Journal of Econometrics 170 (2012) 383–398

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

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

# A regularization approach to the many instruments problem

Marine Carrasco

*Université de Montréal, Departement de Sciences Economiques, CP 6128, succ Centre Ville, Montreal, QC H3C3J7, Canada*

ARTICLE INFO

*Article history:*
Available online 2 June 2012

*Keywords:*
Many instruments
Mean square error
Regularization methods

ABSTRACT

This paper focuses on the estimation of a finite dimensional parameter in a linear model where the number of instruments is very large or infinite. In order to improve the small sample properties of standard instrumental variable (IV) estimators, we propose three modified IV estimators based on three different ways of inverting the covariance matrix of the instruments. These inverses involve a regularization or smoothing parameter. It should be stressed that no restriction on the number of instruments is needed and that all the instruments are used in the estimation. We show that the three estimators are asymptotically normal and attain the semiparametric efficiency bound. Higher-order analysis of the MSE reveals that the bias of the modified estimators does not depend on the number of instruments. Finally, we suggest a data-driven method for selecting the regularization parameter. Interestingly, our regularization techniques lead to a consistent nonparametric estimation of the optimal instrument.



## 1. Introduction

This paper considers the efficient estimation of a finite dimensional parameter in a linear model where the number of potential instruments is very large or infinite. The large number of moment conditions may stem from taking interactions between various exogenous variables or from nonlinear transformations of the exogenous variables. For example, Dagenais and Dagenais (1997) estimate a model with errors in variables using valid instruments obtained from higher-order moments of available variables. In principle, the asymptotic efficiency of the instrumental variable (IV) estimator improves when more moment conditions are used. However, it was observed that in finite samples, the inclusion of an excessive number of moments may be harmful (see Angrist and Krueger, 1991; Andersen and Sorensen, 1996). The poor performance of the IV estimator is due to the fact that its bias increases with the number of moment conditions (Bekker, 1994; Newey and Smith, 2004). It could be tempting to solve the problem of many instruments by taking a few instruments instead. However, an ad hoc selection of instruments will inevitably lead to a loss of efficiency. It may even lead to a loss of identification as pointed out by Dominguez and Lobato (2004). Therefore, it is desirable to have a method that would allow us to keep all the moment conditions available.

The main originality of our estimation procedure is that we do not need to restrict the number of instruments, which may be smaller or larger than the sample size, or even infinite. No instruments are discarded a priori. Our estimators are based on three ways to compute a regularized inverse of the (possibly infinite dimensional) covariance matrix of the instruments. These three regularizations are taken from the literature on inverse problems, see Kress (1999) and Carrasco et al. (2007). The first estimator based on Tikhonov (ridge) regularization was first proposed by Carrasco and Florens (2000). The other two estimators are new. The second estimator is based on an iterative method called Landweber–Fridman. This method is particularly useful in the case of many instruments as it is less computationally intensive than the other two. The third estimator is based on the principal components associated with the largest eigenvalues. All these methods involve a regularization parameter, which is the counterpart of the smoothing parameter that appears in the nonparametric literature. Following the same approach as Nagar (1959) and Donald and Newey (2001, DN henceforth), we compute the approximate mean square error (MSE) of the estimators and suggest selecting the smoothing parameter that minimizes the MSE. Our estimators can be thought of as alternatives to Empirical Likelihood estimators to solve the problem of many instruments. Some limitations apply. When the number of instruments is infinite, our asymptotic theory relies on the assumption that the instruments are sufficiently correlated with each other. This condition seems plausible in practical applications. Moreover, all instruments are assumed to be valid and we do not address the issue of weak instruments.

Interestingly, our regularization techniques have two very different interpretations depending on whether they are used to handle a large but finite number of orthogonality conditions (as in Angrist and Krueger, 1991), or to estimate efficiently a parameter identified by a conditional moment condition. In the first case, regularization is a way to avoid the bias that arises

*E-mail address:* marine.carrasco@umontreal.ca.

0304-4076/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2012.05.012

when using 2SLS. In the second case, our estimator is found to be equivalent to the IV estimator that uses as instrument a nonparametric estimator of the (unknown) optimal instrument. The regularization parameter, $\alpha$, is then a smoothing parameter that plays the same role as the bandwidth in kernel smoothing. We show that in some cases, our estimator of the instrument actually coincides with the spline smoothing estimator (Eubank, 1988).

Donald and Newey's and our three estimators are asymptotically equivalent since they all attain the semiparametric efficiency bound. We compare their small sample properties using both the theoretical expression of the MSE and Monte Carlo experiments. The theoretical MSE shows that the Tikhonov regularization may not perform as well as the other two regularizations, when the instruments capture well the functional form of the endogenous variable. The simulations show that, as expected, the relative performance of the DN estimator depends on whether the initial guess on the importance of the first instruments is correct or not.

The related literature is vast. Various asymptotically efficient estimators have been previously proposed in the literature. Starting from a conditional moment condition, Newey (1993) shows how to estimate nonparametrically the optimal instrument using a nearest neighbor estimator and hence circumvents the many instruments that arise when considering a series expansion for instance. Linton (2002) derives the higher-order expansion of Newey's (1993) estimator. Then, he derives an optimal bandwidth selection based on this expansion. Linton allows for heteroskedasticity of unknown form, which we do not permit here. However, Linton's approach does not apply to the case where there are many orthogonality conditions. Chen et al. (2009) propose an efficient estimator constructed as the weighted sum of inefficient but consistent estimators. The empirical likelihood (EL) estimator (Owen, 1988) has a bias that does not depend on the number of moment conditions and is therefore an attractive alternative to IV in presence of many instruments. Donald et al. (2003) and Kitamura et al. (2004) construct modified versions of the EL estimator that can handle an increasing number of moment conditions and are asymptotically efficient. Both estimators involve a smoothing parameter (the number of instruments in the first paper, the bandwidth of a kernel estimator in the second) but the authors do not provide a rule for selecting these parameters in practice. Finally, Donald and Newey (2001) propose to select the number of instruments, $L$, that minimizes the mean square error (MSE) of the estimator. As pointed out by the authors, this method will work best if the instruments are ordered so that the first one are the most influential. In contrast, our methods do not require a ranking of the instruments and are simpler to implement than the modified EL estimators of Kitamura et al. (2004) and Donald et al. (2003).

Principal components have been used for a long time as a dimension reduction device. Amemiya (1966) provides a rationale for using principal components regression. This method found a new ground of application in factor models (see Stock and Watson (2002), Bai and Ng (2002, 2010) and references therein). In this literature, it is assumed that there is a fixed number of factors, but this does not have to be the case here. Doran and Schmidt (2006) investigate in a simulation study the use of principal components in panel data models. Their approach is very similar to what we propose here. In an attempt to improve the properties of the generalized method of moments (GMM) in the presence of many moments, Kuersteiner (2006, 2012) proposes a kernel weighted GMM estimator and Okui (2011) introduces a shrinkage parameter to allocate less weight on a subset of instruments.

Section 2 introduces the four regularization methods we consider and the associated estimators. Section 3 derives an expression for the approximate MSE in the four cases. In Section 4, we give a feasible MSE based on cross-validations and Mallows $C_p$ criteria. Section 5 makes a comparison with other familiar estimators. Section 6 presents a limited Monte Carlo experiment. An application to measuring the return to education is examined in Section 7. Section 8 concludes. The proofs are collected in Appendix.

## 2. Regularized versions of 2SLS

### 2.1. Presentation of the estimators

The model is

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

$i = 1, 2, \ldots, n$. $\delta$ is the $p \times 1$ vector of interest. $E(\varepsilon_i|x_i) = E(u_i|x_i) = 0, E\left(\varepsilon_i^2|x_i\right) = \sigma_\varepsilon^2.y_i$ 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. In the following, we adopt some general notations in order to encompass in the same framework the case where the number of moment conditions is finite, countable infinite and the case where there is a continuum of moment conditions. The estimation is conducted using a sequence of instruments $Z_i = Z(\tau, x_i)$. $\tau$ may be an integer or a vector taking its values in a set. The set of values of $\tau$ is denoted $\mathcal{S}$. Let $\pi$ be a positive measure on $\mathcal{S}$ with support equal to $\mathcal{S}$. We denote by $L^2(\pi)$ the Hilbert space of square integrable functions with respect to $\pi$.

We provide here a few examples of $Z$ and $\pi$. More insights on how to choose them will be given in Section 2.4.

- Finite number of moments. $Z_i = \left(Z_{i,1}, Z_{i,2}, \ldots, Z_{i,L}\right)$, we can take $\pi$ the uniform probability measure on $\mathcal{S} = \{1, 2, \ldots, L\}$
- Countable infinite number of moments. $Z_i = \left(Z_{i,1}, Z_{i,2}, \ldots\right)$. For example if $x_i \in (-1, 1)$, we may take $Z_{i,j} = x_i^{j-1}$ and $\pi$ a counting measure on $\mathcal{S} = \mathbb{N}$.
- Continuum of moments. $Z_i = Z(\tau, x_i) = \exp\left(i\tau' x_i\right)$ with $\tau \in \mathcal{S} = \mathbf{R}^{\dim(x)}$, $\pi$ can be taken equal to the density of the standard normal as in Carrasco et al. (2007).

We estimate $\delta$ based on the orthogonality condition

$$E\left(\left(y_i - W_i'\delta\right) Z_i\right) = 0$$

using the extension of the generalized method of moments described in Carrasco and Florens (2000, 2008). Let $h_n(\tau, \delta) = \sum_{i=1}^n \left(y_i - W_i'\delta\right) Z_i/n$ and

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad (n \times 1), \qquad W = \begin{pmatrix} W_1' \\ W_2' \\ \vdots \\ W_n' \end{pmatrix} \quad (n \times p),$$

$$u = \begin{pmatrix} u_1' \\ u_2' \\ \vdots \\ u_n' \end{pmatrix} \quad (n \times p).$$

The covariance operator of the moment functions is the operator which associates with any function $g$ of $L^2(\pi)$ the following function of $L^2(\pi)$:

$$\int E\left(h_n(\tau_1, \delta)\, \overline{h_n(\tau_2, \delta)}\right) g(\tau_2)\, \pi(\tau_2)\, d\tau_2$$

$$= \int E\left(\varepsilon_i^2 Z(\tau_1, x_i)\, \overline{Z(\tau_2, x_i)}\right) g(\tau_2)\, \pi(\tau_2)\, d\tau_2,$$

where $\overline{Z(\tau_2, x_i)}$ denotes the complex conjugate of $Z(\tau_2, x_i)$. Because of the homoscedasticity of the error, the covariance

operator simplifies to

$$\sigma_\varepsilon^2 \int E\left(Z(\tau_1, x_i)\overline{Z(\tau_2, x_i)}\right) g(\tau_2)\pi(\tau_2)\, d\tau_2$$

and hence does not depend on $\delta$. In the following, we use the notation $K$ for the covariance operator divided by $\sigma_\varepsilon^2$. Namely, $K$ is the operator from $L^2(\pi)$ to $L^2(\pi)$ such that

$$(Kg)(\tau_1) = \int E\left(Z(\tau_1, x_i)\overline{Z(\tau_2, x_i)}\right) g(\tau_2)\pi(\tau_2)\, d\tau_2.$$

Note that when $\mathcal{S}$ is discrete, a summation replaces the integral. The operator $K$ is assumed to be a Hilbert–Schmidt operator (for a definition, see Carrasco et al., 2007) and hence has a discrete spectrum. $K$ will be Hilbert–Schmidt if there is a sufficiently strong dependence among the instruments. If there is an infinite number of independent instruments, $K$ is not Hilbert–Schmidt, because the identity operator is not compact. $K$ is also assumed to have only nonzero eigenvalues. Let $(\lambda_j, \phi_j : j = 1, 2, \ldots)$ be the eigenvalues and orthonormal eigenfunctions of $K$. The operator $K$ is estimated by its sample counterpart $K_n$ defined as

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

When the number of moment conditions is infinite, the inverse of $K_n$ needs to be regularized because it is nearly singular. By definition (see Kress, 1999, p. 269), a regularized inverse of an operator $K : L^2(\pi) \to L^2(\pi)$ is a bounded linear operator $R_\alpha : L^2(\pi) \to L^2(\pi)$, $\alpha > 0$, that satisfies the pointwise convergence:

$$\lim_{\alpha\to 0} R_\alpha K\varphi = \varphi, \quad \text{for all } \varphi \in L^2(\pi).$$

Various types of regularization techniques will be discussed shortly, they depend on a regularization parameter $\alpha$, the choice of which is the topic of this paper. Let $\left(K_n^\alpha\right)^{-1}$ denote a regularized inverse of $K_n$ and $\left(K_n^\alpha\right)^{-1/2} = \left(\left(K_n^\alpha\right)^{-1}\right)^{1/2}$. Let $\mathcal{E}$ denotes the space $\mathbf{R}^n$ endowed with the norm $\|v\|^2 = v'v/n$. For convenience, we use the same notation $\langle , \rangle$ for the inner product in $L^2(\pi)$ and in $\mathcal{E}$. The regularized 2SLS estimator of $\delta$ is defined as

$$\hat\delta = \arg\min_\delta \left\langle \left(K_n^\alpha\right)^{-1/2} h_n(\cdot, \delta), \left(K_n^\alpha\right)^{-1/2} h_n(\cdot, \delta)\right\rangle.$$

Solving explicitly the minimization problem gives

$$\hat\delta = \left(W' P_n^\alpha W\right)^{-1} W' P_n^\alpha y$$
$$= \left(\hat W' W\right)^{-1} \hat W' y,$$

where $\hat W = P_n^\alpha W$, $P_n^\alpha$ is a $n \times n$ matrix defined as

$$P_n^\alpha = T_n\left(K_n^\alpha\right)^{-1} T_n^*$$

and $T_n : L^2(\pi) \to \mathcal{E}$ such that

$$T_n g = \begin{bmatrix} \langle Z_1, g\rangle \\ \vdots \\ \langle Z_n, g\rangle \end{bmatrix}$$

for any $g \in L^2(\pi)$ and $T_n^* : \mathcal{E} \to L^2(\pi)$ is the adjoint of $T_n$, it satisfies

$$T_n^* v = \frac{1}{n}\sum_{i=1}^{n} Z_i v_i \equiv \hat E(Z_i v_i)$$

for any $v = (v_1, v_2, \ldots, v_n)' \in \mathbf{R}^n$. It is easy to check that $K_n = T_n^* T_n$ and $T_n T_n^*$ is a $n \times n$ matrix with $(i, j)$ element $\langle Z_i, Z_j\rangle/n$. Let $\hat\lambda_1 \geq \hat\lambda_2 \geq \cdots \hat\lambda_n > 0$ and $\hat\phi_j, j = 1, 2, \ldots, n$, be the nonzero eigenvalues and the corresponding orthonormalized eigenfunctions of $K_n$. We use a hat because $\hat\lambda_j$ and $\hat\phi_j$ are consistent estimators of the corresponding eigenfunctions and eigenvalues of $K$, $\lambda_j$ and $\phi_j$. Let $\hat\psi_j$ be the eigenvectors of the $n \times n$ matrix $T_n T_n^*$. Then, we have $T_n\widehat{\phi}_j = \sqrt{\hat\lambda_j}\hat\psi_j$ and $T_n^*\hat\psi_j = \sqrt{\hat\lambda_j}\widehat{\phi}_j$. It may be useful to see what the formulas are when there is a finite number of instruments $L$ and $\pi = 1$.

$$Z_i = \begin{pmatrix} Z_{i,1} \\ Z_{i,2} \\ \vdots \\ Z_{i,L}\end{pmatrix} \quad (L \times 1), \qquad \underline{Z} = \begin{pmatrix} Z_1' \\ Z_2' \\ \vdots \\ Z_n'\end{pmatrix} \quad (n \times L).$$

Then, $K_n$ is a $L \times L$ matrix:

$$K_n = \frac{1}{n}\underline{Z}'\underline{Z}.$$

Note that if no regularization is applied, $P = T_n(K_n)^{-1} T_n^* = T_n\left(T_n^* T_n\right)^{-1} T_n^* = \underline{Z}\left(\underline{Z}'\underline{Z}\right)^{-1}\underline{Z}'$ is the projection matrix on the vector of instruments. Then $\hat\delta$ is the usual IV estimator of $\delta$ using all the instruments.

The eigenfunctions $\hat\psi_j, j = 1, 2, \ldots, L$ can be computed from $\hat\psi_j = T_n\widehat{\phi}_j \Big/ \sqrt{\hat\lambda_j} = \left(Z_1'\widehat{\phi}_j \Big/ \sqrt{\hat\lambda_j}, \ldots, Z_n'\widehat{\phi}_j \Big/ \sqrt{\hat\lambda_j}\right)'$. For $n > L$, the operator $T_n T_n^*$ has the zero eigenvalue associated with $n - L$ eigenfunctions. Note that when $n > L$, it is easier to compute the eigenfunctions of $K_n$ and then infer those of $T_n T_n^*$. In contrast, when $L > n$ or $L$ is infinite, it is easier to compute the $n$ eigenvalues and eigenfunctions of $T_n T_n^*$ and infer the eigenfunctions of $K_n$ by using the formula $\widehat{\phi}_j = T_n^*\hat\psi_j \Big/ \sqrt{\hat\lambda_j}, j = 1, 2, \ldots, n$.

We consider four regularization schemes. The first three are traditionally applied in statistics (Kress, 1999), the fourth one is commonly applied in factor models (Stock and Watson, 2002). We first define the regularized inverse of $K$. To obtain the regularized inverse of $K_n$, it suffices to replace $\phi_j$ by $\widehat{\phi}_j$ and $\lambda_j$ by $\hat\lambda_j$.

(1) *Tikhonov (T)*

This regularization scheme is closely related to the ridge regression.

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

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

(2) *Landweber–Fridman (LF)*

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

$$\begin{cases} \hat\varphi_l = \left(1 - cK^2\right)\hat\varphi_{l-1} + cKr, l = 1, 2, \ldots, 1/\alpha - 1 \\ \hat\varphi_0 = cKr, \end{cases}$$

where $1/\alpha - 1$ is some positive integer. Alternatively, we have

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

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

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

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

for some $\alpha > 0$.

(4) *Principal components (PC)*

This method is a variation around SC and consists in using the first $1/\alpha$ eigenfunctions:

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

where $1/\alpha$ is some positive integer. PC and SC are perfectly equivalent, only the definition of the regularization term $\alpha$ differs. In practice, both methods will give the same estimator. We will study the properties of SC in detail.

These four regularized inverses can be rewritten as

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

where for LF: $q\left(\alpha, \lambda_j^2\right) = 1 - \left(1 - c\lambda_j^2\right)^{1/\alpha}$, for SC: $q\left(\alpha, \lambda_j^2\right) = I\left(\lambda_j^2 \geq \alpha\right)$, for T: $q\left(\alpha, \lambda_j^2\right) = \frac{\lambda_j^2}{\lambda_j^2 + \alpha}$, for PC: $q\left(\alpha, \lambda_j^2\right) = I\left(j \leq 1/\alpha\right)$. Note that $0 \leq q\left(\alpha, \lambda_j^2\right) \leq 1$. Using this notation, we have

$$P_n^\alpha v = \sum_{j=1}^{n} q\left(\alpha, \hat{\lambda}_j^2\right) \left\langle v, \hat{\psi}_j \right\rangle \hat{\psi}_j = \frac{1}{n} \sum_{j=1}^{n} q\left(\alpha, \hat{\lambda}_j^2\right) \left(\hat{\psi}_j' v\right) \hat{\psi}_j$$

for any $v \in \mathbf{R}^n$ and $\operatorname{tr}\left(P_n^\alpha\right) = \sum_{j=1}^{n} q\left(\alpha, \hat{\lambda}_j^2\right)$. The matrix $P_n^\alpha$ is idempotent for SC and PC but not for LF and T. In the case of PC, $P_n^\alpha$ is the projection matrix on the space spanned by $\hat{\psi}_j, j = 1, \ldots, n$ associated with the largest (positive) eigenvalues.

The four regularization methods involve a regularization term, $\alpha$. The set of possible values for $\alpha$ is continuous in the case of T, but is discrete for the three other methods. To see this for SC, observe that the value of $\left(K_n^\alpha\right)^{-1} r$ will vary only for values of $\alpha$ that are equal to the eigenvalues. As there are at most $n$ nonzero eigenvalues, the set of $\alpha$ has at most cardinal $n$. We will choose $\alpha$ so that it minimizes the mean square error (MSE) of $\hat{\delta}$.

### 2.2. Asymptotic properties of the regularized 2SLS estimators

This section establishes that the regularized 2SLS estimators reach the semiparametric efficiency bound under some standard assumptions. It is well known (see for instance Newey, 1993) that the semiparametric efficiency bound is the asymptotic variance of the unfeasible IV estimator that would use the unknown $f(x)$ as instrument. Let $f_a(x)$ be the $a$th element of $f(x)$.

**Proposition 1.** *Assume* (i) $\{y_i, W_i, x_i\}$ *are iid,* $E\left(\varepsilon_i^2\right) = \sigma_\varepsilon^2$, $E\left[f(x_i) f(x_i)'\right]$ *exists and is nonsingular, K is compact, and* $\alpha$ *goes to zero as n goes to infinity. Moreover,* (ii) $f_a(x)$ *belongs to the closure of the linear span of* $\{Z(\cdot, x)\}$ *for* $a = 1, \ldots, p$. *Then, the T, LF, and SC estimators satisfy:*

(1) *Consistency:* $\hat{\delta} \to \delta_0$ *in probability as n and* $n\alpha^{1/2}$ *go to infinity.*
(2) *Asymptotic normality: If moreover, each element of* $E(Z(., x_i) W_i)$ *belongs to the range of K, then*

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

*as n and* $n\alpha$ *go to infinity.*

The compactness of $K$ implies a functional central limit theorem (see van der Vaart and Wellner, 1996), namely $\sum_{i=1}^{n} \frac{Z(.,x_i)\varepsilon_i}{\sqrt{n}}$ converges in $L^2(\pi)$ to a mean zero Gaussian process with covariance operator $\sigma_\varepsilon^2 K$. Condition (ii) is equivalent to the condition that $f$ can be approached by a linear combination of the instruments and corresponds to Assumption 2(ii) of DN:

*For each L there exists* $\pi_L$ *such that* $E\left(\left\|f(x) - \pi_L Z^L(x)\right\|^2\right) \to 0$ *as* $L \to \infty$, where $Z^L$ is a subset of the instruments.

This condition is necessary to achieve efficiency. Note that this property implies that $E(Z(\cdot, x_i) f_a(x_i))$ belongs to the range of $K^{1/2}$. A sufficient condition for the stronger requirement that $E(Z(\cdot, x_i) f_a(x_i))$ belongs to the range of $K$ is that for each $a = 1, \ldots, p$, there exists a function $g_a$ of $L^2(\pi)$ such that $f_a = Tg_a$, that is $f_a(x_i) = \langle Z_i, g_a \rangle, i = 1, 2, \ldots, n$.

We illustrate Condition (ii) by three examples where this condition is satisfied.

(a) Assume the vector of instruments, $Z$, is finite and $E(W|Z) = \Pi' Z$. Then, efficiency is achieved by using all the instruments.
(b) Assume that $E(W|x) = f(x)$ is a smooth function of $x \in \mathbb{R}$. One could use power functions of $x: 1, x, x^2, \ldots$
(c) The same assumption as in (b). Another way to obtain efficiency is to consider $Z(\tau, x) = \exp(i\tau x)$ for $\tau \in \mathbb{R}$.

The choice of instruments to obtain efficiency is discussed in Newey (1993) and Carrasco and Florens (2008) among others.

As the three estimators have the same asymptotic distribution, it is necessary to rely on a second order analysis of their MSE to be able to discriminate between them. This will be done in Section 3.

### 2.3. Interpretation as nonparametric estimation of the optimal instrument

Under some conditions, $\hat{W} = P_n^\alpha W$ can be interpreted as a nonparametric estimator of the unknown function $f$. In this case, our estimator is the 2SLS estimator obtained by replacing the optimal instrument $f$ by its estimator. Our estimation of $f$ is actually standard in the machine learning and inverse literatures (see Vapnik, 1998; Van Rooij and Ruymgaart, 1999). We show how this estimate could be obtained directly. Consider the regression

$$W_i = f(x_i) + u_i \tag{1}$$

and assume that $f$ can be written as $f(x_i) = \langle Z(\cdot, x_i), \varphi(\cdot) \rangle$ for some unknown function $\varphi$ in $L^2(\pi)$. This representation of $f$ is very general. If $Z(\cdot, x_i) = \exp\left(i\tau' x_i\right)$ then $f$ admits such a representation provided that it is continuous and square integrable. Dropping the error term in (1), we look for the solution in $\varphi$ to the equation

$$W = \langle Z, \varphi \rangle,$$

where $W$ is the $n \times 1$ vector of $W_i$ and $Z$ is the $n \times 1$ vector of $Z(\cdot, x_i)$. Let $T$ be as before the operator that associates with functions of $L^2(\pi)$ elements of $R^n$ such that $T\varphi = \langle Z, \varphi \rangle$. We have to solve the inverse problem

$$W = T\varphi.$$

By preconditioning by $T^*$, the adjoint of $T$, we obtain

$$T^* W = T^* T \varphi. \tag{2}$$

As we saw in Section 2.1, $T^* T = K_n$. A solution to (2) is given by

$$\hat{\varphi} = \left(K_n^\alpha\right)^{-1} T^* W.$$

Consequently, an estimator of $f$ is obtained by

$$\hat{W} = T\hat{\varphi} = P_n^\alpha W = \frac{1}{n} \sum_{j=1}^{n} q\left(\alpha, \hat{\lambda}_j^2\right) \left(\psi_j' W\right) \psi_j. \tag{3}$$

More insights on this estimator are provided in the next subsection.

### 2.4. Choice of the instruments and inner product

We now discuss the choice of the instruments $Z$ and the weight $\pi$ that appears in the inner product. Our discussion below borrows from the machine learning literature (see for instance Vapnik, 1998; Hofmann et al., 2008). It is important to outline the role played by $\pi$ and $Z$. They affect the estimation of $\delta$ only through the determination of the eigenvalues and eigenfunctions of the $n \times n$ matrix $TT^*$ with principal element proportional to

$$\tilde{k}(x_i, x_j) = \langle Z(\cdot, x_i), Z(\cdot, x_j) \rangle = \int Z(\tau, x_i) \overline{Z(\tau, x_j)} \pi(\tau) d\tau. \quad (4)$$

As they enter jointly in the calculation of $\tilde{k}(x_i, x_j)$, we can not completely disentangle the role of each of them. This raises the possibility of choosing $\tilde{k}$ a priori without specifying the mappings $Z(\tau, x_i)$ and $\pi$. However, $\tilde{k}$ can not be chosen completely arbitrarily as the corresponding matrix must be definite positive so that it defines a reproducing kernel Hilbert space. Let us define $\widetilde{K} = TT^*$. In the case of Tikhonov regularization, $\hat{W}$ can be rewritten as

$$\hat{W} = \left(\widetilde{K}^2 + \alpha I\right)^{-1} \widetilde{K}^2 W = \widetilde{K}^2 \left(\widetilde{K}^2 + \alpha I\right)^{-1} W. \quad (5)$$

It follows that $\hat{W}_i = \hat{f}(x_i)$ where $\hat{f}(x)$ takes the form of a linear combination of the $\tilde{k}(x, x_i)$:

$$\hat{f}(x) = \sum_{i=1}^{n} \omega_i \tilde{k}(x, x_i). \quad (6)$$

We give below various examples of kernel $\tilde{k}$. The choice of $Z_i = Z(x_i, \cdot)$ is mainly dictated by efficiency considerations. As we see in Proposition 1, the estimator $\hat{\delta}$ is asymptotically efficient (in the sense of the semiparametric efficiency bound) if the space spanned by $\{Z(x_i, \cdot)\}$ is sufficiently rich to encompass the unknown function $f$. A natural choice is to use $Z(\tau, x_i) = \exp(i\tau' x_i)$ for $x_i$, $\tau \in \mathbf{R}^d$. This choice of $Z(\tau, x_i)$ has the advantage of being bounded and the approximation of $f(x_i)$ by a linear combination of $Z(\tau \cdot, x_i)$ has an interpretation in terms of Fourier series expansion. This function is particularly favored in machine learning. Then, $\tilde{k}$ takes a simple form

$$\tilde{k}(x_i, x_j) = \int \exp\left(i(x_i - x_j)'\tau\right) \pi(\tau) d\tau \equiv h(x_i - x_j). \quad (7)$$

Hence, if $\pi$ is a density, $h$ is a characteristic function. On the other hand, if $\pi$ is a characteristic function, $h$ is proportional to a density as a result of Fourier inversion formula. It follows from (6) and (7) that $\hat{f}(x)$ takes the form of a kernel estimator with specific weights $\omega_i$, which differ in general from the usual form $W_i / \sum_{i=1}^{n} \tilde{k}(x, x_i)$. We describe two choices of $\pi$ that have been successful in recovering functions in the machine learning literature.

(a) *Gaussian kernel*

Let $\pi$ be the density of a multivariate normal such that $\pi(\tau) = \frac{\sigma^d}{\sqrt{2\pi}^d} \exp\left(-\frac{\sigma^2 \|\tau\|^2}{2}\right)$, then

$$\tilde{k}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right).$$

The parameter $\sigma^2$ may be set equal to the variance of $W_i$ or may be selected by cross-validation by including $\sigma^2$ along $\alpha$ in the criterion to minimize. We do not pursue the issue of the selection of $\sigma^2$ any further.

(b) *B-spline*

Among the most popular series approximations are those based on splines and B-splines. They are less sensitive to outliers than polynomials. To simplify the exposition, we assume here that the dimension of $x_i$ is $d = 1$. Let $\otimes$ denote the convolution operator, that is $(f \otimes g)(x) = \int f(y) g(x - y) dy$ and $\pi(\tau) = \sin c^{2q+1}(\tau/2)$ where $\sin c(\tau) = \sin(\tau)/\tau$ and $q \in \mathbf{N}$. Then the resulting kernel is

$$\tilde{k}(x_i, x_j) = \otimes^{2q+1} I_{\left[-\frac{1}{2}, \frac{1}{2}\right]}(x_i - x_j) \equiv B_{2q+1}(x_i - x_j).$$

This result follows from the fact that $\sin c(\tau/2)$ is the characteristic function of the uniform distribution on $\left[-\frac{1}{2}, \frac{1}{2}\right]$. Note that $B_{2q+1}$ takes the following form

$$B_{2q+1}(u) = \sum_{j=0}^{2q+1} \frac{(-1)^j}{j!} \binom{2q+1}{j} \max\left(0, u + \frac{2q+1}{2} - j\right)^{2q}.$$

Using this kernel, we get approximations of the form

$$\hat{f}(x) = \sum_{i=1}^{n} \omega_i B_{2q+1}(x - x_i).$$

Below, we provide three examples of kernel $\tilde{k}$ that are not of the form (7).

(c) *Polynomial kernel*

Let $x$ and $y$ be two vectors of $\mathbf{R}^p$, we define $\tilde{k}$ as a homogeneous polynomial kernel i.e.

$$\tilde{k}(x, y) = (x'y)^p = \left(\sum_{j=1}^{d} x_j y_j\right)^p = \sum_{j_1, \ldots, j_p = 1}^{d} x_{j_1} \ldots x_{j_p} y_{j_1} \ldots y_{j_p} = \langle Z_p(x), Z_p(y) \rangle.$$

Here $Z_p$ maps $x$ in $\mathbf{R}^d$ to the vector $Z_p(x)$ whose entries are all possible $p$th degree ordered products of the elements of $x$ and $\langle Z_p(x), Z_p(y) \rangle$ denotes the classical dot product $Z_p(x)' Z_p(y)$. For illustration, let $p = d = 2, x = (x_1, x_2)', y = (y_1, y_2)'$. Then $\tilde{k}(x, y) = (x_1 y_1)^2 + 2x_1 y_1 x_2 y_2 + (x_2 y_2)^2$ and one can choose either $Z_p(x) = \left(x_1^2, x_2^2, x_1 x_2, x_2 x_1\right)'$ or $Z_p(x) = \left(x_1^2, x_2^2, \sqrt{2} x_1 x_2\right)'$.

(d) *Kernel generating expansion on Hermite polynomials*

Let $d = 1$. Consider the Hermite polynomials

$$H_j(x) = \mu_j P_j(x) e^{-x^2},$$

where

$$P_j(x) = (-1)^j e^{x^2} \left(\frac{d}{dx}\right)^j e^{-x^2}$$

and $\mu_j$ are chosen so that $\int H_j(x)^2 e^{-x^2} dx = 1$. We define $\tilde{k}$ as

$$\tilde{k}(x, y) = \sum_{j=1}^{\infty} \rho^j H_j(x) H_j(y) = \frac{1}{\sqrt{\pi(1 - \rho^2)}} \exp\left\{\frac{2xy\rho}{1 + \rho} - \frac{(x - y)^2 \rho^2}{1 - \rho^2}\right\}$$

for some $0 < \rho < 1$. This kernel provides an approximation of $f$ on the basis of Hermite polynomials.

The last example can be generalized to

$$\tilde{k}(x, y) = \sum_{j=1}^{\infty} r_j \psi_j(x) \psi_j(y), \quad (8)$$

where the $\psi_j$ are an orthonormal basis and $r_j$ converges to zero as $j$ goes to infinity, see Vapnik (1998, pp. 461–462). This kernel

provides an expansion of $f$ in terms of $\psi_j$. Then, the principal component estimator of $f$ is nothing but a series (sieve) estimator.

(e) *Kernel generating splines*

Assume that $x$ has a known support $[0, 1]$ and $f$ belongs to

$$W_2^{(m)} = \{f : f, f', \ldots, f^{(m-1)} \text{ absolutely continuous and } f^{(m)} \in L^2[0,1]\}.$$

Consider the following kernel

$$\tilde{k}(x,y) = \omega \sum_{j=0}^{m-1} x^j y^j + c(x,y)$$

with

$$c(x,y) = \frac{1}{((m-1)!)^2} \times \int_0^1 \max(0,(x-u))^{m-1} \max(0,(y-u))^{m-1} du.$$

For multivariate $x$, one can construct $d$-dimensional splines as the product of $d$ one-dimensional kernels (see Vapnik, 1998, p. 465). For comparison purposes, consider the following estimate of $f$:

$$W^* = \widetilde{K}\left(\widetilde{K} + \alpha I\right)^{-1} W.$$

Note that the only difference with (5) is that $\widetilde{K}^2$ has been replaced by $\widetilde{K}$. $W^*$ will have the same asymptotic properties as $\hat{W}$. Denote

$$\hat{f}(x) = \underline{\widetilde{k}}(x,\cdot)\left(\widetilde{K} + \alpha I\right)^{-1} W,$$

where $\underline{\widetilde{k}}(x,\cdot)$ is the $n$-vector of $\widetilde{k}(x, x_i)$. Interestingly, $\hat{f}$ coincides with the Bayes estimator of $f$ in the regression

$$W(x) = f(x) + \varepsilon(x), \quad x \in [0,1],$$

where $\{\varepsilon(x) : x \in [0,1]\}$ is a zero mean normal process with $cov(\varepsilon(x), \varepsilon(y)) = \sigma^2 I(x = y)$. $W(x)$ is to be understood as a stochastic process sampled at the points $W_i = W(x_i)$. A Bayesian structure is obtained by stating that $f(x)$ has the same prior distribution as the stochastic process:

$$\sum_{j=0}^{m-1} \theta_j x^j + bN(x), \quad b > 0,$$

where $\theta = (\theta_0, \ldots, \theta_{m-1})$ has a zero mean normal distribution with covariance matrix $\nu I_m$ and $\{N(x) : t \in [0,1]\}$ is a zero mean normal process with covariance $c(x,y)$. Wahba (1978) (see also Eubank, 1988, Proposition 5.1) showed that

$$\hat{f}(x) = E(f(x)|W_1, \ldots, W_n),$$

where $\omega = \nu/b^2$ and $\alpha = n\lambda = \sigma^2/b^2$. Moreover, if an improper prior distribution ($\nu \to \infty$) is used, then $\hat{f}$ becomes the smoothing spline estimator of $f$. Recall that the smoothing spline[1] estimator of $f$ is the element of $W_2^{(m)}$ which is the solution of

$$\min_f \frac{1}{n}\sum_{i=1}^n (W_i - f(x_i))^2 + \lambda \int_0^1 \left[f^{(m)}(x)\right]^2 dx$$

and notice that $\int_0^1 \left[f^{(m)}(x)\right]^2 dx$ is the norm of $f$ in the RKHS with kernel $c(x,y)$ and represents a measure of roughness of the function.

These few examples illustrate the versatility of our approach as we cover in the same framework, kernel-type, sieve, and spline estimators of $f$. Many other kernels $\widetilde{k}$ can be constructed (see Hofmann et al., 2008, for more examples).

[1] Smoothing spline estimator should not be confused with least-squares spline estimator. The latter is a series estimator where $f$ is approximated by its projection on the first elements of the spline basis.

## 3. Mean square error

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

**Assumption 1.** $\{y_i, W_i, x_i\}$ are iid, $E(\varepsilon_i^2) = \sigma_\varepsilon^2 > 0$, and $E(\|u_i\|^4 |x_i)$, $E(\varepsilon_i^4|x_i)$ are bounded.

**Assumption 2.** (i) $\bar{H} = E\left[f(x_i) f(x_i)'\right]$ exists and is nonsingular, (ii) there is a $\beta \geq 1/2$ such that

$$\sum_{j=1}^{\infty} \frac{\langle E(Z(x_i,\cdot) f_a(x_i)), \phi_j\rangle^2}{\lambda_j^{2\beta+1}} < \infty \tag{9}$$

for all $a = 1, 2, \ldots, p$, where $f_a$ is the $a$th column of $f$.

**Assumption 3.** (i) $E\left[\left((\varepsilon_i, u_i')'(\varepsilon_i, u_i')\right)|x_i\right]$ is constant, (ii) $K$ is a Hilbert–Schmidt operator with nonzero eigenvalues, (iii) $f(x_i)$ is bounded.

Assumptions 1, 2(i) and 3(i) and (iii) are imposed by DN. Assumption 2(ii) is used to derive the rate of convergence of the MSE. More precisely, it guarantees that $\left\|f - P_n^\alpha f\right\|^2 = O_p(\alpha^\beta)$ for LF and PC and $\left\|f - P_n^\alpha f\right\|^2 = O_p(\alpha^{\min(\beta,2)})$ for T. Condition (9) for $\beta = 1/2$ corresponds to the condition $f_a = Tg_a$ needed in Proposition 1. The value of $\beta$ in (9) measures how well the instruments approximate the reduced form, $f$. The larger $\beta$, the better the approximation is. Below Proposition 2, we give an example where $\beta$ is infinite. Now we turn to Assumption 3(ii). A sufficient condition for $K$ to be Hilbert–Schmidt is that its kernel is square integrable:

$$\iint |k(\tau_1,\tau_2)|^2 \pi(\tau_1)\pi(\tau_2) d\tau_1 d\tau_2 < \infty.$$

This implies in particular that its eigenvalues are square summable, $\sum_{j=1}^\infty \lambda_j^2 < \infty$. When $|k(\tau_1,\tau_2)|$ is bounded (as for $Z(\tau,x) = \exp(i\tau x)$) and $\pi$ is a pdf, this condition is automatically satisfied. Again for $Z(\tau,x) = \exp(i\tau x)$ and $\pi$ positive on $\mathbb{R}$, the property of the Fourier transform implies that 0 is not an eigenvalue of $K$. Remark that we do not need Assumption A(iii) of DN, namely $\max_{i\leq n} P_{ii}^\alpha \to 0$.

Denote $\sigma_{u\varepsilon} = E(\varepsilon_i u_i|x_i)$ and $\Sigma_u = E\left(u_i u_i'|x_i\right)$.

**Proposition 2.** *If Assumptions 1–3 are satisfied, $\sigma_{u\varepsilon} \neq 0$, $n\alpha^2 \to \infty$, for LF, SC, PC, and T regularizations, we have*

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

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

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

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

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

As usual in this type of computation, $S(\alpha)$ is composed of a bias term that increases when $\alpha$ goes to zero and a variance term that decreases when $\alpha$ goes to zero. 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. For instance if $f$ were a linear combination of the instruments, $\beta$ would be infinite, and the performance of T would be far worse than that of PC or LF. In order to illustrate this point, consider a factor model with $J$ factors.

$$W_i = \sum_{j=1}^{J} \omega_j f_{ij} + \varepsilon_i,$$

$$x_{ia} = \sum_{j=1}^{J} \gamma_{ija} f_{ij} + \nu_{ia}, \quad a = 1, 2, \ldots, L.$$

Assume that the factors $f_{ij}$ are normalized. It is well known that the $\psi_j, j = 1, \ldots, J$ associated with the $J$ largest eigenvalues are estimators of the factors (Chamberlain and Rothschild, 1983). Hence, Assumption 2(ii) is satisfied for any value of $\beta$. For the PC estimator, $P_n^\alpha f = f$ for $\alpha \leq 1/J$ and the second term in $S(\alpha)$ vanishes. On the other hand, for T, the second term never vanishes completely.

It would be interesting to compare $S(\alpha)$ with the expression of the approximate MSE given by DN for the 2SLS. We briefly review the results of DN. Consider a countable sequence of instruments $(Z_{i,1}, Z_{i,2}, \ldots)$. The DN estimator is based on the first $L$ instruments, where $L$ minimizes the MSE of the 2SLS estimator. Let $P^L$ be the $n \times n$ projection matrix on $(Z_{i,1}, Z_{i,2}, \ldots, Z_{i,L})$. The DN estimator is given by

$$\hat{\delta}_{DN} = \left(W' P^L W\right)^{-1} W' P^L y.$$

Its approximate MSE is

$$S(L) = H^{-1}\left[\sigma_{u\varepsilon}\sigma_{u\varepsilon}' \frac{L^2}{n} + \sigma_\varepsilon^2 \frac{f'\left(I - P^L\right) f}{n}\right] H^{-1}.$$

For DN, the smoothing or regularization parameter is $L$. $S(L)$ is similar to $S(\alpha)$ where $L$ plays the role of $1/\alpha$. There, $P_n^\alpha$ is replaced by $P^L$, which is a projection matrix so that $\left(I - P^L\right)^2 = \left(I - P^L\right)$ and $q_j = I\{j \leq L\}$. The formulas for 2SLS and for PC are exactly the same except that $P_n^\alpha$ is a projection matrix on the first principal components, while $P^L$ is the projection matrix on the first $L$ instruments. Clearly, the relative performance of the two methods will depend on whether the first instruments are informative or not, as illustrated in the Monte Carlo study.

## 4. Estimation of MSE

The aim is to find $\alpha$ that minimizes the conditional MSE of $v'\widehat{\delta}$ for some arbitrary $p \times 1$ vector $v$. The conditional MSE is

$$\begin{aligned} MSE &= E\left[v'\left(\widehat{\delta} - \delta_0\right)\left(\widehat{\delta} - \delta_0\right)' v | X\right] \\ &\sim v' S(\alpha) v \\ &\equiv S_v(\alpha). \end{aligned}$$

$S_v$ involves the function $f$, which is unknown. We need to replace $S_v$ by an estimate. First note that if $\delta \in \mathbf{R}^p$ for $p > 1$, the regression

$$W = f + u$$

involves $n \times p$ matrices. It is possible to reduce the dimension by post-multiplying by $H^{-1}v$:

$$\begin{aligned} &WH^{-1}v = fH^{-1}v + uH^{-1}v \Leftrightarrow \\ &W_v = f_v + u_v \end{aligned} \tag{10}$$

using obvious notation. Then, we are back to a univariate regression where $f_v$ is estimated by $P_n^\alpha W_v$. The results of Li (1986, 1987) apply, except that here $H$ is unknown and needs to be estimated so that $W_v$ itself is not observable.

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 $f'f/n$, $\tilde{H}$ may be $W' P_n^{\tilde{\alpha}} W/n$ 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 the subscript (1) refers to the first column). Let $\tilde{u} = \left(I - P_n^{\tilde{\alpha}}\right) W$, $\tilde{u}_v = \tilde{u}\tilde{H}^{-1}v$.

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

Note that none of these preliminary estimators depend on $\alpha$. Let $\hat{u}^\alpha = \left(I - P_n^\alpha\right) W$ and $\hat{u}_v^\alpha = \hat{u}^\alpha \tilde{H}^{-1} v$. Let $\hat{q}_j = q\left(\alpha, \hat{\lambda}_j^2\right)$.

We consider the following goodness-of-fit criteria:

*Mallows $C_p$* (Mallows, 1973):

$$\hat{R}^m(\alpha) = \frac{\hat{u}_v^{\alpha\prime}\hat{u}_v^\alpha}{n} + 2\hat{\sigma}_{u_v}^2 \frac{\mathrm{tr}\left(P_n^\alpha\right)}{n} = \frac{\hat{u}_v^{\alpha\prime}\hat{u}_v^\alpha}{n} + 2\hat{\sigma}_{u_v}^2 \frac{\sum_j \hat{q}_j}{n}.$$

*Generalized cross-validation* (Craven and Wahba, 1979):

$$\hat{R}^{cv}(\alpha) = \frac{1}{n}\frac{\hat{u}_v^{\alpha\prime}\hat{u}_v^\alpha}{\left(1 - \frac{\mathrm{tr}\left(P_n^\alpha\right)}{n}\right)^2} = \frac{1}{n}\frac{\hat{u}_v^{\alpha\prime}\hat{u}_v^\alpha}{\left(1 - \frac{\sum_j q_j}{n}\right)^2}.$$

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

$$\hat{R}^{lcv}(\alpha) = \sum_{i=1}^{n}\left(\tilde{W}_{v_i} - \hat{f}_{v_{-i}}^\alpha\right)^2 \frac{1}{n},$$

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

As first stage criterion, the two cross-validation methods are preferable to Mallows' $C_p$ because they do not require computing $\hat{\sigma}_{u_v}^2$. The leave-one-out cross-validation is more burdensome to implement than the generalized cross-validation. The second criterion slightly differs from that taken by Donald and Newey (2001) and the third criterion was absent from that paper but they both can be found in Li (1986, 1987).

The *approximate MSE* of $v'\widehat{\delta}$ is given by

$$\hat{S}_v(\alpha) = \hat{\sigma}_{u_v\varepsilon}^2 \frac{\left(\sum_j \hat{q}_j\right)^2}{n} + \hat{\sigma}_\varepsilon^2\left[\hat{R}(\alpha) - \hat{\sigma}_{u_v}^2 \frac{\mathrm{tr}\left(\left(P_n^\alpha\right)^2\right)}{n}\right], \tag{11}$$

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

To see where this expression comes from, note that $S_v(\alpha)$ can be rewritten as

$$\begin{aligned} S_v(\alpha) &= v' H^{-1}\sigma_{u\varepsilon}\sigma_{u\varepsilon}' H^{-1} v \frac{\left(\sum_j \hat{q}_j\right)^2}{n} + \sigma_\varepsilon^2 \frac{f_v'\left(I - P_n^\alpha\right)^2 f_v}{n} \\ &= \sigma_{u_v\varepsilon}^2 \frac{\left(\sum_j \hat{q}_j\right)^2}{n} + \sigma_\varepsilon^2 \frac{f_v'\left(I - P_n^\alpha\right)^2 f_v}{n}, \end{aligned} \tag{12}$$

where $\sigma_{u_v\varepsilon} = E\left[\varepsilon_i u_i' H^{-1} v | x_i\right]$. Using Li's results on $C_p$ and cross-validation procedures for selecting $\alpha$ in the regression (10), $\hat{R}(\alpha)$ approximates to

$$R_v(\alpha) = \frac{f_v'\left(I - P_n^\alpha\right)^2 f_v}{n} + \sigma_{u_v}^2 \frac{\mathrm{tr}\left(\left(P_n^\alpha\right)^2\right)}{n}.$$

Hence replacing $f_v'\left(I-P_n^\alpha\right)^2 f_v/n$ by $\hat{R}(\alpha)-\hat{\sigma}_{u_v}^2 \operatorname{tr}\left(\left(P_n^\alpha\right)^2\right)/n$ in (12) provides an estimate of $S_v(\alpha)$. Note that Li (1987) focuses on discrete index sets. Hence his results apply directly for LF, SC, and PC. For T, where the index set is continuous, we can use results on the ridge regression and spline smoothing (Craven and Wahba, 1979; Golub et al., 1979; Li, 1986). The optimality of this selection rule could be established using the results of Li (1986, 1987).

**Remark.** The quality of the selection of $\alpha$ may be affected by a poor estimation of $\sigma_{u_v\varepsilon}$, $\sigma_\varepsilon^2$, and $\sigma_{u_v}^2$. In particular, it may be desirable to avoid deriving a first-step estimator of $\alpha$, $\tilde{\alpha}$. A solution is to consider the MSE for $v$ such that $H^{-1}v$ equals the unit vector $e$. This choice is perfectly fine as $v$ is arbitrary and permits us to simplify some of the derivations. Indeed, $W_v = We$, $f_v = fe$, and $u_v = ue$. Moreover, $\sigma_{u_v\varepsilon}$ can be estimated[2] by $\hat{\sigma}_{u_v\varepsilon} = W_v'\varepsilon/n$. Finally, the estimation of $\sigma_{u_v}^2$ can be avoided completely by noticing that the term $\sigma_{u_v}^2 \frac{\operatorname{tr}\left(\left(P_n^\alpha\right)^2\right)}{n}$ is negligible with respect to $\sigma_{u_v\varepsilon}^2 \frac{\left(tr\left(P_n^\alpha\right)\right)^2}{n}$. Hence, $\hat{S}_v(\alpha)$ can be replaced by

$$\hat{S}_v(\alpha) = \hat{\sigma}_{u_v\varepsilon}^2 \frac{\left(\sum_j \hat{q}_j\right)^2}{n} + \hat{\sigma}_\varepsilon^2 \hat{R}(\alpha).$$

## 5. Comparison with alternative estimators

In this section, we examine some familiar estimators and try to establish whether they involve some regularization of the covariance matrix.

### 5.1. k-class estimators and LIML

A large class of estimators called the $k$-class was proposed by Theil (1958). The estimators of this class take the following form

$$\hat{\beta} = \left(W'\left(I-\mathrm{k}M_Z\right)W\right)^{-1} W'\left(I-\mathrm{k}M_Z\right)y$$

with $\mathrm{k} \geq 1$ and $M_Z = I - P_Z$ where $P_Z$ is the projection operator on all the instruments $Z$. These estimators are consistent provided $p\lim \sqrt{n}(\mathrm{k}-1) = 0$. Replacing $M_Z$ by $I - P_Z$, we obtain

$$\begin{aligned}\hat{\beta} &= \left(W'\left(P_Z - \frac{\mathrm{k}-1}{\mathrm{k}} I\right)W\right)^{-1} W'\left(P_Z - \frac{\mathrm{k}-1}{\mathrm{k}} I\right)y \\ &= \left(W'P_n^\alpha W\right)^{-1} W'P_n^\alpha y\end{aligned}$$

with $P_n^\alpha = P_Z - \alpha I$ and $\alpha = (\mathrm{k}-1)/\mathrm{k}$. We can see that if $\mathrm{k} \to 1$, then $\alpha \to 0$. The limited information maximum likelihood (LIML) estimator is a member of this class with

$$\alpha = \min \frac{(y-W\delta)' P_Z (y-W\delta)}{(y-W\delta)'(y-W\delta)}.$$

Note that $trace\left(P_n^\alpha\right) = trace(P_Z) - \alpha trace(I_n) = L - \alpha n$. Hence if $\alpha$ is such that $L - \alpha n$ does not depend on $L$, then the higher-order bias of $\hat{\delta}$ does not depend on the number of instruments. This is the case for the bias-adjusted 2SLS estimator and for the LIML (see DN).

An interesting question is the following: Do the estimators of the $k$-class involve a regularization of the covariance matrix $K = \underline{Z}'\underline{Z}/n$ (using the notation of Section 2.1)? More precisely, can we rewrite $P_n^\alpha$ as

$$P_n^\alpha = \underline{Z}\left(\underline{Z}'\underline{Z}\right)^{-1}\underline{Z}' - \alpha I_n = \underline{Z}' Q^\alpha \underline{Z}/n \tag{13}$$

for some $L \times L$ matrix $Q^\alpha$ which could then be interpreted as a regularized inverse of $K$? The answer is no because Eq. (13) does not have a solution when $L < n$. On the other hand, $\left(\underline{Z}'\underline{Z}\right)^{-1}$ can not be computed when $L > n$. So, the $k$-class estimation procedure can not be thought of as a regularization technique.

### 5.2. Methods based on a truncation of the instruments sequence

Consider a countable infinite sequence of instruments $(z_1, z_2, z_3, \ldots)$, then a method based on a truncation exploits only the first $L$ instruments $(z_1, z_2, \ldots, z_L)$. This method is used in DN and is popular when the instruments are lagged values of a variable (see Kuersteiner, 2001, 2006, 2012). Here, we show that this approach corresponds to a regularization scheme called the projection method (see Engl et al., 2000; Böttcher, 1996). First, we outline the projection method as described in Böttcher (1996).

Let $K$ be an infinite dimensional matrix on $l^2$ with $\left(l, l'\right)$ element $E\left(z_l z_{l'}\right)$ (here $z_l$ is assumed to be scalar to simplify notations). We replace the initial problem, $K\varphi = f$ i.e. the infinite system

$$\begin{pmatrix} E\left(z_1^2\right) & E\left(z_1 z_2\right) & \ldots \\ E\left(z_2 z_1\right) & E\left(z_2^2\right) & \ldots \\ \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \end{pmatrix} \tag{14}$$

by a truncated system

$$\begin{pmatrix} E\left(z_1^2\right) & \ldots & E\left(z_1 z_L\right) \\ \vdots & \ddots & \vdots \\ E\left(z_L z_1\right) & \ldots & E\left(z_L z_L\right) \end{pmatrix} \begin{pmatrix} \varphi_1^{(L)} \\ \vdots \\ \varphi_L^{(L)} \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_L \end{pmatrix}. \tag{15}$$

This substitution is valid if the solution $\varphi^{(L)}$ of (15) approximates correctly the solution $\varphi$ of (14). We introduce the projection $P_L$ defined by

$$P_L : \{\varphi_1, \varphi_2, \varphi_3, \ldots\} \to \{\varphi_1, \varphi_2, \ldots \varphi_L, 0, 0, \ldots\}.$$

System (15) can be rewritten as

$$P_L K P_L \varphi^{(L)} = P_L f,$$

where $\varphi^{(L)}$ belongs to the image (range) of $P_L$, denoted $\mathrm{Im}P_L$. Let $K_L$ be the restriction of $P_L K P_L$ to $\mathrm{Im}P_L$. It turns out that if $K$ is a self-adjoint operator with eigenvalues strictly bounded away from zero, then $K_L$ is invertible for all $L$ and $K_L^{-1}$ approaches $K^{-1}$ as $L$ goes to infinity (see Proposition 3.1. and Theorem 4.1 of Böttcher, 1996). On the other hand, if $K$ is a compact operator on $l^2$, $K$ is not invertible (on the whole space $l^2$). In this case, truncation is generally not sufficient to stabilize the inverse, some other form of regularization such as Tikhonov should be combined with the truncation (see Section 5.2 of Engl et al., 2000). In summary, $K_L^{-1}$ satisfies the condition of a regularized inverse when $K$ is not compact, but in this case, the problem (14) is not ill-posed, so that a regularization is not really needed and is used only to approximate the solution.[3]

Kuersteiner (2001) provides an example of a covariance matrix which is not compact. Its instruments are lagged values of $\varepsilon_t$, the error term in an ARMA model. In the simplest case where $\{\varepsilon_t\}$ are independent, $K$ is the identity matrix, then obviously $K$ is invertible with $K^{-1} = K$. $K$ remains invertible for weakly dependent $\{\varepsilon_t\}$.

DN study versions of 2SLS, bias-adjusted 2SLS and LIML estimators where the sequence of instruments is truncated. These estimators involve a projection regularization where the regularization parameter is $\alpha = 1/L$.

[2] The author thanks Whitney Newey for pointing this out.

[3] The projection method may be applied to compact operators if the problem is not too severely ill-posed i.e. if the eigenvalues of $K_L$ do not converge to zero too fast. See Engl et al. (2000).

Now, we turn our attention toward Kuersteiner's kernel weighted GMM estimator. Kuersteiner (2012) considers estimating an MA($\infty$) model using lags of the dependent variable $y_t$ as instruments $(y_{t-1}, y_{t-2}, \ldots)$. To simplify the exposition, we assume $y_t$ scalar. The covariance matrix $K$ is assumed to be invertible. Kuersteiner's estimator is a standard GMM estimator which is based on the first $L$ moment conditions but uses as weighting matrix

$$\Sigma_L = V_L K_L^{-1} V_L,$$

where $K_L$ has been defined above and $V_L$ is the diagonal matrix with $j$th diagonal element $k(j/L)$ where $k$ is a kernel function satisfying $k(0) = 1$. The idea is that each instrument $y_{t-j}$ is weighted by $k(j/L)$. Under the assumptions of Kuersteiner (2012), $\Sigma_L \varphi \to K^{-1}\varphi$ as $L$ goes to infinity. Hence, Kuersteiner's kernel weighted GMM estimator involves a regularized inverse of $K$ with regularization parameter $\alpha = 1/L$.

In conclusion, while LIML is not a regularized estimator, the DN truncation technique and Kuersteiner (2012)'s approach can be regarded as regularizations.

## 6. Monte Carlo experiments

We illustrate the quality of our estimators on a basic model of the form

$$y_i = \delta W_i + \varepsilon_i, \qquad (16)$$
$$W_i = f(x_i) + u_i$$

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

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

*Model* 1. The first specification of (16) is taken from Model 1 of DN[4]

$f(x_i) = x_i'\pi$

with $x_i \sim \mathcal{N}(0, I_L)$, $L = 20$, $R_f^2 = \pi'\pi/(1+\pi'\pi)$ is set equal to 0.1. The $x_i$ are used as instruments so that $z_i = x_i$. As the instruments are independent from each other, this example corresponds to the worst case scenario for our regularized estimators. Indeed, here all the eigenvalues of $K$ are equal to 1, so there is no information contained in the spectral decomposition of $K$. Moreover, if $L$ were infinite, $K$ would not be compact and consequently not Hilbert–Schmidt, hence our method would not apply. However, in practical applications, it is not plausible that a large number of instruments would be uncorrelated with each other.

*Model* 1a. $\pi_l = d(1 - l/(L+1))^4, l = 1, 2, \ldots, L$, where the constant $d$ is chosen so that $\pi'\pi = R_f^2/(1-R_f^2)$. The instruments are ordered in decreasing order of importance. We use this example to compare our estimators with DN. First, we compute the 2SLS estimator with optimal selection of instruments (denoted by DN in the table). We also report a second estimator (DN-R) that uses the same technique of estimation as DN but where the order of the instruments has been reversed. This means that now the worst instruments are selected first. This illustrates the pitfall associated with an a priori ranking of the instruments.

*Model* 1b. $\pi_l = \sqrt{R_f^2/L(1-R_f^2)}, l = 1, 2, \ldots, L$. In this case, there is no reason to prefer one instrument over another.

*Model* 2 (*Factor model*).

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

where $f_i = (f_{i1}, f_{i2}, f_{i3})' \sim iid\mathcal{N}(0, I_3)$ and $x_i$ is a $30 \times 1$ vector of instruments constructed from $f_i$ through

$x_i = Mf_i + v_i,$

where $v_i \sim iid\mathcal{N}(0, \sigma_v^2 I_{30})$ with $\sigma_v = 0.3$ and $M$ is a $30 \times 3$ matrix, the elements of which are independently drawn in a $U[-1, 1]$. In our Monte Carlo experiment, $M$ is the same for all simulations. As mentioned in Section 3, $\{\psi_j\}$ for $j = 1, 2, 3$ are normalized estimates of the factors. Hence, for PC, $P_n^\alpha W$ is an estimator of $(f_{i1} + f_{i2} + f_{i3})$. It is therefore expected that the PC estimator will be close to the instrumental variable estimator that uses the unobservable $(f_{i1} + f_{i2} + f_{i3})$ as instrument. For comparison, we report the unfeasible IV estimator in Table 1.

**Table 1**
Summary Statistics on $\hat{\delta}$.

| | | MSE | Med.bias | Med.abs | Dis. | Cov. |
|---|---|---|---|---|---|---|
| Model 1a | T | 0.0277 | 0.1275 | 0.1306 | 0.2860 | 0.760 |
| | LF | 0.0276 | 0.1231 | 0.1239 | 0.2999 | 0.774 |
| | PC | 9.2223 | 0.1345 | 0.1595 | 0.4054 | 0.778 |
| | DN | 0.0213 | 0.0571 | 0.1024 | 0.3464 | 0.900 |
| | DN-R | 66173 | 0.4391 | 0.6851 | 3.5290 | 0.665 |
| | IV | 0.0200 | 0.0025 | 0.0919 | 0.3545 | 0.957 |
| Model 1b | T | 0.0276 | 0.1257 | 0.1299 | 0.2796 | 0.765 |
| | LF | 0.0274 | 0.1201 | 0.1271 | 0.2952 | 0.779 |
| | PC | 3.9071 | 0.1334 | 0.1587 | 0.4085 | 0.791 |
| | DN | 8.9139 | 0.1644 | 0.1816 | 0.4068 | 0.742 |
| | IV | 0.0197 | -0.0010 | 0.0895 | 0.3461 | 0.959 |
| Model 2 | T | 0.0006 | 0.0012 | 0.0172 | 0.0660 | 0.958 |
| | LF | 0.0006 | 0.0003 | 0.0171 | 0.0661 | 0.957 |
| | PC | 0.0006 | 0.0011 | 0.0173 | 0.0661 | 0.957 |
| | DN | 0.0007 | 0.0029 | 0.0176 | 0.0662 | 0.953 |
| | IV | 0.0006 | 0.0009 | 0.0172 | 0.0653 | 0.957 |
| Model 3 | T | 0.0287 | 0.0396 | 0.1134 | 0.4176 | 0.933 |
| | LF | 0.0290 | 0.0269 | 0.1130 | 0.4216 | 0.946 |
| | PC | 0.0310 | 0.0471 | 0.1178 | 0.4271 | 0.917 |
| | DN | 0.0325 | 0.0617 | 0.1224 | 0.4255 | 0.905 |
| | IV | 0.0314 | 0.0052 | 0.1115 | 0.4352 | 0.954 |

*Model* 3 (*Functional form*).

Now, $f(x)$ in (16) takes the form[5]

$$f(x) = 126\left(\frac{x}{2\pi}\right)^4\left[1 - \left(\frac{x}{2\pi}\right)^4\right]$$

and $x_i$ are iid uniform on $[0, 2\pi]$. We use a continuum of moment conditions $Z(\tau, x_i) = \exp(i\tau x_i)$ and the weight $\pi$ is set equal to the density of a standard normal. The estimation is performed using the eigenvalues and eigenfunctions of the $n \times n$ matrix $TT^*$ and formula (3). We obtain estimators of the function $f$:

$$\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{n} q\left(\alpha, \hat{\lambda}_j^2\right)\left(\psi_j' W\right)\psi_j(x)$$

using samples of size $n = 500$ and Tikhonov regularization. In Fig. 1, we plot the median, the 5th and 95th percentiles of $\hat{f}$ over 1000 simulations. The function appears oversmoothed which is not surprising because we use the value of $\alpha$ that minimizes the MISE of $\hat{\delta}$, it is bigger than the $\alpha$ that minimizes the cross-validation criterion for $f$. We compare our regularized 2SLS estimators with the 2SLS estimator that uses power functions of $x_i$ as instruments, $Z(\tau, x_i) = x_i^\tau, \tau = 0, 1, 2, \ldots$ and where the number of instruments is optimally selected. The results concerning this method are reported in row "DN".

The simulations are performed using 5000 replications of samples of size $n = 500$. We compute the estimators corresponding to Tikhonov (T), Landweber–Fridman (LF), principal components (PC) regularizations, 2SLS with optimal selection of instruments (DN), and unfeasible IV estimator that uses the true $f$ as instrument (IV).

[4] This choice of $f$ does not satisfy the uniform boundedness Assumption 3(iii). However, our estimation procedure is certainly robust to this violation.

[5] This function (with a different scaling factor) is used by Eastwood and Gallant (1991) in their simulations.

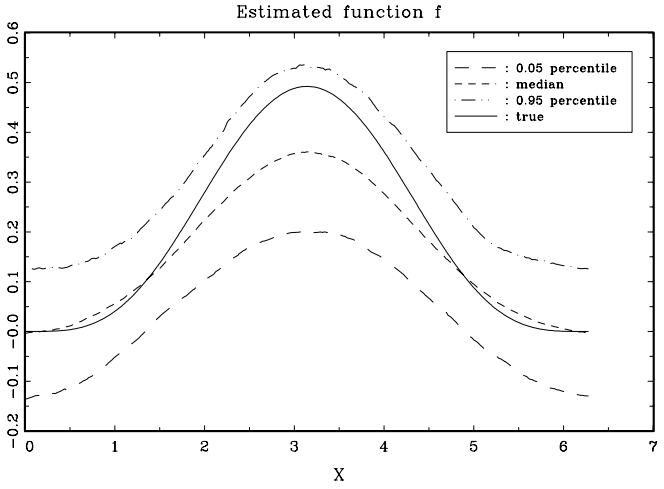


**Fig. 1.** Estimation by Tikhonov regularization of the function $f(x) = 126\left(\frac{x}{2\pi}\right)^4\left[1-\left(\frac{x}{2\pi}\right)^4\right]$. The sample size is $n = 500$ and the number of replications is 1000.

We selected the optimal value of $\alpha$ by minimizing the approximate MSE using the generalized cross-validation criterion and formula (11). Estimates of $\sigma_{u\varepsilon}$ and $\sigma_\varepsilon^2$ are computed using $\tilde{\delta}$ obtained with the value of $\alpha$, which minimizes the first stage cross-validation criterion. For the T regularization, we looked for the values of $\alpha$ in a 10 point grid, ranging from 0 to 0.2 for Models 1a and 1b, from 0.1 to 0.55 for Model 2. For Model 3, we investigate 16 values between 0.001 and 0.4. For LF, we searched among the number of iterations ranging from 1 to 30. Finally for PC, we searched among the number of eigenfunctions from 1 to 20. For LF, we used $c = 0.1/\lambda_1^2$ where $\lambda_1$ is the largest eigenvalue of $K_n$. The 0.1 coefficient is arbitrary, the theory just says that $c$ has to be smaller than $1/\|K\|^2$, we chose this value because it seemed to work best. For DN, the number of instruments is chosen between 1 and 20 by minimizing the MSE expression given in DN but using generalized cross-validation for $\hat{R}$ instead of cross-validation as in DN.

In Table 1, we report the mean square error (MSE), 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, and the coverage rate (Cov.) of a nominal 95% confidence interval. To obtain this last value, we compute an estimate of the variance of each estimator using the following formula

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

where $\widehat{W} = P_n^\alpha W$.

Table 2 contains summary statistics for the value of the regularization parameter which minimizes the approximate MSE. This regularization parameter is the number of instruments in DN, $\alpha \in (0, 1)$ for T, the number of iterations for LF, and the number of eigenfunctions for PC. We report the mean, standard error (std), mode, first, second and third quartile of the distribution of the regularization parameter.

First, we examine the results of Model 1a. As expected, the DN estimator using the right order of the instruments dominates all the other estimators. The T estimator always selects $\alpha = 0$. This is because the smallest eigenvalue is quite large and no regularization is needed. The T estimator boils down to a two-stage least-squares estimator using all the instruments. PC is by far the worse regularized estimator, its poor performance is mainly due to a few outliers. When the order of the instruments is reversed, DN does not perform well and is dominated by the regularized estimators. This makes sense since the regularized estimators are not affected by the order of the instruments. From this small experiment, we can conclude that if there is reliable information on the relative importance of the instruments, the DN approach should be preferred but in contrast if there is little information, it is preferable to use the regularized estimators.

**Table 2**
Properties of the distribution of the regularization parameters.

| | | Mean | Std | Mode | $q1$ | median | $q3$ |
|---|---|---|---|---|---|---|---|
| Model 1a | T | 0.000 | 0.000 | 0.00 | 0.00 | 0.00 | 0.00 |
| | LF | 13.83 | 7.09 | 10.0 | 9.00 | 12.0 | 17.0 |
| | PC | 9.588 | 5.325 | 5.00 | 5.00 | 9.00 | 14.0 |
| | DN | 1.958 | 2.208 | 1.00 | 1.00 | 1.00 | 2.00 |
| | DN-R | 2.856 | 3.864 | 1.00 | 1.00 | 1.00 | 3.00 |
| Model 1b | T | 0.000 | 0.000 | 0.00 | 0.00 | 0.00 | 0.00 |
| | LF | 13.82 | 7.070 | 10.0 | 9.00 | 12.0 | 18.0 |
| | PC | 9.585 | 5.311 | 6.00 | 5.00 | 9.00 | 14.0 |
| | DN | 1.971 | 2.294 | 1.00 | 1.00 | 1.00 | 2.00 |
| Model 2 | T | 0.154 | 0.017 | 0.15 | 0.15 | 0.15 | 0.15 |
| | LF | 30.00 | 0.000 | 30.0 | 30.0 | 30.0 | 30.00 |
| | PC | 2.600 | 0.878 | 3.00 | 2.00 | 2.00 | 3.00 |
| | DN | 8.144 | 1.132 | 8.00 | 8.00 | 8.00 | 8.00 |
| Model 3 | T | 0.019 | 0.013 | 0.01 | 0.01 | 0.01 | 0.03 |
| | LF | 28.36 | 3.524 | 30.0 | 30.0 | 30.0 | 30.0 |
| | PC | 3.186 | 1.515 | 3.00 | 3.00 | 3.00 | 3.00 |
| | DN | 4.007 | 1.563 | 3.00 | 3.00 | 3.00 | 5.00 |

For Model 1b where all the instruments have equal weights, DN estimator is worse and LF estimator is best.

Now, we consider Model 2. As expected the IV estimator dominates all the other methods, except in terms of coverage. There is no clear dominance among the other estimators. DN is dominated by PC for all measures, except for coverage.

For PC, in Models 1a and 1b, the selected number of eigenfunctions is large because all eigenfunctions are equally important. In contrast in Model 2, this number is much smaller and close to the number of factors (3).

In Model 3, DN performs worse in terms of MSE, median bias, MAD, and coverage. Tikhonov performs best in terms of MSE. The poor performance of 2SLS based on power series may be due to the fact that the function is "very nearly a finite Fourier series" (using the words of Eastwood and Gallant, 1991). It is an example where our method works better than DN.

## 7. Measuring the return to education

Although the benefit of education may seem obvious, it is difficult to measure it from the data. When regressing earnings on education, the OLS estimator might be biased because of the endogeneity of education. Indeed, education and earnings are likely be influenced by a common omitted variable, often referred to as "ability". Angrist and Krueger (1991) propose using 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. We use the same model and instruments as in Angrist and Krueger (1991, Table VII). Although some authors argued that these instruments were weak, Hansen et al. (2008) show that the poor performance of 2SLS is more likely to be related to a "many instruments" rather than "weak instruments" problem. The model we estimate is

$$logw = \alpha + \delta education + \beta_1' Y + \beta_2' S + \varepsilon, \tag{17}$$

where $logw$ = log of weekly wage, $education$ = years of education, $Y$ = year of birth dummy ($9 \times 1$), $S$ = state of birth dummy ($50 \times 1$). The vector of instruments, $Z$, includes 240 variables: the 60 included exogenous regressors plus 180 extra variables. More precisely, $Z = (1, Y, S, Q, QY, QS)$ where $Q$ = quarters of birth ($3 \times 1$), $QY$ = interaction between quarter of birth and year of

**Table 3**
Estimates of the return to education.

| OLS | 2SLS | Tikhonov | Landweber–Fridman | Principal component |
|---|---|---|---|---|
| 0.06665 | 0.0775 | 0.1027 | 0.1089 | 0.0827 |
| (0.0011) | (0.0136) | (0.0204) | (0.0360) | (0.0144) |
| | | $\alpha$ | Number of iterations | Number of eigenfunctions |
| | | 0.00012 | 700 | 210 |

birth ($27 \times 1$), $QS$ = interaction between quarter of birth and state of birth ($150 \times 1$). Angrist and Krueger (1991) uses a sample from the 1980 US Census that consisted of men born from 1930 to 1939. We use a random subsample of 10% of the original data. Our sample size is $n = 33{,}130$.

In Table 3, we report various estimates of $\delta$ and their standard errors (in parentheses). The regularization parameters selected by generalized cross-validation are reported at the bottom of Table 3. OLS and 2SLS estimates are included for comparison purposes. In Angrist and Krueger (Table VII, columns 1 and 2), these estimates and their standard errors are respectively 0.0673 (0.0003) and 0.0928 (0.0093). The difference from our results may be explained by the difference of samples. The coefficients we obtain by regularized 2SLS are larger than those obtained by OLS and 2SLS suggesting that these methods provide a bias correction. However, their standard errors are also larger. This illustrates the trade-off between bias and variance. There is not much collinearity among the instruments, the eigenvalues of the matrix $\underline{Z}'\underline{Z}$ decline very slowly and actually only five of them are close to zero. This explains why 210 eigenfunctions are retained by cross-validation for the principal component estimate and why this method does not seem to perform as well as the other two (remember that the principal component approach works best when there are only a few factors common to all instruments). The slow decline in the eigenvalues may also explain the large number of iterations needed for LF. Actually, 700 was the maximum number allowed but the estimates and standard errors did not change much between 400 and 700 iterations.

## 8. Conclusion and extensions

The originality of our approach is to give an estimation technique that works for a finite and infinite number of moment conditions. In particular, no assumption on the growth rate of the number of moments is needed. The regularized 2SLS estimate does not suffer from the bias that arises in standard 2SLS in the presence of many orthogonality conditions. Regularized 2SLS has also an interesting interpretation as a nonparametric estimation technique. We show that the GMM estimator that uses a continuum of instruments of the exponential form is actually equivalent to the IV estimator that uses a nonparametric estimator of the optimal instrument $f$. In this paper, we restricted ourselves to an iid homoscedastic setting and various extensions would be of interest.

*Extension to other nonparametric estimators*

We saw that the estimator of $f$ given by (3) is quite general because we have a lot of flexibility in the choice of the kernel $\widetilde{k}$. Moreover, $\hat{W}$ belongs to the class of linear (in $W$) estimators. We believe that the rule for selecting the optimal smoothing parameter derived in Section 4 applies to general linear estimators of the form

$$\hat{W} = P_n^{\alpha} W,$$

where $P_n^{\alpha}$ is an $n \times n$ symmetric matrix. Kernel and series estimators are of this type. We expect to get the same expression for $S(\alpha)$ in Proposition 2 provided $P_n^{\alpha}$ satisfies some minimal conditions.

*Extension to heteroscedasticity*

If the assumption of homoscedastic errors $\varepsilon_i$ is relaxed, the GMM estimator of $\delta$ should use the heteroscedasticity-robust version of the weighting matrix. Let $h_i(\tau, \delta) = \left(y_i - W_i'\delta\right) Z_i$, the covariance operator of $h_i$ is the operator which associates with function $g$ of $L^2(\pi)$ the following function of $L^2(\pi)$:

$$(Cg)(\tau_1) = \int E\left(h_i(\tau_1, \delta)\, h_i(\tau_2, \delta)\right) g(\tau_2)\, \pi(\tau_2)\, d\tau_2.$$

Its sample counterpart

$$(C_n g)(\tau_1) = \frac{1}{n}\sum_{i=1}^{n} \int h_i(\tau_1, \delta)\, h_i(\tau_2, \delta)\, g(\tau_2)\, \pi(\tau_2)\, d\tau_2 \tag{18}$$

depends on the parameter $\delta$. One could replace the unknown $\delta$ by a first step consistent estimator $\hat{\delta}^1$ and then minimize the objective function using as weighting operator the regularized inverse of

$$(C_n g)(\tau_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 Z_i(\tau_1) \int Z_i(\tau_2)\, g(\tau_2)\, \pi(\tau_2)\, d\tau_2,$$

where $\hat{\varepsilon}_i = y_i - W_i'\hat{\delta}^1$. The resulting estimator will be asymptotically efficient. However, the study of its MSE will be complicated by the presence of $\hat{\varepsilon}_i$. Alternatively, a continuous-updating estimator (CUE) can be obtained by minimizing

$$\left\langle \left(C_n^{\alpha}\right)^{-1/2} h_n(\cdot, \delta),\, \left(C_n^{\alpha}\right)^{-1/2} h_n(\cdot, \delta) \right\rangle,$$

where $C_n$ defined in (18) depends on $\delta$. Such an estimator is investigated by Chaussé (2009). The GMM estimator is also implementable in a time-series context as shown in Carrasco et al. (2007). But again, the MSE would be difficult to derive.

*Bias correction*

The introduction of the regularization parameter $\alpha$ permits to reduce the bias of $\hat{\delta}$ in comparison to standard GMM. However, a bias remains in small samples. The formula derived in Section 4 could be used to remove the bias from the estimator $\hat{\delta}$. We could also develop versions of LIML that allow for an infinite number of instruments.

## Acknowledgments

An earlier version of this paper was entitled "Instrumental variables estimator based on principal components". The author would like to thank the coeditors of the GMM special issues, two referees, and Pierre Chaussé, Nikolay Gospodinov, and Guido Kuersteiner for helpful comments. She also benefited from comments by the participants of CEME (Cambridge, 2005), ESEM (Vienna, 2006), CESG (Niagara Falls, 2006), Econometric Society Meeting (Chicago, 2007), CIREQ GMM conference (Montreal, 2007), and the seminar participants at Boston University, the University of Chicago, Indiana University, Harvard–MIT, and HEC Genève. She also gratefully acknowledges partial financial support from SSHRC.

## Appendix. Proofs

To prove Proposition 1, we need the following preliminary result.

**Lemma 3.** *Consider $g$ and $g_n$ such that $\|g_n - g\| = O_p\left(1/\sqrt{n}\right)$.*

(i) *If $g \in$ Range $\left(K^{1/2}\right)$, then*

$$\left\|\left(K_n^\alpha\right)^{-1/2} g_n - K^{-1/2} g\right\| \to 0$$

*in probability as $n$, $n\alpha^{1/2}$ go to infinity and $\alpha$ goes to zero.*

(ii) *If $g \in$ Range $(K)$, then*

$$\left\|\left(K_n^\alpha\right)^{-1} g_n - K^{-1} g\right\| \to 0$$

*in probability as $n$, $n\alpha$ go to infinity and $\alpha$ goes to zero.*

**Proof of Lemma 3.** First, we give a detailed proof for (ii). To simplify the notation, we denote $B = K^{-1}$, $B^\alpha = (K^\alpha)^{-1}$ and $B_n^\alpha = \left(K_n^\alpha\right)^{-1}$. We define $K^\alpha$ as the generalized inverse of $(K^\alpha)^{-1}$:

$$(K^\alpha) g = \sum_{j/q\neq 0} \frac{\lambda_j}{q\left(\alpha, \lambda_j^2\right)} \langle g, \phi_j\rangle \phi_j.$$

We have

$$\begin{aligned}
&\left\|B_n^\alpha g_n - Bg\right\| \\
&\le \left\|B_n^\alpha g_n - B_n^\alpha g\right\| && \text{(A.1)}\\
&\quad + \left\|B_n^\alpha g - B^\alpha g\right\| && \text{(A.2)}\\
&\quad + \left\|B^\alpha g - Bg\right\|. && \text{(A.3)}
\end{aligned}$$

Term (A.1). $\left\|B_n^\alpha g_n - B_n^\alpha g\right\| \le \left\|B_n^\alpha\right\| \|g_n - g\|$. Moreover, for any $g$ with $\|g\| \le 1$, we have $\|B^\alpha g\|^2 \le \sum_j \frac{q^2\left(\alpha,\lambda_j^2\right)}{\lambda_j^2} \langle g, \phi_j\rangle^2 \le \sup_\lambda \frac{q^2(\alpha,\lambda^2)}{\lambda^2} \sum_j \langle g, \phi_j\rangle^2 \le \sup_\lambda \frac{q(\alpha,\lambda^2)}{\lambda^2}$ because $|q| \le 1$. Note that for the three regularizations, $q\left(\alpha, \lambda^2\right) \le c\lambda^2/\alpha$ for some positive constant $c$, see Kress (1999) and Carrasco et al. (2007, Section 3.3). Hence, $\|B^\alpha\|^2 \le c/\alpha$. The same inequality holds for $\left\|B_n^\alpha\right\|^2$. Therefore, $\left\|B_n^\alpha g_n - B_n^\alpha g\right\| = O_p\left(1/\left(\sqrt{\alpha n}\right)\right)$.

Term (A.2). $\left\|\left(B_n^\alpha - B^\alpha\right) g\right\| = \left\|B_n^\alpha \left(K_n^\alpha - K^\alpha\right) B^\alpha g\right\| \le \left\|B_n^\alpha\right\| \left\|K_n^\alpha - K^\alpha\right\| \|B^\alpha g\|$ where the first inequality follows from $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$. We examine each term of the right-hand side individually. As before, we have $\left\|B_n^\alpha\right\| \le c/\sqrt{\alpha}$. The second term can be rewritten as follows.

$$\begin{aligned}
&\left(K_n^\alpha - K^\alpha\right) g \\
&= (K_n - K) g + \left(K_n^\alpha - K_n\right) g + (K - K^\alpha) g \\
&= (K_n - K) g + \sum_{j/q\neq 0} \hat\lambda_j \left(\frac{1}{q\left(\alpha, \hat\lambda_j^2\right)} - 1\right) \left\langle g, \hat\phi_j\right\rangle \hat\phi_j \\
&\quad + \sum_{j/q\neq 0} \lambda_j \left(1 - \frac{1}{q\left(\alpha, \lambda_j^2\right)}\right) \langle g, \phi_j\rangle \phi_j.
\end{aligned}$$

$$\begin{aligned}
&\left\|\left(K_n^\alpha - K^\alpha\right) g\right\|^2 \\
&\le 3\left\|(K_n - K) g\right\|^2 \\
&\quad + 3 \sum_{j/q\neq 0} \hat\lambda_j^2 \left(\frac{1 - q\left(\alpha, \hat\lambda_j^2\right)}{q\left(\alpha, \hat\lambda_j^2\right)}\right)^2 \left\langle g, \hat\phi_j\right\rangle^2 \\
&\quad + 3 \sum_{j/q\neq 0} \lambda_j^2 \left(\frac{q\left(\alpha, \lambda_j^2\right) - 1}{q\left(\alpha, \lambda_j^2\right)}\right)^2 \langle g, \phi_j\rangle^2.
\end{aligned}$$

We have

$$\begin{aligned}
&\sum_{j/q\neq 0} \lambda_j^2 \left(\frac{q\left(\alpha, \lambda_j^2\right) - 1}{q\left(\alpha, \lambda_j^2\right)}\right)^2 \langle g, \phi_j\rangle^2 \\
&\le \sup_{j/q\neq 0} \frac{\lambda_j^2}{q\left(\alpha, \lambda_j^2\right)} \sum_j \left(q\left(\alpha, \lambda_j^2\right) - 1\right)^2 \langle g, \phi_j\rangle^2.
\end{aligned}$$

The term $\sum_j \left(q\left(\alpha, \lambda_j^2\right) - 1\right)^2 \langle g, \phi_j\rangle^2$ is the regularization bias and is $O(\alpha)$ by Proposition 3.12 of Carrasco et al. (2007). Because $q\left(\alpha, \lambda_j^2\right)$ converges to 1 as $\alpha$ goes to zero, $q\left(\alpha, \lambda_j^2\right)$ is greater than some positive constant $d$ for $n$ large enough and hence, $\sup_{j/q\neq 0} \frac{\lambda_j^2}{q\left(\alpha,\lambda_j^2\right)} < \frac{1}{d} \sup_j \lambda_j^2$ which is bounded because $K$ is compact. Therefore, $\left\|K_n^\alpha - K^\alpha\right\| \le \|K_n - K\| + O_p(\alpha)$. We now turn our attention to the term $\|B^\alpha g\|$. Because $g$ is in the range of $K$, there exists a function $\varphi$ such that $g = K\varphi$ and $\|\varphi\| < \infty$. Using the fact that $0 \le q \le 1$, $\|B^\alpha g\|^2 = \|B^\alpha K\varphi\|^2 = \sum_j q\left(\alpha, \lambda_j^2\right)^2 \langle\varphi, \phi_j\rangle^2 \le \sum_j \langle\varphi, \phi_j\rangle^2 = \|\varphi\|^2 < \infty$. It follows that Term (A.2) goes to zero.

Term (A.3). $\|B^\alpha g - Bg\| = \|B^\alpha K\varphi - \varphi\| = \|(B^\alpha K - I)\varphi\| \to 0$ as $\alpha$ goes to zero by Kress (1999, Section 15.5) for the three regularizations. This concludes the proof of (ii).

For (i), we define now $B = K^{-1/2} = \left(K^{-1}\right)^{1/2}$, so that $Bg = \sum \frac{1}{\lambda_j^{1/2}} \langle g, \phi_j\rangle \phi_j$. Similarly, we define $B^\alpha = (K^\alpha)^{-1/2}$, $B_n^\alpha = \left(K_n^\alpha\right)^{-1/2}$ and $(K^\alpha)^{1/2}$ as the generalized inverse of $(K^\alpha)^{-1/2}$:

$$(K^\alpha)^{1/2} g = \sum_{j/q\neq 0} \frac{\lambda_j^{1/2}}{q\left(\alpha, \lambda_j^2\right)^{1/2}} \langle g, \phi_j\rangle \phi_j.$$

The proof is similar to that of (ii) with obvious adjustments, for instance $K$ is replaced by $K^{1/2}$. Now, we have $\|B^\alpha\|^2 \le \sqrt{c/\alpha}$, hence the different rate of convergence on $\alpha$. □

**Proof of Proposition 1.** *Consistency.* We can write[6]

$$\begin{aligned}
\hat\delta - \delta_0 &= \left(W' P_n^\alpha W\right)^{-1} W' P_n^\alpha \varepsilon \\
&= \left\langle \left(K_n^\alpha\right)^{-1/2} g_n, \left(K_n^\alpha\right)^{-1/2} g_n' \right\rangle^{-1} \\
&\quad \times \left\langle \left(K_n^\alpha\right)^{-1/2} g_n, \left(K_n^\alpha\right)^{-1/2} \hat E\left(Z_i(\tau)\varepsilon_i\right) \right\rangle,
\end{aligned}$$

where $g_n(\tau) = \hat E\left(Z_i(\tau) W_i\right)$. Consider a typical element of the vector $g_n$, namely $g_{na} = \hat E\left(Z_i(\tau) W_{ia}\right)$ where $W_{ia}$ is the $a$th element of $W_i$. Let $g_a(\tau) = E\left(Z_i(\tau) W_{ia}\right)$. By Lemma 3, $\left\|\left(K_n^\alpha\right)^{-1/2} g_{na}\right\| \to \left\|K^{-1/2} g_a\right\|$ provided $n, n\alpha^{1/2} \to \infty$ and $\alpha \to 0$ and $g_a \in$ *Range* $\left(K^{1/2}\right)$. Note that the range of $K^{1/2}$ is the reproducing kernel Hilbert space (RKHS) with kernel $k(\tau_1, \tau_2) = E\left(Z_i(\tau_1) Z_i(\tau_2)\right)$. For the definitions and properties of RKHS, see Berlinet and Thomas-Agnan (2004) and Carrasco et al. (2007). The RKHS is the Hilbert space of functions of the form $\sum_{j=1}^l \omega_j k\left(\tau_j, \cdot\right)$ and its limit. Assuming that $f_a(\cdot) = \sum_{j=1}^l \omega_j Z\left(\tau_j, \cdot\right)$, we can see that $E[Z(\tau, x) W_a] = E[Z(\tau, x) f_a(x)]$ belongs to the RKHS. We use the notation $\left\|K^{-1/2} g_a\right\| = \|g_a\|_K < \infty$ where $\|g_a\|_K$ denotes the norm of $g_a$ in the RKHS associated with $K$. Similarly, we denote by $\langle g_a, g_b\rangle_K$ the inner product in the RKHS. Under the assumptions of

[6] 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 $\langle g_a, h_b\rangle$, $a, b = 1, \ldots, p$.

Proposition 1, we have

$$\left\langle \left(K_n^{\alpha}\right)^{-1/2} g_n, \left(K_n^{\alpha}\right)^{-1/2} g_n' \right\rangle \xrightarrow{P} \langle g, g' \rangle_K ,$$
$$\left\langle \left(K_n^{\alpha}\right)^{-1/2} g_n, \left(K_n^{\alpha}\right)^{-1/2} \hat{E}\left(Z_i(\tau)\varepsilon_i\right) \right\rangle \xrightarrow{P} 0,$$

where $\langle g, g' \rangle_K$ is the $p \times p$ matrix with $(a, b)$ element $\langle K^{-1/2}E(Z(\cdot, x_i)W_{ia}), K^{-1/2}E(Z(\cdot, x_i)W_{ib})\rangle$, where $W_{ib}$ is the $b$th element of the vector $W_i$. This proves the consistency.

*Asymptotic normality.* We have

$$\begin{aligned}
&\sqrt{n}\left\langle \left(K_n^{\alpha}\right)^{-1/2} g_n, \left(K_n^{\alpha}\right)^{-1/2} \hat{E}\left(Z_i(\tau)\varepsilon_i\right) \right\rangle \\
&= \left\langle \left(K_n^{\alpha}\right)^{-1} g_n, \sqrt{n}\hat{E}\left(Z_i(\tau)\varepsilon_i\right) \right\rangle \\
&= \left\langle \left(K_n^{\alpha}\right)^{-1} g_n - K^{-1}g, \sqrt{n}\hat{E}\left(Z_i(\tau)\varepsilon_i\right) \right\rangle \\
&\quad + \left\langle K^{-1}g, \sqrt{n}\hat{E}\left(Z_i(\tau)\varepsilon_i\right) \right\rangle .
\end{aligned}$$

The first term is negligible because

$$\begin{aligned}
&\left\langle \left(K_n^{\alpha}\right)^{-1} g_n - K^{-1}g, \sqrt{n}\hat{E}\left(Z_i(\tau)\varepsilon_i\right) \right\rangle \\
&\le \left\| \left(K_n^{\alpha}\right)^{-1} g_n - K^{-1}g \right\| \left\| \sqrt{n}\hat{E}\left(Z_i(\tau)\varepsilon_i\right) \right\| \\
&= o_p(1)\, O_p(1) = o_p(1) .
\end{aligned}$$

By the functional central limit theorem, the second term satisfies

$$\left\langle K^{-1}g, \sqrt{n}\hat{E}\left(Z_i(\tau)\varepsilon_i\right) \right\rangle \xrightarrow{d} \mathcal{N}\left(0, \sigma_{\varepsilon}^2 \left\langle K^{-1}g, KK^{-1}g' \right\rangle\right) .$$

Its asymptotic variance can be rewritten as $\langle K^{-1}g, g' \rangle = \langle g, g' \rangle_K$. Hence, the three estimators satisfy

$$\sqrt{n}\left(\hat{\delta} - \delta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_{\varepsilon}^2 \langle g, g' \rangle_K^{-1}\right) . \tag{A.4}$$

*Efficiency.* The result (A.4) assumes only that $g_a = E\left(Z(\cdot, x_i) W_{ia}\right)$ belongs to the range of $K$. Let $L^2(Z)$ be the closure of the space spanned by $\{Z(x, \tau) : \tau \in I\}$, that is an element $g(x)$ of this space can be represented as $\sum_{j=1}^{q} \nu_j Z\left(x, \tau_j\right)$ or its limit. If $f(x)$ belongs to $L^2(Z)$, then we can compute explicitly the inner product in the RKHS and show that

$$\langle g_a, g_b \rangle_K = E(f_a f_b) .$$

To see this, we apply Theorem 6.4 of Carrasco et al. (2007). According to this theorem, $\|g_a\|_K^2 = E\left(G^2\right)$ where $G \in L^2(Z) \cap S$ with

$$S = \{G : g_a(\tau) = E\left(G(x_i) Z_a(\tau, x_i)\right)\} .$$

We see that $G_0(x_i) = f_a(x_i)$ is an element of $S$. Since moreover $f$ belongs to $L^2(Z)$, we have $\|g_a\|_K^2 = E\left(f_a^2\right)$. It follows that the asymptotic variance of $\sqrt{n}\left(\hat{\delta} - \delta_0\right)$ is $\sigma_{\varepsilon}^2 E\left(ff'\right)^{-1}$ which is the semiparametric efficiency bound. □

The proof of Proposition 2 is similar to that of DN. To simplify, we omit the hats on $\lambda_j$ and $\phi_j$ and we denote $P_n^{\alpha}$ and $q\left(\alpha, \lambda_j^2\right)$ by $P$ and $q_j$ in the following. We need the following preliminary results.

**Lemma 4.** *If Assumptions 1–3 are satisfied then:*

(i) $\sum_j q\left(\alpha, \lambda_j^2\right) = O(1/\alpha)$, $\sum_j q^2\left(\alpha, \lambda_j^2\right) = o\left(\left(\sum_j q_j\right)^2\right)$,
(ii) $h = f'\varepsilon/\sqrt{n} = O_p(1)$, $H = f'f/n = O_p(1)$.

**Proof of Lemma 4.** (i) For LF, $\sum_j q_j = \sum_j \left(1 - \left(1 - c\lambda_j^2\right)^{1/\alpha}\right) = O(1/\alpha)$.

For SC, $\sum_j q_j = \sum_j I\left(\lambda_j^2 \ge \alpha\right)\left(\lambda_j^2/\lambda_j^2\right) \le \sum_j \lambda_j^2/\alpha = O(1/\alpha)$ because $K$ is a Hilbert–Schmidt operator and hence $\sum_j \lambda_j^2 < \infty$.

For T, $\sum_j q_j = \sum_j \lambda_j^2/\left(\alpha + \lambda_j^2\right) \le \sum_j \lambda_j^2/\alpha = O(1/\alpha)$.

For PC, $\sum_j q_j = 1/\alpha$.

As $0 \le q_j \le 1$, we have $\sum_j q_j^2 \le \sup q_j \sum_j q_j \le \sum_j q_j$. Hence

$$\frac{\sum_j q_j^2}{\left(\sum_j q_j\right)^2} \le \frac{1}{\sum_j q_j} = o(1) .$$

Hence, $\sum_j q_j^2 = o\left(\left(\sum_j q_j\right)^2\right)$.

(ii) follows from the central limit theorem and the law of large numbers.

Let us denote $e_f(\alpha) = f'(I - P)f/n$, $e_{2f}(\alpha) = f'(I - P)^2 f/n$, $\Delta_{\alpha} = \mathrm{tr}\left(e_{2f}(\alpha)\right)$. □

**Lemma 5.** *If Assumptions 1–3 are satisfied then:*

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

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

(ii) $f'(I - P)\varepsilon/\sqrt{n} = O_p\left(\Delta_{\alpha}^{1/2}\right)$,
(iii) $u'P\varepsilon = O_p(1/\alpha)$,
(iv) $E\left[u'P\varepsilon\varepsilon'Pu|X\right] = \left(\sum_j q_j\right)^2 \sigma_{u\varepsilon}\sigma_{u\varepsilon}' + \left(\sum_j q_j^2\right)\left(\sigma_{u\varepsilon}\sigma_{u\varepsilon}' + \sigma_{\varepsilon}^2 \Sigma_u\right)$
$= \left(\sum_j q_j\right)^2 \sigma_{u\varepsilon}\sigma_{u\varepsilon}' + o_p\left(\left(\sum_j q_j\right)^2\right)$.
(v) $E\left[f'\varepsilon\varepsilon'Pu|X\right] = O_p(1/\alpha)$,
(vi) $\Delta_{\alpha}^{1/2}/\sqrt{\alpha n} \le 1/(2\alpha n) + \Delta_{\alpha}/2$,
(vii) $E\left[hh'H^{-1}u'f/n|X\right] = O_p(1/n)$,
(viii) $E\left[f'(I - P)\varepsilon\varepsilon'Pu/n|X\right] = O_p\left(\Delta_{\alpha}^{1/2}/\sqrt{\alpha n}\right)$.

**Proof of Lemma 5.** (i) Using $\langle f, \psi_j \rangle = \psi_j' f/n$, we have

$$\begin{aligned}
\frac{f'(I - P)f}{n} &= f' \sum_j \frac{(1 - q_j)}{n} \langle f, \psi_j \rangle \psi_j \\
&= \sum_j (1 - q_j) \langle f, \psi_j \rangle \langle f, \psi_j \rangle' .
\end{aligned}$$

Taking the trace, we obtain

$$\begin{aligned}
\mathrm{tr}\left(\frac{f'(I - P)f}{n}\right) &= \sum_{a=1}^{p} \sum_{j=1}^{n} (1 - \hat{q}_j) \left\langle f_a, \hat{\psi}_j \right\rangle^2 \\
&= \sum_{a=1}^{p} \sum_{j=1}^{n} \hat{\lambda}_j^{2\beta} (1 - \hat{q}_j) \frac{\left\langle f_a, \hat{\psi}_j \right\rangle^2}{\hat{\lambda}_j^{2\beta}} \\
&\le \sup_{\lambda_j} \lambda_j^{2\beta} (1 - q_j) \sum_{a=1}^{p} \sum_{j=1}^{n} \frac{\left\langle f_a, \hat{\psi}_j \right\rangle^2}{\hat{\lambda}_j^{2\beta}} .
\end{aligned}$$

Remark that

$$\sum_{j=1}^{n}\frac{\left\langle f_a,\hat{\psi}_j\right\rangle^2}{\hat{\lambda}_j^{2\beta}}=\sum_{j=1}^{n}\frac{\left\langle \hat{E}\left(Z\left(x_i,\cdot\right)f_a\left(x_i\right)\right),\hat{\phi}_j\right\rangle^2}{\hat{\lambda}_j^{2\beta+1}}, \tag{A.5}$$

where $\hat{E}\left(Z\left(x_i,\cdot\right)f_a\left(x_i\right)\right)=\sum_{i=1}^{n}Z\left(x_i,\cdot\right)f_a\left(x_i\right)/n$. At the limit, the sum in (A.5) is finite by Assumption 2(ii). Then, it is also true for $n$ sufficiently large. Hence, the rate of $\mathrm{tr}\left(f'\left(I-P\right)f/n\right)$ is given by $\sup_{\lambda_j}\lambda_j^{2\beta}\left(1-q_j\right)$.

For LF, $\sup_{\lambda_j}\lambda_j^{2\beta}\left(1-q_j\right)=\sup_{\mu}\mu^{\beta}\left(1-c\mu\right)^{1/\alpha}$. The maximum is reached for $\mu=\alpha\beta/\left(c\left(\alpha\beta+1\right)\right)$ and $\sup_{\lambda_j}\lambda_j^{2\beta}\left(1-q_j\right)=O\left(\alpha^{\beta}\right)$.

For SC, $\sup_{\lambda_j}\lambda_j^{2\beta}(1-q_j)=\sup_{\mu}\mu^{\beta}I\left(\mu<\alpha\right)=O(\alpha^{\beta})$.

For T, we need to distinguish between $\beta\leq 1$ and $\beta>1$. For $\beta>1$, the function $\lambda_j^{2\beta}(1-q_j)=\alpha\lambda_j^{2\beta}/(\lambda_j^2+\alpha)$ is increasing in $\lambda_j^2$ and reaches its maximum for the maximal eigenvalue (which is bounded by the Hilbert–Schmidt property of $K$). For $\beta\leq 1$, the function $\mu^{\beta}/\left(\mu+\alpha\right)$ reaches its maximum at $\mu=\alpha\beta/\left(1-\beta\right)$. Hence for any $\beta>0$, $\sup_{\lambda_j}\lambda_j^{2\beta}\left(1-q_j\right)=O\left(\alpha^{\min(\beta,1)}\right)$.

Similarly, the rate of $\mathrm{tr}\left(f'\left(I-P\right)^2f/n\right)$ is given by $\sup_{\lambda_j}\lambda_j^{2\beta}\left(1-q_j\right)^2$. This rate is studied in Carrasco et al. (2007, Proposition 3.11).

(ii)

$$\begin{aligned}\left\|f'\left(I-P\right)\frac{\varepsilon}{\sqrt{n}}\right\|^2&=\varepsilon'\left(I-P\right)ff'\left(I-P\right)\varepsilon/n\\&=\mathrm{tr}\left[\left(I-P\right)ff'\left(I-P\right)\varepsilon\varepsilon'/n\right],\end{aligned}$$

$$\begin{aligned}&E\left[\left\|f'\left(I-P\right)\frac{\varepsilon}{\sqrt{n}}\right\|^2|X\right]\\&=\sigma_{\varepsilon}^2\mathrm{tr}\left[\left(I-P\right)ff'\left(I-P\right)/n\right]\\&=\sigma_{\varepsilon}^2\mathrm{tr}\left[f'\left(I-P\right)^2f/n\right]\\&=\Delta_{\alpha}\sigma_{\varepsilon}^2.\end{aligned}$$

By Markov inequality, $\left\|f'\left(I-P\right)\varepsilon/\sqrt{n}\right\|^2=O_p\left(\Delta_{\alpha}\right)$ and $f'\left(I-P\right)\varepsilon/\sqrt{n}=O_p\left(\Delta_{\alpha}^{1/2}\right)$.

(iii) Let $u_a$ denote the $a$th column of $u$. As $P$ is positive semidefinite for the three regularizations considered here, the Cauchy–Schwarz inequality yields $u_a'P\varepsilon\leq\left((u_a'Pu_a\right)\left(\varepsilon'P\varepsilon\right))^{1/2}$ and

$$\begin{aligned}E\left[\left(\varepsilon'P\varepsilon\right)|X\right]&=\mathrm{tr}\left(PE\left(\varepsilon'\varepsilon|X\right)\right)\\&=\sigma_{\varepsilon}^2\mathrm{tr}\left(P\right)=\sigma_{\varepsilon}^2\sum_j q_j,\end{aligned}$$

so by Markov inequality $\varepsilon'P\varepsilon=O_p\left(\sum_j q_j\right)=O_p\left(1/\alpha\right)$. Similarly $u_a'Pu_a=O_p\left(\sum_j q_j\right)$, giving (iii).

(iv)

$$\begin{aligned}&u'P\varepsilon=n\sum_j q_j\left\langle\varepsilon,\psi_j\right\rangle\left\langle u,\psi_j\right\rangle.\\&\left(u'P\varepsilon\right)\left(\varepsilon'Pu\right)\\&=n^2\sum_{j,l}q_jq_l\left\langle\varepsilon,\psi_j\right\rangle\left\langle u,\psi_j\right\rangle\left\langle\varepsilon,\psi_l\right\rangle\left\langle u,\psi_l\right\rangle'\\&=\frac{1}{n^2}\sum_{j,l}q_jq_l\left(\varepsilon'\psi_j\right)\left(u'\psi_j\right)\left(\varepsilon'\psi_l\right)\left(u'\psi_l\right)'.\end{aligned}$$

By the serial independence of the $\{\varepsilon_i\}$ and $\{u_i\}$, we have

$$\begin{aligned}&E\left[\left(u'P\varepsilon\right)\left(\varepsilon'Pu\right)|X\right]\\&=\frac{1}{n^2}\sum_{j,l}q_jq_lE\left\{\sum_i\varepsilon_iu_i\psi_{i,j}^2\sum_b u_b'\varepsilon_b\psi_{b,j}^2\right.\\&\quad+\sum_c u_c\varepsilon_c\psi_{c,j}\psi_{c,l}\sum_i\varepsilon_iu_i'\psi_{i,j}\psi_{i,l}\\&\quad\left.+\sum_i\varepsilon_i^2\psi_{i,j}\psi_{i,l}\sum_c u_cu_c'\psi_{c,j}\psi_{c,l}\right\}\\&=\left(\sum_{j,l}q_jq_l\right)\sigma_{u\varepsilon}\sigma_{u\varepsilon}'+\sum_j q_j^2\left(\sigma_{u\varepsilon}\sigma_{u\varepsilon}'+\sigma_{\varepsilon}^2\Sigma_u\right)\\&=\left(\sum_j q_j\right)^2\sigma_{u\varepsilon}\sigma_{u\varepsilon}'+o_p\left(\left(\sum_j q_j\right)^2\right)\end{aligned}$$

by the fact that $(u_i,\varepsilon_i)$ are independent across $i$ and the eigenvectors are orthonormal, that is $\sum_i\psi_{i,j}\psi_{i,l}=0$ if $j\neq l$ and $=n\left\|\psi\right\|^2=n$ if $j=l$.

(v) The same proof as in DN (proof of Lemma A3(v)) can be used. $E\left[f'\varepsilon\varepsilon'Pu|X\right]=\sum_i f_iP_{ii}E\left(\varepsilon_i^2u_i'|x_i\right)$ and $\left\|\sum_i f_iP_{ii}E\left(\varepsilon_i^2u_i'|x_i\right)\right\|\leq\sum P_{ii}\left\|f_i\right\|\left\|E\left(\varepsilon_i^2u_i'|x_i\right)\right\|=O_p\left(\mathrm{tr}\left(P\right)\right)=O_p\left(\sum_j q_j\right)=O_p\left(1/\alpha\right)$.

(vi) The same proof as in DN (proof of Lemma A3(vi)) applies here with their $K$ replaced by $1/\alpha$.

(vii) follows from Lemma A3(vii) in DN.

(viii) The same proof as in DN (proof of Lemma A3(viii)) applies here with their $K$ replaced by $1/\alpha$, $\Delta_K$ by $\Delta_{\alpha}$ and $o_p$ replaced by $O_p$. We do not obtain $o$ because we do not impose $\max_i P_{ii}\to 0$ as in DN. □

**Proof of Proposition 2.** The proof is similar to that of DN. We have

$$\sqrt{n}\left(\hat{\delta}-\delta_0\right)=\hat{H}^{-1}\hat{h}$$

with

$$\hat{H}=W'PW,\qquad\hat{h}=W'P\varepsilon.$$

Observe that

$$\begin{aligned}\rho_{\alpha,n}&=\mathrm{tr}\left(S\left(\alpha\right)\right)\\&=\mathrm{tr}\left(H^{-1}\sigma_{u\varepsilon}\sigma_{u\varepsilon}'H^{-1}\right)\frac{\left(\sum_j q\left(\alpha,\lambda_j^2\right)\right)^2}{n}+\Delta_{\alpha}\\&=O_p\left(1/\left(\alpha^2n\right)\right)+\Delta_{\alpha}.\end{aligned}$$

We apply Lemma A1 of DN with

$$\begin{aligned}&h=f'\varepsilon/\sqrt{n},\\&T^h=T_1^h+T_2^h,\\&T_1^h=-f'\left(I-P\right)\varepsilon/\sqrt{n}=O_p\left(\Delta_{\alpha}^{1/2}\right),\\&T_2^h=u'P\varepsilon/\sqrt{n}=O_p\left(1/\left(\alpha\sqrt{n}\right)\right),\\&\hat{H}=H+T^H+Z^H,\qquad T^H=T_1^H+T_2^H\\&T_1^H=-f'\left(I-P\right)f/n=-e_f\left(\alpha\right)=O_p\left(\Delta_{\alpha}\right),\\&T_2^H=\left(u'f+f'u\right)/n=O_p\left(1/\sqrt{n}\right),\\&Z^H=\left(u'Pu-u'\left(I-P\right)f-f'\left(I-P\right)u\right)/n\\&\quad=O_p\left(1/\left(\alpha n\right)+\left(\Delta_{\alpha}/n\right)^{1/2}\right)=o_p\left(\rho_{\alpha,n}\right),\end{aligned}$$

where the last equality follows from Lemma 5(vi). As in DN, the terms $\left\|T_1^H\right\|^2$, $\left\|T_2^H\right\|^2$, $\left\|T_j^H\right\|\left\|T_l^h\right\|$ $(j,l=1,2)$ are $o_p\left(\rho_{\alpha,n}\right)$. We examine in details the term $\left\|T_1^H\right\|\left\|T_2^h\right\|$. We have $\left\|T_1^H\right\|\left\|T_2^h\right\| = O_p\left(\Delta_\alpha/\left(\alpha\sqrt{n}\right)\right)$. To compare $\Delta_\alpha/\left(\alpha\sqrt{n}\right)$ and $\rho_{\alpha,n}$, we use the fact that[7] $\rho_{\alpha,n} \sim 1/\left(\alpha^2 n\right)+\alpha^\beta$ is minimized for $\alpha = n^{-1/(2+\beta)}$, in which case $\rho_{\alpha,n} \sim n^{-\beta/(2+\beta)}$. On the other hand, $\Delta_\alpha/\left(\alpha\sqrt{n}\right) \sim \alpha^{\beta-1}/\sqrt{n} = n^{-3\beta/(2(2+\beta))} = o\left(n^{-\beta/(2+\beta)}\right) = o_p\left(\rho_{\alpha,n}\right)$. For Tikhonov, when $\beta > 2$, we have $\rho_{\alpha,n} \sim 1/\left(\alpha^2 n\right)+\alpha^2$ which is minimized for $\alpha = n^{-1/4}$, in which case $\rho_{\alpha,n} \sim n^{-1/2}$. On the other hand, $\Delta_\alpha/\left(\alpha\sqrt{n}\right) \sim \alpha/\sqrt{n} = n^{-3/4} = o_p\left(\rho_{\alpha,n}\right)$.

$$
\begin{aligned}
E\left[T_1^h T_1^{h'}|X\right] &= E\left[f'(I-P)\varepsilon\varepsilon'(I-P)f/n|X\right] \\
&= \sigma_\varepsilon^2 f'(I-P)^2 f/n = \sigma_\varepsilon^2 e_{2f}(\alpha), \\
E\left[T_1^h h'|X\right] &= E\left[f'(I-P)\varepsilon\varepsilon' f/n|X\right] = \sigma_\varepsilon^2 f'(I-P)f/n \\
&= \sigma_\varepsilon^2 e_f(\alpha) = O_p(\Delta_\alpha), \\
E\left[hh'H^{-1}T_1^H|X\right] &= E\left[f'\varepsilon\varepsilon' f H^{-1} f'(I-P)f/n^2|X\right] \\
&= \sigma_\varepsilon^2 f'(I-P)f/n = \sigma_\varepsilon^2 e_f(\alpha), \\
E\left[T_2^h T_2^{h'}|X\right] &= E\left[u'P\varepsilon\varepsilon' Pu/n|X\right] \\
&= \left(\sum_j q_j\right)^2 \sigma_{u\varepsilon}\sigma_{u\varepsilon}'/n + o_p\left(\left(\sum_j q_j\right)^2/n\right), \\
E\left[hT_2^{h'}|X\right] &= E\left[f'\varepsilon\varepsilon' Pu/n|X\right] = O_p\left(1/(n\alpha)\right), \\
E\left[T_1^h T_2^{h'}|X\right] &= E\left[f'(I-P)\varepsilon\varepsilon' Pu/n|X\right] \\
&= O_p\left(\Delta_\alpha^{1/2}/\sqrt{\alpha n}\right) = o_p\left(\rho_{\alpha,n}\right), \\
E\left[hh'H^{-1}T_2^H|X\right] &= E\left[hh'H^{-1}\left(u'f+f'u\right)/n|X\right] \\
&= O_p(1/n),
\end{aligned}
$$

because of Lemma 5(vi)–(viii).

Let $\hat{Z}^A(\alpha)=0$ and $\hat{A}(\alpha) = \left(h+T^h\right)\left(h+T^h\right)' - hh'H^{-1}T^{H'} - T^H H^{-1}hh'$. We have

$$
\begin{aligned}
&E\left(\hat{A}(\alpha)|X\right) \\
&= \sigma_\varepsilon^2 H + 2\sigma_\varepsilon^2 e_f(\alpha) + \sigma_\varepsilon^2 e_{2f}(\alpha) + \frac{\left(\sum_j q_j\right)^2}{n}\sigma_{u\varepsilon}\sigma_{u\varepsilon}' \qquad \text{(A.6)} \\
&\quad - 2\sigma_\varepsilon^2 e_f(\alpha) + O_p\left(\frac{1}{n}\right) + O_p\left(\frac{1}{n\alpha}\right) + o_p\left(\rho_{\alpha,n}\right) \\
&= \sigma_\varepsilon^2 H + \sigma_\varepsilon^2 e_{2f}(\alpha) + \frac{\left(\sum_j q_j\right)^2}{n}\sigma_{u\varepsilon}\sigma_{u\varepsilon}' + o_p\left(\rho_{\alpha,n}\right) \qquad \text{(A.7)} \\
&= \sigma_\varepsilon^2 H + HS(\alpha)H + o_p\left(\rho_{\alpha,n}\right).
\end{aligned}
$$

It is interesting to notice that the term $e_f(\alpha)$ cancels out, so that all the conditions of Lemma A1 of DN are satisfied. □

## References


Amemiya, T., 1966. On the use of principal components of independent variables in two-stage least-squares estimation. International Economic Review 7, 283–303.

Andersen, T., Sorensen, B., 1996. GMM estimation of a stochastic volatility model: a Monte Carlo study. Journal of Business and Economic Statistics 14, 328–352.

Angrist, T.W., Krueger, A., 1991. Does compulsory school attendance affect schooling and earnings. Quaterly Journal of Economics 106, 979–1014.

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

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

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

Berlinet, A., Thomas-Agnan, C., 2004. Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers, Boston.

Böttcher, A., 1996. Lecture 1: infinite matrices and projection methods. In: Lancaster, Peter (Ed.), Lectures on Operator Theory and Its Applications. American Mathematical Society, Providence.

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

Carrasco, M., Florens, J.P., 2000. Generalization of GMM to a continuum of moment conditions. Econometric Theory 16, 797–834.

Carrasco, M., Florens, J.-P., 2008. On the asymptotic efficiency of GMM, mimeo, University of Montreal.

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

Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure and mean-variance analysis in large asset markets. Econometrica 51, 1305–1324.

Chaussé, P., 2009. Generalized empirical likelihood for a continuum of moment conditions, mimeo UQAM.

Chen, X., Jacho-Chavez, D., Linton, O., 2009. An alternative way of computing efficient instrumental variable estimators, mimeo, Indiana University.

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of the generalized cross-validation. Numerische Mathematik 31, 377–403.

Dagenais, M., Dagenais, D., 1997. Higher moment estimators for linear regression models with errors in variables. Journal of Econometrics 76, 193–221.

Dominguez, M., Lobato, I., 2004. Consistent estimation of models defined by conditional moment restrictions. Econometrica 72, 1601–1615.

Donald, S., Imbens, G., Newey, W., 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. Journal of Econometrics 117, 55–93.

Donald, S., Newey, W., 2001. Choosing the number of instruments. Econometrica 69, 1161–1191.

Doran, H.E., Schmidt, P., 2006. GMM estimators with improved finite sample properties using principal components of the weighting matrix, with an application to the dynamic panel data model. Journal of Econometrics 387–409.

Eastwood, B., Gallant, R., 1991. Adaptive rules for seminonparametric estimators that achieve asymptotic normality. Econometric Theory 7, 307–340.

Engl, H., Hanke, M., Neubauer, A., 2000. Regularization of Inverse Problems. Kluwer Academic Publishers, Dordrecht.

Eubank, R., 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York.

Golub, G., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223.

Hansen, C., Hausman, J., Newey, W., 2008. Estimation with many instrumental variables. Journal of Business and Economic Statistics 26, 398–422.

Hofmann, T., Schölkopf, B., Smola, A., 2008. Kernel methods in machine learning. The Annals of Statistics 36, 1171–1220.

Kitamura, Y., Tripathi, G., Ahn, H., 2004. Empirical likelihood-based inference in conditional moment restriction models. Econometrica 72, 1667–1714.

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

Kuersteiner, G., 2001. Optimal instrumental variables estimation for ARMA models. Journal of Econometrics 104, 359–405.

Kuersteiner, G., 2006. Moment selection and bias reduction for GMM in conditionally heteroskedastic models. In: Corbea, D., Durlauf, S., Hansen, B.E. (Eds.), Econometric Theory and Practice — Frontiers of Analysis and Applied Research. Cambridge University Press.

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

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

Li, K-C., 1987. Asymptotic optimality for $C_p, C_L$, cross-validation and generalized cross-validation: discrete index set. The Annals of Statistics 15, 958–975.

Linton, O., 2002. Edgeworth approximations for semiparametric instrumental variable estimators and test statistics. Journal of Econometrics 106, 325–368.

Mallows, C.L., 1973. Some comments on $C_p$. Technometrics 15, 661–675.

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


[7] For SC, LF and Tikhonov with $\beta < 2$, we have $\rho_{\alpha,n} = O_p\left(1/n\alpha^2+\alpha^\beta\right)$. In the worst case scenario, $\rho_{\alpha,n} \sim 1/n\alpha^2+\alpha^\beta$.

Newey, W., 1993. Efficient estimation of models with conditional moment restrictions. In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics.
Newey, W., Smith, R., 2004. Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica 72, 219–255.
Okui, R., 2011. Instrumental variable estimation in the presence of many moment conditions. Journal of Econometrics 165, 70–86.
Owen, A., 1988. Empirical likelihood ratio confidence regions for a single functional. Biometrika 75, 237–249.
Stock, J., Watson, M., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business and Economic Statistics 20, 147–162.
Stone, C.J., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society 36, 111–147.
Theil, H., 1958. Economic Forecasts and Policy. North Holland, Amsterdam.
van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes. Springer Verlag, New York.
Van Rooij, A., Ruymgaart, F., 1999. On inverse estimation. In: Asymptotics, Nonparametrics, and Time Series. Dekker, NY, pp. 579–613.
Vapnik, V., 1998. Statistical Learning Theory. Wiley & Sons, New York.
Wahba, G., 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. Journal of the Royal Statistical Society, B 364–372.