

TIKHONOV REGULARIZATION FOR NONPARAMETRIC INSTRUMENTAL VARIABLE ESTIMATORS

P. Gagliardini* and O. Scaillet[†]

This version: September 2007 [‡]

(First version: May 2006)

*University of Lugano and Swiss Finance Institute.

[†]HEC Genève and Swiss Finance Institute. Corresponding author: Olivier Scaillet, HEC Genève UNI MAIL, Faculté des SES, Bd Carl Vogt 102, CH-1211 Genève 4, Switzerland. Tel.: ++ 41 22 379 88 16. Fax ++ 41 22 379 81 04. Email: Olivier.Scaillet@hec.unige.ch.

[‡]An earlier version of this paper has circulated under the title “Tikhonov regularization for functional minimum distance estimators”. Both authors received support by the Swiss National Science Foundation through the National Center of Competence in Research: Financial Valuation and Risk Management (NCCR FINRISK). We would like to thank Joel Horowitz for providing the dataset of the empirical section and many valuable suggestions as well as Manuel Arellano, Xiaohong Chen, Victor Chernozhukov, Jean-Pierre Florens, Oliver Linton, seminar participants at the University of Geneva, Catholic University of Louvain, University of Toulouse, Princeton University, Columbia University, ECARES, MIT/Harvard, CREST, Queen Mary’s College, Maastricht University, Carlos III University, ESRC 2006 Annual Conference in Bristol, SSES 2007 Annual Meeting in St. Gallen, the Workshop on Statistical Inference for Dependent Data in Hasselt, ESAM 2007 in Brisbane and ESEM 2007 in Budapest for helpful comments.

Tikhonov Regularization for Nonparametric Instrumental Variable Estimators

Abstract

We study a Tikhonov Regularized (TiR) estimator of a functional parameter identified by conditional moment restrictions in an additively separable model. The nonparametric instrumental variable estimator is based on a minimum distance principle with penalization by the norms of the parameter and its derivative. We derive the asymptotic Mean Integrated Square Error, the rate of convergence and the pointwise asymptotic normality under a regularization parameter depending on sample size. The optimal value of the regularization parameter is characterized. We illustrate our theoretical findings and the small sample properties with simulation results in a numerical example. We also discuss two data driven selection procedures of the regularization parameter via a spectral representation and a subsampling approximation of the MISE. Finally, we provide an empirical application to estimation of an Engel curve.

Keywords and phrases: Minimum Distance, Nonparametric Estimation, Ill-posed Inverse Problems, Tikhonov Regularization, Endogeneity, Instrumental Variable, Generalized Method of Moments, Subsampling, Engel curve.

JEL classification: C13, C14, C15, D12.

AMS 2000 classification: 62G08, 62G20.

1 Introduction

Kernel and sieve estimators provide inference tools for nonparametric regression in empirical economic analysis. Recently, several suggestions have been made to correct for endogeneity in such a context, mainly motivated by functional instrumental variable (IV) estimation of structural equations. Newey and Powell (NP, 2003) consider nonparametric estimation of a function, which is identified by conditional moment restrictions given a set of instruments. Ai and Chen (AC, 2003) opt for a similar approach to estimate semiparametric specifications. Darolles, Florens and Renault (DFR, 2003) and Hall and Horowitz (HH, 2005) concentrate on nonparametric IV estimation of a regression function. Horowitz (2005) shows the pointwise asymptotic normality for an asymptotically negligible bias. Horowitz and Lee (2007) extend HH to nonparametric IV quantile regression. Florens (2003) and Blundell and Powell (2003) give further background on endogenous nonparametric regressions.

There is a growing literature extending the above methods and considering empirical applications. Blundell, Chen and Kristensen (BCK, 2007) investigate application of index models to Engel curve estimation with endogenous total expenditure. As argued, e.g., in Blundell and Horowitz (2007), the knowledge of the shape of an Engel curve is a key ingredient of any consumer behaviour analysis. Further, Chen and Ludvigson (2004) consider asset pricing models with functional specifications of habit preferences; Chernozhukov, Imbens and Newey (2007) estimate nonseparable models for quantile regression analysis; Loubes and Vanhems (2004) discuss the estimation of the solution of a differential equation with endogenous variables for microeconomic applications. Other related works include Newey,

Powell, and Vella (1999), Chernozhukov and Hansen (2005), Florens, Johannes and Van Bellegem (2005), Horowitz (2006), Hu and Schennach (2006), and Hoderlein and Holzmann (2007).

The main theoretical difficulty in nonparametric estimation with endogeneity is overcoming ill-posedness (see Kress (1999), and Carrasco, Florens and Renault (2005) for overviews). It occurs since the mapping of the reduced form parameter (that is, the distribution of the data) into the structural parameter (the instrumental regression function) is not continuous. We need a regularization of the estimation to recover consistency. For instance, DFR and HH adopt a regularization technique resulting in a kind of ridge regression in a functional setting.

The aim of this paper is to introduce a new minimum distance estimator for a functional parameter identified by conditional moment restrictions in an additively separable model. We consider a penalized extremum estimator which minimizes $Q_T(\varphi) + \lambda_T G(\varphi)$, where $Q_T(\varphi)$ is a minimum distance criterion in the functional parameter φ , $G(\varphi)$ is a penalty function, and λ_T is a positive sequence converging to zero. The penalty function $G(\varphi)$ exploits the Sobolev norm of function φ , which involves the L^2 norms of both φ and its derivative $\nabla\varphi$. The basic idea is that the penalty term $\lambda_T G(\varphi)$ damps highly oscillating components of the estimator. These oscillations are otherwise unduly amplified by the minimum distance criterion $Q_T(\varphi)$ because of ill-posedness. Parameter λ_T tunes the regularization. We call our estimator a Tikhonov Regularized (TiR) estimator by reference to the pioneering papers of Tikhonov (1963a,b) where regularization is achieved via a penalty term incorporating the

function and its derivative (Groetsch (1984)).

The main contributions of our paper are the following. First, the introduction of the Sobolev penalty enhances the performance of the regularized estimator, both in a statistical and computational sense. We establish this point in an analytical example through asymptotic arguments, and in finite-sample Monte-Carlo examples with various designs. The nonparametric IV estimator admits a closed form and is numerically tractable. Second, the construction through the Sobolev penalty allows us to define the estimator directly over the function space and thus focus on the penalization parameter as the single regularization parameter. This sharply distinguishes our approach from the sieve approach that uses both the number of sieve terms and the penalization parameter to regularize. Third, we derive pointwise asymptotic normality, give the exact asymptotic expression for the mean integrated squared error (MISE), and provide data-driven procedures for the selection of the regularization parameter. These results are crucial for empirical work and, to our knowledge, they are new for this literature, as of writing this paper. In addition, we provide a consistency result with data-driven regularization parameter, and compute the optimal rates of convergence (including explicit multiplicative constants) in different examples.

Our paper is related to different contributions in the literature. To address ill-posedness NP and AC propose to introduce bounds on the norms of the functional parameter of interest and of its derivatives. This amounts to set compactness on the parameter space. This approach does not yield a closed-form estimator in an additive separable setting, because of the inequality constraint on the functional parameter. In their empirical application,

BCK implement a penalized estimator similar to ours. However they do not examine the theoretical properties of such an implementation. Moreover, in their sieve approach, both the number of sieve terms and the penalization coefficient are regularization parameters. In defining the estimator on a function space, we follow the route of Horowitz and Lee (2007) and the suggestion of NP, p. 1573. Finally, we stress that the regularization approach in DFR and HH can be viewed as a Tikhonov regularization, but with a penalty term involving the L^2 norm of the function only (without any derivative). By construction this penalization dispenses from a differentiability assumption of the function φ . To avoid confusion, we refer to DFR and HH estimators as regularized estimators with L^2 norm. In our Monte-Carlo experiments, we find that the use of the Sobolev penalty substantially enhances the finite-sample performance of the regularized estimator relative to the use of the L^2 penalty.

In Section 2 we discuss ill-posedness in nonparametric IV estimation, and in Section 3 we introduce the TiR estimator. Consistency is proved in Section 4. In Section 5, we derive the exact asymptotic MISE of the TiR estimator, discuss some examples of optimal rates of convergence, and compare with L^2 regularization in an analytic example. The MSE and pointwise asymptotic normality are given in Section 6. In Section 7, we discuss the numerical implementation and in Section 8 we present the Monte-Carlo results. We provide an empirical example in Section 9 where we estimate an Engel curve nonparametrically. Gagliardini and Scaillet (GS, 2006) give further simulation results and implementation details. Proofs are gathered in the Appendices. All omitted proofs of technical Lemmas are collected in a Technical Report, which is available online at our web pages.

2 Ill-posedness in nonparametric estimation

Let $\{(Y_t, X_t, Z_t) : t = 1, \dots, T\}$ be i.i.d. copies of vector (Y, X, Z) , and let the support of (Y, Z) be a subset of $\mathbb{R}^{d_Y} \times \mathbb{R}^{d_Z}$ while the support of X is $\mathcal{X} = [0, 1]$. Suppose that the parameter of interest is a scalar function φ_0 defined on \mathcal{X} , which satisfies:

$$E[g(Y, \varphi_0(X)) \mid Z] = 0, \quad (1)$$

where g is additively separable, namely $-g(y, \varphi(x)) = y - a(w)\varphi(x)$, and a is a known vector function of $w := (y, x)$. Parameter φ_0 belongs to a subset Θ of the Sobolev space $H^2[0, 1]$, i.e., the completion of the linear space $\{\varphi \in C^1[0, 1] \mid \nabla\varphi \in L^2[0, 1]\}$ with respect to the scalar product $\langle \varphi, \psi \rangle_H := \langle \varphi, \psi \rangle + \langle \nabla\varphi, \nabla\psi \rangle$, where $\langle \varphi, \psi \rangle = \int_{\mathcal{X}} \varphi(x)\psi(x)dx$. The Sobolev space $H^2[0, 1]$ is an Hilbert space w.r.t. the scalar product $\langle \varphi, \psi \rangle_H$, and the corresponding Sobolev norm is denoted by $\|\varphi\|_H = \langle \varphi, \varphi \rangle_H^{1/2}$. We use the L^2 norm $\|\varphi\| = \langle \varphi, \varphi \rangle^{1/2}$ as consistency norm. Further, we assume the following identification condition.

Assumption 1: (i) φ_0 is the unique function $\varphi \in \Theta$ that satisfies the conditional moment restriction (1); (ii) set Θ is bounded and closed w.r.t. norm $\|\cdot\|$.

We refer to NP, Theorems 2.2-2.4, for sufficient conditions ensuring Assumption 1 (i) in a linear setting, and Chernozhukov and Hansen (2005) for sufficient conditions in a nonlinear setting. Contrary to the standard parametric case, Assumption 1 (ii) does not imply compactness of Θ in infinite dimensional spaces. See Chen (2006) and Horowitz and Lee (2007) for similar noncompact settings.

The nonparametric minimum distance approach relies on φ_0 minimizing

$$Q_\infty(\varphi) = E [m(\varphi, Z)' \Omega_0(Z) m(\varphi, Z)], \quad \varphi \in \Theta, \quad (2)$$

where $m(\varphi, z) := E[g(Y, \varphi(X)) | Z = z]$, and $\Omega_0(z)$ is a positive definite matrix for any given z . The conditional moment function $m(\varphi, z)$ can be written as:

$$m(\varphi, z) = (A\varphi)(z) - r(z) = (A\Delta\varphi)(z), \quad (3)$$

where $\Delta\varphi := \varphi - \varphi_0$, linear operator A is defined by $(A\varphi)(z) = \int a(w)f(w|z)\varphi(x)dw$ and $r(z) = \int yf(w|z)dw$ where f is the conditional density of $W := (Y, X)$ given Z . Conditional moment restriction (1) identifies φ_0 (Assumption 1 (i)) if and only if operator A is injective. Further, we assume that A is a bounded operator from $L^2[0, 1]$ to $L^2_{\Omega_0}(F_Z)$, where $L^2_{\Omega_0}(F_Z)$ denotes the L^2 space of square integrable vector-valued functions of Z defined by scalar product $\langle \psi_1, \psi_2 \rangle_{L^2_{\Omega_0}(F_Z)} = E [\psi_1(Z)' \Omega_0(Z) \psi_2(Z)]$. The limit criterion (2) becomes

$$Q_\infty(\varphi) = \langle A\Delta\varphi, A\Delta\varphi \rangle_{L^2_{\Omega_0}(F_Z)} = \langle \Delta\varphi, A^*A\Delta\varphi \rangle_H, \quad (4)$$

where A^* denotes the adjoint operator of A w.r.t. the scalar products $\langle \cdot, \cdot \rangle_H$ and $\langle \cdot, \cdot \rangle_{L^2_{\Omega_0}(F_Z)}$.

Assumption 2: *The linear operator A from $L^2[0, 1]$ to $L^2_{\Omega_0}(F_Z)$ is compact.*

Then, A^*A is compact and self-adjoint in $H^2[0, 1]$. We denote by $\{\phi_j : j \in \mathbb{N}\}$ an orthonormal basis in $H^2[0, 1]$ of eigenfunctions of operator A^*A , and by $\nu_1 \geq \nu_2 \geq \dots > 0$ the corresponding eigenvalues (see Kress (1999), Section 15.3, for the spectral decomposition of compact, self-adjoint operators). By compactness of A^*A , the eigenvalues are such that

$\nu_j \rightarrow 0$, and $\nu_j / \|\phi_j\|^2 \rightarrow 0$, as $j \rightarrow \infty$. The limit criterion $Q_\infty(\varphi)$ can be minimized by a sequence in Θ such as

$$\varphi_n = \varphi_0 + \varepsilon \frac{\phi_n}{\|\phi_n\|}, \quad n \in \mathbb{N}, \quad (5)$$

for $\varepsilon > 0$, which does not converge to φ_0 in L^2 -norm $\|\cdot\|$. Indeed, we have $Q_\infty(\varphi_n) = \varepsilon^2 \langle \phi_n, A^* A \phi_n \rangle_H / \|\phi_n\|^2 = \varepsilon^2 \nu_n / \|\phi_n\|^2 \rightarrow 0$ as $n \rightarrow \infty$, but $\|\varphi_n - \varphi_0\| = \varepsilon$, $\forall n$. Since $\varepsilon > 0$ is arbitrary, the usual ‘‘identifiable uniqueness’’ assumption (e.g., White and Wooldridge (1991))

$$\inf_{\varphi \in \Theta: \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) > 0 = Q_\infty(\varphi_0), \quad \text{for } \varepsilon > 0, \quad (6)$$

is *not* satisfied. In other words, function φ_0 is not identified in Θ as an isolated minimum of Q_∞ . This is the identification problem of minimum distance estimation with functional parameter. Failure of Condition (6) despite validity of Assumption 1 comes from 0 being a limit point of the eigenvalues of operator A^*A . The minimum distance estimator which minimizes the empirical counterpart of criterion $Q_\infty(\varphi)$ over the set Θ is not consistent w.r.t. the L^2 -norm $\|\cdot\|$.

3 The Tikhonov Regularized (TiR) estimator

We address ill-posedness by Tikhonov regularization (Tikhonov (1963a,b); see also Kress (1999), Chapter 16). We consider a penalized criterion $Q_T(\varphi) + \lambda_T \|\varphi\|_H^2$, where $Q_T(\varphi)$ is an empirical counterpart of (2) defined by

$$Q_T(\varphi) = \frac{1}{T} \sum_{t=1}^T \hat{m}(\varphi, Z_t)' \hat{\Omega}(Z_t) \hat{m}(\varphi, Z_t), \quad (7)$$

and $\hat{\Omega}(z)$ is a sequence of positive definite matrices converging in probability to $\Omega_0(z)$, for any z . In (7) we estimate the conditional moment nonparametrically with

$$\hat{m}(\varphi, z) = \int a(w) \hat{f}(w|z) \varphi(x) dw - \int y \hat{f}(w|z) dw =: (\hat{A}\varphi)(z) - \hat{r}(z),$$

where $\hat{f}(w|z)$ denotes a kernel estimator of the density of W given $Z = z$ with kernel K and bandwidth h_T .

Definition 1: *The Tikhonov Regularized (TiR) minimum distance estimator is defined by*

$$\hat{\varphi} = \arg \inf_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T \|\varphi\|_H^2, \quad (8)$$

where $Q_T(\varphi)$ is as in (7), and λ_T is a stochastic sequence with $\lambda_T > 0$ and $\lambda_T \rightarrow 0$, *P*-a.s..

Term $\lambda_T \|\varphi\|_H^2$ in (8) penalizes highly oscillating components of the estimated function. These components would be otherwise unduly amplified, since ill-posedness yields a criterion $Q_T(\varphi)$ asymptotically flat along some directions. From (4), these directions are spanned by the eigenfunctions ϕ_n of operator A^*A to eigenvalues ν_n close to zero (cf. (5)). Since A is an integral operator, we expect that $\psi_n := \phi_n / \|\phi_n\|$ is a highly oscillating function and $\|\psi_n\|_H \rightarrow \infty$ as $n \rightarrow \infty$. These directions are penalized by penalty $\|\varphi\|_H^2$ in (8). The tuning parameter λ_T in Definition 1 controls for the amount of regularization, and how this depends on sample size T . Its rate of convergence to zero affects the one of $\hat{\varphi}$.

The TiR estimator admits a closed form expression. The objective function in (8) can be rewritten as (see Appendix 3.1)

$$Q_T(\varphi) + \lambda_T \|\varphi\|_H^2 = \langle \varphi, \hat{A}^* \hat{A} \varphi \rangle_H - 2 \langle \varphi, \hat{A}^* \hat{r} \rangle_H + \lambda_T \langle \varphi, \varphi \rangle_H, \quad \varphi \in H^2[0, 1], \quad (9)$$

up to a term independent of φ , where operator \hat{A}^* is given by

$$\hat{A}^* = \mathcal{D}^{-1} \tilde{\hat{A}}, \quad \left(\tilde{\hat{A}} \psi \right) (x) := \frac{1}{T} \sum_{t=1}^T \hat{f}(x|Z_t) \hat{\Omega}(Z_t) \psi(Z_t), \quad (10)$$

and \mathcal{D}^{-1} denotes the inverse of operator $\mathcal{D} : H_0^2[0, 1] \rightarrow L^2[0, 1]$ with $\mathcal{D} := 1 - \nabla^2$ and $H_0^2[0, 1] := \{\varphi \in H^2[0, 1] : \nabla\varphi(0) = \nabla\varphi(1) = 0\}$. Operators \hat{A}^* and $\tilde{\hat{A}}$ are the empirical counterparts of A^* and \tilde{A} , which are the adjoint operators of A w.r.t. the Sobolev and L^2 scalar products on $H^2[0, 1]$, respectively, and are linked by $A^* = \mathcal{D}^{-1} \tilde{A}$ (see Lemma A.1 in Appendix 3.1). Under the regularity conditions in Appendix 1, Criterion (9) admits a global minimum $\hat{\varphi}$ on $H^2[0, 1]$, which solves the first order condition

$$\left(\lambda_T + \hat{A}^* \hat{A} \right) \varphi = \hat{A}^* \hat{r}. \quad (11)$$

This is an integro-differential Fredholm equation of Type II (see e.g. Linton and Mammen (2005), (2006), Gagliardini and Gouriéroux (2007), and the survey by Carrasco, Florens and Renault (2005) for other examples). The transformation of the ill-posed problem (1) in the well-posed estimating equation (11) is induced by the penalty term involving the Sobolev norm. The TiR estimator is the explicit solution of Equation (11):

$$\hat{\varphi} = \left(\lambda_T + \hat{A}^* \hat{A} \right)^{-1} \hat{A}^* \hat{r}. \quad (12)$$

4 Consistency

First we show consistency of penalized minimum distance estimators with a (possibly) non-linear conditional moment function $\hat{m}(\varphi, z)$ and a general penalty function $G(\varphi)$:

$$\hat{\varphi} = \arg \inf_{\varphi \in \Theta} \frac{1}{T} \sum_{t=1}^T \hat{m}(\varphi, Z_t)' \hat{\Omega}(Z_t) \hat{m}(\varphi, Z_t) + \lambda_T G(\varphi). \quad (13)$$

Then we apply the results with $G(\varphi) = \|\varphi\|_H^2$ and additively separable moment function to prove the consistency of the TiR estimator.

The estimator (13) exists under weak conditions (see Appendix 2.1).

Theorem 1: *Let* (i) $\sup_{\varphi \in \Theta} \frac{1}{T} \sum_{t=1}^T \Delta \hat{m}(\varphi, Z_t)' \hat{\Omega}(Z_t) \Delta \hat{m}(\varphi, Z_t) = O_p(\eta_T^2)$ and $\sup_{\varphi \in \Theta} \left| \frac{1}{T} \sum_{t=1}^T m(\varphi, Z_t)' \hat{\Omega}(Z_t) m(\varphi, Z_t) - Q_\infty(\varphi) \right| = O_p(\eta_T^2)$, for a sequence $\eta_T \rightarrow 0$, where $\Delta \hat{m}(\varphi, z) := \hat{m}(\varphi, z) - m(\varphi, z)$; (ii) *There exists* $a \in (0, 1]$ *such that for any* $\varepsilon > 0$: *either* $\tau := \lim_{\lambda \rightarrow 0} \lambda^{-a} \left(\inf_{\varphi \in \Theta: \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) + \lambda G(\varphi) \right) > 0$, *or* $\tau = +\infty$ *if* $a = 1$.

Then, for any sequence (λ_T) *such that* $\lambda_T > 0$, $\lambda_T \rightarrow 0$, *P-a.s., and*

$$\frac{1}{\lambda_T^a} \eta_T^2 = o_p(1), \quad (14)$$

the estimator $\hat{\varphi}$ *defined in* (13) *is consistent, namely* $\|\hat{\varphi} - \varphi_0\| \xrightarrow{P} 0$.

Proof: See Appendix 2.

Theorem 1 extends standard results on consistency for minimum distance and extremum estimators (e.g., Newey and McFadden (1994), White and Wooldridge (1991), Corollary 2.6) to ill-posed settings in which condition (6) does not hold. The penalty function G overcomes

ill-posedness if $C_\varepsilon(\lambda) := \inf_{\varphi \in \Theta: \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) + \lambda G(\varphi) > 0$ for λ close to 0, for any $\varepsilon > 0$.

Since $C_\varepsilon(\lambda) \rightarrow 0$ as $\lambda \rightarrow 0$, Condition (ii) in Theorem 1 is a lower bound for the rate of convergence of C_ε . Coefficient a can be seen as a measure for the severity of ill-posedness.

In particular, $a = 1$ corresponds to the worst case of ill-posedness, while $a \rightarrow 0$ to the well-posed case. Equation (14) shows the interplay between a and the rate η_T^2 of uniform convergence in Condition (i) to guarantee consistency. The regularization parameter λ_T has to converge a.s. to zero at a rate smaller than $\eta_T^{2/a}$.

Proposition 2 particularizes Theorem 1 in the additively separable case with Sobolev penalty.

Proposition 2: *Suppose there exists $a \in (0, 1]$ such that it holds either $\bar{\tau} := \lim_{\lambda \rightarrow 0} \lambda^{-a} \left(\inf_{\substack{\psi \in H^2 \\ \|\psi\|=1}} \langle \psi, A^* A \psi \rangle_H + \lambda \|\psi\|_H^2 \right) > 0$, or $\bar{\tau} = +\infty$ if $a = 1$. Further, let*

λ_T be such that $\lambda_T > 0$, $\lambda_T \rightarrow 0$, P -a.s. and $\frac{(\log T)^2}{\lambda_T^a} \left(\frac{\log T}{T h_T^{d_Z+1}} + h_T^{2m} + \frac{1}{\sqrt{T}} \right) = o_p(1)$,

where $m \geq 2$ is the order of differentiability of the joint density of (Y, X, Z) . Then, under

Assumptions 1-2 and B in Appendix 1, the TiR estimator $\hat{\varphi}$ in (8) is consistent, namely

$$\|\hat{\varphi} - \varphi_0\| \xrightarrow{P} 0.$$

Proof: See Appendix 2.

Proposition 2 extends the results in DFR and in HH since it allows for a data-driven regularization parameter λ_T and a general weighting matrix. Moreover, in a setting with deterministic $\lambda_T \asymp T^{-\gamma}$, our result applies also for $\gamma \geq 1/2$ when $a < 1$ ($a_T \asymp b_T$ means that a_T/b_T converges to a constant $c > 0$ as $T \rightarrow \infty$). Indeed, if $\frac{\log T}{T h_T^{d_Z+1}} + h_T^{2m} = O\left(\frac{1}{\sqrt{T}}\right)$, we get consistency for any $\gamma < \frac{1}{2a}$. The exponent a is related to the spectrum of operator

A^*A . This is illustrated in the next example with trigonometric eigenfunctions as in the Monte-Carlo section of HH.

Example 1: (i) Suppose that the $\langle \cdot, \cdot \rangle_H$ -orthonormal eigenfunctions of A^*A are $\phi_j(x) = d_j \cos(j\pi x)$, $x \in [0, 1]$, $j = 0, 1, 2, \dots$, $d_0 = 1$, $d_j = \sqrt{2/(1 + \pi^2 j^2)}$ for $j \geq 1$, with hyperbolic decay of the eigenvalues $\nu_j = c j^{-\alpha}$, $j = 0, 1, 2, \dots$ for $\alpha > 2$ and a constant $c > 0$. Functions ϕ_j satisfy $\langle \phi_j, \phi_l \rangle = (1 + \pi^2 j^2)^{-1} \delta_{j,l}$. Thus, using $\langle \psi, A^*A\psi \rangle_H = \sum_j \nu_j \xi_j^2$, $\|\psi\|_H^2 = \sum_j \xi_j^2$ and $\|\psi\|^2 = \sum_j \xi_j^2 / (1 + \pi^2 j^2)$, where $\xi_j := \langle \psi, \phi_j \rangle_H$, we get:

$$\inf_{\substack{\psi \in H^2 \\ \|\psi\|=1}} \langle \psi, A^*A\psi \rangle_H + \lambda \|\psi\|_H^2 = \min_{j \geq 0} (\nu_j + \lambda) (1 + \pi^2 j^2) \asymp \lambda^{\frac{\alpha-2}{\alpha}}, \text{ as } \lambda \rightarrow 0,$$

and $a = \frac{\alpha-2}{\alpha} < 1$ (mild ill-posedness). (ii) Suppose instead the eigenvalues feature a geometric decay: $\nu_j = c e^{-\alpha j}$, $j = 0, 1, 2, \dots$ for $\alpha > 0$ and a constant $c > 0$. Then:

$$\inf_{\substack{\psi \in H^2 \\ \|\psi\|=1}} \langle \psi, A^*A\psi \rangle_H + \lambda \|\psi\|_H^2 = \min_{j \geq 0} (c e^{-\alpha j} + \lambda) (1 + \pi^2 j^2) \asymp \lambda (\log \lambda)^2, \text{ as } \lambda \rightarrow 0,$$

and $a = 1$ with $\bar{\tau} = +\infty$ (severe ill-posedness).

5 Mean Integrated Square Error

Next theoretical results are derived for a deterministic sequence (λ_T) . As in AC, Assumption 4.1, we assume the following choice of the weighting matrix.

Assumption 3: The asymptotic weighting matrix is $\Omega_0(z) = V[g(Y, \varphi_0(X)) | Z = z]^{-1}$.

In a semiparametric setting, AC show that this choice of the weighting matrix yields efficient estimators of the finite-dimensional component. Here, Assumption 3 simplifies the

formula for the asymptotic MISE of the TiR estimator computed in the next proposition.

Proposition 3: *Let $\{\phi_j : j \in \mathbb{N}\}$ be an orthonormal basis in $H^2[0, 1]$ of eigenfunctions of operator A^*A to eigenvalues ν_j , ordered such that $\nu_1 \geq \nu_2 \geq \dots > 0$. Under Assumptions 1-3, Assumptions B in Appendix 1, and the conditions with $\varepsilon > 0$*

$$\frac{(\log T)^2}{Th_T^{dz}} + h_T^m = o(\lambda_T b(\lambda_T)), \quad \frac{1}{Th_T^{1+2dz}} = O(1), \quad \frac{1}{Th_T} + h_T^{2m} = O(\lambda_T^{2+\varepsilon}), \quad (15)$$

the MISE of the TiR estimator $\hat{\varphi}$ with deterministic sequence (λ_T) is given by

$$E [\|\hat{\varphi} - \varphi_0\|^2] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \|\phi_j\|^2 + b(\lambda_T)^2 =: V_T(\lambda_T) + b(\lambda_T)^2 =: M_T(\lambda_T) \quad (16)$$

up to terms which are asymptotically negligible w.r.t. the RHS, where function $b(\lambda_T)$ is

$$b(\lambda_T) = \|(\lambda_T + A^*A)^{-1} A^*A\varphi_0 - \varphi_0\|, \quad (17)$$

and $m \geq 2$ is the order of differentiability of the joint density of (Y, X, Z) .

Proof: See Appendix 3.

The asymptotic expansion (16) of the MISE consists of two components.

(i) The bias function $b(\lambda_T)$ is the L^2 norm of $(\lambda_T + A^*A)^{-1} A^*A\varphi_0 - \varphi_0 =: \varphi_* - \varphi_0$. Function φ_* minimizes the penalized limit criterion $\langle \Delta\varphi, A^*A\Delta\varphi \rangle_H + \lambda_T \|\varphi\|_H^2$ w.r.t. $\varphi \in \Theta$. Thus, $b(\lambda_T)$ is the asymptotic bias arising from introducing the penalty $\lambda_T \|\varphi\|_H^2$ in the criterion. It corresponds to the so-called regularization bias in the theory of Tikhonov regularization (Kress (1999), Groetsch (1984)). Under general conditions on operator A^*A and true function φ_0 , the bias function $b(\lambda)$ is increasing w.r.t. λ and such that $b(\lambda) \rightarrow 0$ as $\lambda \rightarrow 0$.

(ii) The variance term $V_T(\lambda_T)$ involves a weighted sum of the regularized inverse eigenvalues $\nu_j/(\lambda_T + \nu_j)^2$ of operator A^*A , with weights $\|\phi_j\|^2$ (since $\nu_j/(\lambda_T + \nu_j)^2 \leq \nu_j$, the infinite sum converges under Assumption B.12 (i) in Appendix 1). To have an interpretation, note that the inverse of operator A^*A corresponds to the standard asymptotic variance matrix $(J_0'V_0^{-1}J_0)^{-1}$ of the efficient GMM in the parametric setting, where $J_0 = E[\partial g/\partial \theta']$ and $V_0 = V[g]$. In the ill-posed nonparametric setting, the inverse of operator A^*A is unbounded, and its eigenvalues $1/\nu_j \rightarrow \infty$ diverge. The penalty term $\lambda_T \|\varphi\|_H^2$ in the criterion defining the TiR estimator implies that inverse eigenvalues $1/\nu_j$ are “ridged” with $\nu_j/(\lambda_T + \nu_j)^2$.

The variance term $V_T(\lambda_T)$ is a decreasing function of λ_T . Under the assumption that the eigenfunctions ϕ_j and the eigenvalues ν_j of A^*A satisfy $\sum_{j=1}^{\infty} \nu_j^{-1} \|\phi_j\|^2 = \infty$, the series $n_T := \sum_{j=1}^{\infty} \|\phi_j\|^2 [\nu_j/(\lambda_T + \nu_j)^2]$ diverges as $\lambda_T \rightarrow 0$. When $n_T \rightarrow \infty$ such that $n_T/T \rightarrow 0$, the variance term $V_T(\lambda_T)$ converges to zero. The previous assumption rules out the parametric rate $1/T$ for the variance. This smaller rate of convergence typical in nonparametric estimation is not coming from localization as for kernel estimation, but from the ill-posedness of the problem, which implies $\nu_j \rightarrow 0$.

The asymptotic expansion of the MISE given in Proposition 3 does not involve the bandwidth h_T , as long as Conditions (15) are satisfied. The variance term is asymptotically independent of h_T since the asymptotic expansion of $\hat{\varphi} - \varphi_0$ involves the kernel density estimator integrated w.r.t. (Y, X, Z) (see the first term of Equation (34) in Appendix 3, and the proof of Lemma A.3). The integral averages the localization effect of the bandwidth h_T . On the contrary, the kernel estimation in $\hat{m}(\varphi, z)$ does impact on bias. However, the

assumption $h_T^m = o(\lambda_T b(\lambda_T))$, which follows from (15), implies that the estimation bias is asymptotically negligible compared to the regularization bias (see Lemma A.4 in Appendix 3). The other restrictions on the bandwidth h_T in (15) are used to control higher order terms in the MISE (see Lemma A.5).

Finally, it is also possible to derive an exact asymptotic expansion of the MISE for the estimator $\tilde{\varphi}$ regularized by the L^2 norm. A similar formula has been derived by Carrasco and Florens (2005) for the density deconvolution problem. This characterization is new in the nonparametric IV regression setting:

$$E [\|\tilde{\varphi} - \varphi_0\|^2] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\tilde{\nu}_j}{(\lambda_T + \tilde{\nu}_j)^2} + \tilde{b}(\lambda_T)^2 =: \tilde{M}_T(\lambda_T), \quad (18)$$

where $\tilde{\nu}_j$ are the eigenvalues of operator $\tilde{A}A$, \tilde{A} denotes the adjoint of A w.r.t. the scalar products $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle_{L^2_{\Omega_0}(F_Z)}$, and $\tilde{b}(\lambda_T) = \left\| \left(\lambda_T + \tilde{A}A \right)^{-1} \tilde{A}A\varphi_0 - \varphi_0 \right\|$. DFR (see also Johannes and Vanhems (2006)) present an extensive discussion of the bias term under L^2 regularization and the relationship with the smoothness properties of φ_0 , the so-called source condition.

Let us now come back to the MISE $M_T(\lambda_T)$ of Proposition 3 and discuss the optimal choice of the regularization parameter λ_T . Since the bias term is increasing in the regularization parameter, whereas the variance term is decreasing, we face a traditional bias-variance trade-off. The optimal sequence of deterministic regularization parameters is given by $\lambda_T^* = \arg \min_{\lambda > 0} M_T(\lambda)$, and the corresponding optimal MISE by $M_T^* := M_T(\lambda_T^*)$. Their rates of convergence depend on the decay behavior of the eigenvalues ν_j and of the norms $\|\phi_j\|$ of the eigenfunctions, as well as on the bias function $b(\lambda)$ close to $\lambda = 0$.

Example 1 (Cont.): The eigenfunctions satisfy $\|\phi_j\|^2 = \frac{1}{1 + \pi^2 j^2} \asymp j^{-2}$. Using $\langle \varphi_0, \phi_j \rangle_H = (1 + \pi^2 j^2) \langle \varphi_0, \phi_j \rangle$, and $\langle \phi_j, \phi_l \rangle = 0$ for $j \neq l$, we have $b(\lambda)^2 = \lambda^2 \sum_{j=1}^{\infty} \frac{\langle \varphi_0, \psi_j \rangle^2}{(\lambda + \nu_j)^2}$, where $\psi_j := \phi_j / \|\phi_j\|$. Assume that function φ_0 satisfies:

$$\sum_{j=1}^{\infty} \frac{\langle \varphi_0, \psi_j \rangle^2}{(\lambda + \nu_j)^2} \asymp \lambda^{2(\delta-1)}, \text{ as } \lambda \rightarrow 0, \quad (19)$$

for some $0 < \delta \leq 1$. The faster the convergence $\langle \varphi_0, \psi_j \rangle \rightarrow 0$ relative to $\nu_j \rightarrow 0$, the larger parameter δ . This smoothness condition is strongly related to the so-called source condition (e.g., DFR). Then the bias is such that $b(\lambda) \asymp \lambda^\delta$.

The next proposition derives the optimal rates of convergence for the geometric spectrum case when the eigenfunction norms and the bias function behave as in Example 1.

Proposition 4: Suppose $e^{\alpha j} \nu_j \rightarrow C_1$, $j^\beta \|\phi_j\|^2 \rightarrow C_2$ and $\lambda^{-\delta} b(\lambda) \rightarrow C_3$ for some constants $\alpha, \beta, C_1, C_2, C_3 > 0$ and $0 < \delta < 1$. Then, under the Assumptions of Proposition 3:

(i) The MISE is $M_T(\lambda) = c_1 \frac{1 + c(\lambda)}{T \lambda [\log(1/\lambda)]^\beta} + c_2 \lambda^{2\delta}$, up to terms which are negligible when $\lambda \rightarrow 0$ and $T \rightarrow \infty$, where $c_1 = \left(\frac{1}{\alpha}\right)^{1-\beta} C_2$, $c_2 = C_3^2$, function $c(\lambda)$ is given in Appendix 4 and is such that $1 + c(\lambda)$ is bounded and bounded away from zero as $\lambda \rightarrow 0$.

(ii) The optimal sequence of regularization parameters is

$$\log \lambda_T^* = \log c^* - \frac{1}{1 + 2\delta} \log T, \quad T \in \mathbb{N}, \quad (20)$$

up to terms which are negligible when $T \rightarrow \infty$, where $\log c^* = \frac{1}{1 + 2\delta} \log \left(\frac{c_1}{2c_2\delta} \right)$.

(iii) The optimal MISE is $M_T^* = c_T^* T^{-\frac{2\delta}{1+2\delta}} (\log T)^{-\frac{2\delta\beta}{1+2\delta}}$, where

$$c_T^* = \frac{c_2}{(1 + 2\delta)^{-\frac{2\delta\beta}{1+2\delta}}} \left(\frac{c_1}{2c_2\delta} (1 + c_T) \right)^{\frac{2\delta}{1+2\delta}} \left(2\delta + \frac{1 + \bar{c}_T}{1 + c_T} \right) + o(1) \quad (21)$$

with $\frac{1}{1+e^{\alpha/2}} - \frac{1}{1+e^{-\alpha}} \leq \bar{c}_T \leq \frac{\alpha}{4}$ and $\frac{1}{(1+\frac{1}{2}e^\alpha)^2} - \frac{1}{(1+\frac{1}{2}e^{-\alpha})^2} \leq c_T \leq \frac{8\alpha}{27}$.

Proof: See Appendix 4.

From (20) the log of the optimal regularization parameter is asymptotically linear in the log sample size with given explicit constants. The slope coefficient $\gamma := 1/(1+2\delta)$ depends on the convexity parameter δ of the bias function close to $\lambda = 0$. The third condition in (15) forces γ to be smaller than $1/2$, which is a condition also used in HH and DFR. In the Monte-Carlo experiments in GS, the asymptotic expansion in Proposition 3 provides a good approximation of the MISE in finite samples also when $\gamma \geq 1/2$. The optimal MISE converges to zero as a power of T and of $\log T$. The negative exponent of the dominant term T is $2\delta/(1+2\delta)$ (c_T^* in (21) is bounded and bounded away from 0). This rate of convergence is smaller than $2/3$ and is increasing w.r.t. δ . Given the geometric spectrum, the convergence of the MISE as a power of T is a consequence of the smoothness assumption on function φ_0 which yields $b(\lambda) \asymp \lambda^\delta$. Proposition 3 can also be applied under a different assumption on $b(\lambda)$, and may lead to a convergence of the MISE as a power of $\log T$, as in BCK, or Hoderlein and Holzmann (2007).

With hyperbolic spectrum, a result similar to Proposition 4 can be obtained. Hereafter we compare analytically in Example 1 the optimal MISE M_T^* of the TiR estimator, and of the L^2 regularized estimator, denoted \tilde{M}_T^* . This comparison is made possible because we have derived the exact asymptotic expansions $M_T(\lambda)$ and $\tilde{M}_T(\lambda)$.

Example 1 (Cont.): Let $\nu_j \asymp j^{-\alpha}$, $\alpha > 3$, and let function φ_0 be such that $\langle \varphi_0, \psi_j \rangle^2 \asymp j^{-2\eta}$, $1/2 < \eta < \alpha - 3/2$. Then, $\tilde{\nu}_j \asymp j^{-\tilde{\alpha}}$, $\tilde{\alpha} = \alpha - 2$. It is possible to show that Condition

(19) is satisfied with $2\delta = (2\eta - 1)/\alpha$, $M_T^* = cT^{-\varrho}$ and $\tilde{M}_T^* = \tilde{c}T^{-\varrho}$, up to asymptotically negligible terms, where $\varrho = \frac{2\delta}{1 + 2\delta - 1/\alpha} = \frac{2\eta - 1}{2\eta + \tilde{\alpha}}$ and the constants c, \tilde{c} are such that

$$\frac{c}{\tilde{c}} = \left(\frac{\alpha - 2}{\alpha}\right)^2 \left(\frac{\sin\left(\frac{\pi}{\alpha - 2}\right)}{\sin\left(\frac{\pi}{\alpha}\right)}\right)^\varrho \left(\frac{\alpha - 2\eta + 1}{\alpha - 2\eta - 1} \frac{\sin\left(\pi \frac{2\eta - 1}{\alpha - 2}\right)}{\sin\left(\pi \frac{2\eta - 1}{\alpha}\right)}\right)^{1-\varrho}.$$

The two estimators feature the same rate of convergence ϱ , which is the optimal rate given in HH, Theorem 4.1, and in BCK, Theorem 3 (with their $r = (2\eta - 1)/2$ and $s = \tilde{\alpha}/2$). The rate of convergence in Proposition 4 is recovered when $\alpha \rightarrow \infty$. The ratio c/\tilde{c} yields the relative efficiency of the TiR estimator compared to the L^2 regularized estimator. For any $\eta > 1/2$, the ratio c/\tilde{c} is a monotonically increasing function of α with range $(0, 1)$. In particular, $c/\tilde{c} < 1$. Moreover, there exist models for which the TiR estimator is arbitrarily more efficient ($c/\tilde{c} \rightarrow 0$) compared to the L^2 regularized estimator.

The common rate of convergence for Sobolev and L^2 penalization in Example 1 is a consequence of operators A^*A and $\tilde{A}A$ admitting a common basis of eigenfunctions. The analytic comparison of the rates of convergence and the discussion of the relative efficiency of the estimators in the general case is still an open question. In the Monte-Carlo analysis in Section 8 we find a smaller MISE for the TiR estimator in finite samples.

6 Mean Square Error and asymptotic normality

The asymptotic MSE at a point $x \in \mathcal{X}$ can be computed along the same lines as the asymptotic MISE, and we only state the result without proof. It is immediately seen that

the integral of the MSE below over the support $\mathcal{X} = [0, 1]$ gives the MISE in (16).

Proposition 5: *Under the Assumptions of Proposition 3, the MSE of the TiR estimator $\hat{\varphi}$ with deterministic sequence (λ_T) is given by*

$$E [(\hat{\varphi}(x) - \varphi_0(x))^2] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j^2(x) + \mathcal{B}_T(x)^2 =: \frac{1}{T} \sigma_T^2(x) + \mathcal{B}_T(x)^2, \quad (22)$$

up to terms which are asymptotically negligible w.r.t. the RHS, where the bias term is

$$\mathcal{B}_T(x) = (\lambda_T + A^*A)^{-1} A^*A\varphi_0(x) - \varphi_0(x). \quad (23)$$

The asymptotic variance $\sigma_T^2(x)/T$ of $\hat{\varphi}(x)$ depends on $x \in \mathcal{X}$ through the eigenfunctions ϕ_j , whereas the asymptotic bias of $\hat{\varphi}(x)$ as a function of $x \in \mathcal{X}$ is given by $\mathcal{B}_T(x)$. Not only the scale but also the rate of convergence of the MSE may differ across the points of the support \mathcal{X} . Hence a locally optimal sequence minimizing the MSE at a given point $x \in \mathcal{X}$ may differ from the globally optimal one minimizing the MISE in terms of rate of convergence (and not only in terms of a scale constant as in usual kernel regression). These features result from our ill-posed setting. Finally, under a regularization with L^2 norm:

$$E [(\tilde{\varphi}(x) - \varphi_0(x))^2] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\tilde{\nu}_j}{(\lambda_T + \tilde{\nu}_j)^2} \tilde{\phi}_j^2(x) + \tilde{\mathcal{B}}_T(x)^2, \quad (24)$$

where $\tilde{\mathcal{B}}_T(x) = (\lambda_T + \tilde{A}A)^{-1} \tilde{A}A\varphi_0(x) - \varphi_0(x)$ and $\tilde{\phi}_j$ denotes an orthonormal basis in $L^2[0, 1]$ of eigenvectors of $\tilde{A}A$ to eigenvalues $\tilde{\nu}_j$.

In the next proposition we establish pointwise asymptotic normality of the TiR estimator.

Proposition 6: *Suppose Assumptions 1-3 and B hold, $\frac{(\log T)^2}{Th_T^{dz}} + h_T^m = O\left(\lambda_T^{1+\varepsilon/2} b(\lambda_T)\right)$,*

$$\frac{1}{Th_T^{1+2dz}} = O(1), \quad \frac{1}{Th_T} + h_T^{2m} = O(\lambda_T^{2+\varepsilon}),$$

$$\frac{M_T(\lambda_T)}{\sigma_T^2(x)/T} = o(\lambda_T^{-\varepsilon}), \quad (25)$$

for a $\varepsilon > 0$. Further, suppose that for a $\bar{\varepsilon} > 0$ we have

$$\frac{1}{T^{1/3}} \frac{\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j^2(x) \|g_j\|_3^2 j^{1+\bar{\varepsilon}}}{\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j^2(x)} = o(1), \quad (26)$$

where $\|g_j\|_3 := E[g_j(Y, X, Z)^3]^{1/3}$, $g_j(y, x, z) := (A\phi_j)(z)' \Omega_0(z)g(y, \varphi_0(x)) / \sqrt{\nu_j}$. Then the TiR estimator is asymptotically normal: $\sqrt{T/\sigma_T^2(x)}(\hat{\varphi}(x) - \varphi_0(x) - \mathcal{B}_T(x)) \xrightarrow{d} N(0, 1)$.

Proof: See Appendix 5.

Condition (25) requires that the local rate of convergence at $x \in \mathcal{X}$ of the variance is not too large compared to the global rate of convergence of the MISE. Condition (26) is used to apply a Lyapunov CLT. When $\|g_j\|_3^2 j^{1+\bar{\varepsilon}}$ diverges with j , Condition (26) is an upper bound on the rate of convergence of λ_T . Under an assumption of geometric spectrum for the eigenvalues ν_j , and hyperbolic decay for the eigenfunction values $\phi_j^2(x)$ and $\|g_j\|_3$, Lemma A.6 in Appendix 4 implies that (26) is satisfied whenever $\lambda_T \geq cT^{-\gamma}$ for some $c, \gamma > 0$. Finally, when λ_T is such that the regularization bias is asymptotically negligible, a natural candidate for a $N(0, 1)$ pivotal statistic is $\sqrt{T/\hat{\sigma}_T^2(x)}(\hat{\varphi}(x) - \varphi_0(x))$, where $\hat{\sigma}_T^2(x)$ is obtained by replacing ν_j and $\phi_j^2(x)$ with consistent estimators (see Darolles, Florens, Gouriéroux (2004) and Carrasco, Florens, Renault (2005) for the estimation of the spectrum of a compact operator). Since $\sigma_T(x)$ depends on T and diverges, the usual argument using Slutsky Theorem does not apply. Instead the condition $[\hat{\sigma}_T(x) - \sigma_T(x)]/\hat{\sigma}_T(x) \xrightarrow{p} 0$ is

required. For the sake of space, we do not discuss here regularity assumptions for this condition to hold, nor the issue of bias reduction (see Horowitz (2005) for the discussion of a bootstrap approach).

7 Numerical implementation

To compute numerically the estimator we solve Equation (11) on the subspace spanned by a finite-dimensional basis of functions $\{P_j : j = 1, \dots, k\}$ in $H^2[0, 1]$ and use the numerical approximation

$$\varphi \simeq \sum_{j=1}^k \theta_j P_j =: \theta' P, \quad \theta \in \mathbb{R}^k. \quad (27)$$

The $k \times k$ matrix corresponding to operator $\hat{A}^* \hat{A}$ on this subspace is given by $\langle P_i, \hat{A}^* \hat{A} P_j \rangle_H = \langle P_i, \tilde{\hat{A}} \hat{A} P_j \rangle = \frac{1}{T} \sum_{t=1}^T \left(\hat{A} P_i \right) (Z_t) \hat{\Omega} (Z_t) \left(\hat{A} P_j \right) (Z_t) = \frac{1}{T} \left(\hat{P}' \hat{P} \right)_{i,j}$, $i, j = 1, \dots, k$, where \hat{P} is the $T \times k$ matrix with rows $\hat{P} (Z_t)' = \hat{\Omega} (Z_t)^{1/2} \int a(w) P(x)' \hat{f} (w|Z_t) dw$, $t = 1, \dots, T$ (we have used (10), and we have set $d_Y = 1$ to simplify the exposition). Matrix \hat{P} is the matrix of the weighted “fitted values” in the regression of $P(X)$ on Z at the sample points. Then, by projection on the k -dimensional linear subspace of $H^2[0, 1]$ spanned by $\{P_j : j = 1, \dots, k\}$, Equation (11) reduces to a matrix equation $\left(\lambda_T D + \frac{1}{T} \hat{P}' \hat{P} \right) \theta = \frac{1}{T} \hat{P}' \hat{R}$, where $\hat{R} = \left(\hat{\Omega} (Z_1)^{1/2} \hat{r} (Z_1), \dots, \hat{\Omega} (Z_T)^{1/2} \hat{r} (Z_T) \right)'$, and D is the $k \times k$ matrix of Sobolev scalar products $D_{i,j} = \langle P_i, P_j \rangle_H$, $i, j = 1, \dots, k$. The solution is given by $\hat{\theta} = \left(\lambda_T D + \frac{1}{T} \hat{P}' \hat{P} \right)^{-1} \frac{1}{T} \hat{P}' \hat{R}$, which yields the approximation of the TiR estimator $\hat{\varphi} \simeq \hat{\theta}' P$. It only asks for inverting a $k \times k$ matrix, which is found to be of small dimension in our economic application ($k = 6$).

The matrix D is by construction positive definite, since its entries are scalar products of linearly independent basis functions. Hence, $\lambda_T D + \frac{1}{T} \widehat{P}' \widehat{P}$ is non-singular, P -a.s..

Estimator $\widehat{\theta}$ is a 2SLS estimator with optimal instruments and a ridge correction term. It is also obtained if we replace (27) in Criterion (9) and minimize w.r.t. θ . This route is followed by NP, AC, and BCK, who use sieve estimators and let $k = k_T \rightarrow \infty$ with T to regularize the estimation. In our setting, the introduction of a series of basis functions as in (27) is simply a method to compute numerically the original TiR estimator (12). The latter is a well-defined estimator on the function space $H^2[0, 1]$, and we do not need to tie down the numerical approximation to sample size. In practice we can use an iterative procedure to verify whether k is large enough to yield a small numerical error. We can start with an initial number of polynomials, and then increment until the absolute or relative variations in the optimized objective function become smaller than a given tolerance level. This mimicks stopping criteria implemented in numerical optimization routines. A visual check of the behavior of the optimized objective function w.r.t. k is another possibility (see the empirical section). Alternatively, we could simply take an a priori large k for which matrix inversion in computing $\widehat{\theta}$ is still numerically feasible.

Finally, a similar approach can be followed under an L^2 regularization by replacing matrix D with matrix B of L^2 scalar products $B_{i,j} = \langle P_i, P_j \rangle$, $i, j = 1, \dots, k$. DFR follow a different approach to compute exactly the estimator (see DFR, Appendix C). Their method requires solving a $T \times T$ linear system of equations. For univariate X and Z , HH implement an estimator which uses the same basis for estimating conditional expectation $m(\varphi, z)$ and for

approximating function $\varphi(x)$.

8 A Monte-Carlo study

Following NP, the errors U and V and the instrument Z are jointly normally distributed, with zero means, unit variances and correlation coefficient $\rho = 0.5$ between U and V . We take $X^* = Z + V$ and build $X = \Phi(X^*)$. The function Φ denotes the cdf of a standard Gaussian variable, and is assumed to be known. To generate Y , we examine the design $Y = \sin(\pi X) + U$. Then $E[Y - \varphi_0(X) | Z] = 0$ and the functional parameter is $\varphi_0(x) = \sin(\pi x)$, $x \in [0, 1]$. This function resembles the shape of the Engel curve in our empirical application. GS also analyze the case of a monotone increasing Engel curve. Qualitatively they find similar results. This reinforces the evidence presented in this section.

As $\mathcal{X} = [0, 1]$, we use a series approximation based on standardized shifted Chebyshev polynomials of the first kind (Abramowitz and Stegun (1970)). We take orders 0 to 5 ($k = 6$) in (27), and matrices D and B are explicitly computed with a symbolic calculus package (see GS). The kernel estimator $\hat{m}(\varphi, z)$ of the conditional moment is approximated through $\theta' \hat{P}(z) - \hat{r}(z)$, where $\hat{P}(z)$ and $\hat{r}(z)$ are standard kernel regressions with Gaussian kernel. This estimator is asymptotically equivalent to the one described above. It avoids a bivariate numerical integration and the choice of two additional bandwidths. The bandwidth is selected via the standard rule of thumb (Silverman (1986)). This choice is motivated by ease of implementation. Moderate deviations from this simple rule do not seem to affect estimation results significantly. The weighting function $\Omega_0(z)$ is taken equal to unity, satisfying

Assumption 3, and assumed to be known.

The sample size is fixed at $T = 1000$. In Figures 1 (TiR estimator) and 2 (L^2 regularized estimator), the left panel plots the MISE on a grid of lambda, the central panel the Integrated Squared Bias (ISB), and the right panel the mean estimated functions and the true function on the unit interval. Mean estimated functions correspond to averages over 1000 repetitions obtained either from regularized estimates with a lambda achieving the lowest MISE or from OLS estimates (standard sieve estimators with six polynomials). Several remarks can be made. First, the endogeneity bias of the OLS estimator is large. Second, the MISE under a Sobolev penalization is more convex and much smaller. Hence the Sobolev norm should be strongly favored in order to recover the shape of the true functions in our design. Third, examining the ISB for λ close to 0 shows that the estimation part of the bias of the TiR estimator is negligible w.r.t. the regularization part. For $T = 100$ and $T = 400$ (see GS) as well as number k of polynomials up to 16 the above conclusions on both estimators remain qualitatively unaffected. This suggests that as soon as the order of the polynomials is sufficiently large to numerically approximate the underlying function, there is no gain by linking it with sample size (cf. Section 7).

In Figure 3 we display ν_j and $\|\phi_j\|^2$ computed by Monte-Carlo integration. The eigenvalues ν_j feature a geometric decay $\nu_j \asymp e^{-\alpha j}$, whereas the decay of $\|\phi_j\|^2 \asymp j^{-\beta}$ is of an hyperbolic type. This is conform to the assumptions in Proposition 4. A linear fit gives 2.254, 2.911 for α, β . In Figure 4 we check whether $\log \lambda_T^* = \log c - \gamma \log T$ holds (cf. Proposition 4 (ii)). We observe a linear relationship between the logarithm of the regularization param-

eter minimizing the average MISE and the logarithm of sample size ranging from $T = 50$ to $T = 1000$. An OLS estimation delivers .012, .428 for c, γ . Inverting $\gamma = 1/(1 + 2\delta)$ yields .668 for δ .

Finally let us discuss two data driven selection procedures of the regularization parameter λ_T . The first one aims at estimating directly the asymptotic spectral representation (16). A similar approach has been successfully applied in Carrasco and Florens (2005) for density deconvolution. Unreported results based on Monte-Carlo integration show that the asymptotic MISE, ISB, and variance are close to the ones exhibited in Figure 1. The asymptotic optimal lambda is equal to .0009, which is of the same magnitude as .0007 in Figure 1.

Algorithm (spectral approach)

- (i) Perform the spectral decomposition of the matrix $D^{-1}\widehat{P}'\widehat{P}/T$ to get eigenvalues $\hat{\nu}_j$ and eigenvectors \hat{w}_j , normalized to $\hat{w}_j'D\hat{w}_j = 1, j = 1, \dots, k$.
- (ii) Get a first-step estimate $\bar{\theta}$ using a pilot regularization parameter $\bar{\lambda}$.
- (iii) Estimate the MISE:

$$\bar{M}(\lambda) = \frac{1}{T} \sum_{j=1}^k \frac{\hat{\nu}_j}{(\lambda + \hat{\nu}_j)^2} \hat{w}_j' B \hat{w}_j + \bar{\theta}' \left[\frac{1}{T} \widehat{P}' \widehat{P} \left(\lambda D + \frac{1}{T} \widehat{P}' \widehat{P} \right)^{-1} - I \right] B \left[\left(\lambda D + \frac{1}{T} \widehat{P}' \widehat{P} \right)^{-1} \frac{1}{T} \widehat{P}' \widehat{P} - I \right] \bar{\theta},$$
 and minimize it w.r.t. λ to get the optimal regularization parameter $\hat{\lambda}$.
- (iv) Compute the second-step TiR estimator with $\hat{\theta}$ using regularization parameter $\hat{\lambda}$.

A second-step estimated MISE viewed as a function of sample size T and regularization

parameter λ can then be estimated with $\hat{\theta}$ instead of $\bar{\theta}$. Besides, if we assume the decay behavior in Proposition 4, the decay factors α and β can be estimated via minus the slopes of the linear fit on the pairs $(\log \hat{\nu}_j, j)$ and on the pairs $(\log \hat{w}_j' B \hat{w}_j, \log j)$, $j = 1, \dots, k$. After getting lambdas minimizing the second-step estimated MISE on a grid of sample sizes we can estimate γ by regressing the logarithm of optimal lambda on the logarithm of sample size.

We use $\bar{\lambda} \in \{.0005, .0001\}$ as pilot regularization parameter. The average (quartiles) of the selected lambda over 1000 simulations is equal to .0009 (.0007, .0008, .0009) when $\bar{\lambda} = .0005$, and .0008 (.0004, .0006, .0009) when $\bar{\lambda} = .0001$. The selection procedure tends to slightly overpenalize on average, but impact on the MISE of the two-step TiR estimator is low. Indeed if we use the optimal data driven regularization parameter at each simulation, the average ISE is equal to .0144 when $\bar{\lambda} = .0005$, resp., .0175 when $\bar{\lambda} = .0001$. These are of the same magnitude as the best MISE .0121 in Figure 1.

We also get average estimated values for the decay factors α and β close to the asymptotic ones. For α the average (quartiles) is equal to 2.2502 (2.1456, 2.2641, 2.3628), and for β it is equal to 2.9222 (2.8790, 2.9176, 2.9619). To estimate γ we use $T \in \{500, 550, \dots, 1000\}$, and get an average (quartiles) of .5597 (.4918, .5333, .5962), when $\bar{\lambda} = .0005$, and .5764 (.4946, .5416, .6203), when $\bar{\lambda} = .0001$.

The second data driven selection procedure builds on the suggestion of Goh (2004) based on a subsampling procedure. Proposition 6 shows that a limit distribution exists, a prerequisite for applying subsampling. Recognizing that asymptotically $\lambda_T^* = cT^{-\gamma}$, we pro-

pose to choose c and γ which minimize the following estimator of the MISE: $\hat{M}(c, \gamma) = \frac{1}{I} \frac{1}{J} \sum_{i,j} \int_0^1 (\hat{\varphi}_{i,j}(x; c, \gamma) - \bar{\varphi}(x))^2 dx$. Here $\hat{\varphi}_{i,j}(x; c, \gamma)$ denotes the estimator based on the j th subsample of size m_i ($m_i \ll T$) with regularization parameter $\lambda_{m_i} = cm_i^{-\gamma}$, and $\bar{\varphi}(x)$ denotes the estimator based on the original sample of size T with a pilot regularization parameter $\bar{\lambda}$ chosen sufficiently small to eliminate the bias. Preliminary results of Monte-Carlo simulations in GS show that the second method is able to pick pairs (c, γ) close to the optimal ones.

9 An empirical example

This section presents an empirical example with the data in Horowitz (2006). We estimate an Engel curve based on the moment condition $E[Y - \varphi_0(X) | Z] = 0$, with $X = \Phi(X^*)$. Variable Y denotes the food expenditure share, X^* denotes the standardized logarithm of total expenditures, and Z denotes the standardized logarithm of annual income from wages and salaries. We have 785 household-level observations from the 1996 US Consumer Expenditure Survey. The estimation procedure is as in the Monte-Carlo study and uses data-driven regularization parameters. We keep six polynomials. Here the value of the optimized objective function stabilizes after $k = 6$ (see Figure 5), and estimation results remain virtually unchanged for larger k . We have estimated the weighting matrix since $\Omega_0(z) = V[Y - \varphi_0(X) | Z = z]^{-1}$ is doubtfully constant in this application. We use a pilot regularization parameter $\bar{\lambda} = .0001$ to get a first step estimator of φ_0 . The estimator $\hat{s}^2(Z_t)$ of the conditional variance $s^2(Z_t) = \Omega_0(Z_t)^{-1}$ is of a kernel regression type. Subsampling relies

on 1000 subsamples ($J = 1000$) for each subsample size $m_i \in \{50, 53, \dots, 200\}$ ($I = 51$), and the extended grid $\{0.005, .01, .05, .1, .25, .5, 1, 2, \dots, 6\} \times \{.3, .35, \dots, .9\}$ for (c, γ) . Estimation with the first, resp. second, data driven selection procedure takes less than 2 seconds, resp. 1 day.

We obtain a selected value of $\hat{\lambda} = .01113$ with the spectral approach, and regression estimates $\hat{\alpha} = 2.05176$, $\hat{\beta} = 3.31044$, $\hat{\gamma} = .90889$, $\hat{\delta} = .05012$. We obtain a value of $\hat{\lambda} = .01240$ from the selected pair (5,.9) for (c, γ) with the subsampling procedure. Figure 6 plots the estimated functions $\hat{\varphi}(x)$ for $x \in [0, 1]$, and $\hat{\varphi}(\Phi(x^*))$ for $x^* \in \mathbb{R}$, using $\hat{\lambda} = .01113$. The plotted shape corroborates the findings of Horowitz (2006), who rejects a linear curve but not a quadratic curve at the 5% significance level to explain $\log Y$. The specification test of Gagliardini and Scaillet (2007) does not reject the null hypothesis of the correct specification of the moment restriction used in estimating the Engel curve at the 5% significance level. Banks, Blundell and Lewbel (1997) consider demand systems that accommodate such empirical Engel curves.

References

- Abramowitz, M. and I. Stegun (1970): *Handbook of Mathematical Functions*, Dover Publications, New York.
- Adams, R. (1975): *Sobolev Spaces*, Academic Press, Boston.
- Ai, C. and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", *Econometrica*, 71, 1795-1843.
- Andrews, D. (1994): "Empirical Process Methods in Econometrics", in the *Handbook of Econometrics*, Vol. 4, Engle, R. and D. McFadden (eds), North Holland, 2247-2294.
- Banks, J., Blundell, R. and A. Lewbel (1997): "Quadratic Engel Curves and Consumer Demand", *Review of Economics and Statistics*, 79, 527-539.
- Blundell, R., Chen, X. and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape Invariant Engel Curves", forthcoming in *Econometrica*.
- Blundell, R. and J. Horowitz (2007): "A Non-parametric Test of Exogeneity", forthcoming in *Review of Economic Studies*.
- Blundell, R. and J. Powell (2003): "Endogeneity in Semiparametric and Nonparametric Regression Models", in *Advances in Economics and Econometrics: Theory and Applications*, Dewatripont, M., Hansen, L. and S. Turnovsky (eds), pp. 312-357, Cambridge University Press.
- Carrasco, M. and J.-P. Florens (2005): "Spectral Method for Deconvolving a Density", Working Paper.
- Carrasco, M., Florens, J.-P. and E. Renault (2005): "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization", forthcoming in the *Handbook of Econometrics*.
- Chen, X. (2006): "Large Sample Sieve Estimation of Semi-Nonparametric Models", forthcoming in the *Handbook of Econometrics*.
- Chen, X. and S. Ludvigson (2004): "Land of Addicts? An Empirical Investigation of Habit-Based Asset Pricing Models", Working Paper.
- Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects", *Econometrica*, 73, 245-271.
- Chernozhukov, V., Imbens, G. and W. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models", *Journal of Econometrics*, 139, 4-14.

Darolles, S., Florens, J.-P. and C. Gouriéroux (2004): "Kernel Based Nonlinear Canonical Analysis and Time Reversibility", *Journal of Econometrics*, 119, 323-353.

Darolles, S., Florens, J.-P. and E. Renault (2003): "Nonparametric Instrumental Regression", Working Paper.

Florens, J.-P. (2003): "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables", in *Advances in Economics and Econometrics: Theory and Applications*, Dewatripont, M., Hansen, L. and S. Turnovsky (eds), pp. 284-311, Cambridge University Press.

Florens, J.-P., Johannes, J. and S. Van Bellegem (2005): "Instrumental Regression in Partially Linear Models", Working Paper.

Gagliardini, P. and C. Gouriéroux (2007): "An Efficient Nonparametric Estimator for Models with Nonlinear Dependence", *Journal of Econometrics*, 137, 189-229.

Gagliardini, P. and O. Scaillet (2006): "Tikhonov Regularization for Functional Minimum Distance Estimators", Working Paper.

Gagliardini, P. and O. Scaillet (2007): "A Specification Test for Nonparametric Instrumental Variable Regression", Working Paper.

Goh, S. (2004): "Bandwidth Selection for Semiparametric Estimators Using the m -out-of- n Bootstrap", Working Paper.

Groetsch, C. W. (1984): *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman Advanced Publishing Program, Boston.

Hall, P. and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables", *Annals of Statistics*, 33, 2904-2929.

Hansen, B. (2007): "Uniform Convergence Rates for Kernel Estimation with Dependent Data", forthcoming in *Econometric Theory*.

Hoderlein, S. and H. Holzmann (2007): "Demand Analysis as an Ill-Posed Inverse Problem with Semiparametric Specification", Working Paper.

Horowitz, J. (2005): "Asymptotic Normality of a Nonparametric Instrumental Variables Estimator", forthcoming in *International Economic Review*.

Horowitz, J. (2006): "Testing a Parametric Model Against a Nonparametric Alternative with Identification Through Instrumental Variables", *Econometrica*, 74, 521-538.

Horowitz, J. and S. Lee (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model", forthcoming in *Econometrica*.

- Hu, Y. and S. Schennach (2004): "Identification and Estimation of Nonclassical Non-linear Errors-in-Variables Models with Continuous Distributions using Instruments", Working Paper.
- Johannes, J. and A. Vanhems (2006): "Regularity Conditions for Inverse Problems in Econometrics", Working Paper.
- Kress, R. (1999): *Linear Integral Equations*, Springer, New York.
- Linton, O. and E. Mammen (2005): "Estimating Semiparametric ARCH(∞) Models by Kernel Smoothing Methods", *Econometrica*, 73, 771-836.
- Linton, O. and E. Mammen (2006): "Nonparametric Transformation to White Noise", forthcoming in *Journal of Econometrics*.
- Loubes, J.-M. and A. Vanhems (2004): "Estimation of the Solution of a Differential Equation with Endogenous Effect", Working Paper.
- Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing", in *Handbook of Econometrics*, Vol. 4, Engle, R. and D. McFadden (eds).
- Newey, W. and J. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models", *Econometrica*, 71, 1565-1578.
- Newey, W., Powell, J. and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models", *Econometrica*, 67, 565-604.
- Reed, M. and B. Simon (1980): *Functional Analysis*, Academic Press, San Diego.
- Silverman, B. (1986): *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- Tikhonov, A. N. (1963a): "On the Solution of Incorrectly Formulated Problems and the Regularization Method", *Soviet Math. Doklady*, 4, 1035-1038 (English Translation).
- Tikhonov, A. N. (1963b): "Regularization of Incorrectly Posed Problems", *Soviet Math. Doklady*, 4, 1624-1627 (English Translation).
- Wahba, G. (1977): "Practical Approximate Solutions to Linear Operator Equations When the Data are Noisy", *SIAM J. Numer. Anal.*, 14, 651-667.
- White, H. and J. Wooldridge (1991): "Some Results on Sieve Estimation with Dependent Observations", in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Proceedings of the Fifth International Symposium in Economic Theory and Econometrics, Cambridge University Press.

Appendix 1: List of regularity conditions

B.1: $\{U_t = (Y_t, X_t, Z_t) : t = 1, \dots, T\}$ is an i.i.d. sample from a distribution admitting a density f with convex support $\mathcal{S} = \mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \subset \mathbb{R}^d$, $\mathcal{X} = [0, 1]$, $d = d_Y + 1 + d_Z$.

B.2: The density f of (Y, X, Z) is in class $C^m(\mathbb{R}^d)$, with $m \geq 2$, and $\nabla^\alpha f$ is uniformly continuous, for any $\alpha \in \mathbb{N}^d$ with $|\alpha| := \sum_{i=1}^d \alpha_i = m$.

B.3: The kernel K on \mathbb{R}^d is such that (i) $\int K(u)du = 1$ and K is bounded; (ii) K has compact support; (iii) K is Lipschitz; (iv) $\int u^\alpha K(u)du = 0$ for any $\alpha \in \mathbb{N}^d$ with $|\alpha| < m$.

B.4: There exists $h > 0$ such that function $q(u) := \sup_{v \in B_h(u)} |\nabla f(v)|$, $u \in \mathcal{S}$, is integrable w.r.t. Lebesgue measure on \mathcal{S} , where $B_h(u)$ denotes the ball in \mathbb{R}^d of radius h around u .

B.5: There exists $h > 0$ such that function $q_\alpha(u) := \sup_{v \in B_h(u)} |\nabla f^\alpha(v)|$, $u \in \mathcal{S}$, satisfies $\int_{\mathcal{S}} \frac{q_\alpha(u)^2}{f(u)} du < \infty$, for any $\alpha \in \mathbb{N}^d$ with $|\alpha| = m$.

B.6: The density f of X given Z is such that (i) $\sup_{x \in \mathcal{X}, z \in \mathcal{Z}} f(x|z) < \infty$;
(ii) $\sup_{x \in \mathcal{X}, z \in \mathcal{Z}} |\nabla_x f(x|z)| < \infty$.

B.7: Function $r(z) := E[Y | Z = z]$ is in class $C^m(\mathcal{Z})$ with uniformly continuous derivatives of order m . Moreover, $E[|Y|^4] < \infty$ and $\sup_{z \in \mathcal{Z}} E[|Y|^s | Z = z] f(z) < \infty$ for $s > 2$.

B.8: Function $\varphi_0 \in H^2[0, 1]$ is such that $E[|\varphi_0(X)|^4] < \infty$.

B.9: Function a is such that $\sup_{w \in \mathcal{Y} \times \mathcal{X}} |a(w)| < \infty$.

B.10: The weighting matrix $\Omega_0(z) = V[g(Y, \varphi_0(X)) | Z = z]^{-1}$ is such that (i) Ω_0 is bounded on \mathcal{Z} ; (ii) $E[|\Sigma_0(Z)|^4] < \infty$, where $\Sigma_0(z) = \Omega_0(z)/f(z)$.

B.11: Estimator $\hat{\Omega}$ of Ω_0 is such that there exists a sequence of sets $\mathcal{Z}_T \subset \mathcal{Z}$, $T \in \mathbb{N}$, with

(i) $\sup_{z \in \mathcal{Z}_T} |z| = O(T^b)$, for $b > 0$, $P[Z \in \mathcal{Z}_T^c] = O(T^{-\bar{b}})$ for any $\bar{b} > 0$, $\sup_{z \in \mathcal{Z}_T} |\Delta \hat{\Omega}(z)| = o_p(1)$

with $\Delta \hat{\Omega} := \hat{\Omega} - \Omega_0$, $\hat{\Omega}(z) = 0$ if $z \in \mathcal{Z}_T^c$ or $\hat{f}(z) < (\log T)^{-1}$, and $\inf_{z \in \mathcal{Z}_T} f(z) \geq 2(\log T)^{-1}$;

(ii) $\frac{1}{\sqrt{T}} \sum_{t=1}^T f(x|Z_t) f(\xi|Z_t) \Delta \hat{\Omega}(Z_t) I(Z_t \in \mathcal{Z}_T) = O_p(1)$, uniformly in $x, \xi \in [0, 1]$;

(iii) $E \left[E \left[\Delta \hat{\Omega}(Z_t)^{2N} | Z_t \right] I(Z_t \in \mathcal{Z}_T) \right] = O \left(\left(\frac{1}{Th_T^{d_Z}} + h_T^{2m} \right)^N \right)$ for any $N \in \mathbb{N}$;

(iv) $T^{-2N} \sum_{t_1, \dots, t_{2N}=1}^T E \left[\left| E \left[\Delta \hat{\Omega}(Z_{t_1}) \cdots \Delta \hat{\Omega}(Z_{t_{2N}}) | Z_{t_1}, \dots, Z_{t_{2N}} \right] \right| I(Z_{t_1} \in \mathcal{Z}_T, \dots, Z_{t_{2N}} \in \mathcal{Z}_T) \right]$
 $= O \left(\left(\frac{1}{Th_T} + h_T^{2m} \right)^N \right)$ for any $N \in \mathbb{N}$.

B.12: The $\langle \cdot, \cdot \rangle_H$ -orthonormal basis of eigenfunctions $\{\phi_j : j \in \mathbb{N}\}$ of operator A^*A satisfies

(i) $\sum_{j=1}^{\infty} \|\phi_j\| < \infty$; (ii) $\sum_{j,l=1, j \neq l}^{\infty} \frac{\langle \phi_j, \phi_l \rangle^2}{\|\phi_j\|^2 \|\phi_l\|^2} < \infty$.

B.13: The eigenfunctions ϕ_j and the eigenvalues ν_j of A^*A are such that

$\sup_{j \in \mathbb{N}} E[\omega(U) |g_j(U)|^2] < \infty$ and $\sup_{j \in \mathbb{N}} E[|g_j(U)|^{\bar{s}}] < \infty$, for $\bar{s} > 2$, where

$g_j(u) := (A\phi_j)(z)' \Omega_0(z) g(y, \varphi_0(x)) / \sqrt{\nu_j}$ and $\omega := q/f$ for q as in Assumption B.4.

B.14: The functions g_j are in class $C^m(\mathbb{R}^d)$ such that (i) $\sup_{j \in \mathbb{N}} E[|\nabla g_j(U)|^2] < \infty$

and $\sup_{j \in \mathbb{N}} E[\omega(U) |\nabla g_j(U)|^2] < \infty$; (ii) $\sup_{j \in \mathbb{N}} E[|\nabla^\alpha g_j(U)|^2] < \infty$ and

$\sup_{j \in \mathbb{N}} E[|\nabla^\alpha g_j(U - v) - \nabla^\alpha g_j(U)|^2] \rightarrow 0$ as $|v| \rightarrow 0$, for any $\alpha \in \mathbb{N}^d$ with $|\alpha| = m$.

In Assumption B.1, the compact support of X is used for technical reasons. Mapping in $[0, 1]$ can be achieved by simple linear or nonlinear monotone transformations. Assuming

univariate X simplifies the exposition. Assumptions B.2 and B.3 are classical conditions in kernel density estimation concerning smoothness of the density and of the kernel. In particular, when $m > 2$, K is a higher order kernel. Moreover, we assume a compact support for kernel K to simplify the set of regularity conditions. Assumptions B.4 and B.5 impose integrability conditions on suitable measures of local variation of density f . These assumptions are used in the proof of Lemmas A.3 and A.4 to bound higher order terms in the asymptotic expansion of the MISE. Assumptions B.6 and B.7 are smoothness conditions on the distributions of X given Z , of Y and of Y given Z , respectively. They are used to apply a weak convergence result for empirical processes, and to show uniform convergence of kernel estimators $\hat{f}(x|z)$ and $\hat{r}(z)$, in the proof of Proposition 2 (consistency). Assumption B.8 on function φ_0 is used to bound the expectation of terms involving powers of $\varphi_0(X)$ in Lemma B.8 (Technical Report). Using Assumption B.9 of bounded function $a(w)$ in the moment condition, the proofs given in Appendix 2-5 for the case $a(w) = 1$ (i.e., linear instrumental variable regression) extend easily to the separable case. Assumption B.10 imposes boundedness and integrability conditions on the weighting function $\Omega_0(z)$. In particular, Assumption B.10 (i) together with Assumptions B.6 (i) and B.9 imply that operator A is compact. Assumption B.11 concerns the estimator $\hat{\Omega}$ of the weighting function. Specifically, Assumption B.11 (i) introduces a trimming for small values of $f(z)$ and $\hat{f}(z)$ in the denominator, while Assumptions B.11 (ii)-(iv) are bounds for suitable sample and population moments of estimator $\hat{\Omega}$. Assumption B.11 covers the trivial case of known weighting function Ω_0 , and the choice $\Omega_0(z) = f(z)$ used by HH. The last three Assumptions B.12-B.14 concern the

spectrum of operator A^*A . Assumption B.12 (i) is used to simplify the proof of Lemmas A.7 and A.9. It is met e.g. when $\|\phi_j\|^2 \asymp j^{-\beta}$, with $\beta > 2$ (see Example 1). Assumption B.12 (ii) requires that the eigenfunctions of operator A^*A , which are orthogonal w.r.t. $\langle \cdot, \cdot \rangle_H$, are sufficiently orthogonal w.r.t. $\langle \cdot, \cdot \rangle$. It is satisfied in Example 1. Under this Assumption, the asymptotic expansion of the MISE in Proposition 3 involves a single sum, and not a double sum, over the spectrum. Assumptions B.13 and B.14 ask for the existence of a uniform bound for moments of derivatives of functions $g_j(u) = \frac{1}{\sqrt{\nu_j}} (A\phi_j)'(z) \Omega_0(z) g(y, \varphi_0(x))$, $j \in \mathbb{N}$, both under density f and under the density defined by function q in Assumption B.4. Functions g_j satisfy $E[g_j(U)^2] = 1$. Thus, Assumptions B.13 and B.14 are met whenever moment function $g(y, \varphi_0(x))$, instrument $\frac{1}{\sqrt{\nu_j}} (A\phi_j)'(z)$, the elements of the weighting matrix $\Omega_0(z)$, and their derivatives, do not exhibit too heavy tails. These assumptions are used to bound higher order terms in the asymptotic expansion of the MISE in Lemma A.3, and in the proof of Lemma A.7.

Appendix 2: Consistency of the TiR estimator

A.2.1 Existence of penalized minimum distance estimators

Since $Q_T(\varphi) := \frac{1}{T} \sum_{t=1}^T \hat{m}(\varphi, Z_t) \hat{\Omega}(Z_t) \hat{m}(\varphi, Z_t)$ is positive, a function $\hat{\varphi} \in \Theta$ is solution of optimization problem in (13) if and only if

$$\hat{\varphi} = \arg \inf_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T G(\varphi), \quad \text{s.t.} \quad \lambda_T G(\varphi) \leq L_T, \quad (28)$$

where $L_T := Q_T(\varphi_0) + \lambda_T G(\varphi_0)$. The solution $\hat{\varphi}$ in (28) exists P -a.s. if

(i) mappings $\varphi \rightarrow G(\varphi)$ and $\varphi \rightarrow Q_T(\varphi)$ are lower semicontinuous on Θ , P -a.s., for any T ,

w.r.t. the L^2 norm $\|\cdot\|$;

(ii) set $\{\varphi \in \Theta : G(\varphi) \leq \bar{L}\}$ is compact w.r.t. the L^2 norm $\|\cdot\|$, for any constant $0 < \bar{L} < \infty$.

We do not address the technical issue of measurability of $\hat{\varphi}$.

A.2.2 Consistency of penalized minimum distance estimators

Proof of Theorem 1: We have for any $\varepsilon > 0$:

$$P[\|\hat{\varphi} - \varphi_0\| \geq \varepsilon] \leq P\left[\inf_{\varphi \in \Theta: \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) + \lambda_T \|\varphi\|_H^2 \leq Q_\infty(\hat{\varphi}) + \lambda_T \|\hat{\varphi}\|_H^2\right]. \quad (29)$$

Let us first derive a probability bound for $Q_\infty(\hat{\varphi}) + \lambda_T \|\hat{\varphi}\|_H^2$. We denote $\langle \varphi, \psi \rangle_T := \frac{1}{T} \sum_{t=1}^T \varphi(Z_t) \hat{\Omega}(Z_t) \psi(Z_t)$ and $\|\varphi\|_T := \langle \varphi, \varphi \rangle_T^{1/2}$. From $\|\hat{m}(\hat{\varphi}, \cdot)\|_T^2 + \lambda_T \|\hat{\varphi}\|_H^2 \leq \|\hat{m}(\varphi_0, \cdot)\|_T^2 + \lambda_T \|\varphi_0\|_H^2$, we get: $\|m(\hat{\varphi}, \cdot)\|_T^2 + \lambda_T \|\hat{\varphi}\|_H^2 + 2\langle m(\hat{\varphi}, \cdot), \Delta \hat{m}(\hat{\varphi}, \cdot) \rangle_T + \|\Delta \hat{m}(\hat{\varphi}, \cdot)\|_T^2 - \|\Delta \hat{m}(\varphi_0, \cdot)\|_T^2 - \lambda_T \|\varphi_0\|_H^2 \leq 0$. Using the Cauchy-Schwarz inequality, we deduce that $\delta_T := \sqrt{\|m(\hat{\varphi}, \cdot)\|_T^2 + \lambda_T \|\hat{\varphi}\|_H^2}$ satisfies $\delta_T^2 - 2d_{1,T}\delta_T + d_{2,T} \leq 0$, where $d_{1,T} := \|\Delta \hat{m}(\hat{\varphi}, \cdot)\|_T$ and $d_{2,T} := \|\Delta \hat{m}(\hat{\varphi}, \cdot)\|_T^2 - \|\Delta \hat{m}(\varphi_0, \cdot)\|_T^2 - \lambda_T \|\varphi_0\|_H^2$. This implies that $\delta_T \leq d_{1,T} + \sqrt{d_{1,T}^2 - 4d_{2,T}}$. From Condition (i) we have $d_{1,T}^2 = O_p(\eta_T^2)$, $d_{1,T}^2 - 4d_{2,T} = O_p(\eta_T^2 + \lambda_T)$ and thus $\|m(\hat{\varphi}, \cdot)\|_T^2 + \lambda_T \|\hat{\varphi}\|_H^2 = O_p(\eta_T^2 + \lambda_T)$. From Condition (i) we get:

$$Q_\infty(\hat{\varphi}) + \lambda_T \|\hat{\varphi}\|_H^2 = O_p(\eta_T^2 + \lambda_T). \quad (30)$$

For $a < 1$, from (29), (30), and Conditions (i)-(ii) we deduce $P[\|\hat{\varphi} - \varphi_0\| \geq \varepsilon] \leq P[\xi_T \geq C(\varepsilon)] \rightarrow 0$, where $\xi_T := \lambda_T^{-a} (Q_\infty(\hat{\varphi}) + \lambda_T \|\hat{\varphi}\|_H^2) = o_p(1)$ from (14) and $C(\varepsilon) > 0$ is a constant.

The argument for $a = 1$ is similar, and the conclusion follows.

A.2.3 Penalization with Sobolev norm

To conclude on existence of the TiR estimator, let us check the assumptions in A.2.1 for the special case $G(\varphi) = \|\varphi\|_H^2$, and additively separable moment function.

(i) The mapping $\varphi \rightarrow \|\varphi\|_H^2$ is lower semicontinuous on $H^2[0,1]$ w.r.t. the norm $\|\cdot\|$ (see Reed and Simon (1980), p. 358). Continuity of $Q_T(\varphi)$, P -a.s., follows from the mapping $\varphi \rightarrow \hat{m}(\varphi, z)$ being continuous for almost any $z \in \mathcal{Z}$, P -a.s.. The latter holds since for any $\varphi_1, \varphi_2 \in \Theta$, $|\hat{m}(\varphi_1, z) - \hat{m}(\varphi_2, z)| \leq \int \left(\int |\hat{f}(w|z)| dy \right) |\varphi_1(x) - \varphi_2(x)| dx \leq \bar{C}_T \|\varphi_1 - \varphi_2\|$, where $\bar{C}_T < \infty$ for almost any $z \in \mathcal{Z}$, P -a.s., by using the mean-value theorem, the Cauchy-Schwarz inequality and Assumption B.3.

(ii) The set $\{\varphi \in \Theta : \|\varphi\|_H^2 \leq \bar{L}\}$ is compact w.r.t. the norm $\|\cdot\|$, for any $0 < \bar{L} < \infty$ (Rellich-Kondrachov Theorem; see Adams (1975)).

Proof of Proposition 2: We have to check Conditions (i) and (ii) in Theorem 1.

Let us first consider Condition (i). We have $\Delta\hat{m}(\varphi, \cdot) = (\hat{A} - A)\varphi - (\hat{r} - r)$. Then, with the notation introduced in the proof of Theorem 1, $\|\Delta\hat{m}(\varphi, \cdot)\|_T \leq \|(\hat{A} - A)\varphi\|_T + \|\hat{r} - r\|_T$. Using $\left|(\hat{A} - A)\varphi(z)\right| \leq \left(\int [\hat{f}(x|z) - f(x|z)]^2 dx\right)^{1/2} \|\varphi\|$, we have

$$\|(\hat{A} - A)\varphi\|_T^2 \leq \left(\frac{1}{T} \sum_{t=1}^T \hat{\Omega}(Z_t) \int [\hat{f}(x|Z_t) - f(x|Z_t)]^2 dx\right) \sup_{\varphi \in \Theta} \|\varphi\|^2.$$

A similar bound holds for $\|\hat{r} - r\|_T$. From Assumptions B.1-B.3, B.7, B.10 (i) and B.11 (i), and from Theorems 6 and 8 in Hansen (2007):

$$\begin{aligned} \sup_{x \in [0,1], z \in \mathcal{Z}_T} [\hat{f}(x|z) - f(x|z)]^2 &= O_p \left((\log T)^2 \left(\frac{\log T}{Th_T^{1+d_Z}} + h_T^{2m} \right) \right), \\ \sup_{z \in \mathcal{Z}_T} [\hat{r}(z) - r(z)]^2 &= O_p \left((\log T)^2 \left(\frac{\log T}{Th_T^{d_Z}} + h_T^{2m} \right) \right). \end{aligned}$$

Thus, we get $\sup_{\varphi \in \Theta} \|\Delta \hat{m}(\varphi, \cdot)\|_T^2 = O_p \left((\log T)^2 \left(\frac{\log T}{T h_T^{1+d_Z}} + h_T^{2m} \right) \right)$. Moreover, we have:

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T m(\varphi, Z_t) \hat{\Omega}(Z_t) m(\varphi, Z_t) - Q_\infty(\varphi) \\ &= \frac{1}{T} \sum_{t=1}^T A \Delta \varphi(Z_t) \hat{\Omega}(Z_t) A \Delta \varphi(Z_t) - E[A \Delta \varphi(Z) \Omega_0(Z) A \Delta \varphi(Z)] \\ &= \left\langle \Delta \varphi, \left(\widehat{\tilde{A}A} - \tilde{A}A \right) \Delta \varphi \right\rangle, \end{aligned}$$

where $\widehat{\tilde{A}A} \varphi(x) := \int \left(\frac{1}{T} \sum_{t=1}^T f(x|Z_t) f(\xi|Z_t) \hat{\Omega}(Z_t) \right) \varphi(\xi) d\xi$ and $\tilde{A}A \varphi(x) = \int \int f(x|z) f(\xi|z) \Omega_0(z) f(z) \varphi(\xi) dz d\xi$. Then we get:

$$\sup_{\varphi \in \Theta} \left| \frac{1}{T} \sum_{t=1}^T m(\varphi, Z_t) \hat{\Omega}(Z_t) m(\varphi, Z_t) - Q_\infty(\varphi) \right| \leq \left\| \widehat{\tilde{A}A} - \tilde{A}A \right\| \sup_{\varphi \in \Theta} \|\Delta \varphi\|^2,$$

$$\text{and } \left\| \widehat{\tilde{A}A} - \tilde{A}A \right\|^2 \leq \int \int \left[\frac{1}{T} \sum_{t=1}^T f(x|Z_t) f(\xi|Z_t) \hat{\Omega}(Z_t) - \int f(x|z) f(\xi|z) \Omega_0(z) f(z) dz \right]^2 dx d\xi.$$

Let us bound the integrand. By empirical process methods, from Assumptions B.1, B.3,

$$\text{B.6 (i)-(ii) and B.10 (i), } \frac{1}{\sqrt{T}} \sum_{t=1}^T \left[f(x|Z_t) f(\xi|Z_t) \Omega_0(Z_t) - \int f(x|z) f(\xi|z) \Omega_0(z) f(z) dz \right] =$$

$O_p(1)$, uniformly in $x, \xi \in [0, 1]$ (use Theorems 1 and 2 in Andrews (1994); Pollard's entropy

condition is satisfied with envelope $\bar{M}(z) = C |\Omega_0(z)|$, for a suitable constant C). Further,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T f(x|Z_t) f(\xi|Z_t) \Omega_0(Z_t) I(Z_t \in \mathcal{Z}_T^c) = O_p \left([TP(Z_t \in \mathcal{Z}_T^c)]^{1/2} \right) = O_p(1), \text{ from Assump-}$$

tion B.11 (i). Then we get $\sup_{\varphi \in \Theta} \left| \frac{1}{T} \sum_{t=1}^T m(\varphi, Z_t) \hat{\Omega}(Z_t) m(\varphi, Z_t) - Q_\infty(\varphi) \right| = O_p \left(\frac{1}{\sqrt{T}} \right)$ from

Assumption B.11 (ii). Thus Condition (i) is satisfied with $\eta_T^2 := (\log T)^2 \left(\frac{\log T}{T h_T^{1+d_Z}} + h_T^{2m} \right) + \frac{1}{\sqrt{T}}$.

Consider now Condition (ii) and the quadratic function

$$\mathcal{L}(\varphi) := Q_\infty(\varphi) + \lambda \|\varphi\|_H^2 = \langle \Delta \varphi, A^* A \Delta \varphi \rangle_H + \lambda \|\varphi\|_H^2.$$

It takes the minimum at $\varphi_* := (\lambda + A^*A)^{-1} A^*A\varphi_0$. Thus $\mathcal{L}(\varphi) = \langle \varphi - \varphi_*, (\lambda + A^*A)(\varphi - \varphi_*) \rangle_H + \mathcal{L}(\varphi_*)$. Using $\varphi_* - \varphi_0 = -\lambda(\lambda + A^*A)^{-1}\varphi_0$ and $A^*A(\varphi_* - \varphi_0) = -\lambda\varphi_*$, we get $\mathcal{L}(\varphi_*) = \lambda \langle \varphi_*, \varphi_0 \rangle_H$ and $\mathcal{L}(\varphi) = \langle \varphi - \varphi_*, (\lambda + A^*A)(\varphi - \varphi_*) \rangle_H + \lambda \langle \varphi_*, \varphi_0 \rangle_H$. Then, $\inf_{\varphi: \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) + \lambda \|\varphi\|_H^2 = \lambda \langle \varphi_*, \varphi_0 \rangle_H + \inf_{\psi: \|\psi + \varphi_* - \varphi_0\| \geq \varepsilon} \langle \psi, (\lambda + A^*A)\psi \rangle_H$. Since $\varphi_* \rightarrow \varphi_0$ as $\lambda \rightarrow 0$:

$$\begin{aligned} \inf_{\varphi \in \Theta: \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) + \lambda \|\varphi\|_H^2 &\geq \inf_{\psi: \|\psi\| \geq \varepsilon/2} \langle \psi, (\lambda + A^*A)\psi \rangle_H + O(\lambda) \\ &\geq \frac{\varepsilon^4}{4} \left(\inf_{\psi: \|\psi\|=1} \langle \psi, A^*A\psi \rangle_H + \lambda \|\psi\|_H^2 \right) + O(\lambda), \end{aligned}$$

and Condition (ii) is satisfied.

Appendix 3: MISE of the TiR estimator

A.3.1 First-order condition

The estimated moment function is $\hat{m}(\varphi, z) = \int \varphi(x) \hat{f}(w|z) dw - \int y \hat{f}(w|z) dw =: (\hat{A}\varphi)(z) - \hat{r}(z)$. The objective function of the TiR estimator becomes

$$Q_T(\varphi) + \lambda_T \|\varphi\|_H^2 = \frac{1}{T} \sum_{t=1}^T \hat{\Omega}(Z_t) \left[(\hat{A}\varphi)(Z_t) - \hat{r}(Z_t) \right]^2 + \lambda_T \langle \varphi, \varphi \rangle_H, \quad (31)$$

and can be written as a quadratic form in $\varphi \in H^2[0, 1]$. To achieve this, we have to give explicitly the adjoint operator A^* and introduce its empirical counterpart \hat{A}^* .

Lemma A.1: (i) For $\phi \in L^2[0, 1]$, there exists a unique twice differentiable function u on $[0, 1]$ such that

$$\mathcal{D}u := u - \nabla^2 u = \phi, \quad \nabla u(0) = \nabla u(1) = 0.$$

It is denoted by $u = \mathcal{D}^{-1}\phi$. (ii) The mapping $\mathcal{D}^{-1} : L^2[0, 1] \rightarrow H^2[0, 1]$ is continuous.

(iii) For any $\varphi \in H^2[0, 1]$ and $\phi \in L^2[0, 1]$ we have $\langle \varphi, \phi \rangle = \langle \varphi, \mathcal{D}^{-1}\phi \rangle_H$.

(iv) We have $A^* = \mathcal{D}^{-1}\tilde{A}$, where $\tilde{A}\psi(x) = \int f(x, z)\Omega_0(z)\psi(z)dz$ for $\psi \in L^2_{\Omega_0}(F_Z)$.

Lemma A.2: Under Assumptions B, the following properties hold P -a.s. :

(i) The linear operator $\hat{A}^* := \mathcal{D}^{-1}\tilde{A}$ from $L^2_{\Omega_0}(F_Z)$ into $H^2[0, 1]$ is such that

$$\left\langle \varphi, \hat{A}^*\psi \right\rangle_H = \frac{1}{T} \sum_{t=1}^T \left(\hat{A}\varphi \right) (Z_t) \hat{\Omega}(Z_t)\psi(Z_t), \text{ for any } \psi \in L^2_{\Omega_0}(F_Z) \text{ and any } \varphi \in H^2[0, 1];$$

(ii) Operator $\hat{A}^*\hat{A} : H^2[0, 1] \rightarrow H^2[0, 1]$ is compact.

Then, from Lemma A.2 (i), Criterion (31) can be rewritten as

$$Q_T(\varphi) + \lambda_T \|\varphi\|_H^2 = \langle \varphi, (\lambda_T + \hat{A}^*\hat{A})\varphi \rangle_H - 2\langle \varphi, \hat{A}^*\hat{r} \rangle_H, \quad (32)$$

up to a term independent of φ . From Lemma A.2 (ii), and since $\hat{A}^*\hat{A}$ is positive, the operator

$\lambda_T + \hat{A}^*\hat{A}$ is invertible (Kress (1999), Theorem 3.4). It follows that the quadratic criterion

function (32) admits a global minimum over $H^2[0, 1]$. It is given by the first-order condition

$(\lambda_T + \hat{A}^*\hat{A})\hat{\varphi} = \hat{A}^*\hat{r}$, that is

$$\hat{\varphi} = (\lambda_T + \hat{A}^*\hat{A})^{-1} \hat{A}^*\hat{r}. \quad (33)$$

A.3.2 Asymptotic expansion of the first-order condition

Let us now expand the estimator in (33). We can write

$$\begin{aligned} \hat{r}(z) &= \int (y - \varphi_0(x)) \frac{\hat{f}(w, z)}{f(z)} dw + \int \varphi_0(x) \hat{f}(w|z) dw + \int (y - \varphi_0(x)) \left[\hat{f}(w|z) - \frac{\hat{f}(w, z)}{f(z)} \right] dw \\ &=: \hat{\psi}(z) + (\hat{A}\varphi_0)(z) + \hat{q}(z). \end{aligned}$$

Hence, $\hat{A}^* \hat{r} = A^* \hat{\psi} + \hat{A}^* \hat{A} \varphi_0 + \left(\hat{A}^* (\hat{q} + \hat{\psi}) - A^* \hat{\psi} \right)$, which yields

$$\hat{\varphi} - \varphi_0 = (\lambda_T + A^* A)^{-1} A^* \hat{\psi} + [(\lambda_T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0] + \mathcal{R}_T =: \mathcal{V}_T + \mathcal{B}_T + \mathcal{R}_T, \quad (34)$$

where the remaining term \mathcal{R}_T is given by

$$\begin{aligned} \mathcal{R}_T &= \left[(\lambda_T + \hat{A}^* \hat{A})^{-1} - (\lambda_T + A^* A)^{-1} \right] A^* \hat{\psi} \\ &\quad + \left[(\lambda_T + \hat{A}^* \hat{A})^{-1} \hat{A}^* \hat{A} - (\lambda_T + A^* A)^{-1} A^* A \right] \varphi_0 + (\lambda_T + \hat{A}^* \hat{A})^{-1} \left(\hat{A}^* (\hat{q} + \hat{\psi}) - A^* \hat{\psi} \right). \end{aligned} \quad (35)$$

The remaining term \mathcal{R}_T accounts for estimation of operator A and density $f(z)$ in the denominator of $\hat{r}(z)$. We prove at the end of this Appendix (Section A.3.5) that \mathcal{R}_T in (34) is asymptotically negligible, i.e. $E [\|\mathcal{R}_T\|^2] = o(E [\|\mathcal{V}_T + \mathcal{B}_T\|^2])$. Then, we deduce $E [\|\hat{\varphi} - \varphi_0\|^2] = E [\|\mathcal{V}_T + \mathcal{B}_T\|^2] + o(E [\|\mathcal{V}_T + \mathcal{B}_T\|^2])$ by applying twice the Cauchy-Schwarz inequality. Since

$$\begin{aligned} E [\|\mathcal{V}_T + \mathcal{B}_T\|^2] &= \left\| (\lambda_T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0 + (\lambda_T + A^* A)^{-1} A^* E \hat{\psi} \right\|^2 \\ &\quad + E \left[\left\| (\lambda_T + A^* A)^{-1} A^* (\hat{\psi} - E \hat{\psi}) \right\|^2 \right], \end{aligned} \quad (36)$$

we get

$$\begin{aligned} E [\|\hat{\varphi} - \varphi_0\|^2] &= \left\| (\lambda_T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0 + (\lambda_T + A^* A)^{-1} A^* E \hat{\psi} \right\|^2 \\ &\quad + E \left[\left\| (\lambda_T + A^* A)^{-1} A^* (\hat{\psi} - E \hat{\psi}) \right\|^2 \right], \end{aligned} \quad (37)$$

up to a term which is asymptotically negligible w.r.t. the RHS. This asymptotic expansion consists of a bias term (regularization bias plus estimation bias) and a variance term, which

will be analyzed separately in Lemmas A.3 and A.4 hereafter. Combining these two Lemmas and the asymptotic expansion in (37) results in Proposition 3.

A.3.3 Asymptotic expansion of the variance term

Lemma A.3: *Under Assumptions B, up to a term which is asymptotically negligible w.r.t.*

the RHS, we have $E \left[\left\| (\lambda_T + A^*A)^{-1} A^* (\hat{\psi} - E\hat{\psi}) \right\|^2 \right] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \|\phi_j\|^2.$

A.3.4 Asymptotic expansion of the bias term

Lemma A.4: *Define* $b(\lambda_T) = \left\| (\lambda_T + A^*A)^{-1} A^* A \varphi_0 - \varphi_0 \right\|.$ *Then, under Assumptions B*

and the bandwidth condition $h_T^m = o(\lambda_T b(\lambda_T)),$ *where* m *is the order of the kernel* $K,$ *we*

have $\left\| (\lambda_T + A^*A)^{-1} A^* A \varphi_0 - \varphi_0 + (\lambda_T + A^*A)^{-1} A^* E\hat{\psi} \right\| = b(\lambda_T),$ *up to a term which is asymptotically negligible w.r.t. the RHS.*

A.3.5 Control of the residual term

Lemma A.5: (i) *Assume the bandwidth conditions* $\frac{(\log T)^2}{Th_T^{d_Z}} + h_T^m = o(\lambda_T b(\lambda_T)),$

$E \left[\left\| \left(1 + S(\lambda_T) \hat{U}\right)^{-1} S(\lambda_T) \hat{U} \right\|^8 \right] = O(1),$ *and* $E \left[\left\| S(\lambda_T) \hat{U} \right\|^8 \right] = o(1),$ *where* m *is*

the order of the kernel $K,$ d_Z *is the dimensions of* $Z,$ $S(\lambda_T) := (\lambda_T + A^*A)^{-1},$ *and*

$\hat{U} := \hat{A}^* \hat{A} - A^* A.$ *Then, under Assumptions B,* $E \left[\|\mathcal{R}_T\|^2 \right] = o \left(E \left[\|V_T + \mathcal{B}_T\|^2 \right] \right).$

(ii) *If* $\left(\frac{1}{Th_T} + h_T^{2m} \right) = O(\lambda_T^{2+\varepsilon}),$ $\varepsilon > 0,$ *and* $\frac{1}{Th_T^{1+2d_Z}} = O(1),$ *then*

$E \left[\left\| \left(1 + S(\lambda_T) \hat{U}\right)^{-1} S(\lambda_T) \hat{U} \right\|^8 \right] = O(1)$ *and* $E \left[\left\| S(\lambda_T) \hat{U} \right\|^8 \right] = o(1).$

Appendix 4: Rate of convergence with geometric spectrum

(i) The first point follows from the next Lemma A.6, which characterizes the variance term (see Wahba (1977) for similar results).

Lemma A.6: *Let $e^{\alpha j} \nu_j \rightarrow C_1$ and $j^\beta \|\phi_j\|^2 \rightarrow C_2$, for constants $\alpha, \beta, C_1, C_2 > 0$, and define function $I(\lambda) = \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda + \nu_j)^2} \|\phi_j\|^2$, $\lambda > 0$. Then, $\lambda [\log(1/\lambda)]^\beta I(\lambda) = \left(\frac{1}{\alpha}\right)^{1-\beta} C_2 [1 + c(\lambda)] + o(1)$, as $\lambda \rightarrow 0$, where $c(\lambda)$ is a function such that $|c(\lambda)| \leq 1/4$ and $\left| \lambda \frac{dc}{d\lambda}(\lambda) \right| \leq 1/4$.*

Function $c(\lambda)$ is given in the Technical Report.

(ii) The optimal sequence λ_T^* is obtained by minimizing function $M_T(\lambda)$ w.r.t. λ . We have

$$\begin{aligned} \frac{dM_T(\lambda)}{d\lambda} &= -\frac{c_1}{T} \frac{1+c(\lambda)}{\lambda^2 [\log(1/\lambda)]^{2\beta}} \left([\log(1/\lambda)]^\beta - \lambda\beta [\log(1/\lambda)]^{\beta-1} \frac{1}{\lambda} \right) + \frac{c_1}{T} \frac{dc/d\lambda}{\lambda [\log(1/\lambda)]^\beta} \\ &\quad + 2c_2\delta\lambda^{2\delta-1} = -\frac{1}{T} \frac{\kappa(\lambda)}{\lambda^2 [\log(1/\lambda)]^\beta} + 2c_2\delta\lambda^{2\delta-1}, \end{aligned}$$

where $\kappa(\lambda) := c_1 [1 + c(\lambda)] \left[1 - \frac{\beta}{\log(1/\lambda)} \right] - \lambda c_1 \frac{dc}{d\lambda}(\lambda)$. From Lemma A.6 function $\kappa(\lambda)$ is positive, bounded and bounded away from 0 as $\lambda \rightarrow 0$. Computation of the second derivative shows that $M_T(\lambda)$ is a convex function of λ , for small λ . We get

$$\frac{dM_T(\lambda_T^*)}{d\lambda} = 0 \iff \frac{1}{T} \frac{1}{2c_2\delta} \frac{\kappa(\lambda_T^*)}{[\log(1/\lambda_T^*)]^\beta} = (\lambda_T^*)^{2\delta+1}. \quad (38)$$

To solve the latter equation for λ_T^* , define $\tau_T := \log(1/\lambda_T^*)$. Then $\tau_T = c_3 + \frac{1}{1+2\delta} \log T + \frac{\beta}{1+2\delta} \log \tau_T - \frac{1}{1+2\delta} \log \kappa(\lambda_T^*)$, where $c_3 = (1+2\delta)^{-1} \log(2c_2\delta)$. It follows that $\tau_T = -\log c^* + \frac{1}{1+2\delta} \log T + \frac{\beta}{1+2\delta} \log \tau_T - \frac{1}{1+2\delta} \log(1+c_T) + o(1)$, where $\log c^* = \frac{1}{1+2\delta} \log\left(\frac{c_1}{2c_2\delta}\right)$

and $c_T := c(\lambda_T^*) - \lambda_T^* \frac{dc}{d\lambda}(\lambda_T^*)$. In the proof of Lemma A.6 in the Technical Report, we show

$$\text{that } \frac{1}{(1 + \frac{1}{2}e^\alpha)^2} - \frac{1}{(1 + \frac{1}{2}e^{-\alpha})^2} \leq c_T \leq \frac{8\alpha}{27}.$$

(iii) Finally, let us compute the MISE corresponding to λ_T^* . We have

$$M_T(\lambda_T^*) = c_1 \frac{1}{T} \frac{1 + c(\lambda_T^*)}{\lambda_T^* [\log(1/\lambda_T^*)]^\beta} + c_2 (\lambda_T^*)^{2\delta} = c_1 \frac{1}{T} \frac{1 + c(\lambda_T^*)}{\lambda_T^* \tau_T^\beta} + c_2 (\lambda_T^*)^{2\delta}.$$

From (38), $\lambda_T^* = \left(\frac{1}{2c_2\delta} \kappa(\lambda_T^*) \right)^{\frac{1}{2\delta+1}} T^{-\frac{1}{2\delta+1}} \left(\frac{1}{\tau_T^\beta} \right)^{\frac{1}{2\delta+1}} =: c_{4,T} T^{-\frac{1}{2\delta+1}} \tau_T^{-\frac{\beta}{2\delta+1}}$. Thus we get

$$\begin{aligned} M_T(\lambda_T^*) &= c_1 \frac{1}{T} \frac{1 + c(\lambda_T^*)}{c_{4,T}} T^{\frac{1}{2\delta+1}} \frac{1}{\tau_T^{-\frac{\beta}{2\delta+1} + \beta}} + c_2 c_{4,T}^{2\delta} T^{-\frac{2\delta}{2\delta+1}} \tau_T^{-\frac{2\delta\beta}{2\delta+1}} \\ &= \left(c_1 \frac{1 + c(\lambda_T^*)}{c_{4,T}} + c_2 c_{4,T}^{2\delta} \right) T^{-\frac{2\delta}{2\delta+1}} \tau_T^{-\frac{2\delta\beta}{2\delta+1}} = c_T^* T^{-\frac{2\delta}{2\delta+1}} (\log T)^{-\frac{2\delta\beta}{2\delta+1}}, \end{aligned}$$

where $c_T^* = \frac{c_2}{(1 + 2\delta)^{-\frac{2\delta\beta}{1+2\delta}}} \left(\frac{c_1}{2c_2\delta} (1 + c_T) \right)^{\frac{2\delta}{1+2\delta}} \left(2\delta + \frac{1 + \bar{c}_T}{1 + c_T} \right) + o(1)$ and $\bar{c}_T := c(\lambda_T^*)$. In

the proof of Lemma A.6 in the Technical Report, we show that $\frac{1}{1 + e^{\alpha/2}} - \frac{1}{1 + e^{-\alpha}} \leq \bar{c}_T \leq \frac{\alpha}{4}$.

Appendix 5: Asymptotic normality of the TiR estimator

From Equation (34) in Appendix 3, we have

$$\begin{aligned} \sqrt{T/\sigma_T^2(x)} (\hat{\varphi}(x) - \varphi_0(x)) &= \sqrt{T/\sigma_T^2(x)} (\lambda_T + A^*A)^{-1} A^* (\hat{\psi} - E\hat{\psi})(x) + \sqrt{T/\sigma_T^2(x)} \mathcal{B}_T(x) \\ &\quad + \sqrt{T/\sigma_T^2(x)} (\lambda_T + A^*A)^{-1} A^* E\hat{\psi}(x) + \sqrt{T/\sigma_T^2(x)} \mathcal{R}_T(x) \\ &=: \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)}, \end{aligned}$$

where $\mathcal{R}_T(x)$ is defined in (35). We now show that the term (I) is asymptotically $N(0, 1)$ distributed and the terms (III) and (IV) are $o_p(1)$, which implies Proposition 6.

A.5.1 Asymptotic normality of (I)

Since $\{\phi_j : j \in \mathbb{N}\}$ is an orthonormal basis w.r.t. $\langle \cdot, \cdot \rangle_H$, we can write:

$$\begin{aligned} (\lambda_T + A^*A)^{-1} A^* (\hat{\psi} - E\hat{\psi})(x) &= \sum_{j=1}^{\infty} \left\langle \phi_j, (\lambda_T + A^*A)^{-1} A^* (\hat{\psi} - E\hat{\psi}) \right\rangle_H \phi_j(x) \\ &= \sum_{j=1}^{\infty} \frac{1}{\lambda_T + \nu_j} \left\langle \phi_j, A^* (\hat{\psi} - E\hat{\psi}) \right\rangle_H \phi_j(x), \end{aligned}$$

for almost any $x \in [0, 1]$. Then, we get

$$\sqrt{T/\sigma_T^2(x)} (\lambda_T + A^*A)^{-1} A^* (\hat{\psi} - E\hat{\psi})(x) = \sum_{j=1}^{\infty} w_{j,T}(x) Z_{j,T}, \quad (39)$$

where $Z_{j,T} := \frac{1}{\sqrt{\nu_j}} \langle \phi_j, \sqrt{T} A^* (\hat{\psi} - E\hat{\psi}) \rangle_H$, $j = 1, 2, \dots$,

and $w_{j,T}(x) := \frac{\sqrt{\nu_j}}{\lambda_T + \nu_j} \phi_j(x) / \left(\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2 \right)^{1/2}$, $j = 1, 2, \dots$.

Note that $\sum_{j=1}^{\infty} w_{j,T}(x)^2 = 1$. Equation (39) can be rewritten (see the proof of Lemma A.3) using

$$\sum_{j=1}^{\infty} w_{j,T}(x) Z_{j,T} = \sqrt{T} \int G_T(u) [\hat{f}(u) - E\hat{f}(u)] du, \quad (40)$$

$u = (w, z)$, $G_T(u) := G_T(x, u) := \sum_{j=1}^{\infty} w_{j,T}(x) g_j(u)$ and $g_j(u) = (A\phi_j)(z) \Omega_0(z) g_{\varphi_0}(w) / \sqrt{\nu_j}$.

Lemma A.7: *Under Assumptions B and $h_T^m = o(\lambda_T)$, $\sqrt{T} \int G_T(u) [\hat{f}(u) - E\hat{f}(u)] du = \frac{1}{\sqrt{T}} \sum_{t=1}^T Y_{tT} + o_p(1)$, where $Y_{tT} := G_T(U_t) = \sum_{j=1}^{\infty} w_{j,T}(x) g_j(U_t)$.*

From Lemma A.7 it is sufficient to prove that $T^{-1/2} \sum_{t=1}^T Y_{tT}$ is asymptotically $N(0, 1)$

distributed. Note that $E[g_j(U)] = \frac{1}{\sqrt{\nu_j}} E[(A\phi_j)(Z) \Omega_0(Z) E[g_{\varphi_0}(W) | Z]] = 0$, and

$$\begin{aligned} \text{Cov}[g_j(U), g_l(U)] &= \frac{1}{\sqrt{\nu_j} \sqrt{\nu_l}} E[(A\phi_j)(Z) \Omega_0(Z) E[g_{\varphi_0}(W)^2 | Z] \Omega_0(Z) (A\phi_l)(Z)] \\ &= \frac{1}{\sqrt{\nu_j} \sqrt{\nu_l}} E[(A\phi_j)(Z) \Omega_0(Z) (A\phi_l)(Z)] = \frac{1}{\sqrt{\nu_j} \sqrt{\nu_l}} \langle \phi_j, A^* A \phi_l \rangle_H = \delta_{j,l}. \end{aligned}$$

Thus $E[Y_{tT}] = 0$ and $V[Y_{tT}] = \sum_{j,l=1}^{\infty} w_{j,T}(x) w_{l,T}(x) \text{Cov}[g_j(U), g_l(U)] = \sum_{j=1}^{\infty} w_{j,T}(x)^2 = 1$.

From application of a Lyapunov CLT, it is sufficient to show that

$$\frac{1}{T^{1/2}} E[|Y_{tT}|^3] \rightarrow 0, \quad T \rightarrow \infty. \quad (41)$$

To this goal, using $|Y_{tT}| \leq \sum_{j=1}^{\infty} |w_{j,T}(x)| |g_j(U_t)|$ and the triangular inequality, we get

$$\begin{aligned} \frac{1}{T^{1/2}} E[|Y_{tT}|^3] &\leq \frac{1}{T^{1/2}} E \left[\left(\sum_{j=1}^{\infty} |w_{j,T}(x)| |g_j(U)| \right)^3 \right] = \frac{1}{T^{1/2}} \left\| \sum_{j=1}^{\infty} |w_{j,T}(x)| |g_j| \right\|_3^3 \\ &\leq \frac{1}{T^{1/2}} \left(\sum_{j=1}^{\infty} |w_{j,T}(x)| \|g_j\|_3 \right)^3 = \frac{1}{T^{1/2}} \frac{\left(\sum_{j=1}^{\infty} \frac{\sqrt{\nu_j}}{\lambda_T + \nu_j} |\phi_j(x)| \|g_j\|_3 \right)^3}{\left(\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2 \right)^{3/2}}. \end{aligned}$$

Moreover, from the Cauchy-Schwarz inequality we have

$$\sum_{j=1}^{\infty} \frac{\sqrt{\nu_j}}{\lambda_T + \nu_j} |\phi_j(x)| \|g_j\|_3 \leq \left(\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2 \|g_j\|_3^2 j^{1+\bar{\varepsilon}} \right)^{1/2} \left(\sum_{j=1}^{\infty} \frac{1}{j^{1+\bar{\varepsilon}}} \right)^{1/2},$$

and $\sum_{j=1}^{\infty} \frac{1}{j^{1+\bar{\varepsilon}}} < \infty$, for any $\bar{\varepsilon} > 0$. Thus, we get

$$\frac{1}{T^{1/2}} E[|Y_{tT}|^3] \leq \left(\sum_{j=1}^{\infty} \frac{1}{j^{1+\bar{\varepsilon}}} \right)^{3/2} \left(\frac{1}{T^{1/3}} \frac{\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2 \|g_j\|_3^2 j^{1+\bar{\varepsilon}}}{\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2} \right)^{3/2},$$

and Condition (41) is implied by Condition (26).

A.5.2 Terms (III) and (IV) are $o(1)$, $o_p(1)$

Lemma A.8: *Under Assumptions B, $h_T^m = O\left(\lambda_T^{1+\varepsilon/2}b(\lambda_T)\right)$ and $\frac{M_T(\lambda_T)}{\sigma_T^2(x)/T} = o(\lambda_T^{-\varepsilon})$, for $a \varepsilon > 0$: $\sqrt{T/\sigma_T^2(x)}(\lambda_T + A^*A)^{-1}A^*E\hat{\psi}(x) = o(1)$.*

Lemma A.9: *Suppose Assumptions B hold, and $\frac{(\log T)^2}{Th_T^{d_Z}} + h_T^m = O\left(\lambda_T^{1+\varepsilon/2}b(\lambda_T)\right)$, $\frac{1}{T^{1+2d_Z}} = O(1)$, $\frac{1}{Th_T} + h_T^{2m} = O(\lambda_T^{2+\varepsilon})$, $\frac{M_T(\lambda_T)}{\sigma_T^2(x)/T} = o(\lambda_T^{-\varepsilon})$, for $a \varepsilon > 0$. Then: $\sqrt{T/\sigma_T^2(x)}\mathcal{R}_T(x) = o_p(1)$.*

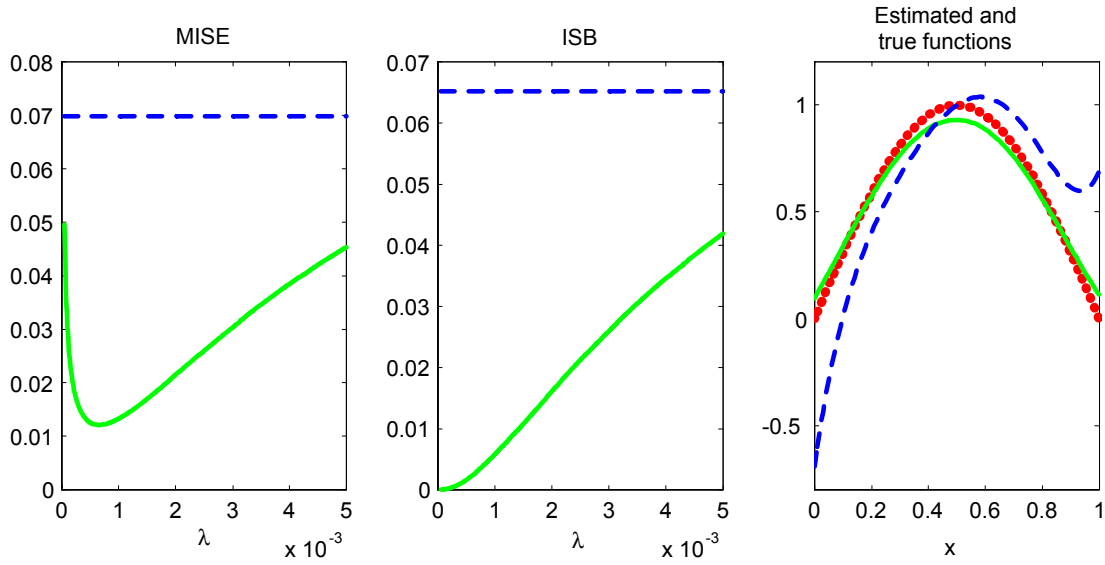


Figure 1: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel. Correlation parameter is $\rho = 0.5$, and sample size is $T = 1000$.

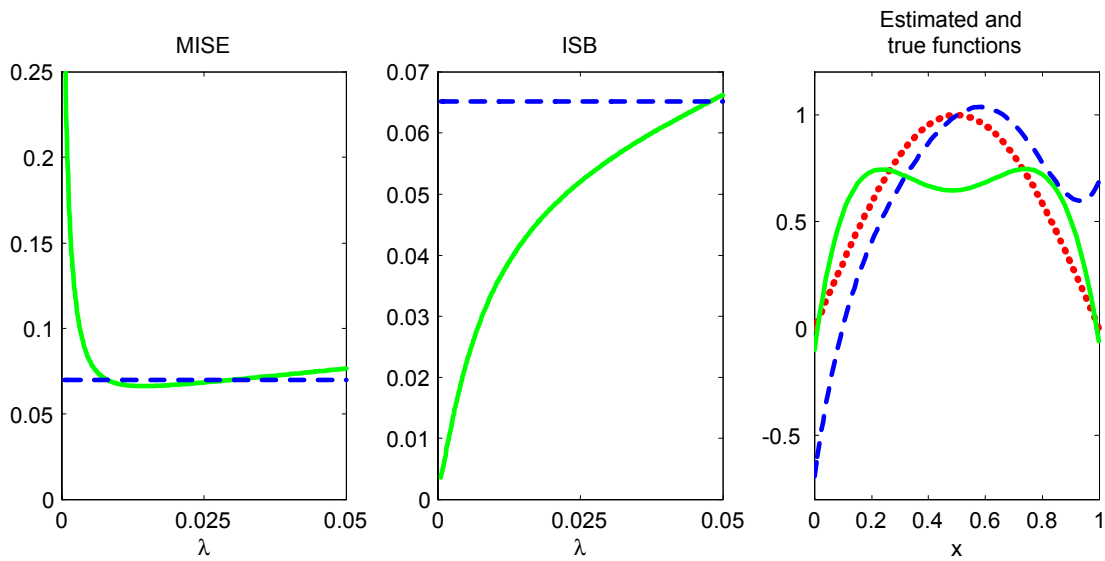


Figure 2: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the regularised estimator using L^2 norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel. Correlation parameter is $\rho = 0.5$, and sample size is $T = 1000$.

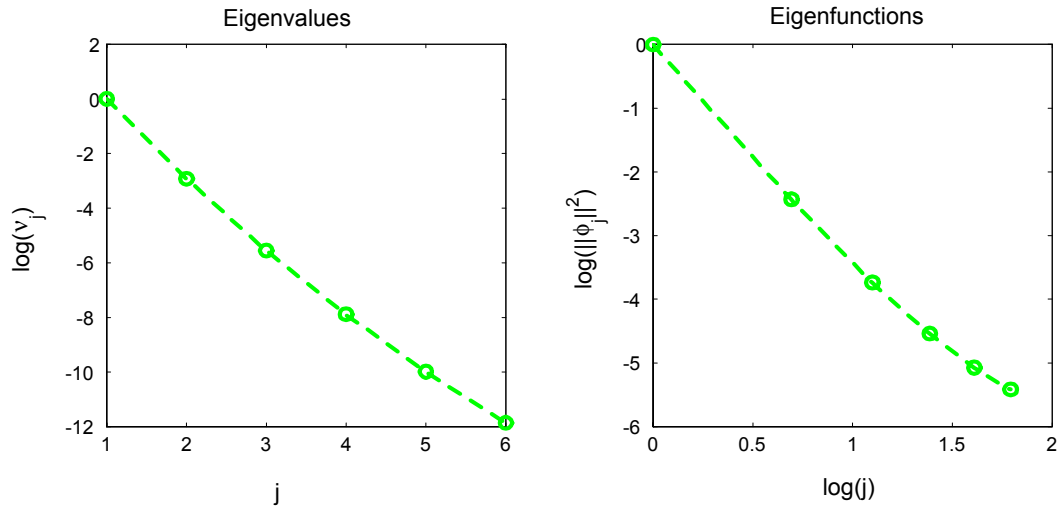


Figure 3: The eigenvalues (left Panel) and the L^2 -norms of the corresponding eigenfunctions (right Panel) of operator A^*A using the approximation with six polynomials.

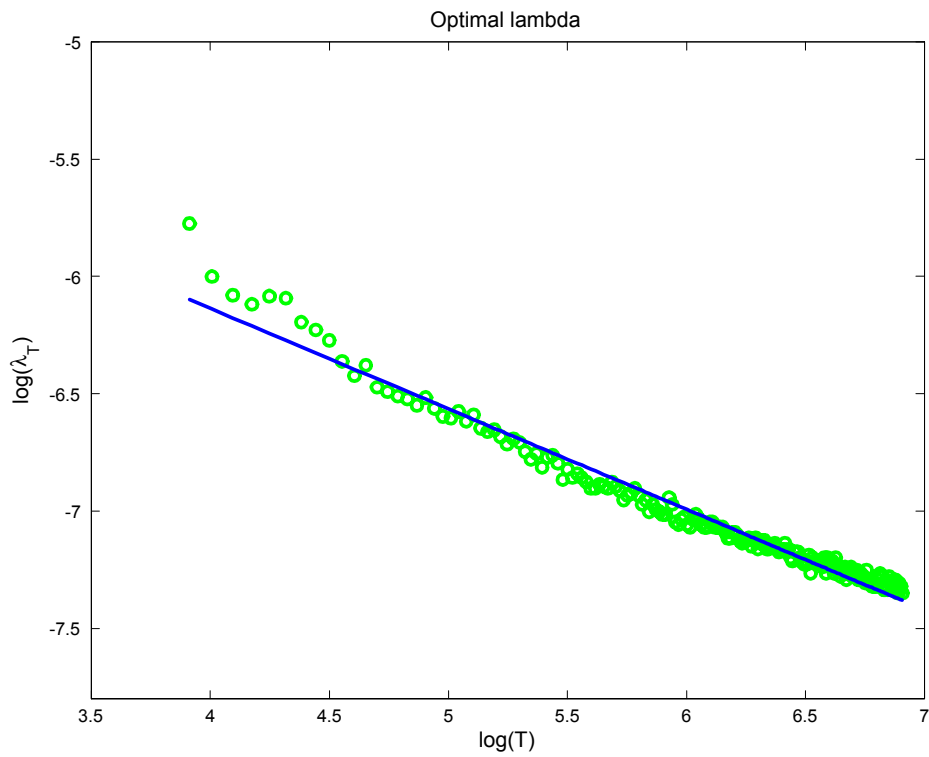


Figure 4: Log of optimal regularization parameter as a function of log of sample size.

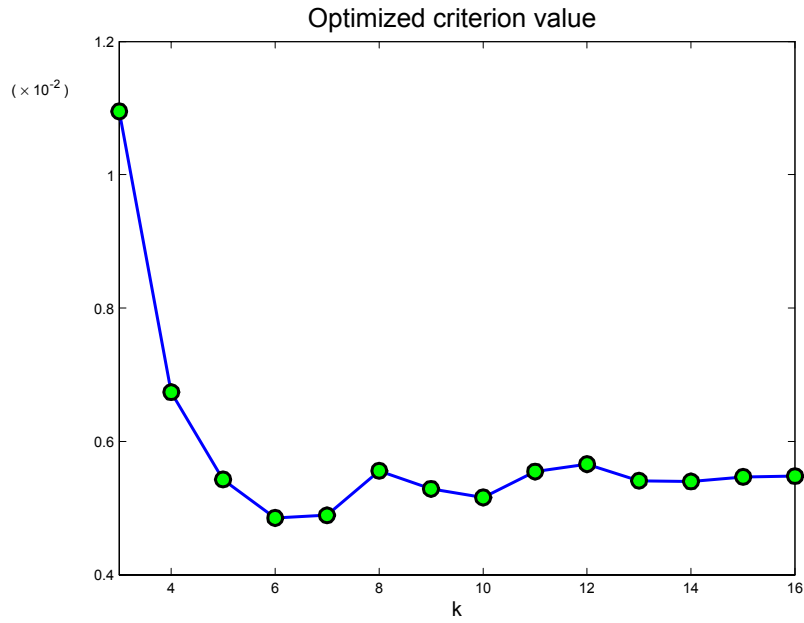


Figure 5: Value of the optimized objective function as a function of the number k of polynomials. The regularization parameter is selected with the spectral approach.

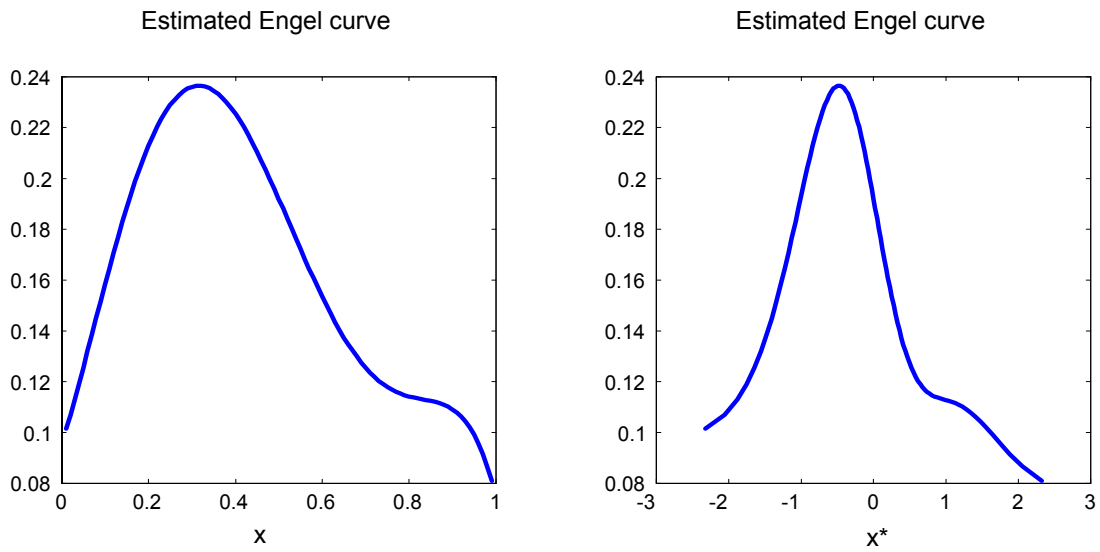


Figure 6: Estimated Engel curves for 785 household-level observations from the 1996 US Consumer Expenditure Survey. In the right Panel, food expenditure share Y is plotted as a function of the standardized logarithm X^* of total expenditures. In the left Panel, Y is plotted as a function of transformed variable $X = \Phi(X^*)$ with support $[0, 1]$, where Φ is the cdf of the standard normal distribution. Instrument Z is standardized logarithm of annual income from wages and salaries.