Université de Montréal

Essais en econometrie et économie de l'éducation

par
Guy Tchuente Nguembu

Département de sciences économiques
Faculté des arts et des sciences

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

Juillet, 2014

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Essais en econometrie et économie de l'éducation

présentée par:

Guy Tchuente Nguembu

a été évaluée par un jury composé des personnes suivantes:

Benoit Perron,	président-rapporteur
Marine Carrasco,	directeur de recherche
Baris Kaymak,	codirecteur
Joshua Lewis,	membre du jury
Jorgen Hansen,	examinateur externe
Jean-Michel Cousineau ,	représentant du doyen de la FES

Thèse acceptée le: 28 août 2014

RÉSUMÉ

Cette thèse est organisée en trois chapitres. Les deux premiers s'intéressent à l'évaluation, par des méthodes d'estimations, de l'effet causal ou de l'effet d'un traitement, dans un environnement riche en données. Le dernier chapitre se rapporte à l'économie de l'éducation. Plus précisément dans ce chapitre j'évalue l'effet de la spécialisation au secondaire sur le choix de filière à l'université et la performance.

Dans le premier chapitre, j'étudie l'estimation efficace d'un paramètre de dimension finie dans un modèle linéaire où le nombre d'instruments peut être très grand ou infini. L'utilisation d'un grand nombre de conditions de moments améliore l'efficacité asymptotique des estimateurs par variables instrumentales, mais accroit le biais. Je propose une version régularisée de l'estimateur LIML basée sur trois méthodes de régularisations différentes, Tikhonov, Landweber Fridman, et composantes principales, qui réduisent le biais.

Le deuxième chapitre étend les travaux précédents, en permettant la présence d'un grand nombre d'instruments faibles. Le problème des instruments faibles est la consequence d'un très faible paramètre de concentration. Afin d'augmenter la taille du paramètre de concentration, je propose d'augmenter le nombre d'instruments. Je montre par la suite que les estimateurs 2SLS et LIML régularisés sont convergents et asymptotiquement normaux.

Le troisième chapitre de cette thèse analyse l'effet de la spécialisation au secondaire sur le choix de filière à l'université. En utilisant des données américaines, j'évalue la relation entre la performance à l'université et les différents types de cours suivis pendant les études secondaires. Les résultats suggèrent que les étudiants choisissent les filières dans lesquelles ils ont acquis plus de compétences au secondaire. Cependant, on a une relation en U entre la diversification et la performance à l'université, suggérant une tension entre la spécialisation et la diversification. Le compromis sous-jacent est évalué par l'estimation d'un modèle structurel de l'acquisition du capital humain au secondaire et de

choix de filière. Des analyses contrefactuelles impliquent qu'un cours de plus en matière quantitative augmente les inscriptions dans les filières scientifiques et technologiques de 4 points de pourcentage.

Mots clés: Modèles de grande dimension, LIML, Variable instrumentale faibles, Erreur Quadratique moyenne, régularisation, capital humain, choix discrets, Choix de filière.

ABSTRACT

This these is organized in three chapter; the first two chapters are in Econometrics and the third in labor Economics. The econometrics chapters focuses on estimating parameters in data rich environments. I investigate how to establish causal effect or treatment effect in high dimensional setting using regularization techniques. The last chapter of this thesis focuses on the outcomes associated with general and specific education. In particular I study the effect of specialization in high school on college major choice and performance in college.

In the first chapter, entitled “Regularized LIML for many instruments ”(joint with Marine Carrasco), I consider the efficient estimation of a finite dimensional parameter in a linear model where the number of potential instruments is very large or infinite. The use of many moment conditions improves the asymptotic efficiency of the instrumental variables estimators. I propose regularized a versions of the limited information maximum likelihood (LIML) based on three different regularizations: Tikhonov, Landweber Fridman, and principal components.

The second chapter, entitled “Efficient estimation with many weak instruments using regularization techniques ”, (Joint with Marine Carrasco), extends the previous works, to allow for the presence of a large number of weak instruments or weak identification. The problem of weak instruments is due to a very small concentration parameter. To boost the concentration parameter, I propose to increase the number of instruments to a large number or even up to a continuum. I show that normalized regularized 2SLS and LIML are consistent and asymptotically normally distributed.

The third chapter of this thesis is entitled “High school human capital portfolio and college outcomes" investigates the trade-off between acquiring specialized skills, in high school, which will be useful for a particular college major and acquiring a package of skills that diversifies risk across majors. Using the 1980 High School and Beyond (HS&B) survey, I study the empirical relationship between college performance and dif-

ferent types of courses taken during formal high school education. This panel shows that students sort into majors according to the subject in which they acquired more skills. However, I find a U-shaped relation between diversification and college performance, suggesting a trad-off between specialization and diversification. The underlying trade-off is assessed by estimating a structural model of high school human capital acquisition and college major choice. Policies experiments suggest that one more high school quantitative course increases enrollment in Science Technology Engineering, and Math (STEM) majors by 4 points percentage.

Keywords: High-dimensional models, LIML, many weak instruments, MSE, regularization methods, Human capital, Discrete choice, College Major.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

A mon épouse Fabiola, Ma fille Maelie, mes parents Nguembu Jean et Kom Rosalie.

REMERCIEMENTS

L'accomplissement des travaux de cette thèse a bénéficié du concours de plusieurs personnes que j'aimerai remercier. Sans pour autant prétendre à des remerciements exhaustifs, ma gratitude est grande à l'endroit de toutes les personnes qui m'ont aidé pendant l'écriture de ce document. J'aimerai, tout d'abord, remercier ma directrice de recherche Marine Carrasco, pour sa présence continue, sa patience, ses multipes conseils en tout genre et pour m'avoir encouragé dans toutes les étapes de cette thèse. Mes remerciements vont également à mon Co-directeur Baris Kaymak pour sa disponibilité, son dynamisme et ses conseils. Sans leurs précieux soutiens, cette thèse ne serait pas arrivée à ce stade.

Cette thèse a bénéficié de commentaires issus de présentations dans des cadres formels (séminaires, conférences) ou dans des cadres moins formels (échanges avec des collègues ou amis). Je tiens à remercier toutes les personnes dont les remarques ont permis d'améliorer ce travail. Merci aux professeurs Benoît perron , Marc Henry, Silvia Gonçalves et Yves Sprimont qui me suivent depuis la première présentation orale et les premières années de cette thèse. Je remercie également les professeurs Andriana Bellou, Joshua Lewis, Julien Bengui et Raphaël Godefroy pour leurs commentaires et conseils sur le dernier chapitre.

Je ne saurai oublier tous les étudiants de doctorat dont les discussions enrichissantes m'ont permis d'approfondir mes connaissances en science économique et d'améliorer le contenu du présent document. Je remercie également mes amis Herman, Ismael , Pierre-Evariste, Théophile, Thierry, Valery dont les conseils et le soutient m'ont été d'une aide précieuse.

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

économiques de l'université de Montréal et à celui du CIREQ pour leur dynamisme et leur efficacité.

J'adresse un merci tout particulier à mon épouse Fabiola et ma fille Maelie pour leur patience et leur soutien malgré les heures passées au bureau et devant mon ordinateur, votre joie de vivre et votre amour m'inspirent tous les jours. A mes parents, Nguembu Jean et Kom Rosalie, pour l'éducation de base et les valeurs que vous m'avez transmis, je vous dis merci. Merci également à mes frères et soeurs, pour leurs conseils et aussi leur patience en ces périodes où nous n'avons pu nous voir aussi souvent qu'on l'aurait souhaité. Enfin Toute ma gratitude va à Dieu, pour toutes les grâces qu'il ne cesse de m'accorder. A tous ceux qui de près ou de loin ont contribué à cette thèse et dont les noms ne figurent pas explicitement ici, je dis merci.

CHAPTER 1

REGULARIZED LIML FOR MANY INSTRUMENTS

1.1 Introduction

The problem of many instruments is a growing part of the econometric literature. [1] 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 estimated 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 reg-

1. This chapter is a joint work with Marine Carrasco. The authors thank the participants of CIREQ conference on High Dimensional Problems in Econometrics (Montreal, May 2012), of the conference in honor of Jean-Pierre Florens (Toulouse, September 2012), of the seminars at the University of Rochester and the University of Pennsylvania for helpful comments.

ularization techniques borrowed from the statistic literature on linear inverse problems (see Kress (1999) and Carrasco et al. (2007a)). 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 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 (2004), 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 regularization 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, grow 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. Recently, some work has been done in the case where the number of instruments exceed the sample size. Kuersteiner (2012) considers a kernel weighted GMM estimator, Okui (2004) uses shrinkage. Bai and Ng (2010) and Kapetanios and Marcellino (2010) assume that the endogenous regressors depend on a small number of factors which are exogenous, they use estimated factors as instruments. Belloni 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 jacknife 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 select the number of instruments by minimizing an approximate MSE. Our method assumes neither a strong factor structure, nor a 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.

1.2 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.

1.2.1 Presentation of the estimators

The model is

$$\begin{cases} y_i = W_i'\delta_0 + \varepsilon_i \\ W_i = f(x_i) + u_i \end{cases} \tag{1.1}$$

$i = 1, 2, ..., n$. 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 $L \times p$ β. 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 $S = \{1,2,....L\}$.
- $Z(\tau;x_i) = (x_i)^{\tau-1}$ with $\tau \in S = \mathbb{N}$, thus we have an 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. (2007b).

We define the covariance operator K of the instruments as

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

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

where $\overline{Z(\tau_2;x_i)}$ denotes the complex conjugate of $Z(\tau_2;x_i)$. K is assumed to be a 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 details 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 $\left\{Z_{ij} : j = 1,2,...,\infty\right\}$ 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 (2004) 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 1.2.2 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 $S = \{1,2,....L\}$. In that case, K is the operator which associates to vector v of $\mathbb{R}^L$, the vector $Kv = E\left(Z_iZ_i'\right)v/L$. The condition "$K$ nuclear" is met if the trace of $E\left(Z_iZ_i'\right)/L$ is finite. This is satisfied if the Z_{il}, $l = 1,2,...,L$ depend 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 $j = 1,2,...$ be respectively the eigenvalues (ordered in decreasing order) and the orthogonal eigenfunctions of K. The operator K can be estimated by K_n defined as:

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

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

If the number of moment conditions is infinite, 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), Landwerber Fridman (LF) and Spectral cut-off (SC). They are defined as follows [2]:

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:

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

   where $\dfrac{1}{\alpha} - 1$ is some positive integer. Equivalently, we have

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

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

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 $q(\alpha, \lambda_j^2) = \dfrac{\lambda_j^2}{\lambda_j^2 + \alpha}$, 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

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

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

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

estimators as

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

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

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

corresponds to the regularized LIML estimator we will study here.

1.2.2 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

$$\tilde{K} : L^2(\pi) \to L^2(\pi)$$

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

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 van der 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* $\left\langle Fg, Fg' \right\rangle$ *is nonsingular. The regularization parameter* α *goes to zero. Then, the T, LF, and SC estimators of LIML 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*
   $\sqrt{n}(\hat{\delta} - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \left\langle Fg, Fg' \right\rangle^{-1} \left\langle Fg, \left(F\tilde{K}F^*\right) Fg \right\rangle \left\langle Fg, Fg' \right\rangle^{-1}\right)$
   as n and $\alpha\sqrt{n}$ *go to infinity.*

The condition $\left\langle Fg, Fg' \right\rangle$ 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 a^{th} 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,..., p. Then, the T, LF, and SC estimators of LIML satisfy:*

1. *Consistency:* $\hat{\hat{\delta}} \to \delta_0$ *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*
   $\sqrt{n}(\hat{\hat{\delta}} - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 [E(f_i f_i')]^{-1}\right)$
   as n and $n\alpha$ *go to infinity.*

Proof In Appendix.

For the asymptotic normality, we need $n\alpha$ go to infinity as in Carrasco (2012) for 2SLS. 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 $a = 1, ..., p$" 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.

1.2.3 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 $X = (x_1, x_2, ..., x_n)$.

Proposition 3. *(Moments of the regularized LIML)*
Assume $\{y_i, W_i', x_i'\}$ *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 r^{th} *moments* $(r = 1, 2, ..)$ *of the regularized LIML estimators are bounded for all n greater than some* $n(r)$*.*

Proof In Appendix.

Proposition 3 assumes that the eigenvalues of K are strictly decreasing which rules out the case where all the eigenvalues are equal [3]. In Proposition 2, we assumed that K

3. Recall that the eigenvalues are ranked in decreasing order by assumption.

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 the 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. The spectral cut-off regularized estimator seems to be more affected than the estimators obtained by T and LF regularizations.

1.3 Mean square error for regularized LIML

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

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

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

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

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

Assumption 3: (i) $E[(\varepsilon_i, u_i')'(\varepsilon_i, u_i')]$ is bounded, (ii) K is a 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 asymptotic MSE. First, this approach makes comparison with DN (2001) and Carrasco (2012) easier since they also use the Nagar expansion. Second, a finite sample parametric approach may not be so convincing as it would rely on a distributional assumption. Finally, the Nagar approximation provides the tools to derive a simple way for selecting the regularization parameter in practice.

Proposition 4. *Let* $\sigma_{u\varepsilon} = E(u_i\varepsilon_i|x_i)$, $\Sigma_u = E(u_i u_i'|x_i)$ *and* $\Sigma_v = E(v_i v_i'|x_i)$ *with* $v_i = u_i - \varepsilon_i \frac{\sigma_{u\varepsilon}}{\sigma_\varepsilon^2}$. *If Assumptions 1 to 3 hold ,* $\Sigma_v \neq 0$, $E(\varepsilon_i^2 v_i) = 0$ *and* $n\alpha \to \infty$ *for LF, SC, T regularized LIML, we have*

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

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

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

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

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 would be far 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:

$$S_{LIML}(\alpha) = \sigma_\varepsilon^2 \bar{H}^{-1} \left[\Sigma_v \frac{[tr((P^\alpha)^2)]}{n} + \frac{f'(I-P^\alpha)^2 f}{n} \right] \bar{H}^{-1} \tag{1.2}$$

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

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

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

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

The MSE of regularized LIML is of smaller order in α than that of the regularized 2SLS because the bias terms for LIML does not depend on α. This is similar to a result found in DN, namely that the 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.

1.4 Data driven selection of the regularization parameter

1.4.1 Estimation of the MSE

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

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

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

$$W = f + u.$$

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

$$W\bar{H}^{-1}\gamma = f\bar{H}^{-1}\gamma + u\bar{H}^{-1}\gamma \Leftrightarrow W_\gamma = f_\gamma + u_\gamma \tag{1.3}$$

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

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

Using (1.2), $S_\gamma(\alpha)$ can be rewritten as

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

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 (1.3).

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

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

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

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

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

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 i^{th} 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 i^{th} observation from the sample. $\tilde{W}_{\gamma_{-i}}$ is the $(n-1)\times 1$ vector constructed by suppressing the i^{th} observation of $\tilde{W}_\gamma$.

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

$$\hat{\alpha} = \arg\min_{\alpha \in M_n} \left[\hat{R}(\alpha) - \frac{\hat{\sigma}^2_{u_\gamma \varepsilon}}{\hat{\sigma}^2_\varepsilon} \frac{tr((P^\alpha)^2)}{n} \right] \tag{1.4}$$

where M_n is the index set of α. M_n is a compact subset of $[0,1]$ for T, M_n is such that $1/\alpha \in \{1,2,...,n\}$ for 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 (1.4) can be computed without relying on any first step estimate of α (except when Mallows C_p is used).

1.4.2 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

$$\frac{S_\gamma(\hat{\alpha})}{\inf_{\alpha \in M_n} S_\gamma(\alpha)} \xrightarrow{P} 1 \tag{1.5}$$

as n and $n\alpha \to \infty$ where $\hat{\alpha}$ is the regularization parameter defined in (1.4). The result (1.5) does not imply that $\hat{\alpha}$ converges to a true α in some sense. Instead, it establishes that using $\hat{\alpha}$ in the criterion $S_\gamma(\alpha)$ delivers the same rate of convergence as if minimizing $S_\gamma(\alpha)$ directly. For each estimator, the selection criteria provide a means to obtain higher

order asymptotically optimal choices for the regularized parameter. It also means that the choice of α using the estimated MSE is asymptotically as good as if the true reduced form were known.

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 (1.5) for SC and LF. Under Assumptions 1-3 and Assumption 4 (i-v), the leave-one out cross validation is asymptotically optimal in the sense of (1.5) for SC and LF.

Optimality of T

Under Assumptions 1-3 and Assumption 4 (i') and (ii), the Mallows C_p is asymptotically optimal in the sense of (1.5) for Tikhonov regularization.

Proof In Appendix.

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

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

1.5 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, Hausman, and Newey (2008) (denoted HHN in the sequel). In all simulations, we set $\pi = 1$.

Consider

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

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

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

In all simulations, we consider large samples of size $n = 500$ and use 1000 replications.

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 $\pi_l = \sqrt{\frac{R_f^2}{1-R_f^2}}$, $l = 1,2,...,L$ 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 estimators, the starting values for the minimization needed in the estimation of v are the 2SLS using all the instruments when $L \leq 50$ or the corresponding regularized 2SLS for $L > 50$. For standard LIML, the starting value is again the 2SLS using all the instruments when $L \leq 50$ or 1 for $L = 400$ and 520. 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 $\hat{W} = (P^{\alpha} - \nu I_n)W$ 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 [4]. 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.I 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

4. The optimal α for Tikhonov is searched over the interval [0.01,0.5] with 0.01 increment for Models 1 and Model 2. The range of values for the number of iterations for LF is from 1 to 10 times the number of instruments and for the number of principal components is from 1 to the number of instruments.

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.

1.6 Empirical applications

1.6.1 Returns to Schooling

A motivating empirical example is provided by the influential paper of Angrist and Krueger (1991). This study has become a benchmark for testing methodologies concerning IV estimation in the presence of many (possibly weak) instrumental variables. The sample drawn from the 1980 U.S. Census consists of 329,509 men born between 1930-1939. Angrist and Krueger (1991) estimate an equation where the dependent variable is the log of the weekly wage, and the explanatory variable of interest is the number of years of schooling. It is obvious that OLS estimate might be biased because of the

Table 1.I: 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	0.033	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	0.030	0.018	0.287	0.382	0.212
	Med.abs	0.411	0.380	0.449	0.594	0.092	0.249	0.264	0.370	0.486	0.428
	Disp	0.128	0.177	2.291	3.116	0.342	1.110	1.231	1.198	1.719	4.373
	MSE	0.171	0.150	763.56	224.83	0.021	6.9e+20	3.1e+23	1.0e+30	Inf	8.7e+27
	Cov	0.000	0.001	0.752	0.795	0.961	0.927	0.948	0.798	0.792	0.838
L=520	Med.bias	0.426	0.418	0.353	0.449	-0.007	0.080	0.106	0.347	0.450	0.594
	Med.abs	0.426	0.418	0.494	0.608	0.098	0.287	0.267	0.431	0.526	0.954
	Disp	0.114	0.123	2.361	2.951	0.365	1.247	1.053	1.357	1.526	54.807
	MSE	0.184	0.178	34.68	639.34	0.021	6.115	4.5e+21	3.1e+29	Inf	6.9e+29
	Cov	0.000	0.000	0.743	0.740	0.961	0.912	0.895	0.803	0.778	0.435

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 1.II: 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	747.443	22.783	23.297	0.326	736.191	31.248	30.903
	sd	0.197	1385.074	87.568	92.740	0.197	1368.495	95.671	99.198
	q1	0.110	73.000	1.000	1.000	0.110	73.000	1.000	1.000
	q2	0.430	153.500	1.000	1.000	0.430	152.000	3.000	3.000
	q3	0.500	522.500	7.000	5.000	0.500	513.500	14.000	10.000

Table 1.III: Simulations results of Model 2, $n = 500$

		T2SL	L2LS	P2LS	D2LS	IV	TLIML	LLIML	PLIML	DLIML	LIML
L=15	Med.bias	0.001	0.001	0.001	0.003	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	0.019	0.018
	Disp	0.068	0.068	0.068	0.068	0.067	0.068	0.068	0.068	0.068	0.069
	MSE	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Cov	0.951	0.951	0.951	0.943	0.952	0.951	0.951	0.951	0.953	0.946
L=30	Med.bias	0.001	0.001	0.001	0.006	0.001	0.000	0.001	0.001	0.002	0.000
	Med.abs	0.017	0.017	0.017	0.018	0.017	0.017	0.017	0.017	0.018	0.018
	Disp	0.067	0.067	0.067	0.066	0.067	0.067	0.067	0.067	0.067	0.066
	MSE	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Cov	0.963	0.961	0.964	0.949	0.958	0.962	0.964	0.965	0.962	0.958
L=50	Med.bias	0.000	0.000	0.000	0.004	0.001	-0.000	-0.001	-0.001	0.001	-0.001
	Med.abs	0.017	0.017	0.017	0.018	0.017	0.017	0.017	0.017	0.018	0.017
	Disp	0.065	0.065	0.065	0.066	0.065	0.065	0.065	0.065	0.066	0.067
	MSE	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Cov	0.955	0.954	0.954	0.945	0.950	0.954	0.953	0.953	0.957	0.952

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 1.IV: Properties of the distribution of the regularization parameters Model 2

		T2SL	L2LS	P2LS	D2LS	TLIML	LLIML	PLIML	DLIML
L=15	Mean	0.330	149.673	3.012	9.936	0.157	149.853	3.012	13.076
	sd	0.085	2.608	0.109	1.203	0.114	1.861	0.109	1.985
	q1	0.290	150.000	3.000	9.000	0.030	150.000	3.000	11.000
	q2	0.345	150.000	3.000	10.000	0.170	150.000	3.000	14.000
	q3	0.390	150.000	3.000	11.000	0.260	150.000	3.000	15.000
L=30	Mean	0.493	299.992	3.011	13.881	0.257	300.000	3.011	24.046
	sd	0.036	0.253	0.114	2.105	0.192	0.000	0.114	4.092
	q1	0.500	300.000	3.000	12.000	0.040	300.000	3.000	22.000
	q2	0.500	300.000	3.000	12.000	0.290	300.000	3.000	23.000
	q3	0.500	300.000	3.000	16.000	0.450	300.000	3.000	28.000
L=50	Mean	0.499	448.503	3.010	13.931	0.305	483.828	3.010	26.343
	sd	0.014	54.664	0.100	0.908	0.204	35.748	0.100	7.897
	q1	0.500	411.000	3.000	14.000	0.070	496.000	3.000	22.000
	q2	0.500	463.000	3.000	14.000	0.410	500.000	3.000	23.000
	q3	0.500	500.000	3.000	14.000	0.500	500.000	3.000	29.000

endogeneity of education. Angrist and Krueger (1991) propose to use the quarters of birth as instruments. Because of the compulsory age of schooling, the quarter of birth is correlated with the number of years of education, while being exogenous. The relative performance of LIML on 2SLS, in presence of many instruments, has been well documented in the literature (DN, Anderson 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):

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

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

Table 1.V reports schooling coefficients generated by different estimators applied to

Table 1.V: 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)
		α= 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)
		α= 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.

the Angrist and Krueger data along with their standard errors [5] in parentheses. Table 1.V shows that all regularized 2SLS and LIML estimators based on the same type of regularization give close results. The coefficients we obtain by regularized LIML are slightly larger than those obtained by regularized 2SLS suggesting that these methods provide an extra bias correction, as observed in our Monte Carlo simulations. Note that the bias reduction obtained by regularized LIML compared to standard LIML comes at the cost of a larger standard error. Among the regularizations, PC gives estimators which are quite a bit smaller than T and LF. However, we are suspicious of PC because there is no factor structure here.

1.6.2 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

5. Our standard errors are not robust to heteroskedasticity.

reciprocal is not different from 1. In this subsection, we follow one of the specifications in Yogo (2004) using quarterly data from 1947.3 to 1998.4 for the United States and compare all the estimators considered in the present paper. The estimated models are given by the following equation:

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

and the "reverse regression":

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

where ψ is the EIS, Δc_{t+1} is the consumption growth at time $t+1$, $r_{f,t+1}$ is the real return on a risk free asset, τ and μ are constants, and ξ_{t+1} and η_{t+1} are the innovations to consumption growth and asset return, respectively.

Yogo (2004) use 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 $Z = [r, i, c, p]$. 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[6] $II = [Z, Z.^2, Z.^3, Z(:,1)*Z(:,2), Z(:,1)*Z(:,3), Z(:,1)*Z(:,4), Z(:,2)*Z(:,3), Z(:,2)*Z(:,4), Z(:,3)*Z(:,4)]$. As a result, the concentration parameters increase in the following way:

Table 1.VI: Concentration parameter μ_n^2 for the reduce form equation.

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

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

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 instrument problem, we expect to obtain reliable point estimates. 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. PC seems to take too many factors, and did not perform well, this is possibly due to the absence of factor structure.

Table 1.VII: 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.

1.7 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.

We restrict our work in this paper to the estimation and asymptotic properties of regularized LIML with many strong instruments. One possible topic for future research would be to extend these results to the case of weak instruments as in Hansen et al. (2008). Another interesting topic is the use of our regularized LIML or 2SLS for inference when facing many instruments or a continuum of instruments. This would enable us to compare our inference results with those of Hansen et al. (2008) and Newey and Windmeijer (2009).

1.8 Appendix

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

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

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)

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

Hence by Markov's inequality, $\varepsilon' P^{\alpha} \varepsilon = O_p\left(\sum_j q_j\right) = O_p\left(1/\alpha\right)$. 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

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

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

In fact

$$\begin{aligned}\hat{A} &= \bar{W}'P^{\alpha}\bar{W}/n &(1.6)\\ &= \frac{D_0'W'P^{\alpha}WD_0}{n} + \frac{D_0'W'P^{\alpha}\varepsilon e}{n} + \frac{e'\varepsilon'P^{\alpha}WD_0}{n} + \frac{e'\varepsilon'P^{\alpha}\varepsilon e}{n}. &(1.7)\end{aligned}$$

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

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

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

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

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

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

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

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

$$\begin{aligned} \tau' A\tau &= \tau' D_0'\left\langle Fg, Fg'\right\rangle D_0\tau && (1.8)\\ &= (\delta_0-\delta)\left\langle Fg, Fg'\right\rangle(\delta_0-\delta)' && (1.9)\end{aligned}$$

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

$$\begin{aligned} \tau' B\tau &= E[(\tau_1 y_i + W_i'\tau_2)^2] && (1.10)\\ &= E[(\tau_1\varepsilon_i + (f_i+u_i)'(\tau_1\delta_0+\tau_2))^2] && (1.11)\\ &= E[(\tau_1\varepsilon_i + u_i'(\tau_1\delta_0+\tau_2))^2] + (\tau_1\delta_0+\tau_2)'H(\tau_1\delta_0+\tau_2) && (1.12)\end{aligned}$$

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

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

Proof of asymptotic normality:

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

The gradient and Hessian are given by

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

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

Then by a standard mean-value expansion of the first-order conditions $\Lambda_{\delta}(\hat{\delta}) = 0$, 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 $B_{\delta\delta}(\tilde{\delta}) = 2W'W/n \xrightarrow{P} 2E(W_i W_i')$, $A_{\delta\delta}(\tilde{\delta}) = 2W' P^{\alpha} W/n \xrightarrow{P} 2\left\langle Fg, Fg' \right\rangle$.

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

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

To obtain the asymptotic normality, note that

$$\frac{W' P^{\alpha} \varepsilon}{\sqrt{n}} = \left\langle (K_n^{\alpha})^{-1} g_n, \frac{\sum_{i=1}^n Z_i(.,x_i)\varepsilon_i}{\sqrt{n}} \right\rangle \tag{1.13}$$

$$= \left\langle K^{-1} g, \frac{\sum_{i=1}^n Z_i(.,x_i)\varepsilon_i}{\sqrt{n}} \right\rangle + \left\langle (K_n^{\alpha})^{-1} g_n - K^{-1} g, \frac{\sum_{i=1}^n Z_i(.,x_i)\varepsilon_i}{\sqrt{n}} \right\rangle \tag{1.14}$$

Moreover, $\{Z_i(.,x_i)\varepsilon_i\}$ is iid with $E\left\|Z_i(.,x_i)\varepsilon_i\right\|^2 < \infty$ (because $E\left(\varepsilon_i^2 | x_i\right)$ is bounded and K is nuclear). It follows from van der 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,

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

As g belongs to the range of K, Lemma 3 of Carrasco (2012) implies that $\left\| (K_n^{\alpha})^{-1} g_n - K^{-1}g \right\| \xrightarrow{P} 0$ and hence the second term of the r.h.s. of (1.14) 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 $v'P^{\alpha}\varepsilon = \frac{1}{n}\sum_j q_j\left(v'\psi_j\right)\left(\varepsilon'\psi_j\right)$

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

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

$$E\left[(v'P^{\alpha}\varepsilon)^2\right] = \frac{1}{n^2}\sum_{j,l} q_j q_l \sum_i E\left(v_i^2\varepsilon_i^2\right)\psi_{ji}^2\psi_{li}^2 + \sum_j q_j^2 E\left(v_i^2\right)E\left(\varepsilon_i^2\right)\left(\frac{\sum_i \psi_{ji}^2}{n}\right)^2. \quad (1.15)$$

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 (1.15) is negligible with respect to the second. By Markov 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 $\left\langle Fg, Fg'\right\rangle = H = E\left(f_i f_i'\right)$ 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, Florens, and Renault (2007). For the asymptotic normality, the beginning of the proof is the same. Let $\hat{\phi}=\dfrac{W'\varepsilon}{\varepsilon'\varepsilon}$, $\phi=\dfrac{\sigma_{u\varepsilon}}{\sigma_\varepsilon^2}$ and $v=u-\varepsilon\phi'$. We have $v'P^\alpha\varepsilon/\sqrt{n}=O_p(1/\sqrt{n\alpha})=o_p(1)$ 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 $\Delta_\alpha = tr(f'(I-P^\alpha)^2 f/n)$.

$$
\begin{aligned}
-\sqrt{n}\sigma_\varepsilon^2\Lambda_\delta(\delta_0)/2 &= (W'P^\alpha\varepsilon-\varepsilon'P^\alpha\varepsilon\frac{W'\varepsilon}{\varepsilon'\varepsilon})/\sqrt{n} && (1.16)\\
&= (f'\varepsilon-f'(I-P^\alpha)\varepsilon+v'P^\alpha\varepsilon-(\hat{\phi}-\phi)\varepsilon'P^\alpha\varepsilon)/\sqrt{n} && (1.17)\\
&= f'\varepsilon/\sqrt{n}+o_p(1)\xrightarrow{d}\mathcal{N}(0,\sigma_\varepsilon^2 H). && (1.18)
\end{aligned}
$$

The conclusion follows from Slutzky'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 follow [7]:

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

where $\nu_\alpha=\min_\delta\dfrac{(y-W\delta)'P^\alpha(y-W\delta)}{(y-W\delta)'(y-W\delta)}$ and $P^\alpha=T(K_n^\alpha)^{-1}T^*$.
The following lemma will be useful in the remaining of the proof.

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

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

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

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

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

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

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

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

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

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

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

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

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

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).

Let us consider $x \in \mathbb{R}^n$. We have

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

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 $q_j \geq v_{\frac{\alpha}{2}}$ for $j \leq j^*_\alpha$ and $q_j < v_{\frac{\alpha}{2}}$ for $j > j^*_\alpha$ and

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

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 $n\alpha \to \infty$, $v_{\frac{\alpha}{2}}$ 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

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

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 $x'\left(P^{\alpha} - v_{\frac{\alpha}{2}} I_n\right) x > 0$.

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,

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

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 $\tilde{\Lambda} = \varepsilon' P\varepsilon/(n\sigma_\varepsilon^2)$ and $\hat{\Lambda} = \Lambda(\hat{\delta})$ with $\Lambda(\delta) = \dfrac{(y - W\delta)'P(y - W\delta)}{(y - W\delta)'(y - W\delta)}$. If the assumptions of Proposition 4 are satisfied, then

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

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

$$\begin{aligned} \hat{\Lambda} &= \Lambda(\delta_0) - (\hat{\delta} - \delta_0)'\Lambda_{\delta\delta}(\delta_0)(\hat{\delta} - \delta_0)/2 + O_p(1/n^{3/2}) && (1.19) \\ &= \Lambda(\delta_0) - \Lambda_{\delta}(\delta_0)'[\Lambda_{\delta\delta}(\delta_0)]^{-1}\Lambda_{\delta}(\delta_0)/2 + O_p(1/n^{3/2}). && (1.20) \end{aligned}$$

As in proof of Proposition 1 and in Lemma A.7 of DN
$-\sqrt{n}\hat{\sigma}_{\varepsilon}^2\Lambda_{\delta}(\delta_0)/2 = h + O_p(\Delta_{\alpha}^{1/2} + \sqrt{1/n\alpha})$ with $h = f'\varepsilon/n$. Moreover,

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

And by combining these two equalities, we obtain

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

Note also that

$$\begin{aligned} \Lambda(\delta_0) &= (\sigma_{\varepsilon}^2/\hat{\sigma}_{\varepsilon}^2)\tilde{\Lambda} = \tilde{\Lambda} - (\hat{\sigma}_{\varepsilon}^2/\sigma_{\varepsilon}^2 - 1)\tilde{\Lambda} + \tilde{\Lambda}(\hat{\sigma}_{\varepsilon}^2 - \sigma_{\varepsilon}^2)^2/(\hat{\sigma}_{\varepsilon}^2\sigma_{\varepsilon}^2) && (1.21) \\ &= \tilde{\Lambda} - (\hat{\sigma}_{\varepsilon}^2/\sigma_{\varepsilon}^2 - 1)\tilde{\Lambda} + O_p(\sqrt{1/n^3\alpha}). && (1.22) \end{aligned}$$

$$\begin{aligned}
\rho_{\alpha n} &= tr(S(\alpha)) && (1.23)\\
&= tr\left(\sigma_\varepsilon^2 \bar{H}^{-1}\left[\Sigma_v \frac{tr(P^2)}{n} + \frac{f'(I-P)^2 f}{n}\right]\bar{H}^{-1}\right) && (1.24)\\
&= tr\left(\sigma_\varepsilon^2 \bar{H}^{-1}[\Sigma_v \frac{tr(P^2)}{n}]\bar{H}^{-1}\right) + tr\left(\sigma_\varepsilon^2 \bar{H}^{-1}\left[\frac{f'(I-P)^2 f}{n}\right]\bar{H}^{-1}\right) && (1.25)\\
&= O_p(1/n\alpha) + \Delta_\alpha. && (1.26)
\end{aligned}$$

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

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

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

To show ii) we can notice that

$$\begin{aligned} E(h\tilde{\Lambda}\varepsilon' v/\sqrt{n}|X) &= E(h\varepsilon' P\varepsilon\varepsilon' v/(n\sigma_\varepsilon^2\sqrt{n})|X) && (1.27)\\ &= \sum_{i,j,k,l} E(f_i\varepsilon_i\varepsilon_j P_{jk}\varepsilon_k\varepsilon_l v_l'^2\sigma_\varepsilon^2)|X) && (1.28)\\ &= \sum_i f_i P_{ii} E(\varepsilon_i^4 v_i'|x_i)/n^2\sigma_\varepsilon^2 + 2\sum_{i\neq j} f_i P_{ij} E(\varepsilon_j^2 v_j'|x_j)/n^2 && (1.29)\\ &+ \sum_{i\neq j} f_i P_{jj} E(\varepsilon_i^2 v_i|x_i)/n^2 && (1.30)\\ &= O(1/n) + (tr(P)/n)\sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n. && (1.31) \end{aligned}$$

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)

$$\begin{aligned} E(hh'\bar{H}^{-1}h/\sqrt{n}|X) &= \sum_{i,j,k} E(f_i\varepsilon_i\varepsilon_j f_j'\bar{H}^{-1} f_k\varepsilon_k|X)/n^2 && (1.32)\\ &= \sum_i E(\varepsilon_i^3|x_i) f_i f_i'\bar{H}^{-1} f_i/n^2 && (1.33)\\ &= O(1/n). && (1.34) \end{aligned}$$

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

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

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

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

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

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 $T^h = \sum_{j=1}^{5} T_j^h$, $T^H = \sum_{j=1}^{3} T_j^H$,

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

9. The expression of T_5^h, Z^h and Z^H below correct 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 $e_f(\alpha) = \dfrac{f'(I-P)f}{n}$ and $e_{2f}(\alpha) = \dfrac{f'(I-P)^2 f}{n}$. By Lemma 3 (ii) $E(hT_4^{h'}|X) = \dfrac{tr(P)}{n}\sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n + O\left(\dfrac{1}{n^2\alpha}\right)$.

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

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

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

By Lemma 5 (iii), $E(hT_5^{h'}) = O_p(1/n)$.
For $\hat{\xi} = \sum_i P_{ii} f_i E(\varepsilon_i^2 v_i'|x_i)/n - \dfrac{tr(P)}{n}\sum_i f_i E(\varepsilon_i^2 v_i'|x_i)/n - \sum_i P_{ii}(1-P_{ii}) f_i E(\varepsilon_i^2 v_i'|x_i)/n$, $\hat{A}(\alpha)$ satisfies

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

We can also show that $\|T_1^h\|\|T_j^h\| = o_p(\rho_{n\alpha})$, $\|T_2^h\|\|T_j^H\| = o_p(\rho_{n\alpha})$ 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 $R(\alpha) = \dfrac{f_\gamma'(I-P)^2 f_\gamma}{n} + \sigma_{u_\gamma}^2 \dfrac{tr(P^2)}{n}$ be the risk approximated by $\hat{R}^m(\alpha)$, $\hat{R}^{cv}(\alpha)$, or $\hat{R}^{lcv}(\alpha)$, and $S_\gamma(\alpha) = \sigma_\varepsilon^2 \left[\dfrac{f_\gamma'(I-P)^2 f_\gamma}{n} + \sigma_{v_\gamma}^2 \dfrac{tr(P^2)}{n} \right]$. For notational convenience, we henceforth drop the γ subscript on S and R. For Mallows C_p, generalized cross-validation and leave one out cross-validation criteria, we have to prove that

$$\sup_{\alpha \in M_n} \left(|\hat{R}(\alpha) - R(\alpha)| / R(\alpha) \right) \to 0 \tag{1.35}$$

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

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

hence Equation 2.6 of Li (1987) holds and Assumption (A.3) can be replaced by (A.3'). The justification for replacing $\sigma_{u_\gamma\varepsilon}^2$, $\sigma_{u_\gamma}^2$ 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 $M_n = \left\{\alpha : \alpha > C/n \text{ with } C > \max(C_1, C_1^2/C_2)\right\}$. This is not very restrictive since α has to satisfy $n\alpha \to \infty$. It follows that

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

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

In the case of leave-one-out cross-validation criterion, we need to verify the assumptions of Theorem 5.1 of Li (1987). Assumption (A.1) to (A.4) still hold as before. Assumptions (A.8), (A.9), and (A.10) hold by Assumption 4 (iii) to (v) of this paper, respectively. This ends the proof of (1.35) for SC and LF.

Continuous index set

The T regularization is a case where the index set is continuous. We apply Li's (1986) results on the optimality of Mallows C_p in the ridge regression. We need to check the Assumption (A.1) of Theorem 1 in Li (1986). (A.1) $\inf_{\alpha \in M_n} nR(\alpha) \to \infty$ holds using the same proof as for SC and LF. It follows that (1.35) holds for T under Assumption 4 (i').

We have proved that (1.35) 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 $R(\alpha) \leq C S_\gamma(\alpha)/\sigma_\varepsilon^2$, we have

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

It follows from (1.35) 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.

CHAPTER 2

EFFICIENT ESTIMATION WITH MANY WEAK INSTRUMENTS USING REGULARIZATION TECHNIQUES

2.1 Introduction

The problem of weak instruments or weak identification has recently received considerable attention [1] in both theoretical and applied econometrics. [2] Empirical examples include Angrist and Krueger (1991) who measure return to schooling, Eichenbaum et al. (1988) who consider consumption asset pricing models. Theoretical literature on weak instruments includes papers by Staiger and Stock (1997), Zivot et al. (1998), Guggenberger and Smith (2005), Chao and Swanson (2005), Han and Phillips (2006), Hansen et al. (2008), and Newey and Windmeijer (2009) among others [3]. Staiger and Stock (1997) proposed an asymptotic framework with local-to-zero parametrization of the coefficients of the instruments in the first-stage regression. They show that with the number of instruments fixed, the two-stage least squares (2SLS) and limited information maximum likelihood (LIML) estimators are not consistent and converge to nonstandard distributions. Subsequent work focused on situations where the number of instruments is large, using an asymptotic framework that lets the number of instruments go to infinity as a function of sample size. In these settings, the use of many moments can improve estimator accuracy. Unfortunately, usual Gaussian asymptotic approximation can be poor and IV estimators can be biased.

---

1. This chapter is a joint work with Marine Carrasco. Carrasco gratefully acknowledges financial support from SSHRC.

2. Hahn and Hausman (2003) define weak instruments, by two features: (i) two-stage least squares (2SLS) analysis is badly biased toward the ordinary least-squares (OLS) estimate, and alternative unbiased estimators such as limited-information maximum likelihood (LIML) may not solve the problem; and (ii) the standard (first-order) asymptotic distribution does not give an accurate framework for inference. Weak instrument may also be an important cause of the finding that the often-used test of over identifying restrictions (OID test) rejects too often.

3. Section 4 discusses related literature in more details.

Carrasco (2012) and Carrasco and Tchuente (2013) proposed respectively regularized versions of 2SLS and LIML estimators for many strong instruments. The regularization permits to address the singularity of the covariance matrix resulting from many instruments. These papers use three regularization methods borrowed from inverse problem literature. The first estimator is based on Tikhonov (ridge) regularization, the second estimator is based on an iterative method called Landweber-Fridman (LF), the third estimator is based on the principal components associated with the largest eigenvalues. We extend these previous works to allow for the presence of a large number of weak instruments or weak identification. We consider a linear model with homoskedastic error and allow for weak identification as in Hansen et al. (2008) and Newey and Windmeijer (2009). This specification helps us to have different types of weak instruments sequences, including the many instruments sequence of Bekker (1994) and the many weak instruments sequence of Chao and Swanson (2005). We impose no condition on the number of moment conditions since our framework allows for an infinite countable or even a continuum of instruments. The advantage of regularization is that all available moments can be used without discarding any a priori. We show that regularized 2SLS and LIML are consistent in the presence of many weak instruments. If properly normalized, the regularized 2SLS and LIML are asymptotically normal and reach the semiparametric efficiency bounds. Therefore, their asymptotic variance is smaller than that of Hansen et al. (2008) and Newey and Windmeijer (2009). All these methods involve a regularization parameter, which is the counterpart of the smoothing parameter that appears in the nonparametric literature. A data driven method was developed in Carrasco (2012) and Carrasco and Tchuente (2013) to select the best regularization parameter when the instruments are strong. We use these methods in our simulations for selecting the regularization parameter when the instruments are weak but we do not prove that these methods are valid in this case. A related paper is that of Hansen and Kozbur (2014) who propose a regularized jackknife instrumental variables estimator in a strong instruments setting where the design is not sparse.

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

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

2.2 Presentation of the regularized 2SLS and LIML estimators

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

Our model is inspired by Hausman et al. (2012). The model is

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

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

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

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

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

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

The estimate δ is based on the orthogonality condition.

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

where the vector of instruments Z_i has dimension L.

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

Let $\mathbf{Z}$ denote the $n \times L$ matrix having rows corresponding to Z_i'. Denote ψ_j the eigen-

vectors of the $n \times n$ matrix $\mathbf{Z}\mathbf{Z}'/n$ associated with eigenvalues λ_j. Recall that two-stage least squares (2SLS) and LIML estimators involve a projection matrix

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

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

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

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

- The Tikhonov (T) regularization: $q\left(\alpha, \lambda_j^2\right) = \dfrac{\lambda_j^2}{\lambda_j^2 + \alpha}$.
- The Landweber-Fridman (LF) regularization: $q(\alpha, \lambda_j^2) = [1-(1-c\lambda_j^2)^{1/\alpha}]$, where c is a constant such that $0 < c < 1/\left\|Z'Z/n\right\|^2$ and $\left\|Z'Z/n\right\|$ denotes the largest eigenvalue of $Z'Z/n$.
- The Spectral Cut-off (SC): $q(\alpha, \lambda_j^2) = I(\lambda_j^2 \geq \alpha)$.

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

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

Consider regularized k-class estimators defined as follows:

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

where ν is either a constant term or a random variable. The case where $\nu = 0$ corre-

sponds to regularized 2SLS estimator studied in Carrasco (2012):

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

and the case $\nu = \nu_{\alpha} = \min_{\delta} \frac{(y-W\delta)'P^{\alpha}(y-W\delta)}{(y-W\delta)'(y-W\delta)}$ corresponds to the regularized LIML studied in Carrasco and Tchuente (2012). We denote $\hat{\delta}$ the regularized 2SLS estimator and $\hat{\delta}_L$ the regularized LIML estimators.

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

2.3 Asymptotic properties

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

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

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

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

more instruments in the first stage equation:

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

the concentration parameter

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

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

Assumption 1:

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

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

These conditions allow for many weak instruments. If $\mu_{jn} = \sqrt{n}$ this leads to asymptotic theory like in Kunitomo (1980), Morimune (1983), and Bekker (1994), but here we use regularization parameter instead of having an increasing sequence of instruments. For μ_n^2 growing slower than n, the convergence rate will be slower that $\sqrt{n}$, leading to an asymptotic approximation as that of Chao and Swanson (2007). This is the case where

we have many instruments without strong identification. Assumption 1 also allows for some components of the reduced form to give only weak identification (corresponding to $\mu_{jn}/\sqrt{n} \to 0$ which allows the concentration parameter to grow slower than $\sqrt{n}$), and other components (corresponding to $\mu_{jn} = \sqrt{n}$) to give strong identification for some coefficients of the reduced form. In particular, this condition allows for fixed constant coefficients in the reduced form. This specification of weak instruments can also be viewed as a generalization of Chao and Swanson (2007) but differs from that of Antoine and Lavergne (2012) who define the identification strength through the conditional moments that flatten as the sample size increases. To illustrate Assumption 1, let us consider the following example.

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

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

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

We also have

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

Assumption 2:

(i) The operator K is nuclear.

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

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

Condition (i) refers to the covariance operator K defined in Appendix A. K is nuclear provided its trace is finite, see for instance Carrasco et al. (2007a). This assumption is

trivially satisfied if L is finite but may or may not be satisfied when L is infinite. This assumption implies in particular that the smallest eigenvalues decrease to zero sufficiently fast. For this to be true, the Z_i have to be correlated with each other. If $E\left(Z_i Z_i'\right) = I_L$ as in Assumption 5 of Newey and Windmeijer (2009), all the eigenvalues of the operator K equal 1 and hence K is not nuclear when L goes to infinity. To see whether Condition (i) is realistic, we examine the properties of the sample counterpart of K, namely $K_n = Z'Z/n$, in three applications: the return to schooling using 240 instruments from Angrist and Krueger (1991) (see also Carrasco and Tchuente (2012)), the elasticity of intertemporal substitution (see Carrasco and Tchuente (2012)), and the application on the effect of institutions on growth (see Section 6 of this paper). In the table below, we report the smallest eigenvalue, the largest eigenvalue, the condition number (which is the ratio of the largest eigenvalue on the smallest eigenvalue) and the trace of $Z'Z/n$ in two cases: raw data and standardized instruments. In the standardized case, the instruments are divided with their standard deviation. This standardization has no impact on 2SLS and LIML estimators which are scale invariant. However, our estimators are not scale invariant and standardization may improve the results. Such standardizations are customary whenever regularizations are used, see for instance De Mol et al. (2008) and Stock and Watson (2012). We observe that in all applications, the smallest eigenvalue is close to zero so the instruments are strongly correlated [4]. The condition number - which is scale invariant - is an indicator on how ill-posed the matrix K_n is. The higher the condition number, the more imprecise the inverse of K_n will be. The smallest possible condition number is 1 (which corresponds to the identity matrix). Here, the condition numbers are all very large which suggests that regularization will be helpful to improve the reliability of the estimate of K^{-1}. The trace of K_n appears to be finite throughout the applications.

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

Table 2.I: Properties of $Z'Z/n$

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

Condition (ii) guarantees that the optimal instrument f can be approached by a sequence of instruments. It is similar to Assumption 4 in Hansen et al. (2008). Condition (iii) is a technical assumption which can also be found in Carrasco (2012). Assumptions 2(ii) and (iii) are needed only for efficiency.

Proposition 6. *(Asymptotic properties of regularized 2SLS with many weak instruments)*

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

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

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

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

Proof In Appendix.

The first point of Proposition 1 implies the consistency of the estimator, namely $(\hat{\delta} - \delta_0) \to 0$ (see the proof of Theorem 1 in Hansen and Kozbur (2014)). Moreover, Proposition 1 shows that the three estimators have the same asymptotic distribution. Instead of restricting the number of instruments (which may be very large or infinite), we impose restrictions on the regularization parameter which goes to zero. This insures

us that all available and valid instruments are used in an efficient way even if they are weak. To obtain consistency, the condition on α is $n\alpha^{\frac{1}{2}}$ go infinity, whereas for the asymptotic normality, we need $n\alpha$ go to infinity. This means that α is allowed to go to zero at a slower rate. However, this rate does not depend on the weakness of the instruments.

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

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

Proposition 7. *(Asymptotic properties of regularized LIML with many weak instruments)*

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

1. *Consistency:* $S_n'(\hat{\delta}_L - \delta_0)/\mu_n \to 0$ *in probability as n,* μ_n *and* $\mu_n^2\alpha$ *go to infinity.*
2. *Asymptotic normality:*
   $S_n'(\hat{\delta}_L - \delta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 \left[E(f_i f_i')\right]^{-1}\right)$
   as n, μ_n *and* $\mu_n^2\alpha$ *go to infinity where* $E(f_i f_i')$ *is a nonsingular* $p \times p$ *matrix.*

Proof In Appendix.

Again, Proposition 1 implies the consistency of the estimator, namely $(\hat{\delta} - \delta_0) \to 0$. Interestingly we obtain the same asymptotic distribution as in the many strong instruments case (with a slower rate of convergence). We also find the same speed of convergence as in Hansen et al. (2008) and Newey and Windmeijer (2009). For the consistency and asymptotic normality, $\mu_n^2\alpha$ needs to go to infinity, which means that the regularization parameter should go to zero at a slower rate than the concentration parameter. The asymptotic variance of regularized LIML corresponds to the lower bound and is smaller than that obtained in Hansen et al. (2008). We believe that the reason, why Hansen et al.

(2008) obtain a larger asymptotic variance than us, is that they use the number of instruments as regularization parameter. As a result, they can not let L grow fast enough to reach efficiency. Our estimator involves the extra tuning parameter α which is selected so that extra terms in the variance vanish asymptotically. Moreover, we assume that the set of instruments is sufficiently rich to span the optimal instrument (Assumption 2(ii)).

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

Now, as in Newey and Windmeijer (2009), we consider a t-ratios for a linear combination $c'\delta$ of the parameter of interest. The following proposition is a corollary of Proposition 6 and 7.

Proposition 8. *Under the assumptions of Proposition 2 and if we assume that there exist r_n, c and $c^* \neq 0$ such that $r_n S_n^{-1} c \to c^*$ and $S_n'\hat{\Phi}S_n/n \to \Phi$ in probability with $\Phi = \sigma_\varepsilon^2 \left[E(f_i f_i')\right]^{-1}$.*

Then,

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

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

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

2.4 Efficiency and Related Literature

2.4.1 Efficiency

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

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

Hence,

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

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

So the lowest asymptotic variance that can be obtained is $\sigma_\varepsilon^2\left[E\left(f_i f_i'\right)\right]^{-1}$. We will refer to this as the semiparametric efficiency bound [5].

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

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

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

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

2.4.2 Related Literature

In the literature on many weak instruments, the asymptotic behavior of estimators depends on the relation between the number of moment conditions L and sample size n. For the CUE, L and n need to satisfy $L^2/n \to 0$ for consistency and $L^3/n \to 0$ for asymptotic normality. Under homoskedasticity, Andrews and Stock (2005) require $L^2/n \to 0$. Hansen, Hausman, and Newey (2008) allowed L to grow at the same rate as n, but restricted L to grow slower than the square of the concentration parameter, for the consistency of LIML and FULL. Andrews and Stock (2006) require $L^3/n \to 0$ when normality is not imposed.

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

Belloni et al. (2012b) propose to use an alternative regularization named lasso in the IV context. This regularization imposes a l_1 type penalty on the first stage coeffi-

cient. Assuming that the first stage equation is approximately sparse, they show that the postlasso estimator reaches the asymptotic efficiency bound.

Just as 2SLS is not consistent if L is too large relative to n, LIML estimator is not feasible if $L > N$ because the matrix $Z'Z$ is not invertible. Therefore, some form of regularization needs to be implemented to obtain consistent estimators when the number of instruments is really large. Regularization has also the advantage to deliver an asymptotically efficient estimator.

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

Table 2.II: Comparison of different IV asymptotics

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

2.5 Monte Carlo study

We now carry out a Monte Carlo simulation for the simple linear IV model where the disturbances and instruments have a Gaussian distribution and the instruments are independent from each other as in Newey and Windmeijer (2009). The parameters of

this experiment are the correlation coefficient ρ between the structural and reduced form errors, the concentration parameter (CP), and the number of instruments L.

The data generating process is given by:

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

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

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

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

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

where ι_L is an L-vector of ones. The sample size is $n = 500$. The instruments are $Z_i = x_i$ and the number of instruments L equals 30 and 50. Note that this setting is not favorable for us because the eigenvalues of the matrix $Z'Z/n$ are all equal to 1. If L were infinite, the matrix $Z'Z/n$ would become an infinite dimensional identity matrix which is not nuclear. Therefore, our basic assumption of compactness of the operator K would not satisfied if we would let L go to infinity. However, here L being no larger than 50, K is nuclear.

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

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

We report the median bias (Med.bias), the median of the absolute deviation of the estimator from the true value (Med.abs), the difference between the 0.1 and 0.9 quantiles

(dis) of the distribution of each estimator, and the coverage rate (Cov.) of a nominal 95% confidence interval for unfeasible instrumental variable regression (IV), regularized two-stage least squares (T2SLS (Tikhonov), L2SLS (Landweber Fridman), P2SLS (Principal component)), LIML and regularized LIML (TLIML (Tikhonov), LLIML (Landweber Fridman), PLIML (Principal component)) and Donald and Newey's (2001) 2SLS and LIML (D2SLS and DLIML). For confidence intervals, we compute the coverage probabilities using the following estimate of asymptotic variance as in Donald and Newey (2001) and Carrasco (2012).

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

where $\hat{W} = P^{\alpha}W$ for 2SLS and $\hat{W} = (P^{\alpha} - \nu I_n)W$ for LIML. Note that the formulaes for the confidence intervals is the same as for strong instruments (see Carrasco and Tchuente (2012)).

Table 2.III reports simulation results. We use different strength (measured by the concentration parameter) of instruments and number of instruments. We investigate the case of very weak instruments for example, when $CP = 8$ and $L = 50$, the first stage F-statistic equals $\frac{CP}{L} + 1 = 1.16$.

We observe that

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

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

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

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

		T2SLS	**L2SLS**	**P2SLS**	**D2SLS**	**IV**	**TLIML**	**LLIML**	**PLIML**	**DLIML**
L=30										
CP=8	Med.bias	0.3909	0.3825	0.3271	0.3520	-0.0325	0.0559	0.0648	0.3228	0.3513
	Med.abs	0.3909	0.3825	0.4947	0.5372	0.2337	0.4286	0.4243	0.4511	0.4598
	Disp	0.3690	0.4148	2.5366	2.8523	1.0464	2.3226	2.1875	1.4859	1.6547
	Cov	0.2550	0.3170	0.7570	0.7710	0.9600	0.9300	0.9230	0.8030	0.7930
CP=35	Med.bias	0.2276	0.2211	0.2245	0.2540	-0.0099	-0.0127	-0.0071	0.0858	0.1125
	Med.abs	0.2276	0.2211	0.2623	0.2846	0.1099	0.1499	0.1434	0.1652	0.1785
	Disp	0.2990	0.3124	0.6175	0.6887	0.4200	0.6014	0.6237	0.6382	0.6592
	Cov	0.4860	0.5220	0.7100	0.6540	0.9660	0.9580	0.9570	0.8590	0.8470
CP=65	Med.bias	0.1499	0.1443	0.1548	0.1905	-0.0097	-0.0076	-0.0050	0.0194	0.0178
	Med.abs	0.1499	0.1446	0.1826	0.2065	0.0836	0.0987	0.0955	0.0972	0.1016
	Disp	0.2500	0.2538	0.4139	0.4502	0.3231	0.3822	0.3870	0.3825	0.3755
	Cov	0.6480	0.6810	0.7560	0.6900	0.9560	0.9670	0.9670	0.9130	0.9090
L=50										
CP=8	Med.bias	0.4220	0.4155	0.3969	0.4103	0.0044	0.0865	0.1483	0.3860	0.4096
	Med.abs	0.4220	0.4155	0.5418	0.5984	0.2450	0.4988	0.4697	0.4730	0.5230
	Disp	0.2999	0.3468	2.4463	2.8784	1.0486	2.7924	2.8911	1.5870	1.7716
	Cov	0.0650	0.1420	0.7450	0.7670	0.9520	0.9160	0.9240	0.7960	0.8030
CP=35	Med.bias	0.2865	0.2761	0.2654	0.2789	0.0056	0.0105	0.0123	0.1475	0.2113
	Med.abs	0.2865	0.2761	0.2961	0.3243	0.1117	0.1719	0.1796	0.2103	0.2495
	Disp	0.2359	0.2624	0.8402	1.1239	0.4450	0.7549	0.7242	0.7554	0.8485
	Cov	0.1670	0.2410	0.6730	0.6360	0.9530	0.9520	0.9520	0.8420	0.7920
CP=65	Med.bias	0.2155	0.2080	0.2020	0.2448	0.0036	0.0051	0.0059	0.0695	0.0789
	Med.abs	0.2155	0.2080	0.2214	0.2701	0.0829	0.1054	0.1129	0.1212	0.1250
	Disp	0.2228	0.2413	0.4840	0.5105	0.3169	0.4208	0.4085	0.4261	0.4568
	Cov	0.3170	0.3650	0.6840	0.6450	0.9590	0.9520	0.9510	0.8520	0.8450

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

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

2.6 Empirical application: Institution and Growth

This section revisits Hall and Jones (1999) empirical work. Hall and Jones (1999) argue that the difference between output per worker across countries is mainly due to the differences in institution and government policies - the so-called social infrastructure. They write "Countries with corrupt government officials, severe impediments to trade, poor contract enforcement, and government interference in production will be unable to achieve levels of output per worker anywhere near the norms of western Europe, northern America, and eastern Asia." To quantify the effect of social infrastructure on per capita income, they use two-stage least squares (2SLS) with four instruments: the fraction of population speaking English at birth (EnL), the fraction of population speaking one of the five major European languages at birth (EuL), the distance from the equator [6] (latitude , Lt) and Romer and Frankel (1999) geography-predicted trade intensity (FR). The linear IV regression model is given by:

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

where y is an $n \times 1$ vector of log income per capita, S is an $n \times 1$ vector which is the proxy for social infrastructure, θ is an $L \times 1$ vector, c and δ are scalars. Dmitriev (2013) points out the fact that the instruments [7] $X = [EnL, EuL, Lt, FR]$ are weak. To address this issue, we increased the number of instruments from 4 to 18. The 18 instruments in our regression are derived from X and are given by [8]

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

7. This correspond to the specification (iv) of Dmitriev (2013).

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

$Z = [X, X.^2, X.^3, X(:,1)*X(:,2), X(:,1)*X(:,3), X(:,1)*X(:,4), X(:,2)*X(:,3), X(:,2)*X(:,4), X(:,3)*X(:,4)]$ where all instruments are divided by their standard deviation.

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

We use a sample of 79 countries for which no data were imputed [9]. The results are reported in Table 2.IV below. Robust to heteroskedasticity standard errors are given in parentheses. They are computed using the formula of Carrasco and Tchuente (2012):

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

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

Table 2.IV: Institutions and growth

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

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

Our findings suggest that "social infrastructure" has a significant causal effect on long-run economic performance throughout the world. The use of many instruments

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

first increase the bias as illustrated by the fact that the distance between 2SLS and LIML is larger when 18 instruments are used. When the regularization is introduced, this gap shrinks. For instance, the regularized LF, LIML and 2SLS are very close, this may be due to bias correction. But, for the PC regularization, the gap remains wide. The reason may be due to the lack of factor structure in the instruments set.

2.7 Conclusion

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

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

2.8 Appendix

2.8.1 General notation

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

We define the covariance operator K of the instruments as

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

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

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

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

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

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

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

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

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

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

1. **Tikhonov(T)**

   This regularization scheme is related to the ridge regression.

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

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

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

2. **Landweber Fridman (LF)**

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

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

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

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

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

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

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

   for $\alpha > 0$.

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

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

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

These regularized inverses can be rewritten in common notation as:

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

where for T: $q(\alpha, \lambda_j^2) = \dfrac{\lambda_j^2}{\lambda_j^2 + \alpha}$,
for LF: $q(\alpha, \lambda_j^2) = [1-(1-c\lambda_j^2)^{1/\alpha}]$,

for SC: $q(\alpha,\lambda_j^2) = I(\lambda_j^2 \geq \alpha)$, for PC $q(\alpha,\lambda_j^2) = I(j \leq 1/\alpha)$.

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

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

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

and

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

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

such that $K_n = T^*T$ and TT^* is an $n \times n$ matrix with typical element $\dfrac{\langle Z_i, Z_j \rangle}{n}$. Let $\hat{\phi}_j$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... > 0$, $j = 1,2,...$ be the orthonormalized eigenfunctions and eigenvalues of K_n. $\hat{\lambda}_j$ are consistent estimators of λ_j the eigenvalues of TT^*. We then have $T\hat{\phi}_j = \sqrt{\lambda_j}\psi_j$ and $T^*\psi_j = \sqrt{\lambda_j}\hat{\phi}_j$.

For $v \in \mathbf{R}^n$, $P^\alpha v = \sum_{j=1}^{\infty} q(\alpha,\lambda_j^2)\langle v, \psi_j \rangle \psi_j$. It follows that for any vectors v and w of $\mathbf{R}^n$:

$$\begin{aligned} v'P^\alpha w &= v'T(K_n^\alpha)^{-1}T^*w & (2.1) \\ &= \left\langle (K_n^\alpha)^{-1/2}\sum_{i=1}^{n} Z_i(.)\, v_i, (K_n^\alpha)^{-1/2}\frac{1}{n}\sum_{i=1}^{n} Z_i(.)\, w_i \right\rangle. & (2.2) \end{aligned}$$

2.8.2 Proofs

Proof of Proposition 1:

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

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

We have $S_n'(\hat{\delta} - \delta_0)/\mu_n = [S_n^{-1} W' P^{\alpha} W S_n^{-1'}]^{-1} [S_n^{-1} W' P^{\alpha} \varepsilon/\mu_n]$ and by construction [10] of P^{α} :

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

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

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

Hence,

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

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

$$\left\langle (K_n^{\alpha})^{-\frac{1}{2}} g_{n1}, (K_n^{\alpha})^{-\frac{1}{2}} g'_{n1} \right\rangle \to \left\langle g_1, g'_1 \right\rangle_K$$

in probability as n and $n\alpha^{\frac{1}{2}}$ go to infinity, with $\left\langle g_1, g'_1 \right\rangle_K$ a $p \times p$ matrix with (a,b) element $\left\langle K^{-\frac{1}{2}} E(Z(.,x_i) f_{ia}), K^{-\frac{1}{2}} E(Z(.,x_i) f_{ib}) \right\rangle$ which is assumed to be nonsingular.

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

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

For the asymptotic normality we write

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

We then have

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

Moreover,

$$\left\langle (K_n^{\alpha})^{-1/2}g_{n1}, (K_n^{\alpha})^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\varepsilon_i \right\rangle \tag{2.3}$$

$$= \left\langle (K_n^{\alpha})^{-1}g_{n1} - K^{-1}g_1, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\varepsilon_i \right\rangle \tag{2.4}$$

$$+ \left\langle K^{-1}g_1, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\varepsilon_i) \right\rangle. \tag{2.5}$$

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

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

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

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

we can compute the inner product in the RKHS and show that

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

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

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

This completes the proof of Proposition 1.

Proof of Proposition 2:

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

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

Proof: The proof of this lemma is the same as in Hansen et al. (2008).

Lemma 2: Let us assume that there exists a constant C such that $E(\|\varepsilon_i\|^4|X) \leq C$ and $E(\|u_{ai}\|^4|X) \leq C$ for all i. Then,

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

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

Proof:

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

$\sigma_{\varepsilon u_a}$ and $E(u'_{ai}u_{ai}) = \sigma^2_{u_a}$,

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

Using the spectral decomposition of P^α, we have

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

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

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

Thus

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

The second conclusion follows by Markov inequality.

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

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

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

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

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

Proof: By the definition of P^{α}, we have (see Equation (2.2)):

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

with

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

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

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

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

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

with probability one.

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

parameter and e is the first unit vector.

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

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

Proof of consistency

Let us consider

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

$$\hat{\delta}_L = argmin Q(\delta).$$

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

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

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

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

Let us show that

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

Let $D(\delta) = \mu_n^{-2}(y - W\delta)' P^{\alpha}(y - W\delta) = \mu_n^{-2}(1, -\delta')\bar{W}' P^{\alpha} \bar{W}(1, -\delta')'$. Moreover, $D(\delta) = \mu_n^{-2}(1, -\delta') D_0' S_n \dfrac{f' P^{\alpha} f}{n} S_n' D_0 (1, -\delta')' + o_p(1) = \mu_n^{-2}(1, -\delta') D_0' S_n E\left(f f'\right) S_n' D_0 (1, -\delta')' +$

$o_p(1)$. It follows from lemma 3 that

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

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

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

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

Now let us prove the asymptotic normality.

Proof of asymptotic normality

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

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

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

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

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

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

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

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

It then follows that

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

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

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

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

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

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

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

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

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

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

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

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

Hence

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

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

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

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

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

The conclusion follows from Slutzky's theorem.

CHAPTER 3

HIGH SCHOOL HUMAN CAPITAL PORTFOLIO AND COLLEGE OUTCOMES

3.1 Introduction

In modern labor markets, workers specialize in specific occupations. For many professional occupations, specialization begins when individuals choose their field of study in university. [1] Before entering the university or college, individuals may acquire particular skills in high school. Every field of study requires a specific set of skills. Conversely, many skills are useful, to different degrees, in a wide variety of fields. A psychology student, a law student and a biology student apply similar skills involving reading, writing and arithmetic albeit in different amounts. Moreover, some majors appear to more heavily emphasize a small subset of particular skills whereas other majors more or less weigh skills evenly. Engineering and mathematics students, for instance, are likely to use more mathematics than literature students.

In high school, individuals are uncertain about their future college major and performance. As a result, a high school graduate may study natural sciences courses and end up majoring in an unrelated field. Faced with uncertainty, high school students may want to balance their efforts in case their initial target does not work out. However, if they specialize in a particular skill they will be more productive in a related field. Therefore, they will want to choose the composition of their high school courses to acquire a set of skills based on their inherent abilities and on their prospective field of study in college.

To investigate the tension between specialization and diversification in high school and its effects on college major choice and performance, I first establish panel data ev-

1. I thank the participants of 9th CIREQ Ph.D. students conference (Montreal, June 2013), of the seminars at the University of Montreal for helpful comments. Comments from Andriana Bellou, Marc Henry and Joshua Lewis are gratefully acknowledged. I am much indebted to Marine Carrasco and Baris Kaymak for intensive discussions and advices.

idence linking individual high school transcript courses with college major choice and performance. I then propose a structural dynamic model of high school skills acquisition and college major choice under uncertainty. This model not only sheds light on the patterns of skill acquisition in high school, and then in university, but also allows for a quantitative evaluation of policies geared towards encouraging certain majors, such as those related to science and technology.

To assess the relationship between courses taken in high school and college major choice, I use the 1980 High School and Beyond (HS&B) survey, which has detailed information on transcripts. High school transcripts in HS&B allow for the construction of empirical measures of human capital portfolios that help investigate the underlying relation between the skill set, the innate ability and college major choice and performance. I find that students choose their majors according to the subject they concentrated in high school. However, there is an U-shaped relation between diversification and college performance: students that are heavily specialized in a particular subject and those that are broadly diversified across subjects have, on average, higher grade point averages (GPA) in college. As a result, students targeting the same first major are likely to acquire different portfolios of high school courses for their intended major as well as their back-up plan.

To investigate the implications of this tension on college major choice, I propose and estimate a structural model of high school human capital acquisition and college major choice. By explicitly modeling the educational decision process, I both disentangle the heterogeneous effects of specialization and control for the self-selection inherent in educational outcomes. In the model, students are endowed with different innate abilities and have two decision periods. In these periods they chooses which high school course to attend and their college major. In the first period students choose the high school courses that maximize their expected discounted utility across majors. Upon graduation from high school, in the second period, students choose their major and observe their major-specific preferences.

Estimation results suggest that quantitative majors are preferred by specialized students even after controlling for selection. High school courses also play an important role in determining college major choice. More quantitative courses in high school increase the likelihood of majoring in natural science, maths & physics and engineering, while more humanities courses are better for social science & humanities majors, business & communication. These results suggest that an appropriate high school quantitative curriculum can increase enrollments in Science Technology Engineering, and Math (STEM) [2] majors.

To confirm this intuition, I perform two types counterfactuals experiments. The first one allows for one more high school course requirement while the second experiment imposes the same high-school curriculum for everyone, completely eliminating specialization. Both experiments show substantial effect on enrollment in STEM majors. Increase of enrollments in STEM is an issue of interest in many countries. In the U.K., the Royal Academy of Engineering reported that the nation will have to graduate 100 000 STEM majors every year until 2020. President of the US Council of Advisors on STEM, stated that over the next decade, 1 million additional STEM graduates will be needed. Moreover, one more quantitative course increases enrollment in STEM majors by 4 to 5 percentage points.

There is a huge literature on college major choice [3]. Most of the theoretical framework in this literature implies that college major choice is influenced by expectations of future earnings, preferences, ability, and preparation (see Altonji et al. (2012) for more details). Turner and Bowen (1999) document the sorting that occurs across majors by SAT math and verbal scores. Arcidiacono (2004) finds that the differences in monetary returns explain little of the ability sorting across majors, and concludes that virtually all ability sorting is because of preferences for particular majors in college and the workplace, with the former being larger than the latter. The present model extends

2. Increase of enrollments in STEM majors is an issue of considerable interest in many rich countries.

3. See Montmarquette et al. (2002), Zafar (2009), Stinebrickner and Stinebrickner (2011), Arcidiacono (2005), Arcidiacono et al. (2013)

Arcidiacono (2004) to add college preparation where students can have a specialization or diversification strategy.

A related strand of the literature studies the causal effect of high school curriculum on labor market outcomes (Altonji (1995) and followed by Levine and Zimmerman (1995) and Rose and Betts (2004)). More recently Joensen and Nielsen (2009) and Goodman (2009) rely on quasi-experiments to estimate the effect of math coursework on earnings. The main research question these studies aim to answer is: do skills accumulated in high school matter for college performance and labor market outcomes? Relative to these papers, I investigate the effect of the *composition* of skills acquired in high school on college performance. It therefore contributes the existing studies by introducing multi-dimensional endowments of skills and by studying the tension between specialization and diversity. In this sense, this paper is closer to Malamud (2010), Smith (2010) and Malamud (2012) who examine the trade-off between specialized and diversified human capital in college and its effect on labor market outcomes. Silos and Smith (2013) study how diversification and specialization strategies in college influence income dynamics. They find that diversification generates higher income for individuals who switch occupations whereas specialization benefits those who stick with one type of job. This paper considers this issue one step back and investigate how specialization in high school affect college major choice and performance.

The paper proceeds as follows. Section 2 provides a brief overview of the US high school system and explains why the US system is a unique opportunity to investigate the effect of high school courses choice on college outcome. Section 3 describes the data and the sample used in the empirical analysis. It also discuss some data regularities and provides a reduced form analysis of the relationship between diversification in high school on college performance. The dynamic model of college and major choice as well as the econometric techniques used to estimate the model are described in Section 4. Section 5 provides the empirical and simulations results. Section 6 concludes.

3.2 Background: High school course choice in US

The US high school education system provides a particularly appropriate, if not unique, setting in which one can examine the effect of specialization in high school system. In the US schools, students have much more control of their education, and are even allowed to choose their core classes. The control given to students varies from state to state [4] and from school to school. It therefore results in a substantial variation in students academic experiences, within schools and across them (Lee et al. (1997), Allensworth et al. (2009)). There is a wide variance in the curriculum required each year but many schools require that courses in the "core "areas of English, science, social studies, and mathematics be taken by the students every year although other schools set the required number of credits and allow the student to choose when the courses will be taken.

The availability of courses depends upon each particular school's financial situation and school staff decisions or preferences. This affects the possibility of specialization in particular set of skills [5]. The degree of flexibility in choice in high school is also a direct function of the preferences of school teachers, these preferences are usually idiosyncratic. Furthermore, inducements for students to take a particular set of skills may change between school as certain teachers are hired or school administrators decide to place greater emphasis on these skills. Thus, there is a substantial element of exogenous variation course choice across schools due to idiosyncracies among teachers, school administrators and state. I will take advantage of these exogenous variations to identify how the specialization in high school affects college performance.

4. See for example Goodman (2009) Figure 2 for difference in state requirement in maths.

5. State requirements for graduation can be found on web pages of States Department of education see for example http://www.azed.gov/hsgraduation/ (view on 2013/09/12)

3.3 Data and Empirical regularities

3.3.1 Data

To investigate the empirical relationship between portfolios of courses acquired through formal high school education and post-secondary education outcome, I use the 1980 High School and Beyond (HS&B) survey data. HS&B has detailed information from high school to post-secondary. I study This panel dataset contains a combination of information on courses taken in different subject of study as well as information on post-secondary education. The HS&B survey was conducted by the National Center for Education Statistics. A nationally representative sample of high school students who were sophomores in 1980 were interviewed once every two years between 1980 and 1986 and once again in 1992. High school usually runs either from grades 9 to 12 or from grades 10 to 12. In this work I restrict high school from 10 to 12 grade because of the availability of data for all high schools in the sample. These interviews recorded detailed informations on student courses in various dimensions of skills when they were in high school and high school transcripts. These high quality data provide the measures of human capital. The Post-Secondary Education Data System (PETS) contains institutional transcripts from all post-secondary institutions attended for a sub-sample of students present in the survey and will be used to have college performance of student.The estimations are performed using data from the 1980, 1982, 1984 and 1986 surveys.

HS&B survey initially contains 14,825 students. A sub-sample of 5,533 students had their transcripts encoded for high school and college. Dropping those who do not have SAT data reduced the sample to 1810 individuals. Taking into account others controls variables reduced the sample to 1265. Cleaning of the data yields a final sample of 1112 students for structural model estimation. Table 1 shows average characteristics for the unrestricted and restricted samples and in almost all cases, there is no significant difference in mean values between the two samples. This is somewhat suggesting that sample selection issues may not be a big problem. The Appendix provides a step-by-step

description of construction of human capital portfolios and college major aggregation.

3.3.2 Empirical regularity

I first provide empirical motivation using data from HS&B. Human capital portfolios are calculated using high school formal courses and contain seven categories of study. Components of high school study are grouped into: (i) Quantitative (mathematics and physics); (ii) Reading and writing; (iii) Social science and Humanities (iv) Life sciences; (v) Business and communication; (vi) Arts; (vii) Others.

Given courses taken in each field or type of human capital $k = 1, ...K$, the weights in the human capital portfolio of an individual i are:

$$\omega_{i,k} = \frac{course_{i,k}}{\sum_{j=1}^{K} course_{i,j}},$$

where $K = 7$ and $course_{i,k}$ is the number of courses taken in subject k. I use the share of each subject to concentrate not on the number of courses but on the distribution. Table 3.II displays these portfolio weights by major across the population. For each major, the table displays the mean, across individuals, of the weights in each of the seven subjects.

The mean weight on quantitative subjects varies from low values in Education (0.165) and Business and communication (0.169) majors to higher values for Engineering and Science major. It is not surprising that Humanities majors have the highest mean weight in humanities subject (0.258). Business and Communication major have also the highest shares of business and communications subject, allocating about 10% of total course on average to this component, this large given high requirement in other type of skills. Although the difference in mean in some subject appears small, the two last lines of Table 3.II shows that these differences are statistically significant.

Each student i has a vector of human capital weights $\omega_{i,k}$ components measuring the weight of skill of type k in the overall portfolio. A skewed or balanced portfolio does not necessarily imply specialization or diversification of human capital investments. Some

students may choose a uniform allocation of course across fields to self-insure against shocks or because a particular major explicitly rewards balanced skills. To evaluate the level of specialization I follow Silos and Smith (2012). I therefore assess how well tailored an individual's acquired skill set is for a particular college field by viewing human capital investments relative to a benchmark in that major [6].

Let us define the measure of diversification as

$$\rho_{i,m} = \sqrt{\sum_{k=1}^{K} (\omega_{i,k} - \bar{\omega}_{k,m})^2}$$

where $\bar{\omega}_{k,m}$ denotes the typical (or average) portfolio for major m observed in Table 3.II. To interpret this measure of specialization. I assume that, a portfolio is chosen for a given major if that portfolio is "close"to the average portfolio of that major. Self-insurance against shocks is simply the distance between the portfolio weights and the typical portfolio of the college major. It is then possible to hedge with respect to your major by diversifying your portfolio (with respect to major related subject) or specializing yourself in major related subjects.

3.3.2.1 Estimation results

Table 3.III presents regression estimates linking the observed college performance measure by grade (GPA) and the portfolio distance measure, ρ. This helps us investigate the empirical regularities beyond raw mean difference [7].

I estimate the following reduced form equation

$$G_{imh} = \alpha_0 + \alpha_1 \rho_{im} + \alpha_2 \rho_{im}^2 + \alpha_3 X_i + \alpha_m + \alpha_h + \varepsilon_{ih}$$

6. This measure is related to Krugman (1992) diversification index in trade where absolute distance instead of square root. See also Palan (2010) for review of specialization index in trade.

7. The division of human capital into seven types of skills is obviously not the only one possible. I consider different division. The results are very similar to those obtained with seven types of skills. I also consider other diversification measure as the Gini index and other relatives measures and the results were qualitatively the same

where α_m and α_h are respectively major and high school fixed effects. G_{mh} is the college GPA of individual i in major m from high school h, X represents other variables such as SAT scores, Socioeconomic status (SES) and gender.

Table 3.III shows the relation between GPA and the measure of diversification ρ is quadratic, large and significant. This results is robust to control for demographics characteristic gender and race, background: Socioeconomic status (SES), parents education, ability by SAT Math and SAT Verbal and number of course taken in high school for each type of human capital. It is also robust to regional disparities by including of dummy variable for living in south. The major specific effect is controlled by including dummy for each major. Furthermore the inclusion of more control variables increased the effect of specialization on college performance. This suggests that the effect of specialization on performance may be larger than the estimates reported here.

3.3.2.2 U-shaped: a comment and robustness check

According to Table 3.III result, the relation between college performance measured by GPA and diversification measured by ρ is a U-shaped relation. This shows trade-off between specialization and diversification. The trade-off is driven by two mains opposite forces implied by diversification strategy. Diversification reduces the human capital in the targeted college major, it also increases skills in other human capitals. As level of diversification increases, the negative effect first dominates since enough knowledge is not acquired to compensate the losses in major specific skills. However, there comes a point where performance increases with the level of diversification. Others skills acquired can now compensate the losses through complementarity.

This result shows that high school human capital plays an important role for college outcomes. The results discussed here are based on the presence of a U-shaped relation. To test properly for the presence of a U-shape I use the procedure proposed by Lind and Mehlum (2010). Results in Appendix Table 3.IV show the presence of a U-shaped relationship. I also perform a non parametric robustness check. I run a regression on all

control variables used in our best regression in Table 3.III. I perform a non parametric regression of residual on ρ. The predict values of residual of the non parametric regression show a U-shaped relationship between diversification and performance. The results are in Figure 3.1.

In this model it is possible to have, despite all controls used in the regression, some selection issues due to the presence of unobserved characteristics. I, therefore, propose and estimate a structural model of high school human capital acquisition and college major choice. This enables me not only to control for potential selection on unobserved variables but also to conduct counterfactual experiments to study the potential effects of various curriculum policies in high school.

3.4 Structural model of high school human capital choice

In the Model, individuals differ in both their innate abilities to learn and in their preferences for different college majors. The choice of subject in high school is based on these differences. I assume that they know their abilities to acquire imperfectly substitutable skills. They choose their high school courses to maximize their expected utility across majors. While they graduate the choose to enter a major or not enroling into college.

Initial ability drawn and college major targeted provide an incentive for individuals to specialize by acquiring skills that reflect their personal circumstances. In contrast, the risk of low utility draws in each college major provides an incentive to acquire a more widely applicable portfolio of human capital skills.

I suppose that individuals with discount factor $\beta \in (0,1)$ live for a finite number of discrete periods, $t = 0,1,2,...T$. Individuals choose their human capital investments, i.e. their set of individually distinct courses to attaint, in the initial period $(t = 0)$ to optimize expected discounted utility.

There are three types of skills which can be employed all majors. High school skills

are useful to college but their importance differs from one major to another.[8] Denote an individual's portfolio of human capital by $s = (s_Q, s_H, s_{NS})$, where s_Q is quantitative human capital and s_H is humanities human capital and s_{NS} is natural sciences human capital.[9] Individuals can choose their portfolio composition by investment through more HS courses in a particular major (or more homework or tutoring this share of investment is not observed).

Before choosing s, individuals draw abilities $\tau = (\tau_Q, \tau_H, \tau_{NS})$ from distribution $H(\tau)$, where τ_{NS} represents the ability to accumulate natural science human capital.

The cost of accumulating s with ability τ is $c^h(s, \tau)$ with c^h convex and twice differentiable. Individual knows how different human capital are used in different majors. But they are unsure about an idiosyncratic component of college preference.

Once an individual has acquired the skill set s, they enter the college in the next period $t = 1$. Individuals can choose in which major to study. Although individuals have a general idea before they invest in their portfolio of skills of how well they are likely to fit into a given major, it is only after they complete HS and enter college that their true fit in that major is known. Actual experience in a major reveals an individual's true preference for that major[10]. They may also choose not to enrol.

The timing of the model is the following:

- **In period 1**: individuals draw abilities τ from distribution $H(\tau)$, and after that choose the amount of each human capital to invest in.
- **In period 2**: individuals choose a major and receive new information about their abilities and preference in a particular major and accumulate human capital (GPA).

8. In the descriptive statistics I assume seven field of skills I reduce the number here to three for computational reason and also to focus on the most important skills.

9. Quantitative human capital is measured by the number of HS course in Maths and Physics. Humanities human capital is measured by HS courses in reading and writing, humanities, business and communication. Natural sciences human capital is the number of HS life science courses.

10. I assume that students make a one-time decision about their college major. In fact, students may change their majors over the course of their college careers. About 55% to 60% college students change their major at least once according to the NELA http://www.nela.net/Centers/pages/collegemajors.aspx. But I am not able to investigate this issue because of data limitation. Second, I ignore the possibility that students may continue their education by seeking post-baccalaureate degrees our drooping out.

3.4.1 High school and college stages

Due to data limitation I was not able to use wage in the model. Given that the grade point averages has a positive effect on future earnings (see Arcidiacono (2004)). I will therefore use it as proxy of future wage. I assume that college grade (G) is a function of the individual abilities, as well as X_G which represent other demographic characteristics such as gender, Socioeconomic status (SES). Specifically, performance in college takes the following form:

$$G = \eta_0 + \eta_1 \rho + \eta_2 \rho^2 + \eta_3' s + \eta_4' X_G + \varepsilon_m + \varepsilon_1$$

The major specific fixed effect is ε_m. The idiosyncratic shocks (the ε_1's) are assumed to be distributed $\mathcal{N}(0; \sigma_G^2)$.

The utility of choosing the college major m is given by

$$u_m^c = \vartheta_{0c}' s + \vartheta_{1c}' X_{cm} - c_m(s, G) + v_m + \varepsilon_m$$

ε_m is generalized extreme value (GEV) distribution. The fixed intercept (v_m) represents the combined effect of all omitted major-specific covariates that causes some students to be more prone to a particular major. Utility of being in high school is given by

$$u^h = -c^h(\tau, s) + \varepsilon$$

where ε has normal distribution. High school and college cost functions are respectively $c_m(s, G)$ and $c^h(\tau, s)$.

I assume that marginal cost for human capital k is:

$$Mc_k^h(\tau, s) = \vartheta_{4hk} s_k + \vartheta_{5hk} \tau$$

$$Mc_{mk}(s, G) = \vartheta_{4mk} + \vartheta_{5mk} G$$

where Mc_k^h is the marginal of acquiring skill k in high school, Mc_{mk} is the marginal of acquiring skill k in major m and ϑ_{4mk} and ϑ_{5mk} are cost elasticity contribution of producing grade in major m of human capital type k and $\vartheta_{.mk}$ is observed with error; that is why I have major fixed effect v_m. Integrating on different dimension of human capital will give the effort cost function. This cost of effort may imply that even if an individual was allowed to attend all majors, the individual may not choose to attend the highest paying major because of the effort required. Individuals also have the option not to attend college, with the utility given by u_o where the o subscript indicates that the individual chose the outside option.

Individuals then choose the major with the highest u_m^c ie which yields the highest utility. I assume that ε_m follows a generalized extreme value distribution. Special cases of the generalized extreme value distribution lead to multinomial logit and nested logit models. I use a nested logit model; this generalized extreme value distribution allows for errors to be correlated across multiple nests while still being consistent with random utility maximization consistent with McFadden (1978) framework. [11]

I assume that majors are grouped in four nests:

-Nest 1 Quantitative major (Maths, Physics and engineering)

-Nest 2 Business & Communication Humanities major and Education and military

-Nest 3 Health and Natural science major

-Nest 4 No college

11. McFadden (1978) framework is as follows. Let $r = 1...R$ index all possible choices. Define a function $G(y_1,...,y_R)$ on y_r for all r. If G is nonnegative, homogeneous of degree 1, approaches $+\infty$ as one of its arguments approaches $+\infty$, has nonnegative n^{th} cross-partial derivatives for odd n, and nonpositive cross-partial derivatives for even n, then McFadden (1978) showed that

$$F(\varepsilon_1,...,\varepsilon_R) = exp\{-G(e^{-\varepsilon_1},...,e^{-\varepsilon_R})\}$$

is the cumulative distribution function for a multivariate extreme value distribution. Furthermore, the probability of choosing the r^{th} alternative conditional on the observed characteristics of the individual is given by

$$P(r) = \frac{y_r G_r(y_1,...,y_R)}{G(y_1,...,y_R)}$$

where G_r is the partial derivative of G with respect to the r^{th} argument. This is the same in Arcidiacono (2005)

Let $u_m^{c'}$ be the net present value of indirect utility for attending major m.

$$F\left(e^{u^{c'}}\right) = \sum_m \left(\sum_N exp\left(\frac{u_{mn}^{c'}}{\eta}\right)\right)^{\eta} + exp\,(u_o)$$

The error terms are known to the individual, but they are not observed by the econometrician. Therefore, from the econometrician's perspective, the probability of choosing the major m then given by

$$Pr(m) = \frac{exp\left(\frac{u_{mn}^{c'}}{\eta}\right)\left(\sum_N exp\left(\frac{u_{mn}^{c'}}{\eta}\right)\right)^{\eta-1}}{F\left(e^{u^{c'}}\right)}$$

Before choosing major individuals first choose their HS human capital. The net utility of outside option, which is not going into college is normalized to zero.

3.4.2 Choice of high school human capital

After making the major choice decision, there are no decisions left. Let u_1^c indicates the best option in the college. Individuals need now to choose how much of different human capital to accumulate in high school. They choose the s which yields the highest utility $V_0(s,\tau)$ where $V_0(s,\tau)$ is given by:

$$V_0(s,\tau) = u^h + \beta E_0(u_1^c|\tau)$$

For each type of human capital, s_k^* is the optimal value of s_k that solves the Euler equation

$$Mc_k = \beta_1 E_0(u_{1k}^c|\tau).$$

If I apply envelope theorem on u_1^c; I get $E_0(u_{1m}^c|\tau) = \beta\,\vartheta_{0ck} - \beta E_0(MC_k(s,G)))$

$$\vartheta_{4hk}s_k + \vartheta_{5hk}\tau_k = \beta\,\vartheta_{0ck} - \beta\,(MC_k(s,G))$$

thus

$$s_k^* = \theta_{0\hat{m}k} + \theta_{1\hat{m}k}\tau_k + \theta_{2\hat{m}k}G$$

Let $\tilde{s}_j^*$ be latent variable with

$$\tilde{s}_k^* = s_k^* + \varepsilon_k = \theta_{0\hat{m}k} + \theta_{1k}\tau + \theta_{2\hat{m}k}G + \varepsilon_k$$

with ε_k normal forecast error.

The observed chosen level of course s_k is

$$s_k = \begin{cases} \tilde{s}_k^* & if \quad \tilde{s}_k^* > C \\ 0 & if \quad s_k^* \leq C \end{cases}$$

The forecast error, ε_k, is independent of τ, G and m. I estimate the coefficients of the model with a Tobit model.

3.4.3 Identification and estimation strategy

In this section, I discuss how several key parameters of the model are identified.

3.4.3.1 Identification without unobservables

All characteristics of the individuals are taken as exogenous, including such things as test scores and GPA in college, high school courses and 10^{th} grade standardized test score. One of the main advantage of HS&B data is that for all individual in the sample there are base year test score in different subject. These score are in math, science, civic, reading and writing and are assume to our main exogenous variables. I assume no correlations across the various stages of the model, selection into majors is then controlled for by these exogenous characteristics.

3.4.3.2 Identification with unobservables

Assuming that preference parameters are uncorrelated over time is particularly unreasonable. That is, if one has a strong preference for high school initially, he is just as likely as someone who has a small preference for high school to choose any major in college. I would suspect that this is not the case. Furthermore, it is unreasonable to assume that there is no unobserved (to the econometrician) ability that is known to the individual. Some variables can be used to identify types: initial ability (here measure by base year standardized test scores), relation in the level of human capital and college major choice.

3.4.3.3 Estimation method

I first estimate a model with independent errors across grades and choice processes. The log likelihood function is the sum of three pieces:

- $L_1(\eta)$ the log likelihood contribution of grade point averages,
- $L_2(\vartheta_c, \eta)$ the log likelihood contribution of major decisions,
- $L_3(\vartheta_h, \vartheta_c, \eta)$ the log likelihood contribution of high school human capital decisions,

The total log-likelihood function is then $L = L_1 + L_2 + L_3$.

Consistent estimates of η can be found by maximizing L_1 separately. Then η are replaced by consistent estimates in L_2 and consistent estimate of ϑ_c can be obtained by maximizing L_2. I estimate ϑ_h using L_3 and all other estimates.

Following Arcidiacono (2004, 2005) I assume that there are $R = 2$ types of people. To account for selection on unobservables into majors, I use a mixture distribution that allows errors to be correlated across the various stages.

Types remain the same throughout all stages, individuals know their type. Preferences and abilities may vary across types.

The log likelihood function for a data set with N observations is then given by

$$L(\eta, \vartheta) = \sum_{i=1}^{N} ln(\sum_{r=1}^{R} \pi_r \mathscr{L}_{ir1} \mathscr{L}_{ir2} \mathscr{L}_{ir3})$$

π_r is the proportion of type r in the data and $\mathscr{L}_{ir.}$ refers to the likelihood (as opposed to the log likelihood L). The log likelihood function is no longer additively separable. I use the Expectation-Maximization (EM) algorithm to solve the problem. The EM algorithm has two steps:

- **first** calculate the expected log likelihood function given the conditional probabilities at the current parameter estimates,
- **second** maximize the expected likelihood function holding the conditional probabilities fixed.

These steps are repeated until convergence . The expected log likelihood function is:

$$L(\eta, \vartheta) = \sum_{i=1}^{N} \sum_{r=1}^{R} P_i(r|X_i, \alpha, \eta, \vartheta)[L_{ir1}(\eta) + L_{ir2}(\eta, \vartheta_c) + L_{ir3}(\eta, \vartheta_{c,h})]$$

with $P_i(r|X_i, \eta, \vartheta) = \dfrac{\pi_r \mathscr{L}_{ir1} \mathscr{L}_{ir2} \mathscr{L}_{ir3}}{\sum_{r=1}^{R} \pi_r \mathscr{L}_{ir1} \mathscr{L}_{ir2} \mathscr{L}_{ir3}}$

Using the EM algorithm helps us recover additivity of log likelihood. And parameters can be estimated at each step as in the case without unobservable heterogeneity. Note that all pieces of the likelihood are still linked through the conditional probabilities where the conditional probabilities are updated at each iteration of the EM algorithm.

3.5 Structural model estimations results

This section presents and discusses the results from estimating the parameters of the performance equations, the structural parameters of the utility function and high school courses choice equations. Results of the model with unobserved heterogeneity are presented in the estimation of each equation separately.

3.5.1 College performance regressions

Estimates of the college period performance equation are given in Table 3.V. The first column displays the coefficient estimates without unobserved heterogeneity, while the second presents estimates with unobserved heterogeneity approximated by two types.

There is a U-shape relationship between college performance and diversification in high school. The size of the coefficients are the same with or without unobserved heterogeneity. Females receive higher grades than their males counterparts. All of the ability coefficients are positive, with smaller coefficients for verbal. Without unobserved heterogeneity, math ability is particularly useful. Once the mixture distribution is added, the differences in ability coefficients dissipate. The results with unobserved heterogeneity show that type 2's receive substantially higher grades.

3.5.2 Estimate of the utility function parameters

I use the estimates of the performance to obtain the second stage maximum likelihood estimates of the utility function parameters. Table 3.VII displays the maximum likelihood estimates for the parameters of the utility function.

The first three sets of rows of Table 3.VII display the differences in preferences across high school courses individuals have for each of the fields. More quantitative courses are more attractive for natural science, maths & physics and engineering, while more humanities courses are better for social science & humanities majors, business & communication. Controlling for unobserved heterogeneity did not change these results.

Diversification effect size differ with the major. Having large diversification index is better for business & communication than for maths & physics majors. This effect is the same with and without unobserved heterogeneity. Diversification has lager negative effect in quantitative majors (maths & physics and Engineering). Females are more likely to be in education or health and less likely in quantitative majors. Taking unobserved heterogeneity into account does not change the results.

Types 1's are more likely to be in quantitative major in model with mixture. Ability measure (SATM, SATV), *GPA* and $GPA \times HScourses$ interacts with major, along with major specific constant terms, also were included. Consistent with Arcidiacono (2004), I also find that students' comparative advantage in their abilities in different majors plays a very important role in the choice of a major.

The nesting parameters are both relatively small for all models. The estimates that are less than one suggest that the preferences for major options are correlated. Indeed, this nesting parameter measures the cross-school component of the variance. In particular, had these coefficients been estimated to be one, then a multinomial logit would have resulted.

3.5.3 Courses choice equations regressions

Estimates of courses equations Tobit model are in Table 3.VIII, 3.IX and 3.X. As in performance results, adding controls for unobserved heterogeneity did not significantly affect the other parameter estimates. Those who have high math and science scores in the 10^{th} grade standardized test have higher probabilities to accumulate more skills in quantitative and life science subject. Those with high score in Civic and writing are more likely to accumulate humanities skills. Types 1's [12] are more likely take more life science and quantitative courses than humanities in high school.

3.5.4 Model Fit

Table 3.XI displays the actual data and the predictions of the model. I use parameters to see how the model matches some key features and trend of the data. For example the number of quantitative courses choose in HS, in the data, is very close to what is predicted by the model. The models with and without the mixture distribution predict the trends in the data very well. The models often hit the observed mean almost exactly. The prediction with mixture model are better than those without.

12. The proportion of type 1 individual is 33.16% of the population.

3.5.5 Simulations

Since the model matches the data reasonably well, I can use the model to simulate how the major choice would vary given a different environment. The purpose of the simulation is to compare policies which aim to increase enrollments in Science Technology Engineering, and Math (STEM) majors. The first policy is an increase in high school Maths course requirements (which imply more specialization in math and science). The second experiment is an increase in high school humanities course requirements while the third simulation increases high school life science course. The last simulation assume that there is no specialization in high school.

Increase of enrollments in STEM majors is an issue of considerable interest in many countries, given that the economy is increasingly driven by complex knowledge and advanced cognitive skills. Thus, STEM is one of the key components to keeping competitive in a global economy. The U.S. President's Council of Advisors on Science and Technology, in a 2012 report, suggested that the number of STEM majors needed to increase significantly to meet the demand for STEM professionals. In UK, a lack of workers in scientific occupations is a recurrent issue (Chevalier (2012)). The shortage of STEM majors occurs despite STEM majors earning substantially more than other college degrees with the exception of perhaps business (Arcidiacono (2004), Pavan and Kinsler (2012), Arcidiacono et al. (2013)).

The first to third simulations assumes respectively that one more quantitative course, one more humanities course and one more life science course, in high school is required. These simulations are designed to answer the question: how much STEM major choice is due to high school courses choice?

The last simulation eliminates specialization in HS. The results of the simulation will then show how much specialization in HS affect STEM major enrollment.

Note that these simulations are not taking into account general equilibrium effects; the simulations are only designed to illustrate how much of the current major choice is

due to high school courses or specialization.

Table 3.7.1 shows that quantitative course and specialization affect STEM majors choice. When there is one more high school quantitative courses, the share of people in STEM (maths & physics and engineering) and natural science increases (see Simulation 1). One more high school quantitative course increases enrollment in STEM by 4 point percentage. But in the same time we have a reduction in overall enrolment. It is interesting to note that when I use the model without unobserved heterogeneity one more high school quantitative course increases enrollment in STEM by 5 points percentage. When the model with unobserved heterogeneity the enrolment in STEM increases only by 4 points, this suggest a correction of the unobserved ability bias. Increase by one of high school humanities course did not heavily decrease enrollment in STEM. One more course of life science increases enrollment in Natural science major by 0.015 point percentage and reduces enrollment in STEM majors by the same amount.

Force every student to have the portfolio (see Simulation 4) also boosts enrollment. The share of students choosing STEM major moves up by 18 to 20 points percentage. This suggests that high school specialization play a key role in major choice.

These results suggest that increasing high school quantitative course requirement will affect enrollment more in STEM major. Moreover, uniform curriculum in high school can also largely increase enrollment in STEM. However, this policy is less realistic. Increasing high school quantitative courses requirement is, therefore, the most appealing policy to reduce the preparation gap in Maths.

3.6 Conclusion

This paper investigates how specialization and diversification in high school influence future college choice and performance. I establish panel data evidence linking individual's high school skill sets with college major choice. I find that students usu-

ally choose major in which they acquired more skills related to this major, suggesting specialization in HS. However, I find a U-shaped relation between diversification and college performance. This result suggests a trade-off between specialization and diversification. This trade-off is assessed through a model of human capital acquisition and college major choice with different skills, abilities and uncertainty regarding college. Estimation of structural parameters of the model suggests that quantitative majors are preferred by specialized students. I also find that high school courses also play an important role in determining college major choice. More quantitative high school courses makes more attractive natural science, maths & physics and engineering, while more humanities courses are better for social science, humanities , business & communication majors. Moreover, the estimated model remarkably match some central tendencies in the data. I then exploited this model to evaluate and quantify the impact of economic policies on enrollment in STEM majors. Policy experiments suggest that increasing high school quantitative courses requirement, by one, will boost STEM enrollment by 4 to 5 percentage points.

In this paper, I restrict my attention to the role played by high specialization on college major choice and performance. Possible future research could be to investigate the effect of high school specialization on labor market outcomes (unemployment, income). It will be also interesting to compare system with forced specialization (European style) and system with chosen specialization (USA style).

3.7 Appendix

3.7.1 Data

Merging the PETS, Sophomores in 1980 - HS&B and High school transcript data sets yields an initial sample of 5,533 students. Dropping those who do not have SAT data reduced to 1185 individuals for regression. Taking into account others controls variables reduced the sample to 1112.

To find portfolios, human capital is partitioned into seven broad areas of knowledge using the Classification of Secondary School Courses (CSSC) in US. Each of these areas is the sum of course taken in areas of study belonging to that area of knowledge [13].

- Quantitative: mathematics and physics: 27, 11, 41, 40, 15, 14, 04
- Reading and writing: 23, 16
- Social science and Humanities: 45, 44, 43, 42, 39, 38, 37, 24, 19, 13, 05
- Natural life science: 17, 18, 26, 34, 02
- Business and communication: 22, 10, 09, 08, 07, 06, 01
- Art: 50, 21
- Others: 55, 51, 49, 48, 47, 46, 29, 28, 20, 03, 56, 54, 36, 35, 33, 32, 31, 30, 25, 12

I also aggregate majors into seven categories: Math and Physics, Engineering, Business and Communication, Social Science & Humanities, Natural Science, Education, Health. The criteria for aggregation was the degree of similarity in field topics.

- Math and Physics:Physics, Science Technologies, Mathematics, Calculus, Communication Technologies, Computer & Information Sciences Computer Programming.
- Engineering: Engineering, Civil Engineering, Electric & Communications Engineering, Mechanical Engineering, Architecture & Environmental Design.
- Business and Communication: Construction Trades,Business & Management, Accounting, Banking & Finance, Business & Office, Secretarial & Related Programs, Marketing & Distribution, Communications, Journalism,Precision Production, Transportation & Material Moving,.
- Natural life sciences: Geology, Life Sciences, Geography, Renewable Natural Resources,.
- Social Science & Humanities: Area & Ethnic Studies, Foreign Languages, German, French, Spanish, Home Economics, Vocational Home Economics, Law, Letters, Composition, American Literature, English Literature, Philosophy & Reli-

13. Number for each field are CSSC codes

gion, Theology, Psychology, Protective Services, Public Affairs, Social Work, Social Sciences, Anthropology, Economics, Geography, History, Political Science & Government, Sociology, Visual & Performing Arts, Dance, Fine Arts, Music, Liberal/General Studies.

- Education: Education, Adult & Continuing Education, Elementary Education, Junior High Education, Pre-Elementary Education, Secondary Education.
- Health: Allied Health, Practical Nursing, Health Sciences, Nursing.

Table 3.I: Summary Statistics

	Unrestricted sample					Restricted sample				
	Mean	SD	SD in HS	Frac in HS	Obs.	Mean	SD	SD in HS	Frac. in HS	Obs.
Female	0.541	0.498	0.211	0.821	5072	0.530	0.499	0.266	0.716	1265
Black	0.125	0.331	0.210	0.597	5072	0.089	0.284	0.151	0.717	1265
SATM	477.747	115.141	47.416	0.830	2064	483.075	110.838	43.172	0.848	1265
SATV	440.612	107.143	43.714	0.834	2042	447.249	103.189	40.650	0.845	1265
College GPA	2.316	0.803	0.211	0.931	4686	2.453	0.689	0.190	0.924	1265
SES	0.224	0.738	0.412	0.688	4912	0.403	0.687	0.398	0.664	1265
Share of Course										
Reading and writing	0.232	0.066	0.045	0.530	5072	0.246	0.062	0.046	0.455	1265
Maths	0.122	0.041	0.023	0.686	5072	0.132	0.037	0.022	0.640	1265
Life science	0.168	0.067	0.055	0.325	5072	0.168	0.066	0.059	0.221	1265
Physics	0.054	0.041	0.021	0.746	5072	0.065	0.041	0.022	0.718	1265
Humanities	0.186	0.074	0.065	0.235	5072	0.200	0.079	0.071	0.195	1265
Business and communication	0.074	0.065	0.031	0.768	5072	0.060	0.054	0.027	0.744	1265
Art	0.070	0.071	0.036	0.747	5072	0.059	0.066	0.035	0.709	1265
Others	0.050	0.058	0.044	0.422	5072	0.042	0.054	0.047	0.260	1265

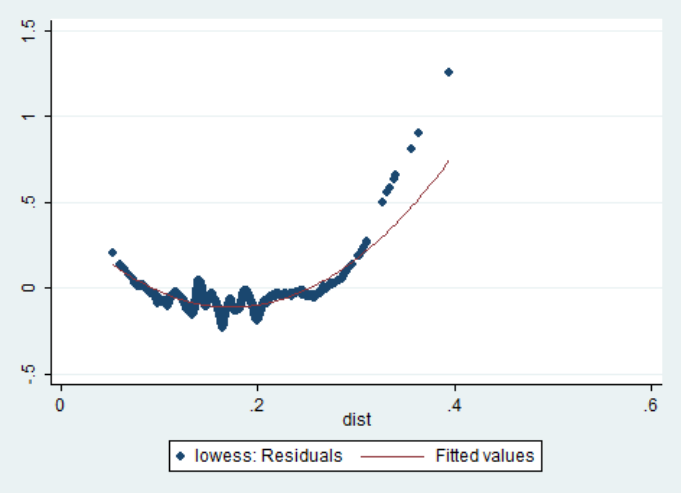

Figure 3.1: U-shaped between residual and ρ by college major

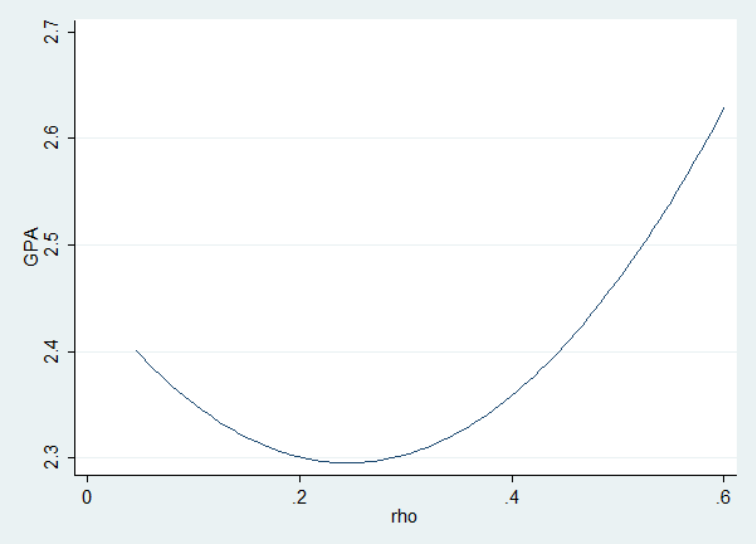

Figure 3.2: U-shaped between GPA and ρ

Table 3.II: High school Human Capital Portfolios by college major

College Major \ Share	Quant.	R. and W.	Life sc.	Hum.	Com./Bus.	Arts	Others
Bus. & Com.	0,169	0,236	0,167	0,190	**0,095**	0,063	0,081
Natural science	0,219	0,251	0,186	0,178	0,040	0,065	0,061
Math and Physical	**0,225**	0,245	0,167	0,187	0,056	0,056	0,064
Education	0,165	0,232	0,169	0,180	0,075	**0,092**	0,088
Engineer	**0,227**	0,227	0,171	0,172	0,050	0,063	0,089
Social/hum.	0,185	**0,258**	0,163	**0,198**	0,056	0,066	0,074
Health	0,172	0,232	0,181	0,188	0,075	0,073	0,078
Others	0,169	0,228	0,170	0,176	0,066	**0,093**	0,098
F	50,218	12,651	3,385	5,174	37,233	9,784	6,410
P-value	,000	,000	,001	,000	,000	,000	0,000

Table 3.III: Estimation results of college performance: GPA as the dependent variable

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ρ	-6.834***	-5.222***	-5.824***	-5.766***	-5.861***	-5.923***	-7.154***
	(2.08)	(1.98)	(1.90)	(1.90)	(1.89)	(1.96)	(2.59)
ρ^2	17.477***	14.707***	16.128***	15.985***	16.195***	16.740***	21.376***
	(5.49)	(5.25)	(5.05)	(5.05)	(5.03)	(5.34)	(6.87)
Female		0.127***	0.179***	0.179***	0.178***	0.164***	0.072
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
Black		-0.281***	-0.141**	-0.136**	-0.141**	-0.114*	-0.047
		(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.09)
SES		0.039	-0.057**	-0.050	-0.054	-0.045	-0.051
		(0.03)	(0.03)	(0.05)	(0.05)	(0.05)	(0.06)
SAT Maths			0.109***	0.109***	0.111***	0.126***	0.096***
			(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
SAT verbal			0.127***	0.126***	0.126***	0.126***	0.144***
			(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Father education				0.002	0.002	-0.000	-0.005
				(0.01)	(0.01)	(0.01)	(0.01)
Mother education				-0.006	-0.005	-0.006	0.007
				(0.01)	(0.01)	(0.01)	(0.01)
Plan college Dad					0.034	0.034	0.093
					(0.08)	(0.08)	(0.08)
Plan college Mom					-0.037	-0.036	0.009
					(0.08)	(0.08)	(0.09)
Dummy for south					0.047	0.071	
					(0.04)	(0.04)	
Plan HS Dad						-0.018	-0.001
						(0.07)	(0.07)
Plan HS Mom						-0.134	-0.096
						(0.10)	(0.10)
Majors		Yes	Yes	Yes	Yes	Yes	Yes
High school courses						Yes	Yes
Constant	3.067***	2.591***	1.604***	1.612***	1.607***	1.646***	1.864***
	(0.19)	(0.18)	(0.20)	(0.21)	(0.22)	(0.24)	(0.37)
Observations	1265	1265	1265	1265	1265	1265	1265
R2	0.01	0.11	0.20	0.20	0.20	0.21	0.19
Number of groups							389
R2 overall							0.16

NB: *** Significant at 1%; ** Significant at 5%; and * Significant at 10% Heteroskedasticity robust standard errors are clustered by high school in parentheses for column 1 to 6. Column 7 estimates OLS with high school fixed effect. Background characteristics include parents education, parents participation to college enrollment decision, High school courses are formal courses taken in HS form HS transcripts.

Table 3.IV: Lind and Mehlum (2010) test for U-shape

Specification: $f(x) = x^2$		
Extreme Point: 0.1699134		
	H1: U shape vs H0 Monotone or inverse Ushape	
	Lower bound	Upper bound
Interval	.04625	.6006787
Slope	-4.066305	14.16445
t-value	-2.559242	3.20296
P>|t|	.0054755	.0007488
Overall test of presence of a U shape:		
	t-value=2.56	
	P>|t|=0.0054	

Table 3.V: Performance regressions

	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
ρ	-7.6077	2.2671	-8.3000	2.0323
ρ^2	22.2215	6.0855	27.6217	5.4638
Female	0.1563	0.0483	0.1418	0.0433
Dummy South	-0.0105	0.0159	0.0014	0.0143
SATM	0.1192	0.0279	0.1110	0.0250
SATV	0.0859	0.0283	0.0929	0.0254
SES	-0.0326	0.0333	-0.0688	0.0300
Black	-0.0814	0.0748	-0.0604	0.0671
Business and Communication	0.1205	0.1111	0.1174	0.0995
Natural science	-0.0906	0.1334	-0.1240	0.1196
Maths & Physics	0.1316	0.1088	0.1569	0.0975
Education and Military	0.0155	0.0956	-0.0269	0.0857
Engineering	0.0962	0.0703	0.0518	0.0631
Health	-0.0176	0.0114	-0.0115	0.0102
Type 1			-0.7128	0.0430
const	2.1672	0.3110	1.9749	0.2790
Variance	0.7225	0.0153	0.6476	0.0137

NB:major-specific constant terms were also included

Table 3.VI: Utility parameters estimates

		One type		Two types	
		Coefficient	Stand. Error	Coefficient	Stand. Error
Life science courses					
	Business and Communication	0.1041	0.0979	0.1204	0.0665
	Natural science	0.2068	0.1301	0.1694	0.0996
	Maths & Physics	0.0511	0.1247	-0.0242	0.0813
	Education and Military	0.1038	0.1316	0.1256	0.0705
	Engineering	0.0658	0.1217	-0.0147	0.0812
	Humanities	0.0648	0.0994	0.0764	0.0687
	Health	0.1603	0.1199	0.1600	0.0917
Quantitative courses					
	Business and Communication	4.8964	1.0766	-0.0443	0.0879
	Natural science	5.0620	1.0837	0.0421	0.1277
	Maths & Physics	5.2565	1.0777	0.2146	0.1057
	Education and Military	4.8938	1.0670	-0.0359	0.0945
	Engineering	5.2959	1.0805	0.2558	0.1058
	Humanities	4.8925	1.0742	-0.0770	0.0907
	Health	5.0410	1.0862	0.0662	0.1180
Humanities courses					
	Business and Communication	0.1223	0.0763	0.1031	0.0369
	Natural science	0.0548	0.1064	-0.0086	0.0717
	Maths & Physics	0.0597	0.1026	-0.0464	0.0499
	Education and Military	0.0657	0.1123	0.0764	0.0417
	Engineering	0.0538	0.0984	-0.0413	0.0496
	Humanities	0.0793	0.0777	0.0600	0.0391
	Health	0.0385	0.0965	0.0118	0.0613
GPA					
	Business and Communication	1.5436	0.9165	0.9756	0.3705
	Natural science	0.8338	1.1941	-0.4789	0.8508
	Maths & Physics	2.0673	1.1337	0.7189	0.4828
	Education and Military	1.7819	1.2066	1.1058	0.3971
	Engineering	2.0723	1.1096	0.7165	0.4878
	Humanities	1.1457	0.9292	0.5134	0.4223
	Health	0.9436	1.1081	0.1823	0.6781
ρ					
	Business and Communication	-12.9912	2.6876	-4.2077	2.8318
	Natural science	-17.4797	3.7017	-4.6484	3.5321
	Maths & Physics	-21.0977	3.3094	-14.1069	3.3499
	Education and Military	-14.8482	3.1440	-4.5911	2.9716
	Engineering	-20.5266	3.2969	-13.5531	3.3480
	Humanities	-13.4522	2.7094	-4.4163	2.8372
	Health	-11.8757	3.3616	-1.4088	3.4043

Table 3.VII: Utility parameters estimates (cont.)

		One type		Two types	
		Coefficient	Stand. Error	Coefficient	Stand. Error
GPA × HS courses					
	Business and Communication	-0.0279	0.0353	-0.0147	0.0147
	Natural science	-0.0145	0.0451	0.0274	0.0328
	Maths & Physics	-0.0487	0.0436	-0.0041	0.0192
	Education and Military	-0.0305	0.0472	-0.0162	0.0163
	Engineering	-0.0561	0.0427	-0.0097	0.0194
	Humanities	-0.0162	0.0359	0.0018	0.0169
	Health	-0.0100	0.0421	0.0132	0.0273
SATM					
	Business and Communication	0.2981	0.1769	0.2984	0.1721
	Natural science	0.5880	0.2189	0.5464	0.2084
	Maths & Physics	0.7487	0.1986	0.7520	0.1914
	Education and Military	0.1307	0.2003	0.2090	0.1798
	Engineering	0.6788	0.1981	0.7199	0.1916
	Humanities	0.1598	0.1771	0.2223	0.1719
	Health	0.1296	0.2111	0.2267	0.2015
SATV					
	Business and Communication	0.4005	0.1723	0.4078	0.1660
	Natural science	0.5737	0.2123	0.4268	0.1988
	Maths & Physics	0.1994	0.1945	0.1636	0.1848
	Education and Military	0.3943	0.1949	0.4086	0.1734
	Engineering	0.2815	0.1937	0.1999	0.1845
	Humanities	0.6591	0.1729	0.5390	0.1659
	Health	0.3215	0.2049	0.3063	0.1963
Constant					
	Business and Communication	-4.9800	1.1411	-3.0227	1.2116
	Natural science	-2.7318	2.1281	-4.1982	2.1996
	Maths & Physics	-5.8052	3.0440	-2.3435	1.5126
	Education and Military	-4.8387	2.8206	-3.3495	1.3289
	Engineering	-2.7981	3.0284	-2.2728	1.5035
	Humanities	-4.4636	2.6980	-2.0683	1.2663
	Health	-2.1956	2.1706	-4.0940	1.9494
Type 1					
	Business and Communication			-0.4613	0.3148
	Natural science			-0.8189	0.3875
	Maths & Physics			-0.2081	0.3527
	Education and Military			-0.4716	0.3296
	Engineering			-0.3901	0.3543
	Humanities			-0.4986	0.3154
	Health			-0.2955	0.3712
Nesting Parameter		0.4834	0.0111	0.2653	0.0088

Table 3.VIII: High school courses choices estimations

	Humanities courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	0.0610	0.0172	0.0568	0.0175
Reading	0.0104	0.0166	-0.0005	0.0168
Math	0.0031	0.0194	-0.0377	0.0193
Science	-0.0357	0.0171	-0.0402	0.0173
Writing	0.0587	0.0184	0.0410	0.0184
Civic	0.0137	0.0142	0.0193	0.0143
Expected GPA	2.8965	0.5722	4.4575	0.4860
Expected GPA interacted by major:				
Business and Communication	-6.9553	0.8609	-4.3865	0.6309
Natural science	-13.1673	2.0744	-8.5054	1.5168
Maths & Physics	-9.1344	2.4134	-3.0334	0.9578
Education and Military	-4.6856	1.4340	-5.5758	1.4804
Engineering	-6.4705	1.3250	-5.0364	0.9241
Humanities	-6.9276	0.9197	-5.7341	0.6865
Health	-8.4818	1.9966	-4.2283	1.0902
Major:				
Business and Communication	17.7216	2.1358	11.1098	1.5040
Natural science	32.4560	5.1415	21.3923	3.7467
Maths & Physics	20.9101	5.7522	6.6542	2.5020
Education and Military	11.1049	3.7968	12.6932	3.5860
Engineering	14.7227	3.2620	11.5303	2.2438
Humanities	17.5508	2.2610	14.6686	1.6314
Health	20.3698	4.7245	10.3561	2.5678
Type 1			0.6906	0.3993
Variance	3.3723	0.0712	3.4201	0.0723

Table 3.IX: High school courses choices estimations

	Life science courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	-0.0311	0.0107	-0.0318	0.0108
Reading	-0.0124	0.0103	-0.0077	0.0104
Math	-0.0337	0.0121	-0.0247	0.0119
Science	0.0142	0.0107	0.0166	0.0107
Writing	-0.0223	0.0115	-0.0179	0.0114
Civic	0.0119	0.0089	0.0160	0.0088
Expected GPA	3.7839	0.3567	2.8776	0.2998
Expected GPA interacted by major:				
Business and Communication	-0.3033	0.5366	-0.2213	0.3892
Natural science	0.6705	1.2927	-1.0865	0.9356
Maths & Physics	3.0931	1.5043	0.1327	0.5908
Education and Military	-1.3200	0.8938	1.3837	0.9131
Engineering	-1.2856	0.8259	-0.4680	0.5700
Humanities	-0.9882	0.5733	-1.1866	0.4235
Health	-0.2333	1.2446	-2.3845	0.6725
Major:				
Business and Communication	0.4262	1.3312	0.7125	0.9277
Natural science	-0.5294	3.2042	4.1990	2.3111
Maths & Physics	-7.0656	3.5854	-0.6851	1.5433
Education and Military	2.7386	2.3666	-2.7487	2.2120
Engineering	3.1075	2.0332	1.4399	1.3841
Humanities	2.2836	1.4094	3.2095	1.0063
Health	1.2013	2.9451	6.5287	1.5839
Type 1			1.4862	0.2463
Variance	2.1020	0.0445	2.1096	0.0447

Table 3.X: High school courses choices estimations

	Quantitative courses			
	One type		Two types	
	Coefficient	Stand. Error	Coefficient	Stand. Error
Base year test score				
Vocabulary	-0.0055	0.0084	-0.0031	0.0083
Reading	0.0112	0.0081	0.0068	0.0080
Math	0.0531	0.0095	0.0451	0.0092
Science	0.0289	0.0084	0.0263	0.0082
Writing	-0.0162	0.0090	-0.0208	0.0087
Civic	-0.0046	0.0070	-0.0051	0.0068
Expected GPA	0.8426	0.2795	0.9916	0.2301
Expected GPA interacted by major:				
Business and Communication	-0.1068	0.4205	0.8343	0.2987
Natural science	1.3458	1.0131	0.9615	0.7181
Maths & Physics	-2.3925	1.1788	-0.2132	0.4534
Education and Military	-1.0858	0.7004	-1.6626	0.7008
Engineering	1.2274	0.6471	0.6087	0.4375
Humanities	-0.0793	0.4492	0.6151	0.3250
Health	-0.3827	0.9752	1.0640	0.5161
Major:				
Business and Communication	-0.1636	1.0431	-2.3126	0.7120
Natural science	-2.9809	2.5111	-1.6040	1.7737
Maths & Physics	5.2358	2.8096	1.2428	1.1844
Education and Military	3.3402	1.8545	3.8587	1.6976
Engineering	-2.1985	1.5932	-0.3277	1.0622
Humanities	-0.0869	1.1043	-1.5347	0.7723
Health	0.9063	2.3076	-2.3652	1.2156
Type 1			1.1915	0.1890
Variance	1.6471	0.0349	1.6191	0.0343

Table 3.XI: Comparing model predictions of individual choices with the data

	HS Quantitative		
	Data	One type	Two types
Business and Communication	5.5044	5.5068	5.4890
Natural science	6.7018	6.7304	6.6787
Education and Military	6.7570	6.7757	6.7561
Maths & Physics	5.1628	5.1810	5.1848
Engineering	7.1130	7.1161	7.1001
Humanities	5.6754	5.6703	5.6627
Health	5.6582	5.6384	5.6624
	HS Humanities		
	Data	One type	Two types
Business and Communication	13.7609	13.9761	13.8079
Natural science	13.7368	13.5100	13.5951
Education and Military	12.8505	12.8022	12.7235
Maths & Physics	12.5349	12.6713	12.5517
Engineering	12.3826	12.4544	12.4572
Humanities	14.0623	13.8463	13.9706
Health	13.4177	13.3346	13.4069
	HS Life sciences		
	Data	One type	Two types
Business and Communication	5.0612	5.0585	5.0410
Natural science	5.9298	5.9030	6.0344
Education and Military	4.8879	4.9064	4.9184
Maths & Physics	5.0465	5.0528	5.0619
Engineering	4.8261	4.8231	4.7790
Humanities	4.9377	4.9399	4.9283
Health	5.5063	5.5292	5.4045

Table 3.XII: Simulations of the change in major choice distribution

		Simulations			
		(1)	(2)	(3)	(4)
One type	STEM Majors	0.036	-0.014	-0.015	0.186
	Natural Science	0.011	0.000	0.015	0.000
	Humanities	-0.05	0.015	0.002	-0.127
	No College	0.003	0.001	-0.002	-0.059
Two types	STEM Majors	0.027	-0.017	-0.021	0.231
	Natural Science	0.012	0.009	0.033	-0.039
	Humanities	-0.05	0.009	-0.01	-0.168
	No College	0.01	-0.001	0.0019	-0.023

BIBLIOGRAPHY

Allensworth, Elaine, Takako Nomi, Nicholas Montgomery, and Valerie E. Lee, "College Preparatory Curriculum for All: Academic Consequences of Requiring Algebra and English I for Ninth Graders in Chicago," *Educational Evaluation and Policy Analysis*, 2009, *31* (4), 367–391.

Altonji, Joseph G., "The Effects of High School Curriculum on Education and Labor Market Outcomes," *Journal of Human Resources*, 1995, *30* (3), 409–438.

_ , Erica Blom, and Costas Meghir, "Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers," *Annual Review of Economics*, 07 2012, *4* (1), 185–223.

Andersen, Torben G and Bent E Sorensen, "GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study," *Journal of Business & Economic Statistics*, July 1996, *14* (3), 328–52.

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

_ , Naoto Kunitomo, and Yukitoshi Matsushita, "On the asymptotic optimality of the LIML estimator with possibly many instruments," *Journal of Econometrics*, August 2010, *157* (2), 191–204.

Andrews, Donald W. K. and James H. Stock, Cambridge University Press, 2005.

Andrews, Donald W.K. and James H. Stock, "Inference with Weak Instruments," in R. Blundell, W. Newey, and T. Persson, eds., *Advances in Economics and Econometrics*, Vol. 3, Cambridge University Press, 2006.

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

Antoine, Bertille and Pascal Lavergne, "Conditional Moment Models under Semi-Strong Identification," Discussion Papers dp11-04, Department of Economics, Simon Fraser University December 2012.

Arcidiacono, Peter, "Ability sorting and the returns to college major," *Journal of Econometrics*, 2004, *121* (1-2), 343–375.

_ , "Affirmative Action in Higher Education: How Do Admission and Financial Aid Rules Affect Future Earnings?," *Econometrica*, 09 2005, *73* (5), 1477–1524.

_ **, Esteban Aucejo, and V. Joseph Hotz**, "University Differences in the Graduation of Minorities in STEM Fields: Evidence from California," IZA Discussion Papers 7227, Institute for the Study of Labor (IZA) February 2013.

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

_ **and** _ , "Instrumental Variable Estimation in a Data Rich Environment," *Econometric Theory*, 2010, *26*, 1577–1606.

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

Belloni, A, D Chen, V Chernozhukov, and C Hansen, "Sparse models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 2012, *80*, 2369–2429.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen, "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 2012, *80* (6), 2369–2429.

Campbell, John Y. and Luis M. Viceira, "Consumption And Portfolio Decisions When Expected Returns Are Time Varying," *The Quarterly Journal of Economics*, May 1999, *114* (2), 433–495.

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

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

_ **and Guy Tchuente**, "Regularized LIML for many instruments," Technical Report, University of Montreal 2012.

_ **and Jean-Pierre Florens**, "Generalization Of Gmm To A Continuum Of Moment Conditions," *Econometric Theory*, December 2000, *16* (06), 797–834.

_ **and** _ , "On the Asymptotic Efficiency of GMM," Econometric Society 2004 North American Winter Meetings 436, Econometric Society August 2004.

_ **and** _ , "ON THE ASYMPTOTIC EFFICIENCY OF GMM," *Econometric Theory*, 4 2014, *30*, 372–406.

_ **,** _ **, and Eric Renault**, "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, Vol. 6 of *Handbook of Econometrics*, Elsevier, January 2007, chapter 77.

_ **, Mikhail Chernov, Jean-Pierre Florens, and Eric Ghysels**, "Efficient estimation of general dynamic models with a continuum of moment conditions," *Journal of Econometrics*, October 2007, *140* (2), 529–573.

Castro, Rui, Gian Luca Clementi, and Glenn Macdonald, "Legal Institutions, Sectoral Heterogeneity, and Economic Development," *Review of Economic Studies*, 04 2009, *76* (2), 529–561.

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

Chao, John and Norman R. Swanson, "Alternative approximations of the bias and MSE of the IV estimator under weak identification with an application to bias correction," *Journal of Econometrics*, April 2007, *137* (2), 515–555.

Chao, John C. and Norman R. Swanson, "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 09 2005, *73* (5), 1673–1692.

_ , _ , Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen, "Asymptotic Distribution of JIVE in a Heteroskedastic Regression with Many Instruments," *Econometric Theory*, 2012, *28*, 42–86.

Chevalier, Arnaud, "To Be or Not to Be... a Scientist?," IZA Discussion Papers 6353, Institute for the Study of Labor (IZA) February 2012.

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

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

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

Donald, Stephen G and Whitney K Newey, "Choosing the Number of Instruments," *Econometrica*, September 2001, *69* (5), 1161–91.

Eichenbaum, Martin S, Lars Peter Hansen, and Kenneth J Singleton, "A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice

under Uncertainty," *The Quarterly Journal of Economics*, February 1988, *103* (1), 51–78.

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

Goodman, Joshua, "The labor of division: returns to compulsory math coursework," Working Paper Series, Harvard University, John F. Kennedy School of Government 2009.

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

_ **and Richard J. Smith**, "Generalized Empirical Likelihood Estimators And Tests Under Partial, Weak, And Strong Identification," *Econometric Theory*, August 2005, *21* (04), 667–709.

Hahn, Jinyong and Atsushi Inoue, "A Monte Carlo Comparison Of Various Asymptotic Approximations To The Distribution Of Instrumental Variables Estimators," *Econometric Reviews*, 2002, *21* (3), 309–336.

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

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

Han, Chirok and Peter C. B. Phillips, "GMM with Many Moment Conditions," *Econometrica*, 01 2006, *74* (1), 147–192.

Hansen, Christian and Damian Kozbur, "Instrumental Variables Estimation with Many Weak Instruments Using Regularized JIVE," *working paper*, 2014.

_ , Jerry Hausman, and Whitney Newey, "Estimation With Many Instrumental Variables," *Journal of Business & Economic Statistics*, 2008, *26*, 398–422.

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

Hausman, Jerry, Randall Lewis, Konrad Menzel, and Whitney Newey, "Properties of the CUE estimator and a modification with moments," *Journal of Econometrics*, 2011, *165* (1), 45 – 57. Moment Restriction-Based Econometric Methods.

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

Joensen, Juanna Schrøter and Helena Skyt Nielsen, "Is there a Causal Effect of High School Math on Labor Market Outcomes?," *Journal of Human Resources*, 2009, *44* (1).

Kapetanios, G and M Marcellino, "Factor-GMM estimation with large sets of possibly weak instruments," *Computational Statistics and Data Analysis*, 2010, *54*, 2655–2675.

Kleibergen, Frank, "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, September 2002, *70* (5), 1781–1803.

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

Krugman, Paul, *Geography and Trade*, Vol. 1 of *MIT Press Books*, The MIT Press, July 1992.

Kuersteiner, Guido, "Kernel-weighted GMM estimators for linear time series models," *Journal of Econometrics*, 2012, *170*, 399–421.

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

Lee, Valerie E., Robert G. Croninger, and Julia B. Smith, "Course-Taking, Equity, and Mathematics Learning: Testing the Constrained Curriculum Hypothesis in U.S. Secondary Schools," *Educational Evaluation and Policy Analysis*, 1997, *19* (2), 99–121.

Levine, Phillip B and David J Zimmerman, "The Benefit of Additional High-School Math and Science Classes for Young Men and Women," *Journal of Business & Economic Statistics*, April 1995, *13* (2), 137–49.

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

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

Lind, Jo Thori and Halvor Mehlum, "With or Without U? The Appropriate Test for a U-Shaped Relationship*," *Oxford Bulletin of Economics and Statistics*, 2010, *72* (1), 109–118.

Malamud, Ofer, "Breadth versus Depth: The Timing of Specialization in Higher Education," *LABOUR*, December 2010, *24* (4), 359–390.

_ , "The Effect of Curriculum Breadth and General Skills on Unemployment," Technical Report, University of Chicago and NBER 2012.

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

McFadden, D. L., "Modelling the Choice of Residential Location," *Spatial Interaction Theory and Planning Models, ed. by A. Karlqvist, L. Lundqvist, F. Snikcars, and J.Weibull. New York: North-Holland,*, 1978, pp. 75–96.

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

Montmarquette, Claude, Kathy Cannings, and Sophie Mahseredjian, "How do young people choose college majors?," *Economics of Education Review*, December 2002, *21* (6), 543–556.

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

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

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

_ , "Efficient Estimation of Models with Conditional Moment Restrictions," in G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., *Handbook of Statistics*, Vol. 11, Elsevier, 1993, pp. 419–454.

_ **and Frank Windmeijer**, "Generalized Method of Moments With Many Weak Moment Conditions," *Econometrica*, 05 2009, *77* (3), 687–719.

Okui, Ryo, "Shrinkage methods for instrumental variable estimation," Econometric Society 2004 Far Eastern Meetings 678, Econometric Society August 2004.

Palan, Nicole, "Measurement of Specialization - The Choice of Indices," FIW Working Paper series 062, FIW December 2010.

Pavan, Ronni and Josh Kinsler, "The Specificity of General Human Capital: Evidence from College Major Choice," Technical Report 2012.

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

Rose, Heather and Julian R. Betts, "The Effect of High School Courses on Earnings," *The Review of Economics and Statistics*, May 2004, *86* (2), 497–513.

Silos, Pedro and Eric Smith, "Human capital portfolios," Technical Report 2012.

Smith, Eric, "Sector-Specific Human Capital and the Distribution of Earnings," *Journal of Human Capital*, 2010, *4* (1), 35–61.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, May 1997, *65* (3), 557–586.

Stinebrickner, Todd R. and Ralph Stinebrickner, "Math or Science? Using Longitudinal Expectations Data to Examine the Process of Choosing a College Major," NBER Working Papers 16869, National Bureau of Economic Research, Inc March 2011.

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

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

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

Turner, Sarah E. and William G. Bowen, "Choice of major: The changing (unchanging) gender gap," *Industrial and Labor Relations Review*, January 1999, *52* (2), 289–313.

van der Vaart, Aad W. and Jon A. Wellner, *Weak Convergence and Empirical Processes*, Springer, 1996.

Yogo, Motohiro, "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak," *The Review of Economics and Statistics*, 03 2004, *86* (3), 797–810.

Zafar, Basit, "College major choice and the gender gap," Technical Report 2009.

Zivot, Eric, Richard Startz, and Charles R Nelson, "Valid Confidence Intervals and Inference in the Presence of Weak Instruments," *International Economic Review*, November 1998, *39* (4), 1119–46.