Asymptotic Normal Inference in Linear Inverse Problems*

Marine Carrasco[†]
Université de Montréal

Jean-Pierre Florens[‡]
Toulouse School of Economics

Eric Renault[§]
Brown University

January 2013

*This chapter has been prepared for the *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, edited by Aman Ullah, Jeff Racine, and Liangjun Su and to be published by Oxford University Press of New York. We thank the editors and one referee for helpful comments.

[†]CIREQ, CIRANO, University of Montreal, Departement de Sciences Economiques, CP 6128, succ Centre Ville, Montreal, QC H3C3J7, Canada. Email: marine.carrasco@umontreal.ca. Carrasco thanks SSHRC for partial financial support.

[‡]Toulouse School of Economics, IDEI, 21 allée de Brienne, 31000 Toulouse, France. Email: florens@cict.fr.

[§]Department of Economics, Box B, Brown University, Providence, RI 02912. Email: Eric_Renault@brown.edu

Abstract

This chapter studies the estimation of φ in linear inverse problems $T\varphi = r$ where r is only observed with error and T may be given or estimated. The unknown element φ belongs to an Hilbert space $\mathcal{E}$. Four examples are relevant for econometrics: the density estimation, the deconvolution problem, the linear regression with an infinite number of possibly endogenous explanatory variables and the nonparametric instrumental variables estimation. In the first two cases T is given, whereas it is estimated in the two other cases, respectively at a parametric or non parametric rate. The paper will recall the main results on these models: concepts of degree of ill-posedness, regularity of φ, regularized estimation, and the rates of convergence usually obtained. The main contributions are moreover related to the asymptotic normality of the regularized solution $\hat{\varphi}_\alpha$ obtained with a regularization parameter α. If $\alpha \to 0$, we particularly consider the asymptotic normality of inner products $\langle \hat{\varphi}_\alpha, \phi \rangle$ where ϕ is an element of $\mathcal{E}$. These results can be used to construct (asymptotic) tests on φ.

Key words: Deconvolution, functional linear regression, nonparametric instrumental regression, ill-posed inverse problem, Tikhonov regularization, Hilbert scales, asymptotic normality.

1. Introduction

At least since Hansen's (1982) seminal paper on Generalized Method of Moments (GMM), econometricians have been used to make inference on an object of interest defined by a family of orthogonality conditions. While Hansen's GMM is focused on inference on a finite dimensional vector θ of structural unknown parameters, our object of interest in this chapter will typically be a function φ element of some Hilbert space $\mathcal{E}$.

While Hansen (1982) acknowledged upfront that "identification requires at least as many orthogonality conditions as they are coordinates in the parameter vector to be estimated", we will be faced with two dimensions of infinity. First, the object of interest, the function φ, is of infinite dimension. Second, similarly to above, identification will require a set of orthogonality conditions at least as rich as the infinite dimension of φ.

Then, a convenient general framework is to describe the set of orthogonality conditions through a linear operator T from the Hilbert space $\mathcal{E}$ to some other Hilbert space $\mathcal{F}$ and a target vector r given in $\mathcal{F}$. More precisely, the testable implications of our structural model will always be summarized by a linear inverse problem:

$$T\varphi = r \tag{1.1}$$

which will be used for inference about the unknown object φ based on a consistent estimator $\hat{r}$ of r. Similarly to the Method of Moments, the asymptotic normality of estimators $\hat{\varphi}$ of φ will be derived from asymptotic normality of the sample counterpart $\hat{r}$ of the population vector r.

However, it is worth realizing that the functional feature of r introduces an additional degree of freedom that is not common for GMM with a finite number of unknown parameters, except in the recent literature on many weak instruments asymptotics. More precisely, the accuracy of estimators of r, namely the rate of convergence of $\hat{r}$ for asymptotic normality heavily depends on the "choice of instruments" namely on the choice of the inverse problem (1.1) to solve. It must actually be kept in mind that this choice is to some extent arbitrary since (1.1) can be transformed by any operator K to be rewritten:

$$KT\varphi = Kr \tag{1.2}$$

An important difference with (semi)parametric settings is that even the transformation by a one-to-one operator K may dramatically change the rate of convergence of the estimators of the right-hand side (r.h.s.) of the equation. Some operators (as integration or convolution) are noise-reducing whereas some others (as differentiation or deconvolution) actually magnify the estimation error.

A maintained assumption will be that some well-suited linear transformation Kr allows us to get root-n asymptotically normal estimator $K\hat{r}$ of Kr. Then, the key issue to address is the degree of ill-posedness of the inverse problem (1.2), that is precisely to what extent the estimation error in $K\hat{r}$ is magnified by the (generalized) inverse operator of (KT).

Because of the ill-posedness of the inverse problem, we need a regularization of the estimation to recover consistency. Here, we consider a class of regularization techniques which includes Tikhonov, iterated Tikhonov, spectral cut-off, and Landweber-Fridman regularizations. For the statistical properties of these methods, see Engl, Hanke, and Neubauer (2000). For a review of the econometric aspects, we refer the reader to Florens (2003) and Carrasco, Florens, and Renault (2007).

The focus of this chapter is the asymptotic normality of the estimator $\hat{\varphi}$ of φ in the Hilbert space $\mathcal{E}$. Normality in the Hilbert space $\mathcal{E}$ being defined through all linear functionals $< \hat{\varphi}, \delta >$ (see e.g. Chen and White (1998)), it is actually the rate of convergence of such functionals that really matters. In the same way as going from (1.1) to (1.2) may modify the rate of convergence of sample counterparts of the r.h.s, rates of convergence of linear functionals $< \hat{\varphi}, \delta >$ will depend on the direction δ we consider. There may exist in particular some Hilbert subspace of directions warranting root-n asymptotic normality of our estimator $\hat{\varphi}$. However, it is worth stressing that focusing only on such directions amounts to overlooking the information content of other test functions and as such, yields to suboptimal inference. It is then worth characterizing the rate of convergence to normality of estimators $\hat{\varphi}$ of φ in any possible direction δ of interest. Since this rate actually depends on the direction, we do not get a functional asymptotic normality result as in other settings put forward in Chen and White (1998).

The paper is organized as follows. Section 2 presents the model and examples. Section 3 describes the estimation method. Section 4 investigates the normality for fixed regularization parameter α. This result is used in the tests described in Section 5. Section 6 establishes asymptotic normality when α goes to zero. Section 7 discusses the practical selection of α and Section 8 describes the implementation. Section 9 concludes.

In the sequel, $\mathcal{D}$ and $\mathcal{R}$ denote the domain and range of an operator. Moreover, $t \wedge s = \min(t, s)$ and $t \vee s = \max(t, s)$.

2. Model and examples

A wide range of econometric problems are concerned with estimating a function φ from a structural model

$$r = T\varphi \tag{2.1}$$

where T is a linear operator from a Hilbert (L^2 or Sobolev) space $\mathcal{E}$ into a Hilbert space $\mathcal{F}$. The function r is estimated by $\hat{r}$ and the operator T is either known or estimated. We present four leading examples.

2.1. Density

We observe data x_1, x_2,..., x_n of unknown density f we wish to estimate. The density f is related to the distribution function F through

$$F(t) = \int_{-\infty}^{t} f(s)\, ds = (Tf)(t).$$

In this setting $r = F$ and the operator T is a known integral operator. F can be estimated by $\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq t)$ where $\widehat{F}(t)$ converges at a parametric rate to F.

2.2. Deconvolution

Assume we observe n i.i.d. realizations $y_1, y_2, ..., y_n$ of a random variable Y with unknown density h. Y is equal to a unobservable variable X plus an error ε where X and ε are mutually independent with density functions f and g respectively so that $h = f * g$. The aim is to estimate f assuming g is known. The problem consists in solving for f the equation

$$h(y) = \int g(y-x) f(x)\, dx.$$

In this setting, the operator T is known and defined by $(Tf)(y) = \int g(y-x) f(x)\, dx$ whereas $r = h$ can be estimated but a slower rate than the parametric rate.

Here, the choice of the spaces of reference is crucial. If T is considered as an operator from $L^2(\mathbb{R})$ into $L^2(\mathbb{R})$ provided with Lebesgue measure, then T has a continuous spectrum. Carrasco and Florens (2011) chose spaces of reference for which T is compact and hence has a discrete spectrum. Let π_X and π_Y be two nonnegative weighting functions. Denote $L^2_{\pi_Y}$ the space of square integrable real-valued functions with respect to π_Y :

$$L^2_{\pi_Y} = \left\{ \psi(y) \text{ such that } \int \psi(y)^2 \pi_Y(y)\, dy < \infty \right\}.$$

$L^2_{\pi_X}$ is defined similarly. We formally define T as the operator from $L^2_{\pi_X}$ into $L^2_{\pi_Y}$ which associates to any function $\phi(x)$ of $L^2_{\pi_X}$ a function of $L^2_{\pi_Y}$ as:

$$(T\phi)(y) = \int g(y-x)\phi(x)\, dx. \tag{2.2}$$

Provided π_X and π_Y are such that

$$\int\int (g(y-x))^2 \frac{\pi_Y(y)}{\pi_X(x)} dxdy < \infty,$$

T is a Hilbert-Schmidt operator and hence has a discrete spectrum. We define the adjoint, T^*, of T, as the solution of $\langle T\varphi, \psi\rangle = \langle \varphi, T^*\psi\rangle$ for all $\varphi \in L^2_{\pi_X}$ and $\psi \in L^2_{\pi_Y}$. It associates to any function $\psi(y)$ of $L^2_{\pi_Y}$ a function of $L^2_{\pi_X}$:

$$(T^*\psi)(x) = \int \frac{g(y-x)\pi_Y(y)}{\pi_X(x)} \psi(y)\, dy.$$

For convenience, we denote its kernel

$$\pi_{Y|X}(y|x) = \frac{g(y-x)\pi_Y(y)}{\pi_X(x)}.$$

2.3. Functional linear regression with possibly endogenous regressors

We observe i.i.d. data (Y_i, Z_i, W_i) where each explanatory variable Z_i is a random function, element of a Hilbert space $\mathcal{E}$ and W_i is a random function in a Hilbert space $\mathcal{F}$. Let $\langle ., . \rangle$ denote the inner product in $\mathcal{E}$. The response Y_i is generated by the model

$$Y_i = \langle Z_i, \varphi \rangle + u_i \tag{2.3}$$

where $\varphi \in \mathcal{E}$ and u_i is i.i.d. with zero mean and finite variance. $Y_i \in \mathbb{R}$ and $u_i \in \mathbb{R}$. The regressors are endogenous, but we observe a function W_i which plays the role of instruments so that φ is identified from

$$E(u_i W_i) = 0$$

or equivalently

$$E(Y_i W_i) = E(\langle Z_i, \varphi \rangle W_i).$$

As X_i and W_i are functions, one can think of them as real random variables observed in continuous time. In this setting, $r = E(Y_i W_i)$ is unknown and needs to be estimated, the operator T, defined by $T\varphi = E(\langle Z_i, \varphi \rangle W_i)$, needs to be estimated also. Both estimators converge at a parametric rate to the true values.

This model is considered in Florens and Van Bellegem (2012). In the case where the regressors are exogenous and $W = Z$, this model has been studied by Ramsay and Silverman (1997), Ferraty and Vieu (2000), Cardot and Sarda (2006), and Hall and Horowitz (2007).

2.4. Nonparametric instrumental regression

We observe an i.i.d. sample $(Y_i, Z_i, W_i) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ where the relationship between the response Y_i and the vector of explanatory variable Z_i is represented by the equation

$$Y_i = \varphi(Z_i) + u_i. \tag{2.4}$$

We wish to estimate the unknown function φ using as instruments the vector W_i. We assume that

$$E(u_i | W_i) = 0$$

or equivalently

$$E(Y_i | W_i) = E(\varphi(Z_i) | W_i). \tag{2.5}$$

In this setting, $r(w) = E(Y_i | W_i = w)$ is estimated at a slow nonparametric rate (even for a given w) and the operator T defined by $(T\varphi)(w) = E(\varphi(Z) | W = w)$ is also estimated at a slow rate. The identification and estimation of φ has been studied in many recent papers, e.g. Newey and Powell (2003), Darolles, Fan, Florens, and Renault (2011), Hall and Horowitz (2005), Blundell, Chen, and Kristensen (2007), Chen and Reiss (2011) and references below. While some authors considered orthogonal series, Darolles et al. (2011) consider a kernel estimator of the conditional expectation. There, the spaces of reference are $\mathcal{E} = L^2_Z$, the space of functions that are square integrable with respect to the true density of Z, similarly $\mathcal{F} = L^2_W$. For these spaces, the adjoint T^* of T is a conditional expectation operator: $(T^*\psi)(z) = E(\psi(W) | Z = z)$ which can also be estimated by kernel. While Darolles et al. (2011) use a Tikhonov regularization where the penalty is on the L_2 norm of φ, Florens, Johannes, and Van Bellegem (2011) and Gagliardini and Scaillet (2012) consider a Tikhonov regularization where the penalty is on a Sobolev norm of φ, i.e. the L^2 norm of its derivatives.

3. Assumptions and estimation method

3.1. Ill-posedness

First we impose some identification conditions on Equation (2.1).

Assumption 1. The solution φ of (2.1) exists and is unique.

The uniqueness condition is equivalent to the condition that T is one-to-one, i.e. the null space of T is reduced to zero. As discussed in Newey and Powel (2003), this identification condition in the case of nonparametric IV is, $E(\varphi(Z)|W=w)=0$ for all w implies $\varphi=0$, which is equivalent to the completeness in w of the conditional distribution of Z given $W=w$. Interestingly this condition is not testable, see Canay, Santos, and Shaikh (2011).

Assumption 2. T is a linear bounded operator from a Hilbert space $\mathcal{E}$ to a Hilbert space $\mathcal{F}$. Moreover, T is a Hilbert-Schmidt operator.

T is an Hilbert-Schmidt operator if for some (and then any) orthonormal basis $\{e_k\}$, we have $\sum \|Te_k\|^2 < \infty$. It means in particular that its singular values are square summable. It implies that T is compact. As T is compact, its inverse is unbounded so that the solution φ does not depend continuously on the data. Indeed, if r is replaced by a noisy observation $r+\varepsilon$, then $T^{-1}(r+\varepsilon)$ may be very far from the true $\varphi=T^{-1}r$. Therefore, the solution needs to be stabilized by regularization.

First, we need to define certain spaces of reference to characterize the properties of T and φ.

3.2. Hilbert scales

To obtain the rates of convergence, we need assumptions on φ and T in terms of Hilbert scales. For a review on Hilbert scales, see Krein and Petunin (1966) and Engl, Hanke, and Neubauer (2000). We define L an unbounded selfadjoint strictly positive operator defined on a dense subset of the Hilbert space $\mathcal{E}$. Let $\mathcal{M}$ be the set of all elements ϕ for which the powers of L are defined, i.e. $\mathcal{M} := \bigcap_{k=0}^{\infty} \mathcal{D}(L^k)$ where $\mathcal{D}$ denotes the domain. For all $s \in \mathbb{R}$, we introduce the inner product and norm:

$$\begin{aligned} \langle \phi, \psi \rangle_s &= \langle L^s \phi, L^s \psi \rangle , \\ \|\phi\|_s &= \|L^s \phi\| \end{aligned}$$

where ϕ, $\psi \in \mathcal{M}$. The Hilbert space $\mathcal{E}_s$ is defined as the completion of $\mathcal{M}$ with respect to the norm $\|.\|_s$. $(\mathcal{E}_s)_{s\in\mathbb{R}}$ is called the Hilbert scale induced by L. If $s \geq 0$, $\mathcal{E}_s = \mathcal{D}(L^s)$. Moreover, for $s \leq s'$, $\mathcal{E}_{s'} \subset \mathcal{E}_s$.

A typical example is the case where L is a differential operator. Let $\mathcal{E}$ be the set of complex-valued functions ϕ such that $\int_0^1 |\phi(s)|^2 ds < \infty$. Define the operator I on $\mathcal{E}$ by

$$(I\phi)(t) = \int_0^t \phi(s)\, ds.$$

Let I^* be the adjoint of I, I^* is such that

$$(I^* g)(t) = \int_t^1 g(s)\, ds.$$

Let L be such that $L^{-2} = I^* I$. Then, for $b > 0$, $\phi \in \mathcal{D}\left(L^b\right)$ is equivalent to say that ϕ is b differentiable and satisfies some boundary conditions (e.g. $\phi \in \mathcal{D}(L^2)$ means that ϕ is twice differentiable and $\phi(0) = \phi'(0) = 0$). Note that we could not define $L\phi = \phi'$ because the derivative is not self-adjoint. The construction above gives heuristically $L\phi = \sqrt{-\phi''}$. Indeed, since $L^{-2} = I^* I$, we have $L^2 (I^* I) \phi = \phi$. This is satisfied for $L^2 \phi = -\phi''$.

The degree of ill-posedness of T is measured by the number a in the following assumption.

Assumption 3. T satisfies

$$\underline{m} \|\phi\|_{-a} \leq \|T\phi\| \leq \overline{m} \|\phi\|_{-a}$$

for any $\phi \in \mathcal{E}$ and some $a > 0$, $0 < \underline{m} < \overline{m} < \infty$.

Assumption 4. $\varphi \in \mathcal{E}_b$, for some $b > 0$.

In our example of a differential operator, Assumption 4 is equivalent to φ is b differentiable.

Let $B = TL^{-s}$, $s \geq 0$. According to Corollary 8.22 of Engl et al. (2000), for $|\nu| < 1$,

$$\underline{c}(\nu) \|\phi\|_{-\nu(a+s)} \leq \left\| (B^* B)^{\nu/2} \phi \right\| \leq \bar{c}(\nu) \|\phi\|_{-\nu(a+s)} \tag{3.1}$$

for any $\phi \in \mathcal{D}\left((B^*B)^{\nu/2}\right)$, with $\underline{c}(\nu) = \min(\underline{m}^\nu, \overline{m}^\nu)$ and $\bar{c}(\nu) = \max(\underline{m}^\nu, \overline{m}^\nu)$. Moreover,

$$\mathcal{R}\left((B^*B)^{\nu/2}\right) = \mathcal{E}_{\nu(a+s)} \tag{3.2}$$

where $(B^*B)^{\nu/2}$ has to be replaced by its extension to $\mathcal{E}$ if $\nu < 0$.

It is useful to make the link between Assumptions 3 and 4 and the source condition given in Carrasco, Florens and Renault (2007, Definition 3.4.). This condition is written in terms of the singular system of T denoted $\left(\lambda_j, \phi_j, \psi_j\right)$:

$$\sum_{j=1}^{\infty} \frac{\left\langle \varphi, \phi_j \right\rangle^2}{\lambda_j^{2\beta}} < \infty. \tag{3.3}$$

This means that $\varphi \in \mathcal{R}\left((T^*T)^{\beta/2}\right)$ or equivalently $\varphi \in \mathcal{D}\left((T^*T)^{-\beta/2}\right)$. If we let $L = (T^*T)^{-1/2}$, we see that Assumption 3 holds with $a = 1$. Then Assumption 4 is equivalent to (3.3) with $\beta = b$. Another interpretation is the following. Using (3.2), we see that $\mathcal{R}\left((T^*T)^{\beta/2}\right) = \mathcal{E}_{\beta a}$. Hence, Assumptions 3 and 4 with $b = \beta a$ imply the source condition (3.3). While the condition (3.3) relates the properties of φ and T directly, Assumptions 3 and 4 characterize the properties of φ and T with respect to an auxiliary operator L.

3.3. Regularization and estimation

As the inverse of T is not continuous, some regularization is needed. The most common one is Tikhonov regularization which consists in penalizing the norm of φ :

$$\min_{\varphi} \|T\varphi - \hat{r}\|^2 + \alpha \|\varphi\|^2 .$$

We will consider a more general case where we penalize the $\mathcal{E}_s$ norm of φ :

$$\min_{\varphi \in \mathcal{E}_s} \|T\varphi - \hat{r}\|^2 + \alpha \|\varphi\|_s^2 . \tag{3.4}$$

The reason to do this is twofold. Assuming L is a differential operator and φ is known to be s times differentiable, we may want to dampen the oscillations of $\hat{\varphi}$ by penalizing its derivatives. Second, if we are interested in estimating $L^c\varphi$ for some $0 < c < s$, then we immediately obtain an estimator $\widehat{L^c\varphi} = L^c\hat{\varphi}$ and its rate of convergence.

The solution to (3.4) is given by

$$
\begin{aligned}
\hat{\varphi} &= \left(\alpha L^{2s} + T^{*}T\right)^{-1} T^{*}\hat{r} \\
&= L^{-s}\left(\alpha I + L^{-s}T^{*}TL^{-s}\right)^{-1} L^{-s}T^{*}\hat{r} \\
&= L^{-s}\left(\alpha I + B^{*}B\right)^{-1} B^{*}\hat{r} \qquad (3.5)
\end{aligned}
$$

where $B = TL^{-s}$.

We also consider other regularization schemes. Let us define the regularized solution to (2.1) as

$$\hat{\varphi} = L^{-s} g_{\alpha}\left(B^{*}B\right) B^{*}\hat{r}. \qquad (3.6)$$

where $g_{\alpha} : \left[0, \|B\|^{2}\right] \to \mathbb{R}$, $\alpha > 0$, is a family of piecewise continuous functions and

$$
\begin{aligned}
\lim_{\alpha \to 0} g_{\alpha}(\lambda) &= \frac{1}{\lambda},\ \lambda \neq 0, \\
\left|g_{\alpha}(\lambda)\right| &\leq \widehat{c}\alpha^{-1}, \qquad (3.7) \\
\lambda^{\mu}\left|1 - \lambda g_{\alpha}(\lambda)\right| &\leq c_{\mu}\alpha^{\mu},\ 0 \leq \mu \leq \mu_{0}, \qquad (3.8)
\end{aligned}
$$

with $\widehat{c}$ and $c_{\mu} > 0$ independent of α and $\mu_0 \geq 1$. The main examples of functions g_{α} are the following.

1. The Tikhonov regularization is given by $g_{\alpha}(\lambda) = 1/(\lambda + \alpha)$.

2. The iterated Tikhonov regularization of order m is given by
   $g_{\alpha}(\lambda) = \left(1 - \left(\alpha/(\lambda + \alpha)\right)^{m}\right)/\lambda$. The solution is obtained after m iterative minimizations:

$$\hat{\varphi}_{j} = \arg\min_{\phi \in \mathcal{E}_{s}} \|T\phi - \hat{r}\|^{2} + \alpha \left\|\phi - \hat{\varphi}_{j-1}\right\|_{s}^{2},\ j = 1, ..., m,\ \hat{\varphi}_{0} = 0.$$

3. The spectral cut-off considers $g_{\alpha}(\lambda) = 1/\lambda$ for $\lambda \geq \alpha$.

4. The Landweber Fridman regularization takes $g_{\alpha}(\lambda) = \left(1 - (1 - \lambda)^{1/\alpha}\right)/\lambda$.

When B is unknown, we replace B by a consistent estimator $\hat{B}$ and B^{*} by $\left(\hat{B}\right)^{*}$. The convergence of $\hat{\varphi}$ is studied in Engl et al. (2000), Carrasco et al. (2007), Chen and Reiss (2011), and Johannes, Van Bellegem and Vanhems (2011).

3.4. Rate of convergence of MSE

Here we study the mean square error (MSE) of $\hat{\varphi}$ when B is known. When B is estimated, the error due to its estimation usually goes to zero faster than the other terms and does not affect the convergence rate of the bias (see Carrasco et al., 2007).

To simplify the exposition, we first let $s = c = 0$ and consider Tikhonov regularization. The general case is discussed at the end. The difference $\hat{\varphi} - \varphi$ can be decomposed as the following sum

$$\hat{\varphi} - \varphi = \hat{\varphi} - \varphi_\alpha + \varphi_\alpha - \varphi$$

where

$$\varphi_\alpha = (\alpha I + T^*T)^{-1} T^*T\varphi.$$

The term $\hat{\varphi} - \varphi_\alpha$ corresponds to an estimation error whereas the term $\varphi_\alpha - \varphi$ corresponds to a regularization bias. We first examine the latter (see Groetsch, 1993).

$$\begin{aligned}\varphi_\alpha - \varphi &= \sum_j \frac{\lambda_j^2}{\lambda_j^2 + \alpha} \langle \varphi, \varphi_j \rangle \varphi_j - \sum_j \langle \varphi, \varphi_j \rangle \varphi_j \\ &= -\alpha \sum_j \frac{1}{\lambda_j^2 + \alpha} \langle \varphi, \varphi_j \rangle \varphi_j.\end{aligned}$$

Given

$$\begin{aligned}\|\varphi_\alpha - \varphi\|^2 &= \alpha^2 \sum_j \frac{1}{\left(\lambda_j^2 + \alpha\right)^2} \langle \varphi, \varphi_j \rangle^2 \\ &\leq \sum_j \langle \varphi, \varphi_j \rangle^2 < \infty,\end{aligned} \tag{3.9}$$

we may in passing to the limit as α goes to zero in (3.9), interchange the limit and the summation yielding

$$\lim_{\alpha \to 0} \|\varphi_\alpha - \varphi\|^2 = 0.$$

From this result, we understand that we can not obtain a rate of convergence for $\|\varphi_\alpha - \varphi\|^2$ unless we impose more restrictions on φ. Assume that φ satisfies the source condition (3.3) for some $\beta > 0$, then

$$\begin{aligned}
\|\varphi_\alpha - \varphi\|^2 &\leq \sup_\lambda \frac{\alpha^2 \lambda^{2\beta}}{\left(\lambda^2 + \alpha\right)^2} \sum_j \frac{\left\langle \varphi, \varphi_j \right\rangle^2}{\lambda_j^{2\beta}} \\
&= O\left(\alpha^{\beta \wedge 2}\right)
\end{aligned}$$

by Kress (1999) and Carrasco and al. (2007).

We now turn to the estimation error. There are two ways to characterize the rate of convergence of $\|\hat{\varphi} - \varphi_\alpha\|^2$ depending on whether we have an assumption on $\|r - \hat{r}\|^2$ or $\|T^* (r - \hat{r})\|^2$. First we consider the rate of $\|\hat{\varphi} - \varphi_\alpha\|^2$ in terms of $\|r - \hat{r}\|^2$. We have

$$\begin{aligned}
\hat{\varphi} - \varphi_\alpha &= (\alpha I + T^* T)^{-1} T^* (T\varphi - \hat{r}), \\
&= T^* (\alpha I + T T^*)^{-1} (T\varphi - \hat{r})
\end{aligned}$$

$$\begin{aligned}
\|\hat{\varphi} - \varphi_\alpha\|^2 &= \left\langle T^* (\alpha I + T T^*)^{-1} (T\varphi - \hat{r}), T^* (\alpha I + T T^*)^{-1} (T\varphi - \hat{r}) \right\rangle \\
&= \left\langle (\alpha I + T T^*)^{-1} (T\varphi - \hat{r}), T T^* (\alpha I + T T^*)^{-1} (T\varphi - \hat{r}) \right\rangle.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\left\| (\alpha I + T T^*)^{-1} \right\| &\leq \frac{1}{\alpha}, \\
\left\| T T^* (\alpha I + T T^*)^{-1} \right\| &\leq 1.
\end{aligned}$$

Hence,

$$\|\hat{\varphi} - \varphi_\alpha\|^2 \leq \frac{1}{\alpha} \|r - \hat{r}\|^2.$$

In summary, the MSE of $\hat{\varphi}$ is bounded in the following way:

$$E\left(\|\hat{\varphi} - \varphi\|^2\right) \leq \frac{1}{\alpha} E\left(\|r - \hat{r}\|^2\right) + C\alpha^{\beta \wedge 2} \tag{3.10}$$

for some constant C.

Second, we consider the rate of $\|\hat{\varphi} - \varphi_\alpha\|^2$ in terms of $\|T^* (\hat{r} - r)\|^2$.

$$\begin{aligned}
\|\hat{\varphi} - \varphi_\alpha\|^2 &\leq \left\| (\alpha I + T^* T)^{-1} \right\|^2 \|T^* (T\varphi - \hat{r})\|^2 \\
&\leq \frac{1}{\alpha^2} \|T^* (r - \hat{r})\|^2.
\end{aligned}$$

The MSE of $\hat{\varphi}$ is bounded in the following way:

$$E\left(\|\hat{\varphi}-\varphi\|^2\right) \leq \frac{1}{\alpha^2} E\left(\|T^*(r-\hat{r})\|^2\right) + C\alpha^{\beta\wedge 2} \tag{3.11}$$

In both expressions (3.10) and (3.11), there is a trade-off between the regularization bias which declines as α goes to zero and the variance which increases as α goes to zero. The optimal α is selected so that the rate of the regularization bias equals that of the variance.

These results generalize to the other three regularization techniques described earlier. In the case of Spectral cut-off, Landweber-Fridman, and iterated Tikhonov regularizations, the rate of $\|\varphi_\alpha-\varphi\|^2$ is $O\left(\alpha^\beta\right)$. In the case of Tikhonov with $\beta<2$, it is also $O\left(\alpha^\beta\right)$. So the rates given below apply to the four methods. The optimal α is chosen so that $\alpha^{\beta+1}=E\left(\|r-\hat{r}\|^2\right)$ or $\alpha^{\beta+2}=E\left(\|T^*(r-\hat{r})\|^2\right)$, hence

$$E\left(\|\hat{\varphi}-\varphi\|^2\right) = O\left(\min\left(E\left(\|r-\hat{r}\|^2\right)^{\beta/(\beta+1)}, E\left(\|T^*(r-\hat{r})\|^2\right)^{\beta/(\beta+2)}\right)\right). \tag{3.12}$$

We can see that, for the optimal α, $\sqrt{n}\|\varphi_\alpha-\varphi\|$ diverges so that there is an asymptotic bias remaining when studying the asymptotic distribution of $\sqrt{n}(\hat{\varphi}-\varphi)$.

We can analyze the rate of (3.12) in different scenarios.

- If $r-\hat{r}$ converges at a parametric rate $\sqrt{n}$ then $T^*(r-\hat{r})$ also converges at a parametric rate and the first term of the r.h.s of (3.12) converges to 0 faster than the second term. So that the rate of the MSE is given by $n^{-\beta/(\beta+1)}$.

- If $r-\hat{r}$ converges at a nonparametric rate so that $\|r-\hat{r}\|^2=O_p\left(n^{-2\nu}\right)$ with $\nu<1/2$ and $\|T^*(T\varphi-\hat{r})\|^2=O_p\left(n^{-1}\right)$, if moreover $2\nu<(\beta+1)/(\beta+2)$ then the second term in the r.h.s of (3.12) converges to 0 faster than the first term. So that the rate of the MSE is given by $n^{-\beta/(\beta+2)}$. This is encountered in nonparametric IV, see e.g. Darolles et al. (2011). There, $r=E(Y|W)$ and $\nu=d/(2d+q)$ where q is the dimension of W and d is the number of derivatives of $E(Y|W)$. If $\beta=2$, $d=2$ and $q\geq 2$ then the condition $2\nu<(\beta+1)/(\beta+2)$ holds. See also Chen and Reiss (2011), Johannes et al. (2011).

So far, we derived the rate of convergence of the MSE using a source condition (3.3). Now we establish the results using assumptions on the degree of ill-posedness of T. Suppose moreover that we are interested in estimating the derivative of φ, $L^c\varphi$.

Proposition 3.1. *Assume that T satisfies Assumption 3, φ satisfies Assumption 4 with $b \leq a+2s$, and $\hat{\varphi}$ is defined as in (3.6). Then, for the optimal α,*

$$\begin{aligned} E\left(\|L^c\hat{\varphi} - L^c\varphi\|^2\right) &= O\left(\min\left(E\left(\|r-\hat{r}\|^2\right)^{(b-c)/(a+b)},\right.\right. \\ &\qquad \left.\left. E\left(\|B^*(r-\hat{r})\|^2\right)^{(b-c)/(b+2a+s)}\right)\right). \end{aligned}$$

Setting $c = s = 0$, we see that this result is the same as the rates (3.12) obtained with the source condition (3.3) and $\beta = b/a$.

Proof of Proposition 3.1. We follow the steps of the proof of Engl, Hanke, and Neubauer (2000, Theorem 8.23). Note that by (3.7) and (3.8)

$$\lambda^t |g_\alpha(\lambda)| \leq C\alpha^{t-1} \tag{3.13}$$

where C denotes a generic constant. We have

$$\begin{aligned} \|L^c(\hat{\varphi} - \varphi_\alpha)\| &= \left\|L^{(c-s)} g_\alpha(B^*B) B^*(\hat{r} - r)\right\| \\ &= \|g_\alpha(B^*B) B^*(\hat{r}-r)\|_{c-s} \\ &\leq C\left\|(B^*B)^{\frac{s-c}{2(a+s)}} g_\alpha(B^*B) B^*(\hat{r}-r)\right\|, \end{aligned} \tag{3.14}$$

where the inequality follows from Inequality (3.1) with $\nu = (s-c)/(a+s)$ and $\phi = g_\alpha(B^*B) B^*(\hat{r}-r)$. Note that

$$\begin{aligned} &\left\|(B^*B)^{\frac{s-c}{2(a+s)}} g_\alpha(B^*B) B^*(\hat{r}-r)\right\|^2 \\ =\ &\left\langle (B^*B)^{\frac{s-c}{2(a+s)}} g_\alpha(B^*B) B^*(\hat{r}-r), (B^*B)^{\frac{s-c}{2(a+s)}} g_\alpha(B^*B) B^*(\hat{r}-r)\right\rangle \\ \leq\ &\|BB^* g_\alpha(BB^*)(\hat{r}-r)\| \left\|(BB^*)^{\frac{s-c}{(a+s)}} g_\alpha(BB^*)(\hat{r}-r)\right\| \\ \leq\ &C\alpha^{-(a+c)/(a+s)} \|\hat{r}-r\|^2 \end{aligned}$$

where the last inequality follows from (3.13). Hence,

$$\|\hat{\varphi} - \varphi_\alpha\| \leq C\|T\varphi - \hat{r}\| \alpha^{-(a+c)/(2(a+s))}$$

Another majoration follows from (3.14) and (3.13):

$$\begin{aligned} \|L^c(\hat{\varphi} - \varphi_\alpha)\| &\leq C\left\|(B^*B)^{\frac{s-c}{2(a+s)}} g_\alpha(B^*B)\right\| \|B^*(\hat{r}-r)\| \\ &\leq C\alpha^{-(c+2a+s)/(2(a+s))} \|B^*(\hat{r}-r)\|. \end{aligned}$$

We turn our attention to the bias term. Note that $L^s\varphi \in \mathcal{E}_{b-s}$. By Equation (3.2), there is a function $\rho \in \mathcal{E}$ such that

$$L^s\varphi = (B^*B)^{(b-s)/(2(a+s))}\,\rho.$$

We have

$$\begin{aligned}
\|L^c(\varphi_\alpha - \varphi)\| &= \left\|L^{(c-s)}\left(g_\alpha(B^*B)B^*B - I\right)L^s\varphi\right\| \\
&= \left\|\left(g_\alpha(B^*B)B^*B - I\right)(B^*B)^{(b-s)/(2(a+s))}\,\rho\right\|_{c-s} \\
&\le \left\|(B^*B)^{(s-c)/(2(a+s))}\left(g_\alpha(B^*B)B^*B - I\right)(B^*B)^{(b-s)/(2(a+s))}\,\rho\right\| \\
&= \left\|(B^*B)^{(b-c)/(2(a+s))}\left(g_\alpha(B^*B)B^*B - I\right)\rho\right\| \\
&\le C'\alpha^{(b-c)/(2(a+s))}\,\|\rho\|,
\end{aligned}$$

for some constant C', where the first inequality follows from (3.1) with $\nu = (s-c)/(a+s)$ and $\phi = \left(g_\alpha(B^*B)B^*B - I\right)(B^*B)^{(b-s)/(2(a+s))}\rho$ and the second inequality follows from (3.8) with $\mu = (b-c)/(2(a+s))$. Then using the optimal α, we obtain the rates given in Proposition 3.1.■

4. Asymptotic normality for fixed α

Let φ_0 be the true value of φ. As seen in Section 3, the estimator $\hat{\varphi}$ defined in (3.6) has a bias which does not vanish. For testing, it is useful to fix α and use $\hat{\varphi}$ minus a regularized version of φ_0 :

$$\varphi_{0\alpha} = L^{-s}g_\alpha(B^*B)B^*T\varphi_0 = L^{-s}g_\alpha(B^*B)B^*r. \tag{4.1}$$

Then, we have

$$\hat{\varphi} - \varphi_{0\alpha} = L^{-s}g_\alpha(B^*B)B^*(\hat{r} - r).$$

Depending on the examples, we will assume either Assumption 5a or Assumption 5b below.

Assumption 5a. $\sqrt{n}(\hat{r} - r) \Rightarrow \mathcal{N}(0,\Omega)$ in $\mathcal{F}$.

Under Assumption 5a, we have for a fixed α,

$$\sqrt{n}(\hat{\varphi} - \varphi_\alpha) \Rightarrow \mathcal{N}(0,\Sigma) \tag{4.2}$$

with $\Sigma = L^{-s}g_\alpha(B^*B)B^*\Omega Bg_\alpha(B^*B)L^{-s}$.

Assumption 5b. $\sqrt{n}B^*\left(\hat{r}-r\right) \Rightarrow \mathcal{N}\left(0,\Omega\right)$ in $\mathcal{F}$.

Under Assumption 5b, we have for a fixed α,

$$\sqrt{n}\left(\hat{\varphi}-\varphi_\alpha\right) \Rightarrow \mathcal{N}\left(0,\Sigma\right) \tag{4.3}$$

with $\Sigma = L^{-s}g_\alpha\left(B^*B\right)\Omega g_\alpha\left(B^*B\right)L^{-s}$.

The results (4.2) and (4.3) are the basis to construct the test statistics of the next section. If T is unknown, we have an extra term corresponding to $\hat{T}-T$ which is negligeable provided $\hat{T}$ converges sufficiently fast. We can check that either Assumption 5a or 5b is satisfied and the asymptotic variance Ω (and hence Σ) is estimable in all the examples considered here.

Example 1: Density

We have

$$\begin{aligned} \hat{r}-r &= \hat{F}-F=\frac{1}{n}\sum_{i=1}^{n}\left[I\left(x_i\leq t\right)-F\left(t\right)\right], \\ \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left[I\left(x_i\leq t\right)-F\left(t\right)\right] &\Rightarrow \mathcal{N}\left(0,F\left(t\wedge s\right)-F\left(t\right)F\left(s\right)\right). \end{aligned}$$

This example satisfies Assumption 5a. Here the asymptotic variance of $\hat{r}-r$ can be estimated using the empirical cumulative distribution function.

Example 2: Deconvolution

Following Carrasco and Florens (2011), we have

$$\widehat{T^*r}-T^*r=\frac{1}{n}\sum_{i=1}^{n}\left(\pi_{Y|X}\left(y_i|x\right)-E\left(\pi_{Y|X}\left(Y|x\right)\right)\right).$$

Here a slight modification of Assumption 5b is satisfied. Since $\pi_{Y|X}$ is known, the variance of $\widehat{T^*r}-T^*r$ can be estimated using the empirical variance.

Example 3: Functional linear regression

We have

$$\begin{aligned} \hat{r} &= \frac{1}{n}\sum_{i=1}^{n}Y_iW_i, \\ E\left(\widehat{r}\right) &= r. \end{aligned}$$

So that Assumption 5a holds and

$$V\left(\hat{r}-r\right)=\frac{1}{n}V\left(Y_iW_i\right)$$

can be estimated using the sample variance of Y_iW_i.

Example 4: Nonparametric instrumental regression

Following Darolles, Florens, and Renault (2002, Assumption A7), we assume that

$$\sqrt{n}\left(\widehat{T}^{*}\widehat{r}-\widehat{T}^{*}\widehat{T}\varphi\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_{i}-\varphi\left(Z_{i}\right)\right)\frac{f_{Z,W}\left(Z,W_{i}\right)}{f_{Z}\left(Z\right)f_{W}\left(W_{i}\right)}+h_{n}^{\rho}\Gamma \tag{4.4}$$

where the term $h_n^{\rho}\Gamma$ is negligible provided the bandwidth h_n is sufficiently small which is consistent with Assumption 5b. We denote the leading term in (4.4) by $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i$. We have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_{i}\Rightarrow N\left(0,\sigma^{2}T^{*}T\right)$$

where $\sigma^2=V(Y-\varphi(Z)\,|W)$. An estimate of σ^2 can be obtained using a first step estimator of φ.

5. Test statistics

5.1. Case where φ_0 is fully specified

We want to test $H_0:\varphi=\varphi_0$ where φ_0 is fully specified. A test can be based on the difference between $\hat{\varphi}$ and $\varphi_{0\alpha}$ defined in (4.1). We can construct a Kolmogorov-Smirnov test

$$\sup_{z}\sqrt{n}\left|\hat{\varphi}\left(z\right)-\varphi_{0\alpha}\left(z\right)\right|$$

or a Cramer-Von Mises test

$$\left\|\sqrt{n}\left(\hat{\varphi}-\varphi_{0\alpha}\right)\right\|^{2}.$$

Using (4.3), we have

$$\left\|\sqrt{n}\left(\hat{\varphi}-\varphi_{0\alpha}\right)\right\|^{2}\Rightarrow\sum_{j=1}^{\infty}\tilde{\lambda}_{j}\chi_{j}^{2}\left(1\right)$$

where χ_j^2 are independent Chi-square random variables and $\tilde{\lambda}_j$ are the eigenvalues of Σ. As Σ is estimable, $\tilde{\lambda}_j$ can be estimated by the eigenvalues of the estimate of Σ, see for instance Blundell and Horowitz (2007).

Another testing strategy consists in using a test function δ and base the test on a rescaled version of $\sqrt{n}\left\langle \hat{\varphi}-\varphi_{0\alpha},\delta\right\rangle$ to obtain a standard distribution.

$$\xi_n = \frac{\sqrt{n}\left\langle \hat{\varphi}-\varphi_{0\alpha},\delta\right\rangle}{\left\langle \widehat{\Sigma}\delta,\delta\right\rangle^{1/2}} \xrightarrow{d} \mathcal{N}\left(0,1\right). \tag{5.1}$$

A more powerful test can be obtained by considering a vector

$$\begin{pmatrix} \sqrt{n}\left\langle \hat{\varphi}-\varphi_{0\alpha},\delta_1\right\rangle \\ \vdots \\ \sqrt{n}\left\langle \hat{\varphi}-\varphi_{0\alpha},\delta_q\right\rangle \end{pmatrix}$$

for a given family δ_l, $l=1,2,...,q$ of linearly independent test functions of $\mathcal{E}$. This vector converges to a q dimensional normal distribution. The covariance between the various components of the vector can be easily deduced from (5.1) since it holds for any linear combinations of test functions δ_l and δ_h, $l \neq h$ chosen in the same space. Then, the appropriately rescaled statistic asymptotically follows a Chi-square distribution with q degrees of freedom.

5.2. Case where φ_0 is parametrically specified

We want to test $H_0 : \exists\theta \in \Theta, \varphi\left(.\right) = h\left(.,\theta\right)$ where h is a known function. Assume that we have an estimator of θ, $\hat{\theta}$, such that $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} \mathcal{N}\left(0,V\right)$. Then, a test statistics can be based on $\sqrt{n}\left(\hat{\varphi}-h_\alpha\left(\hat{\theta}\right)\right)$ where h_α is a regularized version of $h\left(\hat{\theta}\right)$, namely

$$h_\alpha\left(\hat{\theta}\right) = L^{-s}g_\alpha\left(B^*B\right)B^*Th\left(\hat{\theta}\right).$$

This testing strategy permits to eliminate the bias and is very similar in spirit to the twin-smoothing first proposed by Härdle and Mammen (1993) to test a parametric regression model against a nonparametric alternative (see also Fan (1994) and Altissimo and Mele (2009)).

Because θ is estimated, the asymptotic variance of $\sqrt{n}\left(\hat{\varphi}-h_\alpha\left(\hat{\theta}\right)\right)$ will differ from that of $\sqrt{n}\left(\hat{\varphi}-\varphi_{0\alpha}\right)$. We illustrate this point on two examples. In both examples, h is specified as $h\left(\theta\right) = \sum_{d=1}^{D}\theta_d e_d$ where e_d, $d=1,...,D$ are known functions and θ_d, $d=1,...,D$ are unknown scalars.

Functional linear regression

Consider the model (2.3) with exogenous regressors and homoskedastic error, $E\left(u_i Z_i\right)=0$ and $\mathrm{V}(u_i|Z_i)=\sigma^2$. Replacing φ by $h\left(\theta\right)$ in (2.3), we obtain

$$Y_i=\sum_{d=1}^{D}\theta_d\left\langle Z_i,e_d\right\rangle+u_i.$$

Denote $x_{i,d}=\left\langle Z_i,e_d\right\rangle$, X the $n\times D$ matrix of $x_{i,d}$, e the $D\times 1$ vector of e_d, and Y the $n\times 1$ vector of Y_i. Then, θ can be estimated by the OLS estimator, $\hat{\theta}=\left(X'X\right)^{-1}X'Y$. The estimator of $h\left(\theta\right)$ is given by

$$h\left(\hat{\theta}\right)=e'\left(X'X\right)^{-1}X'Y.$$

Consider standard Tikhonov regularization

$$\begin{aligned}\hat{\varphi}-h_\alpha\left(\hat{\theta}\right) &= \left(\alpha I+\widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\hat{r}\\ &\quad-\left(\alpha I+\widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\widehat{T}h\left(\hat{\theta}\right).\end{aligned}$$

Replacing $\widehat{r}$ by $\frac{1}{n}\sum_{i=1}^n Z_iY_i=\frac{1}{n}\sum_{i=1}^n Z_i\left\langle Z_i,\varphi\right\rangle+\frac{1}{n}\sum_{i=1}^n Z_iu_i=\widehat{T}\varphi+\frac{1}{n}\sum_{i=1}^n Z_iu_i$ and $h\left(\hat{\theta}\right)=e'\theta+e'\left(X'X\right)^{-1}X'u=\varphi+e'\left(X'X\right)^{-1}X'u$ (under H_0), we have

$$\begin{aligned}\hat{\varphi}-h_\alpha\left(\hat{\theta}\right) &= \left(\alpha I+\widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\left(\frac{1}{n}\sum_{i=1}^n Z_iu_i\right)\\ &\quad-\left(\alpha I+\widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\widehat{T}e'\left(\frac{X'X}{n}\right)^{-1}\frac{1}{n}X'u.\end{aligned}$$

Let us denote $A_n=\left(\alpha I+\widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*$ and $B_n=-\left(\alpha I+\widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\widehat{T}e'\left(\frac{X'X}{n}\right)^{-1}$. We obtain

$$\hat{\varphi}-h_\alpha\left(\hat{\theta}\right)=\left[A_n\ B_n\right]\frac{1}{n}\sum_{i=1}^n\begin{pmatrix}Z_i\\X_i\end{pmatrix}u_i.$$

Provided $E\left\|\begin{pmatrix}Z\\X\end{pmatrix}u\right\|^2<\infty$, we know from van der Vaart and Wellner (1996) that a central limit theorem holds, so that

$$\frac{\sqrt{n}}{n}\sum_{i=1}^n\begin{pmatrix}Z_i\\X_i\end{pmatrix}u_i\Rightarrow\mathcal{N}\left(0,\Gamma\right).$$

If moreover, $\|A_n - A\| \xrightarrow{P} 0$ and $\|B_n - B\| \xrightarrow{P} 0$, we have

$$\sqrt{n}\left(\hat{\varphi} - h_\alpha\left(\hat{\theta}\right)\right) \Rightarrow \mathcal{N}\left(0, [A\ B]\,\Gamma\begin{bmatrix} A^* \\ B^* \end{bmatrix}\right).$$

Nonparametric instrumental regression

Consider the model (2.4). The null hypothesis of interest is again $H_0 : \exists\theta \in \Theta$, $\varphi(.) = h(., \theta) = \sum_{d=1}^{D} \theta_d e_d$ for some known functions e_d. The finite dimensional parameter θ can be estimated by two-stage least squares. Denote W the $n \times D$ matrix $(W_1', ..., W_n')'$, Y the $n \times 1$ matrix $(Y_1, ..., Y_n)'$ and E the $n \times D$ matrix with (i, d) elements $e_d(Z_i)$. Denote P_W the projection matrix on W, $P_W = W(W'W)^{-1}W'$. Then the two-stage least-squares estimator of θ is

$$\hat{\theta} = \left(E'P_W E\right)^{-1} E'P_W Y \equiv M\frac{W'Y}{n}.$$

Using the notation e for the $D \times 1$ vector of functions $(e_1, ..., e_d, ..., e_D)'$, $h\left(., \hat{\theta}\right)$ takes the simple form

$$h\left(., \hat{\theta}\right) = e'\hat{\theta}.$$

Similarly to the previous example, we have

$$\begin{aligned}\sqrt{n}\left(\hat{\varphi} - h_\alpha\left(\hat{\theta}\right)\right) &= \left(\alpha I + \widehat{T}^*\widehat{T}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{f_{Z,W}(Z, W_i)}{f_Z(Z)\,f_W(W_i)}u_i\right. \\ &\qquad \left. - \widehat{T}^*\widehat{T}e'M\left(\frac{1}{n}\sum_{i=1}^{n}W_i u_i\right)\right] + o_p(1).\end{aligned}$$

Under some mild conditions (see van der Vaart and Wellner, 1996),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix}\frac{f_{Z,W}(Z,W_i)}{f_Z(Z)f_W(W_i)} \\ W_i\end{bmatrix}u_i$$

converges to a Gaussian process, which permits to establish the asymptotic variance of $\sqrt{n}\left(\hat{\varphi} - h_\alpha\left(\hat{\theta}\right)\right)$.

This test procedure can be related to that of Horowitz (2006). The test proposed by Horowitz (2006) is based on $\left\|\widehat{T}^*\left(\hat{r} - \widehat{T}h\left(\hat{\theta}\right)\right)\right\|^2$ while our test is based on $\left\|\left(\alpha I + \widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\left(\hat{r} - \widehat{T}h\left(\hat{\theta}\right)\right)\right\|^2$ with a fixed α.

6. Asymptotic normality for vanishing α

In this section, we are looking at conditions under which $\hat{\varphi} - \varphi$ is asymptotically normal when α goes to zero. There are various ways to state the results.

Carrasco and Florens (2011) and Horowitz (2007) prove a pointwise convergence:

$$\frac{\hat{\varphi}(z) - \varphi(z)}{\sqrt{\hat{V}(z)}} \xrightarrow{d} \mathcal{N}(0,1)$$

where typically the rate of convergence depends on z. Another possibility is to focus on the convergence of inner products:

$$\frac{\sqrt{n}\,\langle \hat{\varphi} - \varphi - b_n, \delta \rangle}{\langle \Sigma\delta, \delta \rangle^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)$$

where b_n is the bias corresponding to $\varphi_\alpha - \varphi$ and $\langle \Sigma\delta, \delta \rangle$ may be finite or infinite depending on the regularity of φ and δ.

We are going to focus on the second case.

6.1. Asymptotic normality with known operator

Here we consider the case of the Tikhonov regularization where T (hence B) is known. The case where B is estimated is studied in the next subsection.

We want to prove the asymptotic normality of $\sqrt{n}\,\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta \rangle$ where $c < s$ and $\hat{\varphi}$ is defined in (3.5):

$$\begin{aligned} \varphi_\alpha &= L^{-s}\left(\alpha I + B^*B\right)^{-1} B^* r, \\ \hat{\varphi} - \varphi_\alpha &= L^{-s}\left(\alpha I + B^*B\right)^{-1} L^{-s} T^*\left(\hat{r} - T\varphi\right). \end{aligned}$$

The following assumption will be used to strengthen Assumption 5a.

Assumption 6. η_i, $i = 1, 2, ..., n$ are iid with mean 0 and variance Ω and satisfy a functional CLT:

$$\frac{\sum_{i=1}^n \eta_i}{\sqrt{n}} \Rightarrow \mathcal{N}(0, \Omega) \text{ in } \mathcal{F}.$$

Define M such that $M = L^{c-s}\left(\alpha I + B^*B\right)^{-1} L^{-s}$.

Proposition 6.1. *Suppose that $\hat{\varphi}$ is as in (3.5). Assume that $T^*(\hat{r} - T\varphi) = \sum_{i=1}^n \eta_i/n$ where η_i satisfies Assumption 6. If $\delta \in \mathcal{E}$ satisfies*

$$\frac{E\left[|\langle M\eta_i, \delta\rangle|^{2+\varepsilon}\right]}{\left\|\Omega^{1/2}M^*\delta\right\|^{2+\varepsilon}} = O(1) \tag{6.1}$$

for some $\varepsilon > 0$, then

$$\frac{\sqrt{n}\,\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\rangle}{\left\|\Omega^{1/2}M^*\delta\right\|} \xrightarrow{d} \mathcal{N}(0,1). \tag{6.2}$$

Proof of Proposition 6.1. We have

$$\sqrt{n}\,\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\rangle = \frac{\sqrt{n}}{n}\sum_{i=1}^n \langle M\eta_i, \delta\rangle.$$

It follows from Assumption 6 that $\langle M\eta_i, \delta\rangle$ are iid with $\mathrm{Var}(\langle M\eta_i, \delta\rangle) = \frac{1}{n}\langle M\Omega M^*\delta, \delta\rangle = \frac{1}{n}\left\|\Omega^{1/2}M^*\delta\right\|^2$. A sufficient condition for the asymptotic normality is Lyapunov condition (Billingsley, 1995, (27.16)):

$$\lim_n \sum_{i=1}^n \frac{E\left[\left(\frac{\sqrt{n}}{n}\left|\langle M\eta_i, \delta\rangle\right|\right)^{2+\varepsilon}\right]}{\left\|\Omega^{1/2}M^*\delta\right\|^{2+\varepsilon}} = 0$$

for some $\varepsilon > 0$. By the stationarity, this relation simplifies to

$$\lim_n \frac{E\left[|\langle M\eta_i, \delta\rangle|^{2+\varepsilon}\right]}{n^{\varepsilon/2}\left\|\Omega^{1/2}M^*\delta\right\|^{2+\varepsilon}} = 0. \tag{6.3}$$

A sufficient condition for (6.3) is given by (6.1). The result follows.■

The rate of convergence of $\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\rangle$ will be slower than $\sqrt{n}$ if $\left\|\Omega^{1/2}M^*\delta\right\|$ diverges (which is the usual case). Moreover, the rate of convergence depends on the regularity of δ. The case of a $\sqrt{n}$ rate of convergence is discussed in Section 6.3. We see that condition (6.1) imposes in general restrictions on both η and δ.

First, we are going to investigate cases where condition (6.1) is satisfied for all δ. Assume there exists μ_i such that

$$L^{-s}\eta_i = B^*B\mu_i. \tag{6.4}$$

This is equivalent to say that $L^{-s}\eta_i \in \mathcal{R}(B^*B) = \mathcal{D}\left(L^{2(a+s)}\right)$ or equivalently $\eta_i \in \mathcal{D}\left(L^{2a+s}\right)$. Under Assumption (6.4), we have

$$\begin{aligned}
& \left|\langle M\eta_i, \delta\rangle\right| \\
= & \left|\left\langle L^{c-s}\left(\alpha I + B^*B\right)^{-1} B^*B\mu_i, \delta\right\rangle\right| \\
\leq & \left\|L^{c-s}\left(\alpha I + B^*B\right)^{-1} B^*B\mu_i\right\| \|\delta\| \\
\leq & C\|\mu_i\| \|\delta\|
\end{aligned}$$

for some constant C. If moreover $E\left(\|\mu_i\|^{2+\varepsilon}\right) < \infty$, Lyapunov condition (6.1) is satisfied for all δ such that $\|\delta\| < \infty$.

Now, we consider a more general case.

Assumption 7. $L^{-s}\eta_i \in \mathcal{R}\left(\left(B^*B\right)^{\nu/2}\right) = \mathcal{D}\left(L^{\nu(a+s)}\right)$ for some $0 \leq \nu \leq 2$.

Assumption 8. $L^{c-s}\delta \in \mathcal{R}\left(\left(B^*B\right)^{\tilde{\nu}/2}\right) = \mathcal{D}\left(L^{\tilde{\nu}(a+s)}\right)$ for some $\tilde{\nu} \geq 0$.

By a slight abuse of notation, we introduce the following μ_i and ρ:

$$\begin{aligned}
L^{-s}\eta_i &= \left(B^*B\right)^{\nu/2}\mu_i, \\
L^{c-s}\delta &= \left(B^*B\right)^{\tilde{\nu}/2}\rho.
\end{aligned}$$

We have

$$\begin{aligned}
& \left|\langle M\eta_i, \delta\rangle\right| \\
= & \left|\left\langle \left(\alpha I + B^*B\right)^{-1} L^{-s}\eta_i, L^{c-s}\delta\right\rangle\right| \\
= & \left|\left\langle \left(\alpha I + B^*B\right)^{-1} \left(B^*B\right)^{\nu/2}\mu_i, \left(B^*B\right)^{\tilde{\nu}/2}\rho\right\rangle\right| \\
= & \left|\left\langle \left(\alpha I + B^*B\right)^{-1} \left(B^*B\right)^{(\nu+\tilde{\nu})/2}\mu_i, \rho\right\rangle\right|.
\end{aligned}$$

If $\nu + \tilde{\nu} \geq 2$, this term is bounded by $\|\mu_i\| \|\rho\|$. If moreover, $E\left(\|\mu_i\|^{2+\varepsilon}\right) < \infty$, the condition (6.1) is satisfied and the asymptotic normality (6.2) holds for this specific δ.

6.2. Case where the operator is estimated

Let

$$\begin{aligned}
\hat{\varphi} &= L^{-s}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\hat{r} \\
\tilde{\varphi}_\alpha &= L^{-s}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\widehat{T}\varphi.
\end{aligned}$$

We want to study the asymptotic normality of

$$\langle L^c (\hat{\varphi} - \tilde{\varphi}_\alpha), \delta \rangle .$$

Assumption 9. $L^{c-s}\delta = (B^*B)^{d/2} \rho$ for some ρ with $\|\rho\| < \infty$.

Proposition 6.2. *Suppose that $\hat{\varphi}$ is as in (3.5). Assume that $\widehat{T}^* \left(\hat{r} - \widehat{T}\varphi \right) = \sum_{i=1}^n \eta_i / n$ for some η_i satisfying Assumption 6 and δ satisfy Assumption 9 and (6.1). If*

$$\frac{\sqrt{n}}{\alpha^{\frac{(3-d)\vee 1}{2}}} \frac{\left\| \hat{B} - B \right\|}{\|\Omega^{1/2} M^* \delta\|} \to 0 \tag{6.5}$$

and

$$\frac{\sqrt{n}}{\alpha^{\frac{(3-d)\vee 2}{2}}} \frac{\left\| \hat{B}^* - B^* \right\|}{\|\Omega^{1/2} M^* \delta\|} \to 0, \tag{6.6}$$

then

$$\frac{\sqrt{n} \langle L^c (\hat{\varphi} - \tilde{\varphi}_\alpha), \delta \rangle}{\|\Omega^{1/2} M^* \delta\|} \xrightarrow{d} \mathcal{N}(0,1) .$$

The notation $a \vee b$ means $\max(a,b)$. In the IV example, $\left\| \hat{B} - B \right\|$ depends on a bandwidth h_n. By choosing h_n in an appropriate way, Conditions (6.5) and (6.6) will be satisfied.

Proof of Proposition 6.2. We have

$$\begin{aligned} & L^c (\hat{\varphi} - \tilde{\varphi}_\alpha) \\ = \; & L^{c-s} (\alpha I + B^*B)^{-1} L^{-s} \widehat{T}^* \left(\hat{r} - \widehat{T}\varphi \right) && (6.7) \\ & + L^{c-s} \left\{ \left(\alpha I + \hat{B}^* \hat{B} \right)^{-1} - (\alpha I + B^*B)^{-1} \right\} L^{-s} \widehat{T}^* \left(\hat{r} - \widehat{T}\varphi \right). && (6.8) \end{aligned}$$

Using the fact that $\widehat{T}^* \left(\hat{r} - \widehat{T}\varphi \right) = \sum_{i=1}^n \eta_i / n$ for some η_i satisfying Assumption 6, we can establish the asymptotic normality of $\sqrt{n} \left\langle L^{c-s} (\alpha I + B^*B)^{-1} L^{-s} \widehat{T}^* \left(\hat{r} - \widehat{T}\varphi \right), \delta \right\rangle$ using the same proof as in Proposition 6.1.

Now we show that the term (6.8) is negligible. By Assumption 9, we have

$$
\begin{aligned}
&\left|\left\langle L^{c-s}\left\{\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}-\left(\alpha I+B^{*}B\right)^{-1}\right\}L^{-s}\widehat{T}^{*}\left(\hat{r}-\widehat{T}\varphi\right),\delta\right\rangle\right| \\
=\ &\left|\left\langle L^{-s}\frac{\sum_{i=1}^{n}\eta_i}{\sqrt{n}},\left\{\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}-\left(\alpha I+B^{*}B\right)^{-1}\right\}L^{c-s}\delta\right\rangle\right| \\
\leq\ &\left\|\frac{\sum_{i=1}^{n}L^{-s}\eta_i}{\sqrt{n}}\right\|\left\|\left\{\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}-\left(\alpha I+B^{*}B\right)^{-1}\right\}\left(B^{*}B\right)^{d/2}\rho\right\|
\end{aligned}
$$

The first term on the r.h.s is $O(1)$. We focus on the second term:

$$
\begin{aligned}
&\left\|\left\{\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}-\left(\alpha I+B^{*}B\right)^{-1}\right\}\left(B^{*}B\right)^{d/2}\rho\right\| \\
=\ &\left\|\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}\left(B^{*}B-\hat{B}^{*}\hat{B}\right)\left(\alpha I+B^{*}B\right)^{-1}\left(B^{*}B\right)^{d/2}\rho\right\| \\
=\ &\left\|\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}\left(\hat{B}^{*}\left(B-\hat{B}\right)+\left(B^{*}-\hat{B}^{*}\right)B\right)\left(\alpha I+B^{*}B\right)^{-1}\left(B^{*}B\right)^{d/2}\rho\right\| \\
\leq\ &\left\|\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}\hat{B}^{*}\left(B-\hat{B}\right)\left(\alpha I+B^{*}B\right)^{-1}\left(B^{*}B\right)^{d/2}\rho\right\| \qquad \text{(term1)} \\
&+\left\|\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}\left(B^{*}-\hat{B}^{*}\right)B\left(\alpha I+B^{*}B\right)^{-1}\left(B^{*}B\right)^{d/2}\rho\right\| \qquad \text{(term2)}
\end{aligned}
$$

Term1: We have $\left\|\left(\alpha I+\hat{B}^{*}\hat{B}\right)^{-1}\hat{B}^{*}\right\|^{2}\leq 1/\alpha$ and

$$
\left\|\left(\alpha I+B^{*}B\right)^{-1}\left(B^{*}B\right)^{d/2}\right\|^{2}\leq C\frac{1}{\alpha^{(2-d)}}
$$

for $d\leq 2$ (see Carrasco et al. 2007). If $d>2$, this term is bounded. So that

$$
\begin{aligned}
(\text{term1})^{2} &\leq \frac{1}{\alpha}\left\|\hat{B}-B\right\|^{2}\frac{1}{\alpha^{(2-d)\vee 0}} \\
&= \left\|\hat{B}-B\right\|^{2}\frac{1}{\alpha^{(3-d)\vee 1}}.
\end{aligned}
$$

Term2:

$$
\begin{aligned}
(\text{term2})^{2} &\leq \frac{1}{\alpha^{2}}\left\|\hat{B}^{*}-B^{*}\right\|^{2}\left\|B\left(\alpha I+B^{*}B\right)^{-1}\left(B^{*}B\right)^{d/2}\right\| \\
&= \frac{1}{\alpha^{2}}\left\|\hat{B}^{*}-B^{*}\right\|^{2}\frac{1}{\alpha^{(1-d)\vee 0}} \\
&= \left\|\hat{B}^{*}-B^{*}\right\|^{2}\frac{1}{\alpha^{(3-d)\vee 2}}.
\end{aligned}
$$

Under the assumptions of Proposition 6.2, $\sqrt{n}(6.8)/\left\|\Omega^{1/2}M^*\delta\right\|$ is negligible.■

6.3. Root n rate of convergence

The rate of convergence of $\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\rangle$ is $\sqrt{n}$ if $\left\|\Omega^{1/2}M^*\delta\right\|$ is bounded. A $\sqrt{n}$ rate of convergence may sound strange in a nonparametric setting. However, it should be noted that taking the inner product has a smoothing property. Moreover, a $\sqrt{n}$ rate will in general be obtained only for functions δ which are sufficiently smooth.

We can illustrate this point in the context of IV estimation where we set $s = c = 0$ to facilitate the exposition. In this case, $\Omega = T^*T$. Assuming δ satisfies Assumption 8, we have

$$\left\|\Omega^{1/2}M^*\delta\right\| = \left\|(T^*T)^{1/2}(T^*T + \alpha I)^{-1}(T^*T)^{\tilde{\nu}/2}\rho\right\|$$

which is finite if $\tilde{\nu} > 1$. Here it is always possible to choose ρ and then δ so that the inner product $\langle\hat{\varphi} - \varphi_\alpha, \delta\rangle$ converges at a $\sqrt{n}$ rate.

The root n rate of convergence of inner products has been discussed in various papers, e.g. Carrasco et al (2007, p.57) and Ai and Chen (2007, 2012) where an efficiency bound is derived. Severini and Tripathi (2012) derive the efficiency bound for estimating inner products of φ which remains valid when φ is not identified.

7. Selection of the regularization parameter

Engl et al. (2000) propose to select α using the following criterion:

$$\min_\alpha \frac{1}{\sqrt{\alpha}}\left\|\hat{r} - T\widehat{\varphi}_\alpha\right\|$$

and show that the resulting α has the optimal rate of convergence when T is known.

Darolles et al (2011) suggest a slightly different rule. Let $\widehat{\varphi}_{\alpha(2)}$ be the iterated Tikhonov estimator of order 2. Then α is chosen to minimize

$$\frac{1}{\alpha}\left\|\hat{T}^*\hat{r} - \hat{T}^*\hat{T}\widehat{\varphi}_{\alpha(2)}\right\|.$$

They show that this selection rule delivers an α with optimal speed of convergence for the model (2.4). See Fève and Florens (2010 and 2011) for the practical implementation of this method.

Other adaptive selection rules have been proposed for the IV model (2.4) but using different estimators than Darolles et al. Loubes and Marteau (2009) consider a spectral cut-off estimator and give a selection criterion of α such that the mean square error of the resulting estimator of φ achieves the optimal bound up to a $ln(n)^2$ factor. They assume that the eigenfunctions are known but the eigenvalues are estimated. Johannes and Schwarz (2010) consider an estimator combining spectral cut-off and thresholding. They show that their data-driven estimator can attain the lower risk bound up to a constant, provided the eigenfunctions are known trigonometric functions.

Recently, Horowitz (2011) proposed a selection rule which does not require the knowledge of the eigenfunctions and/or eigenvalues. The estimator considered in Horowitz (2011) is a modification of Horowitz's (2012) estimator. Let us briefly explain how to construct such an estimator. Multiply the left-hand and right-hand sides of Equation (2.5) by $f_W(w)$ to obtain

$$E\left(Y_i|W_i=w\right)f_W\left(w\right)=E\left(\varphi\left(Z_i\right)|W_i=w\right)f_W\left(w\right).$$

Now define $r(w)=E\left(Y_i|W_i=w\right)f_W(w)$ and $(T\varphi)(z)=\int\varphi(z)\,f_{Z,W}(z,w)\,dz$. Assume the support of Z and W is $[0,1]$. Let $\left\{\psi_j : j=1,2,...\right\}$ be a given complete orthonormal basis for $L^2[0,1]$. Contrary to Darolles et al., the ψ_j are not related to the eigenfunctions of T^*T. Then, T and r are approximated by a series expansion on this basis:

$$\begin{aligned}\hat{r}\left(w\right) &= \sum_{k=1}^{J_n}\hat{r}_k\psi_k\left(w\right),\\ \widehat{f}_{Z,W}\left(z,w\right) &= \sum_{j=1}^{J_n}\sum_{k=1}^{J_n}\hat{c}_{jk}\psi_j\left(z\right)\psi_k\left(w\right),\end{aligned}$$

where J_n is a nonstochastic truncation point and $\hat{r}_k$ and $\hat{c}_{jk}$ are estimated Fourier coefficients:

$$\begin{aligned}\hat{r}_k &= \frac{1}{n}\sum_{i=1}^{n}Y_i\psi_k\left(w_i\right),\\ \hat{c}_{jk} &= \frac{1}{n}\sum_{i=1}^{n}\psi_j\left(z_i\right)\psi_k\left(w_i\right).\end{aligned}$$

For any function $\nu : [0,1] \to \mathbb{R}$, define $D_j \nu(z) = d^j \nu(z)/dz^j$. Let

$$\mathcal{H}_{Js} = \left\{ \nu = \sum_{j=1}^{J} \nu_j \psi_j : \sum_{0 \le j \le s} \int_0^1 [D_j \nu(z)]^2 \, dz \le C_0 \right\}$$

for some finite $C_0 > 0$. Then Horowitz's (2011) sieve estimator is defined as

$$\tilde{\varphi} = \arg \min_{\nu \in \mathcal{H}_{J_n s}} \left\| \hat{T}\nu - \hat{r} \right\|.$$

For $j = 1, 2, ..., J_n$, define $\tilde{b}_j = \langle \tilde{\varphi}, \psi_j \rangle$. Let $J \le J_n$ be a positive integer, the modified estimator of φ considered in Horowitz (2012) is

$$\hat{\varphi}_J = \sum_{j=1}^{J} \tilde{b}_j \psi_j.$$

The optimal J, denoted J_{opt}, is defined as the value that minimizes the asymptotic mean square error (AIMSE) of $\hat{\varphi}_J$. The AIMSE is $E_A \|\hat{\varphi}_J - \varphi\|^2$ where $E_A(.)$ denotes the expectation of the leading term of the asymptotic expansion of (.). The selection rule is the following:

$$\hat{J} = \arg \min_{1 \le J \le J_n} \widehat{T}_n(J)$$

with

$$\widehat{T}_n(J) = \frac{2}{3} \frac{\ln(n)}{n^2} \sum_{i=1}^{n} \left\{ (Y_i - \tilde{\varphi}(W_i))^2 \sum_{j=1}^{J} \left(\left(\hat{T}^{-1} \right)^* \psi_j(W_i) \right)^2 \right\} - \|\hat{\varphi}_J\|^2.$$

For this $\hat{J}$,

$$E_A \left\| \hat{\varphi}_{\hat{J}} - \varphi \right\|^2 \le \left(2 + \frac{4}{3} \ln(n) \right) E_A \left\| \hat{\varphi}_{J_{opt}} - \varphi \right\|^2.$$

So $\hat{J}$ is not strictly speaking optimal but the rate of convergence in probability of $\left\| \hat{\varphi}_{\hat{J}} - \varphi \right\|^2$ is within a factor of $\ln(n)$ of the asymptotically optimal rate.

8. Implementation

We discuss the implementation in the four examples studied in Section 2.

8.1. Case where T is known

When T is known, the implementation is relatively simple.

Example 1: Density (continued):

The Tikhonov estimator of the density is given by the solution of

$$\min_{f} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t} f(u)\, du - \hat{F}(t) \right)^2 dt + \alpha \int_{-\infty}^{\infty} f^{(s)}(u)^2\, du$$

where f possesses s derivatives. This problem has a closed form solution (Vapnik, 1998, pages 309-311):

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} G_{\alpha}(x - x_i)$$

which is a kernel estimator with kernel

$$G_{\alpha}(x) = \int_{-\infty}^{\infty} \frac{e^{ix\omega}}{1 + \alpha \omega^{2(s+1)}} d\omega.$$

This formula simplifies when $s = 0$ (the desired density belongs to L^2):

$$G_{\alpha}(x) = \frac{1}{2\sqrt{\alpha}} \exp\left\{ -\frac{|x|}{\sqrt{\alpha}} \right\}.$$

Example 2: Deconvolution (continued):

We describe the estimator of Carrasco and Florens (2011). Given T and T^* are known, their spectral decompositions are also known (or can be approximated arbitrarily well by simulations). The solution f of $(\alpha I + T^* T) f = T^* h$ is given by

$$f(x) = \sum_{j=0}^{\infty} \frac{1}{\alpha + \lambda_j^2} \langle T^* h, \phi_j \rangle \phi_j(x).$$

The only unknown is $(T^* h)(x) = \int h(y) \pi_{Y|X}(y|x)\, dy = E\left[\pi_{Y|X}(Y|x)\right]$. It can be estimated by

$$\left(\widehat{T^* h}\right)(x) = \frac{1}{n} \sum_{i=1}^{n} \pi_{Y|X}(y_i|x)$$

so that the Tikhonov estimator of f is given by

$$\hat{f}(x) = \sum_{j=0}^{\infty} \frac{1}{\alpha + \lambda_j^2} \left\langle \frac{1}{n} \sum_{i=1}^{n} \pi_{Y|X}(y_i|.), \phi_j(.) \right\rangle \phi_j(x).$$

8.2. Case where T is estimated

Given that the number of observations is n, the estimated operators $\hat{T}$ and $\hat{T}^*$ are necessarily finite dimensional operators of dimensions that cannot exceed n. Assume that the operators $\hat{T}$ and $\hat{T}^*$ take the following forms:

$$\begin{aligned} \hat{T}\varphi &= \sum_{i=1}^{n} a_i(\varphi) f_i, \\ \hat{T}^*\psi &= \sum_{i=1}^{n} b_i(\psi) e_i, \end{aligned}$$

where f_i and e_i are elements of $\mathcal{F}$ and $\mathcal{E}$ respectively and a_i and b_i are linear functions. Assume r takes the form

$$\hat{r} = \sum_{i=1}^{n} c_i f_i.$$

Then, $\hat{T}^*\hat{T}\varphi + \alpha\varphi = \hat{T}^*\hat{r}$ can be rewritten as

$$\sum_{i=1}^{n} b_i\left(\sum_{j=1}^{n} a_j(\varphi) f_j\right) e_i + \alpha\varphi = \sum_{i=1}^{n} b_i\left(\sum_{j=1}^{n} c_j f_j\right) e_i \tag{8.1}$$

$$\sum_{i,j=1}^{n} b_i(f_j) a_j(\varphi) e_i + \alpha\varphi = \sum_{i,j=1}^{n} b_i(f_j) c_j e_i. \tag{8.2}$$

Now, applying a_l on the r.h.s. and l.h.s of (8.2) and using the linearity of the function a_l yields

$$\sum_{i,j=1}^{n} b_i(f_j) a_j(\varphi) a_l(e_i) + \alpha a_l(\varphi) = \sum_{i,j=1}^{n} b_i(f_j) c_j a_l(e_i). \tag{8.3}$$

We obtain n equations with n unknowns $a_j(\varphi)$, $j = 1, 2, ..., n$. We can solve this system and then replace $a_j(\varphi)$ by its expression in (8.1) to obtain φ. We illustrate this method in two examples.

Example 3: Functional linear regression (continued):
To simplify, let $\mathcal{E} = \mathcal{F} = L^2[0,1]$. We have

$$\begin{aligned}
\hat{T}\varphi &= \frac{1}{n}\sum_{i=1}^{n} \langle Z_i, \varphi \rangle W_i, \\
\hat{T}^*\psi &= \frac{1}{n}\sum_{i=1}^{n} \langle W_i, \psi \rangle Z_i, \\
\hat{r} &= \frac{1}{n}\sum_{i=1}^{n} Y_i W_i.
\end{aligned}$$

Then $f_i = W_i/n$, $e_i = Z_i/n$, $a_i(\varphi) = \langle Z_i, \varphi \rangle$, $b_i(\psi) = \langle W_i, \psi \rangle$, $c_i = Y_i$. Equation (8.3) gives

$$\alpha \langle \varphi, Z_l \rangle + \frac{1}{n^2}\sum_{i,j=1}^{n} \langle Z_i, Z_l \rangle \langle W_i, W_j \rangle \langle \varphi, Z_j \rangle = \frac{1}{n^2}\sum_{i,j=1}^{n} \langle Z_i, Z_l \rangle \langle W_i, W_j \rangle Y_j, \; l = 1, ..., n.$$

To compute the inner products $\langle Z_i, Z_l \rangle$, Florens and Van Bellegem (2012) propose to discretize the integrals as follows:

$$\langle Z_i, Z_l \rangle = \frac{1}{T}\sum_{t=1}^{T} Z_i\left(\frac{t}{T}\right) Z_l\left(\frac{t}{T}\right)$$

and the same for $\langle W_i, W_j \rangle$. Let Z and W denote the $T \times n$ matrices with (t,i) elements $Z_i(t/T)$ and $W_i(t/T)$ respectively. Let ξ and Y be the $n \times 1$ vectors of $\langle \varphi, Z_i \rangle$ and Y_i. Then, closed form expressions for ξ and φ are given by

$$\begin{aligned}
\xi &= \left(\alpha I + \frac{1}{n^2}\frac{Z'Z}{T}\frac{W'W}{T}\right)^{-1}\left(\frac{1}{n^2}\frac{Z'Z}{T}\frac{W'W}{T}Y\right), \\
\hat{\varphi} &= \frac{1}{\alpha n^2} Z \frac{W'W}{T}(Y - \xi).
\end{aligned}$$

Example 4: Nonparametric instrumental regression (continued):
In Darolles et al. (2002), the conditional expectation operator is estimated by

a kernel estimator with kernel k and bandwith h_n.

$$
\begin{aligned}
\widehat{T}\varphi &= \frac{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right)\varphi\left(z_i\right)}{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right)}, \\
\hat{T}^{*}\psi &= \frac{\sum_{i=1}^{n} k\left(\frac{z-z_i}{h_n}\right)\psi\left(w_i\right)}{\sum_{i=1}^{n} k\left(\frac{z-z_i}{h_n}\right)}, \\
\hat{r} &= \frac{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right)}.
\end{aligned}
$$

So that $f_i = \frac{k\left(\frac{w-w_i}{h_n}\right)}{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right)}$, $e_i = \frac{k\left(\frac{z-z_i}{h_n}\right)}{\sum_{i=1}^{n} k\left(\frac{z-z_i}{h_n}\right)}$, $a_i\left(\varphi\right) = \varphi\left(z_i\right)$, $b_i\left(\psi\right) = \psi\left(w_i\right)$, $c_i = y_i$.

Note that in Darolles et al. (2011), Z and W are assumed to have bounded supports $[0,1]^p$ and $[0,1]^q$ and a generalized kernel is used to avoid having a larger bias at the boundaries of the support.

Now we illustrate the role of L^{-s}. Consider $\mathcal{F}$ the space of square integrable functions defined on $[0,1]$ which satisfy the conditions $\phi\left(0\right) = 0$ and $\phi'\left(1\right) = 0$. The inner product on this space is defined by $\langle\phi,\psi\rangle = \int_0^1 \phi\left(x\right)\psi\left(x\right)dx$. Let $L\phi = -\phi''$ which satisfies all the properties of Hilbert scale (L is self-adjoint, etc). The estimator is given by

$$\hat{\varphi} = L^{-1}\left(\alpha I + L^{-1}T^{*}TL^{-1}\right)^{-1} L^{-1}K^{*}\hat{r}.$$

This approach is particularly useful if one is interested in the second derivative of φ since we have

$$\widehat{\varphi''} = \left(\alpha I + L^{-1}T^{*}TL^{-1}\right)^{-1} L^{-1}K^{*}\hat{r}.$$

Note that even if φ does not satisfy the boundary conditions $\varphi\left(0\right) = 0$ and $\varphi'\left(1\right) = 0$, $\hat{\varphi}$ satisfies these properties. It has no impact on the second derivatives. Moreover, we know that

$$L^{-1}\varphi = \int_0^1 \left(s \wedge t\right)\varphi\left(s\right)ds.$$

Hence L^{-1} can be approximated by a numerical integral:

$$L^{-1}\varphi = \frac{1}{N}\sum_{i=1}^{N}\left(s_i \wedge t\right)\varphi\left(s_i\right).$$

Florens and Racine (2012) propose an estimation procedure of the first partial derivative of φ by Landweber-Fridman. The paper derives the rate of convergence of the estimator, investigates the small-sample performance via Monte Carlo and applies the method to the estimation of the Engel Curve as in Blundell et al. (2007).

9. Concluding remarks

In this chapter, we mainly focused on the asymptotic normality of $\sqrt{n}\langle\hat{\varphi}-\varphi_\alpha,\delta\rangle$ and omitted to study the regularization bias. However, the bias has a form:

$$b_n = \varphi_\alpha - \varphi = -\alpha \sum_j \frac{1}{\lambda_j^2+\alpha}\left\langle \varphi, \varphi_j \right\rangle \varphi_j$$

which is estimable. Given a consistent α, denoted $\tilde{\alpha}$, we can construct a first-step estimator of φ denoted $\hat{\varphi}_{\tilde{\alpha}}$. Then an estimator of the bias is given by

$$\hat{b}_n = -\alpha \sum_j \frac{1}{\hat{\lambda}_j^2+\alpha}\left\langle \hat{\varphi}_{\tilde{\alpha}}, \hat{\varphi}_j \right\rangle \hat{\varphi}_j$$

where $\hat{\varphi}_j$ and $\hat{\lambda}_j$ are consistent estimators of φ_j and λ_j as described in Carrasco et al. (2007). Given this estimator, we can construct a bias-corrected estimator of $\hat{\varphi}$, $\hat{\varphi}-\hat{b}_n$. Although this estimator will have a smaller bias than the original one, it may have a larger variance.

References

[1] Ai, Chunrong, and Xiaohong Chen. 2007. "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables." *Journal of Econometrics*, 141, pp. 5-43.

[2] Ai, Chunrong, and Xiahong Chen. 2012. "Semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions." *Journal of Econometrics*, 170, pp. 442-457.

[3] Altissimo, Filippo, and Antonio Mele. 2009. "Simulated Nonparametric Estimation of Dynamic Models." *Review of Economic Studies,* 76, pp. 413-450.

[4] Billingsley, Patrick 1995. *Probability and Measure*. Wiley & Sons.

[5] Blundell, Richard, Xiaohong Chen, and Dennis Kristensen. 2007. "Semi-nonparametric IV estimation of shape-invariant Engle curves." *Econometrica*, 75, pp. 1613-1669.

[6] Blundell, Richard, and Joel Horowitz. 2007. "A nonparametric test of exogeneity." *Review of Economic Studies*, 74, pp. 1035-1058.

[7] Canay, Ivan, Andrés Santos, and Azeem Shaikh. 2011. "On the Testability of Identification in Some Nonparametric Models with Endogeneity." Mimeo, Northwestern.

[8] Cardot, Hervé, and Pascal Sarda. 2006. "Linear Regression Models for functional data." In *The Art of Semiparametrics*, ed. W. Härdle, pp. 49-66. Heidelberg: Physica-Verlag, Springer.

[9] Carrasco, Marine, and Jean-Pierre Florens. 2011. "Spectral method for deconvolving a density." *Econometric Theory*, 27 (3), pp. 546-581

[10] Carrasco, Marine, Jean-Pierre Florens, and Eric Renault. 2007. "Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization." In *Handbook of Econometrics*, ed. James Heckman and Edward Leamer, vol. 6B, pp. 5633–5746. Amsterdam: North Holland.

[11] Chen, Xiaohong, and Markus Reiss. 2011. "On rate optimality for ill-posed inverse problems in econometrics." *Econometric Theory*, 27, pp. 497-521.

[12] Chen, Xiaohong, and Halbert White. 1998. "Central Limit and Functional Central Limit Theorems for Hilbert Space-Valued Dependent Heterogenous Arrays with Applications." *Econometric Theory,* 14, pp. 260-284.

[13] Darolles, Serge, Jean-Pierre Florens, and Eric Renault. 2002. "Nonparametric Instrumental Regression." CRDE working paper, cahier 05-2002.

[14] Darolles, Serge, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. 2011. "Nonparametric Instrumental Regression." *Econometrica*, 79, pp. 1541-1565.

[15] Engl, Heinz W., Martin Hanke, and Andreas Neubauer. 2000. *Regularization of Inverse Problems*, Dordrecht: Kluwer Academic.

[16] Fan, Yanqin 1994. "Testing the Goodness-of-Fit of a Parameter Density Function by Kernel Method." *Econometric Theory*, 10, pp. 316-356.

[17] Ferraty, Frédéric, and Philippe Vieu. 2000. "Dimension fractale et estimation de la regression dans des espaces vectoriels semi-normes." *Comptes Rendus de l'Académie des Sciences de Paris, Série I, Mathématique*, 330, pp. 139-142.

[18] Fève, Frédéric, and Jean-Pierre Florens. 2010. "The practice of non parametric estimation by solving inverse problems: the example of transformation models." *The Econometrics Journal*, 13, pp. S1-S27.

[19] Fève, Frédéric, and Jean-Pierre Florens. 2011. "Non Parametric Analysis of Panel Data Models with Endogenous Variables." Mimeo, Toulouse School of Economics.

[20] Florens, Jean-Pierre. 2003. "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables." In *Advances in Economics and Econometrics: Theory and Applications*, ed. Mathias Dewatripont, Lars Peter Hansen, and Stephen Turnovsky, 2, pp. 284-311. Cambridge University Press.

[21] Florens, Jean-Pierre, and Sébastien Van Bellegem. 2012. "Functional Linear Instrumental Regression." mimeo.

[22] Florens, Jean-Pierre, Jan Johannes, and Sébastien Van Bellegem. 2011. "Identification and estimation by penalization in nonparametric instrumental regression." *Econometric Theory*, 27, pp. 472-496.

[23] Florens, Jean-Pierre, and Jeffrey Racine. 2012. "Nonparametric Instrumental Derivatives." Mimeo, McMaster University.

[24] Gagliardini, Patrick and Olivier Scaillet. 2012. "Tikhonov regularization for nonparametric instrumental variable estimators." *Journal of Econometrics*, 167, pp. 61-75.

[25] Groetsch, Charles 1993. *Inverse Problems in the Mathematical Sciences.* Braunschweig, Wiesbaden: Vieweg & Sohn.

[26] Hall, Peter and Joel Horowitz. 2005. "Nonparametric methods for inference in the presence of instrumental variables." *Annals of Statistics,* 33, pp. 2904-2929.

[27] Hall, Peter and Joel Horowitz. 2007. "Methodology and Convergence Rates for Functional Linear Regression." *Annals of Statistics*, 35, pp. 70-91.

[28] Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, 50, pp. 1029-1054.

[29] Härdle, Wolfgang, and Enno Mammen. 1993. "Comparing Nonparametric Versus Parametric Regression Fits." *Annals of Statistics*, 21, pp. 1926-1947.

[30] Horowitz, Joel. 2006. "Testing a Parametric Model Against a Nonparametric Alternative with Identification Through Instrumental Variables." *Econometrica*, 74, pp. 521-538.

[31] Horowitz, Joel. 2007. "Asymptotic Normality of a Nonparametric Instrumental Variables Estimator." *International Economic Review*, 48, pp. 1329-1349.

[32] Horowitz, Joel. 2011. "Adaptive nonparametric instrumental variables estimation: empirical choice of the regularization parameter." Mimeo, Northwestern University.

[33] Horowitz, Joel. 2012. "Specification testing in nonparametric instrumental variable estimation." *Journal of Econometrics*, 167, pp. 383-396.

[34] Johannes, Jan, and Maik Schwarz. 2010. "Adaptive nonparametric instrumental regression by model selection." arxiv:1003.3128v1, Université Catholique de Louvain.

[35] Johannes, Jan, Sébastien Van Bellegem, and Anne Vanhems. 2011. "Convergence rates for ill-posed inverse problems with an unknown operator." *Econometric Theory*, 27, 522-545.

[36] Kress, Rainer 1999. *Linear Integral Equations*, Applied Mathematical Sciences, New York: Springer-Verlag.

[37] Krein, S. G. and Yu I. Petunin. 1966. "Scales of Banach Spaces." *Russian Math. Surveys*, 21, pp. 85-160.

[38] Loubes, Jean-Michel, and Clément Marteau. 2009. "Oracle Inequality for Instrumental Variable Regression." arxiv:0901.4321v1, University of Toulouse 3, France.

[39] Newey, Whitney, and James Powell. 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica*, 71, pp. 1565-1578.

[40] Ramsay, James O., and Bernard W. Silverman. 1997. *Functional Data Analysis*. New York: Springer.

[41] Severini Thomas, and Gautam Tripathi. 2012. "Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors." *Journal of Econometrics*, 170, pp. 491-498.

[42] van der Vaart, Aad, and Jon Wellner. 1996. *Weak Convergence and Empirical Processes*. New York: Springer.

[43] Vapnik, Vladimir. 1998. *Statistical Learning Theory*, New York: John Wiley & Sons.