A regularization approach to the many instruments problem

Marine Carrasco *
Universite de Montreal

September 2007
Comments welcome

Abstract

This paper focuses on the efficient estimation of a finite dimensional parameter in a linear model where the number of potential instruments is very large or infinite. It is well-known that the instrumental variables (IV) estimator has poor small sample properties when the number of instruments is large. In order to improve these small sample properties, we propose three modified IV estimators based on three different ways of inverting the covariance matrix of the instruments. These methods are based on the spectral decomposition of the covariance matrix and involve a regularization or smoothing parameter. We show that the three estimators are asymptotically normal and attain the semiparametric efficiency bound under some standard assumptions. Moreover, we derive the analytic expression of the mean square error (MSE) of the estimators and propose to select the value of the regularization parameter that minimizes the approximated MSE. The simulations show that these new estimators have good small sample properties.

Key Words: Factor model, many instruments, mean square error, regularization methods.

---

*An earlier version of this paper was entitled "Instrumental variables estimator based on principal components". The author thanks the participants of CEME (Cambridge, 2005), ESEM (Vienna, 2006), CESG (Niagara Falls, 2006), Econometric Society Meeting (Chicago, 2007), and the seminar participants at the University of Chicago for helpful comments. She also gratefully aknowledges partial financial support from the National Science Foundation, grant # SES-0211418. Address: Universite de Montreal, Departement de Sciences Economiques, CP 6128, succ Centre Ville, Montreal, QC H3C3J7, Canada. Email: marine.carrasco@umontreal.ca, http://www.econ.rochester.edu/Faculty/Carrasco.html

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. Panel data[1] (Arellano and Bond, 1991) offer also a large number of instruments. In principal, the asymptotic efficiency of the instrumental variables (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 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 choice of instruments will inevitably lead to a loss of efficiency. The focus is on providing an estimator that asymptotically reaches the semiparametric efficiency bound. 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. The empirical likelihood 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, Imbens, and Newey (2003) and Kitamura, Tripathi, and Ahn (2004) construct modified versions of 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. This remark leads us to consider alternative methods that do not require a ranking of the instruments and may be simpler

[1]In this paper, we consider a cross-section model. But we could treat the temporal dependence using Carrasco, Chernov, Florens, and Ghysels (2007).

to implement than the modified EL estimators of Kitamura et al (2004) and Donald et al (2003).

In our analysis, we do not restrict the number of instruments, which may be smaller or larger than the sample size, or even infinite. One restriction we impose, when the number of instruments is infinite, is that they are sufficiently correlated with each other. This condition seems plausible in practical applications. Moreover, all instruments are assumed to be valid and we will not address the issue of weak instruments. Our new estimators are based on three ways to compute a regularized inverse of the covariance matrix of the instruments. These three regularizations are taken from the literature on inverse problems, see Kress (1999) and Carrasco, Florens and Renault (2007). A first estimator is based on the principal components associated with the largest eigenvalues. A second estimator is the regularized method of moments adopted by Carrasco and Florens (2000). It is based on the Tikhonov regularization, well-known because it is applied in the ridge regression. A third estimator is based on a different regularization scheme called Landweber-Fridman. This last method is particularly useful in the case of an infinite number of instruments as it is an iterative method and is less computationally intensive than the other two. All these methods involve a regularization parameter, which is the counterpart of the smoothing parameter that appears in the papers cited above. Following the same approach as Nagar (1959) and Donald and Newey (DN), we compute the approximate mean square error (MSE) of the estimators and suggest selecting the smoothing parameter that minimizes the MSE. These MSE do not depend on the number of instruments and therefore, our estimators can be thought as alternatives to EL estimators to solve the problem of many instruments.

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 when using 2SLS. In the second case, our estimator based on Tikhonov regularization is found to be equivalent to the IV estimator that uses as instrument a nonparametric estimator of the optimal (unknown) instrument. The regularization parameter, α, is then a smoothing parameter that plays the same role as the bandwidth in kernel smoothing. We actually show that in some cases, our estimator of the instrument 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 DN estimator depends on whether the initial guess on the importance of the first instruments is correct or not.

The related literature is vast. Principal components have been used for a long time in order to reduce the dimension of the vector of regressors. 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, 2006) 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. In an attempt to improve the properties of the generalized method of moments (GMM) in presence of many moments, Kuersteiner (2002, 2006) proposes a kernel weighted GMM estimator and Okui (2004) introduces a shrinkage parameter to allocate less weight on a subset of instruments. 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.

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 presents a limited Monte Carlo experiment. An application to measuring the return to education is examined in Section 6. Section 7 concludes. The proofs are collected in Appendix.

2 Regularized versions of 2SLS

2.1 Presentation of the estimators

The model is

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

$i = 1, 2, ..., n$. δ is $p \times 1$ vector of interest. $E(\varepsilon_i|x_i) = E(u_i|x_i) = 0$, $E(\varepsilon_i^2|x_i) = \sigma^2$. y_i is a scalar and x_i is a vector of exogenous variables. The estimation is conducted using a sequence of instruments $Z_i = Z(\tau, x_i)$. τ may be an integer or an index taking its values in an interval. The set of values of τ is denoted $\mathcal{S}$. Let π be a positive weight function on $\mathcal{S}$. We denote $L^2(\pi)$ the Hilbert space of square integrable functions with respect to π.

We provide here a few examples of Z and π. More insights on how to choose them will be given in Section 2.3.

- Finite number of moments. $Z_i = (Z_{i,1}, Z_{i,2}, ..., Z_{i,L})$, we can take $\pi = 1$ on $\mathcal{S} = \{1, 2, ..., L\}$
- Countable infinite number of moments. $Z_i = (Z_{i,1}, Z_{i,2}, ...)$. For example if $x_i \in (-1, 1)$, we may take $Z_{i,j} = x_i^{j-1}$ and $\pi = 1$ on $\mathcal{S} = \mathbb{N}$.

- Continuum of moments. $Z_i = Z(\tau, x_i) = \exp(i\tau' x_i)$ with $\tau \in \mathcal{S} = \mathbf{R}^{\dim(x)}$, π can be taken equal to the density of the standard normal as in Carrasco, Chernov, Florens, and Ghysels (2007).

We estimate δ 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, 2004). Let $g_n(\delta) = \sum_{i=1}^{n} \left(y_i - W_i'\delta\right) Z_i / n$ and

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} (n \times 1),\ W = \begin{pmatrix} W_1' \\ W_2' \\ \vdots \\ W_n' \end{pmatrix} (n \times p),\ u = \begin{pmatrix} u_1' \\ u_2' \\ \vdots \\ u_n' \end{pmatrix} (n \times p).$$

The covariance operator K of the instruments is given by

$$\begin{aligned} K &: L^2(\pi) \to L^2(\pi) \\ (Kg)(\tau_1) &= \int E\left(Z(\tau_1, x_i)\overline{Z(\tau_2, x_i)}\right) g(\tau_2)\,\pi(\tau_2)\,d\tau_2 \end{aligned}$$

where $\overline{Z(\tau_2, x_i)}$ denotes the complex conjugate of $Z(\tau_2, x_i)$. 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, Florens, Renault, 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 $(\lambda_j, \phi_j : j = 1, 2, ...)$ be the eigenvalues and orthonormal eigenfunctions of K. The operator K is estimated by its sample counterpart K_n defined as

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

When the number of moment conditions is infinite, the inverse of K_n needs to be regularized because it is nearly singular. Various types of regularization techniques will be discussed shortly, they depend on a smoothing parameter α which choice is the topic of this paper. Let $(K_n^\alpha)^{-1}$ denote a regularized inverse of K_n and $(K_n^\alpha)^{-1/2} = \left((K_n^\alpha)^{-1}\right)^{1/2}$.

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 δ is defined as

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

Solving explicitly the minimization problem gives

$$\begin{aligned} \hat{\delta} &= (W'P^\alpha W)^{-1} W'P^\alpha y \\ &= \left(\hat{W}'W\right)^{-1} \hat{W}'y \end{aligned}$$

where $\hat{W} = P^\alpha W$, P^α is a $n \times n$ matrix defined as

$$P^\alpha = T (K_n^\alpha)^{-1} T^*$$

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

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

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

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

for any $v = (v_1, v_2, ..., v_n)' \in \mathbf{R}^n$. It is easy to check that $K_n = T^*T$ and TT^* is a $n \times n$ matrix with (i,j) element $\langle Z_i, Z_j \rangle / n$. Let $\hat{\phi}_j$, $j = 1, 2, ..., n$, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... > 0$ be the orthonormalized eigenfunctions and eigenvalues of K_n. We use a hat because $\hat{\phi}_j$ and $\hat{\lambda}_j$ are consistent estimators of the corresponding eigenfunction and eigenvalue of K, ϕ_j, λ_j. Let ψ_j be the eigenfunctions of TT^*. Then, we have $T\widehat{\phi_j} = \sqrt{\lambda_j}\psi_j$ and $T^*\psi_j = \sqrt{\lambda_j}\widehat{\phi_j}$. It may be useful to see what the formulas are when there is a finite number of instruments L and $\pi = 1$.

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

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, $P = T(K_n)^{-1} T^* = T(T^*T)^{-1} T^* = \underline{Z}(\underline{Z}'\underline{Z})^{-1}\underline{Z}'$ is the projection matrix on the vector of instruments. Then $\hat{\delta}$ is the usual IV estimator of δ using all the instruments.

The eigenfunctions ψ_j, $j = 1, 2, ..., L$ can be computed from $\psi_j = T\widehat{\phi}_j/\sqrt{\lambda_j} = \left(Z_1'\widehat{\phi}_j/\sqrt{\lambda_j}, ..., Z_n'\widehat{\phi}_j/\sqrt{\lambda_j}\right)'$. For $n > L$, the operator TT^* 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 TT^*. On the contrary, when $L > n$ or L is infinite, it is easier to compute the n eigenvalues and eigenfunctions of TT^* and infer the eigenfunctions of K_n by using the formula $\widehat{\phi}_j = T^*\psi_j/\sqrt{\lambda_j}$, $j = 1, 2, ..., n$.

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 ϕ_j by $\widehat{\phi}_j$ and λ_j by $\hat{\lambda}_j$.

1) **Tikhonov (T)**

This regularization scheme is closely related to the ridge regression.

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

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

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

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 α differs. In practice, both methods will give the same estimator. We will study the properties of SC in details.

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

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

for any $v \in \mathbf{R}^n$ and $\operatorname{tr}(P^\alpha) = \sum_{j=1}^{n} q\left(\alpha, \hat{\lambda}_j^2\right)$. The matrix P^α is idempotent for SC and PC but not for LF and T. In the case of PC, P^α is the projection matrix on the space spanned by ψ_j, $j = 1, ..., n$ associated with the largest (positive) eigenvalues. Note that for the four regularization methods, P^α can be written in two different ways

$$P^\alpha = T\left(K_n^\alpha\right)^{-1} T^* = \left(TT^*\right)\left(TT_\alpha^*\right)^{-1}$$

where $\left(TT_\alpha^*\right)^{-1}$ denotes the regularized inverse of TT^*. This remark is particularly useful for LF as it means that when $L > n$, one can regularize TT^* instead of T^*T.

The four regularization methods involve a regularization term, α. The set of possible values for α 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 $(K_n^\alpha)^{-1} r$ will vary only for values of α that are equal to the eigenvalues. As there are at most n nonzero eigenvalues, the set of α has at most cardinal n. We will choose α so that it minimizes the mean-square error (MSE) of $\hat{\delta}$.

2.2 Interpretation as nonparametric estimation of the optimal instrument

Under some conditions, $\hat{W} = P^\alpha W$ 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 $f(x_i) = \langle Z(., x_i), \varphi(.)\rangle$ for some unknown function φ in $L^2(\pi)$. This representation of f is very general. If $Z(., x_i) = \exp(i\tau' x_i)$ 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 φ 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 $Z(., x_i)$. Let T be as before the operator that associates to 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

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

Consequently, an estimator of f is obtained by

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

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

2.3 Choice of the instruments and inner product

We now discuss the choice of the instruments Z and the weight π that appears in the inner product. Our discussion below borrows from the machine learning literature (see for instance Vapnik (1998), Hofman, Scholkopf, and Smola (2007)). It is important to outline the role played by π and Z. They affect the estimation of δ only through the determination of the eigenvalues and eigenfunctions of the $n \times n$ matrix TT^* with principal element proportional to

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

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 $\widetilde{k}$ a priori without specifying the mappings $Z(\tau, x_i)$ and π. However, $\widetilde{k}$ can not been chosen completely arbitrary 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

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

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

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

We give below various examples of kernel $\tilde{k}$. The choice of $Z_i = Z(x_i, .)$ is mainly dictated by efficiency consideration. As we will see in Proposition 1, the estimator $\hat{\delta}$ is asymptotically efficient (in the sense of the semiparametric efficiency bound) if the space spanned by $\{Z(x_i, .)\}$ 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 $Z(\tau., x_i)$ has an interpretation in terms of Fourier series expansion. This function is particularly favored in machine learning. Then, $\tilde{k}$ takes a simple form

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

Hence, if π is a density, h is a characteristic function. On the other hand, if π 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 ω_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 π that have been successful in recovering functions in the machine learning literature.

(a) *Gaussian kernel*

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

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

The parameter σ^2 may be set equal to the variance of W_i or may be selected by cross-validation by including σ^2 along α in the criterion to minimize. We do not pursue the issue of the selection of σ^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

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

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.

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

Here Z_p maps x in $\mathbf{R}^d$ to the vector $Z_p(x)$ whose entries are all possible pth 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 $Z_p(x) = (x_1^2, x_2^2, x_1 x_2, x_2 x_1)'$ 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 μ_j are chosen so that $\int H_j(x)^2 e^{-x^2} dx = 1$. We define $\tilde{k}$ as

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

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

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

where ψ_j are orthonormal basis and r_j converges to zero as j goes to infinity, see Vapnik (1998, p. 461-462). This kernel provides an expansion of f in terms of ψ_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

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

Consider the following kernel

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

with

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

For multivariate x, one can construct d-dimensional splines as the product of d one-dimensional kernels (see Vapnik, 1998, p.465). For comparison purpose, 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

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

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

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

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:

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

where $\theta = (\theta_0, ..., \theta_{m-1})$ has a zero mean normal distribution with covariance matrix νI_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

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

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[2] estimator of f is the element of $W_2^{(m)}$ which is the solution of

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

[2]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.

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 a same framework, kernel-type, sieves, and splines estimators of f. Many other kernels $\widetilde{k}$ can be constructed (see Hofman, Scholkopf, and Smola (2007) for more examples).

2.4 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 ath element of $f(x)$.

Proposition 1 *Assuming $\{y_i, W_i, x_i\}$ are iid, $E(\varepsilon_i^2) = \sigma_\varepsilon^2$, $E\left[f(x_i) f(x_i)'\right]$ exists and is nonsingular, K is compact, and α goes to zero as n goes to infinity, the T, LF, and SC estimators satisfy:*

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

2) Asymptotic normality: If moreover, $f_a(x)$ belongs to the closure of the linear span of $\{Z(.,x)\}$ for $a = 1, ..., p$, 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 $n\alpha^2$ go to infinity.

The compactness of K implies a functional central limit theorem (see van der Vaart and Wellner, 1996)

$$\sum_{i=1}^n \frac{Z(.,x_i)\varepsilon_i}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 K\right)$$

in $L^2(\pi)$. Condition (2) 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 K there exists π_K such that $E\left(\left\|f(x) - \pi_K Z^K(x)\right\|^2\right) \to 0$ as $K \to \infty$, where Z^K is a subset of the instruments.

Note that this property is also equivalent to the condition that $E(Z(.,x_i) f_a(x_i))$ belongs to the range of K or for n large, $\hat{E}(Z(.,x_i) f_a(x_i))$ belongs to the range of K_n:

$$\sum_j \frac{\left\langle \hat{E}\left(Z(.,x_i) f_a(x_i)\right), \hat{\phi}_j \right\rangle^2}{\hat{\lambda}_j^2} < \infty,$$

or equivalently

$$\sum_j \frac{\langle f_a, \psi_j \rangle^2}{\hat{\lambda}_j} < \infty. \tag{9}$$

Indeed, (9) is equivalent to "there exists a function $g \in L^2(\pi)$ such that $f_a = Tg$, that is $f_a(x_i) = \langle Z_i, g \rangle$, $i = 1, 2, ..., n$." (9) is also equivalent to say that f_a belongs to the reproducing kernel Hilbert space associated with $\tilde{k}$ defined in (4).

We illustrate Condition (ii) by four 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 $X : 1, X, X^2, ...$

(c) Same assumption as in (b). Another way to obtain efficiency is to consider $Z(\tau, x) = \exp(i\tau x)$ for $\tau \in \mathbb{R}$.

(d) If f is known to belong to some class of functions that can be represented as an expansion

$$f(x) = \sum_{j=1}^{\infty} \langle f, \psi_j \rangle \psi_j(x),$$

then a natural choice is to select $\tilde{k}$ as in (8).

The choice of instruments to obtain efficiency is discussed in Newey (1993) and Carrasco and Florens (2004) 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.

3 Mean square error

We impose some regularity conditions. Let $\|A\|$ be the Euclidean norm of a matrix A. f is the $n \times p$ matrix, $f = (f(x_1), f(x_2), ..., f(x_n))'$. Let H be the $p \times p$ matrix $H = f'f/n$ and $X = (x_1, ..., x_n)$.

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

$$\sum_{j=1}^{n} \frac{\langle f_a, \psi_j \rangle^2}{\hat{\lambda}_j^{2\beta}} < \infty \tag{10}$$

or equivalently

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

for all $a = 1, 2, ..., p$, where f_a is the ath column of f and $\hat{E}\left(Z\left(x_i,.\right) f_a\left(x_i\right)\right) = \sum_{i=1}^{n} Z\left(x_i,.\right) f_a\left(x_i\right)/n$

Assumption 3. (i) $\mathrm{E}\left[\left(\left(\varepsilon_i, u_i'\right)'\left(\varepsilon_i, u_i'\right)\right)|x_i\right]$ is constant, (ii) K is a Hilbert-Schmidt operator with nonzero eigenvalues, (iii) $\max_{i \le n} P_{ii}^{\alpha} \to 0$, (iv) $f\left(x_i\right)$ is bounded.

Assumptions 1, 2(i) and 3 are imposed by DN. Assumption 2(ii) for $\beta = 1/2$ corresponds to the efficiency condition stated in the previous section. The value of β in (10) measures how well the instruments approximate the reduced form, f. The larger β, the better the approximation is. β will show up in the rate of convergence of the MSE.

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

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

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

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

As usual in this type of computation, $S\left(\alpha\right)$ is composed of a bias term that increases when α goes to zero and a variance term that decreases when α goes to zero. Remark that for $\beta \le 2$, 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, β 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.

$$\begin{aligned} W_i &= \sum_{j=1}^{J} \omega_j f_{ij} + \varepsilon_i, \\ x_{ia} &= \sum_{j=1}^{J} \gamma_{ija} f_{ij} + \nu_{ia}, \ a = 1, 2, ...L. \end{aligned}$$

Assume that the factors f_{ij} are normalized. It is well-known that the ψ_j, $j = 1, ..., J$ associated with the J largest eigenvalues are estimators of the factors (Chamberlain and Rothschild, 1983). Hence, condition (1) is satisfied for any value of β. For the PC estimator, $P^{\alpha}f = f$ 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.

If we use the kernel generating splines to approximate the function f (see Section 2.3) and apply the spectral cut-off, then we can characterize the rate of convergence of the MSE in terms of the smoothness of f. Assume that x is scalar with bounded support and f belongs to $W_2^{(1)}$, then $\beta = 1$ (see Eubank, 1998, Chapter 6) and hence $S(\alpha) = O\left(1/\left(\alpha^2 n\right) + \alpha\right) = O\left(n^{-1/3}\right)$ for the optimal α.

It would be interesting to compare $S(\alpha)$ with the expression of the approximate MSE given by DN for the 2SLS. We review briefly the results of DN. Consider a countable sequence of instruments $(Z_{i,1}, Z_{i,2}, ...)$. 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 $(Z_{i,1}, Z_{i,2}, ...Z_{i,L})$. 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^{α} 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^{α} 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. Assume that x has bounded support of dimension 1 and f is one time continuously differentiable, then the rate for DN estimator based on polynomials is $S(L) = O\left(L^2/n + L^{-2}\right) = O\left(n^{-1/2}\right)$ for the optimal L. This rate is faster than that obtained above for the spline estimator, partly because the rate for the variance of the spline estimator could be improved upon by exploiting the specific structure of the splines.

4 Estimation of MSE

The aim is to find α 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 $H^{-1}v$:

$$\begin{aligned} WH^{-1}v &= fH^{-1}v + uH^{-1}v \Leftrightarrow \\ W_v &= f_v + u_v \end{aligned} \tag{11}$$

using obvious notation. Then, we are back to a univariate regression where f_v is estimated by $P^\alpha W_v$. The results of Li (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 $W'P^{\tilde{\alpha}}W/n$ 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 $\tilde{u} = \left(I - P^{\tilde{\alpha}}\right)W$, $\tilde{u}_v = \tilde{u}\tilde{H}^{-1}v$.

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

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

We consider the following goodness-of-fit criteria:

Mallows $\mathbf{C}_p$ (Mallows (1973)):

$$\hat{R}^m(\alpha) = \frac{\hat{u}_v^{\alpha\prime}\hat{u}_v^\alpha}{n} + 2\hat{\sigma}_{u_v}^2\frac{\operatorname{tr}(P^\alpha)}{n} = \frac{\hat{u}_v^{\alpha\prime}\hat{u}_v^\alpha}{n} + 2\hat{\sigma}_{u_v}^2\frac{\sum_j q_j}{n}.$$

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

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

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

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

where $\tilde{W}_v = W\widetilde{H}^{-1}v$, $\tilde{W}_{v_i}$ is the ith element of the vector $\tilde{W}_v$, $\hat{f}^{\alpha}_{v_{-i}} = P^{\alpha}_{-i}\tilde{W}_{v_{-i}}$ where P^{α}_{-i} and $\tilde{W}_{v_{-i}}$ are obtained by suppressing the ith observation from the sample.

As first stage criterion, the two cross-validation methods are preferable to Mallows' C_p because they do not require computing $\hat{\sigma}^2_{u_v}$. The leave-one-out cross-validation is more burdensome to implement than the generalized cross-validation. The second criterion 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

$$\hat{S}_v(\alpha) = \hat{\sigma}^2_{u_v\varepsilon} \frac{\left(\sum_j q_j\right)^2}{n} + \hat{\sigma}^2_{\varepsilon} \left[\hat{R}(\alpha) - \hat{\sigma}^2_{u_v} \frac{\operatorname{tr}\left((P^{\alpha})^2\right)}{n} \right]$$

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

$$\begin{aligned} S_v(\alpha) &= v'H^{-1}\sigma_{u\varepsilon}\sigma'_{u\varepsilon}H^{-1}v \frac{\left(\sum_j q_j\right)^2}{n} + \sigma^2_{\varepsilon} \frac{f'_v (I - P^{\alpha})^2 f_v}{n} \\ &= \sigma^2_{u_v\varepsilon} \frac{\left(\sum_j q_j\right)^2}{n} + \sigma^2_{\varepsilon} \frac{f'_v (I - P^{\alpha})^2 f_v}{n} \end{aligned} \quad (12)$$

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 α in the regression (11), $\hat{R}(\alpha)$ approximates

$$R_v(\alpha) = \frac{f'_v (I - P^{\alpha})^2 f_v}{n} + \sigma^2_{u_v} \frac{\operatorname{tr}\left((P^{\alpha})^2\right)}{n}.$$

Hence replacing $f'_v (I - P^{\alpha})^2 f_v / n$ by $\hat{R}(\alpha) - \hat{\sigma}^2_{u_v} \operatorname{tr}\left((P^{\alpha})^2\right)/n$ into (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, Heath, and Wahba, 1979, Li, 1986). The optimality of this selection rule needs to be established (to be completed).

5 Monte Carlo experiments

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

$$\begin{aligned} y_i &= \delta W_i + \varepsilon_i, \\ W_i &= f(x_i) + u_i \end{aligned} \tag{13}$$

for $i = 1, 2, ..., n$. $\delta = 0.1$ 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 (13) is taken from Model 1 of DN

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

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 worse case scenario for our regularized estimators. Indeed, here all the eigenvalues of K are equal to 1, so there is no information contained in the spectral decomposition of K. Moreover, if L were infinite, K would not be compact 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. $\pi_l = d(1 - l/(L+1))^4$, $l = 1, 2, ..., L$, where the constant d is chosen so that $\pi'\pi = R_f^2/\left(1 - R_f^2\right)$. The instruments are ordered in decreasing order of importance. We use this example to compare our estimators with DN. First, we compute the same estimator as described in DN (denoted 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 worse instruments are selected first. This illustrates the pitfall associated with an a priori ranking of the instruments.

Model 1b. $\pi_l = \sqrt{R_f^2/L\left(1 - R_f^2\right)}$, $l = 1, 2, ..., L$. 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 30×1 vector of instruments constructed from f_i through

$$x_i = M f_i + \nu_i$$

where $\nu_i \sim iid\mathcal{N}(0, I_{30})$ and M is a 30×3 matrix, which elements 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, $P^\alpha W$ 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.

The simulations are performed using 1000 replications of samples of size $n = 500$. We compute the estimators corresponding to Tikhonov (T), Landweber-Fridman (LF) and Principal components (PC) regularizations. We selected the optimal value of α by minimizing the approximate MSE using the generalized cross-validation criterion. Estimates of $\sigma_{u\varepsilon}$ and σ_ε^2 are computed using $\tilde{\delta}$ obtained with the value of α, which minimizes the first stage cross-validation criterion. For the T regularization, we looked for the values of α in a 10 point equispaced grid, ranging from 0.00001 to 0,0090 for Models 1a and 1b, and from 0.001 to 0.451 for Model 2. For LF, we searched among the number of iterations ranging from 1 to 5. Finally for PC, we searched among the number of eigenfunctions from 1 to the maximum (20 in Models 1a, 1b, and 30 in Model 2). For LF, we used $c = 0.1/\lambda_1^2$ where λ_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.

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 $\widehat{W} = P^\alpha W$.

First, we examine the results of Table 1, Model 1a. As expected, DN estimator using the right order of the instruments dominates all the other estimators. T estimator is not too far behind in terms of MSE. PC is by far the worse regularized estimator. 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 a reliable information of the relative importance of the instruments, DN approach should be preferred but on the contrary if there is little information, it is preferable to use the regularized estimators.

For Model 1b, 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.

Table 2 contain summary statistics for the value of the regularization parameter which minimizes the approximate MSE. This regularization parameter is $\alpha \in (0,1)$ for T, the number of iterations in $\{1,2,...,5\}$ 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.

For T, we see that the value of α is the smallest allowed value for Models 1a and 1b and the largest allowed value for Model 2. This makes sense because the smallest eigenvalue is large in Models 1a, 1b and only a small regularization is needed, while in Model 2, the smallest eigenvalue is close to zero and a strong regularization is necessary

For LF, we see that the minimization of the MSE leads to select one iteration only in all cases.

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

6 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, Hausman, and Newey (2006) 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

$$logw = \alpha + \delta education + \beta_1' Y + \beta_2' S + \varepsilon \tag{14}$$

where *logw* = log of weekly wage, *education* = years of education, Y = year of birth dummy (9×1), S = state of birth dummy (50×1). 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 (3×1), QY =

interaction between quarter of birth and year of birth (27×1), QS = interaction between quarter of birth and state of birth (150×1). 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 $n = 33130$.

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)
		α	Number of iterations	Number of eigenfunctions
		0.00012	700	210

In Table 3, we report various estimates of δ 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 purpose. 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 with 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 $Z'Z$ 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 in LF. Actually, 700 was the maximum number allowed but the estimates and standard errors did not change much between 400 and 700 iterations.

7 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 presence of many orthogonality conditions. Regularized 2SLS has also an interesting interpretation as 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

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

where P^{α} is $n \times n$ 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^{α} satisfies some minimal conditions.

Extension to heteroscedasticity

If the assumption of homoscedastic errors ε_i is relaxed, the GMM estimator of δ should use the heroscedasticity-robust version of the weighting matrix. Namely, the kernel of K_n becomes

$$k_n(\tau_1, \tau_2) = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 Z_i(\tau_1) Z_i(\tau_2)$$

where $\hat{\varepsilon}_i$ is an estimator of ε_i. If Z satisfies the condition of Proposition 1, the resulting estimator will be asymptotically efficient. The study of its MSE will be complicated by the presence of $\hat{\varepsilon}_i$. The GMM estimator is also implementable in a time-series context as shown in Carrasco, Chernov, Florens, and Ghysels (2007). But again, the MSE would be difficult to derive.

Bias correction

The introduction of the regularization parameter α 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 $\hat{\delta}$. Instead of selecting the optimal α, we could take the route that consists in selecting α at a rate that guarantees asymptotic efficiency and apply a bias correction. The resulting estimator will be asymptotically efficient and will have smaller bias than the current estimator. We may lose in terms of efficiency in small samples.

A Appendix. Proofs

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

Lemma 3 *Consider g and g_n such that $\|g_n - g\| = O_p(1/\sqrt{n})$.*
(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, $n\alpha$ go to infinity and α 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, $n\alpha^2$ go to infinity and α goes to zero.

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

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

We have

$$\begin{aligned} & \|B_n^\alpha g_n - Bg\| \\ \leq \; & \|B_n^\alpha g_n - B_n^\alpha g\| && (1) \\ & + \|B_n^\alpha g - B^\alpha g\| && (2) \\ & + \|B^\alpha g - Bg\| && (3) \end{aligned}$$

Term (1). $\|B_n^\alpha g_n - B_n^\alpha g\| \leq \|B_n^\alpha\| \|g_n - g\|$. Note that for the three regularizations, we have $\|B_n^\alpha\| \leq c/\alpha$ for some positive constant c, see Kress (1999) and Carrasco, Florens, and Renault (2007). Hence, $\|B_n^\alpha g_n - B_n^\alpha g\| = O_p\left(1/\left(\alpha\sqrt{n}\right)\right)$.

Term (2). $\|\left(B_n^\alpha - B^\alpha\right) g\| = \|B_n^\alpha \left(K_n^\alpha - K^\alpha\right) B^\alpha g\| \leq \|B_n^\alpha\| \|K_n^\alpha - K^\alpha\| \|B^\alpha g\|$ 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 $\|B_n^\alpha\| \leq c/\alpha$. The second term can be rewritten as follows.

$\left(K_n^\alpha - K^\alpha\right) g = \left(K_n - K\right) g + \left(K_n^\alpha - K_n\right) g + \left(K - K^\alpha\right) g$

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

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

We have

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

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, Florens, and Renault (2007). Because $q\left(\alpha, \lambda_j^2\right)$ converges to 1 as α 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, $\|K_n^\alpha - K^\alpha\| \leq \|K_n - K\| + o_p(1)$ and hence goes to zero. We now turn our attention to the third term. Because g is in the range of K, there exists a function φ such that $g = K\varphi$ and $\|\varphi\| < \infty$. Using the fact that $0 \leq q \leq 1$, $\|B^\alpha g\|^2 = \|B^\alpha K \varphi\|^2 = \sum_j q\left(\alpha, \lambda_j^2\right)^2 \langle \varphi, \phi_j \rangle^2 \leq \sum_j \langle \varphi, \phi_j \rangle^2 = \|\varphi\|^2 < \infty$. It follows that Term (2) goes to zero.

Term (3). $\|B^\alpha g - Bg\| = \|B^\alpha K \varphi - \varphi\| = \|(B^\alpha K - I)\, \varphi\| \to 0$ as α 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 $B^\alpha = (K^\alpha)^{-1/2}$, $B_n^\alpha = (K_n^\alpha)^{-1/2}$ 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 $\|B_n^\alpha\| \leq \sqrt{c/\alpha}$, hence the different rate of convergence on α.

Proof of Proposition 1.

Consistency. We can write[3]

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

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

where $g_n\left(\tau\right)=\hat{E}\left(Z_i\left(\tau\right)W_i\right)$. Consider a typical element of the vector g_n, namely $g_{na}=\hat{E}\left(Z_i\left(\tau\right)W_{ia}\right)$ where W_{ia} is the ath element of W_i. Let $g_a\left(\tau\right)=E\left(Z_i\left(\tau\right)W_{ia}\right)$. By Lemma 3, $\left\|\left(K_n^{\alpha}\right)^{-1/2}g_{na}\right\| \to \left\|K^{-1/2}g_a\right\|$ provided $n, n\alpha \to \infty$ 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 $k\left(\tau_1,\tau_2\right)=E\left(Z_i\left(\tau_1\right)Z_i\left(\tau_2\right)\right)$. For the definitions and properties of RKHS, see Berlinet and Thomas-Agnan (2004) and Carrasco, Florens, and Renault (2007). We use the notation $\left\|K^{-1/2}g_a\right\|=\left\|g_a\right\|_K<\infty$ where $\left\|g_a\right\|_K$ denotes the norm of g_a in the RKHS associated with K. Similarly, we denote $\left\langle g_a,g_b\right\rangle_K$ the inner product in the RKHS. Under the assumptions of Proposition 1, we have

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

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

Asymptotic normality. We have

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

The first term is negligeable because

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

By the functional central limit theorem, the second term satisfies

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

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

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

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

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

To see this, we apply Theorem 6.4 of Carrasco, Florens, and Renault (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 $\|g_a\|_K^2 = E\left(f_a^2\right)$. 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 λ_j and ϕ_j and we denote P^α and $q\left(\alpha, \lambda_j^2\right)$ by P and q_j in the sequel. We need the following preliminary results.

Lemma 4 *If Assumptions 1-3 are satisfied then with probability 1:*
(i) $\sum_j q\left(\alpha, \lambda_j^2\right) = O(1/\alpha)$, $\sum_j \left(q\left(\alpha, \lambda_j^2\right)\right)^2 = O(1/\alpha)$,
(ii) $h = f'\varepsilon/\sqrt{n} = O(1)$, $H = f'f/n = O(1)$.

Proof of Lemma 4.

(i) As $0 \leq q_j \leq 1$, we have $\sum_j q_j^2 \leq \sup q_j \sum_j q_j \leq \sum_j q_j$. Hence, $\sum_j \left(q\left(\alpha, \lambda_j^2\right)\right)^2 = O\left(\sum_j q\left(\alpha, \lambda_j^2\right)\right)$.

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, $\sum_j q_j = \sum_j I\left(\lambda_j^2 \geq \alpha\right)\left(\lambda_j^2/\lambda_j^2\right) \leq \sum_j \lambda_j^2/\alpha = O(1/\alpha)$ because K is a Hilbert-Schmidt operator and hence $\sum_j \lambda_j^2 < \infty$.

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

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

(ii) follows from Lemma A2(v) of DN.

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

Lemma 5 *If Assumptions 1-3 are satisfied then with probability 1:*

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

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

(ii) $f'(I-P)\varepsilon/\sqrt{n} = O\left(\Delta_\alpha^{1/2}\right)$,

(iii) $u'P\varepsilon = O(1/\alpha)$,

(iv) $E[u'P\varepsilon\varepsilon'u|X] = \left(\sum_j q_j\right)^2 \sigma_{u\varepsilon}\sigma'_{u\varepsilon} + \left(\sum_j q_j^2\right)(\sigma_{u\varepsilon}\sigma'_{u\varepsilon} + \sigma_\varepsilon^2\Sigma_u) = \left(\sum_j q_j\right)^2 \sigma_{u\varepsilon}\sigma'_{u\varepsilon} + o\left(\left(\sum_j q_j\right)^2\right)$.

(v) $E[f'\varepsilon\varepsilon'Pu|X] = O(1/\alpha)$,

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

(vii) $E[hh'H^{-1}u'f/n|X] = O(1/n)$,

(viii) $E[f'(I-P)\varepsilon\varepsilon'Pu/n|X] = o\left(\Delta_\alpha^{1/2}/\sqrt{\alpha n}\right)$.

Proof of Lemma 5.

(i)

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

Taking the trace, we obtain

$$\begin{aligned} \text{tr}\left(\frac{f'(I-P)f}{n}\right) &= \sum_{a=1}^{p}\sum_j (1-q_j)\langle f_a,\psi_j\rangle^2 \\ &= \sum_{a=1}^{p}\sum_j \lambda_j^{2\beta}(1-q_j)\frac{\langle f_a,\psi_j\rangle^2}{\lambda_j^{2\beta}} \\ &\leq \sum_{a=1}^{p}\sup_{\lambda_j}\lambda_j^{2\beta}(1-q_j)\sum_j \frac{\langle f_a,\psi_j\rangle^2}{\lambda_j^{2\beta}}. \end{aligned}$$

Under Assumption 2, the rate of $\text{tr}(f'(I-P)f/n)$ is given by $\sup_{\lambda_j}\lambda_j^{2\beta}(1-q_j)$.

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

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

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

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

(ii)

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

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

(iii) Following DN, we have $u' P \varepsilon \leq (\mathrm{tr} (u' P u) (\varepsilon' P \varepsilon))^{1/2}$ and $E \left[(\varepsilon' P \varepsilon) |X\right] = \mathrm{tr}(P E (\varepsilon' \varepsilon | X)) = \sigma_\varepsilon^2 \mathrm{tr}(P) = \sigma_\varepsilon^2 \sum_j q_j$, so by Markov inequality $\varepsilon' P \varepsilon = O \left(\sum_j q_j\right) = O (1/\alpha)$. Similarly $u' P u = O \left(\sum_j q_j\right)$, giving (iii).

(iv)

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

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

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

by the fact that the eigenvectors are orthonormal, that is $\sum_i \psi_{i,j}\psi_{i,l} = 0$ if $j \neq l$ and $= n\|\psi\|^2 = n$ if $j = l$. (iv) follows from Lemma 2(i).

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

(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 $1/\alpha$ and Δ_K by Δ_α.

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

$$\hat{H} = W'P^\alpha W, \;\; \hat{h} = W'P^\alpha \varepsilon.$$

Observe that

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

First, we prove for LF, SC, PC, and T (with $\beta \leq 1$). We apply Lemma A1 of DN with

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

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 $o(\rho_{\alpha,n})$.

$E\left[T_1^h T_1^{h\prime}|X\right] = E\left[f'(I-P)\varepsilon\varepsilon'(I-P)f/n|X\right] = \sigma_\varepsilon^2 f'(I-P)^2 f/n = \sigma_\varepsilon^2 e_{2f}(\alpha)$.

$E\left[T_1^h h'|X\right] = E\left[f'(I-P)\varepsilon\varepsilon' f/n|X\right] = \sigma_\varepsilon^2 f'(I-P)f/n = \sigma_\varepsilon^2 e_f(\alpha) = O(\Delta_\alpha)$.

$E\left[hh'H^{-1}T_1^H|X\right] = E\left[f'\varepsilon\varepsilon' f H^{-1} f'(I-P)f/n^2|X\right] = \sigma_\varepsilon^2 f'(I-P)f/n = \sigma_\varepsilon^2 e_f(\alpha)$.

$E\left[T_2^h T_2^{h\prime}|X\right] = E\left[u'P\varepsilon\varepsilon' Pu/n|X\right] = \left(\sum_j q_j\right)^2 \sigma_{u\varepsilon}\sigma_{u\varepsilon}'/n.$

$E\left[hT_2^{h\prime}|X\right] = E\left[f'\varepsilon\varepsilon' Pu/n|X\right] = O(1/(n\alpha))$.

$E\left[T_1^h T_2^{h\prime}|X\right] = E\left[f'(I-P)\varepsilon\varepsilon' Pu/n|X\right] = o\left(\Delta_\alpha^{1/2}/\sqrt{\alpha n}\right) = o(\rho_{\alpha,n})$

because of Lemma 5 (vi) and (viii). Finally by Lemma 5 (vii), $E\left[hh'H^{-1}T_2^H|X\right] = E\left[hh'H^{-1}(u'f + f'u)/n|X\right] = O(1/n)$.

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

$$\begin{aligned}
E\left(\hat{A}(\alpha)|X\right) &= \sigma_\varepsilon^2 H + 2\sigma_\varepsilon^2 e_f(\alpha) + \sigma_\varepsilon^2 e_{2f}(\alpha) + \frac{\left(\sum_j q_j\right)^2}{n}\sigma_{u\varepsilon}\sigma_{u\varepsilon}' \\
&\quad -2\sigma_\varepsilon^2 e_f(\alpha) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n\alpha}\right) \\
&= \sigma_\varepsilon^2 H + HS(\alpha)H + o(\rho_{\alpha,n}).
\end{aligned} \tag{5}$$

So that all the conditions of Lemma A1 of DN are satisfied.

Now we turn our attention to the case of T with $\beta > 1$. In this case, the conditions of Lemma A1 of DN are not satisfied because

$$\operatorname{tr}(e_{2f}(\alpha)) = o(\operatorname{tr}(e_f(\alpha))).$$

So in fact, the term in $e_f(\alpha)$ should dominate, but we observe from (5), that $e_f(\alpha)$ cancels out. Note that if the rate of the MSE[4] is given by $\alpha^{\min(\beta,2)} + 1/(n\alpha^2)$, it is minimized for

[4]In Proposition 1, we write MSE= $O\left(\alpha^{\min(\beta,2)} + 1/(n\alpha^2)\right)$. In the worse case scenario, MSE$\sim$ $\alpha^{\min(\beta,2)} + 1/(n\alpha^2)$.

$\alpha = n^{-1/4}$. Hence, the optimal rate of the MSE is $n^{-1/2}$. $E\left[T_1^h T_2^{h\prime}|X\right] = o\left(\Delta_\alpha^{1/2}/\sqrt{\alpha n}\right) = o\left(\alpha^{\min(\beta,1)/2}/\sqrt{\alpha n}\right) = o\left(n^{-1/2}\right)$, so this term is negligeable. An inspection of all the terms show that the dominant terms are again $\sigma_\varepsilon^2 H + \sigma_\varepsilon^2 e_{2f}\left(\alpha\right) + \frac{\left(\sum_j q_j\right)^2}{n}\sigma_{u\varepsilon}\sigma_{u\varepsilon}'$.

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. and B. Sorensen (1996) "GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study", *Journal of Business and Economic Statistics*, 14, 328-352.

Angrist, T.W. and A. Krueger (1991) "Does Compulsory School Attendance Affect Schooling and Earnings", *Quaterly Journal of Economics*, 106, 979-1014.

Arellano, M. and S. Bond (1991) "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations", *Review of Economic Studies*, 58, 277-297.

Bai, J. and S. Ng (2002) "Determining the number of factors in approximate factor models", *Econometrica*, 70, 191-221.

Bai, J. and S. Ng (2006) "Instrumental Variable Estimation in a Data Rich Environment", mimeo, University of Michigan.

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

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

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

Carrasco, M. and J. P. Florens (2000) "Generalization of GMM to a continuum of moment conditions", *Econometric Theory*, 16, 797-834.

Carrasco, M. and J.-P. Florens (2004) "On the Asymptotic Efficiency of GMM", mimeo, University of Rochester.

Carrasco, M., J. P. Florens, and E. Renault (2007) "Linear Inverse Problems in Structural Econometrics: Estimation based on spectral decomposition and regularization", forthcoming in the *Handbook of Econometrics*, Vol. 6B, edited by J.J. Heckman and E.E. Leamer.

Chamberlain, G. and M. Rothschild (1983) "Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets", *Econometrica*, 51, 1305-1324.

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

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

Donald, S., G. Imbens, and W. Newey (2003) "Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions", *Journal of Econometrics*, 117,

55-93.
Donald, S. and W. Newey (2001) "Choosing the number of instruments", *Econometrica*, 69, 1161-1191.
Doran, H.E. and P. Schmidt (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.
Eubank, R. (1988) *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.
Golub, G., M. Heath, and G. Wahba (1979) "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter", *Technometrics*, 21, 215-223.
Hansen, C., J. Hausman, and W. Newey (2006) "Estimation with Many Instrumental Variables", mimeo, Chicago GSB.
Hofmann, T., B. Scholkopf, and A. Smola (2007) "Kernel Methods in Machine Learning", mimeo, available on arXiv:math.ST/0701907v2.
Kitamura, Y., G. Tripathi, and H. Ahn (2004) "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models", *Econometrica*, 72, 1667-1714.
Kress, R. (1999), *Linear Integral Equations,* Springer.
Kuersteiner, G. (2002) "Mean squared error prediction for GMM estimators of linear time series models", mimeo, UC Davis.
Kuersteiner, G. (2006) "Moment Selection and Bias Reduction for GMM in Conditionally Heteroskedastic Models", in *Econometric Theory and Practice - Frontiers of Analysis and Applied Research*, D. Corbea, S. Durlauf and B.E. Hansen eds. Cambridge University Press.
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.
Newey, W. (1993) "Efficient estimation of Models with Conditional Moment Restrictions", in *Handbook of Statistics*, G.S. Maddala, C.R. Rao, and H.D. Vinod, eds.
Newey, W. and R. Smith (2004) "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators", *Econometrica*, 72, 219-255.
Okui, R. (2004) "Instrumental variable estimation in the presence of many moment conditions", mimeo, University of Pennsylvania.

Owen, A. (1988) “Empirical likelihood ratio confidence regions for a single functional”, *Biometrika*, 75, 237-249.

Stock, J. and M. Watson (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.

van der Vaart, A. and J. Wellner (1996) *Weak Convergence and Empirical Processes*, Springer Verlag, New York.

Van Rooij, A., F. Ruymgaart (1999) “On Inverse Estimation”, in *Asymptotics, Nonparametrics, and Time Series*, 579-613, Dekker, NY.

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.

Table 1. Summary Statistics

		MSE	Med.bias	Med.abs	Dis.	Cov.
Model 1a	DN	0.0256	0.0507	0.1053	0.3925	0.911
	DN-R	121.2187	0.5695	0.8789	5.2462	0.861
	T	0.0296	0.1197	0.1302	0.3249	0.826
	LF	0.0485	0.0168	0.1437	0.5530	0.964
	PC	6.2758	0.1276	0.1515	0.4019	0.850
Model 1b	DN	0.0477	0.1499	0.1675	0.3839	0.819
	T	0.0293	0.1162	0.1278	0.3394	0.838
	LF	0.0215	0.0048	0.0969	0.3752	0.958
	PC	0.0659	0.1207	0.1486	0.3968	0.846
Model 2	DN	0.0015	0.0062	0.0274	0.0959	0.937
	T	0.0009	0.0043	0.0209	0.0749	0.938
	LF	0.0038	-0.0006	0.0431	0.1569	0.954
	PC	0.0009	-0.0005	0.0204	0.0764	0.943
	IV	0.0008	-0.0000	0.0192	0.0711	0.935

Table 2. Properties of the distribution of the regularization parameters

		mean	std	mode	q1	median	q3
Model 1a	DN	5.992	1.995	4.000	5.000	6.000	7.000
	DN-R	2.917	3.910	1.000	1.000	1.000	3.000
	T	.00001	0.000	.00001	.00001	.00001	.00001
	LF	1.000	0.000	1.000	1.000	1.000	1.000
	PC	11.791	5.136	10.000	8.000	12.000	16.000
Model 1b	DN	13.412	5.470	20.000	9.000	14.000	19.000
	T	.00001	0.000	.00001	.00001	.00001	.00001
	LF	1.000	0.000	1.000	1.000	1.000	1.000
	PC	12.029	5.150	20.000	8.000	12.000	16.000
Model 2	DN	12.694	5.185	9.000	9.000	11.000	17.000
	T	0.451	0.000	0.451	0.451	0.451	0.451
	LF	1.000	0.000	1.000	1.000	1.000	1.000
	PC	3.511	1.193	2.000	3.000	3.000	4.000