Journal of Econometrics 186 (2015) 427–442

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Regularized LIML for many instruments☆

Marine Carrasco [a], Guy Tchuente [b,*]

[a] University of Montreal, CIREQ, CIRANO, Canada

[b] University of Kent, School of Economics, Keynes College, Canterbury, Kent CT2 7NP, United Kingdom

ARTICLE INFO

Article history:
Available online 5 March 2015

JEL classification:
C13

Keywords:
Heteroskedasticity
High-dimensional models
LIML
Many instruments
MSE
Regularization methods

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 asymptotically normal under heteroskedastic error. Moreover, they reach the semiparametric efficiency bound assuming homoskedastic error. We show that the regularized LIML estimators possess 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 and two empirical applications illustrate the relevance of our estimators.

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. 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 estimate returns 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 et al. (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). 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 a regularized version is not available for the limited information maximum likelihood (LIML). Providing such an estimator is desirable, given LIML has better properties than 2SLS (see e.g. Hahn and Inoue (2002), Hahn and Hausman (2003), and Hansen et al., 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 et al. (2007)). The three regularization techniques were also used 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 spectral cut-off or principal components, 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. We restrict our attention to the case where the parameters are strongly identified and the estimators converge at the usual $\sqrt{n}$ rate. However, a subset of instruments may be irrelevant.

We show that the regularized LIML estimators are consistent and asymptotically normal under heteroskedastic error. Moreover, they reach the semiparametric efficiency bound in presence of homoskedastic error. We show that the regularized LIML has finite

☆ The authors thank a co-editor and two referees for insightful comments. They 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, the University of Pennsylvania, and Queen's University for helpful comments.

* Corresponding author. Tel.: +44 1227827249.
E-mail addresses: marine.carrasco@umontreal.ca (M. Carrasco), g.tchuente@kent.ac.uk (G. Tchuente).

http://dx.doi.org/10.1016/j.jeconom.2015.02.018

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 Donald and Newey (2001), Okui (2011), and Carrasco (2012), we propose a data-driven method for selecting the regularization parameter, α, based on a cross-validation approximation of the MSE. We show that this selection method is optimal in the sense of Li (1986, 1987), meaning that the choice of α using the estimated MSE is asymptotically as good as if minimizing the true unknown MSE.

The simulations show that the regularized LIML is better than the regularized 2SLS in almost every case. Simulations show that the LIML estimator based on Tikhonov and Landweber–Fridman regularizations often have smaller median bias and smaller MSE than the LIML estimator based on principal components and than the 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, L, grows with the sample size, n, but remains smaller than n. In this case, the 2SLS estimator is inconsistent while LIML is consistent (see Bekker (1994), Chao and Swanson (2005), Hansen et al. (2008), among others). Hausman et al. (2012) and Chao et al. (2012) give modified LIML estimators which are robust to heteroskedasticity in the presence of many weak instruments. Modifications of GMM have been considered by Canay (2010) and Kuersteiner (2012) who consider kernel weighted GMM estimators and Okui (2011) who uses shrinkage. Recently, some work has been done in the case where the number of instruments exceeds the sample size. 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 et al. (2012a) assume the approximate sparsity of the first stage equation and apply an instrument selection based on Lasso. Recently, Hansen and Kozbur (2014) propose a ridge regularized jackknife instrumental variable estimator in the presence of heteroskedasticity 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 selects the number of instruments by minimizing an approximate MSE. Our method assumes neither a strong factor structure, nor an exactly sparse first stage equation. However, it assumes that the instruments are sufficiently correlated among themselves so that the trace of the instruments covariance matrix is finite and hence the eigenvalues of the covariance matrix decrease to zero sufficiently fast.

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 show that the regularized LIML estimators are consistent and asymptotically normal in presence of heteroskedastic error and they reach the semiparametric efficiency bound assuming homoskedasticity. Moreover, we establish that, under some conditions, they have finite moments.

2.1. Presentation of the estimators

The model is

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

$i = 1, 2, \ldots, n$. The main focus is the estimation of the $p \times 1$ vector δ_0. y_i is a scalar and x_i is a vector of exogenous variables. W_i is correlated with ε_i so that the ordinary least-squares estimator is not consistent. Some rows of W_i may be exogenous, with the corresponding rows of u_i being zero. A set of instruments, Z_i, is available so that $E(Z_i\varepsilon_i) = 0$. The estimation of δ is based on the orthogonality condition:

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

Let $f(x_i) = E(W_i|x_i) \equiv f_i$ denote the $p \times 1$ reduced form vector. The notation $f(x_i)$ covers various cases. $f(x_i)$ may be a linear combination of a large dimensional (possibly infinite dimensional) vector x_i. Let $Z_i = x_i$, then $f(x_i) = \beta' Z_i$ for some $L \times p\beta$. Some of the coefficients β_j may be equal to zero, in which case the corresponding instruments Z_j are irrelevant. In that sense, $f(x_i)$ may be sparse as in Belloni et al. (2012b). The instruments have to be strong as a whole but some of them may be irrelevant. We do not consider the case where the instruments are weak (case where the correlation between W_i and Z_i converges to zero at the $\sqrt{n}$ rate) and the parameter δ is not identified as in Staiger and Stock (1997). We do not allow for many weak instruments (case where the correlation between W_i and Z_i declines to zero at a faster rate than $\sqrt{n}$ and the number of instruments Z_i grows with the sample size) considered by Newey and Windmeijer (2009) among others.

The model allows for x_i to be a few variables and Z_i to approximate the reduced form $f(x_i)$. For example, Z_i could be a power series or splines (see Donald and Newey, 2001).

As in Carrasco (2012), we use a general 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 set of instruments $Z_i = \{Z(\tau; x_i) : \tau \in S\}$ where S is an index set. Examples of Z_i are the following.

- Assume $Z_i = x_i$ where x_i is a L-vector with a fixed L. Then $Z(\tau; x_i)$ denotes the τth element of x_i and $S = \{1, 2, \ldots, L\}$.
- $Z(\tau; x_i) = (x_i)^{\tau-1}$ with $\tau \in S = \mathbb{N}$, thus we have infinite countable instruments.
- $Z(\tau; x_i) = exp(i\tau' x_i)$ where $\tau \in S = \mathbb{R}^{dim(x_i)}$, thus we have a continuum of moments.

It is important to note that throughout the paper, the number of instruments, L, of Z_i is either fixed or infinite and L is always independent of T. We view L as the number of instruments available to the econometrician and the econometrician uses all these instruments to estimate the parameters. We need to define a space of reference in which elements such that $E(W_i Z(\tau; x_i))$ are supposed to lie. We denote $L^2(\pi)$ the Hilbert space of square integrable functions with respect to π where π is a positive measure on S. $\pi(\tau)$ attaches a weight to each moments indexed by τ. π permits to dampen the effect of some instruments. For instance, if $Z(\tau; x_i) = exp(i\tau' x_i)$, it makes sense to put more weight on low frequencies (τ close to 0) and less weight on high frequencies (τ large). In that case, a π equal to the standard normal density works well as shown in Carrasco et al. (2007).

We define the covariance operator K of the instruments as

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

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

where $\overline{Z(\tau_2; x_i)}$ denotes the complex conjugate of $Z(\tau_2; x_i)$. K is assumed to be a nuclear (also called trace-class) operator which

is satisfied if and only if its trace is finite. This assumption and the role of π are discussed in detail in Carrasco and Florens (2014). This is trivially satisfied if the number of instruments is finite. However, when it is infinite, this condition requires that the eigenvalues of K decline to zero sufficiently fast which implies some strong colinearity among the instruments. If the instruments $\{Z_{ij} : j = 1, 2, \ldots, \infty\}$ are independent from each other, then K is the infinite dimensional identity matrix which is not nuclear. However, Section 2.3 of Carrasco and Florens (2014) shows that an appropriate choice of π makes such a matrix nuclear. The weight π gives an extra degree of freedom to the econometrician to meet some of our assumptions. We will see in Section 2.2 that the asymptotic distribution of our estimator does not depend on the choice of π. In the case where the vector of instruments Z_i has a finite dimension L (potentially very large), we can select π as the uniform density on $S = \{1, 2, \ldots, L\}$. In that case, K is the operator which associates to vector v of $\mathbb{R}^L$, the vector $Kv = E\left(Z_i Z_i'\right) v/L$. The condition "$K$ nuclear" is met if the trace of $E\left(Z_i Z_i'\right)/L$ is finite. This is satisfied if the Z_{il}, $l = 1, 2, \ldots, L$ depends on a few common factors (see for instance Bai and Ng, 2002). It may be satisfied also if the eigenvalues continuously decline without having a factor structure.

Let λ_j and ϕ_j $j = 1, 2, \ldots$ 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:

$$K_n : L^2(\pi) \to L^2(\pi)$$
$$(K_n g)(\tau_1) = \int \frac{1}{n} \sum_{i=1}^{n} Z(\tau_1; x_i)\overline{Z(\tau_2; x_i)} g(\tau_2)\pi(\tau_2) d\tau_2.$$

If the number of moment conditions is infinite, inverting K is an ill-posed problem in the sense that its inverse is not continuous, moreover its sample counterpart, K_n, is singular. Consequently, the inverse of K_n needs to be stabilized via regularization. By definition (see Kress, 1999, page 269), a regularized inverse of an operator K is $R_\alpha : L^2(\pi) \to L^2(\pi)$ such that $\lim_{\alpha \to 0} R_\alpha K\varphi = \varphi, \forall\, \varphi \in L^2(\pi)$.

As in Carrasco (2012), we consider three different types of regularization schemes: Tikhonov (T), Landweber–Fridman (LF) and Spectral cut-off (SC). They are defined as follows:[1]

1. **Tikhonov(T)**
   This regularization inverse is defined as $(K^\alpha)^{-1} = (K^2 + \alpha I)^{-1} K$ or equivalently
   $$(K^\alpha)^{-1} r = \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j^2 + \alpha} \langle r, \phi_j \rangle \phi_j$$
   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:
   $$\begin{cases} \hat{\varphi}_l = (1 - cK^2)\hat{\varphi}_{l-1} + cKr, \quad l = 1, 2, \ldots, \dfrac{1}{\alpha} - 1; \\ \hat{\varphi}_0 = cKr, \end{cases}$$
   where $\frac{1}{\alpha} - 1$ is some positive integer. Equivalently, we 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 $\alpha > 0$. As the ϕ_j are related to the principal components of Z, this method is also called principal components (PC).

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

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

$$Tg = \left(\langle Z_1, g \rangle', \langle Z_2, g \rangle', \ldots, \langle Z_n, g \rangle'\right)'$$

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

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

such that $K_n = T^* T$ and TT^* is an $n \times n$ matrix with typical element $\frac{\langle Z_i, Z_j \rangle}{n}$. Let $\hat{\phi}_j$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots > 0, j = 1, 2, \ldots$ be the orthonormalized eigenfunctions and eigenvalues of K_n and ψ_j the eigenfunctions of TT^*. We then have $T\hat{\phi}_j = \sqrt{\hat{\lambda}_j}\psi_j$ and $T^*\psi_j = \sqrt{\hat{\lambda}_j}\hat{\phi}_j$. 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 $W = \left(W_1', W_2', \ldots, W_n'\right)'$ $n \times p$ and $y = \left(y_1', y_2', \ldots, y_n'\right)'$ $n \times p$. Let us define k-class estimators as

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

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

$$\nu = \nu_\alpha = \min_{\delta} \frac{(y - W\delta)' P^\alpha (y - W\delta)}{(y - W\delta)'(y - W\delta)} \tag{2}$$

corresponds to the regularized LIML estimator we will study here.

2.2. Asymptotic properties of the regularized LIML

First, we establish the asymptotic properties of the regularized LIML estimators when the errors are heteroskedastic. Next, we will consider the special case where the errors are homoskedastic and the reduced form f can be approached by a sequence of instruments. We will focus on the case where the regularization parameter, α, goes to zero. If α were bounded away from zero, our estimators would remain consistent and asymptotically normal but would be less efficient.

One of the drawbacks of LIML in the many-instruments setting is that it fails to even be consistent in presence of heteroskedasticity. We will show that the regularized LIML estimators remain consistent and asymptotically normal. Here, we assume that (ε_i, u_i') are iid but conditionally heteroskedastic. We define the covariance operator $\tilde{K}$ of the moments $\{\varepsilon_i Z_i\}$ as

$$\tilde{K} : L^2(\pi) \to L^2(\pi)$$
$$(\tilde{K} g)(\tau_1) = \int E(\varepsilon_i^2 Z(\tau_1; x_i)\overline{Z(\tau_2; x_i)}) g(\tau_2)\pi(\tau_2) d\tau_2$$

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

where $\overline{Z(\tau_2; x_i)}$ denotes the complex conjugate of $Z(\tau_2; x_i)$. K nuclear, together with the assumption $E(\varepsilon_i^2|x_i) = \sigma_i^2 < C$, implies that the operator $\widetilde{K}$ is nuclear. This, in turn, implies that a functional central limit theorem holds (see vander Vaart and Wellner, 1996, p. 50), namely $\sum_{i=1}^n Z(.; x_i)\varepsilon_i/\sqrt{n}$ converges in $L^2(\pi)$ to a mean zero Gaussian process with covariance operator $\tilde{K}$. Let g denote $E(Z(.; x_i)W_i)$ and $F = K^{-1/2}$.

Proposition 1 (*Case with Heteroskedasticity*). *Assume* $\left(y_i, W_i', x_i'\right)$ *are iid,* $E(\varepsilon_i|x_i) = E(u_i|x_i) = 0$. *Var* $\left(\left(\varepsilon_i, u_i'|x_i\right)\right)$ *depends on i.* $E(\varepsilon_i^2|x_i) = \sigma_i^2$, *where* σ_i^2 *is bounded, the operator K is nuclear, the* $p \times p$ *matrix* $\langle Fg, Fg'\rangle$ *is nonsingular. The regularization parameter* α *goes to zero. Then, the T, LF, and SC LIML estimators satisfy:*

1. *Consistency: Assume that each element of g belongs to range of* $K^{1/2}$. *Then* $\hat{\delta} \to \delta_0$ *in probability as n and* $n\alpha^{1/2}$ *go to infinity.*
2. *Asymptotic normality: If moreover, each element of g belongs to the range of K, then*
$$\sqrt{n}(\hat{\delta} - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \langle Fg, Fg'\rangle^{-1} \langle Fg, \left(F\widetilde{K}F^*\right) Fg\rangle \langle Fg, Fg'\rangle^{-1}\right)$$
as n and $\alpha\sqrt{n}$ *go to infinity.*

The condition $\langle Fg, Fg'\rangle$ nonsingular is an identification assumption. It would be interesting to compare this result with the asymptotic distribution of the regularized 2SLS estimator of Carrasco (2012). Using Theorem 2 of Carrasco and Florens (2000), it can be shown that they have the same asymptotic distribution. Hence, both types of estimators are robust to heteroskedasticity.

A consistent estimator of the asymptotic variance is given by

$$\left(W'P^\alpha W\right)^{-1}\left(W'P^\alpha \hat{\Omega} P^\alpha W\right)\left(W'P^\alpha W\right)^{-1}$$

where $\hat{\Omega}$ is $n \times n$ diagonal matrix with $\hat{\varepsilon}_i^2$ on the diagonal with $\hat{\varepsilon}_i = y_i - W_i'\tilde{\delta}$ and $\tilde{\delta}$ a consistent estimator of δ. An alternative consistent estimator is given by

$$\left(\hat{W}'W\right)^{-1}\left(\hat{W}'\hat{\Omega}\hat{W}\right)\left(W'\hat{W}\right)^{-1}$$

where $\hat{W} = (P^\alpha - \nu I_n)W$.

Next, we turn to the homoskedastic case and establish that the regularized LIML estimators asymptotically reach the semiparametric efficiency bound. Let $f_a(x)$ be the ath element of $f(x)$.

Proposition 2 (*Case with Homoskedasticity*). *Assume* $\left(y_i, W_i', x_i'\right)$ *are iid,* $E(\varepsilon_i^2|x_i) = \sigma_\varepsilon^2$, $E(f_i f_i')$ *exists and is nonsingular, K is nuclear,* α *goes to zero.* $E\left(\varepsilon_i^4|x_i\right) < C$ *and* $E\left(\|u_i\|^4|x_i\right) < C$, *for some constant C. Moreover,* $f_a(x)$ *belongs to the closure of the linear span of* $\{Z(.; x)\}$ *for* $a = 1, \ldots, p$. *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 g 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. See Appendix.

For the asymptotic normality, we need $n\alpha$ go to infinity as in Carrasco (2012) for 2SLS. It means that α is allowed to go to zero faster than for the heteroskedastic case. Indeed, in Proposition 1, the condition was $\alpha\sqrt{n}$. This improved rate for α has a cost which is the condition that the fourth moments of ε_i and u_i are bounded. We did not need this condition in Proposition 1 because a slightly different proof was used.

The assumption "$f_a(x)$ belongs to the closure of the linear span of $\{Z(.; x)\}$ for $a = 1, \ldots, p$" 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.

2.3. 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 has finite moments. In a recent paper, Hausman et al. (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 estimators possess finite moments provided the sample size is large enough. Let $X = (x_1, x_2, \ldots, x_n)$.

Proposition 3 (*Moments of the Regularized LIML*). *Assume* $\left\{y_i, W_i', x_i'\right\}$ *are iid,* $\varepsilon_i \sim iid\mathcal{N}(0, \sigma_\varepsilon^2)$ *and assume that the vector* u_i *is independent of X, independently normally distributed with mean zero and variance* Σ_u. *Assume that the eigenvalues of K are strictly decreasing. Let* α *be a positive decreasing function of n with* $n\alpha \to \infty$ *as* $n \to \infty$. *Moreover, assume that the regularized LIML estimators based on T, LF, and SC are consistent.*

Then, the rth moments $(r = 1, 2, \ldots)$ *of the regularized LIML estimators are bounded for all n greater than some* $n(r)$.

Proof. See Appendix.

Proposition 3 assumes that the eigenvalues of K are strictly decreasing which rules out the case where all the eigenvalues are equal.[2] In Proposition 2, we assumed that K was nuclear. If the number of instruments is infinite, K nuclear implies that the eigenvalues of K decline to zero fast. However, if the number of instruments is finite, K is a finite dimensional matrix and it is automatically nuclear. To make Proposition 3 hold for both cases with finite and infinite number of moments, we have added the requirement that the eigenvalues strictly decline. The case where the eigenvalues are equal is not covered by our proposition. In this case, the moments of the regularized LIML may not be bounded. This is easy to see for spectral cut-off regularization. Assume that K is the identity matrix and hence the λ_j are all equal to 1. For n large enough, the estimated $\hat{\lambda}_j$ will also be close to 1. For α small, the $q_j = I\left(\hat{\lambda}_j > \alpha\right)$ will be all equal to 1, hence the P^α is the projection matrix on all the instruments and the regularized LIML is nothing but the usual LIML estimator which is known to have no moments. Of course, in practice, with a relatively small sample, the $\hat{\lambda}_j$ may be far from being equal to each other but we may still retain a large number of principal components yielding large moments. This is well illustrated by the simulations of Model 1 in Section 5.

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, $f = (f(x_1), f(x_2), \ldots, f(x_n))'$. Let $\bar{H}$ be the $p \times p$ matrix $\bar{H} = f'f/n$ and $X = (x_1, \ldots, x_n)$.

[2] Recall that the eigenvalues are ranked in decreasing order by assumption.

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

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

where f_a is the ath element of f for $a = 1, 2, \ldots, p$.

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 nuclear 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 SC 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 approximate 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 4. *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 $v_i = u_i - \varepsilon_i\frac{\sigma_{u\varepsilon}}{\sigma_\varepsilon^2}$. If Assumptions 1–3 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*

$$n(\hat\delta - \delta_0)(\hat\delta - \delta_0)' = \hat Q(\alpha) + \hat r(\alpha),$$
$$E(\hat Q(\alpha)|X) = \sigma_\varepsilon^2\bar H^{-1} + S(\alpha) + T(\alpha),$$
$$[\hat r(\alpha) + T(\alpha)]/tr(S(\alpha)) = o_p(1),$$
$$S(\alpha) = \sigma_\varepsilon^2\bar H^{-1}\left[\Sigma_v\frac{[tr((P^\alpha)^2)]}{n} + \frac{f'(1-P^\alpha)^2 f}{n}\right]\bar H^{-1}.$$

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 term, $S(\alpha)$, 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 finite linear combination of the instruments, β would be infinite, and the performance of T is expected to be worse than that of SC 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:

$$S_{LIML}(\alpha) = \sigma_\varepsilon^2\bar H^{-1}\left[\Sigma_v\frac{[tr((P^\alpha)^2)]}{n} + \frac{f'(I-P^\alpha)^2 f}{n}\right]\bar H^{-1} \tag{3}$$

for LIML and

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

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

$$S_{LIML}(\alpha) \sim \frac{1}{n\alpha} + \alpha^\beta,$$
$$S_{2SLS}(\alpha) \sim \frac{1}{n\alpha^2} + \alpha^\beta$$

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

$$S_{LIML}(\alpha) \sim \frac{1}{n\alpha} + \alpha^2,$$
$$S_{2SLS}(\alpha) \sim \frac{1}{n\alpha^2} + \alpha^2.$$

The approximate 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 bias 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 SC LIML are better than T, LF and SC 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 approximate MSE

In this section, we show how to select the regularization parameter α. The aim is to find the α that minimizes the conditional approximate 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 $\bar H^{-1}\gamma$:

$$W\bar H^{-1}\gamma = f\bar H^{-1}\gamma + u\bar H^{-1}\gamma \Leftrightarrow W_\gamma = f_\gamma + u_\gamma \tag{4}$$

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

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

Using (3), $S_\gamma(\alpha)$ can be rewritten as

$$\sigma_\varepsilon^2\left[\sigma_{v_\gamma}^2\frac{[tr((P^\alpha)^2)]}{n} + \frac{f_\gamma'(I-P^\alpha)^2 f_\gamma}{n}\right].$$

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 (4).

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

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

As in Carrasco (2012), the results of Li (1986, 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 $\tilde{u} = (I - P^{\tilde{\alpha}})W$, $\hat{u}_\gamma = \tilde{u}\tilde{H}^{-1}\gamma$,

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

We consider the following goodness-of-fit criteria:

Mallows C_p (Mallows, 1973)

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

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)

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

where $\tilde{W}_\gamma = W\tilde{H}^{-1}\gamma$, $\tilde{W}_{\gamma i}$ is the ith element of $\tilde{W}_\gamma$ and $\hat{f}^\alpha_{\gamma_{-i}} = P^\alpha_{-i}\tilde{W}_{\gamma_{-i}}$. The $n \times (n-1)$ matrix P^α_{-i} is such that $P^\alpha_{-i} = T(K^\alpha_{n-i})T^*_{-i}$ are obtained by suppressing the ith observation from the sample. $\tilde{W}_{\gamma_{-i}}$ is the $(n-1) \times 1$ vector constructed by suppressing the ith observation of $\tilde{W}_\gamma$.

Noting that $\sigma^2_{v_\gamma} - \sigma^2_{u_\gamma} = -\sigma^2_{u_\gamma\varepsilon}/\sigma^2_\varepsilon$ where $\sigma_{u_\gamma\varepsilon} = E\left(u_{\gamma i}\varepsilon_i\right)$. The approximate MSE of $\gamma'\hat{\delta}$ is given by:

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

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

Since $\hat{\sigma}^2_\varepsilon$ does not depend on α, the regularization parameter is selected as

$$\hat{\alpha} = \arg\min_{\alpha\in M_n}\left[\hat{R}(\alpha) - \frac{\hat{\sigma}^2_{u_\gamma\varepsilon}}{\hat{\sigma}^2_\varepsilon}\frac{tr((P^\alpha)^2)}{n}\right] \tag{5}$$

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 $1/\alpha \in \{1, 2, \ldots, n\}$ for SC, and M_n is such that $1/\alpha$ is a positive integer no larger than some finite multiple of n.

Remark 1. This selection is cumbersome because it depends on a first step estimator of α, $\tilde{\alpha}$. Moreover, 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 or any other vector denoted μ. Given the choice of μ is arbitrary and for each μ corresponds a γ, we believe the resulting criterion is a valid way for selecting α. In this case, $W_\gamma = W\mu$, $f_\gamma = f\mu$, $u_\gamma = u\mu$ and $\hat{\sigma}^2_{u_\gamma\varepsilon}$ can be estimated by $u'_\gamma\tilde{\varepsilon}/n$. As a result, the criterion (5) can be computed without relying on any first step estimate of α (except when Mallows C_p is used).

4.2. Optimality

In this section, we will restrict ourselves to the case described in Remark 1 where γ is such that $\bar{H}^{-1}\gamma = \mu$ and μ is an arbitrary vector chosen by the econometrician.

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

$$\frac{S_\gamma(\hat{\alpha})}{\inf_{\alpha\in M_n} S_\gamma(\alpha)} \xrightarrow{P} 1 \tag{6}$$

as n and $n\alpha \to \infty$ where $\hat{\alpha}$ is the regularization parameter defined in (5). The result (6) 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.

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

(iv) $\sum_\alpha(n\tilde{R}(\alpha))^{-2} \xrightarrow{P} 0$ 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 5. ***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 (6) 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 (6) 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 (6) for Tikhonov regularization.

Proof. See 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.

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), DN estimators, and LIML estimator with all the instruments and using the many instrument standard error proposed by Hansen et al. (2008) (denoted HHN in the sequel). In all simulations, we set $\pi = 1$ and we consider large samples of size $n = 500$ and use 1000 replications.

Consider

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

for $i = 1, 2, \ldots, n$, $\delta = 0.1$ 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 two models.

Model 1 (Linear model).

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 $R_f^2 = \pi'\pi/(1+\pi'\pi)$.

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 nuclear, hence our method would not apply.

We set $\pi_l = \sqrt{\frac{R_f^2}{1-R_f^2}}$, $l = 1, 2, \ldots, L$ with $R_f^2 = 0.1$. As all the instruments have the same weight, there is no reason to prefer an instrument over another instrument.

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

$$x_i = Mf_i + v_i$$

where $v_i \sim \mathcal{N}(0, \sigma_v^2 I_3)$ with $\sigma_v = 0.3$, 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 or spectral cut-off), Donald and Newey's (2001) LIML (DLIML), and finally the usual LIML with all instruments and HHN standard errors. When L exceeds n, LIML is computed using a Moore Penrose generalized inverse for the inverse of $Z'Z$. For each regularized and DN estimator, the optimal tuning parameter is selected using generalized cross-validation. For all the regularized LIML, DLIML, and standard LIML estimators, the starting values for the minimization needed in the estimation of v (see Eq. (2)) are the 2SLS using all the instruments when $L \leq 50$ or the corresponding regularized 2SLS for $L > 50$. 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:

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

where $\hat{W} = P^{\alpha}W$ for 2SLS and $\hat{W} = (P^{\alpha} - vI_n)W$ for LIML.

Tables 2 and 4 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 Model 1 are summarized in Tables 1 and 2. In Model 1, 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 from $L = 15$ to $L = 50$, the MSE of regularized LIML becomes smaller than those of regularized 2SLS. We observe that the MSE of regularized LIML, DLIML and standard LIML tend to be very large for $L = 400$ and 520. However, the median bias and dispersions of these remain relatively small suggesting that the large values of the MSE are due to a few outliers. The large MSE of the regularized estimators can be explained by the fact that all eigenvalues of K (in the population) are equal to each other and consequently the assumptions of Proposition 3 are not satisfied. For PC, the cross-validation tends to select either very few or a large number of principal components (see Table 2). In that latter case, the PC LIML is close to the standard LIML estimator which is known for not having any moments. It is important to note that the MSE is sensitive to the starting values used for computing v. For some starting values, explosive behaviors will appear more frequently yielding larger MSE. However, the other statistics reported in the table are not very sensitive to the starting values. We see that HHN standard errors for LIML give an excellent coverage for moderately large values of L ($L \leq 50$) but this coverage deteriorates as L grows much larger.

Now, we turn to Model 2 which is a factor model. From Table 3, we see that there is no clear dominance among the regularized LIML as they all perform very well. Standard LIML is also very good. From Table 4, 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. LIML based estimators have smaller bias than 2SLS based methods. Selection methods as DN are recommended when the rank ordering of the strength of the instruments is clear, otherwise regularized methods are preferable. Among the three regularizations, LLIML and TLIML have smaller bias and better coverage than PLIML in absence of factor structure. Overall, TLIML performs the best across the different values of L. It seems to be the most reliable method.

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 US Census consists of 329, 509 men born between 1930 and 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

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

Table 1
Simulation results of Model 1 with $R_f^2 = 0.1$, $n = 500$.

		T2SL	L2LS	P2LS	D2LS	IV	TLIML	LLIML	PLIML	DLIML	LIML
$L = 15$	Med.bias	0.099	0.096	0.112	0.128	−0.006	−0.001	−0.001	0.015	0.011	−0.002
	Med.abs	0.109	0.115	0.141	0.146	0.087	0.103	0.102	0.103	0.101	0.104
	Disp	0.290	0.297	0.372	0.346	0.347	0.390	0.386	0.378	0.380	0.385
	MSE	0.023	0.023	0.059	0.042	0.019	0.024	0.025	0.023	0.023	0.024
	Cov	0.840	0.843	0.837	0.805	0.946	0.953	0.953	0.928	0.929	0.950
$L = 30$	Med.bias	0.172	0.165	0.174	0.219	0.006	0.010	0.011	0.040	0.050	0.010
	Med.abs	0.173	0.165	0.202	0.237	0.091	0.107	0.110	0.110	0.115	0.108
	Disp	0.264	0.277	0.453	0.457	0.355	0.412	0.421	0.409	0.409	0.413
	MSE	0.039	0.038	3.682	907.31	0.020	0.030	0.032	0.031	0.032	0.029
	Cov	0.594	0.643	0.725	0.673	0.952	0.955	0.950	0.892	0.899	0.951
$L = 50$	Med.bias	0.237	0.226	0.214	0.257	−0.004	−0.004	0.000	0.079	0.105	0.001
	Med.abs	0.237	0.226	0.252	0.285	0.089	0.124	0.126	0.136	0.152	0.123
	Disp	0.235	0.259	0.581	0.590	0.353	0.470	0.489	0.477	0.515	0.492
	MSE	0.061	0.058	1.794	4.946	0.020	0.039	0.045	0.050	0.427	0.040
	Cov	0.300	0.406	0.688	0.639	0.951	0.960	0.955	0.866	0.849	0.957
$L = 400$	Med.bias	0.411	0.380	0.314	0.373	0.006	0.029	0.018	0.270	0.367	0.212
	Med.abs	0.411	0.380	0.449	0.594	0.092	0.249	0.264	0.347	0.450	0.428
	Disp	0.128	0.177	2.291	3.116	0.342	1.116	1.237	1.072	1.386	2.177
	MSE	0.171	0.150	763.56	224.83	0.021	1.4e+22	2e+24	3.460	20.247	3.2e+23
	Cov	0.000	0.001	0.752	0.795	0.961	0.927	0.948	0.823	0.817	0.898
$L = 520$	Med.bias	0.426	0.415	0.360	0.449	−0.007	0.093	0.084	0.346	0.441	0.464
	Med.abs	0.426	0.415	0.468	0.608	0.098	0.294	0.281	0.395	0.512	0.888
	Disp	0.114	0.128	2.192	2.951	0.365	1.307	1.216	1.181	1.459	6.106
	MSE	0.184	0.175	42.561	639.34	0.021	2.6e+29	6.8e+28	2.8e+29	Inf	Inf
	Cov	0.000	0.000	0.702	0.740	0.961	0.912	0.914	0.822	0.790	0.059

NB: We report Median Bias (Med.Bias), Median Absolute deviation (Med.abs), the difference between the 0.1 and 0.9 quantiles (Disp) of the distribution of each estimator, the mean square error (MSE) and the coverage rate (Cov) of a nominal 95% confidence interval. We report results for regularized 2SLS: T2SLS (Tikhonov), L2SLS (Landweber–Fridman), P2SLS (Principal component), the unfeasible instrumental variable regression (IV), regularized LIML: TLIML (Tikhonov), LLIML (Landweber–Fridman), PLIML (Principal component), Donald and Newey's (2001) LIML (DLIML), and finally the LIML with HHN standard errors.

Table 2
Properties of the distribution of the regularization parameters Model 1.

		T2SL	L2LS	P2LS	D2LS	TLIML	LLIML	PLIML	DLIML
$L = 15$	Mean	0.437	18.118	8.909	10.021	0.233	32.909	13.053	14.223
	sd	0.115	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	0.060	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	0.044	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
$L = 400$	Mean	0.500	8.581	9.412	6.580	0.500	5.091	15.633	13.063
	sd	0.000	10.174	20.114	15.373	0.000	3.071	26.556	25.520
	q1	0.500	1.000	1.000	1.000	0.500	3.000	1.000	1.000
	q2	0.500	4.000	2.000	1.000	0.500	5.000	4.000	3.000
	q3	0.500	13.000	7.000	4.000	0.500	7.000	14.000	10.000
$L = 520$	Mean	0.326	156.376	38.712	23.297	0.326	156.270	37.160	30.903
	sd	0.197	106.647	107.346	92.740	0.197	106.594	107.341	99.198
	q1	0.110	63.000	1.000	1.000	0.110	62.500	1.000	1.000
	q2	0.430	127.500	2.000	1.000	0.430	127.500	3.500	3.000
	q3	0.500	300.000	10.000	5.000	0.500	300.000	18.000	10.000

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 et al., 2010, and Hansen et al., 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):

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

where $\log w = log$ of weekly wage, $education =$ year of education, $Y =$ year of birth dummy (9), $S =$ state of birth dummy (50). The vector of instruments $Z = (1, Y, S, Q, Q * Y, Q * S)$ includes 240 variables.

Table 5 reports schooling coefficients generated by different estimators applied to the Angrist and Krueger data along with

Table 3
Simulations results of Model 2, $n = 500$.

		T2SL	L2LS	P2LS	D2LS	IV	TLIML	LLIML	PLIML	DLIML	LIML
$L = 15$	Med.bias	0.000	0.000	0.000	0.004	0.001	−0.000	−0.000	0.000	0.000	−0.000
	Med.abs	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.017	0.018
	Disp	0.068	0.067	0.068	0.066	0.067	0.068	0.068	0.067	0.068	0.068
	MSE	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Cov	0.947	0.947	0.947	0.942	0.952	0.948	0.949	0.948	0.949	0.949
$L = 30$	Med.bias	0.002	0.002	0.002	0.005	0.001	0.001	0.001	0.001	0.002	0.001
	Med.abs	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.018
	Disp	0.067	0.067	0.067	0.068	0.067	0.068	0.068	0.068	0.069	0.069
	MSE	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Cov	0.956	0.955	0.955	0.949	0.958	0.956	0.955	0.955	0.956	0.954
$L = 50$	Med.bias	0.000	−0.000	0.000	0.004	0.001	−0.001	−0.001	−0.001	0.001	−0.000
	Med.abs	0.017	0.017	0.017	0.018	0.017	0.017	0.017	0.017	0.017	0.018
	Disp	0.066	0.066	0.066	0.066	0.065	0.065	0.065	0.065	0.065	0.066
	MSE	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Cov	0.947	0.947	0.948	0.949	0.950	0.948	0.949	0.950	0.951	0.945

NB: We report Median Bias (Med.Bias), Median Absolute deviation (Med.abs), the difference between the 0.1 and 0.9 quantiles (Disp) of the distribution of each estimator, the mean square error (MSE) and the coverage rate (Cov) of a nominal 95% confidence interval. We report results for regularized 2SLS: T2SLS (Tikhonov), L2SLS (Landweber–Fridman), P2SLS (Principal component), the unfeasible instrumental variable regression (IV), regularized LIML: TLIML (Tikhonov), LLIML (Landweber–Fridman), PLIML (Principal component), Donald and Newey's (2001) LIML (DLIML) and finally the LIML with HHN standard errors.

Table 4
Properties of the distribution of the regularization parameters Model 2.

		T2SL	L2LS	P2LS	D2LS	TLIML	LLIML	PLIML	DLIML
$L = 15$	Mean	0.303	270.954	3.012	9.440	0.142	287.230	3.012	13.135
	sd	0.079	31.983	0.109	1.531	0.103	24.322	0.109	1.708
	q1	0.270	248.000	3.000	9.000	0.030	284.500	3.000	12.000
	q2	0.320	280.500	3.000	9.000	0.160	300.000	3.000	14.000
	q3	0.360	300.000	3.000	10.000	0.230	300.000	3.000	14.000
$L = 30$	Mean	0.495	175.479	3.009	11.107	0.294	225.436	3.009	22.370
	sd	0.037	35.078	0.130	2.261	0.217	62.243	0.130	5.839
	q1	0.500	152.000	3.000	10.000	0.040	172.500	3.000	17.000
	q2	0.500	173.000	3.000	11.000	0.425	209.000	3.000	21.000
	q3	0.500	195.000	3.000	11.000	0.500	300.000	3.000	28.000
$L = 50$	Mean	0.499	104.679	2.956	9.331	0.321	175.049	2.956	21.934
	sd	0.011	24.185	0.276	2.790	0.214	93.755	0.276	9.062
	q1	0.500	89.000	3.000	7.000	0.060	99.000	3.000	14.000
	q2	0.500	102.000	3.000	10.000	0.500	125.000	3.000	20.000
	q3	0.500	117.000	3.000	11.000	0.500	300.000	3.000	28.000

Table 5
Estimates of the returns to education.

OLS	**2SLS**	**T2SLS**	**L2SLS**	**P2SLS**
0.0683 (0.0003)	0.0816 (0.0106)	0.1237 (0.0482)	0.1295 (0.0309)	0.1000 (0.0411)
		$\alpha = 0.00001$	Nb of iterations 700	Nb of eigenfunctions 81
	LIML	**TLIML**	**LLIML**	**PLIML**
	0.0918 (0.021)	0.1237 (0.0480)	0.1350 (0.0312)	0.107 (0.0184)
		$\alpha = 0.00001$	Nb of iterations 700	Nb of eigenfunctions 239

NB: Standard errors are in parentheses. For LIML, HHN standard errors are given in parentheses. The concentration parameter is equal to 208.61.

their standard errors[4] in parentheses. Table 5 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 (in the case of Landweber–Fridman regularization). 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.

[4] Our standard errors are not robust to heteroskedasticity.

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

Table 6
Concentration parameter μ_n^2 for the reduced form equation.

	$L = 4$	$L = 18$
$1/\psi$	9.66	33.54
ψ	11.05	68.77

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.

Yogo (2004) uses four instruments: the twice lagged, nominal interest rate (r), inflation (i), consumption growth (c) and log dividend–price ratio (p). This set of instruments is denoted by $Z = [r, i, c, p]$. Yogo (2004) argues that the source for the empirical puzzle mentioned earlier is weak instruments. To strengthen the instruments, we increase the number of instruments from 4 to 18 by including interactions and power functions. The 18 instruments used in our regression are derived from Z and are given by[5] $II = [Z, Z.^2, Z.^3, Z(:,1) * Z(:,2), Z(:,1) * Z(:,3), Z(:,1) * Z(:,4), Z(:,2) * Z(:,3), Z(:,2) * Z(:,4), Z(:,3) * Z(:,4)]$. As a result, the concentration parameters increase in the following way: (See Table 6).

According to Hansen et al. (2008), p. 403, the concentration parameter is a better indication of the potential weak instrument problem than the F-statistic. They argue on p. 404 that "the use of LIML or FULL with the CSE and the asymptotically normal approximation should be adequate in situations where the concentration parameter is around 32 or greater". Since the increase of the number of instruments improves efficiency and regularized 2SLS and LIML correct for the bias due to the many instruments problem, we expect to obtain reliable point estimates. Estimation results are reported in Table 7. Interestingly, the point estimates obtained by T and LF regularized estimators are very close to each other and are close to those used for macro calibrations (EIS equal to 0.71 in our estimations and 0.67 in Castro et al. (2009)). Moreover, the results of the two equations are consistent with each other since we obtain the same value for ψ in both equations.[6] However, we cannot reject the null hypothesis $H_0 : \psi = 1$. PC seems to take too many factors, and did not perform well, this is possibly due to the absence of factor structure.

7. Conclusion

In this paper, we propose a new estimator which is a regularized version of LIML estimator. We 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 regularized LIML estimators have finite moments. 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.

In this paper, we restrict our attention to many strong instruments. In a companion paper, Carrasco and Tchuente (forthcoming) investigate the properties of regularized 2SLS and LIML estimators in the case of many weak instruments as in Chao and Swanson (2005) and Hansen et al. (2008).

Acknowledgment

Carrasco gratefully acknowledges financial support from SSHRC (410-2010-1219).

Appendix. Proofs

Proof of Proposition 1. To prove this proposition, we first need the following lemmas.

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, $\tau_0 = \text{argmin}_{\tau_1=1} \frac{\tau' A \tau}{\tau' B \tau}$ exists and is unique (with $\tau = (\tau_1, \tau_2')'$ and $\tau_1 \in \mathbb{R}$) then*

$$\hat{\tau} = \underset{\tau_1=1}{\text{argmin}} \frac{\tau' \hat{A} \tau}{\tau' \hat{B} \tau} \to \tau_0.$$

Lemma 2. *Under the assumptions of Proposition 1, we have*

$$\varepsilon' P^\alpha \varepsilon = O_p(1/\alpha).$$

Proof of Lemma 2. Let Ω be the $n \times n$ diagonal matrix with ith diagonal element σ_i^2 and $\lambda_{\max}(\Omega)$ be the largest eigenvalue of Ω (which is equal to the largest σ_i^2)

$$\begin{aligned} E\left(\varepsilon' P^\alpha \varepsilon | X\right) &= tr\left(P^\alpha E\left(\varepsilon\varepsilon' | X\right)\right) \\ &= tr\left(P^\alpha \Omega\right) \\ &\le \lambda_{\max}(\Omega)\, tr\left(P^\alpha\right) \\ &\le C \sum_j q_j. \end{aligned}$$

Hence by Markov's inequality, $\varepsilon' P^\alpha \varepsilon = O_p\left(\sum_j q_j\right) = O_p(1/\alpha)$. This completes the proof of Lemma 2.

P^α is a symmetric idempotent matrix for SC but not idempotent for T and LF.

We want to show that $\hat{\delta} \to \delta$ as n and $n\alpha^{\frac{1}{2}}$ go to infinity.

We know that

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

where $\hat{A} = \bar{W}' P^\alpha \bar{W}/n$, $\hat{B} = \frac{\bar{W}'\bar{W}}{n}$ 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

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

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

[6] Note that LIML is invariant to reparametrization whereas 2SLS is not.

Table 7
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)	(0.423)	(0.423)	(0.084)
			$\alpha = 0.01$	Nb of iterations 1000	Nb of PC 11
$1/\psi$	0.6833	0.8241	1.406	1.407	0.7890
	(0.4825)	(0.263)	(0.839)	(0.839)	(0.357)
			$\alpha = 0.01$	Nb of iterations 1000	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)	(0.156)	(0.424)	(0.423)	(0.111)
			$\alpha = 0.01$	Nb of iterations 1000	Nb of PC 8
$1/\psi$	34.1128	4.4952	1.407	1.4072	3.8478
	(112.7122)	(4.421)	(0.839)	(0.839)	(3.138)
			$\alpha = 0.01$	Nb of iterations 1000	Nb of PC 17

NB: For LIML with 18 instruments, HHN standard errors are given in parentheses. For the regularized estimators, we provide the heteroskedasticity robust standard errors in parentheses.

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

$$\begin{aligned}\frac{D_0'W'P^\alpha WD_0}{n} &= D_0'\left\langle (K_n^\alpha)^{-\frac{1}{2}}g_n, (K_n^\alpha)^{-\frac{1}{2}}g_n'\right\rangle D_0\\ &= D_0'\left\langle Fg, Fg'\right\rangle D_0 + o_p(1)\\ &\xrightarrow{P} D_0'\left\langle Fg, Fg'\right\rangle D_0\end{aligned}$$

as n and $n\alpha^{\frac{1}{2}}$ go to infinity and $\alpha\to 0$, see the proof of Proposition 1 of Carrasco (2012).

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

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

$$\frac{e'\varepsilon'P^\alpha WD_0}{n} = e'\left\langle (K_n^\alpha)^{-\frac{1}{2}}\frac{1}{n}\sum_{i=1}^n Z(.;x_i)\varepsilon_i, (K_n^\alpha)^{-\frac{1}{2}}g_n'\right\rangle D_0 = o_p(1),$$

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

We can then conclude that $\hat A\xrightarrow{P} A = D_0'\left\langle Fg, Fg'\right\rangle D_0$ as n and $n\alpha^{\frac{1}{2}}$ go to infinity and $\alpha\to 0$ and

$$\hat B\xrightarrow{P} B = E(\bar W_i\bar W_i')$$

by the law of large numbers with $\bar W_i = [y_i\ W_i']'$.

The LIML estimator is given by

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

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

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

$$\begin{aligned}\tau'A\tau &= \tau'D_0'\left\langle Fg, Fg'\right\rangle D_0\tau\\ &= (\delta_0-\delta)\left\langle Fg, Fg'\right\rangle(\delta_0-\delta)'.\end{aligned}$$

Because $\left\langle Fg, Fg'\right\rangle$ 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

$$\begin{aligned}\tau'B\tau &= E[(\tau_1y_i + W_i'\tau_2)^2]\\ &= E[(\tau_1\varepsilon_i + (f_i+u_i)'(\tau_1\delta_0+\tau_2))^2]\\ &= E[(\tau_1\varepsilon_i + u_i'(\tau_1\delta_0+\tau_2))^2] + (\tau_1\delta_0+\tau_2)'H(\tau_1\delta_0+\tau_2).\end{aligned}$$

Then by $H = E\left(f_if_i'\right)$ 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 $\frac{\tau'A\tau}{\tau'B\tau}$.

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

The gradient and Hessian are given by

$$\begin{aligned}&\Lambda_\delta(\delta) = B(\delta)^{-1}[A_\delta(\delta) - \Lambda(\delta)B_\delta(\delta)],\\ &\Lambda_{\delta\delta}(\delta) = B(\delta)^{-1}[A_{\delta\delta}(\delta) - \Lambda(\delta)B_{\delta\delta}(\delta)]\\ &\qquad - B(\delta)^{-1}[B_\delta(\delta)\Lambda_\delta'(\delta) - \Lambda_\delta(\delta)B_\delta'(\delta)].\end{aligned}$$

Then by a standard mean-value expansion of the first-order conditions $\Lambda_\delta(\hat\delta)=0$, we have

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

where $\tilde\delta$ is the mean-value. Because $\hat\delta$ is consistent, $\tilde\delta\xrightarrow{P}\delta_0$.

It then follows 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 $B_{\delta\delta}(\tilde\delta) = 2W'W/n\xrightarrow{P}2E(W_iW_i'), A_{\delta\delta}(\tilde\delta) = 2W'P^\alpha W/n\xrightarrow{P}2\left\langle Fg, Fg'\right\rangle$.

So that $\tilde\sigma^2\Lambda_{\delta\delta}(\tilde\delta)/2\xrightarrow{P}\left\langle Fg, Fg'\right\rangle$ with $\tilde\sigma^2 = \varepsilon'\varepsilon/n$.

By Lemma 2, we have $\varepsilon'P^\alpha\varepsilon/\sqrt n = O_p(1/(\alpha\sqrt n)) = o_p(1)$.

$$\begin{aligned}-\sqrt n\tilde\sigma^2\Lambda_\delta(\delta_0)/2 &= \frac{W'P^\alpha\varepsilon}{\sqrt n} - \frac{\varepsilon'P^\alpha\varepsilon}{\sqrt n}\frac{W'\varepsilon}{\varepsilon'\varepsilon}\\ &= \frac{W'P^\alpha\varepsilon}{\sqrt n} + o_p(1)\xrightarrow{d}\mathcal N\left(0, \left\langle Fg, \left(F\widetilde KF^*\right)Fg'\right\rangle\right).\end{aligned}$$

To obtain the asymptotic normality, note that

$$\begin{aligned}\frac{W'P^\alpha\varepsilon}{\sqrt n} &= \left\langle \left(K_n^\alpha\right)^{-1}g_n, \frac{\sum_{i=1}^n Z_i(.,x_i)\varepsilon_i}{\sqrt n}\right\rangle\\ &= \left\langle K^{-1}g, \frac{\sum_{i=1}^n Z_i(.,x_i)\varepsilon_i}{\sqrt n}\right\rangle\\ &\quad + \left\langle \left(K_n^\alpha\right)^{-1}g_n - K^{-1}g, \frac{\sum_{i=1}^n Z_i(.,x_i)\varepsilon_i}{\sqrt n}\right\rangle. \qquad (7)\end{aligned}$$

Moreover, $\{Z_i(.,x_i)\varepsilon_i\}$ is iid with $E\|Z_i(.,x_i)\varepsilon_i\|^2 < \infty$ (because $E(\varepsilon_i^2|x_i)$ is bounded and K is nuclear). It follows from vander Vaart and Wellner (1996), p. 50 that $\sum_{i=1}^n Z(.;x_i)\varepsilon_i/\sqrt{n}$ converges in $L^2(\pi)$ to a mean zero Gaussian process with covariance operator $\tilde{K}$. Hence,

$$\left\langle K^{-1}g, \frac{\sum_{i=1}^n Z_i(.,x_i)\varepsilon_i}{\sqrt{n}}\right\rangle \xrightarrow{d} N\left(0, \left\langle K^{-1}g, \tilde{K}K^{-1}g\right\rangle\right).$$

As g belongs to the range of K, Lemma 3 of Carrasco (2012) implies that $\left\|\left(K_n^\alpha\right)^{-1}g_n - K^{-1}g\right\| \xrightarrow{P} 0$ and hence the second term of the r.h.s. of (7) is $o_p(1)$. This concludes the proof of Proposition 1.

Proof of Proposition 2.

Lemma 3. *Let $v = u - \varepsilon\phi'$. Under the assumptions of Proposition 2, we have*

$$v'P^\alpha\varepsilon = O_p\left(\frac{1}{\sqrt{\alpha}}\right).$$

Proof of Lemma 3. Using the spectral decomposition of P^α, we have $v'P^\alpha\varepsilon = \frac{1}{n}\sum_j q_j\left(v'\psi_j\right)\left(\varepsilon'\psi_j\right)$

$$\begin{aligned}(v'P^\alpha\varepsilon)^2 &= \frac{1}{n^2}\sum_{j,l} q_jq_l(v'\psi_j)(\varepsilon'\psi_j)(v'\psi_l)(\varepsilon'\psi_l)\\ &= \frac{1}{n^2}\sum_{j,l} q_jq_l\left(\sum_i v_i\psi_{ji}\right)\left(\sum_b v_b\psi_{lb}\right)\\ &\quad\times\left(\sum_c \varepsilon_c\psi_{jc}\right)\left(\sum_d \varepsilon_d\psi_{ld}\right).\end{aligned}$$

Using the fact that $E(\varepsilon_i) = E(v_i) = E(\varepsilon_iv_i) = 0$ and that the eigenvectors are orthonormal, i.e. $\sum_i \psi_{li}\psi_{ji}/n = 1$ if $l = j$ and 0 otherwise, we have

$$\begin{aligned}E\left[(v'P^\alpha\varepsilon)^2\right] &= \frac{1}{n^2}\sum_{j,l} q_jq_l\sum_i E\left(v_i^2\varepsilon_i^2\right)\psi_{ji}^2\psi_{li}^2\\ &\quad+\sum_j q_j^2E\left(v_i^2\right)E\left(\varepsilon_i^2\right)\left(\frac{\sum_i\psi_{ji}^2}{n}\right)^2. \qquad (8)\end{aligned}$$

As ψ_{li}^2 is summable, it is bounded, hence $\sum_i E\left(v_i^2\varepsilon_i^2\right)\psi_{ji}^2\psi_{li}^2/n < C$ and the first term on the r.h.s. of (8) is negligible with respect to the second. By Markov's inequality,

$$v'P^\alpha\varepsilon = O_p\left(\left(\sum_j q_j^2\right)^{1/2}\right) = O_p\left(1/\sqrt{\alpha}\right).$$

This completes the proof of Lemma 3.

The proof of the consistency is the same as that of Proposition 1.

Now $\langle Fg, Fg'\rangle = H = E\left(f_if_i'\right)$ 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 et al. (2007). For the asymptotic normality, the beginning of the proof is the same. Let $\hat{\phi} = \frac{W'\varepsilon}{\varepsilon'\varepsilon}$, $\phi = \frac{\sigma_{u\varepsilon}}{\sigma_\varepsilon^2}$ and $v = u - \varepsilon\phi'$. We have $v'P^\alpha\varepsilon/\sqrt{n} = O_p(1/\sqrt{n\alpha}) = o_p(1)$ by Lemma 3. Moreover, $\hat{\phi} - \phi = O_p(1/\sqrt{n})$ by the Central limit theorem and delta method so that $(\hat{\phi} - \phi)\varepsilon'P^\alpha\varepsilon/\sqrt{n} = O_p(1/n\alpha) = o_p(1)$ by Lemma 2.

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

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

The conclusion follows from Slutsky's theorem. Note that because $v'P^\alpha\varepsilon/\sqrt{n} = O_p(1/\sqrt{n\alpha})$, we get a faster rate for α in the homoskedastic case than in the heteroskedastic case. The proof in the heteroskedastic case relies on $\varepsilon'P^\alpha\varepsilon/\sqrt{n} = O_p(1/\alpha\sqrt{n})$.

Proof of Proposition 3. We want to prove that the regularized LIML estimators have finite moments. These estimators are defined as follows:[7]

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

where $\nu_\alpha = \min_\delta \frac{(y-W\delta)'P^\alpha(y-W\delta)}{(y-W\delta)'(y-W\delta)}$ and $P^\alpha = T(K_n^\alpha)^{-1}T^*$.

The following lemma will be useful in the remaining of the proof.

Lemma 4. *Under the assumptions of Proposition 3, we have*

$$\nu_\alpha = O_p\left(\frac{1}{n\alpha}\right).$$

Proof of Lemma 4.

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

Using $y - W\hat{\delta} = \varepsilon - W\left(\hat{\delta} - \delta_0\right)$ and the consistency of $\hat{\delta}$, we have

$$\frac{(y - W\hat{\delta})'(y - W\hat{\delta})}{n} = \frac{\varepsilon'\varepsilon}{n} + o_p(1) = O_p(1).$$

Moreover, by Lemma 2, $\varepsilon'P^\alpha\varepsilon = O_p(1/\alpha)$. It follows that

$$\begin{aligned}(y - W\hat{\delta})'P^\alpha(y - W\hat{\delta}) &= \varepsilon'P^\alpha\varepsilon + \left(\hat{\delta} - \delta_0\right)'W'P^\alpha W\left(\hat{\delta} - \delta_0\right)\\ &\quad+2\left(\hat{\delta} - \delta_0\right)'W'P^\alpha\varepsilon\\ &= \varepsilon'P^\alpha\varepsilon + O_p\left(\frac{1}{n}\right)\\ &= O_p(1/\alpha)\end{aligned}$$

where the second equality follows from the proof of Proposition 1 in Carrasco (2012). The result of Lemma 4 follows.

Let us define $\hat{H} = W'(P^\alpha - \nu_\alpha I_n)W$ and $\hat{N} = W'(P^\alpha - \nu_\alpha I_n)y$ thus

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

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

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

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

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 ith 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 jth element of $\hat{N}$ and $|.|$ denotes the determinant.

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

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.[8]

We want to prove that $|\hat{H}| \geq |S|$ where S is a positive definite $p \times p$ matrix to be specified later on. The first step consists in showing that $P^\alpha - \nu_{\frac{\alpha}{2}} I_n$ is positive definite. Let us consider $x \in \mathbb{R}^n$. We have

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

For a given α, q_j is a decreasing function of j because λ_j is decreasing in j. Hence, there exists j^*_α such that $q_j \geq \nu_{\frac{\alpha}{2}}$ for $j \leq j^*_\alpha$ and $q_j \langle \nu_{\frac{\alpha}{2}}$ for $j \rangle j^*_\alpha$ and

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

The term (1) is positive and the term (2) is negative. As n increases, α decreases and q_j increases for any given j. On the other hand, when n increases and $n\alpha \to \infty$, $\nu_{\frac{\alpha}{2}}$ decreases by Lemma 3. It follows that j^*_α increases when n goes to infinity.

Consequently, 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, j^*_α is sufficiently large for (2) to be smaller in absolute value than (1) and hence $x'\left(P^\alpha - \nu_{\frac{\alpha}{2}} I_n\right) x > 0$.

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

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

Hence,

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

For n large but finite, $\nu_{\frac{\alpha}{2}} - \nu_\alpha > 0$ 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 4. 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 5. *Let* $\tilde{\Lambda} = \varepsilon' P \varepsilon / (n\sigma^2_\varepsilon)$ *and* $\hat{\Lambda} = \Lambda(\hat{\delta})$ *with* $\Lambda(\delta) = \frac{(y - W\delta)' P (y - W\delta)}{(y - W\delta)'(y - W\delta)}$. *If the assumptions of Proposition 4 are satisfied, then*

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

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

Proof of Lemma 5. 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 $\hat{\delta}$, $\Lambda_{\delta\delta}(\tilde{\delta}) = \Lambda_{\delta\delta}(\delta_0) + O_p(1/\sqrt{n})$. It implies that

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

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

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

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

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

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

By combining these two equalities, we obtain

$$\begin{aligned} &\Lambda_\delta(\delta_0)' [\Lambda_{\delta\delta}(\delta_0)]^{-1} \Lambda_\delta(\delta_0) \\ &= h' \bar{H}^{-1} h / (n\sigma^2_\varepsilon) + O_p(\Delta_\alpha^{1/2}/n + \sqrt{1/(n^3\alpha)}). \end{aligned}$$

[8] 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).

Note also that

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

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

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

$$\sqrt{n}\hat R_\Lambda = o_p(\rho_{\alpha,n}).$$

By using $\tilde\Lambda = O_p(1/n\alpha)$, it is easy to prove that $\hat\Lambda = \tilde\Lambda + o_p(1/n\alpha)$.

Lemma 6. *If the assumptions of Proposition 4 are satisfied, then*

(i) $u'Pu - \tilde\Lambda\Sigma_u = o_p(1/n\alpha)$,
(ii) $E(h\tilde\Lambda\varepsilon'v/\sqrt{n}|X) = (tr(P)/n)\sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n + O(1/(n^2\alpha))$,
(iii) $E(hh'\bar H^{-1}h/\sqrt{n}|X) = O(1/n)$.

Proof of Lemma 6. 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

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

Thus, $(\tilde\Lambda - tr(P)/n)\Sigma_u = o_p(tr(P)/n) = o_p(1/n\alpha)$ by Markov's inequality. $u'Pu - \frac{tr(P)}{n}\Sigma_u = o_p(1/n\alpha)$ such that $u'Pu - \tilde\Lambda\Sigma_u = o_p(1/(n\alpha))$.

To show (ii) we can notice that

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

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 4.

Proof of Proposition 4. 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.[9]
$\hat h = h + \sum_{j=1}^5 T_j^h + Z^h$ with $h = f'\varepsilon/\sqrt{n}$,
$T_1^h = -f'(I-P)\varepsilon/\sqrt{n} = O_p(\Delta_\alpha^{1/2})$,
$T_2^h = v'P\varepsilon/\sqrt{n} = O_p(\sqrt{1/n\alpha})$, $T_3^h = -\tilde\Lambda h' = O(1/n\alpha)$, $T_4^h = -\tilde\Lambda v'\varepsilon/\sqrt{n} = O_p(1/n\alpha)$,
$T_5^h = h'\bar H^{-1}h\sigma_{u\varepsilon}/2\sqrt{n}\sigma_\varepsilon^2 = O_p(1/\sqrt{n})$,
$Z^h = -\hat R_\Lambda W'\varepsilon/\sqrt{n} - (\hat\Lambda - \tilde\Lambda - \hat R_\Lambda)\sqrt{n}(W'\varepsilon/n - \sigma_{u\varepsilon}')$ where $\hat R_\Lambda$ is defined in Lemma 4.

By using the central limit theorem on $\sqrt{n}(W'\varepsilon/n - \sigma_{u\varepsilon}')$ and Lemma 4, $Z^h = O(\rho_{n\alpha})$. The results on the order of T_j^h hold by Lemma 5 of Carrasco (2012).

We also have

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

$$\begin{aligned}T_1^H &= -f'(I-P)f/n = O_p(\Delta_\alpha), \\ T_2^H &= (u'f + f'u)/n = O_p(1/\sqrt{n}), \\ T_3^H &= -\tilde\Lambda\bar H = O_p(1/n\alpha), \\ Z^H &= u'Pu/n - \tilde\Lambda\Sigma_u - \hat\Lambda W'W/n + \tilde\Lambda(\bar H + \Sigma_u) \\ &\quad - u'(I-P)f/n - f'(I-P)u/n.\end{aligned}$$

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

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

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

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

$$\begin{aligned}Z^A &= \left(\sum_{j=3}^5 T_j^h\right)\left(\sum_{j=3}^5 T_j^h\right)' + \left(\sum_{j=3}^5 T_j^h\right)(T_1^h + T_2^h)' \\ &\quad + (T_1^h + T_1^h)\left(\sum_{j=3}^5 T_j^h\right)',\end{aligned}$$

[9] The expression of T_5^h, Z^h and Z^H below corrects some sign errors in DN.

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 $e_f(\alpha) = \frac{f'(I-P)f}{n}$ and $e_{2f}(\alpha) = \frac{f'(I-P)^2 f}{n}$. By Lemma 3(ii) $E(hT_4^{h'}|X) = \frac{tr(P)}{n}\sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n + O\left(\frac{1}{n^2\alpha}\right)$.

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 5(iii), $E(hT_5^{h'}) = O_p(1/n)$.

For $\hat{\xi} = \sum_i P_{ii} f_i E(\varepsilon_i^2 v_i'|x_i)/n - \frac{tr(P)}{n}\sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n - \sum_i P_{ii}(1-P_{ii}) f_i E(\varepsilon_i^2 v_i'|x_i)/n$, $\hat{A}(\alpha)$ 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 $\|T_1^h\|\|T_j^h\| = o_p(\rho_{n\alpha})$, $\|T_2^h\|\|T_j^H\| = o_p(\rho_{n\alpha})$ for each j and $\|T_k^h\|\|T_j^H\| = o_p(\rho_{n\alpha})$ for each j and $k > 2$. Furthermore $\|T_j^H\|^2 = o_p(\rho_{n\alpha})$ for each j. It follows that $Z^A = o_p(\rho_{n\alpha})$. Therefore, 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 4.

To prove Proposition 5, we need to establish the following result.

Lemma 7 (*Lemma A.9 of DN*). *If* $\sup_{\alpha\in M_n}(|\hat{S}_\gamma(\alpha) - S_\gamma(\alpha)|/S_\gamma(\alpha)) \xrightarrow{P} 0$, *then* $S_\gamma(\hat{\alpha})/\inf_{\alpha\in M_n} S_\gamma(\alpha) \xrightarrow{P} 1$ *as* n *and* $n\alpha \to \infty$.

Proof of Lemma 7. 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 SC and LF and by the compactness of the index set for T. Then, the proof of Lemma 7 follows from that of Lemma A.9 of DN.

Proof of Proposition 5. We proceed by verifying the assumption of Lemma 7.

Let $R(\alpha) = \frac{f_\gamma'(I-P)^2 f_\gamma}{n} + \sigma_{u_\gamma}^2 \frac{tr(P^2)}{n}$ be the risk approximated by $\hat{R}^m(\alpha)$, $\hat{R}^{cv}(\alpha)$, or $\hat{R}^{lcv}(\alpha)$, and $S_\gamma(\alpha) = \sigma_\varepsilon^2\left[\frac{f_\gamma'(I-P)^2 f_\gamma}{n} + \sigma_{v_\gamma}^2 \frac{tr(P^2)}{n}\right]$. 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

$$\sup_{\alpha\in M_n}\left(|\hat{R}(\alpha) - R(\alpha)|/R(\alpha)\right) \to 0 \tag{9}$$

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

To establish this result, we need to verify the assumptions of Li's (1986, 1987) theorems. We treat separately the regularizations with a discrete index set and that with a continuous index set.

Discrete index set:

SC and LF have a discrete index set in terms of $1/\alpha$.

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) $E((u_i e)^8) < \infty$;

(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 $nR(\alpha) = f_\gamma'(I-P)^2 f_\gamma + \sigma_{u_\gamma}^2 tr(P^2) = O_p\left(n\alpha^\beta + \frac{1}{\alpha}\right)$.

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) uses 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 Eq. (2.6) of Li (1987) holds and assumption (A.3) can be replaced by (A.3'). The justification for replacing $\sigma_{u_\gamma\varepsilon}^2$, $\sigma_{u_\gamma}^2$ and σ_ε^2 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) $\inf_{\alpha\in M_n} n^{-1}\left\|f_\gamma - PW_\gamma\right\| \to 0$;

(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 $R(\alpha) = En^{-1}\left\|f_\gamma - PW_\gamma\right\| \to 0$ as n and $n\alpha$ go to infinity.

Note that $tr(P) = O\left(\alpha^{-1}\right)$ 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

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

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

In the case of leave-one-out cross-validation criterion, we need to verify the assumptions of Theorem 5.1 of Li (1987). Assumptions (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 (9) 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 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 (9) holds for T under Assumption 4(i').

We have proved that (9) holds for the various regularizations. We proceed to check the condition of Lemma 7. First note that, given $\sigma_\varepsilon^2 \neq 0$, $R(\alpha) \leq CS_\gamma(\alpha)/\sigma_\varepsilon^2$. To see this, replace $R(\alpha)$ and $S_\gamma(\alpha)$ by their expressions in function of $\frac{f_\gamma'(I-P)^2 f_\gamma}{n}$ and use the fact that $\sigma_{u_\gamma}^2 > \sigma_{v_\gamma}^2$ and take $C = \sigma_{u_\gamma}^2/\sigma_{v_\gamma}^2$. Now we have

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

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

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

It follows from (9) 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 7. This ends the proof of Proposition 5.

References

Andersen, T.G., Sorensen, B.E., 1996. GMM estimation of a stochastic volatility Model: A Monte Carlo study. J. Bus. Econom. Statist. 14 (3), 328–352.

Anderson, T., 2010. The LIML estimator has finite moments! J. Econometrics 157 (2), 359–361.

Anderson, T., Kunitomo, N., Matsushita, Y., 2010. On the asymptotic optimality of the LIML estimator with possibly many instruments. J. Econometrics 157 (2), 191–204.

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? Quart. J. Econom. 106 (4), 979–1014.

Bai, J., Ng, S., 2010. Instrumental variable estimation in a sata rich environment. Econometric Theory 26, 1577–1606.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.

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

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012a. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012b. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 (6), 2369–2429.

Campbell, J.Y., Viceira, L.M., 1999. Consumption and portfolio decisions when expected returns are time varying. Quart. J. Econom. 114 (2), 433–495.

Canay, I., 2010. Simultaneous selection and weighting of moments in GMM using a trapezoidal kernel. J. Econometrics 156 (2), 284–303.

Carrasco, M., 2012. A regularization approach to the many instruments problem. J. Econometrics 170 (2), 383–398.

Carrasco, M., Chernov, M., Florens, J.-P., Ghysels, E., 2007. Efficient estimation of general dynamic models with a continuum of moment conditions. J. Econometrics 140 (2), 529–573.

Carrasco, M., Florens, J.-P., 2014. On the asymptotic efficiency of GMM. Econometric Theory 30 (2), 372–406.

Carrasco, M., Florens, J.-P., 2000. Generalization of Gmm to a continuum of moment conditions. Econometric Theory 16 (06), 797–834.

Carrasco, M., Florens, J.-P., Renault, E., 2007. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6. Elsevier, (Chapter 77).

Carrasco, M., Tchuente, G., 2014. Efficient estimation with many weak instruments using regularization techniques. Econometric Rev. forthcoming.

Castro, R., Clementi, G.L., Macdonald, G., 2009. Legal institutions, sectoral heterogeneity, and economic development. Rev. Econom. Stud. 76 (2), 529–561.

Chao, J.C., Swanson, N.R., 2005. Consistent estimation with a large number of weak instruments. Econometrica 73 (5), 1673–1692.

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

Craven, P., Wahba, G., 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., Dagenais, D.L., 1997. Higher moment estimators for linear regression models with errors in the variables. J. Econometrics 76 (1–2), 193–221.

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

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 Rev. 27 (4–6), 526–541.

Hahn, J., Hausman, J., 2003. Weak instruments: Diagnosis and cures in empirical econometrics. Amer. Econ. Rev. 93 (2), 118–125.

Hahn, J., Inoue, A., 2002. A Monte Carlo comparison of various asymptotic approximations to the sistribution of instrumental variables estimators. Econometric Rev. 21 (3), 309–336.

Hansen, C., Hausman, J., Newey, W., 2008. Estimation With Many Instrumental Variables. J. Bus. Econom. Statist. 26, 398–422.

Hansen, C., Kozbur, D., 2014. Instrumental variables estimation with many weak instruments using regularized JIVE, Working Paper.

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

Hausman, J.A., Newey, W.K., Woutersen, T., Chao, J.C., Swanson, N.R., 2012. Instrumental variable estimation with heteroskedasticity and many instruments. Quantitative Economics 3 (2), 211–255.

James, A., 1954. Normal multivariate analysis and the orthogonal group. Ann. Math. Statist. 25, 46–75.

Kapetanios, G., Marcellino, M., 2010. Factor-GMM estimation with large sets of possibly weak instruments. Comput. Statist. Data Anal. 54, 2655–2675.

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

Kress, R., 1999. Linear Integral Equations. Springer.

Kuersteiner, G., 2012. Kernel-weighted GMM estimators for linear time series models. J. 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. Ann. Statist. 14, 1101–1112.

Li, K.-C., 1987. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. Ann. Statist. 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., Windmeijer, F., 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. J. Econometrics 165, 70–86.

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

Stone, C.J., 1974. Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc 36, 111–147.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer.

Yogo, M., 2004. Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak. Rev. Econom. Statist. 86 (3), 797–810.