = B_n + o_p(1)$.

Note that because $\Gamma_n^{-1/2}Y_n \to N(0, I_p)$ and B_n is only a function of Z, we have that:

$B_n\Gamma_n^{-1/2}Y_n \to N(0, I_p)$. then by the slutsky lemma and $\hat{\delta} = \delta_0 + \hat{H}^{-1}\sum_{i\neq j}^n X_i P_{ij}^{\alpha}\xi_j$.

We have:

$V_n^{-1/2}S_n'(\hat{\delta} - \delta_0) = V_n^{-1/2}(S_n^{-1}\hat{H}S_n^{-1'})^{-1}S_n^{-1}\sum_{i\neq j}^n X_i P_{ij}^{\alpha}\xi_j$.

$V_n^{-1/2}S_n'(\hat{\delta} - \delta_0) = V_n^{-1/2}(S_n^{-1}\hat{H}S_n^{-1'})^{-1}(Y_n + o_p(1))$.

$V_n^{-1/2}S_n'(\hat{\delta} - \delta_0) = (B_n + o_p(1))\ (\Gamma_n^{-1/2}Y_n + o_p(1))$.

$V_n^{-1/2}S_n'(\hat{\delta} - \delta_0) = B_n\Gamma_n^{-1/2}Y_n + o_p(1) \to N(0, I_p)$.

Proof of theorem 3.

P is the projection matrix after regularization and C is the matrix defined in section 1.3, (α is hidden for simplicity). The proof is similar to Donald and Newey (2001) (DN). We have: $\sqrt{n}(\hat{\delta}-\delta_0)=\hat{H}^{-1}\hat{h}$
with $\hat{H}=\frac{W'C'W}{n}$, and $\hat{h}=\frac{W'C'\epsilon}{\sqrt{n}}$. First we state a preliminary lemma.

Lemma A7.

Let denote $e_f(\alpha)=f'(I-P)f/n, e_{2f}(\alpha)=f'(I-P)^2f/n, \Delta_\alpha=tr(e_{2f}(\alpha))$
If Assumptions 1-3 are satisfied then:

(i) $tr(f'(I-P)f/n)=\begin{cases} o_p(\alpha^\beta) & for \quad LF,SC, \\ o_p(\alpha^{min(\beta,1)}) & for \quad T. \end{cases}$

$\Delta_\alpha=\begin{cases} O_p(\alpha^\beta) & for \quad LF,SC, \\ O_p(\alpha^{min(\beta,2)}) & for \quad T. \end{cases}$

(ii) $f'(I-C)\epsilon/\sqrt{n}=O_p(\Delta_\alpha^{1/2})$.

(iii) $u'C\epsilon=O_p(1/(\alpha))$.

(iv) $E[u'C\epsilon\epsilon'C'u|\mathcal{Z}]=(\sigma_u^2\sigma_\epsilon^2+\sigma_{u\epsilon}^2)O(1/\alpha)$.

(v) $E[f'\epsilon\epsilon'Cu|\mathcal{Z}]=O_p(1/\alpha)$.

(vi) $\Delta_\alpha^{1/2}/(\sqrt{\alpha n})\leq 1/(2\alpha n)+\Delta_\alpha/2$.

(vii) $E[hh'H^{-1}u'f/n|\mathcal{Z}]=O_p(1/n)$.

(viii) $E[\frac{1}{n}(f'(I-C)\epsilon\epsilon'Cu)|\mathcal{Z}]=O_p(\Delta_\alpha^{1/2}/\sqrt{\alpha n})$.

Proof of Lemma A7.

(i) For the proof of (i) see Lemma5 (i) of Carrasco (2012).

(ii) $f'(I-C)\epsilon/\sqrt{n}=f'(I-P)\epsilon/\sqrt{n}+f'\bar{P}(I-P)\epsilon/\sqrt{n}$ because $C=P-\bar{P}(I-P)$ and $\bar{P}=\tilde{P}(I-\tilde{P})^{-1}$, where $\tilde{P}$ is a diagonal matrix with element P_{ii}, $i=1,...,n$ on the diagonal.

The first term $f'(I-P)\epsilon/\sqrt{n}=O_p(\Delta_\alpha^{1/2})$ follows from Lemma 5(ii) in Carrasco (2012)

$||f'\bar{P}(I-P)\epsilon/\sqrt{n}||^2=\epsilon'(I-P)\bar{P}ff'\bar{P}(I-P)\epsilon/n$
$=tr((I-P)\bar{P}ff'\bar{P}(I-P))\epsilon\epsilon'/n.$
$E[||f'\bar{P}(I-P)\epsilon/\sqrt{n}||^2|\mathcal{Z}]=\sigma_\epsilon^2 tr((I-P)\bar{P}ff'\bar{P}(I-P)).$

$|f'\bar{P}(I-P)(I-P)\bar{P}f| \leq sup(\frac{P_{ii}}{1-P_{ii}})^2(f'(I-P)^2f/n)^{1/2} \leq c\Delta_\alpha$. By Assumption 4 (iii).

By Markov inequality , $||f'(I-C)\epsilon/\sqrt{n}||^2 = (1-c)O_p(\Delta_\alpha) = O_p(\Delta_\alpha)$, and $f'(I-C)\epsilon/\sqrt{n} = O_p(\Delta_\alpha^{1/2})$.

(iii) $u'C\epsilon = O_p(1/(\alpha))$. It follows from (iv) by Markov inequality.

(iv) We have: $E[u'C\epsilon\epsilon'C'u] = E[(\sum_{i,j} u_i C_{ij}\epsilon_j)(\sum_{i,j} u_k C_{k,l}\epsilon_l)|\mathcal{Z}]$.

$E[u'C\epsilon\epsilon'C'u] = \sigma_u^2\sigma_\epsilon^2 \sum_{i\neq j} C_{ij}^2 + \sigma_{u\epsilon}^2 \sum_{i\neq j} C_{ij}C_{ji}$ (because $C_{ii} = 0$).

We have: $Tr(CC\prime) = \sum_{i\neq j} C_{ij}^2$, with $C = P - \bar{P}(I-P)$, and

$\bar{P} = Diag((\frac{P_{11}}{1-P_{11}}), ..., (\frac{P_{nn}}{1-P_{nn}}))$.

$CC' = (P - \bar{P}(I-P))(P - \bar{P}(I-P))'$

$CC' = P^2 - P(I-P)\bar{P}' - \bar{P}(I-P)P + \bar{P}(I-P)^2\bar{P}'$.

We know that $Tr(P(I-P)\bar{P}') = Tr(\bar{P}(I-P)P) = O(\frac{1}{\alpha})$. $Tr(P^2) = O(\frac{1}{\alpha})$, by Lemma A0.

$Tr(\bar{P}(I-P)^2\bar{P}') = \sum_i \bar{P}_{ii}^2(1-P_{ii})^2 = \sum_i P_{ii}^2 \leq Tr(P^2) = O(\frac{1}{\alpha})$, by Lemma A0.

So $\sum_{i\neq j} C_{ij}^2 = Tr(CC') = Tr(P^2) - Tr(P(I-P)\bar{P}') - Tr(\bar{P}(I-P)P) + Tr(\bar{P}(I-P)^2\bar{P}') = O(\frac{1}{\alpha})$.

Now, we show that, $\sum_{i\neq j} C_{ij}C_{ji} = O(\frac{1}{\alpha})$. With: $\sum_{i\neq j} C_{ij}C_{ji} = Tr(C^2)$.

We have $C^2 = (P - \bar{P}(I-P))(P - \bar{P}(I-P)) = P^2 - P\bar{P}(I-P) - \bar{P}(I-P)P + \bar{P}(I-P)\bar{P}(I-P)$

So, $Tr(C^2) = Tr(P^2) - Tr(P\bar{P}(I-P)) - Tr(\bar{P}(I-P)P) + Tr(\bar{P}(I-P)\bar{P}(I-P))$,

. $Tr(P^2) = O(\frac{1}{\alpha})$. $Tr(P\bar{P}(I-P)) = Tr(\bar{P}(I-P)P) = O(\frac{1}{\alpha})$,

. $Tr(\bar{P}(I-P)\bar{P}(I-P)) = Tr(\bar{P}^2 - \bar{P}^2P - \bar{P}P\bar{P} + \bar{P}P\bar{P}P)$,

. We have: $Tr(\bar{P}(I-P)\bar{P}(I-P)) = Tr(\bar{P}^2) + Tr(\bar{P}^2P) - Tr(\bar{P}P\bar{P}) + Tr(\bar{P}P\bar{P}P)$,

. $Tr(\bar{P}^2) = \sum_i \frac{P_{ii}^2}{(1-P_{ii})^2} \leq \sum_i P_{ii}^2 = O(\frac{1}{\alpha})$,

. $Tr(\bar{P}^2P) = \sum_i \frac{P_{ii}^3}{(1-P_{ii})^2} = O(\frac{1}{\alpha})$,
. $Tr(\bar{P}P\bar{P}) = Tr(\bar{P}^2P) = O(\frac{1}{\alpha})$,
. $Tr(\bar{P}P\bar{P}P) =\leq sup\lambda(\bar{P}P)Tr(\bar{P}P) \leq cst\sum_i P_{ii}^2 \leq cstO(\frac{1}{\alpha})$.
Finally, $\sum_{i\neq j} C_{ij}C_{ji} = Tr(C^2) = O(\frac{1}{\alpha})$.
Then, $E[u'C\epsilon\epsilon'C'u] = O(\frac{1}{\alpha})$.
(v) The same proof as in DN (proof of lemma A3(vi) applies with P replaced by C.
(vi) The same proof as in DN (proof of lemma A3(vi)) applies with K replaced by $1/\alpha$ and P by C.
(vii) The proof follows from lemma A3(vii) in DN.
(viii) The same proof as in DN (proof of lemma Aa3 (viii)) applies here with their K replaced by $1/\alpha$ and Δ_K by Δ_α.
Proof of Theorem 3. (continued).
Let $S(\alpha) = H^{-1}\left[\Sigma_u^2\sigma_\varepsilon^2\frac{tr(CC')}{n} + \sigma_{u\varepsilon}\sigma_{u\varepsilon}'\frac{tr(C^2)}{n} + \sigma_\varepsilon^2\frac{f'(I-C')(I-C)f}{n}\right]H^{-1}$.
By lemma A7, we have $\rho_{\alpha,n} = tr(S(\alpha)) = O\left(\frac{1}{\alpha n}\right) + O\left(\alpha^\beta\right)$ for SC and LF, and $O\left(\frac{1}{\alpha n}\right) + O\left(\alpha^{\beta\wedge 2}\right)$ for T.
Now we check the conditions of Lemma A1 of DN with :
$\hat{h} = h + T^h + Z^h$ with $h = f'\epsilon/\sqrt{n}$,
$Z^h = 0$,
$T^h = T_1^h + T_2^h$,
$T_1^h = -f'(I-C)\epsilon/\sqrt{n}$,
$T_2^h = u'C'\epsilon/\sqrt{n}$,
$\hat{H} = H + T^H + Z^H$, with $H = ff'/n$, $T^H = T_1^H + T_2^H$,
$T_1^H = -f'(I-C)'f/n$,
$T_2^H = (u'f + f'u)/n$,
$Z^H = (u'C'u - u'(I-C)'f - f'(I-C)'u)/n$.
1) $\left\|Z^H\right\| = o(\rho_{\alpha,n})$ by the triangular inequality and Markov inequality.
2) $E[T_1^hT_1^{h\prime}|\mathcal{Z}] = E[f'(I-C')\epsilon\epsilon'(I-C)f/n|\mathcal{Z}] = \sigma_\epsilon^2(f'(I-C')(I-C)f/n)$.
3) $E[T_1^hh'|\mathcal{Z}] = E[f'(I-C')\epsilon\epsilon'f/n|\mathcal{Z}] = \sigma_\epsilon^2 f'(I-C)f/n$.

4) $E[hh'H^{-1}T_1^H|\mathcal{Z}] = E[f'\epsilon\epsilon' f H^{-1} f'\,(I-C)'\,f/n^2|\mathcal{Z}] = \sigma_\epsilon^2 f'(I-C')f/n$
$= E[T_1^h h'|\mathcal{Z}]$.
5) $E[T_2^h T_2^{h\prime}|\mathcal{Z}] = E[u'C\epsilon\epsilon' C'u|\mathcal{Z}] = \Sigma_u^2\sigma_\varepsilon^2 \frac{tr(CC')}{n} + \sigma_{u\varepsilon}\sigma_{u\varepsilon}' \frac{tr(C^2)}{n}$.
6) $E[hT_2^{h'}|\mathcal{Z}] = E[f'\epsilon\epsilon' Cu/n|\mathcal{Z}] = E[f'\epsilon\left(\sum_{i\neq j}\epsilon_i C_{ij} u_j\right)/n|\mathcal{Z}] = 0$.
7) $E[T_1^h T_2^{h'}|\mathcal{Z}] = E[f'(I-C)\epsilon\epsilon' Cu/n|\mathcal{Z}] = 0$.
8) $E[hh'H^{-1}T_2^H|\mathcal{Z}] = E[hh'H^{-1}\left(u'f + f'u\right)/n|\mathcal{Z}] = O_p(1/n)$.
Let $\hat{Z}^A(\alpha) = 0$ and $\hat{A}(\alpha) = (h+T^h)(h+T^h)' - hh'^{-1}T^{H'} - T^H H^{-1} hh'$. We have:
$E(\hat{A}(\alpha)|\mathcal{Z}) = \sigma_\epsilon^2 H + \Sigma_u^2\sigma_\varepsilon^2 \frac{tr(CC')}{n} + \sigma_{u\varepsilon}\sigma_{u\varepsilon}' \frac{tr(C^2)}{n} + \sigma_\varepsilon^2 \frac{f'(I-C')(I-C)f}{n} + o_p\left(\rho_{\alpha,n}\right)$.
$E(\hat{A}(\alpha)|X) = \sigma_\epsilon^2 H + HS(\alpha)H$.

Chapter 2

Testing overidentifying restrictions with many instruments and heteroskedasticity using regularized Jackknife IV.

2.1 Introduction

When the number of the instruments grows, it is well known that the conventional J test for overidentifying restrictions performs poorly[1]. It was shown that the asymptotic behavior of the conventional J test of Hansen (1982) gives a limit distribution which is not standard when the number of instruments or moment conditions is very large (see Kunitomo, Morimune, and Tsukuda (1983) and Burnside and Eichenbaum (1996)).

We propose a modified version of the J test when the number of moment conditions increases and the error is heteroskedastic, we also allow for the presence of weak identification. There is no assumptions or restrictions imposed on the number of instruments L, which can be larger or smaller than the sample size n. We construct our proposed test by using regularization to compute the inverse involved in the projection matrix P, instead of using the projection matrix (see Carrasco, Florens, and Renault (2007) for a review of inverse problems). For that purpose, we apply the Tikhonov regularization method, which is also known as the ridge regression to stabilize the inverse in the projection matrix P. Our method involves a tuning (regularization) parameter, α, which needs to be selected. As in Carrasco and Doukali (2016) we choose α that minimizes the cross-validation approximation of the mean squared error (MSE). Our Monte Carlo study shows that our proposed J test, using the regularized Jackknife IV estimator (RJIVE), performs favorably compared to other existing J tests.

Other regularization techniques could be used in this framework such as the Landweber-Fridman technique which is an iterative method or the principal component which consists in selecting the eigenvectors associated with the largest eigenvalues. Carrasco (2012) used those regularization techniques to

[1]This chapter is a joint work with Marine Carrasco. The authors thank the participants of the CIREQ Econometrics Conference in honor of Jean-Marie Dufour for helpful comments. Carrasco thanks SSHRC for partial financial support.

estimate a linear model in the presence of many instruments in a consistent and efficient way. Carrasco and Doukali (2016) proposed a new estimator which they called the regularized Jackknife instrumental variable estimator (RJIVE) when the number of available instruments is very large in linear models. In fact, there is a long history concerning the many instruments theory, Chao and Swanson (2005) proposed a consistent estimation of various k-class IV estimators when the number of instruments grows but at a slower rate of convergence than the sample size. Bai and Ng (2008) and Kapetanios and Marcellino (2010) proposed to use a variable selection method by using the estimated factors as instruments, but they make the assumption that the endogenous variables depend on a small number of factors which are exogenous. Donald and Newey (2001) selected the number of instruments by minimizing an approximation of the mean squared error. Regularization has been also introduced in the context of times series and forecasting macroeconomic variables using a large number of predictors, see Stock and Watson (2002), Bai and Ng (2008) and De Mol, Giannone, and Reichlin (2008).

There are many studies related to this paper. Lee and Okui (2012) proposed a modification of the Sargan (1958)'s test of overidentifying restrictions in a homoskedastic framework when the number of instruments L grows with the sample size n. They established the asymptotic null distribution of their proposed test statistic and studied its local power under some regularity conditions. Anatolyev and Gospodinov (2011) proposed a modification of the Anderson-Rubin (AR) test and of the conventional J test for overidentifying restrictions in linear models with homoskedasticity assumption under many instruments asymptotics. They consider an alternative way to compute the critical values of the chi-squared distribution. In a recent paper, Carrasco and Tchuente (2016a) propose to use regularization techniques to construct a robust Anderson Rubin (AR) test in linear models when the number of instruments is large. Their inference relies on a new restricted

efficient boostrap method and simulated Monte Carlo test. The closest paper to our approach is Chao, Hausman, Newey, Swanson, and Woutersen (2014), where they propose a new version of the J test that is robust to many instruments and heteroskedasticity. Their test is based on subtracting out the diagonal terms in the numerator of the test statistic. They consider the heteroskedasticity robust version of the Fuller (1977) estimator of Hausman, Newey, Woutersen, Chao, and Swanson (2012). Here, we consider instead the regularized Jackknife instrumental variable estimator (RJIVE). We choose this estimator because of its good properties (see Carrasco and Doukali (2016) for more details) and we implement the Tikhonov technique to stabilize the projection matrix P that appears in the numerator of the test statistic in order to improve the accuracy of the overidentifying restrictions test for linear models in presence of a large number of instruments. The advantage of the regularization is that it permits to handle the case where the number of instruments exceeds the sample size.
The remainder of this paper is organized as follows. Section 2 describes the model and the test statistic. Section 3 establishes asymptotic results. Section 4 reports Monte Carlo simulation results. Empirical applications are illustrated in section 5. Section 6 concludes. All of the proofs are provided in the appendix.

2.2 Model, estimator, and test statistic

This section presents the model, the estimator and the test statistic. We note that, unlike the other existing test statistics, the number of moment conditions is not restricted and may be smaller or larger than the sample size. We propose a regularized J test.

Consider the linear IV regression model:

$$y_i = X_i'\delta_0 + \epsilon_i \tag{2.2.1}$$

$$X_i = \Upsilon_i + u_i \tag{2.2.2}$$

$i = 1, \ldots, n$. The vector of interest is δ_0 which is a $p \times 1$ vector. y_i is the scalar outcome variable. The vector Υ_i is the optimal instrument which is typically unknown. We assume that y_i and X_i are observed but the Υ_i is not and $E(X_i\epsilon_i) \neq 0$. The estimation will be based on a sequence of instruments $Z_i = Z(\tau; \nu_i)$ where ν_i is a vector of exogenous variables and τ is an index taking countable values.

For the estimation of δ_0, we consider the Tikhonov Jackknife estimator proposed in Carrasco and Doukali (2016) because of its good properties relative to other existing IV estimators in the presence of many instruments. First we recall the expression of the Jackknife estimator (JIVE) proposed by Angrist, Imbens, and Krueger (1999) when the number of instruments is finite.

$$\begin{aligned} \hat{\delta} &= (\hat{\Upsilon}'X)^{-1}(\hat{\Upsilon}'Y) & (2.2.3) \\ &= (\sum_{i=1}^{n} \hat{\Upsilon}_i X_i')^{-1} \sum_{i=1}^{n} \hat{\Upsilon}_i y_i & (2.2.4) \end{aligned}$$

The leave-one-out estimator $\hat{\Upsilon}_i$ is defined as $\hat{\Upsilon}_i = Z_i'\hat{\pi}_{-i}$, where $\hat{\pi}_{-i} = (Z'Z - Z_iZ_i')^{-1}(Z'X - Z_iX_i')$ is the OLS coefficient from running a regression of X on Z using all but the i^{th} observation.

The JIVE estimator can alternatively be written as:

$$\hat{\delta} = (\sum_{i=1}^{n} \hat{\pi}'_{-i} Z_i X_i')^{-1} \sum_{i=1}^{n} \hat{\pi}'_{-i} Z_i y_i \tag{2.2.5}$$

with

$$\hat{\pi}'_{-i}Z_i = (X'Z(Z'Z)^{-1}Z_i - P_{ii}X_i)/(1-P_{ii}) = \sum_{j\neq i}^{n} P_{ij}X_j/(1-P_{ii})$$

where P is a $n \times n$ matrix defined as $P = Z(Z'Z)^{-1}Z'$ and P_{ij} denotes the *(i,j)*th element of P.
Then, the JIVE estimator is given by:

$$\hat{\delta} = \hat{H}^{-1}\sum_{i\neq j}^{n} X_i P_{ij}(1-P_{jj})^{-1}y_j,$$

where $\hat{H} = \sum_{i\neq j}^{n} X_i P_{ij}(1-P_{jj})^{-1}X'_j$, and $\sum_{i\neq j}$ denotes the double sum $\sum_i \sum_{j\neq i}$. When the number of the instruments is large, the inverse of Z'Z needs to be regularized because it is singular or nearly singular.
Now let us suppose that the number of moment conditions is finite or countable infinite. Here are some examples of Z_i.
- If $Z_i = \nu_i$ where ν_i is a L-vector of exogenous variables with a fixed L, then $Z(\tau;\nu_i)$ denotes the τth element of ν_i.
- $Z(\tau;\nu_i) = (\nu_i)^{\tau-1}$ with $\tau \in N$, thus we have an infinite countable sequence of instruments.
The expression of the Tikhonov Jackknife IV estimator $\hat{\delta}$ is:

$$\hat{\delta} = \hat{H}^{-1}\sum_{i\neq j}^{n} X_i P_{ij}^{\alpha}(1-P_{jj}^{\alpha})^{-1}y_j, \tag{2.2.6}$$

$$\hat{H} = \sum_{i\neq j}^{n} X_i P_{ij}^{\alpha}(1-P_{jj}^{\alpha})^{-1}X'_j \tag{2.2.7}$$

where P^{α} is a $n \times n$ matrix after regularization defined as:

$$P^{\alpha} = Z(Z'Z+\alpha I)^{-1}Z',$$

and P^α_{ij} denotes the (i,j)th element of P^α. The Tikhonov Jackknife estimator depends on a regularization term α. In practice, we choose α that minimizes the mean square error (MSE) as in Carrasco and Doukali (2016).

Remark:

It is useful to write the RJIVE as:

$$\hat{\delta} = \hat{H}^{-1} \sum_{i,j=1}^{n} X_i C^\alpha_{ij} y_j, \tag{2.2.8}$$

where $\hat{H} = \sum_{i,j=1}^n X_i C^\alpha_{ji} X'_j$, and $C^\alpha = (C^\alpha_{ij}) = \begin{cases} \frac{P^\alpha_{ij}}{1-P^\alpha_{ii}} & if \quad i \neq j \\ C^\alpha_{ii} = 0 & if \quad i = j \end{cases}$. Then, we obtain:

$$\sqrt{n}(\hat{\delta} - \delta_0) = \frac{(X'C^{\alpha\prime}X)^{-1}}{n} \frac{(X'C^{\alpha\prime}\epsilon)}{\sqrt{n}}. \tag{2.2.9}$$

The test statistic.

Chao et al. (2014) proposed a modified J statistic with many instruments based on the heteroskedasticity-robust version of the Fuller (1977) estimator, which is known as HFUL estimator. Their test statistic takes the form:

$$J_{CHNSW} = \frac{\hat{\epsilon}' P \hat{\epsilon} - \sum_{i=1}^n P_{ii} \hat{\epsilon}_i^2}{\sqrt{\hat{V}}} + L \tag{2.2.10}$$

with

$$\hat{V} = \frac{\hat{\epsilon}(2)' P(2) \hat{\epsilon}(2) - \sum_{i=1}^n P_{ii}^2 \hat{\epsilon}_i^4}{tr(P)} = \frac{\sum_{i\neq j}^n \hat{\epsilon}_i^2 P_{ij}^2 \hat{\epsilon}_j^2}{L}$$

where L is the number of instruments, P is the projection matrix, $\hat{\epsilon}_i = y_i - X'_i\hat{\delta}$, $\hat{\epsilon}(2) = (\hat{\epsilon}_1^2, ..., \hat{\epsilon}_n^2)$, $P(2)$ is the n-dimensional square matrix with ijth component equal to P_{ij}^2. Note that the numerator of the test statistic, $\sum_{i\neq j}^n \hat{\epsilon}_i P_{ij} \hat{\epsilon}_j$, is the numerator of the traditional Sargan test without the observation i. The denominator is a heteroskedastic consistent estimator of the variance of $\sum_{i\neq j}^n \hat{\epsilon}_i P_{ij} \hat{\epsilon}_j$. The test rejects the null hypothesis when

J_{CHNSW} is greater than the critical value of a chi-squared distribution with $L-p$ degrees of freedom. Chao et al. (2014), Anatolyev and Gospodinov (2011) and Lee and Okui (2012) have proposed tests that allow for many instruments but they impose that the number of moment conditions L can not be larger than n, which is not the case in our present work.

In this paper, we assume that the number of moment conditions L is large relatively to n. The inverse of $Z'Z$ needs to be stabilized because it is nearly singular or even not invertible whenever $L \geq n$. The main contribution is the use of the Tikhonov regularization method to stabilize the inverse of $(Z'Z)$ in presence of many instruments. For an arbitrary $n \times 1$ vector v, we define the $n \times n$ matrix P^{α} [2] as:

$$P^{\alpha} v = \sum_{j=1}^{\infty} \frac{\lambda_j^2}{\lambda_j^2 + \alpha} \langle v, \psi_j \rangle \psi_j$$

where ψ_j denotes the eigenvector of the $n \times n$ matrix ZZ'/n associated with the jth eigenvalue λ_j. We note here that the Tikhonov technique involves a tuning parameter α. The case $\alpha = 0$ corresponds to the case without regularization. We obtain $P^0 = P = Z(Z'Z)^{\dagger} Z$, where $\dagger$ denotes the Moore-Penrose generalized inverse. The regularization parameter needs to go to zero at a certain rate characterized in Section 2.3.

To describe our proposed test statistic, let $P^{\alpha}(2)$ be the n-dimensional square matrix after regularization with $ijth$ component equal to $(P_{ij}^{\alpha})^2$.

The test statistic we propose is:

$$J_{Tikh} = \frac{\hat{\epsilon}' P^{\alpha} \hat{\epsilon} - \sum_{i=1}^{n} P_{ii}^{\alpha} \hat{\epsilon}_i^2}{\sqrt{\hat{V}}} + tr(P^{\alpha}) \qquad (2.2.11)$$

[2] Appendix A gives a detailed definition of the Tikhonov method and the definition of P^{α}.

with

$$\hat{V} = \frac{\hat{\epsilon}(2)'P^{\alpha}(2)\hat{\epsilon}(2) - \sum_{i=1}^{n}\left(P_{ii}^{\alpha}\right)^2\hat{\epsilon}_i^4}{tr(P^{\alpha})} = \frac{\sum_{i\neq j}^{n}\hat{\epsilon}_i^2\left(P_{ij}^{\alpha}\right)^2\hat{\epsilon}_j^2}{tr(P^{\alpha})}, \qquad (2.2.12)$$

where $\hat{\epsilon}_i = y_i - X_i'\hat{\delta}$ where $\hat{\delta}$ is the regularized Jackknife estimator of Carrasco and Doukali (2016).

It will be shown in the next section that J_{Tikh} follows asymptotically a chi-squared with $tr(P^{\alpha}) - p$ degrees of freedom. Let $q_r(\tau)$ be the τth quantile of chi-squared distribution with r degrees of freedom. We reject the null hypothesis of our test with the asymptotic rejection frequency β if $J_{Tikh} \geq q_{tr(P^{\alpha})-p}(1-\beta)$.

Our test has the same form as Chao et al. (2014) test with the the projection matrix P replaced by the regularized projection matrix P^{α} and the number of instruments L replaced by the trace of P^{α}, i.e $tr(P^{\alpha})$.

2.3 Asymptotic distribution

This section presents the asymptotic theory under which we establish the limiting behaviour of our proposed test statistic in the presence of many moment conditions. We consider many weak instruments asymptotic as in Chao et al. (2014).

Let K be the covariance operator defined in Appendix A. For a finite number of instruments, $K = Z'Z/n$.

Assumption 1. *(i) The operator K is nuclear. (ii) There exists a constant $\bar{C}$ such that $P_{ii}^{\alpha} \leq \bar{C} < 1$, $i = 1, ..., n$.*

Assumption 1 (i) is the same as in Carrasco (2012). Condition (i) means that the eigenvalues of the covariance operator K are summable. Condition (ii) is reminiscent of Assumption 1 in Chao et al. (2014): "for some

$\bar{C} < 1$, $P_{ii} < \bar{C}$, $i = 1, ..., n$". However it is much less restrictive. Indeed, $P_{ii} < \bar{C} < 1$ implies that $\sum_i \frac{P_{ii}}{n} = \frac{L}{n} < 1$, L =rank(Z), which restricts the number of instruments. Our condition $P_{ii}^{\alpha} \leq \bar{C} < 1$ implies that trace(P^{α}) $= \sum_i q_i < n$, which implies a condition on α. Recall from Carrasco (2012) that $\sum_i q_i = O(\frac{1}{\alpha})$. So Assumption (ii) implies $\frac{1}{\alpha n} < 1$.
The next assumption allows for the presence of many weak instruments. A measure of the strength of the instruments is the concentration parameter, which can be seen as a measure of the information contained in the instruments. If one could approximate the reduced form Υ by a sequence of instruments Z, so that $X = Z'\pi + u$ where $E[u^2|Z] = \sigma_u^2$, the concentration parameter would be given by

$$CP = \frac{\pi' Z' Z \pi}{\sigma_u^2}.$$

The following assumption generalizes this notion.

Assumption 2. $\Upsilon_i = S_n f_i/\sqrt{n}$ *where* $S_n = \hat{S}_n diag(\mu_{1n}, \ldots, \mu_{pn})$ *such that* $\hat{S}_n$ *is* $p \times p$ *bounded, the smallest eigenvalue of* $\hat{S}_n \hat{S}_n'$ *is bounded away from zero; for each* j*, either* $\mu_{jn} = \sqrt{n}$ *(strong identification) or* $\frac{\mu_{jn}}{\sqrt{n}} \to 0$ *(weak identification). Moreover* $\mu_n = \underset{1<j<p}{min}\, \mu_{jn} \to \infty$ *and* $1/(\sqrt{\alpha}\mu_n^2) \to 0$, $\alpha \to 0$. *Also there is a constant* $\bar{C}$ *such that* $|| \sum_{i=1}^n f_i f_i'/n|| \leq \bar{C}$ *and* $\lambda_{min}(\sum_{i=1}^n f_i f_i'/n)$ $\geq 1/\bar{C}$*, a.s.n.*

Assumption 2 allows for both strong and weak instruments. If $\mu_{jn} = \sqrt{n}$, the instrument is strong. If μ_{jn}^2 is growing slower than n, this leads to a weak identification as that of Chao and Swanson (2005). f_i defined in Assumption 2 is unobserved and has the same dimension as the infeasible optimal instrument, Υ_i. Then f_i can be seen as a rescaled version of this optimal instrument.
An illustration of assumption 2 is as follows. Let us consider the simple linear model $y_i = z_{i1}\delta_1 + \delta_{0p} x_{i2} + \epsilon_i$, where z_{i1} is an included instruments and

x_{i2} is an endogenous variable. Suppose that x_{i2} is a linear combination of the included instrumental z_{i1} and an unknown excluded instruments z_{ip}, i.e $x_{i2} = \pi_1 z_{i1} + (\frac{\mu_n}{\sqrt{n}})z_{ip}$. The reduced form is:

$$\Upsilon_i = \begin{pmatrix} z_{i1} \\ x_{i2} \end{pmatrix} = \begin{pmatrix} z_{i1} \\ \pi_1 z_{i1} + (\frac{\mu_n}{\sqrt{n}})z_{ip} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \pi_1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{\mu_n}{\sqrt{n}} \end{pmatrix} \begin{pmatrix} z_{i1} \\ z_{ip} \end{pmatrix}$$

with

$$\hat{S}_n = \begin{pmatrix} 1 & 0 \\ \pi_1 & 1 \end{pmatrix}, \ \mu_{jn} = \begin{cases} \sqrt{n} & , \ j = 1 \\ \mu_n & , \ j = 2 \end{cases}, \text{ with } \frac{\mu_n}{\sqrt{n}} \to 0, \text{ and } f_i = \nu_i = \begin{pmatrix} z_{i1} \\ z_{ip} \end{pmatrix}.$$

Assumption 3. *There is a constant $C > 0$ such that $(\epsilon_1, U_1), ..., (\epsilon_n, U_n)$ are independent, with $E[\epsilon_i] = 0$, $E[U_i] = 0$, $E[\epsilon_i^2] \le C$, $E[||U_i||^2] \le C$, $Var((\epsilon_i, U_i')') = diag(\Omega_i, 0)$, and $\lambda_{min}(\sum_{i=1}^n \Omega_i/n) \ge 1/C$.*

This assumption requires the second conditional moments of the disturbances to be bounded. It also imposes uniform nonsingularity of the variance of the reduced form disturbances.

Assumption 4. *There exists a π_L such that $\sum_{i=1}^n ||f_i - \pi_L Z_i||^2/n \to 0$.*

Assumptions 1 and 4 imply that the structural parameters are identified asymptotically. Assumption 4 implies that f_i belongs to the closure of the linear span of instruments. It does not imply that f_i is a finite linear combination of the instruments.

Assumption 5. *There is a constant $C > 0$ such that, with probability one, $\sum_{i=1}^n ||f_i||^4/n^2 \to 0$, $E[\epsilon_i^4] \le C$ and $E[||U_i||^4] \le C$.*

Assumption 5 can be found in Chao et al. (2014). It simplifies the asymptotic theory in the sense that certain terms converge.
Let $\sigma_i^2 = E[\epsilon_i^2]$,$\gamma_n = \sum_{i=1}^n E[U_i\epsilon_i]/\sum_{i=1}^n \sigma_i^2$, $\tilde{U} = U - \epsilon\gamma_n'$ having ith row $\tilde{U}_i'$and let $\tilde{\Omega}_i = E[\tilde{U}_i\tilde{U}_i']$.

Assumption 6. *$\mu_n S_n^{-1} \to S_0$ and $1/(\alpha\mu_n^2) \to C$ for finite C. Also each of the following exists:*

$H_p = \lim_{n\to\infty} \sum_{i=1}^{n} (1 - P_{ii}^{\alpha}) f_i f_i'/n$,
$\Sigma_p = \lim_{n\to\infty} \sum_{i=1}^{n} (1 - P_{ii}^{\alpha})^2 f_i f_i' \sigma_i^2/n$,
$\Psi = \lim_{n\to\infty} \alpha \sum_{i\neq j} (P_{ij}^{\alpha})^2 (\sigma_i^2 E[\tilde{U}_j \tilde{U}_j'] + E[\tilde{U}_i \epsilon_i] E[\tilde{U}_j' \epsilon_j])$.

Note that Assumptions 1, 2, and 6 imply some restrictions on α, namely α needs to go to zero but not too fast.

Theorem 1. Let $q_{tr(P^{\alpha})-p}(1-\beta)$ be the $(1-\beta)$ quantile of a chi-square distribution with $tr(P^{\alpha}) - p$ degrees of freedom. If assumptions 1-6 are satisfied then $Pr(\hat{T} \geq q_{tr(P^{\alpha})-p}(1-\beta)) \to \beta$.

Proof: In appendix B.

Theorem 1 shows that, under the many instruments asymptotic condition, our modified J test achieves the correct asymptotic critical value β. We can see this test as a specification test for the linear instrumental variables regression (see Hansen (1982)). If the model is correctly specified, all the moment conditions (including the overidentifying restrictions) should be close to zero. The novelty of our proposed test is that it is robust to many instruments in the sense that we do not make any assumption on the number of instruments.

Related Literature.

In the literature on testing overidentifying restrictions in linear models with many instruments, the J test performs poorly when we increase the number of the instruments. To deal with this problem, Anatolyev and Gospodinov (2011) proposed a new J test that guarantee the asymptotical sizes, but their test is valid only under the homoskedasticity case and when the number of instruments is a fraction of the sample size $0 < \frac{L}{n} < 1$. Lee and Okui (2012) proposed a modification of the Sargan (1958) test in the presence of a large number of instruments. They gave the limiting behavior of their proposed test statistic when the number of instruments and the sample size go to infinity, but they still maintained the assumption $0 < \frac{L}{n} < 1$. Donald, Imbens, and Newey (2003) established the asymptotic distribution of some parameter and specification tests in models when the number of instruments L increases asymptotically, but again slowly relative to the sample size n. They called this assumption a moderately many instruments, but the validity of their test fails in the case of the many instruments theory of Bekker (1994). Hahn and Hausman (2002) developed a new specification test for the validity of instrumental variables in linear models. They compared the difference of the forward (conventional) 2SLS estimator with the reserve 2SLS estimator under the assumption $0 < \frac{L}{n} < 1$.

In this paper, we consider the case when the number of instruments is potentially very large. The matrix $Z'Z$ may be nearly singular or possibly not invertible, so the projection matrix $P = Z(Z'Z)^{-1}Z'$ that appears in the numerator of the J test may affect the precision of the test statistic. Inverting $Z'Z$ can be seen as solving an ill-posed problem. We implement the Tikhonov technique to stabilize the projection matrix. The advantage of the regularization is that we can use all the available information and we do not need to discard some instruments a priori. This yields an improved performance of the J test as illustrated in the simulation study.

2.4 Simulation study

The goal of our simulation study is to demonstrate the finite-sample performance of the proposed J test and compare it to other existing J tests. We consider a linear model with one regressor and L instruments. The J statistic is interpreted as a test of the validity of the $L-1$ overidentifying restrictions. We investigate two cases: the homoskedastic and heteroskedastic case.

Homoskedastic case.

The data generating process (DGP) is generated as follows:

$$y_i = \delta X_i + \epsilon_i$$
$$X_i = z_i'\pi + u_i,$$

where $(\epsilon_i, u_i) \overset{iid}{\sim} N(0, \sum)$ and $\sum = \begin{pmatrix} 0.25 & 0.20 \\ 0.20 & 0.25 \end{pmatrix}$, $z_i \overset{iid}{\sim} N(0, I_L)$, $\delta = 1$, and $\pi = \frac{1}{\sqrt{L}}\iota_L$, where ι_L is an L-vector of ones.

Heteroskedastic case.

Now the error is allowed to be heteroskedastic. We keep the same DGP except that the errors are now generated as follows:

$\epsilon_i = \rho u_i + \sqrt{\frac{1-\rho^2}{\phi^2+0.86^4}}(\phi v_{1i} + 0.86 v_{2i})$, where $v_{1i} \overset{iid}{\sim} N(0, z_{1i}^2)$ and $v_{2i} \overset{iid}{\sim} N(0, (0.86)^2)$.

We choose $\rho = 0.3$, $\phi = 0.2$.

Tables 1 and 2 present the empirical size at 5% nominal level of J, J_{Corr}, J_{CHNSW} and J_{Tikh} tests which denote respectively the conventional J test, the modified J test proposed in Anatolyev and Gospodinov (2011), the modified J test proposed in Chao et al. (2014), and the Tikhonov J test proposed in this paper. These results are based on 5000 Monte Carlo replications. We consider values of $\lambda = \frac{L}{n}$ equal to 0.2, 0.5, 0.8, 0.95, and 1.1. The values of λ are used in combination with sample sizes of 100, 200 and 500. For the Tikhonov J test, the regularization parameter α [3] is chosen by minimizing the cross-validation approximation of the mean squared error (MSE) as in Carrasco and Doukali (2016).

Description of the other tests.

Hansen-Sargan J test.

Let $\hat{\delta}_{2SLS} = (X'PX)^{-1}X'Py$ be the two stage least-squared estimators. Let $\hat{\epsilon} = y - X\hat{\delta}_{2SLS}$. The Hansen-Sargan J test takes the following form:

$$J = \frac{\hat{\epsilon}'P\hat{\epsilon}}{\hat{\sigma}^2}, \tag{2.4.1}$$

with $\hat{\sigma}^2 = \hat{\epsilon}'\hat{\epsilon}/(n-L)$.

The decision rule of Hansen-Sargan J test consists in rejecting the null hypothesis if J exceeds the critical value given by the chi-square distribution

[3] The regularization parameter α is searched over the interval [0.01,0.5] with 0.01 increment.

with $L-p$ degrees of freedom.

Anatolyev and Gospodinov (2011)'s J test.

They suggest to use the same J statistic as in (2.4.1) with $\hat{\epsilon} = y - X\hat{\delta}_{LIML}$, where $\hat{\delta}_{LIML}$ is the limited information maximum likelihood estimator of δ but the critical value is modified. The decision rule consists in rejecting H_0 at the level β if J exceeds the quantile of a chi-square distribution with $L-p$ degrees of freedom and probability $\Phi(\sqrt{1-\frac{L}{n}}\Phi^{-1}(\beta))$, where Φ is the distribution function of the standard normal.

Chao et al. (2014)'s J test.

J_{CHNSW} uses the test described in equation (2.2.10) with $\hat{\epsilon} = y - X\hat{\delta}_{HFULL}$, where $\hat{\delta}_{HFULL}$ is the heteroskedasticity-robust version of the Fuller (1977) estimator of Hausman et al. (2012).

Tables 2.1 and 2.2 report the empirical sizes of the four tests in the homoskedastic and the heteroskedastic cases respectively. We remark that the performance of the conventional J test is sensitive to the number of instruments, ie the rejection frequencies for the J test is not close to the nominal value 5% throughout these tables. We also remark that Anatolyev and Gospodinov (2011)'s J test, the J_{CHNSW} and the J_{Tikh} perform very well when the number of instruments increase. However, J, J_{Corr}, and J_{CHNSW} tests exhibit a large size distortion when λ is close to 1 (i,e. λ= 0.95), which is worse in the heteroskedastic case. Our regularized J_{Tikh} has almost correct size results even with very large number of instruments. When the number of instruments is larger than the sample size, the J, J_{Corr}, and J_{CHNSW} can not be computed. Table 2.1 and 2.2 show also that our proposed regularized J test performs well when $L > n$, in the sense that the empirical rejection rates are close to the nominal value 5%.

To compare the powers of the different J tests, we consider the same design as before, but the structural error is giving by: $\epsilon_i = u_i + \rho_z z_{1i}$. We allow

the correlation ρ_z between structural error and instrument to vary between 0 and 1. We choose $n = 500$ and $\lambda = 0.8$.
The rejection frequencies under the null hypothesis (ρ_z=0) are 0.048, 0.049, 0.051 respectively for J_{Corr} J_{CHNSW} and the J_{Tikh} for homoskedastic case. For the heteroskedastic case they are 0.046, 0.045, 0.050. The power curves (rejection frequencies) are plotted in Figures 2.1 and 2.2. We see that J_{Tikh} statistic has typically better power properties than the J_{Corr} and J_{CHNSW}.
In conclusion, simulations suggest that the implementation of the Tikhonov regularization can increase the power, while controlling for the size. Thus, the regularization provides a correction to size distortions for the J test arising from the use of many instruments.

Table 2.1: Empirical rejection rates at 0.05 nominal level of the J test - homoskedastic case

λ	0.2	0.5	0.8	0.95	1.1
$n = 100$					
J	0.06	0.017	0	0	NA
J_{Corr}	0.069	0.055	0.049	0.048	NA
J_{CHNSW}	0.071	0.057	0.049	0.056	NA
J_{Tikh}	0.072	0.065	0.070	0.061	0.059
$n = 200$					
J	0.054	0.026	0	0	NA
J_{Corr}	0.057	0.059	0.045	0.04	NA
J_{CHNSW}	0.062	0.057	0.042	0.029	NA
J_{Tikh}	0.063	0.062	0.058	0.056	0.053
$n = 500$					
J	0.059	0.039	0	0	NA
J_{Corr}	0.054	0.057	0.046	0.038	NA
J_{CHNSW}	0.055	0.056	0.047	0.031	NA
J_{Tikh}	0.056	0.059	0.053	0.050	0.050

Table 2.2: Empirical rejection rates at 5% nominal level of the J test - heteroskedastic case

λ	0.2	0.5	0.8	0.95	1.1
$n = 100$					
J	0.044	0.007	0	0	NA
J_{Corr}	0.072	0.079	0.141	0.040	NA
J_{CHNSW}	0.074	0.080	0.144	0.09	NA
J_{Tikh}	0.070	0.056	0.047	0.049	0.045
$n = 200$					
J	0.044	0.008	0	0	NA
J_{Corr}	0.064	0.060	0.104	0.113	NA
J_{CHNSW}	0.066	0.059	0.102	0.125	NA
J_{Tikh}	0.064	0.052	0.046	0.047	0.046
$n = 500$					
J	0.036	0.008	0	0	NA
J_{Corr}	0.056	0.050	0.052	0.126	NA
J_{CHNSW}	0.058	0.0492	0.051	0.126	NA
J_{Tikh}	0.056	0.053	0.048	0.048	0.049

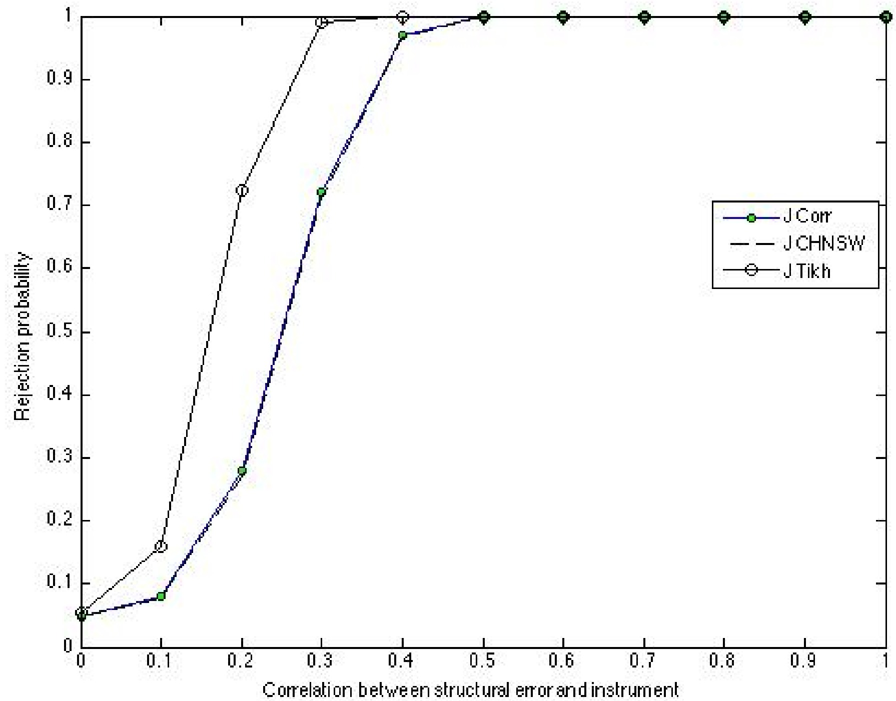

Figure 2.1: Power curves of J tests, n=500, $\lambda = 0.8$, homoskedastic case

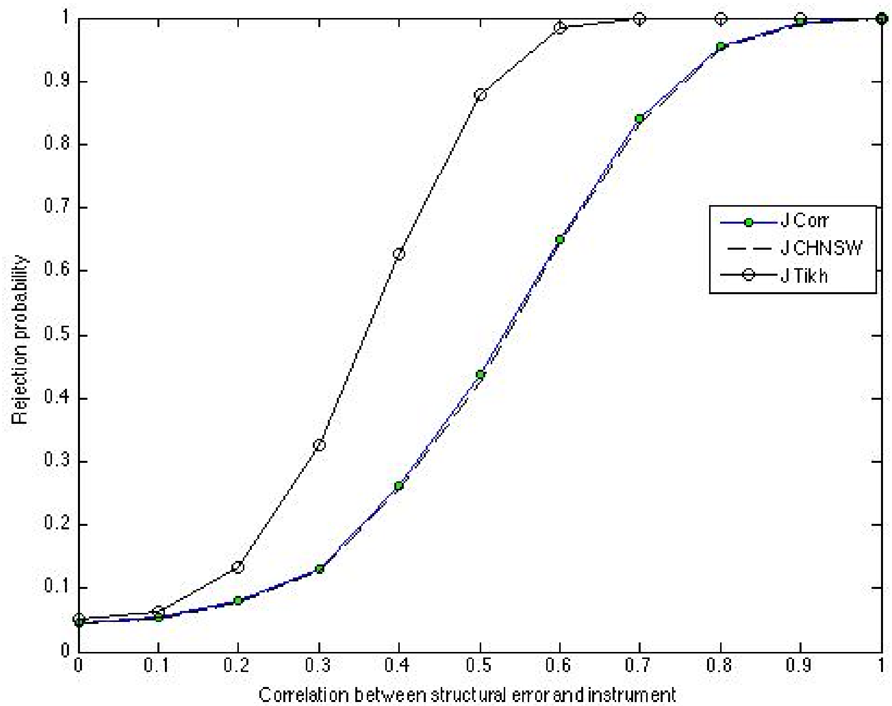

Figure 2.2: Power curves of J tests, n=500, $\lambda = 0.8$, heteroskedastic case

2.5 Empirical applications.

2.5.1 Elasticity of intertemporal substitution.

The elasticity of intertemporal substitution (EIS) in consumption is crucial in macroeconomics and finance. We follow the specification in Yogo (2004) [4] who analyzes the problem of the estimation of the EIS using the linearized Euler equation.
The estimated model is as follows:

$$\Delta c_{t+1} = \tau + \psi r_{f,t+1} + \xi_{t+1} \tag{2.5.1}$$

$$r_{f,t+1} = \mu + \frac{1}{\psi}\Delta c_{t+1} + \eta_{t+1}, \tag{2.5.2}$$

where ψ is the EIS, Δc_{t+1} is the consumption growth at time $t+1$, $r_{f,t+1}$ is the real return on a risk free asset, τ and μ are constants, and ξ_{t+1} and η_{t+1} are the innovations to consumption growth and asset return respectively.
Yogo (2004) explains how weak instruments have been the cause of the EIS empirical puzzle. He shows that, using conventional IV methods, the estimated EIS, ψ, is significantly less than 1 but its reciprocal is not different from 1. Carrasco and Tchuente (2015) estimate EIS using many instruments and regularization. They increase the number of instruments [5] from 4 to 18 by including interactions and power functions. Their point estimates are similar to those used for macro calibrations.
Tables 2.3 reports the test statistics corresponding to the four J tests based

[4] Yogo (2004) used quarterly data from 1947.3 to 1998.4 for the United States.

[5] The instruments that Yogo (2004) uses are: the twice lagged, nominal interest rate (r), inflation (i), consumption growth (c) and log dividend rate (p). We denote this bloc of instruments by Z=[r, i, c, p]. The 18 instruments used in our regression are derived from Z and are given by $\underline{Z} = [Z, Z.^2, Z.^3, Z(:,1) \star Z(:,2), Z(:,1) \star Z(:,3), Z(:,1) \star Z(:,4), Z(:,2) \star Z(:,3), Z(:,2) \star Z(:,4), Z(:,3) \star Z(:,4)]$.

on the 18 instruments of Carrasco and Tchuente (2015). We find that the J statistic of the conventional J test, the J_{corr}, and J_{CHNSW} are larger than chi-square critical value, which means that they reject the null hypothesis. However, the J statistic of our proposed Tikhonov test is smaller than the chi-square critical value. We can conclude that the model is correctly specified according to our proposed test, so the instruments used in the model seem to be exogenous.

It may seem surprising that the J_{Tikh} is so much smaller than other J tests. One possible explanation is the presence of heteroskedasticity. The errors in equations (2.5.1) and (2.5.2) are found to be heteroskedastic according to the F test (p-value= 0.073). The J and J_{Corr} are not robust to heteroskedasticity which may explain the difference of conclusions. However, J_{CHNSW} is robust to heteroskedasticity. An explanation for the difference between J_{CHNSW} and J_{Tikh} may be that the matrix $\underline{Z'Z}$ is very ill-conditioned. The condition number [6], which is the ratio of the largest eigenvalue on the smallest eigenvalue of $\underline{Z'Z}/n$, is an indicator on how ill-posed the matrix $\underline{Z'Z}/n$. The higher the condition number, the more imprecise the inverse of $\underline{Z'Z}/n$ will be. The smallest possible condition number is 1 (which corresponds to the identity matrix). In this application, the condition number is equal to $5.06\ 10^5$, and the smallest eigenvalue is equal to $2.35\ 10^{-5}$. This suggests that regularization is necessary to stabilize the inverse. Finaly, we note here that the instruments are standardized, which means that the instruments are divided with their standard deviation. This standardization has no impact on 2SLS and LIML estimators which are scale invariant and used to construct the J, J_{Corr}. However, our Tikhonov J test, and J_{CHNSW} 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).

[6] The condition number is scale invariant.

2.5.2 New-Keynesian Phillips Curve.

We follow the specification in Galı and Gertler (1999) and estimate the following model, representing a New-Keynesian Phillips Curve:

$$\begin{aligned} \pi_t &= \gamma + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} + \lambda mc_t \\ \pi_{t+1} &= E_t \pi_{t+1} + \nu_{t+1}, \end{aligned}$$

where π_t is the inflation rate at time t, mc_t is the marginal cost, $E_t\pi_{t+1}$ is the expectation of future inflation conditional on the information set at time t, and ν_{t+1} is a white noise error term. From the stochastic process describing how actual inflation and inflation expectations are related, it follows that $E_t\pi_{t+1} = \pi_{t+1} - \nu_{t+1}$ and then:

$$\begin{aligned} \pi_t &= \gamma + \gamma_f \pi_{t+1} + \gamma_b \pi_{t-1} + \lambda mc_t - \gamma_f \nu_{t+1} \\ &= \gamma + \gamma_f \pi_{t+1} + \gamma_b \pi_{t-1} + \lambda mc_t + \epsilon_t. \end{aligned}$$

We use the orthogonality condition $E_t[(\pi_t - \gamma - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1} - \lambda mc_t) Z_t] = 0$ [7], where Z_t is a matrix of instruments including variables that are orthogonal to ν_{t+1}. We use the set of instruments $\{\pi_{t-i}\}_{i=1}^{k}$ and $\{mc_{t-i}\}_{i=1}^{k}$ plus a constant term, we estimate the model over the sample 1959.4−2007.4 for US quarterly data [8], and we get the results in the table 2.4 below for a value of k equal to 10. Note that, in this application, the total number of instruments is $L = 2k + 1$.

Table 2.4 reports the test statistics corresponding to different J tests. We find that the conventional J test and the J_{corr} are larger than chi-square critical value, which means that the null hypothesis is rejected. However,

[7] The instruments are standardized as in the previous application.

[8] The data were downloaded from the Federal Reserve Bank of St Louis' website: https://www.stlouisfed.org/.

J_{CHNSW} and the J statistic of our proposed Tikhonov test [9] are smaller than the chi-square critical value, then we can conclude that the model is correctly specified. The difference in the conclusion may be due to the presence of heteroskedasticity (the p-value of the F test is 0.061). Indeed, J and J_{corr} tests are not robust to heteroskedasticity.

Table 2.3: Estimated J statistics for the EIS Model.

	J	J_{Corr}	J_{CHNSW}	J_{Tikh}
ψ	34.46	34.84	32.68	1.10
1/ ψ	53.23	61.42	41.48	0.510

$*$ The chi-square critical value= 26.29 (level=5% and the degree of freedom=16).
$*$ Critical value of the J_{corr}= 25.72 (level=5% and the degree of freedom=16).
$*$ $tr(P^{\alpha}) = 11.89$, the critical value for the $J_{Tikh} = 18.15$.

[9] For the New-Keynesian Phillips Curve' application, the condition number is equal to 1.53 10^5, and the smallest eigenvalue is equal to 3.73 10^{-6}, which suggests that regularization will be helpful to improve the J test.

Table 2.4: Estimated J statistics for the New-Keynesian Phillips Curve Model.

	J	J_{Corr}	J_{CHNSW}	J_{Tikh}
J **statistic**	39.60	34.92	26.09	21.65

∗ The chi-square critical value= 27,58 (level=5% and the degree of freedom=17).
∗ Critical value of the J_{corr}= 26.97 (level=5% and the degree of freedom=17).
∗ $tr(P^{\alpha}) = 18.51$, the critical value for the $J_{Tikh} = 24.39$.

2.6 Conclusion

To increase the accuracy of the conventional J test of overidentifying restrictions in linear models, researchers may use many instruments or moment conditions. However, many existing J tests perform poorly when the number of instruments is large. This paper proposes a new J test, based on Tikhonov regularization and studies its properties under many possibly weak instruments and heteroskedasticity. Simulations results show that the proposed test performs well. Its empirical size is close to the theoretical size and its power is greater than that of competing tests. We recommend the use of this modified J test in applied studies because of its ease of implementation and its robustness.

2.7 Appendix

2.7.1 Presentation of the Tikhonov Regularization.

Here we consider the general case where the estimation is based on a sequence of instruments $Z_i = Z(\tau; \nu_i)$ with $\tau \in N$. Assume τ lies in a space Ξ ($\Xi = \{1, .., L\}$ or $\Xi = \mathbb{N}$) and let π be a positive measure on Ξ. Let K be the covariance operator for instruments from $L^2(\pi)$ to $L^2(\pi)$ such that:

$$(Kg)(\tau) = \sum_{l=1}^{L} E(Z(\tau, \nu_i)Z(\tau_l, \nu_i))g(\tau_l)\pi(\tau_l).$$

where $L^2(\pi)$ denotes the Hilbert space of square integrable functions with respect to π. K is supposed to be a nuclear operator which means that its trace is finite. Let λ_j and ψ_j, $j = 1...$ be respectively the eigenvalues (ordered in decreasing order) and the orthogonal eigenfunctions of K. The operator can be estimated by K_n defined as:

$$K_n : L^2(\pi) \to L^2(\pi)$$
$$(K_n g)(\tau) = \sum_{l=1}^{L} \frac{1}{n} \sum_{i=1}^{n} (Z(\tau, \nu_i)Z(\tau_l, \nu_i))g(\tau_l)\pi(\tau_l).$$

If the number of instruments L is large relatively to n, inverting the operator K is considered as an ill-posed problem which means that the inverse is not continuous and its sample counterpart, K_n, is singular or nearly singular. To solve this problem we need to stabilize the inverse of K_n using regularization. A regularized inverse of an operator K is defined as: $R_\alpha : L^2(\pi) \to L^2(\pi)$ such that $lim_{\alpha \to 0} R_\alpha K \rho = \rho, \forall \rho \in L^2(\pi)$, where α is the regularization parameter (see Kress (1999) and Carrasco, Florens, and Renault (2007)).

Tikhonov regularization

We consider the Tikhonov regularization scheme.

$$\begin{aligned}(K^{\alpha})^{-1} &= (K^2+\alpha I)^{-1}K.\\ (K^{\alpha})^{-1}r &= \sum_{j=1}^{\infty}\frac{\lambda_j}{\lambda_j^2+\alpha}\langle r,\psi_j\rangle\,\psi_j.\end{aligned}$$

where $\alpha > 0$ and I is the identity operator. For the asymptotic efficiency α has to go to zero at a certain rate. The Tikhonov regularization is related to ridge regularization. Ridge method was first proposed in the presence of many regressors. The aim was to stabilize the inverse of XX' by replacing XX' by $XX' + \alpha I$. 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 implementation of the Tikhonov regularization and the selection of an appropriate ridge parameter for the first step regression helps to reduce this bias.
Let $(K_n^{\alpha})^{-1}$ be the regularized inverse of K_n and P^{α} a $n \times n$ matrix as defined in Carrasco (2012) by $P^{\alpha} = T(K_n^{\alpha})^{-1}T^*$ where $T : L^2(\pi) \to R^n$ with
$Tg = (< Z_1, g >, < Z_2, g >', ..., < Z_n, g >')'$
and $T^* : R^n \to L^2(\pi)$ with
$T^*v = \frac{1}{n}\sum_j^n Z_j v_j$
such that $K_n = T^*T$ and TT^* is a $n \times n$ matrix with typical element $\frac{<Z_i,Z_j>}{n}$.
Let $\hat{\phi}_j$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ..... \geq 0$, $j = 1, 2, ...$ be the orthonormalized eigenfunctions and eigenvalues of K_n and ψ_j the eigenfunctions 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 $v \in R^n$, $P^{\alpha}v = \sum_j^{\infty} q(\alpha, \lambda_j^2) < v, \psi_j > \psi_j$ where $q(\alpha, \lambda_j^2) = \frac{\lambda_j^2}{\lambda_j^2+\alpha}$.
Remark that the case when $\alpha = 0$ corresponds to no regularization Thus we have $q(\alpha, \lambda_j^2) = 1$ and $P^0 = Z(Z'Z)^{-1}Z'$.

2.7.2 Proofs

Lemma A0.

If Assumptions 1-3 are satisfied, then :

i) $P^{\alpha}_{ii} < 1$ *for* $\alpha > 0$,

ii) $\sum_{i \neq j} (P^{\alpha}_{ij})^2 = O(1/\alpha)$,

iii) $\sum_{i \neq j} P^{\alpha}_{ij} = O(1/\alpha)$.

iv) $\sum_{i,l,k,r} P^{\alpha}_{ik} P^{\alpha}_{kl} P^{\alpha}_{lr} P^{\alpha}_{ri} = O(1/\alpha)$.

v) $\sum_{i,j} (P^{\alpha}_{ij})^4 = O(1/\alpha)$.

vi) $\sum_{i,j,k} (P^{\alpha}_{ij})^2 (P^{\alpha}_{j,k})^2 = O(1/\alpha)$.

Proof of Lemma A0.

This is Lemma A0 of Carrasco and Doukali (2016).

Let us define some notations that will be used in the following Lemmas. For random variables[10] W_i, Y_i, η_i, let $\bar{w}_i = E[W_i]$, $\bar{y}_i = E[Y_i]$, $\bar{\eta}_i = E[\eta_i]$, $\tilde{W}_i = W_i - \bar{w}_i$ and $\tilde{Y}_i = Y_i - \bar{y}_i$, $\tilde{\eta}_i = \eta_i - \bar{\eta}_i$, $\bar{w}_n = E[(W_1, ..., W_n)']$, $\bar{y}_n = E[(Y_1, ..., Y_n)']$, $\bar{\mu}_W = max_{i \leq n} |\bar{w}_i|$, $\bar{\mu}_Y = max_{i \leq n} |\bar{y}_i|$, $\bar{\mu}_\eta = max_{i \leq n} |\bar{\eta}_i|$, $\bar{\sigma}^2_{W_n} = max_{i \leq n} var(W_i)$, $\bar{\sigma}^2_{Y_n} = max_{i \leq n} var(Y_i)$, $\bar{\sigma}^2_{\eta} = max_{i \leq n} var(\eta_i)$.

Define the norm: $||W||^2_{L_2} = \sqrt{E[W^2]}$, and let M, CS, T denote the Markov inequality, the Cauchy-Schwarz inequality, and the triangle inequality, respectively.

Lemma A1.

Suppose the following conditions hold:

(i) $P^{\alpha} v = Z(Z'Z + \alpha I)^{-1} Z' v$,

(ii)$(W_{1n}, U_1, \epsilon_1), ..., (W_{nn}, U_n, \epsilon_n)$ *are independent, and* $D_{1,n} := \sum_{i=1}^{n} E[W_{in} W'_{in}]$ *satisfies* $||D_{1,n}|| < C$,

(iii) $E[W'_{in}] = 0$, $E[U_i] = 0$, $E[\epsilon_i] = 0$, *and there is a constant* C *such that* $E[||U_i||^4] \leq C$ *and* $E[\epsilon_i^4] \leq C$,

(iv) $\sum_{i=1}^{n} E[||W_{in}||^4] \to 0$ *a.s.*

[10] Note that here W_i and η_i are arbitrary scalar variables that will take various forms in the sequel.

(v) $\alpha \to 0$ *as* $n \to \infty$.

Then for:

$D_{2,n} := \alpha \sum_{i \neq j}^{n} (P_{ij}^{\alpha})^2 (E[U_i U_i']E[\epsilon_j^2] + E[U_i \epsilon_i]E[U_j' \epsilon_j])$

and any sequences c_{1n} *and* c_{2n} *with* $||c_{1n}|| \leq C$, $||c_{2n}|| \leq C$, *and* $\sum_n = c_{1n}' D_{1,n} c_{1n} + c_{2n}' D_{2,n} c_{2n} > 1/C$, *it follows that:*

$\bar{Y}_n = \sum_n{}^{-1/2} (c_{1n}' \sum_{i=1}^{n} W_{i,n} + \sqrt{\alpha} c_{2n}' \sum_{i \neq j}^{n} U_i (P_{ij}^{\alpha})^2 \epsilon_j) \xrightarrow{d} N(0,1)$.

Proof of Lemma A1.

This is Lemma A2 of Carrasco and Doukali (2016). when Z and Υ are not random.

Lemma A2.

If assumptions 1-3 are satisfied then:

(i) $S_n^{-1} \sum_{i \neq j}^{n} X_i P_{ij}^{\alpha} X_j' S_n^{-1} = O_p(1)$.

(ii) $S_n^{-1} \sum_{i \neq j}^{n} X_i P_{ij}^{\alpha} \epsilon_j = O_p(1 + \frac{1}{\sqrt{\alpha} \mu_n})$.

Proof of Lemma A2.

(ii) holds by Lemma A5 of Carrasco and Doukali (2016) and (i) of Lemma A0.

Now we turn our attention to (i).

We have $S_n^{-1} \sum_{i \neq j}^{n} X_i P_{ij}^{\alpha} X_j' S_n^{-1} = \sum_{i \neq j}^{n} f_i P_{ij}^{\alpha} f_j'/n + o_p(1)$.

We also have $\sum_{i \neq j}^{n} f_i P_{ij}^{\alpha} f_j'/n = f' P^{\alpha} f/n - \sum_{i}^{n} f_i f_i' P_{ii}^{\alpha}/n$, and both $f' P^{\alpha} f/n \leq f'f/n$ and

$\sum_{i}^{n} f_i f_i' P_{ii}^{\alpha}/n \leq f'f/n$.

Lemma A3.

If $\hat{\delta} \to \delta$, $E[||X_i||^2] \leq C$, $E[\epsilon_i^4] \leq C$, $\epsilon_1, ..., \epsilon_n$ *are mutually independent, and either* $\alpha \to 0$ *or* $max_{i \leq n}$ $P_{ii}^{\alpha} \to 0$ *then:*

$\alpha \sum_{i \neq j}^{n} (P_{ij}^{\alpha})^2 \hat{\epsilon}_i^2 \hat{\epsilon}_j^2 - \alpha \sum_{i \neq j}^{n} (P_{ij}^{\alpha})^2 \sigma_i^2 \sigma_j^2 \to 0$.

Proof of Lemma A3.

By $\hat{\delta} \xrightarrow{p} \delta$ we have $||\hat{\delta} - \delta||^2 \leq ||\hat{\delta} - \delta||$ with probability one. Denote

$d_i = 2|\epsilon_i|||X_i|| + |||X_i||^2$, we have:

$$\begin{aligned}\hat{\epsilon}_i &= y_i - X_i'\hat{\delta} \\ &= X_i'\delta + \epsilon_i - X_i'\hat{\delta} \\ &= \epsilon_i - X_i'(\hat{\delta} - \delta).\end{aligned}$$

It follows that:
$\hat{\epsilon}_i^2 = \epsilon_i^2 - 2\epsilon_i X_i'(\hat{\delta} - \delta) + (\hat{\delta} - \delta)'X_i X_i'(\hat{\delta} - \delta)$.
Then:
$\hat{\epsilon}_i^2 - \epsilon_i^2 = -2\epsilon_i X_i'(\hat{\delta} - \delta) + (\hat{\delta} - \delta)'X_i X_i'(\hat{\delta} - \delta)$.
$|\hat{\epsilon}_i^2 - \epsilon_i^2| \leq 2|\epsilon X_i'(\hat{\delta} - \delta)| + |(\hat{\delta} - \delta)'X_i X_i'(\hat{\delta} - \delta)|$.
$|\hat{\epsilon}_i^2 - \epsilon_i^2| \leq 2|\epsilon_i| \; ||X_i|| \; ||\hat{\delta} - \delta|| + |||X_i||^2||\hat{\delta} - \delta||^2 \leq d_i||\hat{\delta} - \delta||$.
Also by (ii) of Lemma A0, $\sum_{i\neq j}^n (P_{ij}^\alpha)^2 = O(1/\alpha)$,
$\alpha E[\sum_{i\neq j}^n (P_{ij}^\alpha)^2 d_i d_j] \leq \alpha C \sum_{i\neq j}^n P_{ij}^2 \leq C$,
$\alpha E[\sum_{i\neq j}^n (P_{ij}^\alpha)^2 \epsilon_i^2 d_j] \leq C$.
Then by M,
$\alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 d_i d_j = O_p(1)$, $\alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 \epsilon_i^2 d_j = O_p(1)$,
Therefore, for $\hat{V}_n = \alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 \hat{\epsilon}_i^2 \hat{\epsilon}_j^2$, $\tilde{V}_n = \alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 \epsilon_i^2 \epsilon_j^2$, we have
$|\hat{V}_n - \tilde{V}_n| \leq \alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 |\hat{\epsilon}_i^2 \hat{\epsilon}_j^2 - \epsilon_i^2 \epsilon_j^2|$
$|\hat{V}_n - \tilde{V}_n| \leq \alpha ||\hat{\delta} - \delta||^2 \sum_{i\neq j}^n (P_{ij}^\alpha)^2 d_i d_j + 2\alpha ||\hat{\delta} - \delta|| \sum_{i\neq j}^n (P_{ij}^\alpha)^2 \epsilon_i^2 d_j \to 0$.
Let $V_n = \alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 \sigma_i^2 \sigma_j^2$ and $v_i = \epsilon_i^2 - \sigma_i^2$. We have:
$\sum_{i\neq j}^n (P_{ij}^\alpha)^2 \epsilon_i^2 \epsilon_j^2 - \sum_{i\neq j}^n (P_{ij}^\alpha)^2 \sigma_i^2 \sigma_j^2 = 2\sum_{i\neq j}^n (P_{ij}^\alpha)^2 v_i \sigma_j^2 + \sum_{i\neq j}^n (P_{ij}^\alpha)^2 v_i v_j$.
We note that $E[v_i^2] \leq E[\epsilon_i^4] \leq C$, so we have:
$E[(\alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 v_i \sigma_j^2)^2] = \alpha^2 \sum_i \sum_{i\neq j} \sum_{k\neq i} (P_{ij}^\alpha)^2 (P_{ik}^\alpha)^2 E[v_i^2] \sigma_i^2 \sigma_k^2$
$E[(\alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 v_i \sigma_j^2)^2] \leq C\alpha^2 \sum_i \sum_j (P_{ij}^\alpha)^2 \sum_k (P_{ik}^\alpha)^2$
We note that $P_{ij}^\alpha = P_{ji}^\alpha$, and $\sum_i \sum_j (P_{ij}^\alpha)^2 \sum_k (P_{ik}^\alpha)^2 = O(1/\alpha)$ by LemmaA0 (vi). So:
$E[(\alpha \sum_{i\neq j}^n (P_{ij}^\alpha)^2 v_i \sigma_j^2)^2] = C\alpha \to 0$.
Also by CS, $max_{ij}(P_{ij}^\alpha)^2 \leq max_{ii}(P_{ii}^\alpha)^2$, so that:

$E[(\alpha \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^2 v_i v_j)^2] = 2\alpha^2 \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^4 E[v_i^2]E[v_j^2] \leq C\alpha^2 \sum_{i\neq j}^{n} (P_{ij}^{\alpha})^4 \leq C\alpha^2 O(1/\alpha) \to 0$. Because of (v) of Lemma A0.
Then by T and M we have $\tilde{V}_n - V_n \xrightarrow{\text{p}} 0$. The conclusion then follows by T.

Proof of Theorem 1.

$$\begin{aligned}
\sqrt{\alpha}\sum_{i\neq j}^{n} \hat{\epsilon}_i P_{ij}^{\alpha} \hat{\epsilon}_j &= \sqrt{\alpha}\sum_{i\neq j}^{n}[\epsilon_i - X_i'(\hat{\delta}-\delta)]P_{ij}^{\alpha}[\epsilon_j - X_j'(\hat{\delta}-\delta)] \\
&= \sqrt{\alpha}\sum_{i\neq j}^{n} \epsilon_i P_{ij}^{\alpha}\epsilon_j + (\hat{\delta}-\delta)'S_n \times \sqrt{\alpha}[S_n^{-1}\sum_{i\neq j}^{n} X_i P_{ij}^{\alpha} X_j' S_n'^{-1}]S_n'(\hat{\delta}-\delta) \\
&+ 2\sqrt{\alpha}(\hat{\delta}-\delta)'S_n[S_n^{-1}\sum_{i\neq j}^{n} X_i P_{ij}^{\alpha}\epsilon_j]
\end{aligned}$$

If $1/(\alpha\mu_n^2) \to C < \infty$, then by Theorem 2 of Carrasco and Doukali (2016) we have $S_n'(\hat{\delta}-\delta) = O_p(1)$. Then by LemmaA2 we have:

$$\sqrt{\alpha}\sum_{i\neq j}^{n} \hat{\epsilon}_i P_{ij}^{\alpha} \hat{\epsilon}_j = \sqrt{\alpha}\sum_{i\neq j}^{n} \epsilon_i P_{ij}^{\alpha}\epsilon_j + o_p(1)$$

Next, note that $\sigma_i^2 \geq C$ by assumption 3 and $P_{ii}^{\alpha} \leq C < 1$ by Assumption 1 so that:

$$\begin{aligned}
V_n &= \alpha\sum_{i\neq j}^{n} \sigma_i^2 (P_{ij}^{\alpha})^2 \sigma_j^2 > C(\alpha\sum_{i,j}^{n} (P_{ij}^{\alpha})^2 - \sum_{i}^{n} (P_{ii}^{\alpha})^2) \\
&= C\alpha\sum_{i}^{n} P_{ii}^{\alpha}(1-P_{ii}^{\alpha}) > C(1-C) > 0.
\end{aligned}$$

Moreover, $E[\epsilon_i^4] \leq C$ and,

$$E[\sum_{i\neq j}^{n}(\epsilon_i P_{ij}^{\alpha}\epsilon_j)^2] = E[\sum_{i\neq j}\sum_{k\in\{i,j\}} P_{ik}^{\alpha}P_{jk}^{\alpha}\epsilon_i\epsilon_j'\epsilon_k^2 + \sum_{i\neq j}^{n} P_{ij}^{\alpha}\epsilon_i^2\epsilon_j^2$$
$$= E[2\sum_{i\neq j}^{n}(P_{ij}^{\alpha}\epsilon_i^2\epsilon_j^2)] = 2\sum_{i\neq j}^{n} P_{ij}^{\alpha}\sigma_i^2\sigma_j^2 = 2tr(P^{\alpha})V_n$$

It follows from Lemma A1 with $W_{in} = 0$, $c_{1n} = 0$, $c_{2n} = 1$, $u_i = \epsilon_i$ that : $\frac{\sum_{i\neq j}^{n}\epsilon_i P_{ij}^{\alpha}\epsilon_j}{\sqrt{2tr(P^{\alpha})V_n}} \xrightarrow{\text{d}} N(0,1)$.

Next by Theorem 1 Carrasco and Doukali (2016) we have $\hat{\delta} \xrightarrow{\text{p}} \delta$. Moreover by Lemma A0 (iii), $tr(P^{\alpha}) = O(\frac{1}{\alpha})$. Hence, by Lemma A3, $\hat{V}_n - V_n \xrightarrow{\text{p}} 0$. Then by V_n bounded and bounded away from zero, $\sqrt{\frac{V_n}{\hat{V}_n}} \to 1$. Therefore by slutsky theorem,

$\frac{\sum_{i\neq j}^{n}\hat{\epsilon}_i P_{ij}^{\alpha}\hat{\epsilon}_j}{\sqrt{2tr(P^{\alpha})\hat{V}_n}} = \frac{\sum_{i\neq j}^{n}\epsilon_i P_{ij}^{\alpha}\epsilon_j}{\sqrt{2tr(P^{\alpha})\hat{V}_n}} + \frac{o_p(1)}{2\hat{V}_n} = \sqrt{\frac{V_n}{\hat{V}_n}}\frac{\sum_{i\neq j}^{n}\epsilon_i P_{ij}^{\alpha}\epsilon_j}{\sqrt{2tr(P^{\alpha})V_n}} + o_p(1) \xrightarrow{\text{d}} N(0,1)$

Next note that $\hat{T} \geq q_{(tr(P^{\alpha})-p)}(1-\beta)$ if and only if

$\frac{\sum_{i\neq j}^{n}\hat{\epsilon}_i P_{ij}^{\alpha}\hat{\epsilon}_j}{\sqrt{2tr(P^{\alpha})\hat{V}_n}} \geq \frac{q_{(tr(P^{\alpha})-p}(1-\beta)-tr(P^{\alpha})}{\sqrt{2tr(P^{\alpha})}}$

Using the fact that $tr(P^{\alpha}) = O(\frac{1}{\alpha})$, we have, as $\alpha \to 0$, $q_{(tr(P^{\alpha})-p)}(1-\beta) - (tr(P^{\alpha})-p)/\sqrt{2(tr(P^{\alpha})-p)} \to q(1-\beta)$ where $q(1-\beta)$ is the $1-\beta$ quantile of the standard normal distribution, also, we have

$= \sqrt{\frac{(tr(P^{\alpha}))-p}{tr(P^{\alpha})}}(\frac{q_{(tr(P^{\alpha}))-p}(1-\beta)-(tr(P^{\alpha})-p)}{\sqrt{2tr(P^{\alpha})-p}}) - \frac{p}{\sqrt{2tr(P^{\alpha})}} \to q(1-\beta)$.

The conclusion now follows.

Chapter 3

Jackknife LIML estimator with many instruments using regularization techniques

3.1 Introduction

This paper considers an estimation of a finite dimensional parameter in a linear model in the presence of many weak instruments and heteroskedastic data[1]. The use of many instruments is useful in practice in the sense that it permits one to improve the efficiency of instrumental variables (IV) estimates. However, the cost of the improvement in efficiency is that the IV estimators, including the two-stage least square (2SLS), may suffer from a substantial bias when the number of moment conditions or instruments is very large; see among others Newey and Smith (2004) and Bekker (1994).
This paper proposes new estimators for IV models based on a Jackknife regularized version of the Limited Information Maximum Likelihood estimator (LIML) when the number of instruments, L, is not restricted and may be smaller or larger than the sample size, n, or even infinite. Our proposed estimators are robust to heteroskedasticity because of their Jackknife form. The main innovation of our approach is the use of regularization methods at each iteration of the Jackknife to reduce the bias in the many-instruments setting. The regularization permits one to solve the problem of the singularity of the covariance matrix resulting from many instruments (see Carrasco, Florens, and Renault (2007) and Kress (1999) for a review of regularization schemes). The first estimator is based on the Tikhonov (T) method, also called the ridge regularization, the second estimator is based on an iterative method called Landweber-Fridman (LF), and the third estimator is based on the principal components (PC) associated with the largest eigenvalues. In this paper, we focus on the asymptotic theory that allows for both, the many strong instruments, as in Kress (1999) and Kunitomo (1980), and the many weak instruments, as in Chao and Swanson (2005), Stock and Yogo (2005), and Han and Phillips (2006). We show that the regularized JLIML estimators

[1] I am much indebted to Marine Carrasco for intensive comments and advices.

are consistent and asymptotically normal under the heteroskedastic assumption. In the homoskedastic case, our estimators reach the semiparametric efficiency bound of Chamberlain (1992). The regularization methods used in our paper involve a regularization parameter that needs to be selected. We propose a data-driven method was developed to select this regularization parameter. We use this method in our simulations to select the regularization parameter.

Our Monte Carlo study shows that the regularized JLIML (T and LF JLIML) performs well (i.e, are nearly median unbiased) compared to other competing IV estimators in the literature.

Our work complements existing approaches to providing efficient estimation that is robust to many possibly weak instruments when heteroskedasticity is present. Hausman, Newey, Woutersen, Chao, and Swanson (2012) propose a Jackknife version of the LIML and the Fuller (1977)' estimator where the number of instruments, L, is smaller than the sample size n (i.e, $\frac{L}{n} \to \rho$ with $1 \leq \rho \leq 1$). Newey and Windmeijer (2009) show that the continuously updated estimator (CUE) is consistent and asymptotically efficient compared to others IV estimators, but L and n are needed to satisfy $\frac{L^2}{n} \to C$ (C is a constant) for consistency and $\frac{L^3}{n} \to 0$ for asymptotic normality. They propose a new variance estimator that is consistent under many weak-moment conditions. However, this variance depends on using a heteroskedasticity-consistent weighting matrix that can degrade the finite sample performance of CUE with many instruments. In a recent paper, Carrasco and Tchuente (2016b) propose a regularized version of the LIML estimator with many weak instruments; their proposed estimator is consistent and asymptotically normally distributed, but they assume that structured errors are allowed to be only homoskedastic. Regularization also has been introduced in the context of forecasting using a large data set of predictors. Carrasco and Rossi (2016) propose several dimension-reductions devices: principal components, Ridge,

LF, and partial least squares to improve in-sample prediction and out-of-sample forecasting in regressions with many exogenous predictors. Their methods differ from those used in traditional factor models (see Bai and Ng (2002) and Stock and Watson (2002)) in the sense that they use all the available predictors instead of summarizing the information from the large data set of predictors into a low-dimensional vector of latent factors.
The rest of the paper is organized as follows. In Section 2, we present the model and the three regularized Jackknife LIML estimators. Section 3 derives the asymptotic properties of our proposed estimators. Section 4 presents a data-driven method for selecting the tuning parameter involved in the regularization methods we consider. Section 5 reports the Monte Carlo results. An empirical application is illustrated in Section 6. Section 7 concludes. All proofs are collected in Appendix.

3.2 Presentation of the regularized Jackknife LIML estimators

In this section, we present the regularized JLIML estimators. We show that our proposed JLIML estimators are consistent and asymptotically normal in the presence of heteroskedastic error and weak instruments and they reach the semiparametric efficiency bound under the homoskedasticity assumption.
Consider the linear IV regression model:

$$y_i = X_i'\delta_0 + \epsilon_i \tag{3.2.1}$$

$$X_i = \Upsilon_i + u_i \tag{3.2.2}$$

$i = 1, \ldots, n$. The vector of interest is δ_0, which is a $p \times 1$ vector. y_i is a scalar. The vector Υ_i is the optimal instrument, which is typically unknown.

We assume that y_i and X_i are observed but the Υ_i is not. $E(X_i\epsilon_i) \neq 0$, as a result, X_i is endogenous and the OLS estimator of δ_0 is not consistent. The estimation will be based on a sequence of instruments, $Z_i = Z(\tau; \nu_i)$, where ν_i is a vector of exogenous variables and τ is an index taking countable values. Such a situation can arise by taking interactions between some exogenous variables, as in Angrist and Krueger (1991), by non-linear transformations of an exogenous variable, as in Dagenais and Dagenais (1997), or also by allowing lagged dependent variables, as in Arellano and Bond (1991).

Assumption 1. *y_i, X_i and ν_i are iid, $E(u_i|\nu_i) = E(\epsilon_i|\nu_i) = 0$; $\Upsilon_i = E(X_i|\nu_i)$ denote the $p \times 1$ reduced-form vector with $E(\Upsilon_i\epsilon_i) = 0$.*

The estimation will be based on a sequence of instruments, $Z_i = Z(\tau; \nu_i)$; τ may be an integer or take its values in an interval. Thus, the model allows that the number of moments conditions is finite, or countable infinite. Here some examples of Z_i.

- If $Z_i = \nu_i$, where ν_i is an L-vector of exogenous variables with a fixed L, then $Z(\tau; \nu_i)$ denotes the τth element of ν_i.
- $Z(\tau; \nu_i) = (\nu_i)^{\tau-1}$ with $\tau \in N$; thus, we have an infinite countable sequence of instruments.

We will present the JLIML estimators in the case where, Z_i is an $L \times 1$ vector of instruments, where L may be a larger or smaller integer than the sample size n.

The estimation of δ is based on the orthogonality condition.

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

where Z_i is a $L \times 1$ vector of instruments. Let Z denote the $n \times L$ matrix having rows corresponding to Z_i'. Denote ψ_j the eigenvectors of the $n \times n$ matrix, ZZ'/n, associated with eigenvalues λ_j. Recall the expression of the

Jackknife k-class estimators:

$$\hat{\delta} = (X'PX - \sum_{i=1}^{n} P_{ii}X_iX_i' - \alpha X'X)^{-1}(X'Py - \sum_{i=1}^{n} P_{ii}X_iy_i - \alpha X'y)$$

where α is either a constant term or a random variable. The case $\alpha = 0$ corresponds to the Jackknife IV estimator (RJIVE) studied in Chao et al. (2012a), and the case $\hat{\alpha} = \underset{\delta}{min}\frac{\sum_{i\neq j}^{n}(y_i - X_i'\delta)'P_{ij}(y_j - X_j'\delta)}{(y-X\delta)'(y-X\delta)}$ corresponds to HLIML estimator studied in Hausman et al. (2012). We note that those estimators involve the projection matrix $P = Z(Z'Z)^{-1}Z$. When the number of instruments, L, is large relatively to n, inverting the matrix, $Z'Z$, is considered as an ill-posed problem, which means that $Z'Z$ is singular or nearly singular. To address this issue, we propose a regularized version of the inverse of the matrix $Z'Z$. We apply the same regularizations methods [2] as in Carrasco (2012). For an arbitrary $n \times 1$ vector, d, we define the $n \times n$ matrix, P^r, as

$$P^r d = \frac{1}{n}\sum_{j=1}^{n} q\left(r, \lambda_j^2\right)\langle d, \psi_j\rangle \psi_j$$

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

- $q\left(r, \lambda_j^2\right) = \lambda_j^2 / \left(r + \lambda_j^2\right)$ for T regularization,
- $q\left(r, \lambda_j^2\right) = I\left(\lambda_j^2 \geq r\right)$ for PC regularization,
- $q\left(r, \lambda_j^2\right) = 1 - \left(1 - c\lambda_j^2\right)^{1/r}$ for LF regularization.

We note here that all the regularizations techniques involve a tuning parameter r. The case $r = 0$ corresponds to the case without regularization. We

[2] Appendix A gives a detailed definition of the regularization methods and the definition of P^r.

obtain:

$$P^0 = P = Z(Z'Z)^{-1}Z$$

Consider now the regularized Jackknife k-class estimators defined as follows:

$$\hat{\delta} = (X'P^rX - \sum_{i=1}^{n} P^r_{ii}X_iX_i' - \alpha X'X)^{-1}(X'P^ry - \sum_{i=1}^{n} P^r_{ii}X_iy_i - \alpha X'y)$$

The case $\alpha = 0$ corresponds to the regularized Jackknife 2SLS estimator (RJIVE) studied in Carrasco and Doukali (2016), and the case $\hat{\alpha} = \underset{\delta}{min}\frac{\sum_{i\neq j}^{n}(y_i - X_i'\delta)'P^r_{ij}(y_j - X_j'\delta)}{(y-X\delta)'(y-X\delta)}$ corresponds to our proposed regularized Jackknife estimator JLIML, which we will study here.
The 2SLS estimator suffers from a small-sample bias in the presence of endogeneity that is increased dramatically when many instruments are used and/or the instruments are only weakly correlated with the endogenous variables. The LIML was proposed to correct the bias problem of the 2SLS estimator in the presence of many instruments. It was shown in the literature that the LIML has better small-sample properties than the 2SLS estimator. A drawback of the LIML estimator is that it is not consistent under heteroskedasticity. Hausman et al. (2012) propose the HLIML estimator, which is based on the Jackknife version of the LIML estimator. Their proposed estimator is robust to heteroskedasticity and many instruments because of the Jackknife form. However, they assume that the number of instruments is smaller than the sample size $\frac{L}{n} < 1$. We contribute to this literature by considering cases where $L > n$ and regularization. The advantage of regularization is that all available instruments can be used without discarding any a priori.

3.3 Asymptotic Properties of the regularized Jackknife LIML

In this section, we establish the asymptotic properties of the regularized JLIML estimator when the errors are heteroskedastic. We also allow for the presence of many weak instruments, as in Chao and Swanson (2005). A measure of the strength of the instruments is the concentration parameter, which can be seen as a measure of the information contained in the instruments. If one could approximate the reduced form, Υ, by a sequence of instruments, Z, so that $X = Z'\pi + u$, where $E[u^2|Z] = \sigma_u^2$, then the concentration parameter would be given by:

$$CP = \frac{\pi' Z' Z \pi}{\sigma_u^2}.$$

The following assumption generalizes this notion.

Assumption 2. $\Upsilon_i = S_n f_i/\sqrt{n}$, *where* $S_n = \hat{S}_n diag(\mu_{1n}, \ldots, \mu_{pn})$ *such that* $\hat{S}_n$ *is a* $p \times p$ *bounded matrix, the smallest eigenvalue of* $\hat{S}_n \hat{S}_n'$ *is bounded away from 0; for each j, either* $\mu_{jn} = \sqrt{n}$ *(strong identification) or* $\frac{\mu_{jn}}{\sqrt{n}} \to 0$ *(weak identification). Moreover,* $\mu_n = \underset{1<j<p}{min}\, \mu_{jn} \to \infty$ *and* $1/(\sqrt{r}\mu_n^2) \to 0$, $r \to 0$. *Also there is a constant,* $\bar{C}$, *such that* $|| \sum_{i=1}^n f_i f_i'/n|| \leq \bar{C}$ *and* $\lambda_{min}(\sum_{i=1}^n f_i f_i'/n) \geq 1/\bar{C}$, *a.s.n.*

This condition is similar to Assumption 2 of Hausman et al. (2012). It allows for both strong and weak instruments. The concentration parameter, μ_n^2, will determine the convergence rate of the estimator $\hat{\delta}$. If $\mu_{jn} = \sqrt{n}$, the instrument is strong and the convergence rate will be $\sqrt{n}$. If μ_{jn}^2 is growing slower than n, the convergence rate will be slower than $1/\sqrt{n}$, leading to a weak identification, as in Chao and Swanson (2005). f_i, defined in Assumption 2, is unobserved and has the same dimension as the infeasible optimal instrument, Υ_i. Then, f_i can be seen as a rescaled version of this optimal instrument.

An illustration of Assumption 2 is as follows. Let us consider the simple linear model; $y_i = z_{i1}\delta_1 + \delta_{0p}x_{i2} + \epsilon_i$, where z_{i1} is an included instruments and x_{i2} is an endogenous variable. Suppose that x_{i2} is a linear combination of the included instrumental, z_{i1}, and an unknown excluded instruments z_{ip}, i.e $x_{i2} = \pi_1 z_{i1} + (\frac{\mu_n}{\sqrt{n}})z_{ip}$. The reduced form is:

$$\Upsilon_i = \begin{pmatrix} z_{i1} \\ x_{i2} \end{pmatrix} = \begin{pmatrix} z_{i1} \\ \pi_1 z_{i1} + (\frac{\mu_n}{\sqrt{n}})z_{ip} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \pi_1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{\mu_n}{\sqrt{n}} \end{pmatrix} \begin{pmatrix} z_{i1} \\ z_{ip} \end{pmatrix}$$

with

$$\hat{S}_n = \begin{pmatrix} 1 & 0 \\ \pi_1 & 1 \end{pmatrix}, \ \mu_{jn} = \begin{cases} \sqrt{n} & , \ j = 1 \\ \mu_n & , \ j = 2 \end{cases}, \text{ with } \frac{\mu_n}{\sqrt{n}} \to 0, \text{ and}$$

$$f_i = \nu_i = \begin{pmatrix} z_{i1} \\ z_{ip} \end{pmatrix}.$$

Assumption 3. *There is a constant, $\bar{C}$, such that the observations $(\epsilon_1, u_1), \ldots, (\epsilon_n, u_n)$ are independent, with $E[\epsilon_i] = 0$ for all i, $E[u_i] = 0$ for all i, $E[\epsilon_i^2] \leq \bar{C}$, and $E[||u_i||^2] \leq \bar{C}$, $Var((\epsilon_i, u_i')') = diag(\Omega_i, 0)$, and $\lambda_{min}(\sum_{i=1}^{n} \Omega_i/n) \geq 1/\bar{C}$.*

This assumption is similar to Assumption 3 in Hausman et al. (2012). It requires the second conditional moments of the disturbances to be bounded. This condition also imposes the uniform nonsingularity of the variance of the reduced-form disturbances, which will permit us to prove the consistency of the proposed estimators.

Assumption 4. *(i) The operator, K,* [3] *is nuclear. (ii)* Υ_a *(the ath row of* Υ*) belongs to the closure of the linear span of* $\{Z(.;\nu)\}$ *for* $a = 1, .., p$. *(iii) There exists a constant,* $\bar{C}$*, such that* $P^r_{ii} \leq \bar{C} < 1$, $i = 1, ..., n$.

Assumption 4 is the same as in Carrasco (2012). Condition (i) provides that the smallest eigenvalues of the covariance operator, K, decreases to 0 sufficiently fast. Condition (ii) implies that the optimal instrument, f, can be approached by a sequence of instruments. In a finite case, this condition is equivalent to saying that f_i can be approached by a linear combination of the instruments, where $Z(\nu)^L$ is a subset of the instruments. Thus, there exists a π_L such that $\sum_{i=1}^n ||f_i - \pi_L Z(\nu)^L||^2/n \to 0$. Condition (iii) is reminiscent of Assumption 1 in Hausman et al. (2012): "For some $\bar{C} < 1$, $P_{ii} < \bar{C}$, $i = 1, ..., n$". However it is much less restrictive. Indeed, $P_{ii} < \bar{C} < 1$ implies that $\sum_i \frac{P_{ii}}{n} = \frac{L}{n} < 1$, $L = \text{rank}(Z)$, which restricts the number of instruments. Our condition, $P^r_{ii} \leq \bar{C} < 1$, implies that $\text{trace}(P^r) = \sum_i q_i < n$, which implies a condition on r. Recall from Carrasco (2012) that $\sum_i q_i = O(\frac{1}{r})$. Thus, condition (iii) implies that $\frac{1}{rn} < 1$. Conditions (i) and (ii) imply that the structural parameters are identified asymptotically.

Assumption 5. *There exist a constant* $\bar{C}$, $\bar{C} > 0$ *such that* $\sum_{i=1}^n ||f_i||^4/n \to 0$, $sup_i E[\epsilon_i^4] \leq \bar{C}$, *and* $sup_i E[||u_i||^4] \leq \bar{C}$.

Assumption 5 is a standard condition that assumes that fourth moments are bounded.

Theorem 1. *Suppose that Assumptions 1-4 are satisfied. The T, LF, and PC regularized JLIML estimators satisfy* $\mu_n^{-1} S_n'(\hat{\delta} - \delta_0) \xrightarrow{p} 0$*; as* n, μ_n *go to infinity, and* r *goes to 0.*

Theorem 1 implies the consistency of the estimator, $\hat{\delta} \xrightarrow{p} \delta$. Let us now state the asymptotic normality of the regularized estimators. First, let:

[3] See Appendix A for the definition of the operator K.

$\sigma_i^2 = E[\epsilon_i^2]$ and $\gamma_n = \sum_i^n E[u_i\epsilon_i]/\sum_i^n \sigma_i^2$, $\tilde{u} = u - \epsilon\gamma_n'$, and $\tilde{\Omega}_i = E[\tilde{u}_i\tilde{u}_i']$, $H_n = lim \sum_i^n (1-P_{ii}^r) f_i f_i'/n$, $\Sigma_n = lim \sum_i^n (1-P_{ii}^r)^2 f_i f_i' \sigma_i^2/n$, and $\Psi_n = r\sum_{i\neq j}^n (P_{ij}^r)^2 (E[u_j u_j']\sigma_i^2 + E[\tilde{u}_i\epsilon_i]E[\tilde{u}_j\epsilon_j])$.
Hausman et al. (2012) separate 2 cases: $\frac{L}{\mu_n^2} \to C$, for some constant C, and the case $\frac{L}{\mu_n^2} \to \infty$, and gave the asymptotic variances for both cases (see Hausman et al. (2012) (page 222)):

$$\begin{aligned} A &= H_n^{-1}\Sigma_n H_n^{-1} + \alpha H_n^{-1} S_0 \Psi_n S_0' H_n^{-1} \text{ in case 1,} \\ A &= H_n^{-1} S_0 \Psi_n S_0' H_n^{-1} \text{ in case 2.} \end{aligned}$$

In this paper, we consider a third case where $\frac{1}{r\mu_n^2} \to 0$. r is the regularization parameter that can be controlled such that Ψ_n vanishes asymptotically. Instead of restricting the number of instruments (which may be very large or infinite), we impose restrictions on the regularization parameter such that it goes to zero at a slower rate than μ_m^2. This insures us that all available and valid instruments are used in an efficient way even if they are weak.
The asymptotic variance of our proposed estimator is given by:

$$V = H_n^{-1}\Sigma_n H_n^{-1}.$$

Theorem 2. *Suppose that Assumptions 1-5 are satisfied and n, μ_n, and $r\mu_n^2$ go to infinity. The T, Lf, and PC regularized JLIML estimators satisfy:*

$$V^{-1/2} S_n'(\hat{\delta} - \delta_0) \xrightarrow{d} N(0, I_p).$$

Theorem 2 states the asymptotic normality of our proposed regularized estimators. The asymptotic variance of those estimators is smaller than that obtained in Hausman et al. (2012). It would be interesting to compare the asymptotic variance of the regularized JLIML when the errors are homoskedastic, $E[\epsilon_i^2] = \sigma_\epsilon^2$. With many weak instruments, where $max_i P_{ii}^r \to 0$, we will have $H_n = lim \sum_i^n (1-P_{ii}^r) f_i f_i'/n = lim \sum_i^n f_i f_i'/n = E(f_i f_i')$, and $\Sigma_n = lim \sum_i^n (1-P_{ii}^r)^2 f_i f_i' \sigma_i^2/n = \sigma_\epsilon^2 E(f_i f_i')$, so the asymptotic variance of the regularized JLIML is equal to $\sigma_\epsilon^2\ [E(f_i f_i')]^{-1}$, which corresponds to the semiparametric efficiency bound of Chamberlain (1992). Our asymptotic variance is more efficient than the one obtained by Hausman et al. (2012). They use the number of instruments as a regularization parameter, that is why they obtain a larger asymptotic variance than we did. Moreover, we assume that the set of instruments is sufficiently rich to span the optimal instrument (Assumption 4 (ii)). In the Monte Carlo simulation, we find the relevance of our leading estimators in finite samples (they are nearly median unbiased).

Related Literature.

The problem of many (possibly) weak instruments in heteroskedastic data is a growing part of the econometric literature. While the 2SLS estimator is inconsistent in this framework, others IV estimators are shown to be consistent and asymptotically normal, like HFUL estimator, and continuous updating generalized estimator (CUE); see Angrist, Imbens, and Krueger (1999), Newey and Windmeijer (2009), Chao and Swanson (2005), Chao et al. (2012a), and Hausman et al. (2012). A drawback of all those papers is that they assume the number of instruments, L, should be smaller than the sample size n. We contribute to this literature by studying the case when $L > n$. In fact, when L is very large, the inverse of the covariance matrix of the instruments, $Z'Z$, needs to be regularized because it is singular. The most classical method to regularize the matrix, $Z'Z$, is to reduce its dimension by using variable selection techniques; see Bai and Ng (2008) and Gautier and Tsybakov (2011). Carrasco (2012) proposed a variety of regularization schemes to directly regularize the inverse $Z'Z$. Those regularization procedures involve a regularization parameter r. The main difference, between our work and the other papers, is that we are not making any restrictions on L to get the asymptotic properties of our proposed estimators. The need for the Jackknife form is motivated by the inconsistency of the

IV estimators in the many instruments situation when heteroskedasticity is present. To deal with this issue, Chao et al. (2012a) introduced Jackknife instrumental variables (JIV) estimators under many possibly weak instruments. In another paper, Hausman et al. (2012) propose the Jackknife version of FULL and LIML referred to as HFUL and HLIML respectively, they derive the asymptotic normal distributions for both estimators, and show the dominance of HFUL and HLIM on JIV in the sense that JIV is less efficient. Again, a common condition in those papers is that L is not allowed to be larger than n which is not the case in our work.

3.4 Selection of the regularization parameter

The three regularization schemes involve a regularization (or tuning) parameter r. An important practical issue is how to choose r. First, it is useful to write the regularized JLIML as:

$$\hat{\delta} = (X'C^rX - \hat{\alpha}X'X)^{-1}(X'C^ry - \hat{\alpha}X'y)$$

where $C^r = (C^r_{ij}) = \begin{cases} P^r_{ij} \quad if \quad i \neq j \\ C^r_{ii} = 0 \quad if \quad i = j \end{cases}$.

We propose to select r that minimizes the generalized cross-validation equation (GCV) of Li (1987). We look for the best approximation of the first stage equation: $X_j = f + u$, $j = 1, .., p$. To reduce the problem to a single equation, we take $X_\lambda = \lambda' X_j$, where λ is some given vector, for instance

$\lambda = (1, ..., 1)$.

$$\hat{r} = arg \min_{r \in A_n} \frac{1}{n} \frac{||X_\lambda - \hat{X}_\lambda||^2}{(1 - \frac{1}{n} tr(M_n(r)))}$$

where $\hat{X}_\lambda = M_n(r)\ X_\lambda$, $M_n(r) = (C^r - \hat{\alpha} I_n)$, and A_n is the set within r is selected.
Note that $tr(M_n(r)) = tr(C^r) - tr(\hat{\alpha} I_n) = -\hat{\alpha} n$ does not depend on r, so the regularization parameter r is selected by minimizing the following criteria:

$$\hat{R}(r) = \frac{1}{n} ||X_\lambda - (C^r - \hat{\alpha} I_n) X_\lambda||^2$$

Note also that, because the trace of C^r is equal to zero, generalized cross-validation (GCV) validation, and Cp cross-validation coincide.

3.5 Simulation study

In this Monte Carlo simulation, our aim is to illustrate the performance of our proposed estimators and provide a comparison to the regularized JIVE estimators proposed in Carrasco and Doukali (2016).
The data generating process (DGP) is given by:

$$y_i = X_i' \delta + \epsilon_i \tag{3.5.1}$$

$$X_i = \nu_i' \pi + u_i \tag{3.5.2}$$

$i = 1, \dots, n$. $\delta = 0.1$, $\nu_i \overset{iid}{\sim} N(0, I_L)$, and $(\epsilon_i, u_i) \overset{iid}{\sim} N(0, \sum)$ with

$$\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

For the purpose of comparison, we are going to consider two models.

Model 1 (Linear model).

We set $\pi_l = \sqrt{\frac{R_f^2}{L(1-R_f^2)}}$, l= 1,2,3,.., L with $R_f^2 = 0.1$. In this case, all the instruments have the same weight, so there is no reason to prefer one instrument over another. The ν_i are used as instruments so that $Z_i = \nu_i$.

Model 2 (weak instruments).

We keep the same DGP and we construct π such as: $\pi = \sqrt{\frac{CP}{nL}}\iota_L$, where ι_L is an L-vector of ones, and CP is the concentration parameter (CP=35) and L is the number of instruments equals to 15, 30 and 50.

The simulations are performed using 1000 replications of samples of size $n = 500$. Our proposed estimators depend on a regularization (smoothing) parameter r [4] that needs to be chosen. We select the regularization parameter by minimizing the generalized cross-validation equation (GCV) as described in Section 3.4.

For comparison, we also report the RJIVE estimators of Carrasco and Doukali (2016). In their paper, they proposed a data-driven methods for selecting the regularization parameter r based on an expansion of the MSE to estimate the RJIVE estimator. We use those data-driven methods for the choice of r.

We report summary statistics for each of the following estimators: Carrasco and Doukali (2016)' regularized Jackknife JIVE, Tikhonov (TJack), Landweber-Fridman (LJack), principal component (PJack), Tikhonov Jackknife LIML (TJLIML), Landweber-Fridman Jackknife LIML (LJLIML) and

[4]The regularization parameter r 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.

principal component Jackknife LIML (PJLIML).

We report the median bias (M.bias), the median of the absolute deviations of the estimator from the true value (M.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. To construct the confidence intervals to compute the coverage probabilities, we used the following estimate of asymptotic variance:

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

where $\hat{X} = (C^r - \hat{\alpha}I_n)X$.

Results on Model 1 are summarized in Table 3.1. We remark that, when $L = 15$, the regularized JLIML estimators perform better than the regularized JIVE estimators in terms of the the median bias. We can also notice that, when the number of instruments increases, our regularized JLIML estimators has similar performance as the regularized JIVE. Within the regularized estimators, T and LF perform better than the PC method.

Now, we turn to Model 2 which allows for the presence of weak instruments. We see that, when the number of instruments is small, $L = 15$, the bias of the regularized JLIML is quite a bit smaller than that of the regularized JIVE. However, when we increase the number of instruments ($L = 50$), there is no clear dominance among the regularized JLIML and the regularized JIVE as they all perform very well.

Table 3.1: Simulation results of model 1 with $R^2 = 0.1$, n=500

		TJack	**LJack**	**PJack**	**TJLIML**	**LJLIML**	**PJLIML**
L=15	**Med.bias**	-0.011	-0.009	0.015	-0.001	-0.001	0.014
	Med.abs	0.108	0.107	0.108	0.108	0.105	0.107
	Disp	0.423	0.417	0.408	0.396	0.384	0.369
	Cov	0.924	0.925	0.946	0.951	0.956	0.932
L=30	**Med.bias**	-0.002	-0.004	0.012	0.009	0.009	0.037
	Med.abs	0.126	0.126	0.355	0.109	0.115	0.115
	Disp	0.504	0.485	1.610	0.455	0.441	0.428
	Cov	0.962	0.961	0.891	0.947	0.953	0.914
L=50	**Med.bias**	-0.004	0.000	0.079	-0.002	0.001	0.073
	Med.abs	0.124	0.126	0.136	0.117	0.121	0.144
	Disp	0.470	0.489	0.477	0.473	0.491	0.475
	Cov	0.960	0.955	0.866	0.957	0.957	0.897

Table 3.2: Simulation results of model 2, $CP = 35$, n=500.

		TJack	**LJack**	**PJack**	**TJLIML**	**LJLIML**	**PJLIML**
L=15	**Med.bias**	-0.006	-0.013	0.033	-0.0001	-0.0001	-0.0005
	Med.abs	0.210	0.215	0.123	0.019	0.016	0.019
	Disp	0.722	0.643	0.614	0.071	0.068	0.072
	Cov	0.927	0.937	0.915	0.954	0.954	0.955
L=30	**Med.bias**	-0.020	-0.013	0.128	0.001	0.001	0.002
	Med.abs	0.216	0.181	0.198	0.019	0.019	0.018
	Disp	0.950	0.822	0.753	0.069	0.069	0.068
	Cov	0.927	0.956	0.893	0.955	0.956	0.958
L=50	**Med.bias**	-0.001	-0.001	-0.001	0.000	0.000	0.000
	Med.abs	0.017	0.017	0.017	0.017	0.018	0.017
	Disp	0.065	0.065	0.065	0.069	0.068	0.068
	Cov	0.948	0.948	0.950	0.954	0.944	0.952

3.6 Empirical application: Elasticity of intertemporal substitution (EIS).

We follow the specification in Yogo (2004) who analyzes the problem of the estimation of the EIS using the linearized Euler equation. He explains how weak instruments have been the cause of the EIS empirical puzzle. He shows that, using conventional IV methods, the estimated EIS is significantly less than 1 but its reciprocal is not different from 1.
The estimated model is as follows:

$$\Delta c_{t+1} = \tau + \psi r_{f,t+1} + \xi_{t+1}$$

$$r_{f,t+1} = \mu + \tfrac{1}{\psi}\Delta c_{t+1} + \eta_{t+1}$$

where ψ is the EIS, Δc_{t+1} is the consumption growth at time $t+1$, $r_{f,t+1}$ is the real return on a risk free asset, τ and μ are constants, and ξ_{t+1} and η_{t+1} are the innovations to consumption growth and asset return, respectively.
The instruments that Yogo (2004) used are: the twice lagged, nominal interest rate ($r_{nominal}$), inflation (i), consumption growth (c) and log dividend rate (p). We denote this bloc of instruments by $Z = [r_{nominal}, i, c, p]$. As mentioned earlier, the source of the empirical puzzle is weak instruments. We increase the number of instruments from 4 to 18 by including interactions and power functions. The 18 instruments used in our regression are derived from Z and are given by $II = [Z, Z.^2, Z.^3, Z(:,1) \star Z(:,2), Z(:,1) \star Z(:,3), Z(:,1) \star Z(:,4), Z(:,2) \star Z(:,3), Z(:,2) \star Z(:,4), Z(:,3) \star Z(:,4)]$. Finally, we note that, the instruments are standardized, which means that the instruments are divided with their standard deviation as in Carrasco and Tchuente (2016b).
Estimation results are reported in Table 3.3. Interestingly, the point estimates obtained by LF and T regularized estimators are close to those used for macro calibrations (EIS equal to 0.71 in our estimations and 0.67 in Castro, Clementi, and MacDonald (2009)).

Moreover, the results of the two equations are consistent with each other since we obtain the same value for ψ in both equations. PC seems to take too many factors, and did not perform well, this is possibly due to the lack of factor structure in the instruments.

Table 3.3: Estimates of the EIS

	LIML (4 instr)	**LIML (18 instr)**	**TLIML**	**LLIML**	**PLIML**
ψ	0.0293	0.2225	0.710	0.7104	0.150
	(0.0994)	(0.156)	(0.424)	(0.423)	(0.111)
			r =0.01	1000 iterations	Nb of PC=8
$\mathbf{1}/\psi$	34.1128	4.4952	1.407	1.4072	3.8478
	(112.7122)	(4.421)	(0.839)	(0.735)	(3.138)
			r =0.01	1000 iterations	Nb of PC=17
	JLIML (4 instr)	**JLIML (18 instr)**	**JTLIML**	**JLLIML**	**JPLIML**
ψ	0.0144	0.1591	0.7256	0.7915	0.1324
	(0.1025)	(0.1326)	(0.4528)	(0.1217)	(0.1147)
			r = 0.01	1000 iterations	Nb of PC= 8
$\mathbf{1}/\psi$	69.4234	6.2855	1.3809	1.5567	5.448
	(491.5135)	(5.2376)	(0.8624)	(0.9841)	(1.2027)
			r =0.01	1000 iterations	Nb of PC=17

$*$ NB: We report LIML and JLIML and for 4 and 18 instruments. The regularized estimators are computed for 18 instruments. The standard errors are given in parentheses.

3.7 Conclusion

This paper proposes three new estimators based on the regularized version of Jackknife LIML estimator. We considered the framework when the number of instruments is very large and errors are heteroskedastic. Theoretical results show that the three main estimators are consistent and asymptotically normal and reach the semiparametric effiency bound when errors are homoskedastic. In Monte Carlo experiments, we show that our proposed regularized estimators (LF and T of Jackknife LIML) perform well. They also perform well in the elasticity of intertemporal substitution example.
It would be interesting, for future research, to propose a method for selecting the parameter of regularization that appears in all regularization schemes. Another topic of interest is to study the asymptotic behaviour of the regularized version of the HFUL estimator proposed by Hausman, Newey, Woutersen, Chao, and Swanson (2012) in presence of a large number or a continuum of moment conditions.

3.8 Appendix

3.8.1 Presentation of the Regularization methods.

This section presents the regularization methods. These methods are the same as those used in Carrasco (2012). We use a compact notation which allows us to deal with a finite, countable infinite number of moments, or a continuum of moments. We consider the following sequence of instruments $Z_i = Z(\tau; \nu_i)$ where $\tau \in S$ may be an integer or an index taking its values in an interval and let π be a positive measure on Ξ. Let K be the covariance operator for instruments from $L^2(\pi)$ to $L^2(\pi)$ such that:

$$(Kg)(\tau) = \sum_{l=1}^{L} E(Z(\tau, \nu_i)Z(\tau_l, \nu_i))g(\tau_l)\pi(\tau_l).$$

where $L^2(\pi)$ denotes the Hilbert space of square integrable functions with respect to π. K is supposed to be a nuclear operator which means that its trace is finite. Let λ_j and ψ_j, $j = 1...$ be respectively the eigenvalues (ordered in decreasing order) and the orthogonal eigenfunctions of K. The operator can be estimated by K_n defined as:

$$\begin{aligned} &K_n : L^2(\pi) \to L^2(\pi) \\ &(K_n g)(\tau) = \sum_{l=1}^{L} \frac{1}{n} \sum_{i=1}^{n} (Z(\tau, \nu_i)Z(\tau_l, \nu_i))g(\tau_l)\pi(\tau_l). \end{aligned}$$

If the number of instruments L is large relatively to n, inverting the operator K is considered as an ill-posed problem which means that the inverse is not continuous and its sample counterpart, K_n, is singular or nearly singular. To solve this problem we need to stabilize the inverse of K_n using regularization. A regularized inverse of an operator K is defined as: $R_r : L^2(\pi) \to L^2(\pi)$ such that $lim_{r\to 0} R_r K\rho = \rho, \forall \rho \in L^2(\pi)$, where r is the regularization parameter (see Kress (1999) and Carrasco, Florens, and Renault (2007)).

Three regularization schemes.

We consider the Tikhonov regularization scheme.

[1] Tikhonov (ridge) regularization:

$$\begin{aligned}(K^r)^{-1} &= \left(K^2 + rI\right)^{-1} K. \\ (K^r)^{-1} v &= \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j^2 + r} \langle v, \psi_j \rangle \psi_j.\end{aligned}$$

where $r > 0$, $v \in R^n$ and I is the identity operator. For the asymptotic efficiency r has to go to zero at a certain rate. The Tikhonov regularization is related to ridge regularization. Ridge method was first proposed in the presence of many regressors. The aim was to stabilize the inverse of XX' by replacing XX' by $XX' + rI$. 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 implementation of the Tikhonov regularization and the selection of an appropriate ridge parameter for the first step regression helps to reduce this bias.

[2] Spectral cut-off or principal components:
It consists in selecting the eigenfunctions associated with the eigenvalues greater than some threshold.

$$(K^r)^{-1} v = \sum_{\lambda_j > r} \frac{1}{\lambda_j} \langle v, \psi_j \rangle \psi_j.$$

for some vector $v > 0$.

[3] Landweber-Fridman

This method of regularization is iterative. Let $0 < c < \frac{1}{\lambda_1^2(K)}$ where $\lambda_1(K)$ is the largest eigenvalues of K. define:

$$\begin{aligned}\psi_k &= (1 - cK^2)\psi_{k-1} + cKv,\ k = 1, 2, ...1/r - 1,\\ \psi_0 &= cKv.\end{aligned}$$

where $1/r - 1$ is some positive integer. ψ_k converges to $K^{-1}v$ when the number of iterations k goes to infinity. The earlier we stop the iterations, the more stable is ψ_k. Alternatively, we have:

$$(K^r)^{-1}v = \sum_j^{\infty} \frac{1 - \left(1 - c\lambda_j^2\right)^{1/r}}{\lambda_j} \langle v, \psi_j \rangle \psi_j.$$

These three regularized inverses of K can be rewritten using a common notation as:

$$(K^r)^{-1}v = \sum_{j=1}^{\infty} \frac{q\left(r, \lambda_j^2\right)}{\lambda_j} \langle v, \psi_j \rangle \psi_j$$

where:

- $q\left(r, \lambda_j^2\right) = \lambda_j^2 / \left(r + \lambda_j^2\right)$ for Tikhonov,
- $q\left(r, \lambda_j^2\right) = I\left(\lambda_j^2 \geq r\right)$ for spectral cut-off,
- $q\left(r, \lambda_j^2\right) = 1 - \left(1 - c\lambda_j^2\right)^{1/r}$ for Landweber-Fridman.

Let $(K_n^r)^{-1}$ be the regularized inverse of K_n and P^r a $n \times n$ matrix as defined in Carrasco (2012) by $P^r = T(K_n^r)^{-1}T^*$ where $T : L^2(\pi) \to R^n$ with
$Tg = (< Z_1, g >, < Z_2, g >', ..., < Z_n, g >')'$
and $T^* : R^n \to L^2(\pi)$ with
$T^*v = \frac{1}{n}\sum_j^n Z_j v_j$
such that $K_n = T^*T$ and TT^* is a $n \times n$ matrix with typical element $\frac{<Z_i, Z_j>}{n}$.
Let $\hat{\phi}_j$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ..... \geq 0$, $j = 1, 2, ...$ be the orthonormalized eigenfunctions

and eigenvalues of K_n and ψ_j the eigenfunctions 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 $v \in R^n$, $P^r v = \sum_j^\infty q(r,\lambda_j^2) < v, \psi_j > \psi_j$ where $q(r,\lambda_j^2) = \frac{\lambda_j^2}{\lambda_j^2+r}$.
Remark that the case when $r = 0$ corresponds to no regularization Thus we have $q(r,\lambda_j^2) = 1$ and $P^0 = Z(Z'Z)^{-1}Z'$.

3.8.2 Proofs

In the following, we provide proofs of Theorems 1 and 2. The proofs follow from arguments similar to those of Hausman, Newey, Woutersen, Chao, and Swanson (2012) and Carrasco and Doukali (2016) modified to account for regularization in performing the Jackknife.

Lemma A0. *(Lemma A0 of Hausman et al. (2012)).*
If Assumption 2 is satisfied and $||S_n'(\hat{\delta} - \delta_0)/\mu_n||^2/(1 + ||\ \hat{\delta}||^2) \to 0$, *then* $||S_n'(\hat{\delta} - \delta_0)/\mu_n|| \xrightarrow{p} 0$.
We next give a result from Carrasco and Doukali (2016) that is used in the proof of consistency.
Let us define some notations that will be used in the following Lemmas. For random variables[5] W_i, Y_i, η_i. Let $\bar{w}_i = E[W_i]$, $\bar{y}_i = E[Y_i]$, $\bar{\eta}_i = E[\eta_i]$, $\tilde{W}_i = W_i - \bar{w}_i$ and $\tilde{Y}_i = Y_i - \bar{y}_i$, $\tilde{\eta}_i = \eta_i - \bar{\eta}_i$, $\bar{w}_n = E[(W_1, ..., W_n)']$, $\bar{y}_n = E[(Y_1, ..., Y_n)']$, $\bar{\mu}_W = max_{i\le n}|\bar{w}_i|$, $\bar{\mu}_Y = max_{i\le n}|\bar{y}_i|$, $\bar{\mu}_\eta = max_{i\le n}|\bar{\eta}_i|$, $\bar{\sigma}^2_{W_n} = max_{i\le n}var(W_i)^{1/2}$, $\bar{\sigma}^2_{Y_n} = max_{i\le n}var(Y_i)^{1/2}$.
Throughout, let C denote a generic positive constant that may be different in different uses and let Markov inequality denote the conditional Markov inequality, and define the norm: $||W||^2_{L_2} = \sqrt{E[W^2]}$.
Lemma A1. *(Special Case of Lemma A1 of Carrasco and Doukali (2016)).*
The pairs of scalar random variables (W_i, Y_i) *are independent across i,* P^r *is the regularized projection matrix. Then there is a constant C such that:*

[5] Note that here W_i and η_i are arbitrary scalar variables that will take various forms in the sequel.

$|| \sum_{i\neq j}^{n} P_{ij}^{r} W_i Y_j - \sum_{i\neq j}^{n} P_{ij}^{r} \bar{w}_i \bar{y}_j ||_{L_2}^2 < CB_n$
where $B_n = (1/r)\bar{\sigma}_{W_n}^2 \bar{\sigma}_{Y_n}^2 + \bar{\sigma}_{Y_n}^2 \bar{w}_n' \bar{w}_n + \bar{\sigma}_{W_n}^2 \bar{y}_n' \bar{y}_n$ *and* $\bar{w}_n$ *is defined as* $\bar{w}_n = E[(W_1, ..., W_n)']$, $\bar{y}_n = E[(Y_1, ..., Y_n)']$, $\bar{\sigma}_{W_n}^2 = max_{i\leq n} var(W_i)^{1/2}$, $\bar{\sigma}_{Y_n}^2 = max_{i\leq n} var(Y_i)^{1/2}$.

For the next result, let $\bar{S}_n = diag(\mu_n, S_n)$, $\tilde{X} = [\epsilon, X]\bar{S}_n^{-1\prime}$, and $H_n = \sum_i^n (1 - P_{ii}^r) f_i f_i'/n$.

Lemma A2.

If Assumptions 1-4 are satisfied and $1/(\sqrt{r}\mu_n^2) \to 0, then:$
$\sum_{i\neq j}^{n} \tilde{X}_i P_{ij}^r \tilde{X}_j = diag(0, H_n) + o_p(1).$

Proof.

Note that:
$\tilde{X}_i = \begin{pmatrix} \mu_n^{-1}\epsilon_i \\ S_n^{-1} X_i \end{pmatrix} = \begin{pmatrix} 0 \\ f_i/\sqrt{n} \end{pmatrix} + \begin{pmatrix} \mu_n^{-1}\epsilon_i \\ S_n^{-1} u_i \end{pmatrix}.$
Since $||S_n^{-1}|| \leq C\mu_n^{-1}$, we have $Var(\tilde{X}_{ik}) \leq C\mu_n^{-2}$ for any element $\tilde{X}_{ik}$ of $\tilde{X}_i$. Then applying Lemma A1 to each element of $\sum_{i\neq j}^{n} \tilde{X}_i P_{ij}^r \tilde{X}_j$ gives:
$\sum_{i\neq j}^{n} \tilde{X}_i P_{ij}^r \tilde{X}_j = diag(0, \sum_{i\neq j}^{n} f_i P_{ij}^r f_j'/n) + O_p(\frac{1}{\sqrt{r}\mu_n^2} + \mu_n^{-1}(\sum_i ||f_i||^2/n)^{\frac{1}{2}}) = diag(0, \sum_{i\neq j}^{n} f_i P_{ij}^r f_j'/n) + o_p(1).$
Note that:

$$\begin{aligned}
H_n - \sum_{i\neq j}^{n} f_i P_{ij}^r f_j'/n &= \sum_i f_i f_i'/n - \sum_{i\neq j} P_{ii}^r f_i f_i'/n - \sum_{i\neq j}^{n} f_i P_{ij}^r f_j'/n \\
&= f'(I - P^r) f/n \\
&= (f - Z\pi_n')'(I - Pr)(f - Z\pi_n')/n \\
&\leq (f - Z\pi_n')'(f - Z\pi_n')/n \\
&\leq I_G \sum_i ||f_i - \pi_n Z_i||^2/n \to 0.
\end{aligned}$$

The last inequality follows by $A \leq tr(A)I$ for any positive semidefinite (p.s.d) matrix A. Since this equation shows that $H_n - \sum_{i\neq j}^{n} f_i P_{ij}^r f_j'/n$ is p.s.d and is less than or equal to another p.s.d matrix that converges to zero, it follows that $\sum_{i\neq j}^{n} f_i P_{ij}^r f_j'/n = H_n + o_p(1)$. The conclusion follows by the triangle inequality.

Lemma A3. *(Lemma A3 of Hausman et al. (2012) holds with P^r replacing P).*

If Assumptions 1-4 are satisfied, then $S_n'(\hat{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0$.

Proof.

Let $\bar{Y} = [0, Y]$, $\bar{U} = [\epsilon, U]$, and $\bar{X} = [y, X]$, so that $\bar{X} = (\bar{Y} + \bar{U})D$ for

$$D = \begin{pmatrix} 1 & 0 \\ \delta_0 & I \end{pmatrix}$$

Let $\hat{B} = \bar{X}'\bar{X}/n$. Note that $||S_n/\sqrt{n}|| \leq C$ and, by standar calculations, $f'u/n \xrightarrow{\text{p}} 0$. Then

$||\bar{Y}\bar{U}/n|| = ||(S_n/\sqrt{n})f'u/n \leq C||f'u/n|| \xrightarrow{\text{p}} 0$.

Let $\bar{\Omega}_n = \sum_i E[\bar{U}_i\bar{U}_i']/n = diag(I_{G_2+1}, 0)$ by Assumption 3, where $G_2 + 1$ is the dimension of the number of included endogenous variables. By the Markov inequality, we have $\bar{U}'\bar{U}/n - \bar{\Omega}_n \rightarrow 0$, so it follows that:

$\hat{B} = (\bar{U}'\bar{U} + \bar{Y}'\bar{U} + \bar{U}'\bar{Y} + \bar{Y}'\bar{Y})/n = \bar{\Omega}_n + \bar{Y}'\bar{Y}/n + o_p(1) \geq Cdiag(I_{G-G_2+1}, 0)$

Since $\Omega_n + \bar{Y}'\bar{Y}/n$ is bounded, it follows that ,

$C \leq (1, -\delta')\hat{B}(1, -\delta')' = (y - X\delta)'(y - X\delta)/n \leq C||(1, -\delta')||^2 = C(1 + ||\delta||^2)$.

Next, as defined preceding Lemma A2, let $\bar{S}_n = diag(\mu_n, S_n)$ and $\tilde{X} = [\epsilon, X]\bar{S}_n^{-1}$.

Note that by $P_{ii}^r \leq C < 1$ and uniform nonsingularity of $\sum_i^n f_i f_i'/n$, we have $H_n \geq (1 - C)\sum_i^n f_i f_i'/n \geq CI_G$. Then by Lemma A2,

$\tilde{A} = \sum_{i \neq j}^n P_{ij}^r \tilde{X}_i \tilde{X}_j \geq Cdiag(0, I_G)$

Note that $\bar{S}_n' D(0, = \delta')' = (\mu_n, (\delta_0 - \delta)'S_n)'$, and $\bar{X}_i = D'\bar{S}_n\tilde{X}_i$. Then for all

δ,

$$\begin{aligned}\mu_n^{-2}\sum_{i\neq j}P^r_{ij}(y_i - X_i'\delta)(y_j - X_j'\delta) &= \mu_n^{-2}(1,-\delta')(\sum_{i\neq j}P^r_{ij}\bar{X}_i\bar{X}_j')(1,-\delta') \\ &= \mu_n^{-2}(1,-\delta')D'\bar{S}_n\tilde{A}\bar{S}_n'D(1,-\delta')' \\ &\geq C||S_n'(\delta-\delta_0)/\mu_n||^2\end{aligned}$$

Let $\tilde{Q}_n(\delta_0) = (n/\mu_n^2)\sum_{i\neq j}(y_i - X_i'\delta)P^r_{ij}(y_j - X_j'\delta)/(y-X\delta)'(y-X\delta)$. Then by the upper left element of the conclusion of Lemma A2, $\mu_n^{-2}\sum_{i\neq j}\epsilon_i P^r_{ij}\epsilon_j \xrightarrow{\text{p}} 0$. Then
$\tilde{Q}_n(\delta_0) = |\mu_n^{-2}\sum_{i\neq j}\epsilon_i P^r_{ij}\epsilon_j/\sum_i \epsilon_i^2/n| \xrightarrow{\text{p}} 0$.
Since $\hat{\delta} = argmin_\delta \hat{Q}(\delta)$, we have $\hat{Q}(\hat{\delta}) \leq \hat{Q}(\delta_0)$. Therefore, by $(y-X\delta)'(y-X\delta)/n \leq C(1+||\delta||^2)$, it follows that:

$$0 \leq \frac{||S_n'(\hat{\delta}-\delta_0)/\mu_n)||^2}{1+||\hat{\delta}||^2} \leq C\hat{Q}(\hat{\delta}) \leq C\hat{Q}(\delta_0) \xrightarrow{\text{p}} 0$$

implying $||S_n'(\hat{\delta}-\delta_0)/\mu_n)||^2/(1+||\hat{\delta}||^2) \xrightarrow{\text{p}} 0$. Lemma A0 gives the conclusion.
Lemma A4. *(Lemma A4 of Hausman et al. (2012) holds with P^r replacing P).*
If Assumptions 1-4 are satisfied, $\tilde{\alpha} = o_p(\mu_n^2/n)$ and $S_n'(\bar{\delta}-\delta_0)/\mu_n) \xrightarrow{p} 0$, then for
$H_n = \sum_i^n(1-P^r_{ii})f_i f_i'/n$,

$$S_n^{-1}(\sum_{i\neq j}X_i P^r_{ij}X_j' - \hat{\alpha}X'X)S_n^{-1\prime} = H_n + o_p(1),$$

$$S_n^{-1}(\sum_{i\neq j}X_i P^r_{ij}\hat{\epsilon}_j' - \hat{\alpha}X'\hat{\epsilon})/\mu_n \xrightarrow{p} 0$$

Proof.

We note that $X'X = O_p(n)$ and $X'\hat{\epsilon} = O_p(n)$. Therefore, by $||S_n^{-1}|| = O(\mu_n^{-1})$,

$$\hat{\alpha}\tilde{S}_n^{-1}X'XS_n^{-1\prime} = o_p(\mu_n^2/n)O_p(n/\mu_n^2) \xrightarrow{p} 0$$

$$\hat{\alpha}S_n^{-1}X'\hat{\epsilon}/\mu_n = o_p(\mu_n^2/n)O_p(n/\mu_n^2) \xrightarrow{p} 0$$

Lemma A2 (lower right-hand block) and the triangle inequality then give the first conclusion. By Lemma A2 (off diagonal), we have $S_n^{-1}\sum_{i\neq j} X_i P_{ij}^r \epsilon_j/\mu_n \xrightarrow{p} 0$, so that:

$$S_n^{-1}\sum_{i\neq j} X_i P_{ij}^r \hat{\epsilon}_j/\mu_n = o_p(1) - (S_n^{-1}\sum_{i\neq j} X_i P_{ij}^r X_j' S_n^{-1\prime})S_n'(\hat{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0.$$

Lemma A5. *(Lemma A5 of Hausman et al. (2012) holds with P^r replacing P). If Assumptions 1-4 are satisfied, and $S_n'(\hat{\delta}-\delta_0)/\mu_n \xrightarrow{p} 0$, then $\sum_{i\neq j}\hat{\epsilon}_i P_{ij}^r \hat{\epsilon}_j/\hat{\epsilon}'\hat{\epsilon} = o_p(\mu_n^2/n)$.*

Proof.

Let $\hat{\beta} = S_n'(\bar{\delta} - \delta_0)/\mu_n$ and $\alpha = \sum_{i\neq j}\epsilon_i P_{ij}^r \epsilon_j/\epsilon'\epsilon = o_p(\mu_n^2/n)$. Note that $\hat{\sigma}_\epsilon^2 = \hat{\epsilon}'\hat{\epsilon}/n$ satisfies $1/\hat{\sigma}_\epsilon^2 = O_p(1)$ by the Markov inequality. By Lemma A4 with $\tilde{\alpha} = \alpha$, we have $\tilde{H}_n = S_n^{-1}(\sum_{i\neq j} X_i P_{ij}^r X_j' - \tilde{\alpha}X'X)S_n^{-1\prime} = O_p(1)$ and $W_n = S_n^{-1}(\sum_{i\neq j} X_i P_{ij}^r \epsilon_j' - \tilde{\alpha}X'\epsilon)/\mu_n \xrightarrow{p} 0$, so

$$\begin{aligned}\frac{\sum_{i\neq j}\hat{\epsilon}_i P_{ij}^r \hat{\epsilon}_j}{\hat{\epsilon}'\hat{\epsilon}} - \alpha &= \frac{1}{\hat{\epsilon}'\hat{\epsilon}}(\sum_{i\neq j}\hat{\epsilon}_i P_{ij}^r \hat{\epsilon}_j - \sum_{i\neq j}\epsilon_i P_{ij}^r \epsilon_j - \tilde{\alpha}(\hat{\epsilon}'\hat{\epsilon} - \epsilon'\epsilon)) \\ &= \frac{\mu_n^2}{n}\frac{1}{\hat{\sigma}_\epsilon^2}(\hat{\beta}'\tilde{H}_n\hat{\beta} - 2\hat{\beta}'W_n) = o_p(\mu_n^2/n),\end{aligned}$$

The conclusion follows by the triangle inequality.

Proof of Theorem 1.

First, if $S_n'(\hat{\delta}-\delta_0)/\mu_n \xrightarrow{\text{p}} 0$, then by $\lambda_{min}(S_n S_n'/\mu_n^2) \geq \lambda_{min}(\bar{S}\bar{S}') > 0$, we have:

$$||S_n'(\hat{\delta}-\delta_0)/\mu_n|| \geq \lambda_{min}(S_n S_n'/\mu_n^2)^{1/2}||\hat{\delta}-\delta_0|| \geq C||\hat{\delta}-\delta_0||,$$

implying $\hat{\delta} \to \delta_0$. Therefore, it suffices to show that $S_n'(\hat{\delta}-\delta_0)/\mu_n) \xrightarrow{\text{p}} 0$. This follows from Lemma A3.

Now we move on to asymptotic normality results. The next result is a central limit theorem that was proven in Carrasco and Doukali (2016).

Lemma A6.

Suppose the following:

(i) $P^r v = \sum_j q(r,\lambda_j^2) < v, \psi_j > \psi_j$.

(ii)$(W_{1n}, U_1, \epsilon_1), ..., (W_{nn}, U_n, \epsilon_n)$ *are independent, and* $D_{1,n} := \sum_{i=1}^n E[W_{in}W_{in}']$ *satisfies* $||D_{1,n}|| < C$.

(iii) $E[W_{in}'] = 0$, $E[U_i] = 0$, $E[\epsilon_i] = 0$, *and there is a constant* C *such that* $E[||U_i||^4] \leq C$ *and* $E[\epsilon_i^4] \leq C$.

(iv) $\sum_{i=1}^n E[||W_{in}||^4] \to 0$ *a.s.*

(v) $r \to 0$ *as* $n \to \propto$.

Then for:

$D_{2,n} := r\sum_{i\neq j}^n (P_{ij}^r)^2(E[U_iU_i']E[\epsilon_j^2] + E[U_i\epsilon_i]E[U_j'\epsilon_j])$

and any sequences c_{1n} *and* c_{2n} *with* $||c_{1n}|| \leq C$, $||c_{2n}|| \leq C$, *and* $\sum_n{}^{-1/2} = c_{1n}'D_{1,n}c_{1n} + c_{2n}'D_{2,n}c_{2n} > 1/C$, *it follows that:*

$\bar{Y}_n = \sum_n{}^{-1/2}\sqrt{r}(c_{1n}'\sum_{i=1}^n W_{i,n} + c_{2n}'\sum_{i\neq j}^n U_i(P_{ij}^r)^2\epsilon_j) \to N(0,1)$.

Let $\tilde{\alpha}(\delta) = \sum_{i\neq j}^n \epsilon_i(\delta)P_{ij}^r\epsilon_j(\delta)/\epsilon(\delta)'\epsilon(\delta)$ *and*

$$\begin{aligned}\hat{D}(\delta) &= -[\frac{\epsilon(\delta)'\epsilon(\delta)}{2}]\frac{\partial}{\partial\delta}[\frac{\sum_{i\neq j}^n \epsilon_i(\delta)P_{ij}^r\epsilon_j(\delta)}{\epsilon(\delta)'\epsilon(\delta)}] \\ &= \sum_{i\neq j}^n X_iP_{ij}^r\epsilon_j(\delta) - \tilde{\alpha}(\delta)X'\epsilon(\delta)\end{aligned}$$

Lemma A7. *(Lemma A7 of Hausman et al. (2012) holds with P^r replacing P).*
If Assumptions 1-4 are satisfied and $S_n'(\bar{\delta}-\delta_0)/\mu_n) \xrightarrow{p} 0$, then $-S_n^{-1}[\partial\hat{D}(\bar{\delta})/\partial\delta]S_n^{-1} = H_n + o_p(1)$.
Proof.
Let $\bar{\epsilon} = \epsilon(\bar{\delta}) = y - X\bar{\delta}$, $\bar{\gamma} = X'\bar{\epsilon}/\bar{\epsilon}'\bar{\epsilon}$, and $\bar{\alpha} = \tilde{\alpha}(\bar{\delta})$. Then differentiating gives:

$$\begin{aligned} -\frac{\partial\hat{D}}{\partial}(\bar{\delta}) &= \sum_{i\neq j}^{n} X_i P_{ij}^r X_j' - \bar{\alpha}X'X - \bar{\gamma}\sum_{i\neq j}^{n}\bar{\epsilon}_i P_{ij}^r X_j' - \sum_{i\neq j}^{n} X_i P_{ij}^r \bar{\epsilon}_j \bar{\gamma}' + 2(\bar{\epsilon}'\bar{\epsilon})\bar{\alpha}\bar{\gamma}\bar{\gamma}' \\ &= \sum_{i\neq j}^{n} X_i P_{ij}^r X_j' - \bar{\alpha}X'X + \bar{\gamma}\hat{D}(\bar{\delta})' + \hat{D}(\bar{\delta})\bar{\gamma}', \end{aligned}$$

where the second equality follows by $\hat{D}(\bar{\delta}) = \sum_{i\neq j}^{n} X_i P_{ij}^r \bar{\epsilon}_j - (\bar{\epsilon}'\bar{\epsilon})\bar{\alpha}\bar{\gamma}$. By Lemma A5, we have $\bar{\alpha} = o_p(\mu_n^2/n)$. We have, $\bar{\gamma} = O_p(1)$ so that $S_n^{-1}\bar{\gamma} = O_p(1/\mu_n)$. Then by Lemma A4 and $\hat{D}(\bar{\delta}) = \sum_{i\neq j}^{n} X_i P_{ij}^r \bar{\epsilon}_j - \bar{\alpha}X'\bar{\epsilon}$,

$$S_n^{-1}(\sum_{i\neq j}^{n} X_i P_{ij}^r X_j' - \bar{\alpha}X'X)S_n^{-1'} = H_n + o_p(1), S_n^{-1}\hat{D}(\bar{\delta})\bar{\gamma}'S_n^{-1'} \xrightarrow{\text{p}} 0$$

The conclusion then follows by the triangle inequality.
Lemma A8. *(Lemma A8 of Hausman et al. (2012) holds with P^r replacing P).*
If Assumptions 1-4 are satisfied, then for $\gamma_n = \sum_i E[u_i\epsilon_i]/\sum_i E[\epsilon_i^2]$ and $\tilde{U}_i = U_i - \gamma_n\epsilon_i$,

$$S_n^{-1}\hat{D}(\delta_o) = \sum_i (1-P_{ii}^r)f_i\epsilon_i/\sqrt{n} + S_n^{-1}\sum_{i\neq j}^{n}\tilde{U}_i P_{ij}^r \epsilon_j + o_p(1)$$

Proof.

Note that for $W = f'(P^r - I)\epsilon/\sqrt{n}$, by $E[\epsilon\epsilon'] \leq CI_n$, we have

$$\begin{aligned} E[WW'] &= Cf'(I - P^r)f/n = C(f - Z\pi_{Ln})'(I - P^r)(f - Z\pi_{Ln})/n \\ &= CI_G \sum_i^n ||f_i - \pi_{Ln} Z_i||^2/n \to 0 \end{aligned}$$

So $f'(P^r - I)\epsilon/\sqrt{n} = 0_p(1)$. Also by the Markov inequality,

$$X'\epsilon/n = \sum_i^n E[X_i\epsilon_i]/n + O_p(1/\sqrt{n}), \qquad \epsilon'\epsilon = \sum_i^n \sigma_i^2/n + O_p(1/\sqrt{n}),$$

Also, by Assumption 3, $\sum_i^n \sigma_i^2/n \geq C > 0$, The delta method then gives $\tilde{\gamma} = X'\epsilon/\epsilon'\epsilon = \gamma_n + O_p(1/\sqrt{n})$. Therefore, it follows by Lemma A1 and $\hat{D}(\bar{\delta}) = \sum_{i\neq j}^n X_i P_{ij}^r \epsilon_j - (\epsilon'\epsilon)\tilde{\alpha}(\delta_0)\tilde{\gamma}$ that

$$\begin{aligned} S_n^{-1}\hat{D}(\delta_o) &= \sum_{i\neq j}^n f_i P_{ii}^r \epsilon_j/\sqrt{n} + S_n^{-1}\sum_{i\neq j}^n \tilde{U}_i P_{ii}^r \epsilon_i - S_n^{-1}(\tilde{\gamma} - \gamma_n)\epsilon'\epsilon\tilde{\alpha}(\delta_0) \\ &= f'P^r\epsilon/\sqrt{n} - \sum_i^n P_{ii}^r f_i \epsilon_i/\sqrt{n} + S_n^{-1}\sum_{i\neq j}^n \tilde{U}_i P_{ii}^r \epsilon_i + O_p(1/(\sqrt{n}\mu_n))o_p(\mu_n^2/n) \\ &= \sum_i^n (1 - P_{ii}^r) f_i \epsilon_i/\sqrt{n} + S_n^{-1}\sum_{i\neq j}^n \tilde{U}_i P_{ij}^r \epsilon_j + 0_p(1) \end{aligned}$$

Proof of Theorem 2.

By Theorem 1, $\hat{\delta} \to \delta_0$, First order conditions for JLIML are $\hat{D}(\hat{\delta}) = 0$. Expanding gives

$$0 = \hat{D}(\hat{\delta}) + \frac{\partial \hat{D}}{\partial \delta}(\bar{\delta})(\hat{\delta} - \delta_0),$$

where $\bar{\delta}$ lies on the line joining $\hat{\delta}$ and δ_0, and hence $\bar{\beta} = \mu_n^{-1} S_n'(\bar{\delta} - \delta_0)) \xrightarrow{p} 0$. Then by Lemma A7, $\bar{H}_n = S_n^{-1}[\partial\hat{D}(\bar{\delta})/\partial\delta]S_n^{-1'} = H_p + o_p(1)$. Then

$[\partial \hat{D}(\bar{\delta})/\partial\delta]$ is nonsingular and solving gives

$$S_n'(\hat{\delta}-\delta_0) = -S_n'[\partial \hat{D}(\bar{\delta})/\partial\delta]^{-1}\hat{D}(\delta_0) = -\bar{H}_n^{-1}S_n^{-1}\hat{D}(\hat{\delta}).$$

Next, apply Lemma A6 with $U_i = \tilde{U}_i$ and $W_{in} = (1-P_{ii}^r)f_i\epsilon_i/\sqrt{n}$,

$$\sum_i^n E[||W_{in}||^4] \leq C\sum_i^n ||f_i||^4/n^2 \to 0$$

By Assumption 6, we have $\sum_i^n E[W_{in}W_{in}'] \to \Sigma_p$. Let $\Gamma = diag(\Sigma_p, \Psi)$ and

$$A_n = \begin{pmatrix} \sum_i^n W_{in} \\ \sqrt{r}\sum_{i\neq j}^n \tilde{U}_i P_{ij}^r \epsilon_j \end{pmatrix}$$

Consider c such that $c'\Gamma c > 0$. Then by the conclusion of Lemma A6, we have $c'A_n \to N(0, c'\Gamma c)$. Also if $c'\Gamma c = 0$, then it is straightforward to show that $c'A_n \xrightarrow{\text{p}} 0$. Then it follows by the Cramer-Wold device that:

$$A_n = \begin{pmatrix} \sum_i^n W_{in} \\ \sqrt{r}\sum_{i\neq j}^n \tilde{U}_i P_{ii}^r \epsilon_j \end{pmatrix} \xrightarrow{\text{d}} N(0,\Gamma), \qquad \Gamma = diag(\Sigma_n, \Psi).$$

We consider the case where $\frac{1}{r\mu_n^2} = 0$. In this case $\frac{S_n^{-1}}{\sqrt{\alpha}} \to 0.$ so that
$F_n = [I, \frac{S_n^{-1}}{\sqrt{\alpha}}] \to F_0 = [I, 0], \qquad F_0\Gamma F_0' = \Sigma_n.$
Then by Lemma A8,
$S_n^{-1}\hat{D}(\delta_0) = F_nA_n + o_p(1) \xrightarrow{\text{d}} N(0,\Sigma_n),$
$S_n'^{-1}(\hat{\delta}-\delta_0) = -H_n^{-1}S_n^{-1}\hat{D}(\delta_0) \xrightarrow{\text{d}} N(0,V)$, with $V = H_n^{-1}\Sigma_n H_n^{-1}$.

Conclusion.

In this thesis, we illustrate the usefulness of regularization techniques for linear IV estimation and testing overidentifying restrictions in the many weak instruments framework. In Chapters 1 and 3, we propose new estimators which are the regularized versions of Jackknife 2SLS and LIML estimators. We derived the theoretical properties of these estimators. In addition, our simulations show that the leading regularized estimators perform better than other IV estimators. In Chapter 2, we propose a new J test, based on Tikhonov regularization scheme and we study its properties under heteroskedasticity for data rich environments. We theoretically show that our new test achieves the asymptotically correct size in the presence of many instruments. Two empirical applications illustrate the dominance of our proposed J test: one regarding the New-Keynesian Phillips Curve, and the other regarding the elasticity of intertemporal substitution.

We restricted our investigation to instrumental variables linear models. It would be interesting, for future research, to study the behavior of regularized version of non linear moment conditions estimators, such as continuously updated generalized method of moments estimators or generalized empirical likelihood estimator (see Newey and Windmeijer (2009)), in presence of many weak instruments and heteroscedasticity. Another topic of interest is the use of regularization to provide versions of robust test for weak instruments such as Lagrange Multiplier (LM) or conditional likelihood ratio test (CLR) tests, that can be used with large numbers or a continuum of moments conditions. Finally, the results in this thesis suggest that our proposed methods may provide practitioners with a useful tool when faced with big data in the sense that all available data can be used without discarding any a priori.

Bibliography

Stanislav Anatolyev and Nikolay Gospodinov. Specification testing in models with many instruments. *Econometric Theory*, 27(02):427–441, 2011.

Joshua Angrist, Guido Imbens, and Alan Krueger. Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, (14):57–67, 1999.

Joshua D Angrist and Alan B Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106 (4):979–1014, 1991.

Manuel Arellano and Stephen Bond. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297, 1991.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Jushan Bai and Serena Ng. Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, 1(1), 2008.

Jushan Bai and Serena Ng. Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26(06):1577–1606, 2010.

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

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Craig Burnside and Martin Eichenbaum. Small-sample properties of gmm-based wald tests. *Journal of Business & Economic Statistics*, 14(3):294–308, 1996.

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

Marine Carrasco and Mohamed Doukali. Efficient estimation using regularization jackknife estimator. *Mimeo*, 2016.

Marine Carrasco and Barbara Rossi. In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics*, (just-accepted):1–72, 2016.

Marine Carrasco and Guy Tchuente. Regularized LIML for many instruments. *Journal of Econometrics*, 186(2):427–442, 2015.

Marine Carrasco and Guy Tchuente. Regularization based Anderson Rubin tests for many instruments. *Mimeo*, 2016a.

Marine Carrasco and Guy Tchuente. Efficient estimation with many weak instruments using regularization techniques. *forthcoming in Econometric Reviews*, 2016b.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Rui Castro, Gian Luca Clementi, and Glenn MacDonald. Legal institutions, sectoral heterogeneity, and economic development. *The Review of Economic Studies*, 76(2):529–561, 2009.

Gary Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 567–596, 1992.

John C Chao and Norman R Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005.

John C Chao, Norman R Swanson, Jerry A Hausman, Whitney K Newey, and Tiemen Woutersen. Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory*, 28(01):42–86, 2012a.

John C Chao, Norman R Swanson, Jerry A Hausman, Whitney K Newey, and Tiemen Woutersen. Combining two consistent estimators. *Advances in Econometrics*, 29:33–53, 2012b.

John C Chao, Jerry A Hausman, Whitney K Newey, Norman R Swanson, and Tiemen Woutersen. Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics*, 178:15–21, 2014.

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

Christine De Mol, Domenico Giannone, and Lucrezia Reichlin. Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328, 2008.

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

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

Martin S Eichenbaum, Lars Peter Hansen, and Kenneth J Singleton. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. *The Quarterly Journal of Economics*, 103(1): 51–78, 1988.

Wayne A Fuller. Some properties of a modification of the limited information estimator. *Econometrica: Journal of the Econometric Society*, pages 939–953, 1977.

Jordi Galı and Mark Gertler. Inflation dynamics: A structural econometric analysis. *Journal of monetary Economics*, 44(2):195–222, 1999.

Eric Gautier and Alexandre Tsybakov. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

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

Chirok Han and Peter CB Phillips. Gmm with many moment conditions. *Econometrica*, 74(1):147–192, 2006.

Christian Hansen and Damian Kozbur. Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, 182(2):290–308, 2014.

Christian Hansen, Jerry Hausman, and Whitney Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4), 2008.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

Jerry A Hausman, Whitney K Newey, Tiemen Woutersen, John C Chao, and Norman R Swanson. Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, 3(2):211–255, 2012.

George Kapetanios and Massimiliano Marcellino. Factor-gmm estimation with large sets of possibly weak instruments. *Computational Statistics & Data Analysis*, 54(11):2655–2675, 2010.

Rainer Kress. Linear integral equations, volume 82 of applied mathematical sciences, 1999.

Naoto Kunitomo. 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, 1980.

Naoto Kunitomo, Kimio Morimune, and Yoshihiko Tsukuda. Asymptotic expansions of the distributions of the test statistics for overidentifying restrictions in a system of simultaneous equations. *International Economic Review*, pages 199–215, 1983.

Yoonseok Lee and Ryo Okui. Hahn–hausman test as a specification test. *Journal of Econometrics*, 167(1):133–139, 2012.

Ker-Chau Li. Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, pages 958–975, 1987.

Whitney K Newey and Richard J Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

Whitney K Newey and Frank Windmeijer. Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719, 2009.

Garry David Alan Phillips and C Hale. The bias of instrumental variable estimators of simultaneous equation systems. *International Economic Review*, pages 219–228, 1977.

John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415, 1958.

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

James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179, 2002.

James H Stock and Mark W Watson. Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493, 2012.

James H Stock and Motohiro Yogo. Asymptotic distributions of instrumental variables statistics with many instruments. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 2005.

Motohiro Yogo. Estimating the elasticity of intertemporal substitution when instruments are weak. *Review of Economics and Statistics*, 86(3):797–810, 2004.