

DISCUSSION PAPER

2016/21

Bias-corrected condence intervals in a class
of linear inverse problems

Florens, J-P., Horowitz, J. and I. Van Keilegom

Bias-corrected confidence intervals in a class of linear inverse problems

Jean-Pierre FLORENS ^{*} Joel HOROWITZ [§] Ingrid VAN KEILEGOM ^{**}

April 15, 2016

Abstract

In this paper we propose a novel method to construct confidence intervals in a class of linear inverse problems. First, point estimators are obtained via a spectral cut-off method depending on a regularisation parameter α , that determines the bias of the estimator. Next, the proposed confidence interval corrects for this bias by explicitly estimating it based on a second regularisation parameter ρ , which is asymptotically smaller than α . The coverage error of the interval is shown to converge to zero. The proposed method is illustrated via two simulation studies, one in the context of functional linear regression, and the second one in the context of instrumental regression.

Key Words: Bias-correction; functional linear regression; instrumental regression; inverse problem; regularisation; spectral cut-off.

^{*}Toulouse School of Economics. Email address: jean-pierre.florens@tse-fr.eu.

[§]Northwestern University. Email address: joel-horowitz@northwestern.edu.

^{**}Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium. E-mail address: ingrid.vankeilegom@uclouvain.be. This research was supported by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy), by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 295298, and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain").

1 Introduction

The recent literature in econometrics has shown the importance of the theory of inverse problems as a conceptual framework for numerous questions related to functional estimation, including regression on functional data, or nonparametric instrumental regression (see e.g. Carrasco et al. (2007) for a survey of this approach). The central point of this approach is the inversion of operators (linear for simplicity) by means of regularisation methods (see e.g. Engl et al. (2000)). Examples of common regularisation methods are Tikhonov, Landweber-Fridman, spectral cut-off, etc. The application of these regularisation methods leads to estimators that converge at nonparametric rates, and are asymptotically normal but biased. The existence of this bias complicates the construction of confidence intervals.

The goal of this note is to adapt a procedure proposed by Hall and Horowitz (2013) to linear inverse problems in order to correct confidence intervals by means of an estimator of the bias that is based on a second regularisation parameter, smaller than the first one. This approach is frequent in nonparametric statistics (see e.g. Schucany and Sommers (1977) and Hall (1992)). Our approach is limited to a simple framework : the operator is defined on a function space but with values in $\mathbb{R}^n$, and the operator is supposed to be known. We restrict attention to regularisation via the spectral cut-off method, and our study concentrates on the estimation of a linear function of the functional parameter. This framework is nevertheless appropriate for the case of regression on functional data and the case of nonparametric regression with instrumental variables.

Engl et al. (2000) provide a detailed discussion of methods for inverse problems. For regression on functional data see e.g. Cardot and Johannes (2010) or Florens and Van Bellegem (2015) among many others. The model with instrumental variables is the topic of interest in e.g. Darolles et al. (2011) and Hall and Horowitz (2005).

The paper is organised as follows. In the next section, it is explained how to estimate linear functions of the functional parameter by means of the spectral cut-off method. Section 3 gives the properties of the proposed estimator, whereas bias-corrected confidence intervals are constructed in Section 4. These intervals require certain properties on the bias of the estimator, which are developed in Section 5. The coverage probability of the proposed confidence interval is obtained in Section 6, and the finite sample performance of the interval is studied in Section 7 by means of a simulation study.

2 Spectral cut-off estimation in a class of linear inverse problems

Consider the following linear model :

$$Y = K\varphi + U, \quad (2.1)$$

where φ belongs to a Hilbert space $\mathcal{E}$ equipped with scalar product $\langle \cdot, \cdot \rangle$, $Y = (Y_1, \dots, Y_n)^t$ is an element of $\mathbb{R}^n$ equipped with scalar product $\langle a_1, a_2 \rangle = n^{-1} \sum_{i=1}^n a_{1i} a_{2i}$ for $a_1, a_2 \in \mathbb{R}^n$, K is a linear and compact operator from $\mathcal{E}$ to $\mathbb{R}^n$, and $U \in \mathbb{R}^n$ satisfies $E(U) = 0$ and $\text{Var}(U) = \sigma^2 I_n$. Throughout, we will use the notation $\langle \cdot, \cdot \rangle$ to indicate both the scalar product in $\mathcal{E}$ and in $\mathbb{R}^n$, as it will be clear from the context in which space we are working. We suppose that K and σ^2 are known in order to simplify the presentation. Note that K depends on n , but we do not highlight this dependence, since we will work with fixed n .

Let us illustrate model (2.1) by means of two examples :

1. Example 1 : Functional linear regression

Let $\mathbb{R}$ be equipped with a measure π and define $\mathcal{E} = L^2(\mathbb{R}, \pi) = \{\varphi : \mathbb{R} \rightarrow \mathbb{R}; \int \varphi^2(x) \pi(dx) < \infty\}$. The operator K is based on n fixed elements of $\mathcal{E}$ denoted by $Z_1, \dots, Z_n$:

$$K\varphi = \left(\int_{\mathbb{R}} Z_i(x) \varphi(x) \pi(x) dx \right)_{i=1, \dots, n} = (\langle Z_i, \varphi \rangle)_{i=1, \dots, n}.$$

Hence, the model is in this case

$$Y_i = \langle Z_i, \varphi \rangle + U_i.$$

2. Example 2 : Instrumental regression

Let $(Y_1, Z_1, W_1), \dots, (Y_n, Z_n, W_n)$ be n identically distributed random vectors in $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$, and suppose that the marginal distribution of (Z_i, W_i) is known. Consider the model

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} E(\varphi(Z)|W = W_1) \\ \vdots \\ E(\varphi(Z)|W = W_n) \end{pmatrix} + \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix},$$

where $E(U_i|W_i) = 0$ and $\varphi \in \mathcal{E} = L^2(\mathbb{R}, f_Z)$, where f_Z is the density of Z_i . This defines implicitly the operator $K : L^2(\mathbb{R}, f_Z) \rightarrow \mathbb{R}^n$ by $(K\varphi)_i = E(\varphi(Z)|W = W_i)$, $i = 1, \dots, n$. We condition on the instruments $W_1, \dots, W_n$, so that the operator K is deterministic (but depending on n) as required.

Equation (2.1) defines an ill-posed inverse problem. The operator K is not injective, and hence not invertible in general. Moreover, the minimisation of $\|Y - K\varphi\|^2$ does not solve the problem in general. In fact, this minimisation leads to the normal equation

$$K^*Y = K^*K\varphi,$$

where $K^* : \mathbb{R}^n \rightarrow \mathcal{E}$ is the adjoint operator of K defined by $\langle K\varphi, a \rangle = \langle \varphi, K^*a \rangle$. The operator K^*K is not invertible neither, and its generalised inverse is not continuous, which necessitates the regularisation of K^*K via its inverse. Let us show the calculation of K^* for our two leading examples.

1. Functional linear regression (cont'd)

The operator K^* is given by $K^*a = n^{-1} \sum_{i=1}^n Z_i a_i$ for an arbitrary $a = (a_1, \dots, a_n)^t$. Indeed,

$$\langle K^*a, \varphi \rangle = \langle K\varphi, a \rangle = n^{-1} \sum_{i=1}^n \langle Z_i, \varphi \rangle a_i = \langle n^{-1} \sum_{i=1}^n Z_i a_i, \varphi \rangle.$$

Hence, in this example $K^*K\varphi = n^{-1} \sum_{i=1}^n Z_i \langle Z_i, \varphi \rangle$ is the empirical variance of $Z_1, \dots, Z_n$ applied to φ .

2. Instrumental regression (cont'd)

In this case, $K^*a(z) = n^{-1} \sum_{i=1}^n a_i [f_{Z|W}(z|W_i)/f_Z(z)]$, where $f_{Z|W}$ is the conditional density of Z_i given W_i . This follows from the following calculation :

$$\begin{aligned} \langle K^*a, \varphi \rangle &= \langle K\varphi, a \rangle = n^{-1} \sum_{i=1}^n E(\varphi(Z)|W = W_i) a_i \\ &= n^{-1} \sum_{i=1}^n \int \varphi(z) f_{Z|W}(z|W_i) dz a_i \\ &= \int \varphi(z) f_Z(z) \left[n^{-1} \sum_{i=1}^n a_i \frac{f_{Z|W}(z|W_i)}{f_Z(z)} \right] dz, \end{aligned}$$

where the scalar product in $\mathcal{E}$ is defined by $\langle \varphi_1, \varphi_2 \rangle = \int \varphi_1(z) \varphi_2(z) f_Z(z) dz$. Then, $K^*K\varphi(z) = n^{-1} \sum_{i=1}^n E(\varphi(Z)|W = W_i) [f_{Z|W}(z|W_i)/f_Z(z)]$.

The operator K admits a singular value decomposition $(\lambda_j)_{j=1,2,\dots}, (\varphi_j)_{j=1,2,\dots}, (\psi_j)_{j=1,\dots,n}$, where $\varphi_j \in \mathcal{E}$, $\psi_j \in \mathbb{R}^n$, $\lambda_j \in [0, \infty)$ for all j and $\lambda_j = 0$ for $j > n$. This means that $K^*K\varphi_j = \lambda_j^2 \varphi_j$, $KK^*\psi_j = \lambda_j^2 \psi_j$, $K\varphi_j = \lambda_j \psi_j$ and $K^*\psi_j = \lambda_j \varphi_j$. Moreover, $(\varphi_j)_j$ is an orthonormal basis of $\mathcal{E}$, and $(\psi_j)_j$ is an orthonormal basis of $\mathbb{R}^n$. Hence we can write

$\varphi = \sum_{j=1}^{\infty} \langle \varphi, \varphi_j \rangle \varphi_j$. Note that λ_j , φ_j and ψ_j depend on n (since K depends on n), but to keep the notation simple, we will omit the index n in the notation.

We will use this spectral value decomposition to regularise model (2.1). We restrict attention to the so-called spectral cut-off method :

$$\hat{\varphi}_\alpha = \sum_{\substack{j=1 \\ \lambda_j > \alpha}}^{\infty} \frac{1}{\lambda_j} \langle Y, \psi_j \rangle \varphi_j = \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2} \langle Y, \psi_j \rangle K^* \psi_j. \quad (2.2)$$

In the example on functional linear regression this means that

$$\hat{\varphi}_\alpha(x) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2} Y_i \psi_{ji} \left(\frac{1}{n} \sum_{\ell=1}^n Z_\ell(x) \psi_{j\ell} \right) \right\},$$

where the j -th eigenvector of the matrix KK^* is written as $(\psi_{j\ell})_{\ell=1,\dots,n}$. On the other hand, for the example on instrumental regression, we have

$$\hat{\varphi}_\alpha(z) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2} Y_i \psi_{ji} \left(\frac{1}{n} \sum_{\ell=1}^n \frac{f_{Z|W}(z|W_\ell)}{f_Z(z)} \psi_{j\ell} \right) \right\}.$$

In the first example the estimator is an element of the space generated by the Z_ℓ 's, whereas in the second example the estimator belongs to the space generated by the $f_{Z|W}(\cdot|W_\ell)/f_Z(\cdot)$'s.

We now turn to the estimation of continuous linear functions of φ , i.e. scalar products of the form $\theta = \langle \varphi, \mu \rangle$ thanks to Riesz' Theorem, where $\mu \in \mathcal{E}$ is a known function. In this case we have

$$\hat{\theta}_\alpha = \langle \hat{\varphi}_\alpha, \mu \rangle = \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j} \langle Y, \psi_j \rangle \langle \mu, \varphi_j \rangle = \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2} \langle Y, \psi_j \rangle \langle K\mu, \psi_j \rangle.$$

The advantage of this estimator is that it reduces the problem to an estimation problem in $\mathbb{R}^n$: the ψ_j 's and the λ_j^2 's are the eigenvectors and eigenvalues of KK^* , which is an $n \times n$ matrix.

3 Properties of the estimator

The estimator $\hat{\varphi}_\alpha$ defined in (2.2) is biased, and we can calculate its bias and variance. In fact,

$$\begin{aligned} E(\hat{\varphi}_\alpha) &= \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j} \langle E(Y), \psi_j \rangle \varphi_j = \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j} \langle K\varphi, \psi_j \rangle \varphi_j \\ &= \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j} \langle \varphi, K^* \psi_j \rangle \varphi_j = \sum_{\lambda_j > \alpha} \langle \varphi, \varphi_j \rangle \varphi_j \end{aligned}$$

and $E(\hat{\varphi}_\alpha) - \varphi = -\sum_{\lambda_j \leq \alpha} \langle \varphi, \varphi_j \rangle \varphi_j$. This bias is the rest of the development of a Fourier expansion of φ in the basis of φ_j functions by truncating the development at $\lambda_j > \alpha$.

The variance of $\hat{\varphi}_\alpha$ is obtained as follows : for an arbitrary $g \in \mathcal{E}$ and denoting $V = \text{Var}(\hat{\varphi}_\alpha)$, we have

$$\begin{aligned} V(g) &= E\left[(\hat{\varphi}_\alpha - E(\hat{\varphi}_\alpha))\langle \hat{\varphi}_\alpha - E(\hat{\varphi}_\alpha), g \rangle\right] \\ &= E\left[\left(\sum_{\lambda_j > \alpha} \frac{1}{\lambda_j} \langle U, \psi_j \rangle \varphi_j\right) \left\langle \sum_{\lambda_\ell > \alpha} \frac{1}{\lambda_\ell} \langle U, \psi_\ell \rangle \varphi_\ell, g \right\rangle\right] \\ &= E\left[\sum_{\lambda_j > \alpha, \lambda_\ell > \alpha} \frac{1}{\lambda_j \lambda_\ell} \langle U, \psi_j \rangle \langle U, \psi_\ell \rangle \langle \varphi_\ell, g \rangle \varphi_j\right]. \end{aligned}$$

Since $E[\langle U, \psi_j \rangle \langle U, \psi_\ell \rangle] = n^{-2} \sum_{i, i'=1}^n E(U_i U_{i'}) \psi_{ji} \psi_{\ell i'} = \sigma^2 n^{-2} \sum_{i=1}^n \psi_{ji} \psi_{\ell i} = \sigma^2 n^{-1} \langle \psi_j, \psi_\ell \rangle = \sigma^2 n^{-1} I(j = \ell)$, we have that

$$V(g) = \frac{\sigma^2}{n} \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2} \langle \varphi_j, g \rangle \varphi_j$$

and

$$\text{tr}(V) = \sum_j \langle V \varphi_j, \varphi_j \rangle = \frac{\sigma^2}{n} \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2}.$$

Hence,

$$MSE(\hat{\varphi}_\alpha) = E\|\hat{\varphi}_\alpha - \varphi\|^2 = \|E\hat{\varphi}_\alpha - \varphi\|^2 + \text{tr}(V) = \sum_{\lambda_j \leq \alpha} \langle \varphi, \varphi_j \rangle^2 + \frac{\sigma^2}{n} \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2}.$$

If the vector of errors U is normally distributed, we have

$$\hat{\varphi}_\alpha - \varphi \sim N\left(-\sum_{\lambda_j \leq \alpha} \langle \varphi, \varphi_j \rangle \varphi_j, V\right).$$

For the estimator $\hat{\theta}_\alpha = \langle \hat{\varphi}_\alpha, \mu \rangle$ of $\theta = \langle \varphi, \mu \rangle$ it follows that the bias equals

$$E(\hat{\theta}_\alpha) - \theta = -\sum_{\lambda_j \leq \alpha} \langle \varphi, \varphi_j \rangle \langle \varphi_j, \mu \rangle,$$

the variance equals

$$\text{Var}(\hat{\theta}_\alpha) = \frac{\sigma^2}{n} \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2} \langle \varphi_j, \mu \rangle^2,$$

and hence

$$\hat{\theta}_\alpha - \theta \sim N\left(-\sum_{\lambda_j \leq \alpha} \langle \varphi, \varphi_j \rangle \langle \varphi_j, \mu \rangle, \frac{\sigma^2}{n} \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^2} \langle \varphi_j, \mu \rangle^2\right).$$

Note that K is an operator from $\mathcal{E}$ (of infinite dimension) to $\mathbb{R}^n$, and hence it can only have n singular values that are non-zero. We suppose that these n singular values are strictly positive. The operator K is hence not injective and for fixed n the model is not identified. The identified part of φ and of θ corresponds to the eigenvectors φ_j for which the corresponding λ_j are non-zero. We therefore have to distinguish two parts in the bias b_α of $\hat{\theta}_\alpha$:

$$\begin{aligned} b_\alpha = E(\hat{\theta}_\alpha) - \theta &= - \sum_{\lambda_j \leq \alpha} \langle \varphi, \varphi_j \rangle \langle \varphi_j, \mu \rangle \\ &= - \sum_{0 < \lambda_j \leq \alpha} \langle \varphi, \varphi_j \rangle \langle \varphi_j, \mu \rangle - \sum_{\lambda_j = 0} \langle \varphi, \varphi_j \rangle \langle \varphi_j, \mu \rangle \\ &= b_{\alpha 1} + b_{\alpha 2}. \end{aligned}$$

The term $b_{\alpha 1}$ is the bias due to regularisation, whereas the term $b_{\alpha 2}$ is the bias due to under-identification.

We are now interested in the rate of convergence of the MSE of $\hat{\varphi}_\alpha$ and of $\hat{\theta}_\alpha$. This study necessitates certain regularity conditions on φ and μ with respect to the rate of decrease of the λ_j 's. Suppose for instance that

$$\lim_{n \rightarrow \infty} \sum_{j=1}^n \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_j^{2\beta}} < \infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \sum_{j=1}^n \frac{\langle \varphi_j, \mu \rangle^2}{\lambda_j^{2\gamma}} < \infty, \quad (3.1)$$

for some $0 < \beta, \gamma < \infty$, which are so-called source conditions. We focus attention on the case where $\gamma < 1$. In fact, if $\gamma \geq 1$, then $\mu = K^*v$ for some $v \in \mathbb{R}^n$ and hence $\langle \varphi, \mu \rangle = \langle \varphi, K^*v \rangle = \langle K\varphi, v \rangle = \langle E(Y), v \rangle$. The problem is then well-posed and does not necessitate inversion of K .

If we analyse the problem for n going to infinity, we have to suppose that the series $K = K_n$ converges to an identified limiting operator. We can then write

$$E(\hat{\theta}_\alpha) - \theta = - \sum_{0 < \lambda_j \leq \alpha} \lambda_j^{\beta+\gamma} \frac{\langle \varphi, \varphi_j \rangle}{\lambda_j^\beta} \frac{\langle \varphi_j, \mu \rangle}{\lambda_j^\gamma} - \sum_{\lambda_j = 0} \langle \varphi, \varphi_j \rangle \langle \varphi_j, \mu \rangle. \quad (3.2)$$

The square of the first term of (3.2) is bounded by $c\alpha^{2(\beta+\gamma)}$ and

$$c = \left(\sum_{0 < \lambda_j \leq \alpha} \frac{\langle \varphi, \varphi_j \rangle}{\lambda_j^\beta} \frac{\langle \varphi_j, \mu \rangle}{\lambda_j^\gamma} \right)^2 \leq \sum_{0 < \lambda_j \leq \alpha} \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_j^{2\beta}} \sum_{0 < \lambda_j \leq \alpha} \frac{\langle \varphi_j, \mu \rangle^2}{\lambda_j^{2\gamma}},$$

and this is bounded when n tends to infinity thanks to (3.1). The second term of (3.2) tends to zero for two reasons : (1) the number of λ_j 's that equal zero decreases when n tends to

infinity, and (2) $\langle \varphi, \varphi_j \rangle$ and $\langle \varphi_j, \mu \rangle$ tend to zero. We suppose that K^*K tends to its limit at a sufficiently fast rate so that the second term of (3.2) is asymptotically negligible compared to the first term. This hypothesis is verified if we suppose that K converges to a limit $\tilde{K}$ with eigenvectors $\tilde{\varphi}_j$ and eigenvalues $\tilde{\lambda}_j$, that λ_j converges to $\tilde{\lambda}_j$ uniformly in j , that

$$\sum_{j=1}^{\infty} \frac{\langle \varphi, \tilde{\varphi}_j \rangle^2}{\tilde{\lambda}_j^{2\beta}} < \infty \quad \text{and} \quad \sum_{j=1}^{\infty} \frac{\langle \tilde{\varphi}_j, \mu \rangle^2}{\tilde{\lambda}_j^{2\gamma}} < \infty,$$

and that $\tilde{\lambda}_j < \alpha$ for $j > n$.

This implies that

$$(E(\hat{\theta}_\alpha) - \theta)^2 = \left(\sum_{\lambda_j \leq \alpha} \lambda_j^{\beta+\gamma} \frac{\langle \varphi, \varphi_j \rangle}{\lambda_j^\beta} \frac{\langle \varphi_j, \mu \rangle}{\lambda_j^\gamma} \right)^2 (1 + o(1)) \leq c_1 \alpha^{2(\beta+\gamma)}$$

for some $0 < c_1 < \infty$. In the same way we can bound the variance of $\hat{\theta}_\alpha$:

$$\text{Var}(\hat{\theta}_\alpha) = \frac{\sigma^2}{n} \sum_{\lambda_j > \alpha} \frac{1}{\lambda_j^{2(1-\gamma)}} \frac{\langle \varphi_j, \mu \rangle^2}{\lambda_j^{2\gamma}} \leq \frac{c_2 \sigma^2}{n \alpha^{2(1-\gamma)}}$$

for some $0 < c_2 < \infty$. Hence, the MSE of $\hat{\theta}_\alpha$ is bounded by $c_1 \alpha^{2(\beta+\gamma)} + c_2 \sigma^2 / (n \alpha^{2(1-\gamma)})$, and this is minimal for α^2 proportional to $n^{-1/(\beta+1)}$, which leads to a MSE of the order $n^{-(\beta+\gamma)/(\beta+1)}$.

4 Confidence intervals

If we assume normality of the error U , a naive approach to calculate a confidence interval for θ would be to neglect the bias of $\hat{\theta}_\alpha$, which gives the classical interval

$$\hat{\theta}_\alpha \pm z(\delta) \tau_\alpha,$$

where $\tau_\alpha = (\text{Var}(\hat{\theta}_\alpha))^{1/2}$ is the standard deviation (which is known), $z(\delta)$ satisfies $P(N(0, 1) > z(\delta)) = \delta/2$, and $0 < 1 - \delta < 1$ is the desired probability of the interval. Neglecting the bias leads to an erroneous evaluation of the coverage probability. Following the method proposed by Hall and Horowitz (2013), we can in fact write the following :

$$\begin{aligned} & P(\hat{\theta}_\alpha - t \tau_\alpha \leq \theta \leq \hat{\theta}_\alpha + t \tau_\alpha) \\ &= P\left(-t \leq \frac{\hat{\theta}_\alpha - \theta}{\tau_\alpha} \leq t\right) = \Phi\left(t - \frac{b_\alpha}{\tau_\alpha}\right) - \Phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right), \end{aligned}$$

where $b_\alpha = \theta_\alpha - \theta = E(\hat{\theta}_\alpha) - \theta$. Hence, in order to obtain an interval of coverage probability $1 - \delta$, it suffices to solve

$$\Phi\left(t - \frac{b_\alpha}{\tau_\alpha}\right) - \Phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right) = 1 - \delta$$

with respect to t , where $\Phi(\cdot)$ is the distribution function of a standard normal random variable. This can however not be solved in practice, since the bias b_α of $\hat{\theta}_\alpha$ is unknown. We will therefore estimate this bias, based on a second regularisation parameter $\rho > 0$.

The proposed procedure is as follows :

1. Estimation of φ and of θ based on a regularisation parameter ρ , smaller than α :

$$\hat{\varphi}_\rho = \sum_{\lambda_j > \rho} \frac{1}{\lambda_j} \langle Y, \psi_j \rangle \varphi_j \quad \text{and} \quad \hat{\theta}_\rho = \sum_{\lambda_j > \rho} \frac{1}{\lambda_j} \langle Y, \psi_j \rangle \langle \varphi_j, \mu \rangle.$$

2. Estimation of the bias of $\hat{\theta}_\alpha$ by

$$\hat{b}_{\alpha\rho} = \hat{\theta}_\alpha - \hat{\theta}_\rho.$$

3. Calculation of $\hat{t}$, which is the solution of

$$\Phi\left(t - \frac{\hat{b}_{\alpha\rho}}{\tau_\alpha}\right) - \Phi\left(-t - \frac{\hat{b}_{\alpha\rho}}{\tau_\alpha}\right) = 1 - \delta$$

with respect to t , and calculation of the confidence interval

$$\hat{\theta}_\alpha \pm \hat{t} \tau_\alpha.$$

5 Properties of the estimator of the bias

An easy calculation shows that

$$\hat{b}_{\alpha\rho} = - \sum_{\rho < \lambda_j \leq \alpha} \frac{1}{\lambda_j} \langle Y, \psi_j \rangle \langle \varphi_j, \mu \rangle,$$

from which we derive that

$$E(\hat{b}_{\alpha\rho}) - b_\alpha = \sum_{\lambda_j \leq \rho} \langle \varphi, \varphi_j \rangle \langle \varphi_j, \mu \rangle,$$

and

$$\text{Var}(\hat{b}_{\alpha\rho}) = \frac{\sigma^2}{n} \sum_{\rho < \lambda_j \leq \alpha} \frac{1}{\lambda_j^2} \langle \varphi_j, \mu \rangle^2.$$

Moreover, $\hat{b}_{\alpha\rho} - b_\alpha$ follows a normal distribution if we assume that the error vector U is normal.

In order to justify our estimation procedure of the bias, we will need that $E[(\hat{b}_{\alpha\rho} - b_\alpha)^2] = o(b_\alpha^2)$ for appropriate choices of α and ρ . We will first show why this property holds true. In the next section, we will use this property to study the coverage error of our bias-corrected interval.

Let us suppose that we have a method to choose α^2 proportional to $n^{-1/(\beta+1)}$, such that the squared bias and the variance are exactly proportional to $n^{-(\beta+\gamma)/(\beta+1)}$. We then have the following result.

Proposition 5.1. *Let $\rho^2 = \alpha^2 a(n)$ and let α^2 be proportional to $n^{-1/(\beta+1)}$. Moreover, suppose that*

1. *the source conditions (3.1) hold.*
2. *$a(n) \rightarrow 0$ when n tends to infinity.*
- 3.

$$a(n)^{-(1-\gamma)} \sum_{\rho < \lambda_j \leq \alpha} \frac{1}{\lambda_j^{2\gamma}} \langle \varphi_j, \mu \rangle^2 \rightarrow 0.$$

Then,

$$E[(\hat{b}_{\alpha\rho} - b_\alpha)^2] = o(b_\alpha^2).$$

Proof. Define

$$A = n^{\frac{\beta+\gamma}{\beta+1}} \left(\sum_{\lambda_j \leq \rho} \langle \varphi, \varphi_j \rangle \langle \varphi_j, \mu \rangle \right)^2 \quad \text{and} \quad B = n^{\frac{\beta+\gamma}{\beta+1}-1} \sum_{\rho < \lambda_j \leq \alpha} \frac{1}{\lambda_j^2} \langle \varphi_j, \mu \rangle^2.$$

We need to show that A and B tend to zero as n tends to infinity. First note that

$$A \leq n^{\frac{\beta+\gamma}{\beta+1}} \rho^{2(\beta+\gamma)} \left(\sum_{\lambda_j \leq \rho} \frac{\langle \varphi, \varphi_j \rangle}{\lambda_j^\beta} \frac{\langle \mu, \varphi_j \rangle}{\lambda_j^\gamma} \right)^2.$$

Now, the Cauchy-Schwarz inequality yields that

$$\lim_{n \rightarrow \infty} \left(\sum_{\lambda_j \leq \rho} \frac{\langle \varphi, \varphi_j \rangle}{\lambda_j^\beta} \frac{\langle \mu, \varphi_j \rangle}{\lambda_j^\gamma} \right)^2 \leq \lim_{n \rightarrow \infty} \sum_{j=1}^n \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_j^{2\beta}} \times \lim_{n \rightarrow \infty} \sum_{j=1}^n \frac{\langle \mu, \varphi_j \rangle^2}{\lambda_j^{2\gamma}} < \infty$$

and $n^{\frac{\beta+\gamma}{\beta+1}} \rho^{2(\beta+\gamma)} = a(n)^{2(\beta+\gamma)} \rightarrow 0$.

Next, consider B :

$$B \leq n^{-\frac{1-\gamma}{\beta+1}} \rho^{-2(1-\gamma)} \sum_{\rho < \lambda_j \leq \alpha} \frac{\langle \mu, \varphi_j \rangle^2}{\lambda_j^{2\gamma}} = a(n)^{-(1-\gamma)} \sum_{\rho < \lambda_j \leq \alpha} \frac{\langle \mu, \varphi_j \rangle^2}{\lambda_j^{2\gamma}} \rightarrow 0,$$

which finishes the proof. $\square$

The condition $a(n) \rightarrow 0$ is natural and implies that ρ should be chosen smaller than α . The third condition in the proposition is less obvious and needs some comments. Suppose that $\lambda_j = j^{-k}$ for some $k > 0$ (and for any n) and that $\langle \mu, \varphi_j \rangle^2 / \lambda_j^{2\gamma} \sim 1/j^{1+\varepsilon}$ for some arbitrary small ε . By supposing that $\lambda_j = j^{-k}$ we obtain a mildly ill-posed problem, which is in line with the source conditions that we supposed earlier. Moreover, if γ is chosen as large as possible, so as to maintain the integrability of the series $\sum_j \langle \mu, \varphi_j \rangle^2 / \lambda_j^{2\gamma}$, it is natural to suppose that the series differs very little from the harmonic series : $\sum_{j=1}^{\infty} j^{-1}$ is divergent, but $\sum_{j=1}^{\infty} j^{-1-\varepsilon}$ converges for all $\varepsilon > 0$. We can hence write

$$B = O\left(a(n)^{-(1-\gamma)} \sum_{\rho < \lambda_j \leq \alpha} \frac{1}{j^{1+\varepsilon}}\right) = O\left(a(n)^{-(1-\gamma)} m(\alpha)^{-\varepsilon} \sum_{j=m(\alpha)}^{m(\rho)} \frac{1}{j}\right),$$

where $m(\alpha)$ is the largest integer smaller than $\alpha^{-1/k}$ (and similar for $m(\rho)$). Next, note that $\sum_{j=1}^{L-1} j^{-1} = \Psi(L) + \gamma$, where Ψ is the digamma function and γ is the Euler constant. By approximating the function Ψ by the log-function (see Abramowitz and Stegun (1972)), we have

$$B = O\left(a(n)^{-(1-\gamma)} m(\alpha)^{-\varepsilon} (\log m(\rho) - \log m(\alpha))\right) = O\left(\log(a(n)^{-1}) a(n)^{-(1-\gamma)} \alpha^{\varepsilon/k}\right),$$

and this converges to zero if for instance $a(n)^{-1} = O(\log n)$ or $a(n)^{-1} = O(n^b)$ for sufficiently small b so that $b < \varepsilon/[2k(\beta+1)(1-\gamma)]$. Hence, $a(n)$ should tend to zero, but not faster than $(\log n)^{-1}$ respectively n^{-b} .

6 Coverage error

Recall that the proposed confidence interval is

$$I = [\hat{\theta}_\alpha - \hat{t}\tau_\alpha, \hat{\theta}_\alpha + \hat{t}\tau_\alpha],$$

where $\hat{t}$ is the solution of

$$\Phi\left(t - \frac{\hat{b}_{\alpha\rho}}{\tau_\alpha}\right) - \Phi\left(-t - \frac{\hat{b}_{\alpha\rho}}{\tau_\alpha}\right) = 1 - \delta.$$

Let us now calculate the coverage probability of the interval I :

$$p = P(\hat{\theta}_\alpha - \hat{t}\tau_\alpha \leq \theta \leq \hat{\theta}_\alpha + \hat{t}\tau_\alpha).$$

Some elementary calculations show that

$$p = \Phi\left(\hat{t} - \frac{b_\alpha}{\tau_\alpha}\right) - \Phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right) + o_P(1).$$

We will show that under certain conditions p is close to $1 - \delta$.

Proposition 6.1. *If $E[(\hat{b}_{\alpha\rho} - b_\alpha)^2] = o(b_\alpha^2)$ and if α^2 is proportional to $n^{-1/(\beta+1)}$, then*

$$p = 1 - \delta + o_P(1).$$

Proof. The proof is based on two linear approximations. First of all, we have that

$$\left[\Phi\left(\hat{t} - \frac{\hat{b}_{\alpha\rho}}{\tau_\alpha}\right) - \Phi\left(-\hat{t} - \frac{\hat{b}_{\alpha\rho}}{\tau_\alpha}\right)\right] - \left[\Phi\left(t - \frac{b_\alpha}{\tau_\alpha}\right) - \Phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right)\right] = 0,$$

where t is chosen such that

$$\Phi\left(t - \frac{b_\alpha}{\tau_\alpha}\right) - \Phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right) = 0.$$

The implicit function theorem implies that

$$\hat{t} - t = \frac{\phi\left(t - \frac{b_\alpha}{\tau_\alpha}\right) - \phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right)}{\phi\left(t - \frac{b_\alpha}{\tau_\alpha}\right) + \phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right)} \left(\frac{\hat{b}_{\alpha\rho} - b_\alpha}{\tau_\alpha}\right) + o_P\left(\frac{\hat{b}_{\alpha\rho} - b_\alpha}{\tau_\alpha}\right),$$

where ϕ is the density of the standard normal variable. In addition, a Taylor expansion of order 1 yields :

$$\begin{aligned} p &= (1 - \delta) + \left[\phi\left(t - \frac{b_\alpha}{\tau_\alpha}\right) + \phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right)\right](\hat{t} - t) + o_P(\hat{t} - t) \\ &= (1 - \delta) + \left[\phi\left(t - \frac{b_\alpha}{\tau_\alpha}\right) - \phi\left(-t - \frac{b_\alpha}{\tau_\alpha}\right)\right]\left(\frac{\hat{b}_{\alpha\rho} - b_\alpha}{\tau_\alpha}\right) + o_P\left(\frac{\hat{b}_{\alpha\rho} - b_\alpha}{\tau_\alpha}\right). \end{aligned} \quad (6.1)$$

In addition, note that b_α is of the same order as τ_α if α^2 is proportional to $n^{-1/(\beta+1)}$. Hence the expression between square brackets in (6.1) is $O(1)$. Finally, if $E[(\hat{b}_{\alpha\rho} - b_\alpha)^2] = o(b_\alpha^2) = o(\tau_\alpha^2)$, then $\hat{b}_{\alpha\rho} - b_\alpha = o_P(\tau_\alpha)$, and hence the result follows. $\square$

7 Simulations

We consider our two leading examples (functional linear regression and instrumental regression) in a simulation study, in order to investigate the small sample performance of the proposed bias-corrected confidence intervals.

The regularisation parameter α is chosen as the optimal one :

$$\alpha = \operatorname{argmin}_{\alpha} (\hat{b}_{\alpha\rho}^2 + \tau_{\alpha}^2),$$

where ρ is chosen as $\rho = \alpha n^{-1}$. This empirical choice of ρ satisfies the conditions needed at the end of Section 5 for appropriate values of β, γ, k and ε , and works well in practice for a wide range of models and sample sizes. Note that the above choice of α is legitimate, since $\hat{b}_{\alpha\rho}^2 + \tau_{\alpha}^2$ is asymptotically equivalent to $b_{\alpha}^2 + \tau_{\alpha}^2$ if $\hat{b}_{\alpha\rho} = b_{\alpha} + o(\tau_{\alpha})$, and this is the case thanks to Proposition 5.1.

7.1 Functional linear regression

We consider the following model :

$$Y_i = \langle Z_i, \varphi \rangle + U_i \quad (i = 1, \dots, n),$$

where $U_i \sim N(0, \sigma^2)$, $\langle Z_i, \varphi \rangle = \int Z_i(x) \varphi(x) dx$, $\varphi(x) = \exp(-ax)$ with $a > 0$,

$$Z_i(x) = P_i \frac{B_i^{A_i}}{\Gamma(A_i)} x^{A_i-1} \exp(-B_i x),$$

which corresponds (upto the constant P_i) to the curve of a Gamma density for different values of A_i and B_i . Here we take $A_i \sim \text{Unif}[0.5, 2.5]$, $B_i \sim \text{Unif}[0, 1]$ and $P_i \sim \text{Unif}[1, 3]$ in order to obtain a rich collection of curves of different shapes and amplitudes. The vectors $(A_1, B_1, P_1, U_1), \dots, (A_n, B_n, P_n, U_n)$ are independent. It is easy to show that

$$\langle Z_i, \varphi \rangle = P_i \frac{B_i^{A_i}}{(B_i + a)^{A_i}}.$$

We consider two choices for the function $\mu(\cdot) : \mu(x) = \exp(-0.5x)$ which corresponds to $\theta = \langle \varphi, \mu \rangle = 1/(a + 0.5)$, and $\mu(x) = I(0.5 \leq x \leq 1.5)$, which leads to $\theta = [\exp(-0.5a) - \exp(-1.5a)]/a$. Table 1 (which is based on 500 simulated samples) shows that the bias correction works well in practice. The coverage error is most of the time reduced by at least 50%, depending on the situation.

7.2 Instrumental regression

Next, consider the regression model

$$Y = E(\varphi(Z)|W) + U,$$

a	σ	n	$\mu(x) = \exp(-0.5x)$		$\mu(x) = I(0.5 \leq x \leq 1.5)$	
			Non-corrected	Corrected	Non-corrected	Corrected
0.5	0.2	100	.926	.942	.878	.956
		200	.856	.912	.858	.958
		400	.892	.910	.834	.928
	0.3	100	.926	.942	.908	.972
		200	.882	.920	.896	.976
		400	.916	.950	.888	.966
1.0	0.2	100	.916	.940	.856	.950
		200	.830	.884	.798	.942
		400	.906	.930	.786	.904
	0.3	100	.924	.944	.902	.970
		200	.870	.924	.882	.970
		400	.894	.930	.872	.954

Table 1: Coverage probabilities of the classical (non-corrected) and the bias-corrected 95% confidence intervals for several choices of $\varphi(x) = \exp(-ax)$, σ , n and $\mu(\cdot)$.

where U is independent of (Z, W) and is normally distributed with variance s^2 . The function φ equals $\varphi(z) = az^2$ for some $a \in \mathbb{R}$, the instrument W has density $f_W(w) = 2wI(0 \leq w \leq 1)$ and the endogenous variable Z is defined as $Z = RW$, where R is independent of all other variables and has a uniform distribution on the interval $[1 - b, 1 + b]$ for some $b > 0$. Hence, it can be easily seen that $E(\varphi(Z)|W) = a(1 + b^2/3)W^2$.

We can also write the model as

$$Y = \varphi(Z) + V,$$

where $V = E(\varphi(Z)|W) - \varphi(Z) + U = a(1 + b^2/3)W^2 - aZ^2 + U$ has mean zero and it can be easily seen that its variance is equal to

$$\sigma^2 = \text{Var}(V) = s^2 + a^2 \left(\frac{1}{12} + \frac{b^2}{2} + \frac{7b^4}{180} \right) - \frac{a^2}{12} \left(1 + \frac{b^2}{3} \right)^2.$$

In addition, V is uncorrelated with W ,

$$\rho(Z, V) = \text{Corr}(V, Z) = -\frac{12ab^2}{45\sigma\sqrt{\frac{1}{18} + \frac{b^2}{6}}},$$

and

$$\rho(Z, W) = \text{Corr}(Z, W) = \frac{1}{\sqrt{1 + 3b^2}}.$$

Hence, b determines the strength of the instrument, and a , b and σ determine the degree of endogeneity. We carry out simulations for several choices of a , b and σ , leading to a variety of values for $\rho(Z, V)$ and $\rho(Z, W)$. They can be found in Table 2.

Consider now i.i.d. data (Z_i, W_i, Y_i) , $i = 1, \dots, n$ having the same distribution as (Z, W, Y) . We are interested in constructing a confidence interval for $\theta = \langle \varphi, \mu \rangle$, where $\mu(z) = z^{-2}$. Hence, $\theta = a$. We calculate the proposed bias-corrected confidence interval, as well as the classical confidence interval. The corresponding coverage probabilities (based on 500 simulations) for samples of size 100, 200 and 400 are given in Table 3. The table shows that the proposed interval reduces the coverage error substantially compared to the classical confidence interval, and that the performance is better when the degree of endogeneity is small and/or the instrument is strong, as can be expected.

a	b	$\sigma = 0.3$		$\sigma = 0.4$	
		$\rho(Z, W)$	$\rho(Z, V)$	$\rho(Z, W)$	$\rho(Z, V)$
0.25	0.4	.822	-.124	.822	-.093
	0.6	.693	-.235	.693	-.177
0.5	0.4	.822	-.248	.822	-.186
	0.6	.693	-.471	.693	-.353

Table 2: Correlations $\rho(Z, W)$ and $\rho(Z, V)$ for several choices of $\varphi(z) = az^2$, b and σ .

a	b	n	$\sigma = 0.3$		$\sigma = 0.4$	
			Non-corrected	Corrected	Non-corrected	Corrected
0.25	0.4	100	.922	.952	.866	.934
		200	.908	.980	.870	.954
		400	.900	.974	.864	.944
	0.6	100	.816	.904	.868	.936
		200	.822	.912	.842	.936
		400	.826	.950	.828	.964
0.5	0.4	100	.906	.950	.900	.948
		200	.902	.954	.890	.970
		400	.878	.966	.892	.954
	0.6	100	.822	.866	.780	.858
		200	.766	.872	.792	.878
		400	.696	.836	.764	.900

Table 3: Coverage probabilities of the classical (non-corrected) and the bias-corrected 95% confidence intervals for several choices of $\varphi(z) = az^2$, b , n and σ .

References

- Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Applied Mathematics Series 55. Washington D.C., USA.
- Cardot, H. and Johannes, J. (2010). Thresholding projection estimators in functional linear models. *J. Multiv. Anal.*, **101**, 395-408.
- Carrasco, M., Florens, J.-P. and Renault, E. (2007). Linear inverse problems in structural econometrics : Estimation based on spectral decomposition and regularization. In: J. Heckman and E. Leamer, eds, *Handbook of Econometrics*, Elsevier, **6B**, 5633-5751.
- Darolles, S., Fan, Y., Florens, J.-P. and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, **79**, 1541-1565.
- Engl, H.W., Hanke, M. and Neubauer, A. (2000). *Regularization of Inverse Problems*. Springer.
- Florens, J.-P. and Van Bellegem, S. (2015). Instrumental variable estimation in functional

- linear models. *J. Econometrics*, **186**, 465-476.
- Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *Ann. Statist.*, **20**, 695-711.
- Hall, P. and Horowitz, J. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.*, **33**, 2904-2929.
- Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *Ann. Statist.*, **41**, 1892-1921.
- Schucany, W.R. and Sommers, J.P. (1977). Improvement of kernel density estimators. *J. Amer. Statist. Assoc.*, **72**, 420-423.