2013s-20

Regularized LIML for many instruments

Marine Carrasco, Guy Tchuente

Série Scientifique
Scientific Series

Montréal
Juillet 2013

CIRANO
Allier savoir et décision

Centre interuniversitaire de recherche en analyse des organisations

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l’unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.
This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 1198-8177

Partenaire financier

Enseignement supérieur,
Recherche, Science
et Technologie
Québec

Regularized LIML for many instruments [*]

Marine Carrasco [†], *Guy Tchuente* [‡]

Résumé / *Abstract*

The use of many moment conditions improves the asymptotic efficiency of the instrumental variables estimators. However, in finite samples, the inclusion of an excessive number of moments increases the bias. To solve this problem, we propose regularized versions of the limited information maximum likelihood (LIML) based on three different regularizations: Tikhonov, Landweber Fridman, and principal components. Our estimators are consistent and reach the semiparametric efficiency bound under some standard assumptions. We show that the regularized LIML estimator based on principal components possesses finite moments when the sample size is large enough. The higher order expansion of the mean square error (MSE) shows the dominance of regularized LIML over regularized two-staged least squares estimators. We devise a data driven selection of the regularization parameter based on the approximate MSE. A Monte Carlo study shows that the regularized LIML works well and performs better in many situations than competing methods. Two empirical applications illustrate the relevance of our estimators: one regarding the return to schooling and the other regarding the elasticity of intertemporal substitution.

Mots clés/Keywords : High-dimensional models, LIML, many instruments, MSE, regularization methods.

[*] The authors thank the participants of CIREQ conference on High Dimensional Problems in Econometrics (Montreal, May 2012), of the conference in honor of Jean-Pierre Florens (Toulouse, September 2012), of the seminars at the University of Rochester and the University of Pennsylvania for helpful comments.

[†] University of Montreal, CIREQ, CIRANO, email: marine.carrasco@umontreal.ca. Carrasco gratefully acknowledges financial support from SSHRC.

[‡] University of Montreal, CIREQ, email: guy.tchuente.nguembu@umontreal.ca.

1 Introduction

The problem of many instruments is a growing part of the econometric literature. This paper considers the efficient estimation of a finite dimensional parameter in a linear model where the number of potential instruments is very large or infinite. The relevance of such models is due to the collection of large data sets along with the increased power of computers. Many moment conditions can be obtained from nonlinear transformations of an exogenous variable or from using interactions between various exogenous variables. One empirical example of this kind often cited in econometrics is Angrist and Krueger (1991) who estimated return to schooling using many instruments, Dagenais and Dagenais (1997) also estimate a model with errors in variables using instruments obtained from higher-order moments of available variables. The use of many moment conditions improve the asymptotic efficiency of the instrumental variables (IV) estimators. For example, Hansen, Hausman, and Newey (2008) have recently found that in an application from Angrist and Krueger (1991), using 180 instruments, rather than 3 shrinks correct confidence intervals substantially toward those of Kleibergen (2002). But, it has been observed that in finite samples, the inclusion of an excessive number of moments may result in a large bias (Andersen and Sorensen (1996)).

To solve the problem of many instruments efficiently, Carrasco (2012) proposed an original approach based on regularized two-stage least-squares (2SLS). However, such regularized version is not available for the limited information maximum likelihood (LIML). Providing such estimator is desirable, given LIML has better properties than 2SLS (see e.g. Hahn and Inoue (2002), Hahn and Hausman (2003), and Hansen, Hausman, and Newey (2008)). In this paper, we propose a regularized version of LIML based on three regularization techniques borrowed from the statistic literature on linear inverse problems (see Kress (1999) and Carrasco, Florens, and Renault (2007)). The three regularization techniques were also used as in Carrasco (2012) for 2SLS. The first estimator is based on Tikhonov (ridge) regularization. The second estimator is based on an iterative method called Landweber-Fridman. The third regularization technique called principal components or spectral cut-off is based on the principal components associated with the largest eigenvalues. In our paper, the number of instruments is not restricted and may be smaller or larger than the sample size or even infinite. We also allow for a continuum of moment restrictions. Only strong instruments are considered here.

We show that the regularized LIML estimators are consistent, asymptotically normal, and reach the semiparametric efficiency bound under some standard assumptions. We show that the regularized LIML based on principal components has finite first moments provided the sample size is large enough. This result is in contrast with the fact that standard LIML does not possess any moments in finite sample.

Following Nagar (1959), we derive the higher-order expansion of the mean-square error (MSE) of our estimators and show that the regularized LIML estimators dominate the regularized 2SLS in terms of the rate of convergence of the MSE. Our three estimators involve a regularization or tuning parameter, which needs to be selected in practice. The expansion of the MSE provides a tool for selecting the regularization parameter. Following the same approach as in Carrasco (2012), Okui (2011) and Donald and Newey (2001), we propose a data-driven method for selecting the regularization parameter based on a cross-validation approximation of the MSE and show that this selection method is optimal.

The simulations show that the regularized LIML is better than the regularized 2SLS in almost every case. Simulations show that LIML estimator based on Tikhonov and Landweber-Fridman regularizations have most of the times smaller median bias and smaller MSE than LIML estimator based on principal components and than LIML estimator proposed by Donald and Newey (2001).

There is a growing amount of articles on many instruments and LIML. The first papers focused on the case where the number of instruments grow with the sample size, n, but remains smaller than n. In this case, 2SLS estimator is inconsistent while LIML is consistent (see Bekker (1994), Chao and Swanson (2005), Hansen, Hausman, and Newey (2008), Hausman, Newey, Woutersen, Chao, and Swanson (2012) among others). Recently, some work has been done in the case where the number of instruments exceed the sample size. Kuersteiner (2012) considers a kernel weighted GMM estimator, Okui (2011) uses shrinkage. Bai and Ng (2010) and Kapetanios and Marcellino (2010) assume that the endogenous regressors depend on a small number of factors which are exogenous, they use estimated factors as instruments. Belloni, Chen, Chernozhukov, and Hansen (2012) requires the sparsity of the first stage equation and apply an instrument selection based on Lasso. Recently, Hansen and Kozbur (2013) propose a ridge regularized jacknife instrumental variable estimator which does not require sparsity and provide tests with good sizes. The paper which is the most closely related to ours is that by Donald and Newey (2001) (DN henceforth) which select the number of instruments by minimizing an approximate MSE. Their method relies on an a priori ordering of the instruments in decreasing order of strength. Our method does not require such an ordering of the instruments because all the instruments are taken into consideration. It assumes neither a factor structure, nor a sparse first stage equation. However, it assumes that the instruments are sufficiently correlated among themselves so that the inversion of the covariance matrix is ill-posed. This condition is not very restrictive as discussed in Carrasco and Florens (2012).

The paper is organized as follows. Section 2 presents the three regularized LIML estimators and their asymptotic properties. Section 3 derives the higher order expansion of the MSE of the three estimators. In Section 4, we give a data-driven selection of the regularization parameter. Section 5 presents a Monte Carlo experiment. Empirical applications are examined in Section 6. Section 7 concludes. The proofs are collected in appendix.

2 Regularized version of LIML

This section presents the regularized LIML estimators and their properties. We establish that, under some conditions, the regularized LIML estimator based on principal components has finite moments. We also show that the regularized LIML estimators are consistent and reach the semiparametric efficiency bound under some standard assumptions.

2.1 Presentation of the estimators

The model is

$$\begin{cases} y_i = W_i'\delta_0 + \varepsilon_i \\ W_i = f(x_i) + u_i \end{cases} \tag{1}$$

$i = 1, 2, ..., n$. $E(u_i|x_i) = E(\varepsilon_i|x_i) = 0$, $E(\varepsilon_i^2|x_i) = \sigma_\varepsilon^2 > 0$. y_i is a scalar and x_i is a vector of exogenous variables. Some rows of W_i may be exogenous, with the corresponding rows of u_i being zero. $f(x_i) = E(W_i|x_i) = f_i$ is a $p \times 1$ vector of reduced form values. The main focus is the estimation of the $p \times 1$ vector δ_0.

In Model (1), the asymptotic variance of a $\sqrt{n}$-consistent regular estimators cannot be smaller than $\sigma_\varepsilon^2 H^{-1}$, where $H = E(f_i f_i')$ (Chamberlain (1987)). This lower bound is achieved by standard 2SLS if f_i can be written as a finite linear combination of the instruments. In general, efficiency can be reached only from an infinite number of instruments based on power series or exponential functions of x_i (see Carrasco and Florens (2012)). This observation implies that using many instruments is desirable in terms of asymptotic variance. However, the bias of the instrumental variables estimator increases with the number of instruments. To avoid a large bias, some form of instruments selection or regularization need to be applied. To address this issue, Carrasco (2012) proposed a regularized 2SLS estimator building on former work by Carrasco and Florens (2000). Here, we will apply the same regularizations as in Carrasco (2012) to LIML.

As 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. The estimation is based on a sequence of instruments $Z_i = Z(\tau; x_i)$ where $\tau \in S$ may be an integer or an index taking its values in an interval. Examples of Z_i are the following.

- $S = \{1, 2, ...L\}$ thus we have L instruments.

- $Z_{ij} = (x_i)^{j-1}$ with $j \in S = \mathbb{N}$, thus we have an infinite countable instruments.

- $Z_i = Z(\tau; x_i) = exp(i\tau' x_i)$ where $\tau \in S = \mathbb{R}^{dim(x_i)}$, thus we have a continuum of moments.

The estimation of δ is based on the orthogonality condition:

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

Let π be a positive measure on S. For a detailed discussion on the role of π, see Carrasco (2012). We denote $L^2(\pi)$ the Hilbert space of square integrable functions with respect to π. We define the covariance operator K of the instruments as

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

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

where $\overline{Z(\tau_2; x_i)}$ denotes the complex conjugate of $Z(\tau_2; x_i)$. K is assumed to be a compact operator (see Carrasco, Florens, and Renault (2007) for a definition). Carrasco and Florens (2012) show that π can be chosen so that K is compact so that the compactness assumption is not very restrictive.

Let λ_j and ϕ_j $j = 1, 2, ...$ be respectively the eigenvalues (ordered in decreasing order) and the orthogonal eigenfunctions of K. The operator K can be estimated by K_n defined as:

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

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

If the number of moment conditions is infinite, the inverse of K_n needs to be regularized because it is nearly singular. By definition (see Kress, 1999, page 269), a regularized inverse of an operator K is

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

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

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

1. **Tikhonov(T)**

   This regularization scheme is closely related to the ridge regression[1].

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

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

   where $\alpha > 0$ and I is the identity operator.

2. **Landweber Fridman (LF)**

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

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

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

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

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

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

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

   for $\alpha > 0$.

4. **Principal Components (PC)**

   This method is very close to SC and consists in using the first eigenfunctions:

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

[1] $\langle ., . \rangle$ represents the scalar product in $L^2(\pi)$ and in $\mathbb{R}^n$ (depending on the context).

where $\dfrac{1}{\alpha}$ is some positive integer. As the estimators based on PC and SC are identical, we will use PC and SC interchangeably.

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

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

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

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

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

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

and

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

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

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

Let $W = \left(W_1', \ W_2', \ ..., \ W_n'\right)'$ $n \times p$ and $y = \left(y_1', \ y_2', \ ..., \ y_n'\right)'$ $n \times p$. Let us define k-class estimators as

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

where $\nu = 0$ corresponds to the regularized 2SLS estimator studied in Carrasco (2012) and

$$\nu = \nu_{\alpha} = \min_{\delta} \frac{(y - W\delta)' P^{\alpha} (y - W\delta)}{(y - W\delta)'(y - W\delta)}$$

corresponds to the regularized LIML estimator we will study here.

2.2 Existence of moments

The LIML estimator was introduced to correct the bias problem of the 2SLS in the presence of many instruments. It is thus recognized in the literature that LIML has better, small-sample, properties than 2SLS. However, this estimator has no finite moments. Guggenberger (2008) shows by simulations that LIML and GEL have large standard deviations. Fuller (1977) proposes a modified estimator that has finite moments provided the sample size is large enough. Moreover, Anderson (2010) shows that the lack of finite moments of LIML under conventional normalization is a feature of the normalization, not of the LIML estimator itself. He provides a normalization (natural normalization) under which the LIML have finite moments. In a recent paper, Hausman, Lewis, Menzel, and Newey (2011) propose a regularized version of CUE with two regularization parameters and prove the existence of moments assuming these regularization parameters are fixed. However, to obtain efficiency these regularization parameters need to go to zero. In the following proposition, we give some conditions under which the regularized LIML estimator possesses finite moments provided the sample size is large enough.

Proposition 1. *(Moments of the regularized LIML)*
Assume $\{y_i, W_i', x_i'\}$ *are iid,* $\varepsilon_i \sim iid\mathcal{N}(0, \sigma_{\varepsilon}^2)$*,* $X = (x_1, x_2, ..., x_n)$*. Moreover, assume that the vector* u_i *is independent of* X*, independent normally distributed with mean zero and variance* Σ_u*. Let* α *be a positive decreasing function of* n*.*
The r^{th} *moment* $(r = 1, 2, ..)$ *of the regularized LIML estimator with SC regularization is bounded for all* n *greater than some* $n(r)$*.*

Proof In Appendix.

As explained in the proof, we were not able to establish this result for Tikhonov and LF regularizations, even though it may hold. Indeed, the simulations suggest that the first two moments of the estimators based on T and LF exist.

2.3 Asymptotic properties of the regularized LIML

We establish that the regularized LIML estimators are asymptotically normal and reach the semiparametric efficiency bound. Let $f_a(x)$ be the a^{th} element of $f(x)$.

Proposition 2. *(Asymptotic properties of regularized LIML)*
Assume (y_i, W_i', x_i') are iid, $E(\varepsilon_i^2|X) = \sigma_\varepsilon^2$, $E(f_i f_i')$ exists and is nonsingular, K is compact, α goes to zero and n goes to infinity. Moreover, $f_a(x)$ belongs to the closure of the linear span of $\{Z(.;x)\}$ for a = 1,..., p. Then, the T, LF, and SC estimators of LIML satisfy:

1. *Consistency: $\hat{\delta} \to \delta_0$ in probability as n and $n\alpha^{1/2}$ go to infinity.*
2. *Asymptotic normality: If moreover, each element of $E(Z(.;x_i)W_i)$ belongs to the range of K, then*
   $\sqrt{n}(\hat{\delta} - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2[E(f_i f_i')]^{-1}\right)$
   as n and $n\alpha$ go to infinity.

Proof In Appendix.

For the asymptotic normality, we need $n\alpha$ go to infinity as in Carrasco (2012) for 2SLS.

The assumption "$f_a(x)$ belongs to the closure of the linear span of $\{Z(.;x)\}$ for $a = 1, ..., p$" is necessary for the efficiency but not for the asymptotic normality. We notice that all regularized LIML have the same asymptotic properties and achieve the asymptotic semiparametric efficiency bound, as for the regularized 2SLS of Carrasco (2012). Therefore to distinguish among these different estimators, a higher-order expansion of the MSE is necessary.

3 Mean square error for regularized LIML

Now, we analyze the second-order expansion of the MSE of regularized LIML estimators. First, we impose some regularity conditions. Let $\|A\|$ be the Euclidean norm of a matrix A. f is the $n \times p$ matrix, $f = (f(x_1), f(x_2), ..., f(x_n))'$.

Let $\bar{H}$ be the $p \times p$ matrix $\bar{H} = f'f/n$ and $X = (x_1, ..., x_n)$.

Assumption 1: (i) $H = E(f_i f_i')$ exists and is nonsingular,
(ii) there is a $\beta \geq 1/2$ such that

$$\sum_{j=1}^{\infty} \frac{\left\langle E(Z(.,x_i) f_a(x_i)), \phi_j \right\rangle^2}{\lambda_j^{2\beta+1}} < \infty$$

where f_a is the a^{th} element of f for $a = 1; 2...p$

Assumption 2: $\{W_i, y_i, x_i\}$ iid, $E(\varepsilon_i^2|X) = \sigma_\varepsilon^2 > 0$ and $E(\|u_i\|^5|X)$, $E(|\varepsilon_i|^5|X)$ are bounded.

Assumption 3: (i) $E[(\varepsilon_i, u_i')'(\varepsilon_i, u_i')]$ is bounded, (ii) K is a compact operator with nonzero eigenvalues, (iii) $f(x_i)$ is bounded.

These assumptions are similar to those of Carrasco (2012). Assumption 1(ii) is used to derive the rate of convergence of the MSE. More precisely, it guarantees that $\| f - P^\alpha f \| = O_p(\alpha^\beta)$ for LF and PC and $\| f - P^\alpha f \| = O_p(\alpha^{min(2,\beta)})$ for T. The value of β measures how well the instruments approximate the reduced form, f. The larger β, the better the approximation is. The notion of asymptotic MSE employed here is similar to the Nagar-type asymptotic expansion (Nagar (1959)), this Nagar-type approximation is popular in IV estimation literature. We have several reasons to investigate the Nagar asymptotic MSE. First, this approach makes comparison with DN (2001) and Carrasco (2012) easier since they also use the Nagar expansion. Second, a finite sample parametric approach may not be so convincing as it would rely on a distributional assumption. Finally, the Nagar approximation provides the tools to derive a simple way for selecting the regularization parameter in practice.

Proposition 3. *Let $\sigma_{u\varepsilon} = E(u_i\varepsilon_i|x_i)$, $\Sigma_u = E(u_iu_i'|x_i)$ and $\Sigma_v = E(v_iv_i'|x_i)$ with $v_i = u_i - \varepsilon_i\dfrac{\sigma_{u\varepsilon}}{\sigma_\varepsilon^2}$. If Assumptions 1 to 3 hold , $\Sigma_v \neq 0$, $E(\varepsilon_i^2 v_i) = 0$ and $n\alpha \to \infty$ for LF, SC,T regularized LIML, we have*

$$n(\hat{\delta} - \delta_0)(\hat{\delta} - \delta_0)' = \hat{Q}(\alpha) + \hat{r}(\alpha),$$

$$E(\hat{Q}(\alpha)|X) = \sigma_\varepsilon^2 \bar{H}^{-1} + S(\alpha) + T(\alpha),$$

$$[\hat{r}(\alpha) + T(\alpha)]/tr(S(\alpha)) = o_p(1),$$

$$S(\alpha) = \sigma_\varepsilon^2 \bar{H}^{-1}[\Sigma_v \frac{[tr((P^\alpha)^2)]}{n} + \frac{f'\,(1 - P^\alpha)^2\, f}{n}]\bar{H}^{-1}.$$

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

The MSE dominant terms, $S(\alpha)$, is composed of two variance terms one which increases when α goes to zero and the other term which decreases when α goes to zero corresponding to a better approximation of the reduced form by the instruments. Remark that for $\beta \leq 2$, LF, SC, and T give the same rate of convergence of the MSE. However, for $\beta > 2$, T is not as good as the other two regularization schemes. This is the same result found for the regularized 2SLS of Carrasco (2012). For instance if f were a linear combination of the instruments, β would be infinite, and the performance of T would be far worse than that of PC or LF.

The MSE formulae can be used to compare our estimators with those in Carrasco (2012). As in DN, the comparison between regularized 2SLS and LIML depends on the size of $\sigma_{u\varepsilon}$. For $\sigma_{u\varepsilon} = 0$ where there is no endogeneity, 2SLS has smaller MSE than LIML for all regularization schemes, but in this case OLS dominates 2SLS. In order to do this comparison, we need to be precise about the size of the leading term of our MSE approximation

$$S_{LIML}(\alpha) = \sigma_\varepsilon^2 \bar{H}^{-1}[\Sigma_v \frac{[tr((P^\alpha)^2)]}{n} + \frac{f'\,(I - P^\alpha)^2\, f}{n}]\bar{H}^{-1} \tag{2}$$

for LIML and

$$S_{2SLS}(\alpha) = \bar{H}^{-1}[\sigma_{u\varepsilon}\sigma_{u\varepsilon}' \frac{[tr(P^\alpha)]^2}{n} + \sigma_\varepsilon^2 \frac{f'\,(I - P^\alpha)^2\, f}{n}]\bar{H}^{-1}$$

for 2SLS (see Carrasco (2012)). We know that

$$\begin{aligned} S_{LIML}(\alpha) &\sim \frac{1}{n\alpha} + \alpha^{\beta} \\ S_{2SLS}(\alpha) &\sim \frac{1}{n\alpha^2} + \alpha^{\beta} \end{aligned}$$

for LF, PC and if $\beta < 2$ in the Tikhonov regularization. For $\beta \geq 2$ the leading term of the Tikhonov regularization is

$$\begin{aligned} S_{LIML}(\alpha) &\sim \frac{1}{n\alpha} + \alpha^{2} \\ S_{2SLS}(\alpha) &\sim \frac{1}{n\alpha^2} + \alpha^{2} \end{aligned}$$

The MSE of regularized LIML is of smaller order in α than that of the regularized 2SLS because the bias terms for LIML does not depend on α. This is similar to a result found in DN, namely that the biais of LIML does not depend on the number of instruments. For comparison purpose we minimize the equivalents with respect to α and compare different estimators at the minimized point. We find that T, LF and PC LIML are better than T, LF and PC 2SLS in the sense of having smaller minimized value of the MSE, for large n. Indeed, the rate of convergence to zero of $S(\alpha)$ is $n^{-\frac{\beta}{\beta+1}}$ for LIML and $n^{-\frac{\beta}{\beta+2}}$ for 2SLS. The Monte Carlo study presented in Section 5 reveals that almost everywhere regularized LIML performs better than regularized 2SLS.

4 Data driven selection of the regularization parameter

4.1 Estimation of the MSE

In this section we show how to select the regularization parameter α. The aim is to find the α that minimizes the conditional MSE of $\gamma'\hat{\delta}$ for some arbitrary $p \times 1$ vector

γ. This conditional MSE is:

$$\begin{aligned} MSE &= E[\gamma'(\hat{\delta}-\delta_0)(\hat{\delta}-\delta_0)'\gamma|X] \\ &\sim \gamma' S(\alpha)\gamma \\ &\equiv S_\gamma(\alpha). \end{aligned}$$

$S_\gamma(\alpha)$ involves the function f which is unknown. We will need to replace S_γ by an estimate. Stacking the observations, the reduced form equation can be rewritten as

$$W = f + u$$

This expression involves $n \times p$ matrices. We can reduce the dimension by post-multiplying by $\bar{H}^{-1}\gamma$:

$$W\bar{H}^{-1}\gamma = f\bar{H}^{-1}\gamma + u\bar{H}^{-1}\gamma \Leftrightarrow W_\gamma = f_\gamma + u_\gamma \tag{3}$$

where $u_{\gamma i} = u_i'\bar{H}^{-1}\gamma$ is a scalar. Then, we are back to a univariate equation. Let $v_\gamma = v\bar{H}^{-1}\gamma$ and denote

$$\sigma^2_{v_\gamma} = \gamma'\bar{H}^{-1}\Sigma_v\bar{H}^{-1}\gamma.$$

Using (2), $S_\gamma(\alpha)$ can be rewritten as

$$\sigma^2_\varepsilon[\sigma^2_{v_\gamma}\frac{[tr((P^\alpha)^2)]}{n} + \frac{f_\gamma'(I-P^\alpha)^2 f_\gamma}{n}]$$

We see that S_γ depends on f_γ which is unknown. The term involving f_γ is the same as the one that appears when computing the prediction error of f_γ in (3).

The prediction error $\frac{1}{n}E\left[(f_\gamma - \hat{f}^\alpha_\gamma)'(f_\gamma - \hat{f}^\alpha_\gamma)\right]$ equals to

$$R(\alpha) = \sigma^2_{u_\gamma}\frac{tr((P^\alpha)^2)}{n} + \frac{f_\gamma'(I-P^\alpha)^2 f_\gamma}{n}$$

As in Carrasco (2012), the results of Li (1986) and Li (1987) can be applied. Let $\tilde{\delta}$ be a preliminary estimator (obtained for instance from a finite number of instruments) and $\tilde{\varepsilon} = y - W\tilde{\delta}$. Let $\tilde{H}$ be an estimator of $f'f/n$, possibly $W'P^{\tilde{\alpha}}W/n$ where $\tilde{\alpha}$ is obtained from a first stage cross-validation criterion based on one single endogenous variable, for instance the first one (so that we get a univariate regression $W^{(1)} = f^{(1)} + u^{(1)}$ where (1) refers to the first column).

Let $\tilde{u} = (I - P^{\tilde{\alpha}})W$, $\tilde{u}_\gamma = \tilde{u}\tilde{H}^{-1}\gamma$,

$$\hat{\sigma}_\varepsilon^2 = \tilde{\varepsilon}'\tilde{\varepsilon}/n, \hat{\sigma}_{u_\gamma}^2 = \tilde{u}_\gamma'\tilde{u}_\gamma/n, \hat{\sigma}_{u_v\varepsilon} = \tilde{u}_\gamma'\varepsilon/n.$$

We consider the following goodness-of-fit criteria:

Mallows C_p (Mallows (1973))

$$\hat{R}^m(\alpha) = \frac{\hat{u}_\gamma\hat{u}_\gamma}{n} + 2\hat{\sigma}_{u_\gamma}^2\frac{tr(P^\alpha)}{n}.$$

Generalized cross-validation (Craven and Wahba (1979))

$$\hat{R}^{cv}(\alpha) = \frac{1}{n}\frac{\hat{u}_\gamma'\hat{u}_\gamma}{\left(1 - \frac{tr(P^\alpha)}{n}\right)^2}.$$

Leave-one-out cross-validation (Stone (1974))

$$\hat{R}^{lcv}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(\bar{W}_{\gamma_i} - \hat{f}_{\gamma_{-i}}^\alpha)^2,$$

where $\tilde{W}_\gamma = W\tilde{H}^{-1}\gamma$, $\tilde{W}_{\gamma_i}$ is the i^{th} element of $\tilde{W}_\gamma$ and $\hat{f}_{\gamma_{-i}}^\alpha = P_{-i}^\alpha\tilde{W}_{\gamma_{-i}}$. The $n\times(n-1)$ matrix P_{-i}^α is such that $P_{-i}^\alpha = T(K_{n-i}^\alpha)T_{-i}^*$ are obtained by suppressing i^{th} observation from the sample. $\tilde{W}_{\gamma_{-i}}$ is the $(n-1)\times 1$ vector constructed by suppressing the i^{th} observation of $\tilde{W}_\gamma$.

The approximate MSE of $\gamma'\hat{\delta}$ is given by:

$$\hat{S}_\gamma(\alpha) = \sigma_\varepsilon^2\left[\hat{R}(\alpha) + (\hat{\sigma}_{v_\gamma}^2 - \hat{\sigma}_{u_\gamma}^2)\frac{tr((P^\alpha)^2)}{n}\right]$$

where $\hat{R}(\alpha)$ denotes either $\hat{R}^m(\alpha)$, $\hat{R}^{cv}(\alpha)$, or $\hat{R}^{lcv}(\alpha)$.

Noting that $\sigma^2_{v_\gamma} - \sigma^2_{u_\gamma} = -\sigma^2_{u_\gamma \varepsilon}/\sigma^2_\varepsilon$ where $\sigma_{u_\gamma \varepsilon} = E\left(u_{\gamma i}\varepsilon_i\right)$. We define

$$\hat{S}_\gamma(\alpha) = \sigma^2_\varepsilon \left[\hat{R}(\alpha) - \frac{\hat{\sigma}^2_{u_\gamma \varepsilon}}{\hat{\sigma}^2_\varepsilon} \frac{tr((P^\alpha)^2)}{n} \right].$$

The selected regularization parameter[2] is

$$\hat{\alpha} = \arg\min_{\alpha \in M_n} \left[\hat{R}(\alpha) - \frac{\hat{\sigma}^2_{u_\gamma \varepsilon}}{\hat{\sigma}^2_\varepsilon} \frac{tr((P^\alpha)^2)}{n} \right] \tag{4}$$

where M_n is the index set of α. M_n is a compact subset of $[0,1]$ for T, M_n is such that $1/\alpha \in \{1, 2, ..., n\}$ for PC, and M_n is such that $1/\alpha$ is a positive integer no larger than some finite multiple of n.

4.2 Optimality

The quality of the selection of the regularization parameter $\hat{\alpha}$ may be affected by the estimation of $\bar{H}$. A solution to avoid the estimation of $\bar{H}$ is to select γ such that $\bar{H}^{-1}\gamma$ equals a deterministic vector chosen by the econometrician, for instance the unit vector e. This choice is perfectly fine as γ is arbitrary. In this case, $W_\gamma = We$, $f_\gamma = fe$, and $u_\gamma = ue$. In this section, we will restrict ourselves to this case.

We wish to establish the optimality of the regularization parameter selection criteria in the following sense

$$\frac{S_\gamma(\hat{\alpha})}{\inf_{\alpha \in M_n} S_\gamma(\alpha)} \xrightarrow{P} 1 \tag{5}$$

as n and $n\alpha \to \infty$ where $\hat{\alpha}$ is the regularization parameter defined in (4). The result (5) does not imply that $\hat{\alpha}$ converges to a true α in some sense. Instead it establishes that using $\hat{\alpha}$ in the criterion $S_\gamma(\alpha)$ delivers the same rate of convergence as if minimizing $S_\gamma(\alpha)$ directly. For each estimator, the selection criteria provide a means to obtain higher order asymptotically optimal choices for the regularized parameter. It also means that the choice of α using the estimated MSE is asymptotically as good as if the true reduced form were known.

[2] We drop σ^2_ε because it has no effect on the selection of the regularization parameter.

Assumption 4:

(i) $E[((u_i e)^8)]$ is bounded. (i') u_i iid $\mathcal{N}(0, \Sigma_u)$.

(ii) $\hat{\sigma}^2_{u_\gamma} \xrightarrow{P} \sigma^2_{u_\gamma}$, $\hat{\sigma}^2_{u_\gamma \varepsilon} \xrightarrow{P} \sigma^2_{u_\gamma \varepsilon}$, $\hat{\sigma}^2_{\varepsilon} \xrightarrow{P} \sigma^2_{\varepsilon}$,

(iii) $\lim\limits_{n\to\infty} \sup\limits_{\alpha \in M_n} \lambda(P^\alpha_{-i}) < \infty$ where $\lambda(P^\alpha_{-i})$ is largest eigenvalue of P^α_{-i},

(iv) $\sum\limits_{\alpha} (n\tilde{R}(\alpha))^{-2} \xrightarrow{P} 0$ as $n \to \infty$ with $\tilde{R}$ is defined as R with P^α replaced by P^α_{-i}

(v) $\tilde{R}(\alpha)/R(\alpha) \xrightarrow{P} 1$ if either $\tilde{R}(\alpha) \xrightarrow{P} 0$ or $R(\alpha) \xrightarrow{P} 0$.

Proposition 4. ***Optimality of SC and LF***

Under Assumptions 1-3 and Assumption 4 (i-ii), the Mallows C_p and Generalized cross-validation criteria are asymptotically optimal in the sense of (5) for SC and LF. Under Assumptions 1-3 and Assumption 4 (i-v), the leave-one out cross validation is asymptotically optimal in the sense of (5) for SC and LF.

Optimality of T

Under Assumptions 1-3 and Assumption 4 (i') and (ii), the Mallows C_p is asymptotically optimal in the sense of (5) for Tikhonov regularization.

Proof In Appendix.

In the proof of the optimality, we distinguish two cases: the case where the index set of the regularization parameter is discrete and the case where it is continuous. Using as regularization parameter $1/\alpha$ instead of α, SC and LF regularizations have a discrete index set, whereas T has a continuous index set. We use Li (1987) to establish the optimality of Mallows C_p, generalized cross-validation and leave-one-out cross-validation for SC and LF. We use Li (1986) to establish the optimality of Mallows C_p for T. The proofs for generalized cross-validation and leave-one-out cross-validation for T regularization could be obtained using the same tools but are beyond the scope of this paper.

Note that our optimality results hold for a vector of endogenous regressors W_i whereas DN deals only with the case where W_i is scalar.

We will now provide some simulations to see how well our methods perform.

5 Simulation study

In this section we present a Monte carlo study. Our aim is to illustrate the quality of our estimators and compare them to regularized 2SLS estimators of Carrasco (2012) and DN estimators.

Consider

$$\begin{cases} y_i = W_i'\delta + \varepsilon_i \\ W_i = f(x_i) + u_i \end{cases}$$

for $i = 1, 2, ..., n$, $\delta = 0.1$ and $(\varepsilon_i, u_i) \sim \mathcal{N}(0, \Sigma)$ with

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

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

Model 1.

In this model, f is linear as in DN.

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

with $x_i \sim iid\mathcal{N}(0, I_L)$, $L = 15, 30, 50$.

As shown in Hahn and Hausman (2003), the specification implies a theoretical first stage R-squared that is of the form

$$R_f^2 = \frac{\pi'\pi}{1 + \pi'\pi}.$$

The x_i are used as instruments so that $Z_i = x_i$. We can notice that the instruments are independent from each other, this example corresponds to the worse case scenario for our regularized estimators. Indeed, here all the eigenvalues of 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 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 $\pi'\pi = \dfrac{R_f^2}{1 - R_f^2}$.
Here, the instruments are ordered in decreasing order of importance. This model represents a case where there is some prior information about what instruments are important.

Model 1b. $\pi_l = \sqrt{\dfrac{R_f^2}{1 - R_f^2}}$, $l = 1, 2, ..., L$. Here, there is no reason to prefer an instrument over another instrument as all the instruments have the same weight.

Model 2 (Factor model).

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

where $f_i = (f_{i1}, f_{i2}, f_{i3})' \sim iid\mathcal{N}(0, I_3)$, x_i is a $L \times 1$ vector of instruments constructed from f_i through

$$x_i = M f_i + \nu_i$$

where $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2 I_3)$ with $\sigma_\nu = 0.3$, and M is a $L \times 3$ matrix which elements are independently drawn in a U[-1, 1].

We report summary statistics for each of the following estimators: Carrasco's (2012) regularized two-stage least squares, T2SLS (Tikhonov), L2SLS (Landweber Fridman), P2SLS (Principal component), Donald and Newey's (2001) 2SLS (D2SLS), the unfeasible instrumental variable regression (IV), regularized LIML, TLIML (Tikhonov), LLIML (Landweber Fridman), PLIML (Principal component), and finally Donald and Newey's (2001) LIML (DLIML). For each of these estimators, the optimal regularization parameter is selected using Mallows C_p. We report the median bias (Med.bias), the median of the absolute deviations of the estimator from the true value (Med.abs), the difference between the 0.1 and 0.9 quantiles (dis) of the distribution of each estimator, 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:

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

where $\hat{W} = P^{\alpha}W$.

Tables 2, 4, and 6 contain summary statistics for the value of the regularization parameter which minimizes the approximate MSE. This regularization parameter is the number of instruments in DN, α for T, the number of iterations for LF, and the number of principal components for PC[3]. We report the mean, standard error (std), mode, first, second and third quartile of the distribution of the regularization parameter.

Results on Models 1a and 1b are summarized in Tables 1, 2, 3, and 4. We first start by Model 1a where we can notice that the regularized LIML is better than the regularized 2SLS in almost every case. This dominance becomes clearer when the number of instruments increases. We observe that the coverage of regularized 2SLS is very poor while that for regularized LIML is much better, even though it is quite below 95%. Within the regularized LIML, T is the best especially when the number of instruments is very high, except in terms of coverage. However the DLIML is better than all the regularized LIML except for the coverage rate where PLIML is the best and for the median bias where TLIML and LLIML are better. But PLIML have very large MSE, median bias and dispersion. Overall, the performance of TLIML and LLIML is at par with DLIML even though in Model 1a, the instruments are ordered in decreasing order of importance which puts DN at an advantage.

In Model 1b when all instruments have the same weights and there is no reason to prefer one over another, the regularized LIML strongly dominates the regularized 2SLS. The LF and T LIML dominate the DN LIML with respect to all the criteria. We can then conclude that in presence of many instruments and in absence of a reliable information on the relative importance of the instruments, the regularized LIML approach should be preferred to DN approach. We can also notice that when the number of instruments increases the MSE of regularized LIML becames smaller than those of regularized 2SLS. We observe that the MSE of DLIML explodes while those of TLIML and LLIML are stable, which can be explained by the existence of the moments of the regularized LIML, except for the PLIML which MSE also explodes.

[3] The optimal α for Tikhonov is searched over the interval [0.1,0.5] with 0.01 increment for Models 1a and 1b and the set {0, 0.000000001, 0.00000001, 0.0000001, 0.0000001, 0.000001, 0.00001, 0.01, 0.1, 0.2} for Model 2. The range of values for the number of iterations for LF is from 1 to 10 times the number of instruments and for the number of principal components is from 1 to the number of instruments.

Table 1: Simulations results of Model 1a with $R_f^2 = 0.1$, $n = 500$, 1000 replications

Model 1a		T2SLS	L2SLS	P2SLS	D2SLS	IV	TLIML	LLIML	PLIML	DLIML
L=15	Med.bias	0.101	0.101	0.115	0.052	0.004	0.006	0.007	0.017	0.024
	Med.abs	0.113	0.118	0.139	0.105	0.092	0.103	0.102	0.102	0.096
	Disp	0.305	0.312	0.367	0.360	0.355	0.391	0.391	0.390	0.363
	MSE	0.024	0.024	0.153	0.021	0.019	0.025	0.025	4.5e+28	0.021
	Cov	0.811	0.819	0.818	0.900	0.954	0.898	0.902	0.895	0.919
L=30	Med.bias	0.172	0.167	0.180	0.070	0.005	0.007	0.009	0.046	0.024
	Med.abs	0.172	0.169	0.204	0.109	0.091	0.106	0.108	0.113	0.100
	Disp	0.263	0.291	0.457	0.352	0.370	0.447	0.441	0.433	0.364
	MSE	0.039	0.039	11.038	0.024	0.020	0.032	0.032	Inf	0.023
	Cov	0.586	0.632	0.712	0.872	0.950	0.820	0.812	0.824	0.894
L=50	Med.bias	0.231	0.219	0.217	0.099	0.004	-0.002	0.007	0.083	0.039
	Med.abs	0.231	0.219	0.250	0.123	0.092	0.125	0.128	0.141	0.097
	Disp	0.237	0.268	0.568	0.354	0.341	0.470	0.487	0.496	0.367
	MSE	0.061	0.057	42.529	0.028	0.020	0.039	0.046	1.5e+30	0.022
	Cov	0.294	0.427	0.661	0.833	0.955	0.707	0.732	0.760	0.894

The poor performance of PC 2SLS and LIML in Models 1a and 1b can be explained by the absence of factor structure. Indeed, all eigenvalues of K (in the population) are equal to each other and consequently the Mallows C_p tend to select a large number of principal components (see Tables 2 and 4). The PC LIML is therefore close to the standard LIML estimator which is known for not having any moments.

Now, we turn to Model 2 which is a factor model. From Table 5, we see that LIML dominates the 2SLS for all regularization schemes. But, there is no clear dominance among the regularized LIML as they all perform very well. The DLIML is dominated by regularized LIML for all measures. From Table 6, we can observe that PC selects three principal components in average corresponding to the three factors.

We conclude this section by summarizing the Monte Carlo results. LLIML and TLIML are highly recommended if we are concerned with bias and mean square error. Selection methods as DN are recommended when the rank ordering of the strength of the instruments is clear. Otherwise, regularized methods are preferrable. Among the three regularizations, LF and T are more reliable than PC in absence of factor structure. Moreover, we observe small mean square errors for T and LF regularized LIML estimators which suggests the existence of moments.

Table 2: Properties of the distribution of the regularization parameters Model 1a

Model 1a		T2SLS	L2SLS	P2SLS	D2SLS	TLIML	LLIML	PLIML	DLIML
L=15	Mean	0.433	18.227	9.074	4.759	0.235	32.721	13.061	6.053
	sd	0.114	11.615	4.015	1.816	0.086	9.533	2.374	2.533
	q1	0.400	11.000	6.000	4.000	0.180	26.000	12.000	4.000
	q2	0.500	15.000	9.000	4.000	0.220	31.000	14.000	5.000
	q3	0.500	22.000	13.000	5.000	0.280	37.000	15.000	7.000
L=30	Mean	0.485	12.324	10.490	6.630	0.420	26.669	22.740	9.865
	sd	0.064	12.406	7.573	2.480	0.092	9.568	6.873	4.064
	q1	0.500	6.000	5.000	5.000	0.340	20.000	18.000	7.000
	q2	0.500	9.000	9.000	6.000	0.460	25.000	25.000	9.000
	q3	0.500	14.000	14.000	8.000	0.500	31.000	29.000	11.000
L=50	Mean	0.494	9.772	12.743	8.211	0.492	20.383	26.693	13.100
	sd	0.040	11.356	12.227	3.528	0.030	7.628	14.082	4.945
	q1	0.500	4.000	4.000	6.000	0.500	16.000	15.000	10.000
	q2	0.500	7.000	9.000	8.000	0.500	19.000	27.500	12.000
	q3	0.500	11.000	17.000	10.000	0.500	24.000	38.000	16.000

Table 3: Simulations results of Model 1 b with $R_f^2 = 0.1$, $n = 500$, 1000 replications

Model 1b		T2SL	L2LS	P2LS	D2LS	IV	TLIML	LLIML	PLIML	DLIML
L=15	Med.bias	0.099	0.096	0.112	0.128	-0.006	-0.000	-0.001	0.015	0.011
	Med.abs	0.109	0.115	0.141	0.146	0.087	0.103	0.102	0.103	0.101
	Disp	0.290	0.297	0.372	0.346	0.347	0.390	0.386	0.380	0.380
	MSE	0.023	0.023	0.059	0.042	0.019	0.024	0.025	1.4e+28	0.023
	Cov	0.840	0.843	0.837	0.805	0.946	0.897	0.899	0.891	0.895
L=30	Med.bias	0.172	0.165	0.174	0.219	0.006	0.010	0.011	0.042	0.052
	Med.abs	0.173	0.165	0.202	0.237	0.091	0.107	0.110	0.111	0.115
	Disp	0.264	0.277	0.453	0.457	0.355	0.412	0.421	0.413	0.411
	MSE	0.039	0.038	3.682	907.310	0.020	0.030	0.033	1.3e+29	1.7e+30
	Cov	0.594	0.643	0.725	0.673	0.952	0.829	0.828	0.806	0.801
L=50	Med.bias	0.237	0.226	0.214	0.257	-0.004	-0.004	0.000	0.082	0.107
	Med.abs	0.237	0.226	0.252	0.285	0.089	0.124	0.126	0.137	0.156
	Disp	0.235	0.259	0.581	0.590	0.353	0.470	0.489	0.483	0.526
	MSE	0.061	0.058	1.794	4.946	0.020	0.039	0.045	4.4e+30	3.8e+30
	Cov	0.300	0.406	0.688	0.639	0.951	0.723	0.723	0.752	0.714

Table 4: Properties of the distribution of the regularization parameters Model 1b

Model 1b		T2SL	L2SLS	P2SLS	D2SLS	TLIML	LLIML	PLIML	DLIML
L=15	Mean	0.438	18.118	8.909	10.021	0.233	32.909	13.053	14.223
	sd	0.112	12.273	3.916	3.995	0.085	9.925	2.463	1.460
	q1	0.410	11.000	6.000	7.000	0.170	26.000	12.000	14.000
	q2	0.500	15.000	9.000	11.000	0.210	31.000	14.000	15.000
	q3	0.500	21.000	12.000	14.000	0.270	37.000	15.000	15.000
L=30	Mean	0.486	11.963	10.431	11.310	0.421	26.584	22.636	25.283
	sd	0.059	11.019	7.660	8.634	0.091	9.299	7.160	6.303
	q1	0.500	6.000	4.000	4.000	0.360	20.000	18.000	24.000
	q2	0.500	9.000	9.000	9.000	0.460	25.000	25.000	28.000
	q3	0.500	14.000	15.000	17.000	0.500	31.000	29.000	30.000
L=50	Mean	0.493	10.127	11.911	13.508	0.492	20.146	26.210	29.362
	sd	0.043	13.632	11.605	13.943	0.031	7.537	14.197	16.864
	q1	0.500	4.000	4.000	3.000	0.500	15.000	15.000	13.000
	q2	0.500	7.000	8.000	8.000	0.500	19.000	26.000	33.000
	q3	0.500	11.000	16.000	19.000	0.500	24.000	38.000	46.000

Table 5: Simulations results of Model 2 , $n = 500$, 1000 replications

Model 2		**T2SLS**	**L2SLS**	**P2SLS**	**D2SLS**	**IV**	**TLIML**	**LLIML**	**PLIML**	**DLIML**
L=15	Med.bias	0.00033	0.00014	0.00033	0.0039	0.00067	-0.00033	-0.00023	-0.00025	0.00047
	Med.abs	0.017	0.0174	0.0176	0.0177	0.0179	0.0177	0.0177	0.017	0.017
	Disp	0.066	0.066	0.066	0.0658	0.0670	0.067	0.067	0.067	0.068
	MSE	0.00069	0.00069	0.00069	0.00070	0.00068	0.00069	0.00069	0.00069	0.00070
	cov	0.947	0.946	0.947	0.942	0.952	0.948	0.948	0.948	0.946
L=30	Med.bias	0.0019	0.0016	0.0016	0.0051	0.0013	0.0011	0.00095	0.00095	0.00193
	Med.abs	0.016	0.0170	0.0171	0.0174	0.0171	0.0171	0.0172	0.0172	0.0172
	Disp	0.0668	0.0671	0.0672	0.0676	0.0668	0.0679	0.0677	0.0677	0.0685
	MSE	0.00066	0.00066	0.00066	0.00069	0.00065	0.00066	0.00066	0.00066	0.00068
	cov	0.956	0.955	0.955	0.949	0.958	0.953	0.955	0.955	0.946
L=50	Med.bias	0.0010	0.00036	0.000458	0.00352	0.00062	4.7045e-5	-0.00034	-0.00017	0.00108
	Med.abs	0.0168	0.0165	0.0165	0.0175	0.0171	0.0167	0.0167	0.0167	0.0177
	Disp	0.065	0.0656	0.0655	0.0666	0.0648	0.0656	0.0654	0.065	0.067
	MSE	0.00065	0.00066	0.00066	0.00068	0.00065	0.00066	0.00066	0.00066	0.00067
	cov	0.945	0.946	0.947	0.946	0.95	0.945	0.944	0.945	0.944

Table 6: Properties of the distribution of the regularization parameters Model 2

Model 2		**T2SLS**	**L2SLS**	**P2SLS**	**D2SLS**	**TLIML**	**LLIML**	**PLIML**	**DLIML**
L=15	Mean	0.19431	150	3.012	9.44	0.1334	150	3.0120	13.1350
	Sd	0.023543	0	0.10894	1.5306	0.0776	0	0.1089	1.7081
	q1	0.2	150	3	9	0.1	150	3	12
	q2	0.2	150	3	9	0.2	150	3	14
	q3	0.2	150	3	10	0.2	150	3	14
L=30	mean	0.1998	175.48	3.009	11.107	0.14671	225.44	3.009	22.37
	Sd	0.0044699	35.078	0.13014	2.2607	0.071249	62.243	0.13014	5.8391
	q1	0.2	152	3	10	0.1	172.5	3	17
	q2	0.2	173	3	11	0.2	209	3	21
	q3	0.2	195	3	11	0.2	300	3	28
L=50	Mean	0.2	140.73	3.015	9.57	0.15795	292.1	3.015	22.223
	Sd	2.8325e-15	36.604	0.13709	1.5249	0.061017	175	0.13709	9.1024
	q1	0.2	122	3	9	0.1	140	3	14
	q2	0.2	137	3	9	0.2	177	3	22
	q3	0.2	155	3	11	0.2	500	3	27

6 Empirical applications

6.1 Returns to Schooling

A motivating empirical example is provided by the influential paper of Angrist and Krueger (1991). This study has become a benchmark for testing methodologies concerning IV estimation in the presence of many (possibly weak) instrumental variables. The sample drawn from the 1980 U.S. Census consists of 329,509 men born between 1930-1939. Angrist and Krueger (1991) estimate an equation where the dependent variable is the log of the weekly wage, and the explanatory variable of interest is the number of years of schooling. It is obvious that OLS estimate might be biased because of the endogeneity of education. Angrist and Krueger (1991) propose to use the quarters of birth as instruments. Because of the compulsory age of schooling, the quarter of birth is correlated with the number of years of education, while being exogenous. The relative performance of LIML on 2SLS, in presence of many instruments, has been well documented in the literature (DN, Anderson, Kunitomo, and Matsushita (2010), and Hansen, Hausman, and Newey (2008)). We are going to compute the regularized version of LIML and compare it to the regularized 2SLS in order to show the empirical relevance of our method.

We use the model of Angrist and Krueger (1991):

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

where $logw$ = 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. Table 7 reports schooling coefficients generated by different estimators applied to the Angrist and Krueger data along with their standard errors in parentheses. Table 7 shows that all regularized 2SLS and LIML estimators based on the same type of regularization give close results. The coefficients we obtain by regularized LIML are slightly larger than those obtained by regularized 2SLS suggesting that these methods provide an extra bias correction, as observed in our Monte Carlo simulations. Note that the bias reduction obtained by regularized LIML compared to standard LIML comes at the cost of a larger standard error. Among the regularizations, PC gives estimators which are quite a bit smaller than T and LF. However, we are suspicious of PC because there is no factor structure here.

Table 7: Estimates of the return to education

OLS	**2SLS**	**T2SLS**	**L2SLS**	**P2SLS**
0.0683 (0.0003)	0.0816 (0.0106)	0.1237 (0.0482)	0.1295 (0.0272)	0.1000 (0.0411)
		α= 0.00001	Nb of iterations 700	Nb of eigenfunctions 81
	LIML	**TLIML**	**LLIML**	**PLIML**
	0.0918 (0.0101)	0.1237 (0.0480)	0.1350 (0.0275)	0.107 (0.0184)
		α= 0.00001	Nb of iterations 700	Nb of eigenfunctions 239

6.2 Elasticity of Intertemporal Substitution

In macroeconomics and finance, the elasticity of intertemporal substitution (EIS) in consumption is a parameter of central importance. It has important implications for the relative magnitudes of income and substitution effects in the intertemporal consumption decision of an investor facing time varying expected returns. Campbell and Viceira (1999) show that when the EIS is less (greater) than 1, the investor's optimal consumption-wealth ratio is increasing (decreasing) in expected returns.

Yogo (2004) analyzes the problem of EIS using the linearized Euler equation. He explains how weak instruments have been the source for an empirical puzzle namely that using conventional IV methods the estimated EIS is significantly less than 1 but its reciprocal is not different from 1. In this subsection, we follow one of the specifications in Yogo (2004) using quarterly data from 1947.3 to 1998.4 for the United States and compare all the estimators considered in the present paper. The estimated models are given by the following equation:

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

and the "reverse regression":

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

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.

To solve the empirical puzzle, we increase the number of instruments from 4 to 18 by including interactions and power functions. The four instruments used by Yogo (2004) are the twice lagged, nominal interest rate (r), inflation (i), consumption growth (c) and log dividend-price ratio (p). The set of instruments is then $Z = [r, i, c, p]$. The 18 instruments used in our regression are derived from Z and are given by[4] $II = [Z, Z.^2, Z.^3, Z(:,1) * Z(:,2), Z(:,1) * Z(:,3), Z(:,1) * Z(:,4), Z(:,2) * Z(:,3), Z(:,2) * Z(:,4), Z(:,3) * Z(:,4)]$.

Since the increase of the number of instruments improves efficiency and regularized 2SLS and LIML correct the bias due to the use of many instruments, the increase of the number of instruments will certainly enable us to have better points estimates. 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)). For LF estimator, 2SLS and LIML

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

regularizations give very close estimators.

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 fact that there is no factor structure.

Table 8: Estimates of the EIS

	2SLS (4 instr)	**2SLS** (18 instr)	**T2SLS**	**L2SLS**	**P2SLS**
ψ	0.0597	0.1884	0.71041	0.71063	0.1696
	(0.0876)	(0.0748)	(0.093692)	(0.093689)	(0.077808)
			$\alpha = 0.0001$	Nb of iterations 300	Nb of PC 11
$1/\psi$	0.6833	0.8241	1.406	1.407	0.7890
	(0.4825)	(0.263)	(0.2496)	(0.249)	(0.246)
			$\alpha = 0.01$	Nb of iterations 300	Nb of PC 17
	LIML (4 instr)	**LIML** (18 instr)	**TLIML**	**LLIML**	**PLIML**
ψ	0.0293	0.2225	0.71041	0.71063	0.1509
	(0.0994)	(0.0751)	(0.093692)	(0.093689)	(0.077835)
			$\alpha = 0.01$	Nb of iterations 300	Nb of PC 8
$1/\psi$	34.1128	4.4952	1.407	1.4072	3.8478
	(112.7122)	(0.5044)	(0.249)	(0.249)	(0.37983)
			$\alpha = 0.01$	Nb of iterations 300	Nb of PC 17

7 Conclusion

In this paper, we propose a new estimator which is a regularized version of LIML estimator. Our framework has the originality to allow for a finite and infinite number of moment conditions. We show theoretically that regularized LIML improves upon regularized 2SLS in terms of smaller leading terms of the MSE. All the regularization methods involve a tuning parameter which needs to be selected. We propose a data-driven method for selecting this parameter and show that this selection procedure is optimal. Moreover, we prove that the LIML estimator based on principal components has finite moments. Although, we were not able to prove this result for LIML regularized with T and LF, the simulations suggest that their first two moments exist. Our simulations show that the leading regularized estimators (LF and T of LIML) are nearly median unbiased and dominate regularized 2SLS and standard LIML in terms of MSE.

We restrict our work in this paper to the estimation and asymptotic properties of regularized LIML with many strong instruments. One possible topic for future research would be to extend these results to the case of weak instruments as in Hansen, Hausman, and Newey (2008). Moreover, it would be interesting to study the behavior of other k-class estimators, such as Fuller's (1977) estimator and bias corrected 2SLS estimator, in presence of many (and possibly weak) instruments. Another interesting topic is the use of our regularized LIML or 2SLS for inference when facing many instruments or a continuum of instruments. This would enable us to compare our inference with those of Hansen, Hausman, and Newey (2008) and Newey and Windmeijer (2009).

References

ANDERSEN, T. G., AND B. E. SORENSEN (1996): "GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study," *Journal of Business & Economic Statistics*, 14(3), 328–52.

ANDERSON, T. (2010): "The LIML estimator has finite moments!," *Journal of Econometrics*, 157(2), 359–361.

ANDERSON, T., N. KUNITOMO, AND Y. MATSUSHITA (2010): "On the asymptotic optimality of the LIML estimator with possibly many instruments," *Journal of Econometrics*, 157(2), 191–204.

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

BAI, J., AND S. NG (2010): "Instrumental Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26, 1577–1606.

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

BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): "Sparse models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

CAMPBELL, J. Y., AND L. M. VICEIRA (1999): "Consumption And Portfolio Decisions When Expected Returns Are Time Varying," *The Quarterly Journal of Economics*, 114(2), 433–495.

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

CARRASCO, M., AND J.-P. FLORENS (2000): "Generalization Of GMM To A Continuum Of Moment Conditions," *Econometric Theory*, 16(06), 797–834.

——— (2012): "On the Asymptotic Efficiency of GMM," *forthcoming in Econometric Theory.*

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

Castro, R., G. L. Clementi, and G. Macdonald (2009): "Legal Institutions, Sectoral Heterogeneity, and Economic Development," *Review of Economic Studies*, 76(2), 529–561.

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

Craven, P., and G. Wahba (1979): "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of the generalized cross-validation," *Numer. Math.*, 31, 377–403.

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

Donald, S. G., and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–91.

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

Guggenberger, P. (2008): "Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator," *Econometric Reviews*, 27(4-6), 526–541.

Hahn, J., and J. Hausman (2003): "Weak Instruments: Diagnosis and Cures in Empirical Econometrics," *American Economic Review*, 93(2), 118–125.

HAHN, J., AND A. INOUE (2002): "A Monte Carlo Comparison Of Various Asymptotic Approximations To The Distribution Of Instrumental Variables Estimators," *Econometric Reviews*, 21(3), 309–336.

HANSEN, C., J. HAUSMAN, AND W. NEWEY (2008): "Estimation With Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26, 398–422.

HANSEN, C., AND D. KOZBUR (2013): "Instrumental Variables Estimation with Many Weak Instruments Using Regularized JIVE," *mimeo.*

HAUSMAN, J., R. LEWIS, K. MENZEL, AND W. NEWEY (2011): "Properties of the CUE estimator and a modification with moments," *Journal of Econometrics*, 165(1), 45 – 57.

HAUSMAN, J. A., W. K. NEWEY, T. WOUTERSEN, J. C. CHAO, AND N. R. SWANSON (2012): "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics*, 3(2), 211–255.

JAMES, A. (1954): "Normal Multivariate Analysis and the Orthogonal Group," *Annals of Mathematical Statistics*, 25, 46–75.

KAPETANIOS, G., AND M. MARCELLINO (2010): "Factor-GMM estimation with large sets of possibly weak instruments," *Computational Statistics and Data Analysis*, 54, 2655–2675.

KLEIBERGEN, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70(5), 1781–1803.

KRESS, R. (1999): in *Linear Integral Equations*vol. 82, p. 388. Springer, 2 edn.

KUERSTEINER, G. (2012): "Kernel-weighted GMM estimators for linear time series models," *Journal of Econometrics*, 170, 399–421.

LI, K.-C. (1986): "Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing," *The Annals of Statistics*, 14, 1101–1112.

——— (1987): "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

Mallows, C. L. (1973): "Some Comments on Cp," *Technometrics*, 15, 661–675.

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

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

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

Stone, C. J. (1974): "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, 36, 111–147.

Yogo, M. (2004): "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak," *The Review of Economics and Statistics*, 86(3), 797–810.

A Proofs

Proof of Proposition 1

We want to prove that the regularized LIML estimators have finite moments. These estimators are defined as follow [5]:

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

where $\nu_{\alpha} = \min_{\delta} \dfrac{(y - W\delta)' P^{\alpha} (y - W\delta)}{(y - W\delta)'(y - W\delta)}$ and $P^{\alpha} = T(K_n^{\alpha})^{-1} T^*$.
Let us define $\hat{H} = W' \left(P^{\alpha} - \nu_{\alpha} I_n\right) W$ and $\hat{N} = W' \left(P^{\alpha} - \nu_{\alpha} I_n\right) y$ thus

$$\hat{\delta} = \hat{H}^{-1} \hat{N}.$$

If we denote $W^v = (W_{1v}, W_{2v}, ..., W_{nv})'$, $\hat{H}$ is a $p \times p$ matrix with a typical element

$$\hat{H}_{vl} = \sum_j (q_j - \nu_{\alpha}) \left\langle W^v, \hat{\psi}_j \right\rangle \left\langle W^l, \hat{\psi}_j \right\rangle$$

and $\hat{N}$ is a $p \times 1$ vector with a typical element

$$N_l = \sum_j (q_j - \nu_{\alpha}) \left\langle y, \hat{\psi}_j \right\rangle \left\langle W^l, \hat{\psi}_j \right\rangle .$$

By the Cauchy-Schwarz inequality and because $|\nu_{\alpha}| \leq 1$, $|q_j| \leq 1$, we can prove that $|\hat{H}_{vl}| \leq 2\|W^l\| \|W^v\|$ and $|N_l| \leq 2\|y\| \|W^l\|$.
Under our assumptions, all the moments (conditional on X) of W and y are finite, we can conclude that all elements of $\hat{H}$ and $\hat{N}$ have finite moments.

The i^{th} element of $\hat{\delta}$ is given by:

$$\hat{\delta}_i = \sum_{j=1}^{p} |\hat{H}|^{-1} cof(\hat{H}_{ij}) N_j$$

where $cof(\hat{H}_{ij})$ is the signed cofactor of $\hat{H}_{ij}$, N_j is the j^{th} element of $\hat{N}$ and $|\ .\ |$ denotes

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

the determinant.

$$| \hat{\delta}_i |^r \leq |\hat{H}|^{-r}| \sum_{j=1}^{p} cof(\hat{H}_{ij})N_j|^r$$

Let $\alpha_1 > \alpha_2$ be two regularization parameters. It turns out that $P^{\alpha_1} - P^{\alpha_2}$ is semi definite negative and hence $0 \leq \nu_{\alpha_1} \leq \nu_{\alpha_2}$. This will be used in the proof. [6]

We also have that

$$\begin{aligned} \nu_\alpha &= \min_\delta \frac{(y - W\delta)' P^\alpha (y - W\delta)}{(y - W\delta)'(y - W\delta)} \\ &\leq \max_c \frac{c'^\alpha c}{c'c} = \max_j q_j. \end{aligned}$$

We want to prove that $|\hat{H}| \geq |S|$ where S is a positive definite $p \times p$ matrix to be specified later on.

We want to show that $P^\alpha - \nu_{\frac{\alpha}{2}} I_n$ is positive definite. Let us consider $x \in \mathbb{R}^n$. We have

$$\begin{aligned} x' \left(P^\alpha - \nu_{\frac{\alpha}{2}} I_n\right) x &= \sum_j (q_j - \nu_{\frac{\alpha}{2}}) \left\langle x, \psi_j \right\rangle' \left\langle x, \psi_j \right\rangle \\ &= \sum_j (q_j - \nu_{\frac{\alpha}{2}}) \parallel \left\langle x, \psi_j \right\rangle \parallel^2 \\ &= \sum_{j, q_j > \nu_{\frac{\alpha}{2}}} (q_j - \nu_{\frac{\alpha}{2}}) \parallel \left\langle x, \psi_j \right\rangle \parallel^2 \quad (1) \\ &+ \sum_{j, q_j \leq \nu_{\frac{\alpha}{2}}} (q_j - \nu_{\frac{\alpha}{2}}) \parallel \left\langle x, \psi_j \right\rangle \parallel^2 . \quad (2) \end{aligned}$$

We know that q_j is a decreasing sequence. Hence there exists j^*_α such that $q_j \geq \nu_{\frac{\alpha}{2}}$ for $j^*_\alpha < j$ and $q_j < \nu_{\frac{\alpha}{2}}$ for $j^*_\alpha > j$ and

$$\begin{aligned} x' \left(P^\alpha - \nu_{\frac{\alpha}{2}} I_n\right) x &= \sum_{j \leq j^*_\alpha} (q_j - \nu_{\frac{\alpha}{2}}) \parallel \left\langle x, \psi_j \right\rangle \parallel^2 \quad (1) \\ &+ \sum_{j > j^*_\alpha} (q_j - \nu_{\frac{\alpha}{2}}) \parallel \left\langle x, \psi_j \right\rangle \parallel^2 . \quad (2) \end{aligned}$$

The term (1) is positive and the term (2) is negative. As n increases, α decreases and

[6] Note that if the number of instruments is smaller than n we can compare ν obtained with P^α replaced by P, the projection matrix on the instruments, and ν_α. It turns out that $P^\alpha - P$ is definite negative for fixed α and hence $0 \leq \nu_\alpha \leq \nu$ as in Fuller (1977).

q_j increases for all j while $\nu_{\frac{\alpha}{2}}$ also increases. Note that in the case of SC, q_j switches from 1 to 0 at j_α^* so that the variation in α of q_j at $j = j_\alpha^*$ is greater than the variation of $\nu_{\frac{\alpha}{2}}$. It follows that j_α^* increases when α decreases. We were not able to prove this result in the case of T and LF.

The term (2) goes to zero as n goes to infinity. Indeed when j_α^* goes to infinity, we have

$$\left| \sum_{j>j_{\alpha,}^*} (q_j - \nu_{\frac{\alpha}{2}}) \parallel \left\langle x, \psi_j \right\rangle \parallel^2 \right| \leq \sum_{j>j_\alpha^*} \parallel \left\langle x, \psi_j \right\rangle \parallel^2 = o_p(1).$$

We can conclude that for n sufficiently large, α is small and j_α^* is sufficiently large for (2) to be smaller in absolute value than (1) and hence $x'^\alpha \left(P^\alpha - \nu_{\frac{\alpha}{2}} I_n\right) x > 0$.

Denote $S = (\nu_{\frac{\alpha}{2}} - \nu_\alpha) W'W$ we have

$$\begin{aligned} \hat{H} &= W' \left(P^\alpha - \nu_\alpha I_n\right) W \\ &= W' \left(P^\alpha - \nu_{\frac{\alpha}{2}} I_n\right) W + (\nu_{\frac{\alpha}{2}} - \nu_\alpha) W'W \\ &= W' \left(P^\alpha - \nu_{\frac{\alpha}{2}} I_n\right) W + S. \end{aligned}$$

Hence,

$$\begin{aligned} |\hat{H}| &= |W' \left(P^\alpha - \nu_{\frac{\alpha}{2}} I_n\right) W + S| \\ &= |S||I_p + S^{-1/2} W' \left(P^\alpha - \nu_{\frac{\alpha}{2}} I_n\right) W S^{-1/2}| \\ &\geq |S|. \end{aligned}$$

For n large but not infinite, $\nu_{\frac{\alpha}{2}} - \nu_\alpha > 0$ and $|S| > 0$. As in Fuller (1977) using James (1954), we can show that the expectation of the inverse $2r^{th}$ power of the determinant of S exists and is bounded for n greater than some number $n(r)$, since S is expressible as a product of multivariate normal r.v.. Thus we can apply Lemma B of Fuller (1977) and conclude that the regularized LIML has finite rth moments for n sufficiently large but finite. At the limit when n is infinite, the moments exist by the asymptotic normality of the estimators established in Proposition 2.

Proof of Proposition 2

To prove this proposition we first need the following lemma.

Lemma 1 (Lemma A.4 of DN)
If $\hat{A} \xrightarrow{P} A$ and $\hat{B} \xrightarrow{P} B$. A is positive semi definite and B is positive definite, $\tau_0 = argmin_{\tau_1=1}\dfrac{\tau' A\tau}{\tau' B\tau}$ exists and is unique (with $\tau = (\tau_1, \tau_2')'$ and $\tau_1 \in \mathbb{R}$) then

$$\hat{\tau} = argmin_{\tau_1=1}\frac{\tau'\hat{A}\tau}{\tau'\hat{B}\tau} \rightarrow \tau_0$$

Let $H = E(f_i f_i')$, P^α is a symmetric idempotent matrix for SC and PC matrix but not necessarily for T and LF.
We want to show that $\hat{\delta} \rightarrow \delta$ as n and $n\alpha^{\frac{1}{2}}$ go to infinity.

We know that

$$\begin{aligned}\hat{\delta} &= argmin_\delta \frac{(y - W\delta)'P^\alpha(y - W\delta)}{(y - W\delta)'(y - W\delta)}\\ &= argmin_\delta \frac{(1, -\delta')\widehat{A}(1, -\delta')'}{(1, -\delta')\widehat{B}(1, -\delta')'}\end{aligned}$$

where $\hat{A} = \bar{W}'P^\alpha\bar{W}/n$, $\hat{B} = \dfrac{\bar{W}'\bar{W}}{n}$ and $\bar{W} = [y, W] = WD_0 + \varepsilon e$, where $D_0 = [\delta_0, I]$, δ_0 is the true value of the parameter and e is the first unit vector.
In fact

$$\begin{aligned}\hat{A} &= \bar{W}'P^\alpha\bar{W}/n\\ &= \frac{D_0'W'P^\alpha WD_0}{n} + \frac{D_0'W'P^\alpha\varepsilon e}{n} + \frac{e'\varepsilon'P^\alpha WD_0}{n} + \frac{e'\varepsilon'P^\alpha\varepsilon e}{n}.\end{aligned}$$

Let us define $g_n = \dfrac{1}{n}\sum_{i=1}^{n} Z(.; x_i)W_i$, $g = EZ(.; x_i)W_i$ and $\langle g, g'\rangle_K$ is a $p \times p$ matrix with (a, b) element equal to $\langle K^{-\frac{1}{2}}E(Z(., x_i)W_{ia}), K^{-\frac{1}{2}}E(Z(., x_i)W_{ib})\rangle$ where W_{ia} is the a^{th} element of W_i vector.

$$\begin{aligned}\frac{D_0'W'^\alpha WD_0}{n} &= D_0'\langle (K_n^\alpha)^{-\frac{1}{2}}g_n, (K_n^\alpha)^{-\frac{1}{2}}g_n'\rangle D_0\\ &= D_0'HD_0 + o_p(1)\\ &\rightarrow D_0'\langle g, g'\rangle_K D_0\end{aligned}$$

as n and $n\alpha^{\frac{1}{2}}$ go to infinity and $\alpha \rightarrow 0$, see the proof of Proposition 1 of Carrasco

(2012).

We also have by Lemma 3 of Carrasco (2012):

$$\frac{D_0' W'^{\alpha} \varepsilon e}{n} = D_0' \langle (K_n^{\alpha})^{-\frac{1}{2}} g_n, (K_n^{\alpha})^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} Z(.; x_i)\varepsilon_i \rangle e = o_p(1),$$

$$\frac{e' \varepsilon'^{\alpha} W D_0}{n} = e' \langle (K_n^{\alpha})^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} Z(.; x_i)\varepsilon_i, (K_n^{\alpha})^{-\frac{1}{2}} g_n' \rangle D_0 = o_p(1),$$

$$\frac{e' \varepsilon'^{\alpha} \varepsilon e}{n} = e' \langle (K_n^{\alpha})^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} Z(.; x_i)\varepsilon_i, (K_n^{\alpha})^{-\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} Z(.; x_i)\varepsilon_i' \rangle e = o_p(1).$$

We can then conclude that $\hat{A} \to A = D_0' \langle g, g' \rangle_K D_0$ as n and $n\alpha^{\frac{1}{2}}$ go to infinity and $\alpha \to 0$.

Note that $H = \langle g, g' \rangle_K$ because by assumption $g_a = E(Z(., x_i) f_{ia})$ belongs to the range of K. Let $L^2(Z)$ be the closure of the space spanned by $\{Z(x, \tau), \tau \in I\}$ and g_1 be an element of this space. If $f_i \in L^2(Z)$ we can compute the inner product and show that $\langle g_a, g_b \rangle_K = E(f_{ia} f_{ib})$ by applying Theorem 6.4 of Carrasco, Florens, and Renault (2007). Thus $A = D_0' H D_0$.

$$\hat{B} \to B = E(\bar{W}_i \bar{W}_i')$$

by the law of large numbers with $\bar{W}_i = [y_i \; W_i']'$.

The LIML estimator is given by

$$\hat{\delta} = argmin_{\delta} \frac{(1, -\delta')\widehat{A}(1, -\delta')'}{(1, -\delta')\widehat{B}(1, -\delta')'},$$

so that it suffices to verify the hypotheses of Lemma 1.

For $\tau = (1, -\delta')$

$$\begin{aligned} \tau' A \tau &= \tau' D_0' H D_0 \tau \\ &= (\delta_0 - \delta) H (\delta_0 - \delta)' \\ &= (\delta_0 - \delta) E(f_i f_i') (\delta_0 - \delta)' \end{aligned}$$

Because H is positive definite, we have $\tau' A \tau \geq 0$, with equality if and only if $\delta = \delta_0$.

Also, for any $\tau = (\tau_1, \tau_2')' \neq 0$ partitioned conformably with $(1, \delta')$, we have

$$\begin{aligned}
\tau' B\tau &= E[(\tau_1 y_i + W_i'\tau_2)^2] \\
&= E[(\tau_1\varepsilon_i + (f_i + u_i)'(\tau_1\delta_0 + \tau_2))^2] \\
&= E[(\tau_1\varepsilon_i + u_i'(\tau_1\delta_0 + \tau_2))^2] + (\tau_1\delta_0 + \tau_2)'H(\tau_1\delta_0 + \tau_2)
\end{aligned}$$

Then by H nonsingular $\tau' B\tau > 0$ for any τ with $\tau_1\delta_0 + \tau_2 \neq 0$. If $\tau_1\delta_0 + \tau_2 = 0$ then $\tau_1 \neq 0$ and hence $\tau' B\tau = \tau_1^2\sigma^2 > 0$. Therefore B is positive definite.

It follows that $\delta = \delta_0$ is the unique minimum of $\dfrac{\tau' A\tau}{\tau' B\tau}$.

Now by Lemma 1 we can conclude that $\hat{\delta} \xrightarrow{p} \delta_0$ as n and $n\alpha^{\frac{1}{2}}$ go to infinity.

Proof of asymptotic normality:

Let $A(\delta) = (y - W\delta)' P^\alpha (y - W\delta)/n$, $B(\delta) = (y - W\delta)'(y - W\delta)/n$ and $\Lambda(\delta) = \dfrac{A(\delta)}{B(\delta)}$. We know that the LIML is $\hat{\delta} = argmin\Lambda(\delta)$.

The gradient and Hessian are given by

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

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

Then by a standard mean-value expansion of the first-order conditions $\Lambda_\delta(\hat{\delta}) = 0$ with probability one.

$$\sqrt{n}(\hat{\delta} - \delta_0) = -\Lambda_{\delta\delta}^{-1}(\tilde{\delta})\sqrt{n}\Lambda_\delta(\delta_0)$$

where $\tilde{\delta}$ is the mean-value. By Lemma 1, $\tilde{\delta} \xrightarrow{P} \delta_0$.

It then follows, as in the proof of Lemma 1, that $B(\tilde{\delta}) \xrightarrow{P} \sigma_\varepsilon^2$, $B_\delta(\tilde{\delta}) \xrightarrow{P} -2\sigma_{u\varepsilon}$, $\Lambda(\tilde{\delta}) \xrightarrow{p} 0$, $\Lambda_\delta(\tilde{\delta}) \xrightarrow{P} 0$ where $\sigma_{u\varepsilon} = E(u_i\varepsilon_i)$ and $B_{\delta\delta}(\tilde{\delta}) = 2W'W/n \xrightarrow{P} 2E(W_iW_i')$, $A_{\delta\delta}(\tilde{\delta}) = 2W'P^\alpha W/n \xrightarrow{P} 2H$ with $H = E(f_i f_i')$.

So that $\tilde{\sigma}^2\Lambda_{\delta\delta}(\tilde{\delta})/2 \xrightarrow{P} H$ with $\tilde{\sigma}^2 = \varepsilon'\varepsilon/n$.

Let $\hat{\phi} = \dfrac{W'\varepsilon}{\varepsilon'\varepsilon}$, $\phi = \dfrac{\sigma_{u\varepsilon}}{\sigma_\varepsilon^2}$ and $v = u - \varepsilon\phi'$. We have $v'P^\alpha\varepsilon/\sqrt{n} = O_p(1/\sqrt{n\alpha}) = o_p(1)$ using Lemma 5(iii) of Carrasco (2012) and $E\left(v_i\varepsilon_i\right) = 0$.

$\hat{\phi} - \phi = O_p(1/\sqrt{n})$ by the Central limit theorem and delta method. Also $\varepsilon' P^\alpha\varepsilon = O_p(1/\alpha)$ as in Carrasco (2012). So that $(\hat{\phi} - \phi)\varepsilon' P^\alpha\varepsilon/\sqrt{n} = O_p(1/n\alpha) = o_p(1)$.

Furthermore, $f'\left(I - P^\alpha\right)\varepsilon/\sqrt{n} = O_p(\Delta_\alpha^2) = o_p(1)$ by Lemma 5(ii) of Carrasco

(2012) with $\Delta_\alpha = tr(f'\,(I - P^\alpha)^2\, f/n)$.

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

The conclusion follows from Slutzky theorem.

Proof of Proposition 3

To prove this proposition, we need some preliminary result. To simplify, we omit the hats on λ_j and ϕ_j and we denote P^α and $q(\alpha, \lambda_j)$ by P and q_j in the sequel.

Lemma 2:

Let $\tilde{\Lambda} = \varepsilon' P \varepsilon/(n\sigma_\varepsilon^2)$ and $\hat{\Lambda} = \Lambda(\hat{\delta})$ with $\Lambda(\delta) = \dfrac{(y - W\delta)'P(y - W\delta)}{(y - W\delta)'(y - W\delta)}$. If the assumptions of Proposition 2 are satisfied, then

$$\begin{aligned}
\hat{\Lambda} &= \tilde{\Lambda} - (\widehat{\sigma_\varepsilon^2}/\sigma_\varepsilon^2 - 1)\tilde{\Lambda} - \varepsilon' f (f'f)^{-1} f'\varepsilon/2n\sigma_\varepsilon^2 + \hat{R}_\Lambda \\
&= \tilde{\Lambda} + o(1/n\alpha), \\
\sqrt{n}\hat{R}_\Lambda &= o(\rho_{\alpha,n}),
\end{aligned}$$

where $\rho_{\alpha,n} = trace(S(\alpha))$.

Proof of Lemma 2: It can be shown similarly to the calculations in Proposition 1 that $\Lambda(\delta)$ is three times continuously differentiable with derivatives that are bounded in probability uniformly in a neighborhood of δ_0. For any $\tilde{\delta}$ between δ_0 and $\hat{\delta}$, $\Lambda_{\delta\delta}(\tilde{\delta}) = \Lambda_{\delta\delta}(\delta_0) + O(1/\sqrt{n})$. It implies that

$$\hat{\delta} = \delta_0 + [\Lambda_{\delta\delta}(\delta_0)]^{-1}\Lambda_\delta(\delta_0) + O(1/n).$$

Then expanding $\Lambda(\hat{\delta})$ around δ_0 gives

$$\begin{aligned}
\hat{\Lambda} &= \Lambda(\delta_0) - (\hat{\delta} - \delta_0)'\Lambda_{\delta\delta}(\delta_0)(\hat{\delta} - \delta_0)/2 + O(1/n^{3/2}) \\
&= \Lambda(\hat{\delta_0}) - \Lambda_\delta(\delta_0)'[\Lambda_{\delta\delta}(\delta_0)]^{-1}\Lambda_\delta(\delta_0)/2 + O(1/n^{3/2}).
\end{aligned}$$

As in proof of Proposition 1 and in Lemma A.7 of DN

$-\sqrt{n}\hat{\sigma}_\varepsilon^2\Lambda_\delta(\delta_0)/2 = h + O_p(\Delta_\alpha^{1/2} + \sqrt{1/n\alpha})$ with $h = f'\varepsilon/n$. Moreover,

$$\hat{\sigma}_\varepsilon^2\Lambda_{\delta\delta}(\delta_0)/2 = \bar{H} + O_p(\Delta_\alpha^{1/2} + \sqrt{1/n\alpha})$$

And by combining these two equalities, we obtain

$$\Lambda_\delta(\delta_0)'[\Lambda_{\delta\delta}(\delta_0)]^{-1}\Lambda_\delta(\delta_0) = h'\bar{H}^{-1}h/(n\sigma_\varepsilon^2) + O(\Delta_\alpha^{1/2}/n + \sqrt{1/(n^3\alpha)}).$$

Note also that

$$\begin{aligned}
\Lambda(\delta_0) &= (\sigma_\varepsilon^2/\hat{\sigma}_\varepsilon^2)\tilde{\Lambda} = \tilde{\Lambda} - (\hat{\sigma}_\varepsilon^2/\sigma_\varepsilon^2 - 1)\tilde{\Lambda} + \tilde{\Lambda}(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)^2/(\hat{\sigma}_\varepsilon^2\sigma_\varepsilon^2) \\
&= \tilde{\Lambda} - (\hat{\sigma}_\varepsilon^2/\sigma_\varepsilon^2 - 1)\tilde{\Lambda} + O(\sqrt{1/n^3\alpha})
\end{aligned}$$

$$\begin{aligned}
\rho_{\alpha n} &= tr(S(\alpha)) \\
&= tr(\sigma_\varepsilon^2\bar{H}^{-1}[\Sigma_v\frac{tr(P^2)}{n} + \frac{f'(I-P)^2f}{n}]\bar{H}^{-1}) \\
&= tr(\sigma_\varepsilon^2\bar{H}^{-1}[\Sigma_v\frac{tr(P^2)}{n}]\bar{H}^{-1}) + tr(\sigma_\varepsilon^2\bar{H}^{-1}[\frac{f'(I-P)^2f}{n}]\bar{H}^{-1}) \\
&= O(1/n\alpha) + \Delta_\alpha.
\end{aligned}$$

We then have that $\sqrt{n}\sqrt{1/(n^3\alpha)} = o(\rho_{\alpha n})$ and $\sqrt{n}\Delta_\alpha^{1/2}/n = o(\rho_{\alpha n})$. Using this and combining equations give

$$\hat{\Lambda} = \tilde{\Lambda} - (\hat{\sigma_\varepsilon^2}/\sigma_\varepsilon^2 - 1)\tilde{\Lambda} - \varepsilon'f(f'f)^{-1}f'\varepsilon/2n\sigma_\varepsilon^2 + \hat{R}_\Lambda$$

and

$$\sqrt{n}\hat{R}_\Lambda = o(\rho_{\alpha,n}).$$

By using $\tilde{\Lambda} = O(1/n\alpha)$, it is easy to prove that $\hat{\Lambda} = \tilde{\Lambda} + o(1/n\alpha)$.

Lemma 3: If the assumptions of Proposition 3 are satisfied, then

i) $u'Pu - \tilde{\Lambda}\Sigma_u = o(1/n\alpha)$,

ii) $E(h\tilde{\Lambda}\varepsilon'v/\sqrt{n}|X) = (tr(P)/n)\sum_i f_iE(\varepsilon_i^2v_i'|x_i)/n + O(1/(n^2\alpha))$,

iii) $E(hh'\bar{H}^{-1}h/\sqrt{n}|X) = O(1/n)$.

Proof of Lemma 3: For the proof of i), note that $E(\tilde{\Lambda}/X) = tr(PE(\varepsilon'\varepsilon))/n\sigma_\varepsilon^2 = tr(P)/n$.

Similarly, we have $E(u'Pu|X) = tr(P)\Sigma_u$ and by Lemma 5 (iv) of Carrasco (2012) using ε in place of u we have

$$E[(\tilde{\Lambda} - tr(P)/n)^2|X] = [\sigma_\varepsilon^4 tr(P)^2 + o(tr(P)^2)]/(n^2\sigma_\varepsilon^4) - (tr(P)/n)^2 = o((tr(P)/n)^2).$$

Thus, $(\tilde{\Lambda} - tr(P)/n)\Sigma_u = o(tr(P)/n) = o(1/n\alpha)$ by Markov inequality.

And $u'Pu - \dfrac{tr(P)}{n}\Sigma_u = o(1/n\alpha)$ such that $u'Pu - \tilde{\Lambda}\Sigma_u = o(1/(n\alpha))$.

To show ii) we can notice that

$$\begin{aligned}
E(h\tilde{\Lambda}\varepsilon'v/\sqrt{n}|X) &= E(h\varepsilon'P\varepsilon\varepsilon'v/(n\sigma_\varepsilon^2\sqrt{n})|X) \\
&= \sum_{i,j,k,l} E(f_i\varepsilon_i\varepsilon_j P_{jk}\varepsilon_k\varepsilon_l v_l'^2\sigma_\varepsilon^2)|X) \\
&= \sum_i f_i P_{ii} E(\varepsilon_i^4 v_i'|x_i)/n^2\sigma_\varepsilon^2 + 2\sum_{i\neq j} f_i P_{ij} E(\varepsilon_j^2 v_j'|x_j)/n^2 \\
&+ \sum_{i\neq j} f_i P_{jj} E(\varepsilon_i^2 v_i|x_i)/n^2 \\
&= O(1/n) + (tr(P)/n)\sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n
\end{aligned}$$

This is true because $E(\varepsilon_i^4 v_i'|x_i)$ and $E(\varepsilon_i^2 v_i'|x_i)$ are bounded by Assumption 2 hence $f'P\mu/n$ is bounded for $\mu_i = E(\varepsilon_i^4 v_i'|x_i)$ and $\mu_i = E(\varepsilon_i^2 v_i'|x_i)$.

For iii)

$$\begin{aligned}
E(hh'\bar{H}^{-1}h/\sqrt{n}|X) &= \sum_{i,j,k} E(f_i\varepsilon_i\varepsilon_j f_j'\bar{H}^{-1}f_k\varepsilon_k|X)/n^2 \\
&= \sum_i E(\varepsilon_i^3|x_i)f_i f_i'\bar{H}^{-1}f_i/n^2 \\
&= O(1/n).
\end{aligned}$$

Now we turn to the proof of Proposition 3.

Proof of Proposition 3

Our proof strategy will be very close to those of Carrasco (2012) and DN. To obtain the LIML, we solve the following first order condition

$$W'P(y - W\hat{\delta}) - \hat{\Lambda}W'(y - W\hat{\delta}) = 0$$

with $\hat{\Lambda} = \Lambda(\hat{\delta})$.

Let us consider $\sqrt{n}(\hat{\delta} - \delta) = \hat{H}^{-1}\hat{h}$ with $\hat{H} = W'PW/n - \hat{\Lambda}W'W/n$ and

$$\hat{h} = W'P\varepsilon/\sqrt{n} - \hat{\Lambda}W'\varepsilon/\sqrt{n}.$$

As in Carrasco (2012) we are going to apply Lemma A.1 of DN[7].

$\hat{h} = h + \sum_{j=1}^{5} T_j^h + Z^h$ with $h = f'\varepsilon/\sqrt{n}$,

$T_1^h = -f'(I - P)\varepsilon/\sqrt{n} = O(\Delta_\alpha^{1/2})$

$T_2^h = v'P\varepsilon/\sqrt{n} = O(\sqrt{1/n\alpha})$, $T_3^h = -\tilde{\Lambda}h' = O(1/n\alpha)$, $T_4^h = -\tilde{\Lambda}v'\varepsilon/\sqrt{n} = O(1/n\alpha)$,

$T_5^h = h'\bar{H}^{-1}h\sigma_{u\varepsilon}/2\sqrt{n}\sigma_\varepsilon^2 = O(1/\sqrt{n})$,

$Z^h = -\hat{R}_\Lambda W'\varepsilon/\sqrt{n} - (\hat{\Lambda} - \tilde{\Lambda} - \hat{R}_\Lambda)\sqrt{n}(W'\varepsilon/n - \sigma'_{u\varepsilon})$ where $\hat{R}_\Lambda$ is defined in Lemma 2.

By using the central limit theorem on$\sqrt{n}(W'\varepsilon/n - \sigma'_{u\varepsilon})$ and Lemma 2, $Z^h = O(\rho_{n\alpha})$.

The results on order of T_j^h hold by Lemma 5 Carrasco (2012).

We also have

$\hat{H} = \bar{H} + \sum_{j=1}^{3} T_j^H + Z^H$,

$T_1^H = -f'(I - P)f/n = O(\Delta_\alpha)$, $T_2^H = (u'f + f'u)/n = O(1/\sqrt{n})$,

$T_3^H = -\tilde{\Lambda}\bar{H} = O(1/n\alpha)$,

$Z^H = u'Pu/n - \tilde{\Lambda}\Sigma_u - \hat{\Lambda}W'W/n + \tilde{\Lambda}(\bar{H} + \Sigma_u) - u'(I - P)f/n - f'(I - P)u/n$.

By Lemma 3, $u'Pu/n - \tilde{\Lambda}\Sigma_v = o(1/n\alpha)$. Lemma 5 (ii) of Carrasco (2012) implies $u'(I - P)f/n = O(\Delta_\alpha^{1/2}/\sqrt{n}) = o(\rho_{n\alpha})$. By the central limit theorem, $W'W/n = \bar{H} + \Sigma_u + O(1/\sqrt{n})$.

$$\begin{aligned} \hat{\Lambda}W'W/n - \tilde{\Lambda}(\bar{H} + \Sigma_u) &= (\hat{\Lambda} - \tilde{\Lambda})W'W/n + \tilde{\Lambda}(W'W/n - \bar{H} - \Sigma_u) \\ &= o(1/n\alpha) + O(1/n\alpha)O(1/\sqrt{n}) = o(\rho_{n\alpha}) \end{aligned}$$

thus, $Z^H = o(\rho_{n\alpha})$.

[7] The expression of T_5^h, Z^h and Z^H below correct some sign errors in DN

We apply Lemma A.1 of DN with
$T^h = \sum_{j=1}^{5} T_j^h$, $T^H = \sum_{j=1}^{3} T_j^H$,

$$Z^A = (\sum_{j=3}^{5} T_j^h)(\sum_{j=3}^{5} T_j^h)' + (\sum_{j=3}^{5} T_j^h)(T_1^h + T_2^h)' + (T_1^h + T_1^h)(\sum_{j=3}^{5} T_j^h)',$$

and

$$\hat{A}(\alpha) = hh' + \sum_{j=1}^{5} hT_j^{h'} + \sum_{j=1}^{5} T_j^h h' + (T_1^h + T_2^h)(T_1^h + T_2^h)' - hh'\bar{H}^{-1}\sum_{j=1}^{3} T_j^{H'} - \sum_{j=1}^{3} T_j^H \bar{H}^{-1} hh'.$$

Note that $hT_3^{h'} - hh'\bar{H}^{-1}T_3^{H'} = 0$.

Also we have $E(hh'\bar{H}^{-1}(T_1^H + T_2^H)|X) = -\sigma_\varepsilon^2 e_f(\alpha) + O(1/n)$, $E(T_1^h h') = E(hT_1^{h'}) = -\sigma_\varepsilon^2 e_f(\alpha)$, $E(T_1^h T_1^{h'}) = \sigma_\varepsilon^2 e_{2f}(\alpha)$ where
$e_f(\alpha) = \dfrac{f'(I-P)f}{n}$ and $e_{2f}(\alpha) = \dfrac{f'(I-P)^2 f}{n}$.
By Lemma 3 (ii) $E(hT_4^{h'}|X) = \dfrac{tr(P)}{n} \sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n + O\left(\dfrac{1}{n^2\alpha}\right)$.

By Lemma 5 (iv) of Carrasco (2012), with v in place of u and noting that $\sigma_{v\varepsilon} = 0$, we have

$$E(T_2^h T_2^{h'}|X) = \sigma_\varepsilon^2 \Sigma_v \frac{tr(P^2)}{n},$$

$$E(hT_2^{h'}|X) = \sum_i P_{ii} f_i E(\varepsilon_i^2 v_i'|x_i)/n.$$

By Lemma 3 (iii), $E(hT_5^{h'}) = O(1/n)$.
For $\hat{\xi} = \sum_i P_{ii} f_i E(\varepsilon_i^2 v_i'|x_i)/n - \dfrac{tr(P)}{n} \sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n - \sum_i P_{ii}(1-P_{ii}) f_i E(\varepsilon_i^2 v_i'|x_i)/n$,
$\hat{A}(\alpha)$ satisfies

$$E(\hat{A}(\alpha)|X) = \sigma_\varepsilon^2 \bar{H} + \sigma_\varepsilon^2 \Sigma_v \frac{tr(P^2)}{n} + \sigma_\varepsilon^2 e_{2f} + \hat{\xi} + \hat{\xi}' + O(1/n)$$

We can also show that $\|T_1^h\|\|T_j^h\| = o(\rho_{n\alpha})$, $\|T_2^h\|\|T_j^H\| = o(\rho_{n\alpha})$ for each j and $\|T_k^h\|\|T_j^H\| = o(\rho_{n\alpha})$ for each j and $k > 2$. Furthermore $\|T_j^H\|^2 = o(\rho_{n\alpha})$ for each j. It follows that $Z^A = o(\rho_{n\alpha})$. It can be noticed that all conditions of Lemma A.1 of DN are satisfied and the result follows by observing that $E(\varepsilon_i^2 v_i'|x_i) = 0$. This ends the proof of Proposition 3.

To prove the Proposition 4 we need to establish the following result.

Lemma 4 (Lemma A.9 of DN): If $\sup_{\alpha\in M_n}(|\hat{S}_\gamma(\alpha)-S_\gamma(\alpha)|/S_\gamma(\alpha))\overset{P}{\to}0$, then $S_\gamma(\hat{\alpha})/\inf_{\alpha\in M_n}S_\gamma(\alpha)\overset{P}{\to}1$ as n and $n\alpha\to\infty$

Proof of Lemma 4: We have that $\inf_{\alpha\in M_n}S_\gamma(\alpha)=S_\gamma(\alpha^*)$ for some α^* in M_n by the finiteness of the index set for $1/\alpha$ for PC, SC and LF and by the compactness of the index set for T. Then, the proof of Lemma 4 follows from that of Lemma A.9 of DN.

Proof of Proposition 4

We proceed by verifying the assumption of Lemma 4.

Let $R(\alpha)=\dfrac{f_\gamma'(I-P)^2f_\gamma}{n}+\sigma_{u_\gamma}^2\dfrac{tr(P^2)}{n}$ be the risk approximated by $\hat{R}^m(\alpha)$, $\hat{R}^{cv}(\alpha)$, or $\hat{R}^{lcv}(\alpha)$, and $S_\gamma(\alpha)=\sigma_\varepsilon^2\left[\dfrac{f_\gamma'(I-P)^2f_\gamma}{n}+\sigma_{v_\gamma}^2\dfrac{tr(P^2)}{n}\right]$. For notational convenience, we henceforth drop the γ subscript on S and R. For Mallows C_p, generalized cross-validation and leave one out cross-validation criteria, we have to prove that

$$\sup_{\alpha\in M_n}\left(|\hat{R}(\alpha)-R(\alpha)|/R(\alpha)\right)\to 0 \tag{6}$$

in probability as n and $n\alpha\to\infty$.

To establish this result, we need to verify the assumptions of Li's (1987, 1986) theorems. We treat separately the regularizations with a discrete index set and that with a continuous index set (Tikhonov regularization). SC and LF have a discrete index set in terms of $1/\alpha$.

Discrete index set:

We recall the assumptions of Li (1987) (A.1) to (A.3') for $m=2$.

(A.1) $\lim_{n\to\infty}\sup_{\alpha\in M_n}\lambda(P)<\infty$ where $\lambda(P)$ is the largest eigenvalue of P;

(A.2) $E((u_ie)^8)<\infty$;

(A.3') $\inf_{\alpha\in M_n}nR(\alpha)\to\infty$.

(A.1) is satisfied because for every $\alpha\in M_n$, all eigenvalues $\{q_j\}$ of P are less than or equal to 1.

(A.2) holds by our assumption 4 (i).

For (A.3'), note that $nR(\alpha)=f_\gamma'(I-P)^2f_\gamma+\sigma_{u_\gamma}^2tr(P^2)=O_p\left(n\alpha^\beta+\dfrac{1}{\alpha}\right)$.

Minimizing w.r. to α gives

$$\alpha = \left(\frac{1}{n\beta}\right)^{\frac{1}{1+\beta}}$$

hence $\inf_{\alpha \in M_n} nR(\alpha) \approx n\alpha^{\beta} \to \infty$, therefore the condition (A.3') is satisfied for SC and LF (and T also).

Note that Theorem 2.1 of Li (1987) use assumption (A.3) instead of (A.3'). However, Corollary 2.1 of Li (1987) justifies using (A.3') when P is idempotent which is the case for SC. For LF, P is not idempotent however the proof provided by Li (1987) still applies. Given $tr(P^2) = O_p\left(\frac{1}{\alpha}\right)$ for LF, we can argue that for n large enough, there exists a constant C such that

$$tr(P^2) \geq \frac{C}{n},$$

hence Equation 2.6 of Li (1987) holds and Assumption (A.3) can be replaced by (A.3'). The justification for replacing $\sigma^2_{u_\gamma \varepsilon}$, $\sigma^2_{u_\gamma}$ and σ^2_{ε} by their estimates in the criteria is the same as in the proof of Corollary 2.2 in Li (1987).

For the generalized cross-validation, we need to verify the assumptions of Li's (1987) Theorem 3.2. that are recalled below.

(A.4) $\inf_{\alpha \in M_n} n^{-1} \|f_\gamma - PW_\gamma\| \to 0$;

(A5) For any sequence $\{\alpha_n \in M_n\}$ such that

$$\frac{1}{n} tr(P^2) \to 0,$$

we have $\left(n^{-1}tr(P)\right)^2 / (n^{-1}tr(P^2)) \to 0$;

(A.6) $\sup_{\alpha \in M_n} n^{-1}tr(P) \leq \gamma_1$ for some $0 < \gamma_1 < 1$;

(A.7) $\sup_{\alpha \in M_n} \left(n^{-1}tr(P)\right)^2 / (n^{-1}tr(P^2)) \leq \gamma_2$, for some $0 < \gamma_2 < 1$.

Assumption (A.4) holds for SC and LF from $R(\alpha) = En^{-1}\|f_\gamma - PW_\gamma\| \to 0$ as n and $n\alpha$ go to infinity.

Note that $tr\left(P\right) = O\left(\alpha^{-1}\right)$ and $tr\left(P^2\right) = O\left(\alpha^{-1}\right)$. So that $n^{-1}tr(P^2) \to 0$ if and only if $n\alpha \to \infty$. Moreover $\frac{1}{n}(tr(P))^2/tr(P^2) = O(1/n\alpha) \to 0$ as $n\alpha \to \infty$. This proves Assumption (A.5) for SC and LF.

Now we turn our attention to Assumptions (A.6) and (A.7). By Lemma 4 of Carrasco (2012), we know that $tr(P) \leq C_1/\alpha$ and $tr(P^2) \leq C_2/\alpha$. To establish Assumptions (A.6) and (A.7), we restrict the set M_n to the set $M_n = \{\alpha : \alpha > C/n \text{ with } C > \max(C_1, C_1^2/C_2)\}$. This is not very restrictive since α has to satisfy $n\alpha \to \infty$. It follows that

$$\begin{aligned}
\sup_{\alpha \in M_n} tr(P)/n &= \sup_{\alpha > C/n} tr(P)/n \leq \frac{C_1}{C} < 1, \\
\sup_{\alpha \in M_n} \frac{1}{n}(tr(P))^2/tr(P^2) &= \sup_{\alpha > C/n} \frac{1}{n}(tr(P))^2/tr(P^2) \leq \frac{C_1^2}{CC_2} < 1.
\end{aligned}$$

Thus, Assumptions (A.6) and (A.7) hold.

In the case of leave-one-out cross-validation criterion, we need to verify the sssumptions of Theorem 5.1 of Li (1987). Assumption (A.1) to (A.4) still hold as before. Assumptions (A.8), (A.9), and (A.10) hold by Assumption 4 (iii) to (v) of this paper, respectively. This ends the proof of (6) for SC and LF.

Continuous index set

The T regularization is a case where the index set is continuous. We apply Li's (1986) results on the optimality of Mallows C_p in the ridge regression. We need to check the Assumption (A.1) of Theorem 1 in Li (1986). (A.1) $\inf_{\alpha \in M_n} nR(\alpha) \to \infty$ holds using the same proof as for SC and LF. It follows that (6) holds for T under Assumption 4 (i').

Given $\sigma_\varepsilon^2 \neq 0$ we have $R(\alpha) \leq CS_\gamma(\alpha)/\sigma_\varepsilon^2$. To see this, replace $R(\alpha)$ and $S_\gamma(\alpha)$ by their expressions in function of $\dfrac{f_\gamma' (I-P)^2 f_\gamma}{n}$ and use the fact that $\sigma_{u_\gamma}^2 > \sigma_{v_\gamma}^2$ and take $C = \sigma_{u_\gamma}^2/\sigma_{v_\gamma}^2$.

$$\begin{aligned}
|\hat{S}_\gamma(\alpha) - S_\gamma(\alpha)| &= \sigma_\varepsilon^2 \left| \left(\hat{R}(\alpha) - \frac{\hat{\sigma}_{u_\gamma\varepsilon}^2}{\hat{\sigma}_\varepsilon^2} \frac{tr(P^2)}{n} \right) - \left(\sigma_{v_\gamma}^2 \frac{tr(P^2)}{n} + \frac{f_\gamma' (I-P)^2 f_\gamma}{n} \right) \right| \\
&= \sigma_\varepsilon^2 \left| \hat{R}(\alpha) - \frac{f_\gamma' (I-P)^2 f_\gamma}{n} - \left(\sigma_{v_\gamma}^2 + \frac{\hat{\sigma}_{u_\gamma\varepsilon}^2}{\hat{\sigma}_\varepsilon^2} \right) \frac{tr(P^2)}{n} \right| \\
&= \sigma_\varepsilon^2 \left| \hat{R}(\alpha) - R(\alpha) + \sigma_{u_\gamma}^2 \frac{tr(P^2)}{n} - \left(\sigma_{v_\gamma}^2 + \frac{\hat{\sigma}_{u_\gamma\varepsilon}^2}{\hat{\sigma}_\varepsilon^2} \right) \frac{tr(P^2)}{n} \right| \\
&\leq \sigma_\varepsilon^2 \left| \hat{R}(\alpha) - R(\alpha) \right| + \sigma_\varepsilon^2 \left| \left(\frac{\hat{\sigma}_{u_\gamma\varepsilon}^2}{\hat{\sigma}_\varepsilon^2} - \frac{\sigma_{u_\gamma\varepsilon}^2}{\sigma_\varepsilon^2} \right) \frac{tr(P^2)}{n} \right|
\end{aligned}$$

Using $S_\gamma(\alpha) \geq \sigma_\varepsilon^2 \sigma_{v_\gamma}^2 \dfrac{tr(P^2)}{n}$ and $R(\alpha) \leq C S_\gamma(\alpha)/\sigma_\varepsilon^2$, we have

$$\frac{|\hat{S}_\gamma(\alpha) - S_\gamma(\alpha)|}{S_\gamma(\alpha)} \leq C \frac{|\hat{R}(\alpha) - R(\alpha)|}{R(\alpha)} + \frac{|\frac{\hat{\sigma}_{u_\gamma \varepsilon}^2}{\hat{\sigma}_\varepsilon^2} - \frac{\sigma_{u_\gamma \varepsilon}^2}{\sigma_\varepsilon^2}|}{\sigma_{v_\gamma}^2}.$$

It follows from (6) and Assumption 4(ii) that $\sup_{\alpha \in M_n} |\hat{S}_\gamma(\alpha) - S_\gamma(\alpha)|/S_\gamma(\alpha) \to 0$. The optimality of the selection criteria follows from Lemma 4. This ends the proof of Proposition 4.