TIKHONOV REGULARIZATION FOR FUNCTIONAL MINIMUM DISTANCE ESTIMATORS

P. Gagliardini* **and O. Scaillet†**

This version: November 2006 ‡

(First version: May 2006)

*University of Lugano and Swiss Finance Institute.

†HEC Genève and Swiss Finance Institute.

‡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 many suggestions as well as Xiaohong Chen, 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 and the ESRC 2006 Annual Conference in Bristol for helpful comments.

Tikhonov Regularization for Functional Minimum Distance Estimators

Abstract

We study the asymptotic properties of a Tikhonov Regularized (TiR) estimator of a functional parameter based on a minimum distance principle for nonparametric conditional moment restrictions. The estimator is computationally tractable and takes a closed form in the linear case. We derive its asymptotic Mean Integrated Squared Error (MISE), its rate of convergence and its 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 for two numerical examples. 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 nonparametric 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

Minimum distance and extremum estimators have received a lot of attention in the literature. They exploit conditional moment restrictions assumed to hold true on the data generating process [see e.g. Newey and McFadden (1994) for a review]. In a parametric setting, leading examples are the Ordinary Least Squares estimator and the Nonlinear Least Squares estimator. Correction for endogeneity is provided by the Instrumental Variable (IV) estimator in the linear case and by the Generalized Method of Moments (GMM) estimator in the nonlinear case.

In a functional setting, regression curves are inferred by local polynomial estimators and sieve estimators. A well known example is the Parzen-Rosenblatt kernel estimator. Correction for endogeneity in a nonparametric context is motivated by functional IV estimation of structural equations. Newey and Powell (NP, 2003) consider nonparametric estimation of a regression function, which is identified by a conditional expectation given a set of instruments. Their consistent minimum distance estimator is a nonparametric analog of the Two-Stage Least Squares (2SLS) estimator. The NP methodology extends to the nonlinear case. Ai and Chen (AC, 2003) opt for a similar approach to estimate semiparametric specifications. Although their focus is on the efficient estimation of the finite-dimensional component, AC show that the estimator of the functional component converges at a rate faster than $T^{-1/4}$ in an appropriate metric. Darolles, Florens and Renault (DFR, 2003) and Hall and Horowitz (HH, 2005) concentrate on nonparametric estimation of an instrumental regression function. Their estimation approach is based on the empirical analog of the

conditional moment restriction, seen as a linear integral equation in the unknown functional parameter. HH derive the rate of convergence of their estimator in quadratic mean and show that it is optimal in the minimax sense. Horowitz (2005) shows the pointwise asymptotic normality for an asymptotically negligible bias. For further background, Florens (2003) and Blundell and Powell (2003) present surveys on endogenous nonparametric regressions.

There is a growing literature building on the above methods and considering empirical applications to different fields. Among others, Blundell, Chen and Kristensen (2004), Chen and Ludvigson (2004), Loubes and Vanhems (2004), Chernozhukov, Imbens and Newey (2006). Other related references include Newey, Powell, and Vella (1999), Chernozhukov and Hansen (2005), Carrasco and Florens (2000,2005), Hu and Schennach (2004), Florens, Johannes and Van Bellegem (2005), Horowitz (2006), and Horowitz and Lee (2006).

The main theoretical difficulty in nonparametric estimation with endogeneity is overcoming ill-posedness [see Kress (1999), Chapter 15, for a general treatment, and Carrasco, Florens and Renault (2005) for a survey in econometrics]. 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. A serious potential consequence is inconsistency of the estimators. To address ill-posedness NP and AC propose to introduce bounds on the functional parameter of interest and its derivatives. This amounts to set compacity on the parameter space. In the linear case, DFR and HH adopt a different 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. We consider penalized extremum estimators which minimize $Q_T(\varphi) + \lambda_T G(\varphi)$, 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 amount of 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 (Kress (1999), Groetsch (1984)). 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_0. To avoid confusion, we refer to DFR and HH estimators as regularized estimators with L^2 norm.

Our paper contributes to the literature along several directions. First, we introduce an estimator admitting appealing features: (i) it applies in a general (linear and nonlinear) setting; (ii) the tuning parameter is allowed to depend on sample size and to be stochastic; (iii) it may have a faster rate of convergence than L^2 regularized estimators in the linear case (DFR, HH); (iv) it has a faster rate of convergence than estimators based on bounding the Sobolev norm (NP, AC); (v) it admits a closed form in the linear case. Point (ii) is

crucial to develop estimators with data-driven selection of the regularization parameter. This point is not addressed in the setting of NP and AC, where the tuning parameter is constant. Concerning point (iii), we give in Section 4 several conditions under which this property holds. In our Monte-Carlo experiments in Section 6, we find a clear-cut superior performance of the TiR estimator compared to the regularized estimator with L^2 norm. [1]

Point (iv) is induced by the requirement of a fixed bound on the Sobolev norm in the approach of NP and AC. Point (v) is not shared by NP and AC estimators because of the inequality constraint. We will further explain the links between the TiR estimator and the literature in Section 2.4.

Second, we study in depth the asymptotic properties of the TiR estimator: (a) we prove consistency; (b) we derive the asymptotic expansion of the Mean Integrated Squared Error (MISE); (c) we characterize the MSE, and prove the pointwise asymptotic normality when bias is still present asymptotically. To the best of our knowledge, results (b) and (c), as well as (a) for a sequence of stochastic regularization parameters, are new for nonparametric instrumental regression estimators. In particular, the asymptotic expansion of the MISE allows us to study the effect of the regularization parameter on the variance term and on the bias term of our estimator, to find the optimal sequence of regularization parameters, and to derive the associated optimal rate of convergence. We parallel the analysis for L^2 regularized estimators, and provide a comparison. Finally, the asymptotic expansion of the MISE suggests a quick procedure for the data-driven selection of the regularization

[1] The advantage of the Sobolev norm compared to the L^2 norm for regularization of ill-posed inverse problems is also pointed out in a numerical example in Kress (1999), Example 16.21.

parameter, that we implement in the Monte-Carlo study.

Third, we investigate the attractiveness of the TiR estimator from an applied point of view. In the nonlinear case, the TiR estimator only requires running an unconstrained optimization routine instead of a constrained one. In the linear case it even takes a closed form. Numerical tractability is a key advantage to apply resampling techniques. The finite sample properties are promising from our numerical experiments on two examples mimicking possible shapes of Engel curves and with two data driven selection procedures of the regularization parameter.

The rest of the paper is organized as follows. In Section 2, we introduce the general setting of nonparametric estimation under conditional moment restrictions and the problem of ill-posedness. We define the TiR estimator, and discuss the links with the literature. In Section 3 we prove its consistency through establishing a general result for penalized extremum estimators with stochastic regularization parameter. Section 4 is devoted to the characterization of the asymptotic MISE and examples of optimal rates of convergence for the TiR estimator with deterministic regularization parameter. We compare these results with those obtained under regularization via an L^2 norm. We further discuss the suboptimality of a fixed bounding of the Sobolev norm. We also derive the asymptotic MSE and establish pointwise asymptotic normality of the TiR estimator. Implementation for linear moment restrictions is outlined in Section 5. In Section 6 we illustrate numerically our theoretical findings, and present a Monte-Carlo study of the finite sample properties. We also describe two data driven selection procedures of the regularization parameter, and show that they

perform well in practice. We provide an empirical example in Section 7 where we estimate an Engel curve nonparametrically. Section 8 concludes. Proofs of theoretical results are gathered in the Appendices. All omitted proofs of technical Lemmas are collected in a Technical Report, which is available from the authors on request.

2 Minimum distance estimators under Tikhonov regularization

2.1 Nonparametric minimum distance estimation

Let $\{(Y_t, X_t, Z_t) : t = 1, ..., T\}$ be i.i.d. copies of $d \times 1$ 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 $\mathcal{X} = [0, 1]$. [2] Suppose that the parameter of interest is a function φ_0 defined on $\mathcal{X}$, which satisfies the conditional moment restriction

$$E\left[g\left(Y, \varphi_0\left(X\right)\right) \mid Z\right] = 0, \tag{1}$$

where g is a known function. 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 $\langle\varphi, \psi\rangle = \int_{\mathcal{X}} \varphi(x)\psi(x)dx$ (see Gallant and Nychka (1987) for use of Sobolev spaces as functional parameter set). The Sobolev space $H^2\left[0, 1\right]$ 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

[2] We need compactness of the support of X for technical reasons. Mapping in $[0, 1]$ can be achieved by simple linear or nonlinear monotone transformations. Assuming univariate X simplifies the exposition. Extension of our theoretical results to higher dimensions is straightforward. Then the estimation methodology can also be extended to the general case where X and Z have common elements.

norm. Further, we assume the following identification condition. [3]

Assumption 1 (Identification): (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* $\|.\|$.

The nonparametric minimum distance approach relies on φ_0 minimizing

$$Q_\infty(\varphi) = E\left[m\left(\varphi, Z\right)' \Omega_0(Z) m\left(\varphi, Z\right)\right], \qquad \varphi \in \Theta, \tag{2}$$

where $m\left(\varphi, z\right) := E\left[g\left(Y, \varphi\left(X\right)\right) \mid Z = z\right]$, and $\Omega_0(z)$ is a positive definite matrix for any given z. The criterion (2) is well-defined if $m\left(\varphi, z\right)$ belongs to $L^2_{\Omega_0}(F_Z)$, for any $\varphi \in \Theta$, where $L^2_{\Omega_0}(F_Z)$ denotes the L^2 space of square integrable vector-valued functions of Z defined by scalar product $\langle \psi_1, \psi_2 \rangle_{L^2_{\Omega_0}(F_Z)} = E\left[\psi_1\left(Z\right)' \Omega_0(Z) \psi_2\left(Z\right)\right]$. Then, the idea is to estimate φ_0 by the minimizer of the empirical counterpart of (2). For instance, AC and NP estimate the conditional moment $m\left(\varphi, z\right)$ by an orthogonal polynomial approach, and minimize the empirical criterion over a finite-dimensional sieve approximation of Θ.

The main difficulty in nonparametric minimum distance estimation is that Assumption 1 is not sufficient to ensure consistency of the estimator. This is due to the so-called ill-posedness of such an estimation problem.

[3] See 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 compacity of Θ in infinite dimensional spaces. See Chen (2006), and Horowitz and Lee (2006) for similar noncompact settings.

2.2 Unidentifiability and ill-posedness in minimum distance estimation

The goal of this section is to highlight the issue of ill-posedness in minimum distance estimation (NP; see also Kress (1999) and Carrasco, Florens and Renault (2005)). To explain this, observe that solving the integral equation $E\left[g\left(Y,\varphi\left(X\right)\right)\mid Z\right]=0$ for the unknown function $\varphi\in\Theta$ can be seen as an inverse problem, which maps the conditional distribution $F_0\left(y,x|z\right)$ of (Y,X) given $Z=z$ into the solution φ_0 (cf. (1)). Ill-posedness arises when this mapping is not continuous. Then the estimator $\hat{\varphi}$ of φ_0, which is the solution of the inverse problem corresponding to a consistent estimator $\hat{F}$ of F_0, is not guaranteed to be consistent. Indeed, by lack of continuity, small deviations of $\hat{F}$ from F_0 may result in large deviations of $\hat{\varphi}$ from φ_0. We refer to NP for further discussion along these lines. Here we prefer to develop the link between ill-posedness and a classical concept in econometrics, namely parameter unidentifiability.

To illustrate the main point, let us consider the case of nonparametric linear IV estimation, where $g(y,\varphi(x))=\varphi\left(x\right)-y$, and

$$m(\varphi,z)=\left(A\varphi\right)\left(z\right)-r\left(z\right)=\left(A\Delta\varphi\right)\left(z\right), \tag{3}$$

where $\Delta\varphi:=\varphi-\varphi_0$, operator A is defined by $\left(A\varphi\right)\left(z\right)=\int\varphi(x)f(w|z)dw$ 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. The limit criterion in (2) becomes

$$Q_\infty(\varphi)=E\left[\left(A\Delta\varphi\right)\left(Z\right)'\Omega_0(Z)\left(A\Delta\varphi\right)\left(Z\right)\right]=\langle\Delta\varphi,A^*A\Delta\varphi\rangle_H, \tag{4}$$

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

Under weak regularity conditions, the integral operator A is compact in $L^2[0,1]$. Thus, A^*A is compact and self-adjoint in $H^2[0,1]$. We denote by $\left\{\phi_j : j \in \mathbb{N}\right\}$ an orthonormal basis in $H^2[0,1]$ of eigenfunctions of operator A^*A, and by $\nu_1 \geq \nu_2 \geq \cdots > 0$ 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 $\nu_j \to 0$, and it can be shown that $\nu_j / \left\|\phi_j\right\|^2 \to 0$. The limit criterion $Q_\infty(\varphi)$ can be minimized by a sequence in Θ such as

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

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

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

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.

In the general nonlinear setting (1), we link failure of Condition (6) with compacity of the operator induced by the linearization of moment function $m(\varphi, z)$ around $\varphi = \varphi_0$.

Assumption 2 (Ill-posedness): *The moment function* $m(\varphi, z)$ *is such that* $m(\varphi, z) = (A\Delta\varphi)(z) + R(\varphi, z)$, *for any* $\varphi \in \Theta$, *where*

(i) *the operator* A *defined by* $(A\Delta\varphi)(z) = \int \nabla_v g(y, \varphi_0(x)) f(w|z) \Delta\varphi(x) dw$ *is a compact operator in* $L^2[0,1]$ *and* $\nabla_v g$ *is the derivative of* g *w.r.t. its second argument;*

(ii) *the second-order term* $R(\varphi, z)$ *is such that for any sequence* $(\varphi_n) \subset \Theta$:

$\langle \Delta\varphi_n, A^*A\Delta\varphi_n \rangle_H \to 0 \implies Q_\infty(\varphi_n) \to 0.$

Under Assumption 2, the identification condition (6) is not satisfied, and the minimum distance estimator which minimizes the empirical counterpart of criterion $Q_\infty(\varphi)$ over the set Θ (or a sieve approximation of Θ) is not consistent w.r.t. the L^2-norm $\|.\|$.

In the ill-posed setting, Horowitz and Lee (2006) emphasize that Assumption 2 (ii) is not implied by a standard Taylor expansion argument (see also Chapter 10 in Engl, Hanke and Neubauer (2000)). Indeed, the residual term $R(\varphi, .)$ may well dominate $A\Delta\varphi$ along the directions $\Delta\varphi$ where $A\Delta\varphi$ is small. Assumption 2 (ii) requires that a sequence (φ_n) minimizes Q_∞ if the second derivative $\nabla_t^2 Q_\infty(\varphi_0 + t\Delta\varphi_n)\big|_{t=0} = 2\langle \Delta\varphi_n, A^*A\Delta\varphi_n \rangle_H$ of the criterion Q_∞ at φ_0 in direction $\Delta\varphi_n$ becomes small, i.e., if Q_∞ gets flat in direction $\Delta\varphi_n$. [4] For a moment function $m(\varphi, z)$ linear in φ, Assumption 2 (ii) is clearly satisfied. In the general nonlinear case, it provides a local rule for the presence of ill-posedness, namely compacity of the linearized operator A.

[4] Since $\|\varphi_1 - \varphi_2\|_w^2 := \nabla_t^2 Q_\infty(\varphi_0 + t(\varphi_1 - \varphi_2))\big|_{t=0}$ corresponds to the metric introduced by AC in their Equation (14), Assumption 2 (ii) is tightly related to their Assumption 3.9 (ii).

2.3 The Tikhonov Regularized (TiR) estimator

In this paper, we address ill-posedness by introducing minimum distance estimators based on Tikhonov regularization. We consider a penalized criterion $Q_T(\varphi) + \lambda_T G(\varphi)$. The criterion $Q_T(\varphi)$ is an empirical counterpart of (2) defined by

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

where $\hat{\Omega}(z)$ is a sequence of positive definite matrices converging to $\Omega_0(z)$, P-a.s., for any z. In (7) we estimate the conditional moment nonparametrically with $\hat{m}(\varphi, z) = \int g(y, \varphi(x)) \hat{f}(w|z)\, dw$, where $\hat{f}(w|z)$ denotes a kernel estimator of the density of (Y, X) given $Z = z$ with kernel K, bandwidth h_T, and $w = (y, x)$. Different choices of penalty function $G(\varphi)$ are possible, leading to consistent estimators under the assumptions of Theorem 1 in Section 3 below. In this paper, we focus on $G(\varphi) = \|\varphi\|_H^2$. [5]

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

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

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

The name Tikhonov Regularized (TiR) estimator is in line with the pioneering papers of Tikhonov (1963a,b) on the regularization of ill-posed inverse problems (see Kress (1999), Chapter 16). Intuitively, the presence of $\lambda_T \|\varphi\|_H^2$ in (8) penalizes highly oscillating components of the estimated function. These components would be otherwise unduly amplified,

[5] Instead, we could rely on a generalized Sobolev norm to get $G(\varphi) = \omega \|\varphi\|^2 + (1-\omega)\|\nabla\varphi\|^2$ with $\omega \in (0,1)$. Using $\omega = 0$ yields a penalization involving solely the derivative $\nabla\varphi$ but we loose the interpretation of a norm useful in the derivation of our asymptotic results.

since ill-posedness yields a criterion $Q_T(\varphi)$ asymptotically flat along some directions. In the linear IV case where $Q_\infty(\varphi) = \langle \Delta\varphi, A^*A\Delta\varphi\rangle_H$, 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 $\|\psi_n\|_H \to \infty$ as $n \to \infty$, so that these directions are penalized by $G(\varphi) = \|\varphi\|_H^2$ in (8). In Theorem 1 below, we provide precise conditions under which the penalty function $G(\varphi)$ restores the validity of the identification Condition (6), and ensures consistency. Finally, 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 $\hat{\varphi}$.

2.4 Links with the literature

2.4.1 Regularization by compactness

To address ill-posedness, NP and AC (see also Blundell, Chen and Kristensen (2004)) suggest considering a compact parameter set Θ. In this case, by the same argument as in the standard parametric setting, Assumption 1 (i) implies identification Condition (6). Compact sets in $L^2[0,1]$ w.r.t. the L^2 norm $\|.\|$ can be obtained by imposing a bound on the Sobolev norm of the functional parameter via $\|\varphi\|_H^2 \leq \bar{B}$. Then, a consistent estimator of a function satisfying this constraint is derived by solving minimization problem (8), where λ_T is interpreted as a Kuhn-Tucker multiplier.

Our approach differs from AC and NP along two directions. On the one hand, NP and AC use finite-dimensional sieve estimators whose sieve dimension grows with sample size (see Chen (2006) for an introduction on sieve estimation in econometrics). By contrast, we

define the TiR estimator and study its asymptotic properties as an estimator on a function space. We introduce a finite dimensional basis of functions only to approximate numerically the estimator (see Section 5). [6]

On the other hand, λ_T is a free regularization parameter for TiR estimators, whereas λ_T is tied down by the slackness condition in NP and AC approach, namely either $\lambda_T = 0$ or $\|\hat{\varphi}\|_H^2 = \bar{B}$, P-a.s.. As a consequence, our approach presents three advantages.

(i) Although, for a given sample size T, selecting different λ_T amounts to select different $\bar{B}$ when the constraint is binding, the asymptotic properties of the TiR estimator and of the estimators with fixed $\bar{B}$ are different. Putting a bound on the Sobolev norm independent of sample size T implies in general the selection of a sub-optimal sequence of regularization parameters λ_T (see Section 4.3). Thus, the estimators with fixed $\bar{B}$ share rates of convergence which are slower than that of the TiR estimator with an optimally selected sequence. [7]

(ii) For the TiR estimator, the tuning parameter λ_T is allowed to depend on sample size T and sample data, whereas the tuning parameter $\bar{B}$ is treated as fixed in NP and AC. Thus, our approach allows for regularized estimators with data-driven selection of the tuning parameter. We prove their consistency in Theorem 1 and Proposition 2 of Section 3.

(iii) Finally, the TiR estimator enjoys computational tractability. This is because, for given λ_T, the TiR estimator is defined by an unconstrained optimization problem, whereas

[6] See NP at p. 1573 for such a suggestion as well as Horowitz and Lee (2006), Gagliardini and Gouriéroux (2006). To make an analogy, an extremum estimator is most of the times computed numerically via an iterative optimization routine. Even if the computed estimator differs from the initially defined extremum estimator, we do not need to link the number of iterations determining the numerical error with sample size.

[7] Letting $\bar{B} = \bar{B}_T$ grow (slowly) with sample size T without introducing a penalty term is not equivalent to our approach, and does not guarantee consistency of the estimator. Indeed, when $\bar{B}_T \to \infty$, the resulting limit parameter set Θ is not compact.

the inequality constraint $\|\varphi\|_H \leq \bar{B}$ has to be accounted for in the minimization defining estimators with given $\bar{B}$. In particular, in the case of linear conditional moment restrictions, the TiR estimator admits a closed form (see Section 5), whereas the computation of the NP and AC estimator requires the use of numerical constrained quadratic optimization routines.

2.4.2 Regularization with L^2 norm

DFR and HH (see also Carrasco, Florens and Renault (2005)) study nonparametric linear IV estimation of the single equation model (3). Their estimators are tightly related to the regularized estimator defined by minimization problem (8) with the L^2 norm $\|\varphi\|$ replacing the Sobolev norm $\|\varphi\|_H$ in the penalty term. The first order condition for such an estimator with $\hat{\Omega}(z) = 1$ (see the remark by DFR at p. 20) corresponds to the equation (4.1) in DFR, and to the estimator defined at p. 4 in HH when $\hat{\Omega}(z) = \hat{f}(z)$, the only difference being the choice of the empirical counterparts of the expectation operators in (1) and (2). [8] Our approach differs from DFR and HH by the norm adopted for penalization. Choosing the Sobolev norm allows us to achieve a faster rate of convergence under conditions detailed in Section 4, and a superior finite-sample performance in the Monte-Carlo experiments of Section 6. Intuitively, incorporating the derivative $\nabla\varphi$ in the penalty helps to control tightly the oscillating components induced by ill-posedness.

[8] In particular, HH do not smooth variable Y w.r.t. instrument Z. As in 2SLS, projecting Y on Z is not necessary. In a functional framework, this possibility applies in the linear IV regression setting only and allows avoiding a differentiability assumption on φ_0. Following DFR, we use high-order derivatives of the joint density of (Y, X, Z) to derive our asymptotic distributional results. This implicitly requires high-order differentiability of φ_0.

3 Consistency of the TiR estimator

First we show consistency of penalized extremum estimators as in Definition 1 but with a general penalty function $G(\varphi)$:

$$\hat{\varphi} = \arg \inf_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T G(\varphi). \tag{9}$$

Then we apply the results with $G(\varphi) = \|\varphi\|_H^2$ to prove the consistency of the TiR estimator. The estimator (9) exists under weak conditions (see Appendix 2.1), while the TiR estimator in the linear case exists because of an explicit derivation (see Section 5).

Theorem 1: *Let* (i) $\bar{\delta}_T := \sup_{\varphi \in \Theta} |Q_T(\varphi) - Q_\infty(\varphi)| \xrightarrow{p} 0$; (ii) $\varphi_0 \in \Theta$;
(iii) *For any* $\varepsilon > 0$, $C_\varepsilon(\lambda) := \inf_{\varphi \in \Theta : \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) + \lambda G(\varphi) - Q_\infty(\varphi_0) - \lambda G(\varphi_0) > 0$, *for any* $\lambda > 0$ *small enough;* (iv) $\exists a \geq 0$ *such that* $\lim_{\lambda \to 0} \lambda^{-a} C_\varepsilon(\lambda) > 0$ *for any* $\varepsilon > 0$;
(v) $\exists b > 0$ *such that* $T^b \bar{\delta}_T = O_p(1)$.

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

$$\left(\lambda_T^{a/b} T\right)^{-1} \to 0, \quad P\text{-a.s.}, \tag{10}$$

the estimator $\hat{\varphi}$ *defined in (9) is consistent, namely* $\|\hat{\varphi} - \varphi_0\| \xrightarrow{p} 0$.

Proof: See Appendix 2.

If $G = 0$, Theorem 1 corresponds to a version of the standard result of consistency for extremum estimators (e.g., White and Wooldridge (1991), Corollary 2.6). [9] In this case,

[9] It is possible to weaken Condition (i) in Theorem 1 requiring uniform convergence of $Q_T(\varphi)$ on a sequence of compact sets (see the proof of Theorem 1).

Condition (iii) is the usual identification Condition (6), and Condition (iv) is satisfied. When Condition (6) does not hold (cf. Section 2.2) identification of φ_0 as an isolated minimum is restored through penalization. Condition (iii) in Theorem 1 is the condition on the penalty function $G(\varphi)$ to overcome ill-posedness and achieve consistency. To interpret Condition (iv), note that in the ill-posed setting we have $C_\varepsilon(\lambda) \to 0$ as $\lambda \to 0$, and the rate of convergence can be seen as a measure for the severity of ill-posedness. Thus, Condition (iv) introduces a lower bound a for this rate of convergence. Condition (10) shows the interplay between a and the rate b of uniform convergence in Condition (v) to guarantee consistency. The regularization parameter λ_T has to converge a.s. to zero at a rate smaller than $T^{-b/a}$. Theorem 1 extends currently available results, since sequence (λ_T) is allowed to be stochastic, possibly data dependent, in a fully general way. Thus, this result applies to estimators with data-driven selection of the tuning parameter. Finally, Theorem 1 is also valid when estimator $\hat{\varphi}$ is defined by $\hat{\varphi} = \arg\inf_{\varphi\in\Theta_T} Q_T(\varphi) + \lambda_T G(\varphi)$ and (Θ_T) is an increasing sequence of subsets of Θ (sieve). Then, we need to define $\bar{\delta}_T := \sup_{\varphi\in\Theta_T} |Q_T(\varphi) - Q_\infty(\varphi)|$, and assume that $\cup_{T=1}^{\infty}\Theta_T$ is dense in Θ and that $b > 0$ in Condition (v) is such that $T^b\bar{\rho}_T = O(1)$ for any $\varepsilon > 0$, where $\bar{\rho}_T := \inf_{\varphi\in\Theta_T:\|\varphi-\varphi_0\|\leq\varepsilon} Q_\infty(\varphi) + |G(\varphi) - G(\varphi_0)|$ (see the Technical Report).

The next proposition provides a sufficient condition for the validity of the key assumptions of Theorem 1, that is identification assumptions (iii) and (iv).

Proposition 2: *Assume that the function G is bounded from below. Furthermore, suppose that, for any $\varepsilon > 0$ and any sequence (φ_n) in Θ such that $\|\varphi_n - \varphi_0\| \geq \varepsilon$ for all $n \in \mathbb{N}$,*

$$Q_\infty(\varphi_n) \to Q_\infty(\varphi_0) \quad as\ n \to \infty \quad \Longrightarrow \quad G(\varphi_n) \to \infty \quad as\ n \to 0. \tag{11}$$

Then, Conditions (iii) and (iv) of Theorem 1 are satisfied with $a = 1$.

Proof: See Appendix 2.

Condition (11) provides a simple intuition on why the penalty function $G(\varphi)$ restores identification. It requires that the sequences (φ_n) in Θ, which minimize $Q_\infty(\varphi)$ without converging to φ_0, make the function $G(\varphi)$ to diverge.

When the penalty function $G(\varphi) = \|\varphi\|_H^2$ is used, Condition (11) in Proposition 2 is satisfied, and consistency of the TiR estimator results from Theorem 1 (see Appendix 2.3).

4 Asymptotic distribution of the TiR estimator

Next theoretical results are derived for a deterministic sequence (λ_T). They are stated in terms of operators A and A^* underlying the linearization in Assumption 2. The proofs are derived for the nonparametric linear IV regression (3) in order to avoid the technical burden induced by the second order term $R(\varphi, z)$. As in AC Assumption 4.1, we assume the following choice of the weighting matrix to simplify the exposition.

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

4.1 Mean Integrated Square Error

Proposition 3: *Let* $\left\{\phi_j : j \in \mathbb{N}\right\}$ *be an orthonormal basis in* $H^2[0,1]$ *of eigenfunctions of operator* A^*A *to eigenvalues* ν_j, *ordered such that* $\nu_1 \geq \nu_2 \geq \cdots > 0$. *Under Assumptions*

1-3, Assumptions B in Appendix 1, and the conditions with $\varepsilon > 0$

$$\frac{1}{Th_T^{d_Z+d/2}} + h_T^m \log T = o\left(\lambda_T b\left(\lambda_T\right)\right), \qquad \frac{1}{Th_T^{d+d_Z}} = O(1), \qquad \frac{1}{Th_T} + h_T^{2m} \log T = O\left(\lambda_T^{2+\varepsilon}\right), \tag{12}$$

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

$$E\left[\left\|\hat{\varphi} - \varphi_0\right\|^2\right] = \frac{1}{T}\sum_{j=1}^{\infty} \frac{\nu_j}{\left(\lambda_T + \nu_j\right)^2} \left\|\phi_j\right\|^2 + b\left(\lambda_T\right)^2 =: V_T(\lambda_T) + b\left(\lambda_T\right)^2 =: M_T(\lambda_T) \tag{13}$$

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

$$b\left(\lambda_T\right) = \left\|\left(\lambda_T + A^*A\right)^{-1} A^*A\varphi_0 - \varphi_0\right\|, \tag{14}$$

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

Proof: See Appendix 3.

The asymptotic expansion of the MISE consists of two components.

(i) The bias function $b\left(\lambda_T\right)$ is the L^2 norm of $\left(\lambda_T + A^*A\right)^{-1} A^*A\varphi_0 - \varphi_0 =: \varphi^* - \varphi_0$. To interpret φ^*, recall the quadratic approximation $\langle \Delta\varphi, A^*A\Delta\varphi\rangle_H$ of the limit criterion. Then, function φ^* minimizes $\langle \Delta\varphi, A^*A\Delta\varphi\rangle_H + \lambda_T \left\|\varphi\right\|_H^2$ w.r.t. $\varphi \in \Theta$. Thus, $b\left(\lambda_T\right)$ is the asymptotic bias arising from introducing the penalty $\lambda_T \left\|\varphi\right\|_H^2$ 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 $b\left(\lambda\right)$ is increasing w.r.t. λ and such that $b\left(\lambda\right) \to 0$ as $\lambda \to 0$.

(ii) The variance term $V_T(\lambda_T)$ involves a weighted sum of the regularized inverse eigenvalues $\nu_j / \left(\lambda_T + \nu_j\right)^2$ of operator A^*A, with weights $\left\|\phi_j\right\|^2$. [10] To have an interpretation, note

[10] Since $\nu_j/(\lambda_T + \nu_j)^2 \leq \nu_j$, the infinite sum converges under Assumption B.10 (i) in Appendix 1.

that the inverse of operator A^*A corresponds to the standard asymptotic variance matrix $\left(J_0'V_0^{-1}J_0\right)^{-1}$ of the efficient GMM in the parametric setting, where $J_0 = E\left[\partial g/\partial\theta'\right]$ and $V_0 = V[g]$. In the ill-posed nonparametric setting, the inverse of operator A^*A is unbounded, and its eigenvalues $1/\nu_j \to \infty$ 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 $\nu_j/\left(\lambda_T + \nu_j\right)^2$.

The variance term $V_T(\lambda_T)$ is a decreasing function of λ_T. To study its behavior when $\lambda_T \to 0$, we introduce the next assumption.

Assumption 4: *The eigenfunctions ϕ_j and the eigenvalues ν_j of A^*A satisfy* $\sum_{j=1}^{\infty} \nu_j^{-1} \left\|\phi_j\right\|^2 = \infty$.

Under Assumption 4, the series $n_T := \sum_{j=1}^{\infty} \left\|\phi_j\right\|^2 \left[\nu_j/\left(\lambda_T + \nu_j\right)^2\right]$ diverges as $\lambda_T \to 0$. When $n_T \to \infty$ such that $n_T/T \to 0$, the variance term converges to zero. Assumption 4 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 $\nu_j \to 0$.

The asymptotic expansion of the MISE given in Proposition 3 does not involve the bandwidth h_T, as long as Conditions (12) 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 (35) 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 $h_T^m = o\left(\lambda_T b\left(\lambda_T\right)\right)$, which follows from (12), 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 (12) are used to control higher order terms in the MISE (see Lemma A.5).

Finally, it is also possible to derive a similar asymptotic expansion of the MISE for the estimator $\tilde{\varphi}$ regularized by the L^2 norm. This characterization is new in the nonparametric instrumental regression setting: [11]

$$E\left[\left\|\tilde{\varphi}-\varphi_0\right\|^2\right] = \frac{1}{T}\sum_{j=1}^{\infty}\frac{\tilde{\nu}_j}{\left(\lambda_T+\tilde{\nu}_j\right)^2} + \tilde{b}\left(\lambda_T\right)^2, \tag{15}$$

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 $\langle .,. \rangle$ and $\langle .,. \rangle_{L^2_{\Omega_0}(F_Z)}$, and $\tilde{b}\left(\lambda_T\right) = \left\| \left(\lambda_T + \tilde{A}A\right)^{-1} \tilde{A}A\varphi_0 - \varphi_0 \right\|$. [12]

Let us now come back to the MISE $M_T(\lambda_T)$ of the TiR estimator in 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 $\lambda_T^* = \arg\min_{\lambda>0} M_T\left(\lambda\right)$, and the corresponding optimal MISE by $M_T^* := M_T\left(\lambda_T^*\right)$. Their rate of convergence depends on the decay behavior of the eigenvalues ν_j and of the norms $\left\|\phi_j\right\|$ of the eigenfunctions, as well as on the bias function $b\left(\lambda\right)$ close to

[11] A similar formula has been derived by Carrasco and Florens (2005) for the density deconvolution problem.

[12] The adjoint defined w.r.t. the L^2 scalar product is denoted by a superscripted $*$ in DFR or Carrasco, Florens, and Renault (2005). We stress that in our paper the adjoint A^* is defined w.r.t. a Sobolev scalar product. Besides 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.

$\lambda = 0$. In the next section, we characterize these rates in a broad set of examples.

4.2 Examples of optimal rates of convergence

The eigenvalues ν_j and the L^2-norms $\left\|\phi_j\right\|$ of the eigenfunctions can feature different types of decay as $j \to \infty$. A geometric decay for the eigenvalues is associated with a faster convergence of the spectrum to zero, and to a more serious problem of ill-posedness. We focus on this case. Results for the hyperbolic decay are summarized in Table 1 below.

Assumption 5: *The eigenvalues ν_j and the norms $\left\|\phi_j\right\|$ of the eigenfunctions of operator A^*A are such that, for $j = 1, 2, \cdots$, and some positive constants C_1, C_2,*

(i) $\nu_j = C_1 \exp(-\alpha j)$, $\alpha > 0$, (ii) $\left\|\phi_j\right\|^2 = C_2 j^{-\beta}$, $\beta > 0$.

Assumption 5 (i) is satisfied for a large number of models, including the two cases in our Monte-Carlo analysis below. In general, under appropriate regularity conditions, compact integral operators with smooth kernel induce eigenvalues with decay of (at least) exponential type (see Theorem 15.20 in Kress (1999)). [13] We verify numerically in Section 6 that Assumption 5 (ii) is satisfied in our two Monte-Carlo designs. For this reason and the sake of space we do not develop the example of geometric decay for $\left\|\phi_j\right\|^2$. We are not aware of any theoretical result implying that $\left\|\phi_j\right\|^2$ has an hyperbolic, or geometric, decay.

We further assume that the bias function features a power-law behavior close to $\lambda = 0$.

Assumption 6: *The bias function is such that $b(\lambda) = C_3\lambda^\delta, \delta > 0$, for λ close to 0, where*

[13] In the case of nonparametric linear IV estimation and regularization with L^2 norm, the eigenvalues correspond to the nonlinear canonical correlations of (X, Z). When X and Z are monotone transformations of variables which are jointly normally distributed with correlation parameter ρ, the canonical correlations of (X, Z) are ρ^j, $j \in \mathbb{N}$ (see, e.g., DFR). Thus the eigenvalues exhibit exponential decay.

C_3 is a positive constant.

From (14) we get $b(\lambda)^2 = \lambda^2 \sum_{j,l=1}^{\infty} \frac{\epsilon_j}{\lambda+\nu_j}\frac{\epsilon_l}{\lambda+\nu_l}\langle\phi_j,\phi_l\rangle$, where $\epsilon_j := \langle\varphi_0,\phi_j\rangle_H$, $j \in \mathbb{N}$. Therefore the coefficient δ depends on the decay behavior of eigenvalues ν_j, Fourier coefficients ϵ_j, and L^2-scalar products $\langle\phi_j,\phi_l\rangle$ as $j,l \to \infty$. In particular, the decay of ϵ_j as $j \to \infty$ characterizes the influence of the smoothness properties of function φ_0 on the bias $b(\lambda)$ and on the rate of convergence of the TiR estimator. Given Assumption 5, the decay rate of the Fourier coefficients must be sufficiently fast for Assumption 6 to hold. Besides, the above expression of $b(\lambda)$ implies $\delta \leq 1$.

Proposition 4: *Under the Assumptions of Proposition 3, Assumptions 5 and 6, for some positive constants c_1, c_2 and c^*, we have*

(i) *The MISE is $M_T(\lambda) = c_1\frac{1}{T}\frac{1+c(\lambda)}{\lambda[\log(1/\lambda)]^\beta} + c_2\lambda^{2\delta}$, up to terms which are negligible when $\lambda \to 0$ and $T \to \infty$, where function $c(\lambda)$ is such that $1+c(\lambda)$ is bounded and bounded away from zero.*

(ii) *The optimal sequence of regularization parameters is*

$$\log\lambda_T^* = \log c^* - \frac{1}{1+2\delta}\log T, \qquad T \in \mathbb{N}, \tag{16}$$

up to a term which is negligible w.r.t. the RHS.

(iii) *The optimal MISE is $M_T^* = \overline{c}_T T^{-\frac{2\delta}{1+2\delta}}(\log T)^{-\frac{2\delta\beta}{1+2\delta}}$, up to a term which is negligible w.r.t. the RHS, where sequence $\overline{c}_T$ is bounded and bounded away from zero.*

Proof: See Appendix 4.

The log of the optimal regularization parameter is linear in the log sample size. 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 (12) forces γ to be smaller than $1/2$. This condition is also used in HH and DFR. [14] The optimal MISE converges to zero as a power of T and of $\log T$. The negative exponent of the dominant term T is $2\delta/(1+2\delta)$. This rate of convergence is smaller than $2/3$ and is increasing w.r.t. δ. The decay rate α does not affect neither the rate of convergence of the optimal regularization sequence (up to order $o(\log T)$), nor that of the MISE. The decay rate β affects the exponent of the $\log T$ term in the MISE only. Finally, under Assumptions 5 and 6, the bandwidth conditions (12) are fulfilled for the optimal sequence of regularization parameters (16) if $h_T = CT^{-\eta}$, with $\dfrac{\delta_U}{1+2\delta} > \eta > \dfrac{1}{m}\dfrac{1+\delta}{1+2\delta}$, where $\delta_U := \min\left\{(d_Z + d/2)^{-1}\delta, 2\delta - 1\right\}$. An admissible η exists if and only if $m > \dfrac{1+\delta}{\delta_U}$. This inequality illustrates the intertwining in (12) between the degree m of differentiability, the dimensions d, d_Z, and the decay rate δ.

To conclude this section, we discuss the optimal rate of convergence of the MISE when the eigenvalues have hyperbolic decay, that is $\nu_j = Cj^{-\alpha}$, $\alpha > 0$, or when regularization with L^2 norm is adopted. The results summarized in Table 1 are found using Formula (15) and arguments similar to the proof of Proposition 4. In Table 1, parameter β is defined as in Assumption 5 (ii) for the TiR estimator. Parameters α and $\tilde{\alpha}$ denote the hyperbolic decay

[14] The sufficient condition $\dfrac{1}{Th_T} + h_T^{2m}\log T = O\left(\lambda_T^{2+\varepsilon}\right)$, $\varepsilon > 0$, in (12) is used to prove that some expectation terms are bounded, see Lemma A.5 (ii) in Appendix 3. Although a weaker condition could be found, we do not pursue this strategy. This would unnecessarily complicate the proofs. To assess the importance of this technical restriction, we consider two designs in our Monte-Carlo experiments in Section 6. In Case 1 the condition $\gamma < 1/2$ is not satisfied, and in Case 2 it is. In both settings we find that the asymptotic expansion in Proposition 3 provides a very good approximation of the MISE in finite samples.

rates of the eigenvalues of operator A^*A for the TiR estimator, and of operator $\tilde{A}A$ for L^2 regularization, respectively. We assume $\alpha, \tilde{\alpha} > 1$, and $\alpha > \beta - 1$ to satisfy Assumption 4. Finally, parameters δ and $\tilde{\delta}$ are the power-law coefficients of the bias function $b(\lambda)$ and $\tilde{b}(\lambda)$ for $\lambda \to 0$ as in Assumption 6, where $b(\lambda)$ is defined in (14) for the TiR estimator, and $\tilde{b}(\lambda)$ in (15) for L^2 regularization, respectively. With a slight abuse of notation we use the same greek letters α, $\tilde{\alpha}$, β, δ and $\tilde{\delta}$ for the decay rates in the geometric and hyperbolic cases.

	TiR estimator	L^2 regularization
geometric spectrum	$T^{-\frac{2\delta}{1+2\delta}} (\log T)^{-\frac{2\delta\beta}{1+2\delta}}$	$T^{-\frac{2\tilde{\delta}}{1+2\tilde{\delta}}}$
hyperbolic spectrum	$T^{-\frac{2\delta}{1+2\delta+(1-\beta)/\alpha}}$	$T^{-\frac{2\tilde{\delta}}{1+2\tilde{\delta}+1/\tilde{\alpha}}}$

Table 1: Optimal rate of convergence of the MISE. The decay factors are α and $\tilde{\alpha}$ for the eigenvalues, δ and $\tilde{\delta}$ for the bias, and β for the squared norm of the eigenfunctions.

The rate of convergence of the TiR estimator under an hyperbolic spectrum includes an additional term $(1-\beta)/\alpha$ in the denominator. The rate of convergence with geometric spectrum is recovered letting $\alpha \to \infty$ (up to the $\log T$ term). The rate of convergence with L^2 regularization coincides with that of the TiR estimator with $\beta = 0$, and coefficients α, δ corresponding to operator $\tilde{A}A$ instead of A^*A. When both operators share a geometric spectrum, the TiR estimator enjoys a faster rate of convergence than the regularized estimator with L^2 norm if $\delta \geq \tilde{\delta}$, that is if the bias function of the TiR estimator is more convex.

Conditions under which the inclusion of higher order derivatives of function φ in the penalty improves or not on the optimal rate of convergence are of interest, but we leave this for future research. Finally, we recover the formula derived by HH in their Theorem 4.1 under an hyperbolic spectrum and L^2 regularization. [15]

4.3 Suboptimality of bounding the Sobolev norm

The approach of NP and AC forces compactness by a direct bounding of the Sobolev norm. Unfortunately this leads to a suboptimal rate of convergence of the regularized estimator.

Proposition 5: *Let $\bar{B} \geq \|\varphi_0\|_H^2$ be a fixed constant. Let $\breve{\varphi}$ be the estimator defined by $\breve{\varphi} = \arg\inf_{\varphi\in\Theta} Q_T(\varphi)$ s.t. $\|\varphi\|_H^2 \leq \bar{B}$, and denote by $\breve{\lambda}_T$ the associated stochastic Kuhn-Tucker multiplier. Suppose that:*

(i) Function $b(\lambda)$ in (14) is non-decreasing, for λ small enough;

(ii) The variance term $V_T(\lambda)$ and the squared bias $b(\lambda)^2$ of the TiR estimator in (13) are such that for any deterministic sequence (l_T) : $l_T = o(\lambda_T^*) \implies V_T(l_T)/\, M_T^* \to \infty$ and $\lambda_T^* = o(l_T) \implies b(l_T)^2/\, M_T^* \to \infty$, where λ_T^* is the optimal deterministic regularization sequence for the TiR estimator and $M_T^* = M_T(\lambda_T^*)$;

(iii) $P\left[\lambda_T^l \leq \breve{\lambda}_T \leq \lambda_T^u\right] \to 1$, for two deterministic sequences λ_T^l, λ_T^u such that either $\lambda_T^u = o(\lambda_T^*)$ or $\lambda_T^* = o(\lambda_T^l)$.

Further, let the regularity conditions of the Lemma B.13 in the Technical Report be satisfied. Then: $E\left[\|\breve{\varphi} - \varphi_0\|^2\right]/M_T^* \to \infty$.

[15] To see this, note that their Assumption A.3 implies hyperbolic decay of the eigenvalues and is consistent with $\tilde{\delta} = (2\beta_{HH} - 1)/(2\tilde{\alpha})$, where β_{HH} is the β coefficient of HH (see also the remark at p. 21 in DFR).

This proposition is proved in the Technical Report. It states that, whenever the stochastic regularization parameter $\check{\lambda}_T$ implied by the bound $\bar{B}$ does not exhibit the same rate of convergence as the optimal deterministic TiR sequence λ_T^*, the regularized estimator with fixed bound on the Sobolev norm has a slower rate of convergence than the optimal TiR estimator. Intuitively, imposing a fixed bound offers no guarantee to select an optimal rate for $\check{\lambda}_T$. Conditions (i) and (ii) of Proposition 5 are satisfied under Assumptions 5 and 6 (geometric spectrum; see also Proposition 4 (i)). In the Technical Report, we prove that Condition (iii) of Proposition 5 is also satisfied in such a setting.

4.4 Mean Squared Error and pointwise 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 (13).

Proposition 6: *Under the assumptions of Proposition 3, the MSE of the TiR estimator $\hat{\varphi}$ with deterministic sequence (λ_T) is given by*

$$E\left[(\hat{\varphi}(x) - \varphi_0(x))^2\right] = \frac{1}{T}\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2}\phi_j^2(x) + \mathcal{B}_T(x)^2 =: \frac{1}{T}\sigma_T^2(x) + \mathcal{B}_T(x)^2, \qquad (17)$$

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

$$\mathcal{B}_T(x) = (\lambda_T + A^*A)^{-1} A^*A\varphi_0(x) - \varphi_0(x). \qquad (18)$$

An analysis similar to Sections 4.1 and 4.2 shows that the rate of convergence of the MSE depends on the decay behavior of eigenvalues ν_j and eigenfunctions $\phi_j(x)$ in a given

point $x \in \mathcal{X}$. 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 (even for a sequence of regularization parameters making the bias asymptotically negligible as in Horowitz (2005)).

Finally, under a regularization with an L^2 norm, we get

$$E\left[\left(\tilde{\varphi}(x) - \varphi_0(x)\right)^2\right] = \frac{1}{T}\sum_{j=1}^{\infty} \frac{\tilde{\nu}_j}{\left(\lambda_T + \tilde{\nu}_j\right)^2} \tilde{\phi}_j^2(x) + \tilde{\mathcal{B}}_T(x)^2, \tag{19}$$

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

To conclude we state pointwise asymptotic normality of the TiR estimator.

Proposition 7: *Suppose Assumptions 1-3 and B hold,* $\frac{1}{Th_T^{d_Z+d/2}} + h_T^m \log T = O\left(\frac{b(\lambda_T)}{\sqrt{Th_T}}\right)$, $\left(Th_T^{d+d_Z}\right)^{-1} = O(1)$, $(Th_T)^{-1} + h_T^{2m}\log T = O(\lambda_T^{2+\varepsilon})$, $\varepsilon > 0$, $\frac{M_T(\lambda_T)}{\sigma_T^2(x)/T} = o\left(Th_T\lambda_T^2\right)$. *Further, suppose that for a strictly positive sequence* (a_j) *such that* $\sum_{j=1}^{\infty} 1/a_j < \infty$, *we have*

$$\frac{\sum_{j=1}^{\infty} \frac{\nu_j}{\left(\lambda_T + \nu_j\right)^2} \phi_j^2(x) \left\|g_j\right\|_3^2 a_j}{\sum_{j=1}^{\infty} \frac{\nu_j}{\left(\lambda_T + \nu_j\right)^2} \phi_j^2(x)} = o\left(T^{1/3}\right), \tag{20}$$

where $\left\|g_j\right\|_3 := E\left[g_j(Y,X,Z)^3\right]^{1/3}$, $g_j(y,x,z) := \left(A\phi_j\right)(z)' \Omega_0(z) g\left(y, \varphi_0(x)\right)/\sqrt{\nu_j}$. *Then*

the TiR estimator is asymptotically normal: $\sqrt{T/\sigma_T^2(x)}\left(\hat{\varphi}(x)-\varphi_0(x)-\mathcal{B}_T(x)\right)\xrightarrow{d}N(0,1)$.

Proof: See Appendix 5.

Condition (20) is used to apply a Lyapunov CLT. In general, it is satisfied when λ_T converges to zero not too fast. Under Assumption 5 (i) of geometric spectrum for the eigenvalues ν_j, and an assumption of hyperbolic decay for the eigenvectors $\phi_j^2(x)$ and $\|g_j\|_3$, Lemma A.6 in Appendix 4 implies that (20) is satisfied whenever $\lambda_T \geq cT^{-\gamma}$ for some $c,\gamma>0$. Finally, for an asymptotically negligible bias, a natural candidate for a $N(0,1)$ pivotal statistic is $\sqrt{T/\hat{\sigma}_T^2(x)}\left(\hat{\varphi}(x)-\varphi_0(x)\right)$, 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). [16]

5 The TiR estimator for linear moment restrictions

In this section we develop nonparametric IV estimation of a single equation model as in (3). Then, the estimated moment function is $\hat{m}(\varphi,z)=\int\varphi(x)\,\hat{f}(w|z)\,dw-\int y\hat{f}(w|z)\,dw=:\left(\hat{A}\varphi\right)(z)-\hat{r}(z)$. The objective function in (8) can be rewritten as (see Appendix 3.1)

$$Q_T(\varphi)+\lambda_T\|\varphi\|_H^2=\langle\varphi,\hat{A}^*\hat{A}\varphi\rangle_H-2\langle\varphi,\hat{A}^*\hat{r}\rangle_H+\lambda_T\langle\varphi,\varphi\rangle_H,\quad\varphi\in H^2[0,1], \tag{21}$$

up to a term independent of φ, where $\hat{A}^*$ denotes the linear operator defined by

$$\langle\varphi,\hat{A}^*\psi\rangle_H=\frac{1}{T}\sum_{t=1}^{T}\left(\hat{A}\varphi\right)(Z_t)\,\hat{\Omega}(Z_t)\,\psi(Z_t),\qquad\varphi\in H^2[0,1],\quad\psi\text{ measurable.} \tag{22}$$

[16] Since $\sigma_T(x)$ depends on T and diverges, the usual argument using Slutsky Theorem does not apply. Instead the condition $\left[\hat{\sigma}_T(x)-\sigma_T(x)\right]/\hat{\sigma}_T(x)\xrightarrow{p}0$ 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).

Under the regularity conditions in Appendix 1, Criterion (21) admits a global minimum $\hat{\varphi}$ on $H^2[0,1]$, which solves the first order condition

$$\left(\lambda_T + \hat{A}^*\hat{A}\right)\varphi = \hat{A}^*\hat{r}. \tag{23}$$

This is a Fredholm integral equation of Type II. [17] The transformation of the ill-posed problem (1) in the well-posed estimating equation (23) is induced by the penalty term involving the Sobolev norm. The TiR estimator is the explicit solution of Equation (23):

$$\hat{\varphi} = \left(\lambda_T + \hat{A}^*\hat{A}\right)^{-1}\hat{A}^*\hat{r}. \tag{24}$$

To compute numerically the estimator we solve Equation (23) on the subspace spanned by a finite-dimensional basis of functions $\{P_j : j = 1, ..., k\}$ in $H^2[0,1]$ and use the numerical approximation

$$\varphi \simeq \sum_{j=1}^{k} \theta_j P_j =: \theta' P, \quad \theta \in \mathbb{R}^k. \tag{25}$$

From (22) the $k \times k$ matrix corresponding to operator $\hat{A}^*\hat{A}$ on this subspace is given by $\langle P_i, \hat{A}^*\hat{A}P_j\rangle_H = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{A}P_i\right)(Z_t)\,\hat{\Omega}(Z_t)\left(\hat{A}P_j\right)(Z_t) = \frac{1}{T}\left(\widehat{P}'\widehat{P}\right)_{i,j}, \quad i,j = 1, ..., k$, where $\widehat{P}$ is the $T \times k$ matrix with rows $\widehat{P}(Z_t)' = \hat{\Omega}(Z_t)^{1/2}\int P(x)'\hat{f}(w|Z_t)\,dw$, $t = 1, ..., T$. Matrix $\widehat{P}$ is the matrix of the weighted "fitted values" in the regression of $P(X)$ on Z at the sample points. Then, Equation (23) reduces to a matrix equation $\left(\lambda_T D + \frac{1}{T}\widehat{P}'\widehat{P}\right)\theta = \frac{1}{T}\widehat{P}'\widehat{R}$, where $\widehat{R} = \left(\hat{\Omega}(Z_1)^{1/2}\hat{r}(Z_1), ..., \hat{\Omega}(Z_T)^{1/2}\hat{r}(Z_T)\right)'$, and D is the $k \times k$ matrix of Sobolev scalar products $D_{i,j} = \langle P_i, P_j\rangle_H, \quad i,j = 1, ..., k$. The solution is given by $\widehat{\theta} =$

[17] See e.g. Linton and Mammen (2005), (2006), Gagliardini and Gouriéroux (2006), and the survey by Carrasco, Florens and Renault (2005) for other examples of estimation problems leading to Type II equations.

$\left(\lambda_T D + \frac{1}{T}\widehat{P}'\widehat{P}\right)^{-1} \frac{1}{T}\widehat{P}'\widehat{R}$, which yields the approximation of the TiR estimator $\hat{\varphi} \simeq \hat{\theta}' P$. [18] It only asks for inverting a $k \times k$ matrix, which is expected to be of small dimension in most economic applications.

Estimator $\hat{\theta}$ is a 2SLS estimator with optimal instruments and a ridge correction term. It is also obtained if we replace (25) in Criterion (21) and minimize w.r.t. θ. This route is followed by NP, AC, and Blundell, Chen and Kristensen (2004), who use sieve estimators and let $k = k_T \to \infty$ with T. In our setting the introduction of a series of basis functions as in (25) is simply a method to compute numerically the original TiR estimator $\hat{\varphi}$ in (24). 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, and Formula (24) is akin to the estimator of DFR and HH. The approximation with a finite-dimensional basis

[18] Note that 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..

of functions gives an estimator $\hat{\theta}$ similar to above, with matrix D replaced by matrix B of L^2 scalar products $B_{i,j} = \langle P_i, P_j \rangle, \quad i,j = 1, ..., k.$ [19]

6 A Monte-Carlo study

6.1 Data generating process

Following NP we draw the errors U and V and the instruments Z as

$$\begin{pmatrix} U \\ V \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right), \qquad \rho \in \{0, 0.5\},$$

and build $X^* = Z + V$. Then we map X^* into a variable $X = \Phi(X^*)$, which lives in $[0,1]$. The function Φ denotes the cdf of a standard Gaussian variable, and is assumed to be known. To generate Y, we restrict ourselves to the linear case since a simulation analysis of a nonlinear case would be very time consuming. We examine two designs. **Case 1** is $Y = \mathcal{B}_{a,b}(X) + U$, where $\mathcal{B}_{a,b}$ denotes the cdf of a Beta(a,b) variable. The parameters of the beta distribution are chosen equal to $a = 2$ and $b = 5$. **Case 2** is $Y = \sin(\pi X) + U$. When the correlation ρ between U and V is 50% there is endogeneity in both cases. When $\rho = 0$ there is no need to correct for the endogeneity bias. The moment condition is $E[Y - \varphi_0(X) \mid Z] = 0$, where the functional parameter is $\varphi_0(x) = \mathcal{B}_{a,b}(x)$ in Case 1, and $\varphi_0(x) = \sin(\pi x)$ in Case 2, $x \in [0,1]$. The chosen functions resemble possible shapes of Engel curves, either monotone increasing or concave.

[19] 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 $\varphi(x)$.

6.2 Estimation procedure

Since we face an unknown function φ_0 on $[0,1]$, we use a series approximation based on standardized shifted Chebyshev polynomials of the first kind (see Section 22 of Abramowitz and Stegun (1970) for their mathematical properties). We take orders 0 to 5 which yields six coefficients ($k = 6$) to be estimated in the approximation $\varphi(x) \simeq \sum_{j=0}^{5} \theta_j P_j(x)$, where $P_0(x) = T_0^*(x)/\sqrt{\pi}$, $P_j(x) = T_j^*(x)/\sqrt{\pi/2}$, $j \neq 0$. The shifted Chebyshev polynomials of the first kind are $T_0^*(x) = 1$, $T_1^*(x) = -1 + 2x$, $T_2^*(x) = 1 - 8x + 8x^2$, $T_3^*(x) = -1 + 18x - 48x^2 + 32x^3$, $T_4^*(x) = 1 - 32x + 160x^2 - 256x^3 + 128x^4$, $T_5^*(x) = -1 + 50x - 400x^2 + 1120x^3 - 1280x^4 + 512x^5$. The squared Sobolev norm is approximated by $\|\varphi\|_H^2 = \int_0^1 \varphi^2 + \int_0^1 (\nabla\varphi)^2 \simeq \sum_{i=0}^{5}\sum_{j=0}^{5} \theta_i\theta_j \int_0^1 (P_iP_j + \nabla P_i \nabla P_j)$. The coefficients in the quadratic form $\theta' D\theta$ are explicitly computed with a symbolic calculus package. The squared L_2 norm $\|\varphi\|^2$ is approximated similarly by $\theta' B\theta$. The two matrices take the form:

$$D = \begin{pmatrix} \frac{1}{\pi} & 0 & \frac{-\sqrt{2}}{3\pi} & 0 & \frac{-\sqrt{2}}{15\pi} & 0 \\ \vdots & \frac{26}{3\pi} & 0 & \frac{38}{5\pi} & 0 & \frac{166}{21\pi} \\ & & \frac{218}{5\pi} & 0 & \frac{1182}{35\pi} & 0 \\ & & & \frac{3898}{35\pi} & 0 & \frac{5090}{63\pi} \\ \vdots & & & & \frac{67894}{315\pi} & 0 \\ & \cdots & & & \cdots & \frac{82802}{231\pi} \end{pmatrix}, \qquad B = \begin{pmatrix} \frac{1}{\pi} & 0 & \frac{-\sqrt{2}}{3\pi} & 0 & \frac{-\sqrt{2}}{15\pi} & 0 \\ \vdots & \frac{2}{3\pi} & 0 & \frac{-2}{5\pi} & 0 & \frac{-2}{21\pi} \\ & & \frac{14}{15\pi} & 0 & \frac{-38}{105\pi} & 0 \\ & & & \frac{34}{35\pi} & 0 & \frac{-22}{63\pi} \\ \vdots & & & & \frac{62}{63\pi} & 0 \\ & \cdots & & & \cdots & \frac{98}{99\pi} \end{pmatrix}.$$

Such simple and exact forms ease implementation [20] , and improve on speed. The convexity in θ (quadratic penalty) helps numerical stability of the estimation procedure.

[20] The Gauss programs developed for this section and the empirical application are available on request from the authors.

The kernel estimator $\hat{m}(\varphi, z)$ of the conditional moment is approximated through $\theta' \hat{P}(z) - \hat{r}(z)$ where $\hat{P}(z) \simeq \sum_{t=1}^{T} P(X_t) K\left(\frac{Z_t - z}{h_T}\right) / \sum_{t=1}^{T} K\left(\frac{Z_t - z}{h_T}\right)$, $\hat{r}(z) \simeq \sum_{t=1}^{T} Y_t K\left(\frac{Z_t - z}{h_T}\right) / \sum_{t=1}^{T} K\left(\frac{Z_t - z}{h_T}\right)$, where K is the Gaussian kernel. This kernel estimator is asymptotically equivalent to the one described in the lines above. We prefer it because of its numerical tractability: we avoid bivariate numerical integration and the choice of two additional bandwidths. The bandwidth is selected via the standard rule of thumb $h = 1.06\hat{\sigma}_Z T^{-1/5}$ (Silverman (1986)), where $\hat{\sigma}_Z$ is the empirical standard deviation of observed Z_t. [21] Here the weighting function $\Omega_0(z)$ is taken equal to unity, satisfying Assumption 3, and assumed to be known.

6.3 Simulation results

The sample size is initially fixed at $T = 400$. Estimator performance is measured in terms of the MISE and the Integrated Squared Bias (ISB) based on averages over 1000 repetitions. We use a Gauss-Legendre quadrature with 40 knots to compute the integrals.

Figures 1 to 4 concern Case 1 while Figures 5 to 8 concern Case 2. The left panel plots the MISE on a grid of lambda, the central panel the ISB, and the right panel the mean estimated functions and the true function on the unit interval. Mean estimated functions correspond to averages obtained either from regularized estimates with a lambda achieving the lowest MISE or from OLS estimates (standard sieve estimators with six polynomials). The regularization schemes use the Sobolev norm, corresponding to the TiR estimator (odd

[21] This choice is motivated by ease of implementation. Moderate deviations from this simple rule do not seem to affect estimation results significantly.

numbering of the figures), and the L_2 norm (even numbering of the figures). We consider designs with endogeneity ($\rho = 0.5$) in Figures 1, 2, 5, 6, and without endogeneity ($\rho = 0$) in Figures 3, 4, 7, 8.

Several remarks can be made. First, the bias of the OLS estimator is large under endogeneity. Second, the MISE under a Sobolev penalization is more convex in lambda than under an L_2 penalization, and is much smaller. Hence the Sobolev norm should be strongly favored in order to recover the shape of the true functions in our two designs. Third, the fit obtained by the OLS estimator is almost perfect when endogeneity is absent. Using six polynomials is enough here to deliver a very good approximation of the true functions. Fourth, 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.

We have also examined sample sizes $T = 100$ and $T = 1000$, as well as approximations based on polynomials with orders up to 10 and 15. The above conclusions remain qualitatively unaffected. This suggests that as soon as the order of the polynomials is sufficiently large to deliver a good numerical approximation of the underlying function, it is not necessary to link it with sample size (cf. Section 5). For example Figures 9 and 10 are the analogues of Figures 1 and 5 with $T = 1000$. The bias term is almost identical, while the variance term decreases by a factor about $2.5 = 1000/400$ as predicted by Proposition 3.

In Figure 11 we display the six eigenvalues of operator A^*A and the L^2-norms of the corresponding eigenfunctions when the same approximation basis of six polynomials is used. These true quantities have been computed by Monte-Carlo integration. The eigenvalues ν_j

feature a geometric decay w.r.t. the order j, whereas the decay of $\left\|\phi_j\right\|^2$ is of an hyperbolic type. This is conform to Assumption 5 and the analysis conducted in Proposition 4. A linear fit of the plotted points gives decay values 2.254, 2.911 for α, β.

Figure 12 is dedicated to check whether the line $\log \lambda_T^* = \log c - \gamma \log T$, induced by Proposition 4 (ii), holds in small samples. For $\rho = 0.5$ both panels exhibit a linear relationship between the logarithm of the regularization parameter minimizing the average MISE on the 1000 Monte-Carlo simulations and the logarithm of sample size ranging from $T = 50$ to $T = 1000$. The OLS estimation of this linear relationship from the plotted pairs delivers .226, .752 in Case 1, and .012, .428 in Case 2, for c, γ. Both estimated slope coefficients are smaller than 1, and qualitatively consistent with the implications of Proposition 4. Indeed, from Figures 9 and 10 the ISB curve appears to be more convex in Case 2 than in Case 1. This points to a larger δ parameter, and thus to a smaller slope coefficient $\gamma = 1/(1+2\delta)$ in Case 2. Inverting this relationship yields .165, .668 in Case 1, 2, for δ.

By a similar argument, Proposition 4 and Table 1 support the better performance of the TiR estimator compared to the L^2-regularized estimator. Indeed, by comparing the ISB curves of the two estimators in Case 1 (Figures 1 and 2) and in Case 2 (Figures 5 and 6), it appears that the TiR estimator induces a more convex ISB curve ($\delta > \tilde{\delta}$).

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 (13). [22] Unreported results based on Monte-Carlo integration show that the asymptotic MISE, ISB

[22] A similar approach has been successfully applied in Carrasco and Florens (2005) for density deconvolution.

and variance are close to the ones exhibited in Figures 9 and 10. The asymptotic optimal lambda is equal to .0018, .0009 in Case 1, 2. These are of the same magnitude as .0013, .0007 in Figures 9, 10. We have checked that the linear relationship of Figure 12 holds true when deduced from optimizing the asymptotic MISE. The OLS estimation delivers .418, .795, .129 for c, γ, δ in Case 1, and .037, .546, .418, in Case 2.

The data driven estimation algorithm based on (13) goes as follows:

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 $\hat{w}_j' D \hat{w}_j = 1$, $j = 1, ..., k$.

(ii) Get a first-step TiR estimator $\bar{\theta}$ using a pilot regularization parameter $\bar{\lambda}$.

(iii) Estimate the MISE:

$$\bar{M}(\lambda) = \frac{1}{T}\sum_{j=1}^{k} \frac{\hat{\nu}_j}{(\lambda + \hat{\nu}_j)^2} \hat{w}_j' B \hat{w}_j$$
$$+\bar{\theta}' \left[\frac{1}{T}\widehat{P}'\widehat{P}\left(\lambda D + \frac{1}{T}\widehat{P}'\widehat{P}\right)^{-1} - I\right] B \left[\frac{1}{T}\widehat{P}'\widehat{P}\left(\lambda D + \frac{1}{T}\widehat{P}'\widehat{P}\right)^{-1} - I\right] \bar{\theta},$$

and minimize it w.r.t. λ to get the optimal regularization parameter $\hat{\lambda}$.

(iv) Compute the second-step TiR estimator $\widehat{\theta}$ 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 of Assumptions 5 and 6, 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 $(\log \hat{w}_j' B \hat{w}_j, \log j)$, $j = 1, ..., k$.

After getting lambdas minimizing the second-step estimated MISE on a grid of sample sizes we can estimate γ by regressing the logarithm of lambda on the logarithm of sample size.

We use $\bar{\lambda} \in \{.0005, .0001\}$ as pilot regularization parameter for $T = 1000$ and $\rho = .5$. In Case 1, the average (quartiles) of the selected lambda over 1000 simulations is equal to .0028 (.0014, .0020, .0033) when $\bar{\lambda} = .0005$, and .0027 (.0007, .0014, .0029) when $\bar{\lambda} = .0001$. In Case 2, results are .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, especially in Case 1, 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 MISE based on averages over the 1000 simulations is equal to .0120 for Case 1 and equal to .0144 for Case 2 when $\bar{\lambda} = .0005$ (resp., .0156 and .0175 when $\bar{\lambda} = .0001$). These are of the same magnitude as the best MISEs .0099 and .0121 in Figures 9 and 10. In Case 1, the tendency of the selection procedure to overpenalize without unduly affecting efficiency is explained by flatness of the MISE curve at the right hand side of the optimal lambda.

We also get average estimated values for the decay factors α and β close to the asymptotic ones. For $\hat{\alpha}$ the average (quartiles) is equal to 2.2502 (2.1456, 2.2641, 2.3628), and for $\hat{\beta}$ it is equal to 2.9222 (2.8790, 2.9176, 2.9619). To compute the estimated value for the decay factor γ we use $T \in \{500, 550, ..., 1000\}$ in the variance component of the MISE, together with the data driven estimate of θ in the bias component of the MISE. Optimizing on the grid of sample sizes yields an optimal lambda for each sample size per simulation. The logarithm of the optimal lambda is then regressed on the logarithm of the sample size, and the estimated

slope is averaged over the 1000 simulations to obtain the average estimated gamma. In Case 1, we get an average (quartiles) of .6081 (.4908, .6134, .6979), when $\bar{\lambda} = .0005$, and .7224 (.5171, .6517, .7277), when $\bar{\lambda} = .0001$. In Case 2, we 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. Even if his theoretical results are derived for bandwidth selection in semiparametric estimation, we believe that they could be extended to our case as well. Proposition 7 shows that a limit distribution exists, a prerequisite for applying subsampling. Recognizing that asymptotically $\lambda_T^* = cT^{-\gamma}$, we propose to choose c and γ which minimize the following estimator of the MISE: $\hat{M}(c,\gamma) = \frac{1}{I}\frac{1}{J}\sum_{i,j}\int_0^1 (\hat{\varphi}_{i,j}(x;c,\gamma) - \bar{\varphi}(x))^2 dx$, where $\hat{\varphi}_{i,j}(x;c,\gamma)$ denotes the estimator based on the jth subsample of size m_i ($m_i << T$) 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.

In our small scale study we take 500 subsamples ($J = 500$) for each subsample size $m_i \in \{50, 60, 70, ..., 100\}$ ($I = 6$), $\bar{\lambda} = \{.0005, .0001\}$, and $T = 1000$. To determine c and γ we build a joined grid with values around the OLS estimates coming from Case 1, namely $\{.15, .2, .25\} \times \{.7, .75, .8\}$, and coming from Case 2, namely $\{.005, .01, .015\} \times \{.35, .4, .45\}$. [23] The two grids yield a similar range for λ_T. In the experiments for $\rho = 0.5$ we want to verify whether the data driven procedure is able to pick most of the time c and γ in the first

[23] A full scale Monte-Carlo study based on large J and I and a fine grid for (c,γ) is computationally too demanding because of the resampling nature of the selection procedure.

set of values in Case 1, and in the second set of values in Case 2. On 1000 simulations we have found a frequency equal to 96% of adequate choices in Case 1 when $\bar{\lambda} = .0005$, and 87% when $\bar{\lambda} = .0001$. In Case 2 we have found 77% when $\bar{\lambda} = .0005$, and 82% when $\bar{\lambda} = .0001$. These frequencies are scattered among the grid values.

7 An empirical example

This section presents an empirical example with the data in Horowitz (2006). [24] We estimate an Engel curve based on the moment condition $E[Y - \varphi_0(X) \mid Z] = 0$, 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 regularisation parameters. We keep six polynomials. Here the value of the optimized objective function stabilizes after $k = 6$ (see Figure 13), and estimation results remain virtually unchanged for larger k. We have estimated the weighting matrix since $\Omega_0(z) = V[Y - \varphi_0(X) \mid Z = z]^{-1}$ 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 kernel estimator $\hat{s}^2(Z_t)$ of the conditional variance $s^2(Z_t) = \Omega_0(Z_t)^{-1}$ at observed sample points is of the same type as for the conditional moment restriction. Subsampling relies on 1000 subsamples ($J = 1000$) for each subsample size $m_i \in \{50, 53, ..., 200\}$ ($I = 51$), and the ex-

[24] We would like to thank Joel Horowitz for kindly providing the dataset.

tended grid $\{0.005, .01, .05, .1, .25, .5, 1, 2, ..., 6\} \times \{.3, .35, ..., .9\}$ 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 14 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 $\log Y$. Banks, Blundell and Lewbel (1997) consider demand systems that accommodate such empirical Engel curves.

8 Concluding remarks

We have studied a new estimator of a functional parameter identified by conditional moment restrictions. It exploits a Tikhonov regularization scheme to solve ill-posedness, and is referred to as the TiR estimator. Our framework proves to be (a) numerically tractable, (b) well-behaved in finite samples, (c) amenable to in-depth asymptotic analysis. (a) and (b) are key advantages for finding a route towards numerous empirical applications. (c) paves the way to further extensions: asymptotics for data driven estimation, estimation of average derivatives, estimation of semiparametric models, etc.

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.

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 (2004): "Semi-Nonparametric IV Estimation of Shape Invariant Engel Curves", Working Paper.

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 (2000): "Generalization of GMM to a Continuum of Moment Conditions", Econometric Theory, 16, 797-834.

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*, Vol. 6, Heckman, J. and E. Leamer (eds.).

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 (2006): "Instrumental Variable Identification and Estimation of Nonseparable Models via Quantile Conditions", forthcoming in Journal of Econometrics.

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.

Engl, H., Hanke, M. and A. Neubauer (2000): *Regularisation of Inverse Problems,* Kluwer Academic Publishers, Dordrecht.

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 (2006): "An Efficient Nonparametric Estimator for Models with Nonlinear Dependence", forthcoming in Journal of Econometrics.

Gallant, R. and D. Nychka (1987): "Semi-Nonparametric Maximum Likelihood Estimation", Econometrica, 55, 363-390.

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.

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 (2006): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model", Working Paper.

Hu, Y. and S. Schennach (2004): "Identification and Estimation of Nonclassical Nonlinear 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", Working Paper.

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), North Holland.

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: $\{R_t = (Y_t, X_t, Z_t) : t = 1, ..., T\}$ *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* R *is in class* $C^m\left(\mathbb{R}^d\right)$, *with* $m \geq 2$.

B.3: *The density* f *of* X *given* Z *is such that* $\sup_{x \in \mathcal{X}, z \in \mathcal{Z}} f(x|z) < \infty$.

B.4: *The kernel* K *is a Parzen-Rosenblatt kernel of order* m *on* $\mathbb{R}^d$, *that is* (i) $\int K(u)du = 1$, *and* K *is bounded*; (ii) $\int u^\alpha K(u)du = 0$ *for any multi-index* $\alpha \in \mathbb{N}^d$ *with* $|\alpha| < m$, *and* $\int |u|^m |K(u)| \, du < \infty$.

B.5: *The kernel* K *is such that* $\int |K(u)| \, q(u)du < \infty$ *where* $q(u) = \int |K(u+z)| \, |z|^2 \, dz$.

B.6: *The density* f *of* R *is such that there exists a function* $\omega \in L^2(F)$ *satisfying* $\omega \geq 1$ *and* $\sup_{0 \leq t \leq h} \int \overline{K}(z) \left| \frac{f(r+tz) - f(r)}{f(r)} \right| dz \leq h\omega^2(r)$, $\sup_{0 \leq t \leq h} \int \widetilde{K}(z) \left| \frac{f(r+tz) - f(r)}{f(r)} \right| dz \leq h\omega^2(r)$ $\sup_{0 \leq t \leq h} \int |K(z)| \left| \frac{f(r+tz) - f(r)}{f(r)} \right|^2 dz \leq h^2\omega^2(r)$, *for any* $r \in \mathcal{S}$ *and* $h > 0$ *small, where* $\overline{K}(z) := \int |K(u+z)K(u)| \, du$ *and* $\widetilde{K}(z) := \int |K(u+z)K(u)| \, q(u)du$.

B.7: *The density* f *of* R *is such that there exists a function* $\omega_m \in L^2(F)$ *satisfying* $\sup_{\alpha \in \mathbb{N}^d : |\alpha| = m} \sup_{0 \leq t \leq h} \int |K(u)| \left| \frac{\nabla^\alpha f(r+tu)}{f(r)} \right| |u|^m \, du \leq \omega_m(r)$, *for any* $r \in \mathcal{S}$ *and* $h > 0$ *small.*

B.8: *The moment function* g *is differentiable and such that* $\sup_{u,v} |\nabla_v g(u, v)| < \infty$.

B.9: *The weighting matrix* $\Omega_0(z) = V\left[g\left(Y, \varphi_0\left(X\right)\right) \mid Z = z\right]^{-1}$ *is such that* $E\left[|\Omega_0(Z)|\right] < \infty$.

B.10: *The* $\langle .,. \rangle_H$*-orthonormal basis of eigenvectors* $\left\{\phi_j : j \in \mathbb{N}\right\}$ *of operator* A^*A *satisfies* (i) $\sum_{j=1}^{\infty} \left\|\phi_j\right\| < \infty$; (ii) $\sum_{j,l=1, j\neq l}^{\infty} \frac{\left\langle \phi_j, \phi_l \right\rangle^2}{\left\|\phi_j\right\|^2 \left\|\phi_l\right\|^2} < \infty$.

B.11: *The eigenfunctions* ϕ_j *and the eigenvalues* ν_j *of* A^*A *are such that* $\sup_{j\in\mathbb{N}} E\left[\omega(R)^2 |g_j(R)|^2\right] < \infty$ *and* $\sup_{j\in\mathbb{N}} E\left[\omega(R)^2 |\nabla g_j(R)|^2\right] < \infty$, *where* $g_j(r) := \left(A\phi_j\right)(z)' \Omega_0(z) g(y, \varphi_0(x)) / \sqrt{\nu_j}$ *and* ω *is as in Assumption B.6.*

B.12: *There exists a constant* $C < \infty$ *such that for all* $j \in \mathbb{N}$ *and* $h > 0$ *small:* $\sup_{\alpha\in\mathbb{N}^d : |\alpha| = m} \sup_{0\leq t\leq h} \int |K(u)| \, |u|^m E\left[|\nabla^\alpha g_j(R - tu)|^2\right]^{1/2} du \leq C.$

B.13: *The functions* g_j *are such that* $\sup_{j\in\mathbb{N}} E\left[\chi(R,h)^2 |g_j(R)|^2\right] = o(h)$, *as* $h \to 0$, *where* $\chi(r,h) := \int \overline{K}(z) 1_{\mathcal{S}}(r) 1_{\mathcal{S}^c}(r - hz) dz$ *and* $\overline{K}$ *is as in Assumption B.6.*

B.14: *The estimator* $\hat{\Omega}$ *of* Ω_0 *is such that* $\int \left|g_{\varphi_0}(w)\right| f(w,z)^{1/2} E\left[\left|\Delta\hat{\Sigma}(z)\right|^4\right]^{1/4} f(z) dw dz = O\left(\frac{1}{\sqrt{T h^{2d_Z}}}\right)$, *where* $g_{\varphi_0}(w) = g(y, \varphi_0(x))$, $\Delta\hat{\Sigma}(z) := \hat{\Omega}(z)/\hat{f}(z) - \Omega_0(z)/f_0(z)$.

B.15: *For any* $\bar{\zeta} \in \mathbb{N}$: $E\left[\hat{I}_3(x,\xi)^{2\bar{\zeta}}\right] = O\left(a_T^{\bar{\zeta}}\right)$, *uniformly in* $x, \xi \in [0,1]$, *where* $\hat{I}_3(x,\xi) := \int \hat{f}(x,z) \hat{f}(\xi,z) \Delta\hat{\Sigma}(z) dz$ *and* $a_T := \frac{1}{T h_T} + h_T^{2m} \log T$.

B.16: *The estimator* $\hat{\Omega}$ *is such that* $E\left[\sup_{z\in\mathcal{Z}} \|\nabla_z^\alpha \hat{a}(.,z)\|\right] = O(\log T)$, *for any* $\alpha \in \mathbb{N}^{d_Z}$ *s.t.* $|\alpha| = m$, *where* $\hat{a}(x,z) := \int \hat{\Omega}(z) g_{\varphi_0}(w) \hat{f}(w|z) \hat{f}(x|z) dw$.

B.17: *The estimator* $\hat{\Omega}$ *is such that* $E\left[\sup_{z\in\mathcal{Z}} \left|\nabla_z^\alpha \hat{b}(x,\xi,z)\right|^{2\bar{\zeta}}\right]^{1/\bar{\zeta}} = O(\log T)$, *uniformly in* $x, \xi \in [0,1]$, *for any* $\bar{\zeta} \in \mathbb{N}$ *and any* $\alpha \in \mathbb{N}^{d_Z}$ *s.t.* $|\alpha| = m$, *where* $\hat{b}(x,\xi,z) := \hat{f}(x|z) \hat{f}(\xi|z) \hat{\Omega}(z)$.

Assumption B.1 of i.i.d. data avoids additional technicalities in the proofs. Results can be extended to the time series setting. Assumptions B.2, B.3 and B.4 are classical conditions in kernel density estimation concerning smoothness of the density and of the kernel. Assumptions B.5, B.6 and B.7 require existence of higher order moments of the kernel and a sufficient degree of smoothness of the density. These assumptions are used in the proof of Lemma A.3 to bound higher order terms in the asymptotic expansion of the MISE. Assumption B.8 is a smoothness condition on the moment function g. Assumption B.9, together with Assumptions B.3 and B.8, imply that the operator A is compact. Assumption B.10 (i) is used to simplify the proof of Lemma A.9. It is met under Assumption 5 (ii) with $\beta > 2$. Assumption B.10 (ii) requires that the eigenfunctions of operator A^*A, which are orthogonal w.r.t. $\langle ., . \rangle_H$, are sufficiently orthogonal w.r.t. $\langle ., . \rangle$. 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.11 and B.12 ask for the existence of a uniform bound for moments of derivatives of functions $g_j(r) = \frac{1}{\sqrt{\nu_j}} \left(A\phi_j\right)(z)' \Omega_0(z) g\left(y, \varphi_0(x)\right)$, $j \in \mathbb{N}$. These functions satisfy $E\left[g_j(R)^2\right] = 1$. Assumptions B.11 and B.12 are met whenever moment function $g\left(y, \varphi_0(x)\right)$, instrument $\frac{1}{\sqrt{\nu_j}} \left(A\phi_j\right)(z)$, 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. In Assumption B.13, the support of function $\chi(., h)$ shrinks around the boundary of $\mathcal{S}$ as $h \to 0$. Thus, Assumption B.13 imposes a uniform bound on the behavior of functions $g_j(r)$, $j \in \mathbb{N}$, close to this boundary. It is used in the proof of Lemma

A.3. Assumptions B.14 and B.15 are restrictions on the rate of convergence of $\hat{\Omega}$ and guarantee that estimation of the weighting matrix Ω_0 has no impact on the asymptotic MISE of the TiR estimator. They are used in Lemmas B.11 and B.12 in the Technical Report, respectively. In general managing large values of $\hat{\Omega}(z)/\hat{f}(z)$ requires trimming. Finally, Assumptions B.16 and B.17 control for the residual terms in the asymptotic expansion of the MISE. They are needed since the estimate $\hat{A}^*$ of A^* defined in Lemma A.2 (i) differs from the adjoint $(\hat{A})^*$ of $\hat{A}$ in finite sample (cf. discussion in Carrasco, Florens and Renault (2005)).

Appendix 2

Consistency of the TiR estimator

A.2.1 Existence of penalized extremum estimators

Since Q_T is positive, a function $\hat{\varphi} \in \Theta$ is solution of optimization problem in (9) if and only if it is a solution

$$\hat{\varphi} = \arg \inf_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T G(\varphi), \quad \text{s.t.} \quad \lambda_T G(\varphi) \leq L_T \ , \tag{26}$$

where $L_T := Q_T(\varphi_0) + \lambda_T G(\varphi_0)$. The solution $\hat{\varphi}$ in (26) exists P-a.s. if

(i) mappings $\varphi \to G(\varphi)$ and $\varphi \to Q_T(\varphi)$ are lower semicontinuous on Θ, P-a.s., for any T, w.r.t. the L^2 norm $\|.\|$;

(ii) set $\left\{\varphi \in \Theta : G(\varphi) \leq \bar{L}\right\}$ is compact w.r.t. the L^2 norm $\|.\|$, 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 extremum estimators

Proof of Theorem 1: For any T and any given $\varepsilon > 0$, we have

$$P\left[\|\hat{\varphi} - \varphi_0\| > \varepsilon\right] \leq P\left[\inf_{\varphi\in\Theta:\|\varphi-\varphi_0\|\geq\varepsilon} Q_T(\varphi) + \lambda_T G(\varphi) \leq Q_T(\varphi_0) + \lambda_T G(\varphi_0)\right].$$

Let us bound the probability on the RHS. Denoting $\Delta Q_T := Q_T - Q_\infty$, we get

$$\begin{aligned}
\inf_{\varphi\in\Theta:\|\varphi-\varphi_0\|\geq\varepsilon} Q_T(\varphi) + \lambda_T G(\varphi) &\leq Q_T(\varphi_0) + \lambda_T G(\varphi_0) \\
\Longrightarrow \inf_{\varphi\in\Theta:\|\varphi-\varphi_0\|\geq\varepsilon} Q_\infty(\varphi) + \lambda_T G(\varphi) + \inf_{\varphi\in\Theta} \Delta Q_T(\varphi) &\leq \lambda_T G(\varphi_0) + \sup_{\varphi\in\Theta} |\Delta Q_T(\varphi)| \\
\Longrightarrow \inf_{\varphi\in\Theta:\|\varphi-\varphi_0\|\geq\varepsilon} Q_\infty(\varphi) + \lambda_T G(\varphi) - \lambda_T G(\varphi_0) &\leq 2\sup_{\varphi\in\Theta} |\Delta Q_T(\varphi)| = 2\bar{\delta}_T.
\end{aligned}$$

Thus, from (iii) we get for any $a \geq 0$ and $b > 0$

$$\begin{aligned}
P\left[\|\hat{\varphi} - \varphi_0\| > \varepsilon\right] &\leq P\left[C_\varepsilon(\lambda_T) \leq 2\bar{\delta}_T\right] \\
&= P\left[1 \leq \frac{1}{\lambda_T^{-a} C_\varepsilon(\lambda_T)} \frac{1}{\left(T\lambda_T^{a/b}\right)^b} \left(2T^b \bar{\delta}_T\right)\right] =: P\left[1 \leq \bar{Z}_T\right].
\end{aligned}$$

Since $\lambda_T \to 0$ such that $(T\lambda_T^{a/b})^{-1} \to 0$, P-a.s., for a and b chosen as in (iv) and (v) we have $\bar{Z}_T \xrightarrow{p} 0$, and we deduce $P\left[\|\hat{\varphi} - \varphi_0\| > \varepsilon\right] \leq P\left[\bar{Z}_T \geq 1\right] \to 0$. Since $\varepsilon > 0$ is arbitrary, the proof is concluded. This proof and Equation (26) show that Condition (i) could be weakened to $\bar{\delta}_T := \sup_{\varphi\in\bar{\Theta}_T} |Q_T(\varphi) - Q_\infty(\varphi)| \xrightarrow{p} 0$, where $\bar{\Theta}_T := \{\varphi \in \Theta : G(\varphi) \leq G(\varphi_0) + Q_T(\varphi_0)/\lambda_T\}$.

Proof of Proposition 2: We prove that, for any $\varepsilon > 0$ and any sequence (λ_n) such that $\lambda_n \searrow 0$, we have $\lambda_n^{-1} C_\varepsilon(\lambda_n) > 1$ for n large, which implies both statements of Proposition 2. Without loss of generality we set $Q_\infty(\varphi_0) = 0$. By contradiction, assume that there exists $\varepsilon > 0$ and a sequence (λ_n) such that $\lambda_n \searrow 0$ and

$$C_\varepsilon(\lambda_n) \leq \lambda_n, \qquad \forall n \in \mathbb{N}. \tag{27}$$

By definition of function $C_\varepsilon(\lambda)$, for any $\lambda > 0$ and $\eta > 0$, there exists $\varphi \in \Theta$ such that $\|\varphi - \varphi_0\| \geq \varepsilon$, and $Q_\infty(\varphi) + \lambda G(\varphi) - \lambda G(\varphi_0) \leq C_\varepsilon(\lambda) + \eta$. Setting $\lambda = \eta = \lambda_n$ for $n \in \mathbb{N}$, we deduce from (27) that there exists a sequence (φ_n) such that $\varphi_n \in \Theta, \|\varphi_n - \varphi_0\| \geq \varepsilon$, and

$$Q_\infty(\varphi_n) + \lambda_n G(\varphi_n) - \lambda_n G(\varphi_0) \leq 2\lambda_n. \tag{28}$$

Now, since $Q_\infty(\varphi_n) \geq 0$, we get $\lambda_n G(\varphi_n) - \lambda_n G(\varphi_0) \leq 2\lambda_n$, that is

$$G(\varphi_n) \leq G(\varphi_0) + 2. \tag{29}$$

Moreover, since $G(\varphi_n) \geq G_0$, where G_0 is the lower bound of function G, we get $Q_\infty(\varphi_n) + \lambda_n G_0 - \lambda_n G(\varphi_0) \leq 2\lambda_n$ from (28), that is $Q_\infty(\varphi_n) \leq \lambda_n (2 + G(\varphi_0) - G_0)$, which implies

$$\lim_n Q_\infty(\varphi_n) = 0 = Q_\infty(\varphi_0). \tag{30}$$

Obviously, the simultaneous holding of (29) and (30) violates Assumption (11).

A.2.3 Penalization with Sobolev norm

To conclude on existence and consistency of the TiR estimator, let us check the assumptions in A.2.1 and Proposition 2 for the special case $G(\varphi) = \|\varphi\|_H^2$ under Assumptions 1-2.

(i) The mapping $\varphi \to \|\varphi\|_H^2$ is lower semicontinuous on $H^2[0,1]$ w.r.t. the norm $\|.\|$ (see Reed and Simon (1980), p. 358). Continuity of $Q_T(\varphi)$, P-a.s., follows from the mapping $\varphi \to \hat{m}(\varphi, z)$ being continuous for almost any $z \in \mathcal{Z}$, P-a.s.. The latter holds since for any $\varphi_1, \varphi_2 \in \Theta$, $|\hat{m}(\varphi_1, z) - \hat{m}(\varphi_2, z)| \leq \int \left(\int \sup_v |\nabla_v g(y, v)| \left| \hat{f}(w|z) \right| dy \right) |\varphi_1(x) - \varphi_2(x)| \, dx$ $\leq \bar{C}_T \|\varphi_1 - \varphi_2\|$, 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, Assumptions B.4 and B.8.

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

(iii) The set $\bar{\Theta}_T$ in the proof of Theorem 1 is compact, P-a.s..

(iv) Assumptions of Proposition 2 are satisfied. Clearly function $G(\varphi) = \|\varphi\|_H^2$ is bounded from below by 0. Furthermore Assumption (11) holds.

Lemma A.1: *Assumption 1 implies Assumption (11) in Proposition 2 for* $G(\varphi) = \|\varphi\|_H^2$.

Proof: By contradiction, let $\varepsilon > 0$, $0 < \bar{L} < \infty$ and (φ_n) be a sequence in Θ such that $\|\varphi_n - \varphi_0\| \geq \varepsilon$ for all $n \in \mathbb{N}$,

$$Q_\infty(\varphi_n) \to 0 \text{ as } n \to \infty, \tag{31}$$

and $\|\varphi_n\|_H^2 \leq \bar{L}$ for any n. Then, the sequence (φ_n) belongs to the compact set $\left\{\varphi \in \Theta : \|\varphi\|_H^2 \leq \bar{L}\right\}$. Thus, there exists a converging subsequence $\varphi_{N_n} \to \varphi_0^* \in \Theta$. Since Q_∞ is continuous, $Q_\infty\left(\varphi_{N_n}\right) \to Q_\infty(\varphi_0^*)$. From (31) we deduce $Q_\infty(\varphi_0^*) = 0$, and $\varphi_0^* = \varphi_0$ from identification Assumption 1 (i). This violates the condition that $\|\varphi_n - \varphi_0\| \geq \varepsilon$ for all $n \in \mathbb{N}$.

Appendix 3

The MISE of the TiR estimator

A.3.1 The first-order condition

The estimated moment function is $\hat{m}(\varphi, z) = \int \varphi(x)\,\hat{f}(w|z)dw - \int y\hat{f}(w|z)dw =:$

$\left(\hat{A}\varphi\right)(z) - \hat{r}(z)$. The objective function of the TiR estimator becomes

$$Q_T(\varphi) + \lambda_T \|\varphi\|_H^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{\Omega}(Z_t)\left[\left(\hat{A}\varphi\right)(Z_t) - \hat{r}(Z_t)\right]^2 + \lambda_T \langle\varphi,\varphi\rangle_H, \tag{32}$$

and can be written as a quadratic form in $\varphi \in H^2[0,1]$. To achieve this, let us introduce the empirical counterpart $\hat{A}^*$ of operator A^*.

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

(i) There exists a linear operator $\hat{A}^*$*, such that*
$\left\langle \varphi, \hat{A}^*\psi \right\rangle_H = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{A}\varphi\right)(Z_t)\,\hat{\Omega}(Z_t)\psi(Z_t)$ *, for any measurable* ψ *and any* $\varphi \in H^2[0,1]$*;*
(ii) Operator $\hat{A}^*\hat{A} : H^2[0,1] \to H^2[0,1]$ *is compact.*

Then, from Lemma A.2 (i), Criterion (32) can be rewritten as

$$Q_T(\varphi) + \lambda_T \|\varphi\|_H^2 = \langle\varphi, \left(\lambda_T + \hat{A}^*\hat{A}\right)\varphi\rangle_H - 2\langle\varphi, \hat{A}^*\hat{r}\rangle_H, \tag{33}$$

up to a term independent of φ. From Lemma A.2 (ii), $\hat{A}^*\hat{A}$ is a compact operator from $H^2[0,1]$ to itself. 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 (33) admits a global minimum over $H^2[0,1]$. It is given by the first-order condition $\left(\lambda_T + \hat{A}^*\hat{A}\right)\hat{\varphi} = \hat{A}^*\hat{r}$, that is

$$\hat{\varphi} = \left(\lambda_T + \hat{A}^*\hat{A}\right)^{-1}\hat{A}^*\hat{r}. \tag{34}$$

A.3.2 Asymptotic expansion of the first-order condition

Let us now expand the estimator in (34). We can write

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

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

$$\hat{\varphi} - \varphi_0 = \left(\lambda_T + A^*A\right)^{-1}A^*\hat{\psi} + \left[\left(\lambda_T + A^*A\right)^{-1}A^*A\varphi_0 - \varphi_0\right] + \mathcal{R}_T =: \mathcal{V}_T + \mathcal{B}_T + \mathcal{R}_T, \quad (35)$$

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

$$\begin{aligned}\mathcal{R}_T &= \left[\left(\lambda_T + \hat{A}^*\hat{A}\right)^{-1} - \left(\lambda_T + A^*A\right)^{-1}\right]A^*\hat{\psi} \\ &\quad + \left[\left(\lambda_T + \hat{A}^*\hat{A}\right)^{-1}\hat{A}^*\hat{A} - \left(\lambda_T + A^*A\right)^{-1}A^*A\right]\varphi_0 + \left(\lambda_T + \hat{A}^*\hat{A}\right)^{-1}\left(\hat{A}^*\left(\hat{q}+\hat{\psi}\right) - A^*\hat{\psi}\right).\end{aligned} \quad (36)$$

We prove at the end of this Appendix (Section A.3.5) that the residual term $\mathcal{R}_T$ in (35) is asymptotically negligible, i.e. $E\left[\|\mathcal{R}_T\|^2\right] = o\left(E\left[\|\mathcal{V}_T + \mathcal{B}_T\|^2\right]\right)$. Then, we deduce

$$\begin{aligned}E\left[\|\hat{\varphi} - \varphi_0\|^2\right] &= E\left[\|\mathcal{V}_T + \mathcal{B}_T\|^2\right] + E\left[\|\mathcal{R}_T\|^2\right] + 2E\left[\langle \mathcal{V}_T + \mathcal{B}_T, \mathcal{R}_T\rangle\right] \\ &= E\left[\|\mathcal{V}_T + \mathcal{B}_T\|^2\right] + o\left(E\left[\|\mathcal{V}_T + \mathcal{B}_T\|^2\right]\right),\end{aligned}$$

by applying twice the Cauchy-Schwarz inequality. Since

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

we get

$$\begin{aligned}E\left[\|\hat{\varphi} - \varphi_0\|^2\right] &= \left\|\left(\lambda_T + A^*A\right)^{-1}A^*A\varphi_0 - \varphi_0 + \left(\lambda_T + A^*A\right)^{-1}A^*E\hat{\psi}\right\|^2 \\ &\quad + E\left[\left\|\left(\lambda_T + A^*A\right)^{-1}A^*\left(\hat{\psi} - E\hat{\psi}\right)\right\|^2\right], \quad (38)\end{aligned}$$

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

A.3.4 Asymptotic expansion of the bias term

Lemma A.4: *Define* $b(\lambda_T) = \left\|\left(\lambda_T + A^*A\right)^{-1} A^*A\varphi_0 - \varphi_0\right\|$. *Then, under Assumptions B and the bandwidth condition* $h_T^m = o\left(\lambda_T b(\lambda_T)\right)$, *where* m *is the order of the kernel* K, *we have* $\left\|\left(\lambda_T + A^*A\right)^{-1} A^*A\varphi_0 - \varphi_0 + \left(\lambda_T + A^*A\right)^{-1} A^*E\hat{\psi}\right\| = b(\lambda_T)$, *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* $\frac{1}{Th_T^{d_Z+d/2}} + h_T^m \log T = o\left(\lambda_T b\left(\lambda_T\right)\right)$, $\left(Th_T^{d+d_Z}\right)^{-1} = O(1)$, $E\left[\left\|\left(1 + S\left(\lambda_T\right)\hat{U}\right)^{-1} S\left(\lambda_T\right)\hat{U}\right\|^8\right] = O(1)$, and $E\left[\left\|S\left(\lambda_T\right)\hat{U}\right\|^8\right] = o(1)$, *where* m *is the order of the kernel* K, d_Z *and* d *are the dimensions of* Z *and* (Y, X, Z), *respectively,* $S\left(\lambda_T\right) := \left(\lambda_T + A^*A\right)^{-1}$, *and* $\hat{U} := \hat{A}^*\hat{A} - A^*A$. *Then, under Assumptions B,* $E\left[\left\|\mathcal{R}_T\right\|^2\right] = o\left(E\left[\left\|V_T + \mathcal{B}_T\right\|^2\right]\right)$.

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

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

The second part of Lemma A.5 clarifies the sufficiency of the condition $\left(\frac{1}{Th_T}+h_T^{2m}\log T\right)=O(\lambda_T^{2+\varepsilon})$, $\varepsilon>0$, in the control of the remaining term $\mathcal{R}_T$.

Appendix 4

Rate of convergence with geometric spectrum

i) The next Lemma A.6 characterizes the variance term (see Wahba (1977) for similar results).

Lemma A.6: *Let* ν_j *and* $\left\|\phi_j\right\|^2$ *satisfy Assumption 5, and define the function* $I(\lambda)=\sum_{j=1}^{\infty}\frac{\nu_j}{\left(\lambda+\nu_j\right)^2}\left\|\phi_j\right\|^2,\quad\lambda>0.$ *Then,* $\lambda\left[\log\left(1/\lambda\right)\right]^{\beta}I(\lambda)=\left(\frac{1}{\alpha}\right)^{1-\beta}C_2\left[1+c\left(\lambda\right)\right]$ $+o(1)$, *as* $\lambda\to0$, *where* $c\left(\lambda\right)$ *is a function such that* $\left|c\left(\lambda\right)\right|\leq1/4$ *and* $\left|\lambda\frac{dc}{d\lambda}\left(\lambda\right)\right|\leq1/4.$

From Lemma A.6 and using Assumption 6, we get $M_T(\lambda)=c_1\frac{1}{T}\frac{1+c\left(\lambda\right)}{\lambda\left[\log\left(1/\lambda\right)\right]^{\beta}}+c_2\lambda^{2\delta}$, up to negligible terms for $\lambda\to0$ and $T\to\infty$, where $c_1=\left(\frac{1}{\alpha}\right)^{1-\beta}C_2$, $c_2=C_3^2$.

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

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

where $\kappa\left(\lambda\right):=c_1\left[1+c(\lambda)\right]\left[1-\frac{\beta}{\log\left(1/\lambda\right)}\right]-\lambda c_1\frac{dc}{d\lambda}\left(\lambda\right)$. From Lemma A.6 function $\kappa\left(\lambda\right)$ is positive, bounded and bounded away from 0 as $\lambda\to0$. Computation of the second derivative

shows that $M_T(\lambda)$ is a convex function of λ, for small λ. We get

$$\frac{dM_T(\lambda_T^*)}{d\lambda} = 0 \iff \frac{1}{T}\frac{1}{2c_2\delta}\frac{\kappa\left(\lambda_T^*\right)}{\left[\log\left(1/\lambda_T^*\right)\right]^\beta} = \left(\lambda_T^*\right)^{2\delta+1}. \tag{39}$$

To solve the latter equation for λ_T^*, define $\tau_T := \log\left(1/\lambda_T^*\right)$. Then $\tau_T = c_3 + \frac{1}{1+2\delta}\log T + \frac{\beta}{1+2\delta}\log\tau_T - \frac{1}{1+2\delta}\log\kappa\left(\lambda_T^*\right)$, where $c_3 = (1+2\delta)^{-1}\log(2c_2\delta)$. It follows that $\tau_T = c_4 + \frac{1}{1+2\delta}\log T + \frac{\beta}{1+2\delta}\log\log T + o(\log\log T)$, for a constant c_4, that is $\log\left(\lambda_T^*\right) = -c_4 - \frac{1}{1+2\delta}\log T - \frac{\beta}{1+2\delta}\log\log T + o(\log\log T)$.

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

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

From (39), $\lambda_T^* = \left(\frac{1}{2c_2\delta}\kappa\left(\lambda_T^*\right)\right)^{\frac{1}{2\delta+1}} T^{-\frac{1}{2\delta+1}}\left(\frac{1}{\tau_T^\beta}\right)^{\frac{1}{2\delta+1}} = c_{5,T}T^{-\frac{1}{2\delta+1}}\tau_T^{-\frac{\beta}{2\delta+1}}$, where $c_{5,T}$ is a sequence which is bounded and bounded away from 0. Thus we get

$$\begin{aligned} M_T\left(\lambda_T^*\right) &= c_1\frac{1}{T}\frac{1+c\left(\lambda_T^*\right)}{c_{5,T}}T^{\frac{1}{2\delta+1}}\frac{1}{\tau_T^{-\frac{\beta}{2\delta+1}+\beta}} + c_2c_{5,T}^{2\delta}T^{-\frac{2\delta}{2\delta+1}}\tau_T^{-\frac{2\delta\beta}{2\delta+1}} \\ &= c_{6,T}T^{-\frac{2\delta}{2\delta+1}}\tau_T^{-\frac{2\delta\beta}{2\delta+1}} \quad = \quad c_{7,T}T^{-\frac{2\delta}{2\delta+1}}\left(\log T\right)^{-\frac{2\delta\beta}{2\delta+1}}, \end{aligned}$$

up to a term which is negligible w.r.t. the RHS, where $c_{6,T}$ and $c_{7,T}$ are bounded and bounded away from 0.

Appendix 5

Asymptotic normality of the TiR estimator

From Equation (35) in Appendix 3, we have

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

where $\mathcal{R}_T(x)$ is defined in (36). 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 7.

A.5.1 Asymptotic normality of (I)

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

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

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

$$\sqrt{T/\sigma_T^2(x)}\left(\lambda_T+A^*A\right)^{-1}A^*\left(\hat{\psi}-E\hat{\psi}\right)(x)=\sum_{j=1}^{\infty}w_{j,T}(x)Z_{j,T}, \tag{40}$$

where $Z_{j,T}:=\dfrac{1}{\sqrt{\nu_j}}\langle\phi_j,\sqrt{T}A^*\left(\hat{\psi}-E\hat{\psi}\right)\rangle_H, \quad j=1,2,...,$

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

Note that $\sum_{j=1}^{\infty} w_{j,T}(x)^2 = 1$. Equation (40) can be rewritten (see the proof of Lemma A.3) using

$$\sum_{j=1}^{\infty} w_{j,T}(x) Z_{j,T} = \sqrt{T} \int G_T(r) \left[\hat{f}(r) - E\hat{f}(r) \right] dr, \tag{41}$$

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

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

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 $E[g_j(R)] = \frac{1}{\sqrt{\nu_j}} E\left[\left(A\phi_j\right)(Z)\, \Omega_0(Z) E\left[g_{\varphi_0}(W) | Z\right]\right] = 0$, and

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

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

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

$$\frac{1}{T^{1/2}} E\left[|Y_{tT}|^3\right] \to 0, \qquad T \to \infty. \tag{42}$$

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

$$\frac{1}{T^{1/2}} E\left[|Y_{tT}|^3\right] \leq \frac{1}{T^{1/2}} E\left[\left(\sum_{j=1}^{\infty} |w_{j,T}(x)|\, |g_j(R)|\right)^3\right] = \frac{1}{T^{1/2}} \left\| \sum_{j=1}^{\infty} |w_{j,T}(x)|\, |g_j| \right\|_3^3$$

$$\leq \frac{1}{T^{1/2}}\left(\sum_{j=1}^{\infty}\left|w_{j,T}(x)\right|\left\|g_j\right\|_3\right)^3 = \frac{1}{T^{1/2}}\frac{\left(\sum_{j=1}^{\infty}\frac{\sqrt{\nu_j}}{\lambda_T+\nu_j}\left|\phi_j(x)\right|\left\|g_j\right\|_3\right)^3}{\left(\sum_{j=1}^{\infty}\frac{\nu_j}{\left(\lambda_T+\nu_j\right)^2}\phi_j(x)^2\right)^{3/2}}.$$

Moreover, from the Cauchy-Schwarz inequality we have

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

and $\sum_{j=1}^{\infty}a_j^{-1}<\infty$, $a_j>0$. Thus, we get

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

and Condition (42) is implied by Condition (20).

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

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

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

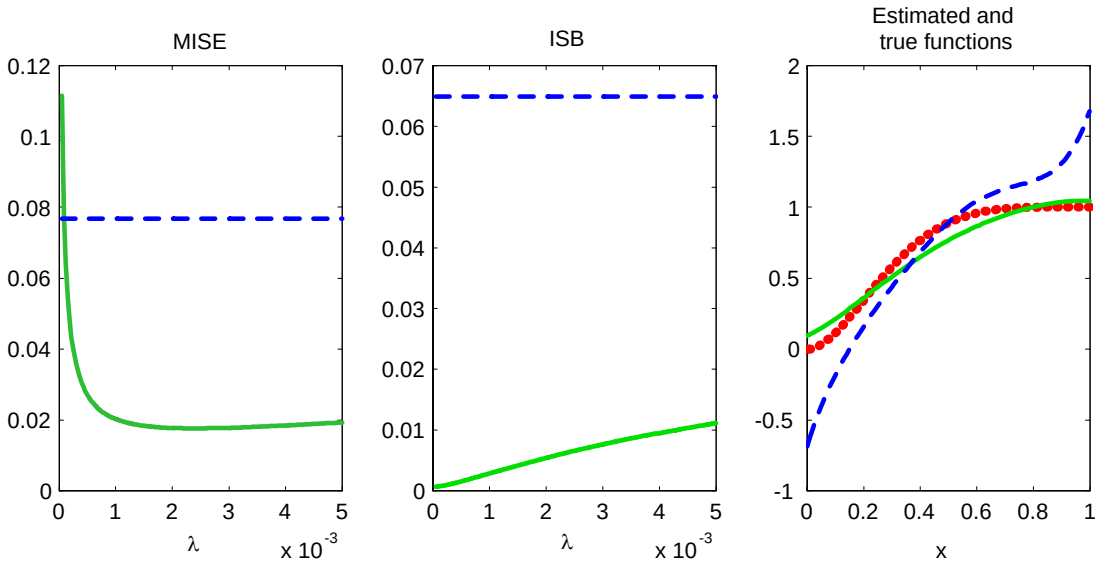

Figure 1: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0.5$, and sample size is $T = 400$.

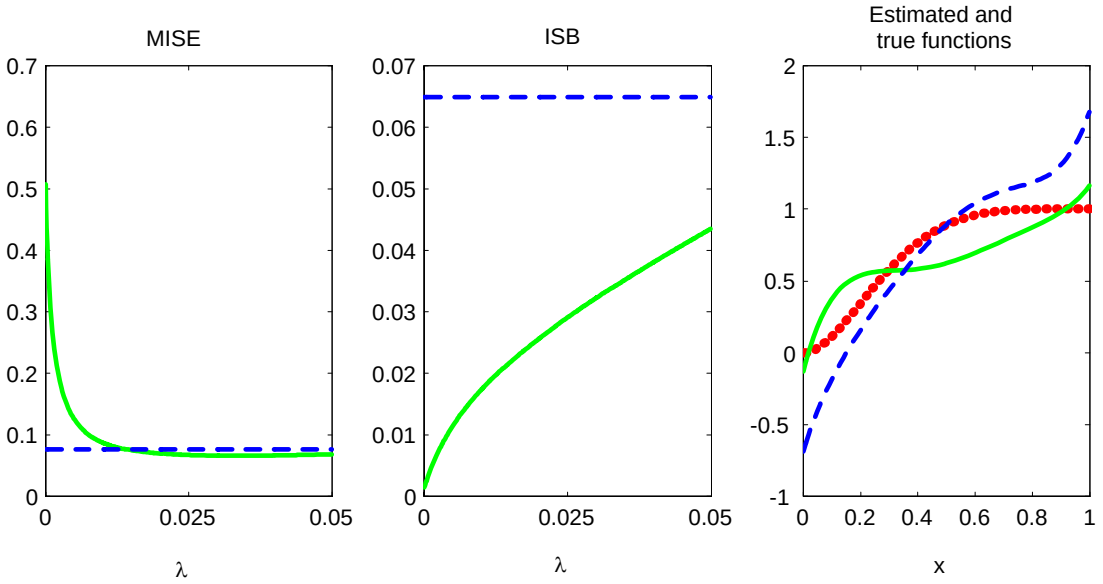

Figure 2: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the regularized estimator using L^2 norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0.5$, and sample size is $T = 400$.

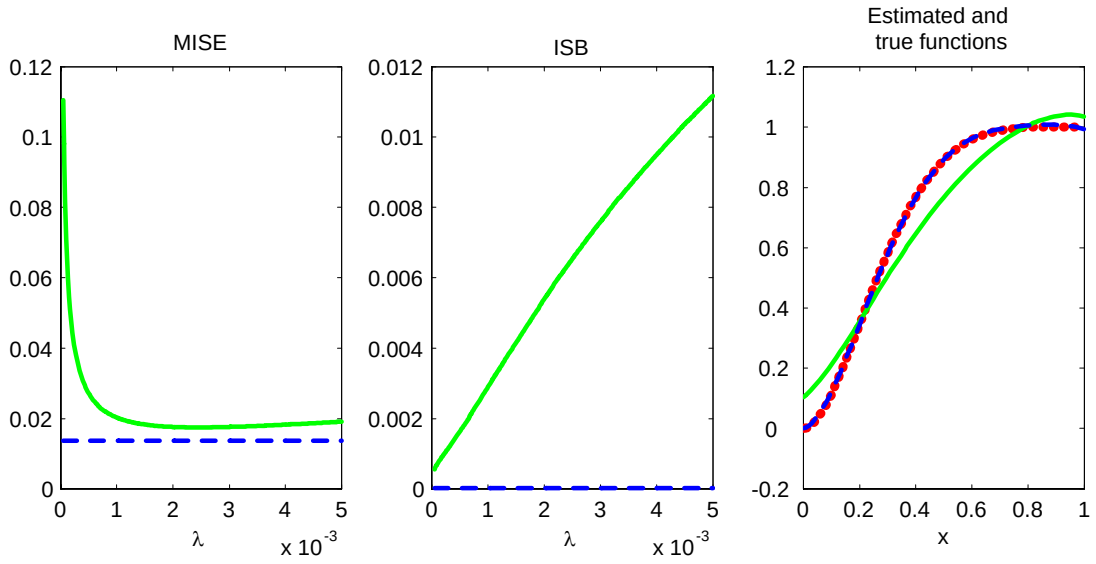

Figure 3: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0$, and sample size is $T = 400$.

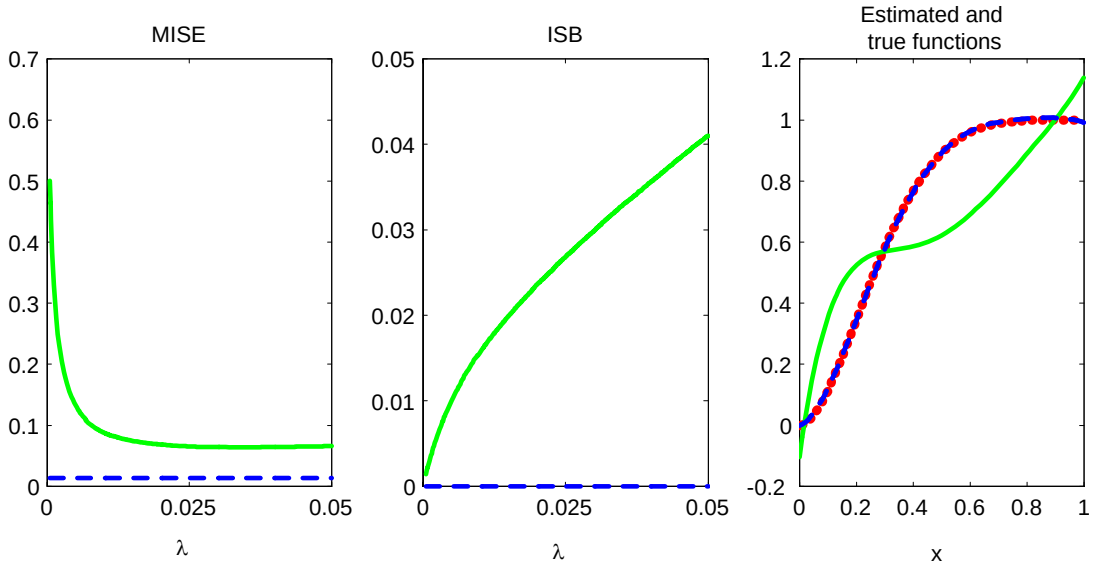

Figure 4: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the regularized estimator using L^2 norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0$, and sample size is $T = 400$.

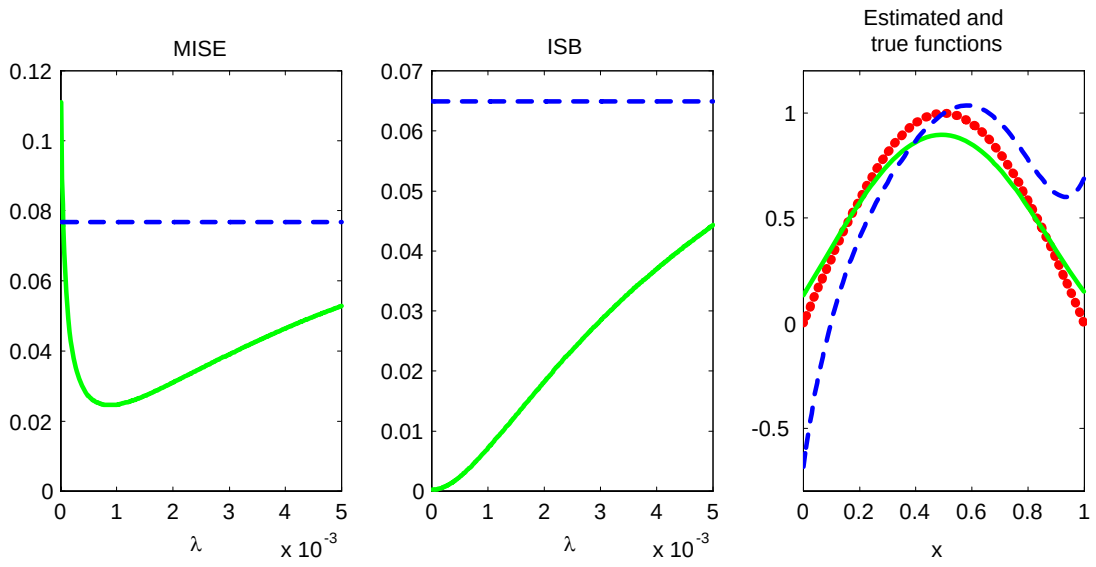

Figure 5: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0.5$, and sample size is $T = 400$.

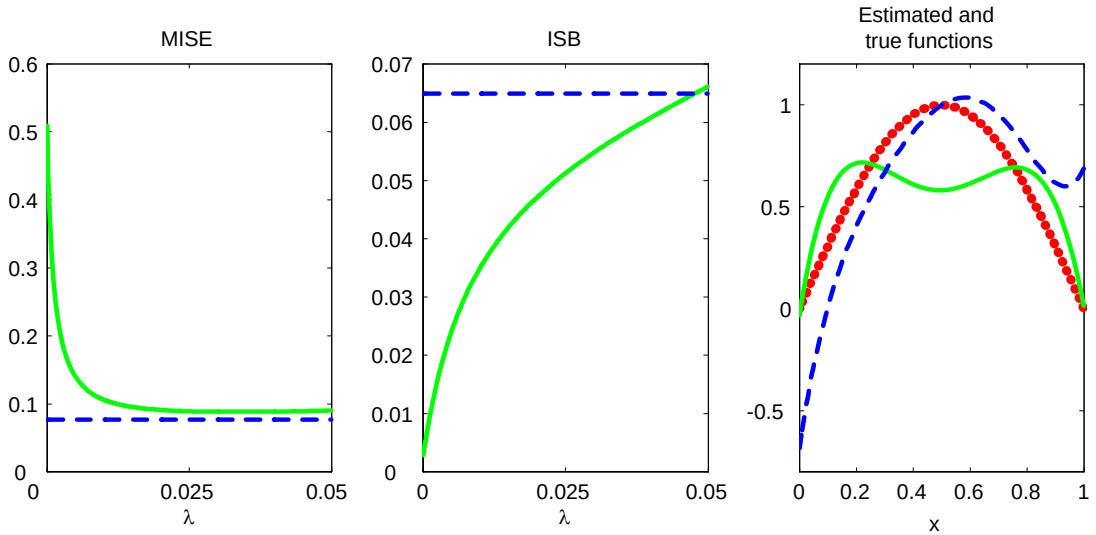

Figure 6: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the regularized estimator using L^2 norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0.5$, and sample size is $T = 400$.

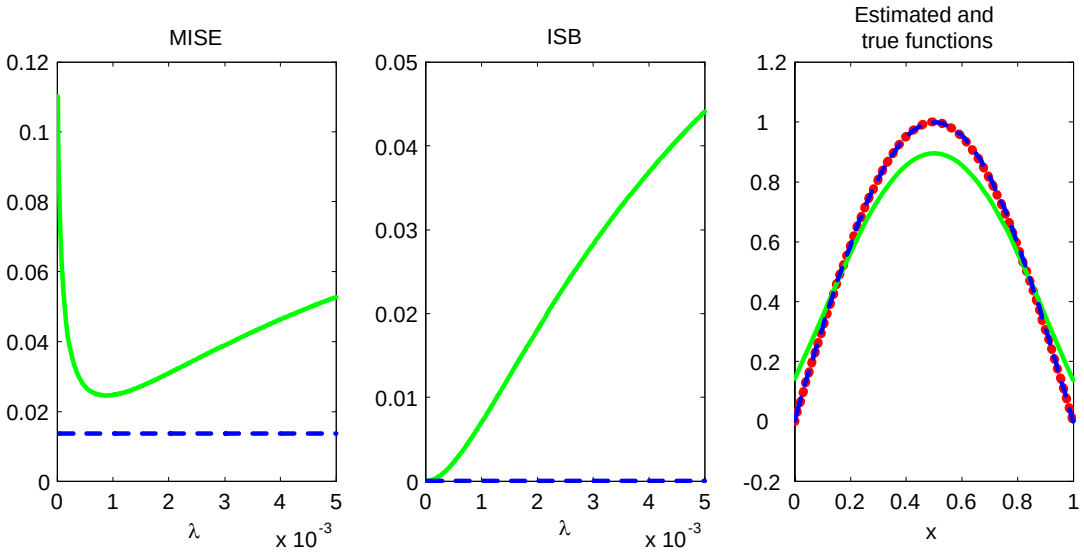

Figure 7: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0$, and sample size is $T = 400$.

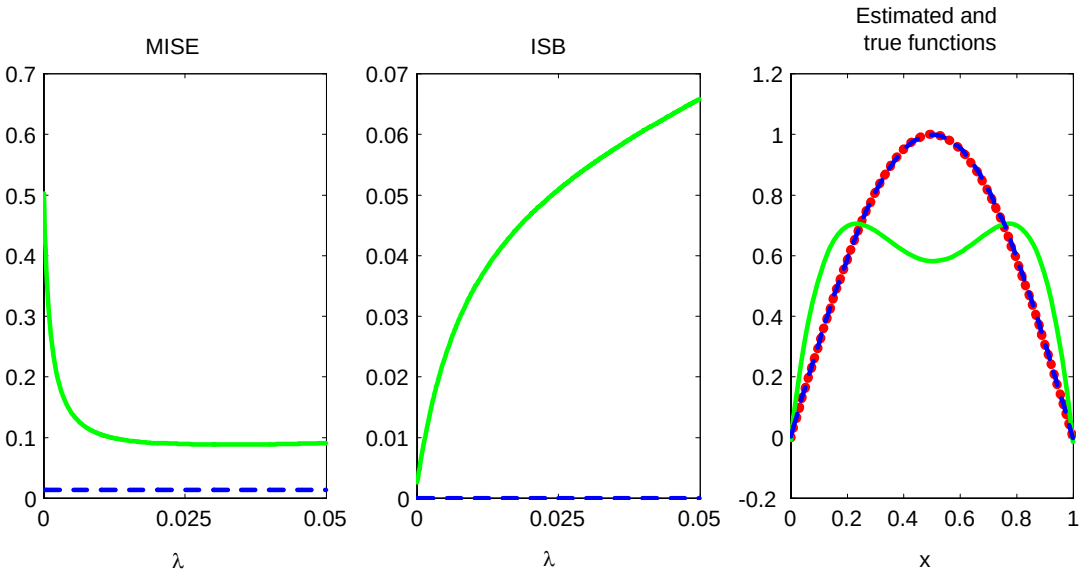

Figure 8: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the regularized estimator using L^2 norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0$, and sample size is $T = 400$.

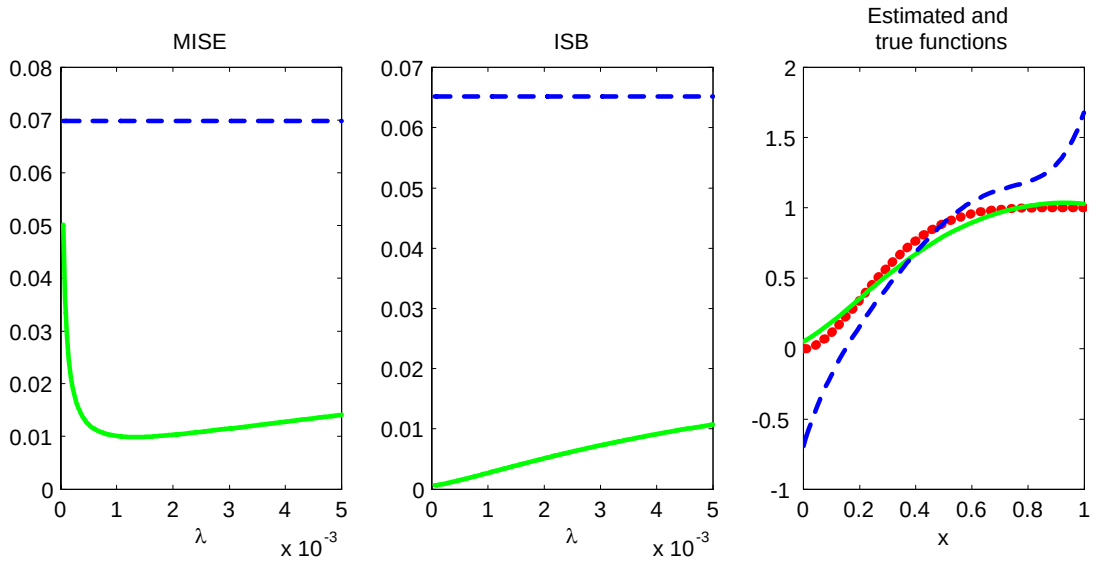

Figure 9: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0.5$, and sample size is $T = 1000$.

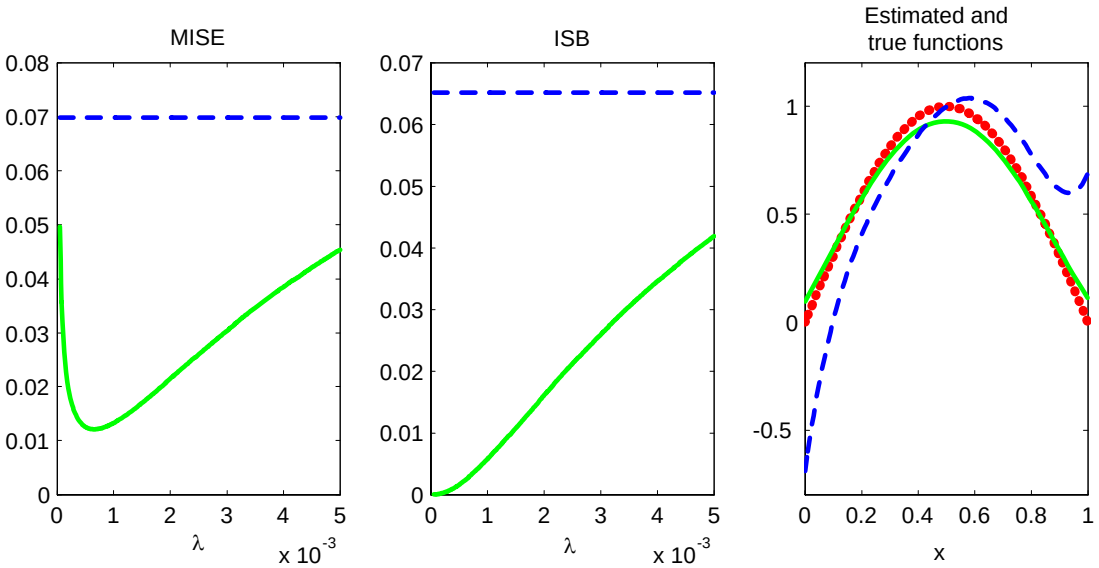

Figure 10: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0.5$, and sample size is $T = 1000$.

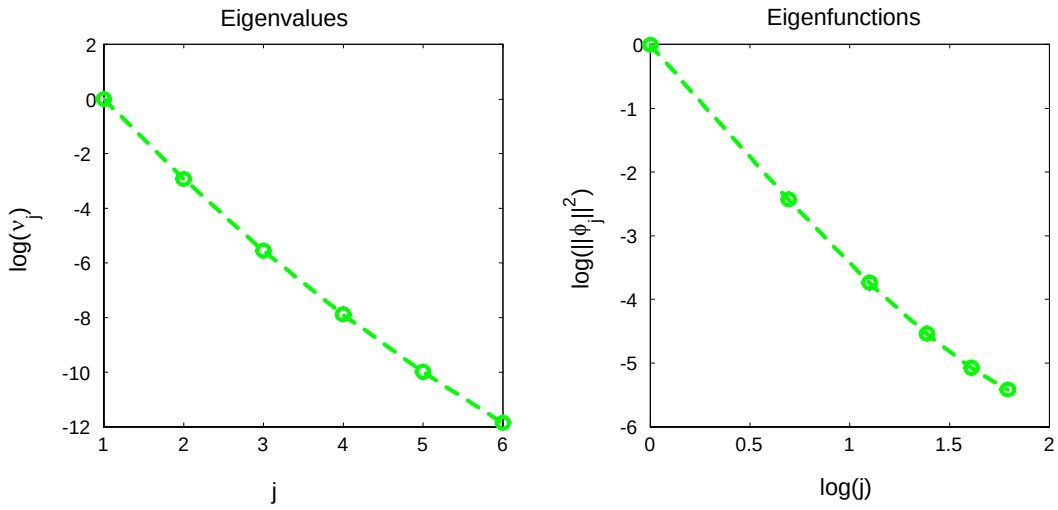

Figure 11: The eigenvalues (left Panel) and the L^2-norms of the corresponding eigenfunctions (right Panel) of operator A^*A using the approximation with six polynomials.

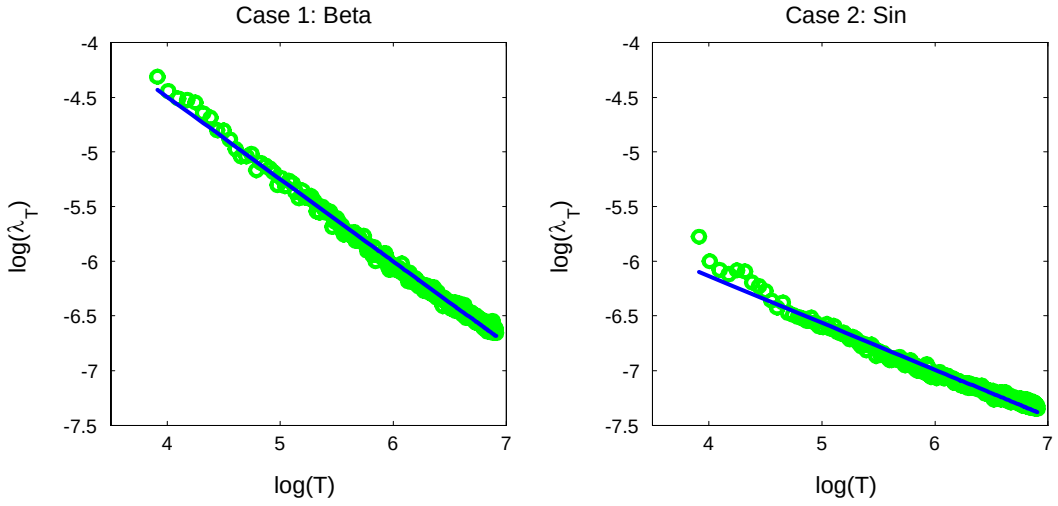

Figure 12: Log of optimal regularization parameter as a function of log of sample size for Case 1 (left panel) and Case 2 (right panel). Correlation parameter is $\rho = 0.5$.

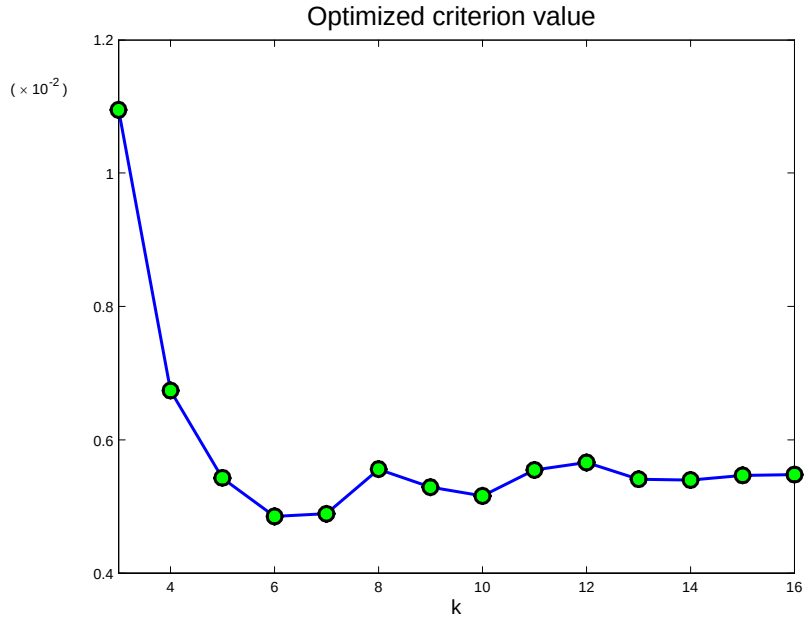

Figure 13: Value of the optimized objective function as a function of the number k of polynomials. The regularization parameter is selected with the spectral approach.

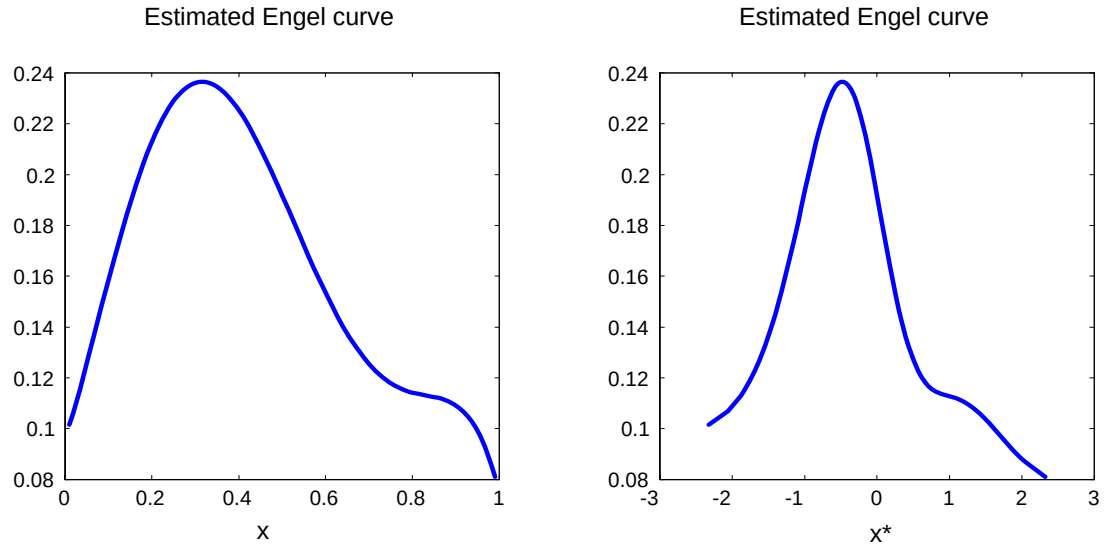

Figure 14: Estimated Engel curves for 785 household-level observations from the 1996 US Consumer Expenditure Survey. In the right Panel, food expenditure share Y is plotted as a function of the standardized logarithm X^* of total expenditures. In the left Panel, Y is plotted as a function of transformed 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.