TIKHONOV REGULARIZATION FOR NONPARAMETRIC INSTRUMENTAL VARIABLE ESTIMATORS

P. Gagliardini[*] and O. Scaillet[†]

This version: August 2011 [‡]

(First version: May 2006)

[*]University of Lugano and Swiss Finance Institute. Corresponding author: Patrick Gagliardini, University of Lugano, Faculty of Economics, Via Buffi 13, CH-6900 Lugano, Switzerland. Tel.: ++ 41 58 666 4660. Fax: ++ 41 58 666 4734. Email: patrick.gagliardini@usi.ch.

[†]Université de Genève and Swiss Finance Institute.

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

Tikhonov Regularization for Nonparametric Instrumental Variable Estimators

Abstract

We study a Tikhonov Regularized (TiR) estimator of a functional parameter identified by conditional moment restrictions in a linear model with both exogenous and endogenous regressors. The nonparametric instrumental variable estimator is based on a minimum distance principle with penalization by the norms of the parameter and its derivatives. After showing its consistency in the Sobolev norm and uniform consistency under an embedding condition, we derive the expression of the asymptotic Mean Integrated Square Error and the rate of convergence. The optimal value of the regularization parameter is characterized in two examples. We illustrate our theoretical findings and the small sample properties with simulation results. Finally, we provide an empirical application to estimation of an Engel curve, and discuss a data driven selection procedure for the regularization parameter.

Keywords and phrases: Nonparametric Estimation, Ill-posed Inverse Problems, Tikhonov Regularization, Endogeneity, Instrumental Variable.

JEL classification: C13, C14, C15, D12.

AMS 2000 classification: 62G08, 62G20.

1 Introduction

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

There is a growing recent literature in econometrics extending the above methods and considering empirical applications. Blundell, Chen and Kristensen (BCK, 2007) investigate application of index models to Engel curve estimation with endogenous total expenditure. As argued, e.g., in Blundell and Horowitz (2007), the knowledge of the shape of an Engel curve is a key ingredient of any consumer behaviour analysis. Chen and Pouzo (2009, 2011) consider a general semiparametric setting including partially linear quantile IV regression, and apply their results to sieve estimation of Engel curves. Further, Chen and Ludvigson (2009) consider asset pricing models with functional specifications of habit preferences; Cher-

nozhukov, Imbens and Newey (2007) estimate nonseparable models for quantile regression analysis; Loubes and Vanhems (2004) discuss the estimation of the solution of a differential equation with endogenous variables for microeconomic applications. Other related works include Chernozhukov and Hansen (2005), Florens, Johannes and Van Bellegem (2005), Horowitz (2006), Hoderlein and Holzmann (2011), and Hu and Schennach (2008).

The main theoretical difficulty in nonparametric estimation with endogeneity is overcoming ill-posedness of the associated inverse problem (see Kress, 1999, and Carrasco, Florens and Renault (CFR), 2007, for overviews). It occurs since the mapping of the reduced form parameter (that is, the distribution of the data) into the structural parameter (the instrumental regression function) is not continuous. We need a regularization of the estimation to recover consistency. For instance, DFFR and HH adopt an L^2 regularization technique resulting in a kind of ridge regression in a functional setting.

The aim of this paper is to introduce a new minimum distance estimator for a functional parameter identified by conditional moment restrictions in a linear model with both exogenous and endogenous regressors. We consider a penalized extremum estimator which minimizes $Q_T(\varphi) + \lambda_T G(\varphi)$, where $Q_T(\varphi)$ is a minimum distance criterion in the functional parameter φ, $G(\varphi)$ is a penalty function, and λ_T is a positive sequence converging to zero. The penalty function $G(\varphi)$ exploits the Sobolev norm of function φ, which involves the L^2 norms of both φ and its derivatives $\nabla^\alpha \varphi$ up to a finite order. The basic idea is that the penalty term $\lambda_T G(\varphi)$ damps highly oscillating components of the estimator. These oscillations are otherwise unduly amplified by the minimum distance criterion $Q_T(\varphi)$ because of

ill-posedness. Parameter λ_T tunes the regularization. We call our estimator a Tikhonov Regularized (TiR) estimator by reference to the pioneering papers of Tikhonov (1963a,b) where regularization is achieved via a penalty term incorporating the function and its derivative (Groetsch, 1984). The TiR estimator admits a closed form and is numerically tractable. Our approach relies on the maintained assumption that the functional parameter lives in some Sobolev space of functions with square integrable derivatives up to a finite order. In many economic applications, differentiability of the parameter of interest is a natural assumption.

The key contribution of our paper is the computation of an explicit asymptotic expression for the mean integrated squared error (MISE) of a Sobolev penalized estimator in an NIVR setting with both exogenous and endogenous regressors. Such a sharp result extends the asymptotic bounds of HH obtained under a L^2 penalty. Our other specific contributions are consistency of the TiR estimator in the Sobolev norm, and as a consequence uniform consistency under an embedding condition, and a detailed analytic treatment of two examples yielding the optimal value of the regularization parameter. The embedding condition states that the order of the Sobolev norm used for penalization is strictly larger than half the number of endogenous regressors.

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

similar to ours. Their estimation relies on series estimators instead of kernel smoothers that we use. Chen and Pouzo (2009, 2011) examine the convergence rate of a sieve approach for an implementation as in BCK.

In defining directly the estimator on a function space, we follow the route of Horowitz and Lee (2007) and the suggestion of NP, p. 1573 (see also Gagliardini and Gouriéroux, 2007, Chernozhukov, Gagliardini, and Scaillet (CGS), 2006). Working directly over an infinite-dimensional parameter space (and not over finite-dimensional parameter spaces of increasing dimensions) allows us to develop a well-defined theoretical framework which uses the penalization parameter as the single regularization parameter. In a sieve approach, either the number of sieve terms, or both the number of sieve terms and the penalization coefficient, are regularization parameters that need to be controlled (see Chen and Pouzo, 2009, 2011, for a detailed treatment). As in the implementation of a sieve approach, our computed estimator uses a projection on a finite-dimensional basis of polynomials. The approximation error is of a purely numerical nature, and not of a statistical nature as in a sieve approach where the number of sieve terms can be used as a regularization parameter. The dimension of the basis should be selected sufficiently large to get a small approximation error. In some cases, for example when the parameter of interest is close to a line, a few basis functions are enough to successfully implement our approach. We cannot see our approach as a sieve approach with an infinite number of terms, and both asymptotic theoretical treatments do not nest each other (see CGS for similar comments in the quantile regression case). However we expect an asymptotic equivalence between our approach and a sieve approach under a

number of sieve terms growing sufficiently fast to dominate the decay of the penalization term. The proof of such an equivalence is left for future research.

While the regularization approach in DFFR and HH can be viewed as a Tikhonov regularization, their penalty term involves the L^2 norm of the function only (without any derivative). By construction this penalization dispenses from a differentiability assumption of the function φ. To avoid confusion, we refer to DFFR and HH estimators as regularized estimators with L^2 norm. In our Monte-Carlo experiments and in an analytic example, we find that the use of the Sobolev penalty substantially enhances the performance of the regularized estimator relative to the use of the L^2 penalty. Another advantage of a Sobolev penalty is in the proof of uniform consistency when embedding holds. Finally CGS focus on a feasible asymptotic normality theorem for a TiR estimator in an NIVQR setting. Their results can be easily specialized to the linear setting of this paper, and are not further considered here.

In Section 2 we discuss ill-posedness in nonparametric IV regression. We introduce the TiR estimator in Section 3. Its consistency in Sobolev norm is proved in Section 4. We further establish uniform consistency under an embedding condition and discuss the convergence rate. In Section 5, we derive the exact asymptotic MISE of the TiR estimator. In Section 6 we discuss optimal rates of convergence in two examples, and provide an analytic comparison with L^2 regularization. We discuss the numerical implementation in Section 7, and we present the Monte-Carlo results in Section 8. In Section 9 we provide an empirical example where we estimate an Engel curve nonparametrically, and discuss a data driven selection procedure for the regularization parameter. Gagliardini and Scaillet (GS, 2006)

give further simulation results and implementation details. The set of regularity conditions and the proofs of propositions are gathered in the Appendices. Omitted proofs of technical Lemmas are collected in a Technical Report, which is available online at our web pages.

2 Ill-posedness in nonparametric regression

Let $\{(Y_t, X_t, Z_t) : t = 1, ..., T\}$ be i.i.d. copies of vector (Y, X, Z), where vectors X and Z are decomposed as $X := (X_1, X_2)$ and $Z := (Z_1, X_1)$. Let the supports of X and Z be $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ and $\mathcal{Z} := \mathcal{Z}_1 \times \mathcal{X}_1$, where $\mathcal{X}_i := [0,1]^{d_{X_i}}$, $i = 1, 2$, and $\mathcal{Z}_1 = [0,1]^{d_{Z_1}}$, while the support of Y is $\mathcal{Y} \subset \mathbb{R}$. The parameter of interest is a function φ_0 defined on $\mathcal{X}$ which satisfies the NIVR:

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

The subvectors X_1 and X_2 correspond to exogenous and endogenous regressors, while Z is a vector of instruments. The conditional moment restriction (1) is equivalent to:

$$m_{x_1}\left(\varphi_{x_1,0}, Z_1\right) := E\left[Y - \varphi_{x_1,0}(X_2) \mid Z_1, X_1 = x_1\right] = 0, \text{ for } \textit{all } x_1 \in \mathcal{X}_1,$$

where $\varphi_{x_1,0}(.) := \varphi_0(x_1, .)$. For any given $x_1 \in \mathcal{X}_1$, the function $\varphi_{x_1,0}$ satisfies a NIVR with endogenous regressors X_2 only. Parameter φ_0 is such that, for all $x_1 \in \mathcal{X}_1$, the function $\varphi_{x_1,0}$ belongs to the Sobolev space $H^l(\mathcal{X}_2)$ of order $l \in \mathbb{N}$, i.e., the completion of the linear space $\left\{\psi \in C^l(\mathcal{X}_2) \mid \nabla^\alpha \psi \in L^2(\mathcal{X}_2), |\alpha| \leq l\right\}$ with respect to the scalar product $\langle \psi_1, \psi_2 \rangle_{H^l(\mathcal{X}_2)} := \sum_{|\alpha| \leq l} \langle \nabla^\alpha \psi_1, \nabla^\alpha \psi_2 \rangle_{L^2(\mathcal{X}_2)}$, where $\langle \psi_1, \psi_2 \rangle_{L^2(\mathcal{X}_2)} := \int_{\mathcal{X}_2} \psi_1(u)\psi_2(u)du$ and $\alpha \in \mathbb{N}^{d_{X_2}}$ is a multi-index. The Sobolev space $H^l(\mathcal{X}_2)$ is an Hilbert space w.r.t. the scalar product $\langle \psi_1, \psi_2 \rangle_{H^l(\mathcal{X}_2)}$,

and the corresponding Sobolev norm is denoted by $\|\psi\|_{H^l(\mathcal{X}_2)} := \langle \psi, \psi \rangle^{1/2}_{H^l(\mathcal{X}_2)}$. We denote the L^2 norm by $\|\psi\|_{L^2(\mathcal{X}_2)} := \langle \psi, \psi \rangle^{1/2}_{L^2(\mathcal{X}_2)}$. The Sobolev embedding theorem (see Adams and Fournier (2003), Theorem 4.12) states that the Sobolev space $H^l(\mathcal{X}_2)$ is embedded in the space of continuous functions on $\mathcal{X}_2$ equipped with the sup norm, and hence in space $L^2(\mathcal{X}_2)$ as well, when $2l > d_{X_2}$. This implies that $\|\psi\|_{L^2(\mathcal{X}_2)} \leq \sup_{x_2 \in \mathcal{X}_2} |\psi(x_2)| \leq C \|\psi\|_{H^l(\mathcal{X}_2)}$ for $\psi \in H^l(\mathcal{X}_2)$ and a constant C, when $2l > d_{X_2}$. With a single endogenous regressor ($d_{X_2} = 1$), the embedding condition is satisfied for any l.

We assume the following identification condition.

Assumption 1: *$\varphi_{x_1,0}$ is the unique function $\varphi_{x_1} \in H^l(\mathcal{X}_2)$ that satisfies the conditional moment restriction $m_{x_1}(\varphi_{x_1}, Z_1) = 0$, for all $x_1 \in \mathcal{X}_1$.*

We refer to NP, Theorems 2.2-2.4, for sufficient conditions ensuring Assumption 1. We work below with a penalized quadratic criterion in the parameter of interest, which yields a closed form expression for the estimator. Hence, we do not need to restrict the parameter set by imposing further assumptions, such as boundedness or compactness. See Chen (2007), Horowitz and Lee (2007), and Chen and Pouzo (2009, 2011) for similar noncompact settings.

Let us now consider a given $x_1 \in \mathcal{X}_1$ and a nonparametric minimum distance approach for $\varphi_{x_1,0}$. This relies on $\varphi_{x_1,0}$ minimizing

$$Q_{x_1,\infty}(\varphi_{x_1}) := E\left[\Omega_{x_1,0}(Z_1) m_{x_1}(\varphi_{x_1}, Z_1)^2 \mid X_1 = x_1\right], \quad \varphi_{x_1} \in H^l(\mathcal{X}_2), \tag{2}$$

where $\Omega_{x_1,0}$ is a positive function on $\mathcal{Z}_1$. The conditional moment function $m_{x_1}(\varphi_{x_1}, z_1)$ can

be written as:

$$m_{x_1}\left(\varphi_{x_1}, z_1\right) = \left(A_{x_1}\varphi_{x_1}\right)(z_1) - r_{x_1}(z_1) = \left(A_{x_1}\Delta\varphi_{x_1}\right)(z_1), \tag{3}$$

where $\Delta\varphi_{x_1} := \varphi_{x_1} - \varphi_{x_1,0}$, linear operator A_{x_1} is defined by $\left(A_{x_1}\varphi_{x_1}\right)(z_1) := \int \varphi_{x_1}(x_2) f_{X_2|Z}(x_2|z)dx_2$ and $r_{x_1}(z_1) := \int y f_{Y|Z}(y|z)dy$, where $f_{X_2|Z}$ and $f_{Y|Z}$ are the conditional densities of X_2 given Z, and Y given Z. Assumption 1 on identification of $\varphi_{x_1,0}$ holds if and only if operator A_{x_1} is injective for all $x_1 \in \mathcal{X}_1$. Further, we assume that A_{x_1} is a bounded operator from $L^2(\mathcal{X}_2)$ to $L^2_{x_1}(\mathcal{Z}_1)$, where $L^2_{x_1}(\mathcal{Z}_1)$ denotes the L^2 space of square integrable functions of Z_1 defined by scalar product $\langle \psi_1, \psi_2 \rangle_{L^2_{x_1}(\mathcal{Z}_1)} = E\left[\Omega_{x_1,0}(Z_1)\psi_1(Z_1)\psi_2(Z_1)|X_1 = x_1\right]$. The limit criterion (2) becomes

$$\begin{aligned} Q_{x_1,\infty}(\varphi_{x_1}) &= \langle A_{x_1}\Delta\varphi_{x_1}, A_{x_1}\Delta\varphi_{x_1}\rangle_{L^2_{x_1}(\mathcal{Z}_1)} \\ &= \langle \Delta\varphi_{x_1}, A^*_{x_1}A_{x_1}\Delta\varphi_{x_1}\rangle_{H^l(\mathcal{X}_2)} = \langle \Delta\varphi_{x_1}, \tilde{A}_{x_1}A_{x_1}\Delta\varphi_{x_1}\rangle_{L^2(\mathcal{X}_2)}, \end{aligned} \tag{4}$$

where $A^*_{x_1}$, resp. $\tilde{A}_{x_1}$, denotes the adjoint operator of A_{x_1} w.r.t. the scalar products $\langle ., . \rangle_{H^l(\mathcal{X}_2)}$, resp. $\langle ., . \rangle_{L^2(\mathcal{X}_2)}$, and $\langle ., . \rangle_{L^2_{x_1}(\mathcal{Z}_1)}$.

Assumption 2: *The linear operator A_{x_1} from $L^2(\mathcal{X}_2)$ to $L^2_{x_1}(\mathcal{Z}_1)$ is compact for all $x_1 \in \mathcal{X}_1$.*

Assumption 2 on compactness of operator A_{x_1} holds under mild conditions on the conditional density $f_{X_2|Z}$ and the weighting function $\Omega_{x_1,0}$ (see Assumptions B.3 (i) and B.6 in Appendix 1). Then, operator $A^*_{x_1}A_{x_1}$ is compact and self-adjoint in $H^l(\mathcal{X}_2)$, while $\tilde{A}_{x_1}A_{x_1}$ is compact and self-adjoint in $L^2(\mathcal{X}_2)$. We denote by $\left\{\phi_{x_1,j} : j \in \mathbb{N}\right\}$ an orthonormal basis in $H^l(\mathcal{X}_2)$ of eigenfunctions of operator $A^*_{x_1}A_{x_1}$, and by $\nu_{x_1,1} \geq \nu_{x_1,2} \geq \cdots > 0$ the corresponding eigenvalues (see Kress, 1999, Section 15.3, for the spectral decomposition of compact, self-adjoint

operators). Similarly, $\left\{\tilde{\phi}_{x_1,j} : j \in \mathbb{N}\right\}$ is an orthonormal basis in $L^2(\mathcal{X}_2)$ of eigenfunctions of operator $\tilde{A}_{x_1}A_{x_1}$ for eigenvalues $\tilde{\nu}_{x_1,1} \geq \tilde{\nu}_{x_1,2} \geq \cdots > 0$. By compactness of $A^*_{x_1}A_{x_1}$ and $\tilde{A}_{x_1}A_{x_1}$, the eigenvalues are such that $\nu_{x_1,j}, \tilde{\nu}_{x_1,j} \to 0$, as $j \to \infty$, for any given $x_1 \in \mathcal{X}_1$. Moreover, under Assumptions B.3 (i) and B.6, $\tilde{\phi}_{x_1,j} \in H^l(\mathcal{X}_2)$ for any j. Then, the limit criterion $Q_{x_1,\infty}(\varphi_{x_1})$ can be minimized by a sequence $\varphi_{x_1,n}$ in $H^l(\mathcal{X}_2)$ such that

$$\varphi_{x_1,n} = \varphi_{x_1,0} + \varepsilon\tilde{\phi}_{x_1,n}, \quad n \in \mathbb{N}, \tag{5}$$

for $\varepsilon > 0$, which does not converge to $\varphi_{x_1,0}$ in L^2-norm $\|.\|_{L^2(\mathcal{X}_2)}$. Indeed, we have $Q_{x_1,\infty}(\varphi_{x_1,n}) = \varepsilon^2\langle\tilde{\phi}_{x_1,n}, \tilde{A}_{x_1}A_{x_1}\tilde{\phi}_{x_1,n}\rangle_{L^2(\mathcal{X}_2)} = \varepsilon^2\tilde{\nu}_{x_1,n} \to 0$ as $n \to \infty$, but $\left\|\varphi_{x_1,n} - \varphi_{x_1,0}\right\|_{L^2(\mathcal{X}_2)} = \varepsilon$, $\forall n$. Since $\varepsilon > 0$ is arbitrary, the usual "identifiable uniqueness" assumption (e.g., White and Wooldridge (1991))

$$\inf_{\varphi_{x_1} \in H^l(\mathcal{X}_2): R \geq \left\|\varphi_{x_1} - \varphi_{x_1,0}\right\|_{L^2(\mathcal{X}_2)} \geq \varepsilon} Q_{x_1,\infty}(\varphi_{x_1}) > 0 = Q_{x_1,\infty}(\varphi_{x_1,0}), \text{ for } R > \varepsilon > 0, \tag{6}$$

is *not* satisfied. In other words, function $\varphi_{x_1,0}$ is not identified as an isolated minimum of $Q_{x_1,\infty}$. This is the identification problem of minimum distance estimation with functional parameter and endogenous regressors. Failure of Condition (6) despite validity of Assumption 1 comes from 0 being a limit point of the eigenvalues of operator $\tilde{A}_{x_1}A_{x_1}$ (and $A^*_{x_1}A_{x_1}$). This shows that the minimum distance problem for any given $x_1 \in \mathcal{X}_1$ is ill-posed. The minimum distance estimator of $\varphi_{x_1,0}$ which minimizes the empirical counterpart of criterion $Q_{x_1,\infty}(\varphi_{x_1})$ over $H^l(\mathcal{X}_2)$, or some bounded noncompact subset of it, is not consistent w.r.t. the L^2-norm $\|.\|_{L^2(\mathcal{X}_2)}$.

To conclude this section, let us further discuss the link between function φ_0 and func-

tions $\varphi_{x_1,0}$, $x_1 \in \mathcal{X}_1$. First, $\varphi_0 \in L^2(\mathcal{X})$. Indeed, the set $\mathcal{P} := \{\varphi : \varphi_{x_1} \in H^l(\mathcal{X}_2), \forall x_1 \in \mathcal{X}_1, \sup_{x_1 \in \mathcal{X}_1} \|\varphi_{x_1}\|_{H^l(\mathcal{X}_2)} < \infty\}$ is a subset of $L^2(\mathcal{X})$, since $\|\varphi\|^2_{L^2(\mathcal{X})} = \int_{\mathcal{X}_1} \|\varphi_{x_1}\|^2_{L^2(\mathcal{X}_2)} dx_1$. Second, Assumption 1 implies identification of $\varphi_0 \in \mathcal{P}$. Third, minimizing $Q_{x_1,\infty}$ w.r.t. $\varphi_{x_1} \in H^l(\mathcal{X}_2)$ for all $x_1 \in \mathcal{X}_1$ is equivalent to minimizing the global criterion $Q_\infty(\varphi) := E\left[\Omega_0(Z) m(\varphi, Z)^2\right] = E\left[Q_{X_1,\infty}(\varphi_{X_1})\right]$, w.r.t. $\varphi \in \mathcal{P}$, where $m(\varphi, z) := E[Y - \varphi(X) \mid Z = z]$ and $\Omega_0(z) = \Omega_{x_1,0}(z_1)$. Under Assumptions B.3 (i), ill-posedness of the minimum distance approach for φ_{x_1}, $x_1 \in \mathcal{X}_1$, transfers by Lebesgue theorem to ill-posedness of the minimum distance approach for φ. Indeed, the sequence φ_n induced by (5) yields $Q_\infty(\varphi_n) \to 0$ and $\varphi_n \nrightarrow \varphi_0$ as $n \to \infty$. Compactness (Assumption 2) cannot hold for the conditional expectation operator of X given Z per se. Indeed, as discussed in DFFR, this operator is not compact in the presence of exogenous regressors treated as random variables and not fixed values. This explains why we work x_1 by x_1 as in HH to estimate φ_0. Finally, we assume the following uniform behaviour needed to show the asymptotic properties of the estimator.

Assumption 3: *The true function φ_0 satisfies* $\sup_{x_1 \in \mathcal{X}_1} \|\varphi_{x_1,0}\|_{H^l(\mathcal{X}_2)} < \infty$.

3 The Tikhonov Regularized (TiR) estimator

We address ill-posedness by Tikhonov regularization (Tikhonov, 1963a,b; see Kress, 1999, Chapter 16). We consider a penalized criterion $L_{x_1,T}(\varphi_{x_1}) := Q_{x_1,T}(\varphi_{x_1}) + \lambda_{x_1,T} \|\varphi_{x_1}\|^2_{H^l(\mathcal{X}_2)}$, where $Q_{x_1,T}(\varphi_{x_1})$ is an empirical counterpart of $Q_{x_1,\infty}(\varphi_{x_1})$ defined by

$$Q_{x_1,T}(\varphi_{x_1}) = \int_{\mathcal{Z}_1} \hat{\Omega}_{x_1}(z_1)\ \hat{m}_{x_1}(\varphi_{x_1}, z_1)^2\ \hat{f}_{Z_1|X_1}(z_1|x_1) dz_1, \tag{7}$$

and $\hat{\Omega}_{x_1}$ is a sequence of positive functions converging in probability to $\Omega_{x_1,0}$. In (7) we estimate the conditional moment nonparametrically with

$$\hat{m}_{x_1}\left(\varphi_{x_1}, z_1\right) = \int \varphi_{x_1}\left(x_2\right) \hat{f}_{X_2|Z}\left(x_2|z\right) dx_2 - \int y \hat{f}_{Y|Z}\left(y|z\right) dy =: \left(\hat{A}_{x_1}\varphi_{x_1}\right)\left(z_1\right) - \hat{r}_{x_1}\left(z_1\right),$$

where $\hat{f}_{X_2|Z}$ and $\hat{f}_{Y|Z}$ denote kernel estimators of the density of X_2 given Z, and Y given Z. We use a common kernel K and two different bandwidths h_T for Y, X_2, Z_1, and $h_{x_1,T}$ for X_1.

Definition 1: *The Tikhonov Regularized (TiR) minimum distance estimator for $\varphi_{x_1,0}$ is defined by*

$$\hat{\varphi}_{x_1} := \underset{\varphi_{x_1} \in H^l(\mathcal{X}_2)}{\operatorname{arginf}} L_{x_1,T}(\varphi_{x_1}), \tag{8}$$

where $\lambda_{x_1,T} > 0$ and $\lambda_{x_1,T} \to 0$, for any $x_1 \in \mathcal{X}_1$. The TiR estimator $\hat{\varphi}$ for φ_0 is defined by $\hat{\varphi}\left(x\right) := \hat{\varphi}_{x_1}(x_2)$, $x \in \mathcal{X}$.

To emphasize the difference between $\hat{\varphi}_{x_1}$ for a given $x_1 \in \mathcal{X}_1$, and $\hat{\varphi}$, we refer to the former as a *local* estimator, and to the latter as a *global* estimator.

To get the intuition on why advocating the Sobolev norm as a penalty, let us consider the case of a single endogenous explanatory variable, i.e. $d_{X_2} = 1$, and let the parameter set be the Sobolev space $H^1\left(\mathcal{X}_2\right)$, i.e. $l = 1$. From the proof of Proposition 1 in CGS, we know that bounded sequences $\left(\varphi_{x_1,n}\right)$ such that $Q_{x_1,\infty}(\varphi_{x_1,n}) \to 0$ and $\varphi_{x_1,n} \nrightarrow \varphi_{x_1,0}$ have the property $\limsup_{n\to\infty} \left\|\nabla\varphi_{x_1,n}\right\|_{L^2(\mathcal{X}_2)} = \infty$. This explains why we prefer in definition (8) to use a Sobolev penalty $\lambda_{x_1,T} \left\|\varphi_{x_1}\right\|^2_{H^1(\mathcal{X}_2)}$ instead of an L^2 penalty $\lambda_{x_1,T} \left\|\varphi_{x_1}\right\|^2_{L^2(\mathcal{X}_2)}$ to dampen the highly oscillating components in the estimated function. Without penalization oscillations are

unduly amplified, since ill-posedness yields a criterion $Q_{x_1,T}(\varphi_{x_1})$ asymptotically flat along some directions. This intuition generalizes to the case where $l > 2d_{X_2}$ because of the Sobolev embedding theorem (see Section 2), and sequences $\left(\varphi_{x_1,n}\right)$ such that $Q_{x_1,\infty}(\varphi_{x_1,n}) \to 0$ and $\varphi_{x_1,n} \nrightarrow \varphi_{x_1,0}$ have the property $\limsup_{n\to\infty} \left\|\varphi_{x_1,n}\right\|_{H^l(\mathcal{X}_2)} = \infty$. The tuning parameter $\lambda_{x_1,T}$ in Definition 1 controls for the amount of regularization, and how this depends on point x_1 and sample size T. Its rate of convergence to zero affects the one of $\hat{\varphi}_{x_1}$.

The TiR estimator admits a closed form expression. The objective function in (8) can be rewritten as (see Lemma A.2 (i) in Appendix 2)

$$L_{x_1,T}(\varphi_{x_1}) = \langle \varphi_{x_1}, \hat{A}^*_{x_1}\hat{A}_{x_1}\varphi_{x_1}\rangle_{H^l(\mathcal{X}_2)} - 2\langle \varphi_{x_1}, \hat{A}^*_{x_1}\hat{r}_{x_1}\rangle_{H^l(\mathcal{X}_2)} + \lambda_{x_1,T}\langle \varphi_{x_1}, \varphi_{x_1}\rangle_{H^l(\mathcal{X}_2)}, \qquad (9)$$

up to a term independent of φ_{x_1}, where operator $\hat{A}^*_{x_1}$ is such that

$$\left\langle \varphi, \hat{A}^*_{x_1}\psi \right\rangle_{H^l(\mathcal{X}_2)} = \int \hat{\Omega}_{x_1}(z_1) \left(\hat{A}_{x_1}\varphi\right)(z_1)\, \psi(z_1)\, \hat{f}_{Z_1|X_1}(z_1|x_1)\, dz_1$$

for any $\varphi \in H^l(\mathcal{X}_2)$ and $\psi \in L^2_{x_1}(\mathcal{Z}_1)$. When $l = 1$ we have

$$\hat{A}^*_{x_1} = \mathcal{D}^{-1}\widetilde{\hat{A}}_{x_1}, \qquad \left(\widetilde{\hat{A}}_{x_1}\psi\right)(x_2) := \int_{\mathcal{Z}_1} \hat{\Omega}_{x_1}(z_1)\, \hat{f}_{X_2,Z_1|X_1}(x_2, z_1|x_1)\psi(z_1)\, dz_1, \qquad (10)$$

where $\mathcal{D}^{-1}$ denotes the inverse of operator $\mathcal{D} : H^2_0(\mathcal{X}_2) \to L^2(\mathcal{X}_2)$ with $\mathcal{D} := 1 - \sum_{i=1}^{d_{X_2}} \nabla^2_i$. Here $H^2_0(\mathcal{X}_2) = \{\psi \in H^2(\mathcal{X}_2) \mid \nabla_i\psi(x_2) = 0 \text{ for } x_{2,i} = 0,\ 1, \text{ and } i = 1, ..., d_{X_2}\}$ is the subspace of $H^2(\mathcal{X}_2)$ consisting of functions with first-order derivatives vanishing on the boundary of $\mathcal{X}_2$. Operators $\hat{A}^*_{x_1}$ and $\widetilde{\hat{A}}_{x_1}$ are the empirical counterparts of $A^*_{x_1}$ and $\tilde{A}_{x_1}$, which are linked by $A^*_{x_1} = \mathcal{D}^{-1}\tilde{A}_{x_1}$. The boundary conditions $\nabla_i\psi(x_2) = 0$ for $x_{2,i} = 0,\ 1$ and $i = 1, ..., d_{X_2}$, in the definition of $H^2_0(\mathcal{X}_2)$ are not restrictive since they concern the estimate $\hat{\varphi}_{x_1}$, whose

properties are studied in L^2 or pointwise, but not the true function $\varphi_{x_1,0}$. The sole purpose of the boundary conditions is to guarantee a unique characterization of operator $\mathcal{D}^{-1}$ yielding the solution of a partial differential equation (PDE; see Appendix 2 for the proof of the above statements and the characterization of $\hat{A}^*_{x_1}$ and $A^*_{x_1}$ in the general case $l \geq 1$). Propositions 1-4 below hold independently whether $\varphi_{x_1,0}$ satisfy these boundary conditions or not (see also Kress (1999), Theorem 16.20). From Lemma A.2 (ii), operator $\hat{A}^*_{x_1}\hat{A}_{x_1}$ is compact, and hence $\lambda_T + \hat{A}^*_{x_1}\hat{A}_{x_1}$ is invertible (Kress (1999), Theorem 3.4). Then, Criterion (9) admits a global minimum $\hat{\varphi}_{x_1}$ on $H^l(\mathcal{X}_2)$, which solves the first order condition

$$\left(\lambda_{x_1,T} + \hat{A}^*_{x_1}\hat{A}_{x_1}\right)\varphi_{x_1} = \hat{A}^*_{x_1}\hat{r}_{x_1}. \tag{11}$$

This is an integro-differential Fredholm equation of Type II (see e.g. Mammen, Linton and Nielsen, 1999, Linton and Mammen, 2005, Gagliardini and Gouriéroux, 2007, Linton and Mammen, 2008, and the survey by CFR for other examples). The transformation of the ill-posed problem (1) in the well-posed estimating equation (11) is induced by the penalty term involving the Sobolev norm. The TiR estimator of $\varphi_{x_1,0}$ is the explicit solution of Equation (11):

$$\hat{\varphi}_{x_1} = \left(\lambda_{x_1,T} + \hat{A}^*_{x_1}\hat{A}_{x_1}\right)^{-1}\hat{A}^*_{x_1}\hat{r}_{x_1}. \tag{12}$$

4 Consistency

Equation (12) can be rewritten as (see Appendix 3):

$$\begin{aligned}\hat{\varphi}_{x_1} - \varphi_{x_1,0} &= \left(\lambda_{x_1,T} + A_{x_1}^* A_{x_1}\right)^{-1} A_{x_1}^* \hat{\psi}_{x_1} + \mathcal{B}_{x_1,T}^r + \left(\lambda_{x_1,T} + A_{x_1}^* A_{x_1}\right)^{-1} A_{x_1}^* \zeta_{x_1} + \mathcal{R}_{x_1,T} \\ &=: \mathcal{V}_{x_1,T} + \mathcal{B}_{x_1,T}^r + \mathcal{B}_{x_1,T}^e + \mathcal{R}_{x_1,T}, \end{aligned} \tag{13}$$

where

$$\begin{aligned}\hat{\psi}_{x_1}(z_1) &:= \int (y - \varphi_{x_1,0}(x_2)) \frac{\hat{f}_{W,Z}(w,z) - E\left[\hat{f}_{W,Z}(w,z)\right]}{f_Z(z)} dw, \\ \zeta_{x_1}(z_1) &:= \int (y - \varphi_{x_1,0}(x_2)) \frac{E\left[\hat{f}_{W,Z}(w,z)\right] - f_{W,Z}(w,z)}{f_Z(z)} dw, \end{aligned} \tag{14}$$

and $W := (Y, X_2) \in \mathcal{W} := \mathcal{Y} \times \mathcal{X}_2$. In Equation (13) the first three terms $\mathcal{V}_{x_1,T}$, $\mathcal{B}_{x_1,T}^r := \left(\lambda_{x_1,T} + A_{x_1}^* A_{x_1}\right)^{-1} A_{x_1}^* A_{x_1} \varphi_{x_1,0} - \varphi_{x_1,0} =: \varphi_{x_1,\lambda} - \varphi_{x_1,0}$, and $\mathcal{B}_{x_1,T}^e$ are the leading terms asymptotically, while $\mathcal{R}_{x_1,T}$ is a remainder term given in (26). The stochastic term $\mathcal{V}_{x_1,T}$ has mean zero and contributes to the variance of the estimator. The deterministic term $\mathcal{B}_{x_1,T}^e$ corresponds to kernel estimation bias. The deterministic term $\mathcal{B}_{x_1,T}^r$ corresponds to the regularization bias in the theory of Tikhonov regularization (Kress, 1999, Groetsch, 1984). Indeed, function $\varphi_{x_1,\lambda}$ minimizes the penalized limit criterion $Q_{x_1,\infty}\left(\varphi_{x_1}\right) + \lambda_{x_1,T} \left\|\varphi_{x_1}\right\|_{H^l(\mathcal{X}_2)}^2$ w.r.t. $\varphi_{x_1} \in H^l(\mathcal{X}_2)$. Thus, $\mathcal{B}_{x_1,T}^r$ is the asymptotic bias term arising from introducing the penalty $\lambda_{x_1,T} \left\|\varphi_{x_1}\right\|_{H^l(\mathcal{X}_2)}^2$ in the criterion. To control $\mathcal{B}_{x_1,T}^r$ we introduce a source condition (see DFFR).

Assumption 4: *The function* $\varphi_{x_1,0}$ *satisfies* $\sum_{j=1}^{\infty} \frac{\left\langle \phi_{x_1,j}, \varphi_{x_1,0} \right\rangle_{H^l(\mathcal{X}_2)}^2}{\nu_{x_1,j}^{2\delta_{x_1}}} < \infty$ *for* $\delta_{x_1} \in (0,1]$.

As in the proof of Proposition 3.11 in CFR, Assumption 4 implies:

$$\left\|\mathcal{B}^r_{x_1,T}\right\|_{H^l(\mathcal{X}_2)} = O\left(\lambda^{\delta_{x_1}}_{x_1,T}\right). \tag{15}$$

By bounding the Sobolev norms of the other terms $\mathcal{V}_{x_1,T}$, $\mathcal{B}^e_{x_1,T}$, and $\mathcal{R}_{x_1,T}$ (see Appendix 3), we get the following consistency result. The relation $a_T \asymp b_T$, for positive sequences a_T and b_T, means that a_T/b_T is bounded away from 0 and ∞ as $T \to \infty$.

Proposition 1: *Let the bandwidths $h_T \asymp T^{-\eta}$ and $h_{x_1,T} \asymp T^{-\eta_{x_1}}$ and the regularization parameter $\lambda_{x_1,T} \asymp T^{-\gamma_{x_1}}$ be such that:*

$$\eta > 0,\ \eta_{x_1} > 0,\ \gamma_{x_1} > 0, \tag{16}$$

$$\gamma_{x_1} + d_{X_1}\eta_{x_1} + (d_{Z_1} + d_{X_2})\,\eta < 1, \tag{17}$$

and:

$$\gamma_{x_1} < \min\left\{m\eta_{x_1}, m\eta, \frac{1 - d_{X_1}\eta_{x_1} - \eta\max\{d_{Z_1}, d_{X_2}\}}{2}\right\}, \tag{18}$$

where $m \geq 2$ is the order of differentiability of the joint density of (W, Z). Under Assumptions 1-4 and B.1-B.3, B.6, B.7 (i)-(ii): $\left\|\hat{\varphi}_{x_1} - \varphi_{x_1,0}\right\|_{H^l(\mathcal{X}_2)} = o_p(1)$.

Proposition 1 shows that the powers γ_{x_1}, η_{x_1}, and η need to be sufficiently small for large dimensions d_{X_1}, d_{X_2}, and d_Z and small order of differentiability m to ensure consistency. An analysis of γ_{x_1}, η_{x_1}, and η close to the origin reveals that conditions (16)-(18) are not mutually exclusive, and that these conditions do not yield an empty region. Consistency of $\hat{\varphi}_{x_1}$ in the Sobolev norm $H^l(\mathcal{X}_2)$ implies consistency of both $\hat{\varphi}_{x_1}$ and $\nabla^\alpha\hat{\varphi}_{x_1}$ for $|\alpha| \leq l$ in

the norm $L^2(\mathcal{X}_2)$. When $2l > d_{X_2}$, the Sobolev embedding theorem (see Section 2) implies uniform consistency of $\hat{\varphi}_{x_1}$, i.e., $\sup_{x_2 \in \mathcal{X}_2} |\hat{\varphi}_{x_1}(x_2) - \varphi_{x_1,0}(x_2)| = o_p(1)$, for a given $x_1 \in \mathcal{X}_1$. When $d_{X_2} = 1$ (single endogenous regressor) uniform consistency is valid for any order l since the embedding condition is always satisfied. Uniform consistency without the embedding condition, i.e., when $2l \leq d_{X_2}$, is a conjecture left for future research.

Building on the bounds for terms $\mathcal{V}_{x_1,T}$, $\mathcal{B}^e_{x_1,T}$, and $\mathcal{R}_{x_1,T}$ in the proof of Proposition 1, we can further derive a result on the consistency rate uniformly in $x_1 \in \mathcal{X}_1$ if we introduce a strengthening of the source condition.

Assumption 4 bis: *The function* φ_0 *satisfies* $\sup_{x_1 \in \mathcal{X}_1} \sum_{j=1}^{\infty} \frac{\left\langle \phi_{x_1,j}, \varphi_{x_1,0} \right\rangle^2_{H^l(\mathcal{X}_2)}}{\nu_{x_1,j}^{2\,\delta_{x_1}}} < \infty$, *for* $\delta_{x_1} \in (0,1]$ *with* $x_1 \in \mathcal{X}_1$, *and* $\underline{\delta} := \inf_{x_1 \in \mathcal{X}_1} \delta_{x_1} > 0$.

Assumption 4 bis implies:

$$\sup_{x_1 \in \mathcal{X}_1} \left\| \mathcal{B}^r_{x_1,T} \right\|^2_{H^l(\mathcal{X}_2)} = O\left(\sup_{x_1 \in \mathcal{X}_1} \lambda_{x_1,T}^{2\delta_{x_1}} \right), \tag{19}$$

and we get the next uniform consistency result.

Proposition 2: *Let the bandwidths* $h_T \asymp T^{-\eta}$ *and* $h_{x_1,T} \asymp T^{-\eta_{x_1}}$ *and the regularization parameter* $\lambda_{x_1,T} \asymp T^{-\gamma_{x_1}}$ *be such that* $\underline{\eta} \leq \eta_{x_1} \leq \bar{\eta}$ *and* $\underline{\gamma} \leq \gamma_{x_1} \leq \bar{\gamma}$ *for all* $x_1 \in \mathcal{X}_1$, *where* $\eta, \underline{\eta}, \underline{\gamma} > 0$, $\bar{\gamma} + d_{X_1}\bar{\eta} + (d_{Z_1} + d_{X_2})\,\eta < 1$, *and* $\bar{\gamma} < \min\left\{ m\underline{\eta}, m\eta, \frac{1 - d_{X_1}\bar{\eta} - \eta \max\{d_{Z_1}, d_{X_2}\}}{2} \right\}$. *Under Assumptions 1-4 bis and B.1-B.3, B.6, B.7 (i)-(ii):* $\sup_{x_1 \in \mathcal{X}_1} \left\| \hat{\varphi}_{x_1} - \varphi_{x_1,0} \right\|_{H^l(\mathcal{X}_2)} = O_p\left((\log T)\, T^{-\varkappa} \right)$ *where* $\varkappa > 0$ *is given by*

$$\varkappa = \min\left\{ \frac{1 - d_{X_1}\bar{\eta} - \bar{\gamma} - \min\{\bar{\gamma}, d_{Z_1}\eta\}}{2}, 1 - d_{X_1}\bar{\eta} - \bar{\gamma} - \eta\,(d_{Z_1} + d_{X_2}), m \min\left\{\underline{\eta}, \eta\right\} - \frac{\bar{\gamma}}{2}, \underline{\delta}\underline{\gamma} \right\}.$$

Again from the Sobolev embedding theorem, when $2l > d_{X_2}$, Proposition 2 yields a uniform consistency rate of the global estimator $\hat{\varphi}$: $\sup_{x\in\mathcal{X}} |\hat{\varphi}(x)-\varphi_0(x)| = O_p\left((\log T)\, T^{-\varkappa}\right)$. This in turn implies the L^2-consistency rate $\|\hat{\varphi} - \varphi_0\|_{L^2(\mathcal{X})} = O_p\left((\log T)\, T^{-\varkappa}\right)$.

5 Mean Integrated Square Error

As in AC, Assumption 4.1, we assume the following choice of the weighting matrix.

Assumption 5: *The asymptotic weighting matrix is* $\Omega_0(z) = V\left[Y - \varphi_0\left(X\right) \mid Z = z\right]^{-1}$.

In a semiparametric setting, AC show that this choice of the weighting matrix yields efficient estimators of the finite-dimensional component. Here, Assumption 5 is used to derive the exact asymptotic expansion of the MISE of the TiR estimator provided in the next proposition.

Proposition 3: *Under Assumptions 1-5, Assumptions B, the conditions (16)-(18) and*

$$\frac{1}{T h_{x_1,T}^{d_{X_1}} h_T^{d_{Z_1}+d_{X_2}}} + h_{x_1,T}^{2m} + h_T^{2m} = o\left(\lambda_{x_1,T} b\left(\lambda_{x_1,T}, h_{x_1,T}\right)\right), \quad \frac{h_T h_{x_1,T}^{m-1} + h_T^m}{\sqrt{\lambda_{x_1,T}}} = o\left(b(\lambda_{x_1,T}, h_{x_1,T})\right), \tag{20}$$

the MISE of $\hat{\varphi}_{x_1}$ *is given by*

$$E\left[\left\|\hat{\varphi}_{x_1} - \varphi_{x_1,0}\right\|^2_{L^2(\mathcal{X}_2)}\right] = M_{x_1,T}(\lambda_{x_1,T}, h_{x_1,T})(1 + o(1)), \tag{21}$$

where

$$M_{x_1,T}(\lambda_{x_1,T}, h_{x_1,T}) := \frac{1}{T h_{x_1,T}^{d_{X_1}}} \sigma^2_{x_1}(\lambda_{x_1,T}) + b_{x_1}\left(\lambda_{x_1,T}, h_{x_1,T}\right)^2, \tag{22}$$

and:

$$\sigma^2_{x_1}(\lambda_{x_1,T}) := \omega^2 f_{X_1}\left(x_1\right) \sum_{j=1}^{\infty} \frac{\nu_{x_1,j}}{\left(\lambda_{x_1,T} + \nu_{x_1,j}\right)^2} \left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)},$$

$$b_{x_1}\left(\lambda_{x_1,T}, h_{x_1,T}\right) := \left\| \mathcal{B}^r_{x_1,T} + h^m_{x_1,T}\left(\lambda_{x_1,T} + A^*_{x_1}A_{x_1}\right)^{-1} A^*_{x_1}\Xi_{x_1} \right\|_{L^2(\mathcal{X}_2)},$$

with $\omega^2 = \int K(x_1)^2 dx_1$ *and* $\Xi_{x_1}(z_1) := \dfrac{1}{m!}\sum_{|\alpha|=m}\int (y - \varphi_{x_1,0}(x_2))\dfrac{\nabla^\alpha_{X_1} f_{W,Z}(w,z)}{f_Z(z)}dw.$

Proof: See Appendix 3.

The asymptotic expansion (22) of the MISE consists of one bias component and one variance component which we comment on.

(i) The bias function $b_{x_1}\left(\lambda_{x_1,T}, h_{x_1,T}\right)$ is the L^2 norm of the sum of two contributions, namely the Tikhonov regularization bias $\mathcal{B}^r_{x_1,T}$ and function $h^m_{x_1,T}\left(\lambda_{x_1,T} + A^*_{x_1}A_{x_1}\right)^{-1} A^*_{x_1}\Xi_{x_1}$. The latter contribution corresponds to a population Tikhonov regression applied to function $h^m_{x_1,T}\Xi_{x_1}$. Function $h^m_{x_1,T}\Xi_{x_1}$ arises from smoothing the exogenous regressors X_1 and is derived by a standard Taylor expansion w.r.t. X_1 of the kernel estimation bias $E\left[\hat{f}_{W,Z}(w,z)\right] - f_{W,Z}(w,z)$ in $\mathcal{B}^e_{x_1,T}$ (see (14)).

(ii) The variance term is $V_{x_1,T} := \dfrac{1}{Th^{d_{X_1}}_{x_1,T}}\sigma^2_{x_1}(\lambda_{x_1,T})$. The ratio $1/\left(Th^{d_{X_1}}_{x_1,T}\right)$ and the multiplicative factor $\omega^2 f_{X_1}(x_1)$ are standard for kernel regression in dimension d_{X_1} and are induced by smoothing X_1. The coefficient $\sigma^2_{x_1}(\lambda_{x_1,T})$ involves a weighted sum of the regularized inverse eigenvalues $\nu_{x_1,j}/\left(\lambda_{x_1,T} + \nu_{x_1,j}\right)^2$ of operator $A^*_{x_1}A_{x_1}$, with weights $\left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)}$ (since $\nu_{x_1,j}/(\lambda_{x_1,T} + \nu_{x_1,j})^2 \leq \nu_{x_1,j}$, the infinite sum converges under Assumption B.8 (ii) in Appendix 1). To have an interpretation, note that the inverse of operator $A^*_{x_1}A_{x_1}$ corresponds to the standard asymptotic variance matrix $\left(Q_{XZ}V_0^{-1}Q_{ZX}\right)^{-1}$ of the 2-Stage Least Square (2SLS) estimator of the finite-dimensional parameter θ in the instrumental regression $Y = X'\theta + U$ with $E\left[U|Z\right] = 0$, where $Q_{ZX} = E\left[ZX'\right]$ and $V_0 = V\left[U^2ZZ'\right]$. In the ill-posed nonparametric setting, the inverse of operator $A^*_{x_1}A_{x_1}$ is unbounded, and its eigenvalues

$1/\nu_{x_1,j} \to \infty$ diverge. The penalty term $\lambda_{x_1,T} \left\|\varphi_{x_1}\right\|^2_{H^l(\mathcal{X}_2)}$ in the criterion defining the TiR estimator implies that inverse eigenvalues $1/\nu_{x_1,j}$ are "ridged" with $\nu_{x_1,j}/\left(\lambda_{x_1,T}+\nu_{x_1,j}\right)^2$.

The coefficient $\sigma^2_{x_1}(\lambda_{x_1,T})$ is a decreasing function of $\lambda_{x_1,T}$. Since $\sum_{j=1}^{\infty} \nu^{-1}_{x_1,j} \left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)} = \infty$, the series $\sigma^2_{x_1}(\lambda_{x_1,T})$ diverges as $\lambda_{x_1,T} \to 0$. When $\sigma^2_{x_1}(\lambda_{x_1,T}) \to \infty$ such that $\frac{1}{Th^{d_{X_1}}_{x_1,T}}\sigma^2_{x_1}(\lambda_{x_1,T}) \to 0$, the variance term $V_{x_1,T}$ converges to zero at a slower rate than the standard nonparametric rate $1/\left(Th^{d_{X_1}}_{x_1,T}\right)$. The slower rate is not coming from smoothing variables (W, Z_1), but from the ill-posedness of the problem, which implies $\nu_{x_1,j} \to 0$. The weighting matrix Ω_0 (Assumption 5) impacts in a non-trivial way both the asymptotic variance and the asymptotic bias of the estimator through the adjoint operator $A^*_{x_1}$. Under a generic weighting matrix Ω_0, the asymptotic variance component $V_{x_1,T}$ involves a double sum over the spectrum of $A^*_{x_1}A_{x_1}$. The notion of efficiency of the functional estimator and the associated optimal choice of Ω_0 is out of the scope of the present paper.

The asymptotic expansion of the MISE of estimator $\hat{\varphi}_{x_1}$ given in Proposition 3 involves the conditional distribution of the endogenous regressor X_2 given Z_1 and $X_1 = x_1$, and the conditional variance of the error $U := Y - \varphi_0(X)$ given Z_1 and $X_1 = x_1$ (see Assumption 5), by means of operator $A^*_{x_1}A_{x_1}$. It also involves the joint distribution of U, X_2 and Z by means of the estimation bias contribution Ξ_{x_1}. The asymptotic expansion of the MISE does not involve the bandwidth h_T for smoothing (W, Z_1), as long as Conditions (20) are satisfied. The variance term is asymptotically independent of h_T since the asymptotic expansion of $\hat{\varphi}_{x_1} - \varphi_{x_1,0}$ involves the kernel density estimator integrated w.r.t. (W, Z_1) (see term $\mathcal{V}_{x_1,T}$ in Equation (13)). The integral averages the localization effect of the bandwidth h_T (but not

that of $h_{x_1,T}$). On the contrary, kernel smoothing for both (W, Z_1) and X_1 does impact on bias. However, the second condition in (20) implies that the estimation bias from smoothing (W, Z_1) is asymptotically negligible compared to the regularization bias (see Lemma A.7 in Appendix 3). The other restrictions on the bandwidth h_T in (20) are used to control higher order terms in the MISE (see Lemma A.5).

The set of Assumptions B in Appendix 1 used to prove Proposition 3 includes regularity conditions on the eigenfunctions of operator $A_{x_1}^* A_{x_1}$ (Assumption B.8), which are more restrictive than the conditions used e.g. in HH, DFFR and BCK. These assumptions are required to derive the sharp asymptotic expansion of the MISE, a result stronger than the rates of convergence derived in HH, DFFR and BCK (see also the discussion in Appendix 1).

When there are no exogenous regressors, the asymptotic MISE of the estimator $\hat{\varphi}$ reduces to:

$$M_T(\lambda_T) = \frac{1}{T}\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \left\|\phi_j\right\|^2_{L^2(\mathcal{X}_2)} + \left\|\left(\lambda_T + A^*A\right)^{-1} A^*A\varphi_0 - \varphi_0\right\|^2_{L^2(\mathcal{X}_2)}. \tag{23}$$

The bias term comes solely from Tikhonov regularization, and no contribution from kernel smoothing appears under the conditions of Proposition 3.

Finally, it is also possible to derive an exact asymptotic expansion of the MISE for the estimator $\tilde{\varphi}_{x_1}$ regularized by the L^2 norm:

$$\begin{aligned} E\left[\left\|\tilde{\varphi}_{x_1} - \varphi_{x_1,0}\right\|^2\right] &= \tilde{M}_{x_1,T}(\lambda_{x_1,T}, h_{x_1,T})(1 + o(1)), \\ \tilde{M}_{x_1,T}(\lambda_{x_1,T}, h_{x_1,T}) &= \frac{\omega^2 f_{X_1}(x_1)}{T h_{x_1,T}^{d_{X_1}}} \sum_{j=1}^{\infty} \frac{\tilde{\nu}_{x_1,j}}{(\lambda_{x_1,T} + \tilde{\nu}_{x_1,j})^2} + \tilde{b}_{x_1}(\lambda_{x_1,T}, h_{x_1,T})^2, \end{aligned} \tag{24}$$

where $\tilde{b}_{x_1}(\lambda_{x_1,T}, h_{x_1,T}) = \left\| \tilde{\mathcal{B}}^r_{x_1,T} + h^m_{x_1,T} \left(\lambda_{x_1,T} + \tilde{A}_{x_1} A_{x_1}\right)^{-1} \tilde{A}_{x_1} \Xi_{x_1} \right\|_{L^2(\mathcal{X}_2)}$ and $\tilde{\mathcal{B}}^r_{x_1,T} := \left(\lambda_T + \tilde{A}_{x_1} A_{x_1}\right)^{-1} \tilde{A}_{x_1} A_{x_1} \varphi_{x_1,0} - \varphi_{x_1,0}$.

This simplifies to $\tilde{M}_T(\lambda_T) = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\tilde{\nu}_j}{(\lambda_T + \tilde{\nu}_j)^2} + \left\| \left(\lambda_T + \tilde{A}^* \tilde{A}\right)^{-1} \tilde{A}^* \tilde{A} \varphi_0 - \varphi_0 \right\|^2_{L^2(\mathcal{X}_2)}$ when only endogenous regressors are present. A similar formula has been derived by Carrasco and Florens (2011) for the density deconvolution problem. Such a characterization is new in the nonparametric IV regression setting.

6 Examples

In the general framework, the derivation of the optimal regularization parameter $\lambda_{x_1,T}$, the optimal bandwidth $h_{x_1,T}$, and the optimal MISE is difficult because Expression (21) involves the spectrum of operator $A^*_{x_1} A_{x_1}$. In this section, for illustrative purpose, we consider two examples where the spectrum of $A^*_{x_1} A_{x_1}$ is characterized explicitly. In Section 6.1 we consider an example with a Gaussian distribution similar to NP yielding a mixed geometric-hyperbolic decay of the spectrum (severe ill-posedness). In Section 6.2 we consider an example with trigonometric eigenfunctions similar to HH yielding an hyperbolic decay of the spectrum (mild ill-posedness). In both examples we work on Sobolev spaces with $l = d_{X_2} = 1$.

6.1 Gaussian distribution

Let the errors U and V, and the instruments X_1 and Z_1, admit a joint Gaussian distribution, with zero means, unit variances and correlation 0.5 between U and V. The endogenous

regressor X_2 is given by $X_2 = \dfrac{Z_1 + V}{\sqrt{2}}$. The variable Y is given by $Y = \bar{\varphi}_0 (X_1 + X_2) + U$, where $\bar{\varphi}_0$ is a differentiable function on $\mathbb{R}$. To ease the notation, the variables X_1 and X_2 are not transformed to have compact support, and we keep $\mathcal{X}_1 = \mathcal{X}_2 = \mathbb{R}$. Accordingly, in this example the L^2 and Sobolev norms are such that $\|\varphi\|^2_{L^2(\mathcal{X}_2)} = \int \varphi(x_2)^2 \phi(x_2)\, dx_2$ and $\|\varphi\|^2_{H^1(\mathcal{X}_2)} = \int \varphi(x_2)^2 \phi(x_2)\, dx_2 + \int (\nabla\varphi(x_2))^2 \phi(x_2)\, dx_2$, where ϕ is the pdf of the standard Gaussian distribution. Similarly, the norm on $\mathcal{Z}_1 = \mathbb{R}$ is such that $\|\psi\|^2_{L^2_{x_1}(\mathcal{Z}_1)} = \int \psi(z_1)^2 \phi(z_1)\, dz_1$, for any $x_1 \in \mathcal{X}_1$. The function $\varphi_{x_1,0}$ is $\varphi_{x_1,0} = \bar{\varphi}_0(\cdot + x_1)$. The operator $A := A_{x_1}$ is independent of x_1, since X_2 is independent of X_1 conditionally on Z_1. The spectrum of the operator $\tilde{A}A$ consists of eigenvalues $\tilde{\nu}_j = \varrho^{2(j-1)} = e^{-\alpha(j-1)}$, with $\varrho = 1/\sqrt{2}$ and $\alpha = -2\log\varrho$, and associated eigenfunctions $\tilde{\phi}_j = H_{j-1}$, $j = 1, 2, ...$, where H_j is the Hermite polynomial of order j (see e.g. CFR). In the Technical Report we show that $A^*A = \mathcal{D}^{-1}\tilde{A}A$, where the differential operator $\mathcal{D}$ is given by $\mathcal{D} = 1 - \nabla^2 - (\nabla \log\phi)\nabla$. From the differential equation of Hermite polynomials (e.g., Abramowitz and Stegun, 1970), we get that the functions $\tilde{\phi}_j$ are eigenfunctions of operator $\mathcal{D}$ with eigenvalues j. This property allows us to derive the spectrum of operator A^*A and to characterize the asymptotic behaviour of the variance $\sigma^2_{x_1}(\lambda_{x_1,T})$ and squared bias $b_{x_1}(\lambda_{x_1,T}, h_{x_1,T})^2$ in the asymptotic MISE of the TiR estimator. This result is stated in the next Proposition 4 for a setting that generalizes our Gaussian example.

Proposition 4: *Let us assume that the spectrum of operator $\tilde{A}_{x_1}A_{x_1}$ consists of eigenvalues $\tilde{\nu}_{x_1,j} \asymp e^{-\alpha j}$, $\alpha > 0$, and eigenfunctions $\tilde{\phi}_{x_1,j}$, and that the functions $\tilde{\phi}_{x_1,j}$ are eigenfunctions of operator $\mathcal{D}$ with eigenvalues $\tau_{x_1,j} \asymp j^\beta$, $\beta \geq 0$. Let further $l = 1$. Then:*

(i) *The spectrum of* $A_{x_1}^* A_{x_1}$ *consists of eigenvalues* $\nu_{x_1,j} = \dfrac{\tilde{\nu}_{x_1,j}}{\tau_{x_1,j}} \asymp j^{-\beta} e^{-\alpha j}$ *and eigenfunctions* $\phi_{x_1 j} = \dfrac{1}{\sqrt{\tau_{x_1,j}}} \tilde{\phi}_{x_1,j}$.

Furthermore, assume that $d_{x_1,j} := \left\langle \phi_{x_1,j}, \varphi_{x_1,0} \right\rangle_{H^1(\mathcal{X}_2)}$ *and* $\xi_{x_1,j} := \left\langle \psi_{x_1,j}, \Xi_{x_1} \right\rangle_{L^2_{x_1}(\mathcal{Z}_1)}$, *where* $\psi_{x_1,j} := A_{x_1}\phi_{x_1 j}/\sqrt{\nu_{x_1,j}}$, *are such that:* (a) $d_{x_1,j}^2 \asymp e^{-2\delta\alpha j}$ *and* $\xi_{x_1,j}^2 \asymp e^{-2\rho\alpha j}$, *for* $\delta \in (0,1)$ *and* $\rho \in (0, 1/2)$; (b) *the number* $n(J)$ *of lags* $1 \leq j \leq J$ *for which* $d_{x_1,j}\xi_{x_1,j} \leq 0$ *is such that* $\lim_{J\to\infty} n(J)/J > 0$. *Then:*

(ii) *Under Conditions (16)-(18) and (20), up to logarithmic terms, we have* $M_{x_1,T}(\lambda_{x_1,T}, h_{x_1,T}) \asymp \dfrac{1}{T h_{x_1,T}^{d_{X_1}} \lambda_{x_1,T}} + \lambda_{x_1,T}^{2\delta} + h_{x_1,T}^{2m} \lambda_{x_1,T}^{2\rho-1}$.

(iii) *Under Conditions (16)-(18) and (20), up to logarithmic terms, the bandwidth* $h_{x_1,T}^* \asymp T^{-\eta_{x_1}^*}$ *and regularization parameter* $\lambda_{x_1,T}^* \asymp T^{-\gamma_{x_1}^*}$ *that optimize the rate of convergence of the MISE are such that* $\eta_{x_1}^* = \dfrac{1}{d_{X_1} + 2\omega}$ *and* $\gamma_{x_1}^* = \dfrac{1}{d_{X_1} + 2\delta} \dfrac{2\omega}{1 + 2\omega}$, *where* $\omega = m \dfrac{1 + 2\delta}{1 + 2(\delta - \rho)}$. *The optimal MISE is such that* $M_{x_1,T}^* \asymp T^{-\varkappa_{x_1}^*}$ *with* $\varkappa_{x_1}^* = \dfrac{2\delta}{1 + 2\delta} \dfrac{2\omega}{1 + 2\omega}$.

Under the assumptions of Proposition 4 (i), the eigenvalues of operator $A_{x_1}^* A_{x_1}$ feature a mixed geometric-hyperbolic decay behaviour. This decay behaviour when the adjoint w.r.t. the Sobolev norm is used, is distinct from the geometric decay obtained when the adjoint w.r.t. the L^2 norm is used in the Gaussian case. Indeed we get $\nu_j = e^{-\alpha(j-1)}/j$, $\phi_j = H_{j-1}/\sqrt{j}$ instead of $\tilde{\nu}_j = e^{-\alpha(j-1)}$, $\tilde{\phi}_j = H_{j-1}$ in our Gaussian example.

Condition (a) requires that the coefficients of functions $\varphi_{x_1,0}$ and Ξ_{x_1} w.r.t. the orthonormal systems $\left\{\phi_{x_1 j} : j \in \mathbb{N}\right\}$ and $\left\{\psi_{x_1,j} : j \in \mathbb{N}\right\}$ in the singular value decomposition of $A_{x_1}^* A_{x_1}$ (see Kress, 1999, Theorem 15.16), feature geometric decay. Under Condition (a), the source condition in Assumption 4 is satisfied for any $\delta_{x_1} < \delta$. Condition (b) re-

quires that the proportion of lags for which either $d_{x_1,j} = 0$, or $\xi_{x_1,j} = 0$, or $d_{x_1,j}$ and $\xi_{x_1,j}$ have opposite sign, is non-zero asymptotically. This condition is used to control a cross-term in the squared bias function and show $b_{x_1}\left(\lambda_{x_1,T}, h_{x_1,T}\right)^2 \asymp \lambda_{x_1,T}^{2\delta} + h_{x_1,T}^{2m}\lambda_{x_1,T}^{2\rho-1}$ in Proposition 4 (ii). By using this result, the conditions (16)-(18) and (20) simplify to $\eta_{x_1} \leq \eta$, $\gamma_{x_1}+\eta_{x_1}+2\eta < 1, \gamma_{x_1} < \min\left\{m\eta_{x_1}, \frac{1-\eta_{x_1}-\eta}{2}, \frac{1-\eta_{x_1}-2\eta}{1+\delta}, \frac{2\left(\eta+(m-1)\eta_{x_1}\right)}{2\delta+1}, \frac{2m\eta}{2\delta+1}\right\}$ when $d_{X_1} = d_{Z_1} = d_{X_2} = 1$.

The optimal convergence rate of the bandwidth in Proposition 4 (iii) is smaller than the standard d_{X_1}-dimensional nonparametric rate with m derivatives. It can be interpreted as the standard nonparametric rate with $\omega > m$ derivatives. The optimal convergence rate of the regularization parameter is smaller than $\frac{1}{1+2\delta}$, that is the optimal rate with no exogenous regressors (i.e. $d_{X_1} = 0$). The optimal convergence rates $\gamma_{x_1}^*$ and $\eta_{x_1}^*$ of the regularization parameter and bandwidth depend on the dimension d_{X_1} of the exogenous variable, but their ratio $\gamma_{x_1}^*/\eta_{x_1}^*$ is independent of d_{X_1}. The optimal convergence rate of the MISE is the product of $\frac{2\delta}{1+2\delta}$, that is the optimal rate with no exogenous regressors, times $\frac{2\omega}{d_{X_1}+2\omega}$, that is the convergence rate of a kernel estimator with ω derivatives. The optimal convergence rates $\gamma_{x_1}^*$, $\eta_{x_1}^*$, and any rate η such that $\frac{1}{1+2\omega}\frac{1+2(\delta-\rho)+2m\rho}{1+2(\delta-\rho)} < \eta < \frac{\omega}{1+2\omega}\frac{\min\{2(2\delta-1),\delta\}}{2\delta+1}$, satisfy Conditions (16)-(18) and (20) if $1 \geq \delta > \frac{1}{2}+\rho$ and $1+2(\delta-\rho)+2m\rho < m\min\{2(2\delta-1),\delta\}$ when $d_{X_1} = d_{Z_1} = d_{X_2} = 1$. Finally, by using Expression (24) it is possible to verify that under the conditions of Proposition 4 the optimal rate of convergence of the L^2 regularized estimator is exactly the same as for the TiR estimator (including logarithmic terms).

6.2 Trigonometric eigenfunctions

Under an hyperbolic spectrum, a result similar to Proposition 4 can be obtained. Hereafter we focus on an example with no exogenous variables and a trigonometric basis of eigenfunctions. Specifically, let the spectrum of $\tilde{A}A$ consist of eigenvalues $\tilde{\nu}_j \asymp j^{-\tilde{\alpha}}$, $\tilde{\alpha} > 1$, and eigenfunctions $\tilde{\phi}_1(x_2) = 1$, $\tilde{\phi}_j(x_2) = \sqrt{2}\cos\left((j-1)\pi x_2\right)$, $j = 2, 3, \cdots$, $x_2 \in \mathcal{X}_2 = [0,1]$. Since the functions $\tilde{\phi}_j$ are eigenfunctions of operator $\mathcal{D} = 1 - \nabla^2$ to eigenvalues $1 + \pi^2 j^2$, the spectrum of operator A^*A for $l = 1$ is such that $\nu_j \asymp j^{-\alpha}$, $\alpha = \tilde{\alpha} + 2$, and $\phi_j = \sqrt{1/\left(1 + \pi^2 j^2\right)}\tilde{\phi}_j$. Moreover, we assume that the function φ_0 is such that $\left\langle \varphi_0, \tilde{\phi}_j \right\rangle^2_{L^2(\mathcal{X}_2)} \asymp j^{-2\rho}$, $1/2 < \rho < 1/2 + \tilde{\alpha}$. Then, the squared bias function is such that $b(\lambda)^2 \asymp \lambda^{2\delta}$ with $2\delta = (2\rho - 1)/\alpha$.

In this second example we compare analytically the optimal MISEs of the TiR estimator, and of the L^2 regularized estimator, denoted M_T^* and $\tilde{M}_T^*$. This comparison is made possible because we have derived the exact asymptotic expansions $M_T(\lambda)$ and $\tilde{M}_T(\lambda)$. The optimal MISEs are $M_T^* = cT^{-\varkappa}(1 + o(1))$ and $\tilde{M}_T^* = \tilde{c}T^{-\varkappa}(1 + o(1))$, where $\varkappa = \dfrac{2\delta}{1 + 2\delta - 1/\alpha} = \dfrac{2\rho - 1}{2\rho + \tilde{\alpha}}$ and the constants $c, \tilde{c}$ are such that

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

The two estimators feature the same rate of convergence $\varkappa$, which is the optimal rate given in HH, Theorem 4.1, and in BCK, Theorem 3 (with their $r = (2\rho - 1)/2$ and $s = \tilde{\alpha}/2$). The rate of convergence in Proposition 4 is recovered when $d_{X_1} = 0$ and $\alpha \to \infty$. The ratio $c/\tilde{c}$ yields the relative efficiency of the TiR estimator compared to the L^2 regularized estimator. For any $\rho > 1/2$, the ratio $c/\tilde{c}$ is a monotonically increasing function of α with range $(0, 1)$. In

particular, $c/\tilde{c} < 1$. Moreover, there exist models for which the TiR estimator is arbitrarily more efficient ($c/\tilde{c} \to 0$) compared to the L^2 regularized estimator.

In the two above examples, the operators A^*A and $\tilde{A}A$ admit a common basis of eigenfunctions and we obtain a common rate of convergence for Sobolev and L^2 penalization. The analytic comparison of the rates of convergence and the discussion of the relative efficiency of the estimators in the general case is still an open question.

7 Numerical implementation

To compute numerically the estimator we solve Equation (11) on the subspace spanned by a finite-dimensional basis of functions $\{P_j : j = 1, ..., k\}$, such as Chebyshev or Legendre polynomials, and use the numerical approximation

$$\varphi_{x_1} \simeq \sum_{j=1}^{k} \theta_{x_1 j} P_j =: \theta'_{x_1} P, \quad \theta_{x_1} \in \mathbb{R}^k. \tag{25}$$

The $k \times k$ matrix corresponding to operator $\hat{A}^*_{x_1}\hat{A}_{x_1}$ on this subspace is given by $\langle P_i, \hat{A}^*_{x_1}\hat{A}_{x_1}P_j\rangle_{H^l(\mathcal{X}_2)} = \langle \hat{A}_{x_1}P_i, \hat{A}_{x_1}P_j\rangle_{L^2_{x_1}(\mathcal{Z}_1)} \simeq \frac{1}{Th_{x_1,T}} \sum_{t=1}^{T} \left(\hat{A}_{x_1}P_i\right)(Z_{1t})\, \hat{\Omega}_{x_1}(Z_{1t}) \left(\hat{A}_{x_1}P_j\right)(Z_{1t})$ $K((X_{1t}-x_1)/h_{x_1,T})/\hat{f}_{X_1}(x_1) \simeq \left(\widehat{P}'_{x_1}\widehat{\Sigma}_{x_1}\widehat{P}_{x_1}\right)_{i,j}$, $\quad i, j = 1, ..., k$, where $\widehat{P}_{x_1}$ is the $T \times k$ matrix with rows $\widehat{P}_{x_1}(Z_{1t})' = \frac{1}{Th_T h_{x_1,T}} \sum_{l=1}^{T} P(X_{2l})' K((Z_{1l}-Z_{1t})/h_T) K((X_{1l}-x_1)/h_{x_1,T})/\hat{f}_{Z_1,X_1}(Z_{1t}, x_1)$ and $\widehat{\Sigma}_{x_1}$ is the $T \times T$ diagonal matrix with diagobal elements $\frac{1}{Th_{x_1,T}\hat{f}_{X_1}(x_1)}\hat{\Omega}_{x_1}(Z_{1t})\, K((X_{1t}-x_1)/h_{x_1,T})$, $t = 1, ..., T$. The use of empirical averages instead of integrals in the definition of the estimator simplifies the implementation and is asymptotically equivalent. It avoids bivariate numerical integration and the choice of two additional bandwidths. Ma-

trix $\widehat{P}_{x_1}$ is the matrix of the "fitted values" in the regression of $P(X_2)$ on Z_1 at the sample points conditionally to $X_1 = x_1$. Then, by projection on the k-dimensional linear subspace of $H^l(\mathcal{X}_2)$ spanned by $\{P_j : j = 1, ..., k\}$, Equation (11) reduces to a matrix equation $\left(\lambda_T D + \widehat{P}'_{x_1}\widehat{\Sigma}_{x_1}\widehat{P}_{x_1}\right)\theta = \widehat{P}'_{x_1}\widehat{\Sigma}_{x_1}\widehat{R}_{x_1}$, where $\left(\widehat{R}_{x_1}\right)_t = \hat{r}_{x_1}(Z_{1t})$ with $\hat{r}_{x_1}(Z_{1t}) = \frac{1}{Th_T h_{x_1,T}}\sum_{l=1}^{T} Y_l K((Z_{1l} - Z_{1t})/h_T)K((X_{1l} - x_1)/h_{x_1,T})/\hat{f}_{Z_1,X_1}(Z_{1t}, x_1)$, and D is the $k \times k$ matrix of Sobolev scalar products $D_{i,j} = \langle P_i, P_j\rangle_{H^l(\mathcal{X}_2)}$, $\quad i, j = 1, ..., k$. The solution is $\widehat{\theta}_{x_1} = \left(\lambda_T D + \widehat{P}'_{x_1}\widehat{\Sigma}_{x_1}\widehat{P}_{x_1}\right)^{-1}\widehat{P}'_{x_1}\widehat{\Sigma}_{x_1}\widehat{R}_{x_1}$, which yields the approximation of the TiR estimator $\hat{\varphi}_{x_1} \simeq \hat{\theta}'_{x_1}P$. It only asks for inverting a $k \times k$ matrix. The matrix D is by construction positive definite, since its entries are scalar products of linearly independent basis functions. Hence, $\lambda_T D + \widehat{P}'_{x_1}\widehat{\Sigma}_{x_1}\widehat{P}_{x_1}$ is non-singular, P-a.s..

Estimator $\hat{\theta}_{x_1}$ is a 2SLS estimator with optimal instruments and a ridge correction term. It is also obtained if we replace (25) in Criterion (9) and minimize w.r.t. θ_{x_1}. This route is followed by NP, AC, and BCK, who use sieve estimators and let $k = k_T \to \infty$ with T to regularize the estimation. 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 (12). The latter is a well-defined estimator on the function space $H^l(\mathcal{X}_2)$, 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 (not too small) 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 a stable behavior of the optimized objective function w.r.t. k is another possibility (see the empirical application). Alternatively, we could simply take an a priori large k as in the next section for which matrix inversion in computing $\widehat{\theta}_{x_1}$ is numerically feasible.

Finally, a similar approach can be followed under an L^2 regularization by replacing matrix D with matrix B of L^2 scalar products $B_{i,j} = \langle P_i, P_j \rangle_{L^2(\mathcal{X}_2)}, \quad i,j = 1, ..., k$. DFFR follow a different approach to compute exactly the estimator (see DFFR, Appendix C). Their method requires solving a $T \times T$ linear system of equations. For univariate X and Z, HH implement an estimator which uses the same basis for estimating conditional expectation $m(\varphi, z)$ and for approximating function $\varphi(x)$.

8 A Monte-Carlo study

We include an exogenous regressor X_1 in a design similar to NP and to the example of Section 6.1. The errors U and V and the instrument Z_1 and X_1^* are jointly normally distributed, with zero means, unit variances and correlation coefficient $\rho = 0.5$ between U and V. We take $X_2^* = (Z_1 + V)/\sqrt{2}$ and build the endogenous regressor $X_2 = \Phi(X_2^*)$ where the function Φ denotes the cdf of a standard Gaussian variable. Similarly we take $X_1 = \Phi(X_1^*)$ for the exogenous regressor. To generate Y, we examine the design $Y = \sin(\pi(X_1 + X_2 - 0.5)) + U$. Then $E[Y - \varphi_0(X) \mid Z] = 0$ and the functional parameter satisfies $\varphi_{x_1,0}(.) = \sin(\pi(x_1 + . - 0.5))$. We work with a Sobolev space of order $l = 1$.

As $\mathcal{X}_2 = [0, 1]$, we use a numerical approximation based on standardized shifted Chebyshev polynomials of the first kind (Abramowitz and Stegun, 1970). We take a large number

$k = 16$ of polynomials from orders 0 to 15 in (25) to match our theory. Matrices D and B are explicitly computed with a symbolic calculus package.

The kernel estimator $\hat{m}_{x_1}\left(\varphi_{x_1}, z_1\right)$ of the conditional moment is approximated through $\theta'_{x_1}\hat{P}_{x_1}(z_1) - \hat{r}_{x_1}(z_1)$, where $\hat{P}_{x_1}(z_1)$ and $\hat{r}_{x_1}(z_1)$ are standard kernel regressions with Gaussian kernel. All bandwidths are selected via the standard rule of thumb (Silverman, 1986). This choice is motivated by ease of implementation. Moderate deviations from this simple rule do not seem to affect estimation results significantly. The weighting function $\Omega_{x_1,0}(z_1)$ is taken equal to unity, satisfying Assumption 5, and assumed to be known.

The sample size is fixed at $T = 1000$. In Figures 1 and 3 (TiR estimator) and Figures 2 and 4 (L^2 regularized estimator), the left panel plots the MISE on a grid of lambda, the central panel the Integrated Squared Bias (ISB), and the right panel the mean estimated functions and the true function on the unit interval. Mean estimated functions correspond to averages over 1000 repetitions obtained from regularized estimates with a lambda achieving the lowest MISE. In each panel, we also display corresponding quantities for a bivariate standard kernel regression. We look at function φ_{x_1} for $x_1 = \Phi(0)$ in the two first figures and $x_1 = \Phi(1)$ in the two next ones. Several remarks can be made. First, the endogeneity bias of the standard kernel estimator is large, and as a consequence its MISE as well. Second, the MISE under a Sobolev penalization is more convex and much smaller than the MISE under an L^2 penalization for the same range of λ. So even if we expect the same optimal convergence rates (cf. Section 6.1), the Sobolev norm should be strongly favored in our Monte-Carlo design in order to recover the shape of the true function. A potential theoretical

explanation is the that multiplicative constants play a crucial role in the MISE behavior (cf. Section 6.2). Third, examining the ISB for λ close to 0 reveals that the estimation part of the bias of the TiR estimator coming from smoothing is negligible w.r.t. the regularization part.

In Figures 5 and 6 we look at the same design as in Figures 1 and 3 except for $\rho = 0$. When we suppress the endogeneity of X_2 the MISE of the TiR estimator is slightly larger than the MISE of the standard kernel regression estimator. We loose in terms of ISB and variance w.r.t. kernel regression as predicted by theory. Moreover, the MISE of the TiR estimator is very close in the endogenous and exogenous designs. This feature is in accordance with Proposition 3, since the spectrum of operator $A_{x_1}^* A_{x_1}$ is the same in the two designs, and the estimation bias contribution is rather small.

For $T = 100$ and $T = 400$ as well as a number k of polynomials as low as 6, our conclusions remain qualitatively unaffected. This suggests that as soon as the order of the polynomials is sufficiently large to numerically approximate the underlying function, there is no gain by linking it with sample size (cf. Section 7).

To summarize our Monte-Carlo findings we get evidence in favor of a Sobolev penalty instead of an L^2 penalty. We also observe that there is little to gain from using kernel regression in the exogenous case but a lot to loose if we neglect endogeneity when it is present.

9 An empirical example

This section presents an empirical example with the data in Horowitz (2006) based on the moment condition $E\left[Y-\varphi_0\left(X_2\right) \mid Z_1\right]=0$, with $X_2=\Phi\left(X_2^*\right)$ and no exogenous regressor. We estimate an Engel curve where variable Y denotes the food expenditure share, X_2^* denotes the standardized logarithm of total expenditures, and Z_1 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 a data-driven regularization parameter.

The data driven selection procedure of the regularization parameter λ_T aims at estimating directly the asymptotic spectral representation (23). A similar heuristic approach has been successfully applied in Carrasco and Florens (2011) for density deconvolution. Theoretical properties of such a selection procedure are still unknown, and beyond the scope of this paper. Preliminary Monte-Carlo results show that the selected parameter is of the same magnitude as the optimal one (see GS). The selection algorithm works as follows.

Algorithm to select the regularization parameter

(i) Perform the spectral decomposition of the matrix $D^{-1}\widehat{P}'\widehat{\Sigma}\widehat{P}$ 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 estimate $\bar{\theta}$ using a pilot regularization parameter $\bar{\lambda}$.

(iii) Estimate the MISE:

$$\bar{M}\left(\lambda\right)=\frac{1}{T}\sum_{j=1}^{k}\frac{\hat{\nu}_j}{\left(\lambda+\hat{\nu}_j\right)^2}\hat{w}_j'B\hat{w}_j$$

$$+\bar{\theta}' \left[\widehat{P}' \widehat{\Sigma} \widehat{P} \left(\lambda D + \widehat{P}' \widehat{\Sigma} \widehat{P} \right)^{-1} - I \right] B \left[\left(\lambda D + \widehat{P}' \widehat{\Sigma} \widehat{P} \right)^{-1} \widehat{P}' \widehat{\Sigma} \widehat{P} - I \right] \bar{\theta},$$

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

(iv) Compute the second-step TiR estimator with $\widehat{\theta}$ using regularization parameter $\hat{\lambda}$.

We take six polynomials. Here the value of the optimized objective function stabilizes after $k = 6$ (see Figure 7), and estimation results remain virtually unchanged for larger k. We have observed a stabilization of the loadings in the numerical series approximation and of the data-driven regularization parameter. We have also observed that higher order polynomials receive loadings which are closer and closer to zero. This suggests that we can limit ourselves to a small number of polynomials in this empirical example.

Since $\Omega_0(z_1) = V\left[Y - \varphi_0\left(X_2\right) \mid Z_1 = z_1\right]^{-1}$ is doubtfully constant in this application we estimate the weighting matrix. We use a pilot regularization parameter $\bar{\lambda} = .0001$ to get a first step estimator of φ_0. The estimator $\hat{s}^2(Z_{1,t})$ of the conditional variance $s^2(Z_{1,t}) = \Omega_0(Z_{1,t})^{-1}$ is of a kernel regression type.

Estimation with the data driven selection procedure takes less than 2 seconds, and we obtain a selected value of $\hat{\lambda} = .01113$. Figure 8 plots the estimated functions $\hat{\varphi}(x_2)$ for $x_2 \in [0,1]$, and $\hat{\varphi}(\Phi\left(x_2^*\right))$ for $x_2^* \in \mathbb{R}$. The plotted shape corroborates the findings of Horowitz (2006), who rejects a linear curve but not a quadratic curve at the 5% significance level to explain $\log Y$. The specification test of Gagliardini and Scaillet (2007) does not reject the null hypothesis of the correct specification of the moment restriction used in estimating the Engel curve at the 5% significance level (p-value = .67). Banks, Blundell and Lewbel

(1997) consider demand systems that accommodate such empirical Engel curves.

Appendix 1: List of regularity conditions

Below we list the additional regularity conditions. For a function f of variable s in $\mathbb{R}^{d_s}$ and a multi-index $\alpha \in \mathbb{N}^{d_s}$, we denote $\nabla^\alpha f := \nabla_{s_1}^{\alpha_1} \cdots \nabla_{s_{d_s}}^{\alpha_{d_s}} f$, $|\alpha| := \sum_{i=1}^{d_s} \alpha_i$, $\|f\|_\infty := \sup_s |f(s)|$ and $\|D^m f\|_\infty := \sum_{\alpha:|\alpha|\leq m} \|\nabla^\alpha f\|_\infty$.

B.1: (i) $\{(Y_t, X_{2,t}, Z_t^*) : t = 1, ..., T^*\}$ *is a sample of i.i.d. observations of random variable* (Y, X_2, Z^*), *where* $Z^* := (Z_1^*, X_1^*)$, *admitting a density* f_{Y,X_2,Z^*} *on the support* $\mathcal{Y}\times\mathcal{X}_2\times\mathcal{Z}^* \subset \mathbb{R}^d$, *where* $\mathcal{Y} \subset\mathbb{R}$, $\mathcal{X}_2 = [0,1]^{d_{X_2}}$, $\mathcal{Z}^* \subset \mathbb{R}^{d_{Z_1}+d_{X_1}}$, $d = d_{X_2} + d_{Z_1} + d_{X_1} + 1$. (ii) *The density* f_{Y,X_2,Z^*} *is in class* $C^m\left(\mathbb{R}^d\right)$, *with* $m \geq 2$, *and* $\nabla^\alpha f_{Y,X_2,Z^*}$ *is uniformly continuous and bounded, for any* $\alpha \in \mathbb{N}^d$ *with* $|\alpha| = m$. (iii) *The random variable* (Y, X_2, Z) *is such that* $(Y, X_2, Z) = (Y, X_2, Z^*)$ *if* $Z^* \in \mathcal{Z}$, *where* $\mathcal{Z} = \mathcal{Z}_1\times\mathcal{X}_1$=$[0,1]^{d_{Z_1}+d_{X_1}}$ *is interior to* $\mathcal{Z}^*$, *and the density* f_Z *of* Z *is such that* $\inf_{z\in\mathcal{Z}} f_Z(z) > 0$.

B.2: *The kernel* K *on* $\mathbb{R}^d$ *is such that* (i) $\int K(u)du = 1$ *and* K *is bounded*; (ii) K *has compact support;* (iii) K *is Lipshitz*; (iv) $\int u^\alpha K(u)du = 0$ *for any* $\alpha \in \mathbb{N}^d$ *with* $|\alpha| < m$.

B.3: (i) *The density* $f_{X_2|Z}$ *of* X_2 *given* Z *is such that* $\left\|D^{m\vee l} f_{X_2|Z}\right\|_\infty < \infty$; (ii) *The function* $\mu(z) = E[Y|Z = z] f_Z(z)$ *is such that* $\|D^m\mu\|_\infty < \infty$. *Moreover,* $E[|Y|^s] < \infty$ *and* $\sup_{z\in\mathcal{Z}} E[|Y|^s \mid Z = z] f_Z(z) < \infty$ *for* $s > 4$.

B.4: *There exists* $h > 0$ *such that function* $q(s) := \sum_{\alpha:|\alpha|\leq m} \sup_{v\in B_h(s)} |\nabla^\alpha f_{Y,X_2,Z}(v)|$, $s \in \mathcal{S}$, is *integrable and satisfies* $\sup_{x_1\in\mathcal{X}_1} \int \frac{q(s)^2}{f_{Y,X_2,Z}(s)} dydx_2dz_1 < \infty$, *where* $B_h(s)$ *denotes the ball in* $\mathbb{R}^d$ *of radius* h *centered at* s.

B.5: *Function* φ_0 *is such that* $\int \varphi_{x_1,0}(x_2)^4 dx_2 < \infty$, *for any* $x_1 \in \mathcal{X}_1$.

B.6: *The weighting function* Ω_0 *is such that* $\sup_{z\in\mathcal{Z}} \Omega_0(z) < \infty$ *and* $\inf_{z\in\mathcal{Z}} \Omega_0(z) > 0$.

B.7: *Estimator* $\hat{\Omega}$ *of* Ω_0 *is such that* (i) $\sup_{z_1\in\mathcal{Z}_1} \hat{\Omega}_{x_1}(z_1) < \infty$, *P-a.s., for any* $x_1 \in \mathcal{X}_1$, (ii) $\sup_{z_1\in\mathcal{Z}_1} \left|\Delta\hat{\Omega}_{x_1}(z_1)\right| = O_p\left(\sqrt{\frac{\log T}{T h_{x_1,T}^{d_{X_1}} h_T^{d_{Z_1}}}} + h_T^m + h_{x_1,T}^m\right)$, *uniformly in* $x_1 \in \mathcal{X}_1$; (iii) $\sup_{z_1\in\mathcal{Z}_1} E\left[\left|\Delta\hat{\Omega}_{x_1}(z)\right|^N\right] = O\left(\left(\frac{1}{\sqrt{T h_{x_1,T}^{d_{X_1}} h_T^{d_{Z_1}}}} + h_T^m + h_{x_1,T}^m\right)^N\right)$, *for any* $x_1 \in \mathcal{X}_1$ *and* $N \in \mathbb{N}$.

B.8: *For any* $x_1 \in \mathcal{X}_1$: (i) $\sum_{j,i=1,j\neq i}^{\infty} \frac{\left\langle \phi_{x_1,j}, \phi_{x_1,i}\right\rangle^2_{L^2(\mathcal{X}_2)}}{\left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)} \left\|\phi_{x_1,i}\right\|^2_{L^2(\mathcal{X}_2)}} < \infty$; (ii) $\sum_{j=1}^{\infty} \nu_{x_1,j} \left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)} < \infty$; (iii) $\sup_{j\in\mathbb{N}} E\left[\left|g_{x_1,j}(U_2)\right|^{\bar{s}} | X_1 = x_1\right] < \infty$, *for* $\bar{s} \geq 4$, *where* $g_{x_1,j}(u_2) := \left(\psi_{x_1,j}\right)(z_1)\Omega_{x_1,0}(z_1)(y - \varphi_{x_1,0}(x_2))$ *and* $\psi_{x_1,j} = A_{x_1}\phi_{x_1,j}/\sqrt{\nu_{x_1,j}}$; (iv) *The functions* $g_{x_1,j}$ *are differentiable such that* $\sup_{j\in\mathbb{N}} E\left[\left|\nabla g_{x_1,j}(U_2)\right|^{\bar{s}} | X_1 = x_1\right] < \infty$.

In Assumption B.1 (i), the compact support of X_2 is used for technical reasons. Mapping in the unit hypercube can be achieved by simple linear or nonlinear monotone transformations. If a nonlinear invertible transform Λ is used to map the observations $X^*_{2,t}$ into $[0,1]^{d_{X_2}}$ through $X_{2,t} = \Lambda\left(X^*_{2,t}\right)$ (e.g. the cdf of the standard Gaussian distribution applied componentwise) then the smoothness assumptions bear on function $\varphi_{x_1,0} = \varphi^*_{x_1,0} \circ \Lambda^{-1}$, since $\varphi_{x_1,0}(x_2) = \varphi^*_{x_1,0}(x^*_2)$. Assumptions B.1 (ii) and B.2 are classical conditions in kernel density estimation concerning smoothness of the density and of the kernel. In particular, when $m > 2$, K is a higher order kernel. Moreover, we assume a compact support for the kernel K to simplify the set of regularity conditions. In Assumption B.1 (iii), variable Z is obtained

by truncating Z^* on the compact set $\mathcal{Z}$, and the density f_Z of Z is bounded from below away from 0 on the support $\mathcal{Z}$. The corresponding observations are Z_t, $t = 1, ..., T$, where $T \leq T^*$. We get the estimator $\hat{f}_{Y,X_2,Z}$ of the density $f_{Y,X_2,Z}$ on $\mathcal{Y} \times \mathcal{X}_2 \times \mathcal{Z}$ from the kernel estimator $\hat{f}_{Y,X_2,Z^*}(y,x,z) = \frac{1}{T^* h_T^d} \sum_{l=1}^{T^*} K\left((Y_l - y)/h_T\right) K\left((X_{2,l} - x)/h_T\right) K\left((Z_l^* - z)/h_T\right)$ of density f_{Y,X_2,Z^*} by normalization and trimming, namely $\hat{f}_{Y,X_2,Z} = \hat{f}_{Y,X_2,Z^*} / \int_{\mathcal{Z}} \hat{f}_{Z^*,\tau}$, where $\hat{f}_{Z^*,\tau} = \max\left\{\hat{f}_{Z^*}, (\log T)^{-1}\right\}$. Similarly, $\hat{f}_{Y,X_2|Z} = \hat{f}_{Y,X_2,Z^*} / \hat{f}_{Z^*,\tau}$. The truncation trick is used to avoid edge effects when smoothing Z while maintaining the assumption $\inf_{z \in \mathcal{Z}} f_Z(z) > 0$, or equivalently $\inf_{z \in \mathcal{Z}} f_{Z^*}(z) > 0$ (since f_Z is a rescaled version of f_{Z^*}). The latter condition is useful to control in probability for small values of the estimator $\hat{f}_{Z^*}$ of density f_{Z^*} appearing in denominators. The additional trimming of $\hat{f}_{Z^*}$ is necessary to control in mean square sense small values of the estimator $\hat{f}_{Z^*}$ of density f_{Z^*} appearing in denominators. The condition $\inf_{z \in \mathcal{Z}} f_{Z^*}(z) > 0$ allows us to select a simple trimming sequence $(\log T)^{-1}$ independent of the density tails. Alternative approaches to address these technical issues consist in using more general forms of trimming (see e.g. Hansen, 2008), boundary kernels or density weighting (see e.g. HH).

Assumptions B.3 (i) and (ii) concern boundedness and smoothness of the p.d.f. of X_2 given Z, and the (conditional) moments of Y (given Z), respectively. Assumption B.4 concerns the joint density $f_{Y,X_2,Z}$, and imposes an integrability condition on a suitable measure of local variation of density $f_{Y,X_2,Z}$ and its derivatives. This assumption is used in the proof of Lemmas A.5-A.7 to bound higher order terms in the asymptotic expansion of the MISE coming from kernel estimation bias. Similarly, Assumption B.5 on function φ_0 is used to

bound the expectation of terms involving powers of $\varphi_0(X_2)$ in the proof of Lemma A.5 (see also Lemmas B.6 and B.7 in the Technical Report). Assumption B.6 imposes boundedness from above and from below on the weighting function Ω_0. In particular, Assumption B.6 together with Assumption B.3 (i) imply that operator A_{x_1} is compact, for any $x_1 \in \mathcal{X}_1$. Assumption B.7 concerns the estimator $\hat{\Omega}_{x_1}$ of the weighting function. Specifically, Assumption B.7 (i) is a uniform a.s. bound for $\hat{\Omega}_{x_1}$, Assumption B.7 (ii) yields a uniform rate of convergence in probability and Assumption B.7 (iii) yields a uniform rate of convergence in mean square. Assumption B.7 covers the trivial case of known weighting function Ω_0, and the choice $\Omega_0 = f_Z$ used by HH. In particular, Assumptions B.1-B.3, B.6, B.7 (i)-(ii), together with Assumptions 1-4, imply the (uniform) consistency of the TiR estimator (Proposition 1 and 2).

Finally, Assumption B.8 concerns the singular system $\left\{\sqrt{\nu_{x_1,j}}, \phi_{x_1,j}, \psi_{x_1,j}; j \in \mathbb{N}\right\}$ of operator A_{x_1} (Kress, 1999, p. 278) and is used to derived the sharp asymptotic expansion of the MISE (Proposition 3). Assumption B.8 (i) requires that the $\langle ., . \rangle_{H^l(\mathcal{X}_2)}$-orthonormal basis of eigenfunctions of $A^*_{x_1} A_{x_1}$ satisfies a summability condition w.r.t. $\langle ., . \rangle_{L^2(\mathcal{X}_2)}$, for any $x_1 \in \mathcal{X}_1$. Assumption B.8 (ii) implies the convergence of the series defining the variance term of the MISE. Assumptions B.8 (iii) and (iv) ask for the existence of bounds for moments of derivatives of functions $g_{x_1,j}$, uniformly $j \in \mathbb{N}$ and for any $x_1 \in \mathcal{X}_1$. These assumptions control for terms of the type $\int g_{x_1,j}(u_2)\, \hat{f}_{Y,X_2,Z_1}(u_2, x_1) du_2$, uniformly in $j \in \mathbb{N}$ and for any $x_1 \in \mathcal{X}_1$, in the proof of Lemmas A.5 and A.6.

Appendix 2: Characterization of the adjoint operators

In this appendix we characterize the adjoint operator $A_{x_1}^*$ and its empirical counterpart $\hat{A}_{x_1}^*$. Let $H_0^{2l}(\mathcal{X}_2) = \left\{\psi \in H^{2l}(\mathcal{X}_2) \mid \nabla^\alpha \psi(x_2) = 0 \text{ for a.e. } x_2 \in \partial\mathcal{X}_2 \text{ and all } |\alpha| < 2l \text{ odd}\right\}$ be the subspace of $H^{2l}(\mathcal{X}_2)$ consisting of functions with odd-order derivatives vanishing on the boundary $\partial\mathcal{X}_2 = \{x_2 \in \mathcal{X}_2 : \text{either } x_{2,i} = 0 \text{ or } x_{2,i} = 1, \text{ for some } i = 1, ..., d_{X_2}\}$ of $\mathcal{X}_2$. Let us define the polynomial $p(z) = \sum_{|\alpha| \leq l} z^\alpha$ and the differential operator $\mathcal{D} = p\left(-\nabla^2\right)$, where $\left(-\nabla^2\right)^\alpha := (-1)^{|\alpha|} \prod_{i=1}^{d_{X_2}} \nabla_i^{2\alpha_i}$. Let us introduce the orthonormal basis $\left\{\chi_j : j = (j_1, ..., j_{d_{X_2}}) \in \mathbb{N}^{d_{X_2}}\right\}$ of $L^2(\mathcal{X}_2)$ given by $\chi_j(x_2) = \prod_{i=1}^{d_{X_2}} \tilde{\chi}_{j_i}(x_{2,i})$, where $\tilde{\chi}_{j_i}(x_{2,i}) = 1$, if $j_i = 1$, and $\tilde{\chi}_{j_i}(x_{2,i}) = \sqrt{2}\cos(\pi(j_i - 1)x_{2,i})$, otherwise. The elements of the basis belong to $H_0^{2l}(\mathcal{X}_2)$ and are eigenfunctions of operator $\mathcal{D}$, that is $\mathcal{D}\chi_j = \xi_j \chi_j$, with eigenvalues $\xi_j = p(z_j)$, where $z_j = \pi^2\left((j_i - 1)^2, i = 1, ..., d_{X_2}\right)$. Define further the linear vector space $\mathcal{S}^l(\mathcal{X}_2) = \left\{\varphi \in L^2(\mathcal{X}_2) \mid \sum_{j \in \mathbb{N}^{d_{X_2}}} \left[\xi_j \left\langle \varphi, \chi_j \right\rangle_{L^2(\mathcal{X}_2)}\right]^2 < \infty\right\}$, which is a linear vector subspace of $L^2[0,1]$ made of functions whose basis coefficients $\left\langle \varphi, \chi_j \right\rangle_{L^2(\mathcal{X}_2)}$ feature rapid decay for large $|j|$ such that $\xi_j \left\langle \varphi, \chi_j \right\rangle_{L^2(\mathcal{X}_2)}$, $j \in \mathbb{N}^{d_{X_2}}$, are square-summable. It is an Hilbert space w.r.t. the scalar product $\langle \varphi, \phi \rangle_{\mathcal{S}} := \sum_{j \in \mathbb{N}^{d_{X_2}}} \xi_j^2 \langle \varphi, \chi_j \rangle_{L^2(\mathcal{X}_2)} \langle \phi, \chi_j \rangle_{L^2(\mathcal{X}_2)}$. We denote by $\|\varphi\|_{\mathcal{S}} := \langle \varphi, \varphi \rangle_{\mathcal{S}}^{1/2}$ the associated norm. The space $\mathcal{S}^l(\mathcal{X}_2)$ is equivalent to $H_0^{2l}(\mathcal{X}_2)$, which therefore is also an Hilbert space equiped with the scalar product $\langle ., . \rangle_{\mathcal{S}}$.

Lemma A.1: (i) *For any $\phi \in L^2(\mathcal{X}_2)$, the PDE $\mathcal{D}u = \phi$, $u \in H_0^{2l}(\mathcal{X}_2)$, admits a unique solution, denoted by $u = \mathcal{D}^{-1}\phi$.* (ii) *The mapping $\mathcal{D}^{-1} : L^2(\mathcal{X}_2) \to H_0^{2l}(\mathcal{X}_2)$ is continuous.* (iii) *There exists a unique linear continuous operator $\mathcal{E} : H_0^{2l}(\mathcal{X}_2) \to H^l(\mathcal{X}_2)$ such*

that $\langle \mathcal{D}u, \varphi \rangle_{L^2(\mathcal{X}_2)} = \langle \mathcal{E}u, \varphi \rangle_{H^l(\mathcal{X}_2)}$, *for any* $u \in H_0^{2l}(\mathcal{X}_2)$ *and* $\varphi \in H^l(\mathcal{X}_2)$. (iv) *We have* $A_{x_1}^* = \mathcal{E}\mathcal{D}^{-1}\tilde{A}_{x_1}$ *where* $\widetilde{A}_{x_1}\psi(x_2) = \int \Omega_{x_1,0}(z_1) f_{X_2,Z_1|X_1}(x_2, z_1|x_1)\psi(z_1)dz_1$ *for* $\psi \in L^2_{x_1}(\mathcal{Z}_1)$.

Operator $\mathcal{E}$ is the identity when $l = 1$. When $l > 1$ and $d_{X_2} = 1$, operator $\mathcal{E}$ is characterized in CGS. In the Technical Report, we discuss the characterization of $\mathcal{E}$ when $l, d_{X_2} \geq 1$, and the embedding condition $2l > d_{X_2}$ is satisfied.

Lemma A.2: *Under Assumptions B.2 and B.7 (i), the following properties hold P-a.s. for any* $x_1 \in \mathcal{X}_1$*:*

(i) *The linear operator* $\hat{A}_{x_1}^* := \mathcal{E}\mathcal{D}^{-1}\widetilde{\hat{A}}_{x_1}$ *from* $L^2_{x_1}(\mathcal{Z}_1)$ *into* $H^l(\mathcal{X}_2)$ *is such that, for any* $\psi \in L^2_{x_1}(\mathcal{Z}_1)$, $\varphi \in H^l(\mathcal{X}_2)$: $\left\langle \varphi, \hat{A}_{x_1}^* \psi \right\rangle_{H^l(\mathcal{X}_2)} = \int \hat{\Omega}_{x_1}(z_1) \left(\hat{A}_{x_1}\varphi\right)(z_1)\, \psi(z_1)\, \hat{f}_{Z_1|X_1}(z_1|x_1)\, dz_1$.

(ii) *Operator* $\hat{A}_{x_1}^* \hat{A}_{x_1} : H^l(\mathcal{X}_2) \to H^l(\mathcal{X}_2)$ *is compact.*

Appendix 3: Proof of Propositions 1-3

This appendix concerns the proof of the consistency and the derivation of the asymptotic MISE of the TiR estimator. The steps are as follows: getting the asymptotic expansion of the estimator in A.3.1, controlling the regularization bias in A.3.2, proving consistency in A.3.3 and finally deriving the asymptotic MISE in A.3.4.

A.3.1 Asymptotic expansion (proof of Equation (13))

We can write

$$\begin{aligned}
\hat{r}_{x_1}(z_1) &= \int \left(y - \varphi_{0,x_1}(x_2)\right) \left[\hat{f}_{W|Z}(w|z) - f_{W|Z}(w|z)\right] dw + \int \varphi_{0,x_1}(x_2) \hat{f}_{W|Z}(w|z) dw \\
&= \int \left(y - \varphi_{0,x_1}(x_2)\right) \frac{\Delta \hat{f}_{W,Z}(w,z)}{f_Z(z)} dw + \int \varphi_{0,x_1}(x_2) \hat{f}_{W|Z}(w|z) dw \\
&\quad - \frac{\Delta \hat{f}_Z(z)}{\hat{f}_Z(z)} \int \left(y - \varphi_{0,x_1}(x_2)\right) \frac{\Delta \hat{f}_{W,Z}(w,z)}{f_Z(z)} dw \\
&= \hat{\psi}_{x_1}(z_1) + \zeta_{x_1}(z_1) + \left(\hat{A}_{x_1} \varphi_{0,x_1}\right)(z_1) + \hat{q}_{x_1}(z_1),
\end{aligned}$$

where $\hat{q}_{x_1}(z_1) := -\frac{\Delta \hat{f}_Z(z)}{\hat{f}_Z(z)} \int \left(y - \varphi_{0,x_1}(x_2)\right) \frac{\Delta \hat{f}_{W,Z}(w,z)}{f_Z(z)} dw$, $\Delta \hat{f}_Z := \hat{f}_Z - f_Z$, and $\Delta \hat{f}_{W,Z} := \hat{f}_{W,Z} - f_{W,Z}$. Hence, $\hat{A}^*_{x_1} \hat{r}_{x_1} = A^*_{x_1} \hat{\psi}_{x_1} + A^*_{x_1} \zeta_{x_1} + \hat{A}^*_{x_1} \hat{A}_{x_1} \varphi_{0,x_1} + \left(\hat{A}^*_{x_1} - A^*_{x_1}\right) \left(\hat{\psi}_{x_1} + \zeta_{x_1}\right) + \hat{A}^*_{x_1} \hat{q}_{x_1}$. By replacing this equation into (12), we get (13) where the remainder term $\mathcal{R}_{x_1,T}$ is given by

$$\begin{aligned}
\mathcal{R}_{x_1,T} &= \left[\left(\lambda_{x_1,T} + \hat{A}^*_{x_1} \hat{A}_{x_1}\right)^{-1} - \left(\lambda_{x_1,T} + A^*_{x_1} A_{x_1}\right)^{-1}\right] A^*_{x_1} \left(\hat{\psi}_{x_1} + \zeta_{x_1}\right) \\
&\quad + \left(\lambda_{x_1,T} + \hat{A}^*_{x_1} \hat{A}_{x_1}\right)^{-1} \left[\left(\hat{A}^*_{x_1} - A^*_{x_1}\right) \left(\hat{\psi}_{x_1} + \zeta_{x_1}\right) + \hat{A}^*_{x_1} \hat{q}_{x_1}\right] \\
&\quad + \left[\left(\lambda_{x_1,T} + \hat{A}^*_{x_1} \hat{A}_{x_1}\right)^{-1} \hat{A}^*_{x_1} \hat{A}_{x_1} - \left(\lambda_{x_1,T} + A^*_{x_1} A_{x_1}\right)^{-1} A^*_{x_1} A_{x_1}\right] \varphi_{0,x_1}. \quad (26)
\end{aligned}$$

The remainder term $\mathcal{R}_{x_1,T}$ accounts for estimation of operator A_{x_1} and its adjoint.

A.3.2 Control of the regularization bias term (proof of (15) and (19))

Similarly to the proof of Proposition 3.11 in CFR, we have

$$\begin{aligned}
\|\mathcal{B}^r_{x_1,T}\|^2_{H^l(\mathcal{X}_2)} &= \sum_{j=1}^{\infty} \frac{\lambda^2_{x_1,T} \langle \phi_{x_1,j}, \varphi_{x_1,0} \rangle^2_{H^l(\mathcal{X}_2)}}{\left(\lambda_{x_1,T} + \nu_{x_1,j}\right)^2} = \lambda^{2\delta_{x_1}}_{x_1,T} \sum_{j=1}^{\infty} \frac{\lambda^{2-2\delta_{x_1}}_{x_1,T} \nu^{2\delta_{x_1}}_{x_1,j}}{\left(\lambda_{x_1,T} + \nu_{x_1,j}\right)^2} \frac{\langle \phi_{x_1,j}, \varphi_{x_1,0} \rangle^2_{H^l(\mathcal{X}_2)}}{\nu^{2\delta_{x_1}}_{x_1,j}} \\
&\leq \lambda^{2\delta_{x_1}}_{x_1,T} \sum_{j=1}^{\infty} \frac{\langle \phi_{x_1,j}, \varphi_{x_1,0} \rangle^2_{H^l(\mathcal{X}_2)}}{\nu^{2\delta_{x_1}}_{x_1,j}}.
\end{aligned}$$

Then, Assumption 4 implies (15), while Assumption 4 bis implies (19).

A.3.3 Consistency (proof of Propositions 1 and 2)

The next lemma gives a bound in probability for the Sobolev norm of the remainder term $\mathcal{R}_{x_1,T}$, uniformly in $x_1 \in \mathcal{X}_1$.

Lemma A.3: *Under Assumptions 3, B.1-B.3, B.6, B.7 (i)-(ii) and if* $\frac{\log T}{T h_{x_1,T}^{d_{X_1}} h_T^{d_{Z_1} \vee d_{X_2}}} + h_T^{2m} + h_{x_1,T}^{2m} = o\left(\lambda_{x_1,T}^2\right)$ *and* $\frac{\log T}{T h_{x_1,T}^{d_{X_1}} h_T^{d_{X_2}+d_{Z_1}}} = O(1)$ *uniformly in* $x_1 \in \mathcal{X}_1$, *we have uniformly in* $x_1 \in \mathcal{X}_1$:

$$\begin{aligned}\left\|\mathcal{R}_{x_1,T}\right\|_{H^l(\mathcal{X}_2)} = o_p\left(\left\|\mathcal{V}_{x_1,T}\right\|_{H^l(\mathcal{X}_2)} + \left\|\mathcal{B}^r_{x_1,T}\right\|_{H^l(\mathcal{X}_2)} + \left\|\mathcal{B}^e_{x_1,T}\right\|_{H^l(\mathcal{X}_2)}\right) \\ + O_p\left(\frac{1}{\lambda_{x_1,T}}\left(\frac{\log T}{T h_{x_1,T}^{d_{X_1}} h_T^{d_{Z_1}+d_{X_2}}} + h_T^{2m} + h_{x_1,T}^{2m}\right)\right).\end{aligned}$$

From Equation (13), the triangular inequality and Lemma A.3 we get:

$$\begin{aligned}\left\|\hat{\varphi}_{x_1} - \varphi_{x_1,0}\right\|_{H^l(\mathcal{X}_2)} &= O_p\left(\left\|\mathcal{V}_{x_1,T}\right\|_{H^l(\mathcal{X}_2)} + \left\|\mathcal{B}^r_{x_1,T}\right\|_{H^l(\mathcal{X}_2)} + \left\|\mathcal{B}^e_{x_1,T}\right\|_{H^l(\mathcal{X}_2)}\right) \\ &\quad + O_p\left(\frac{1}{\lambda_{x_1,T}}\left(\frac{\log T}{T h_{x_1,T}^{d_{X_1}} h_T^{d_{Z_1}+d_{X_2}}} + h_T^{2m} + h_{x_1,T}^{2m}\right)\right),\end{aligned} \tag{27}$$

uniformly in $x_1 \in \mathcal{X}_1$. In order to bound in probability the Sobolev norms of $\mathcal{V}_{x_1,T}$ and $\mathcal{B}^e_{x_1,T}$, we use the next lemma.

Lemma A.4: *Under Assumptions 3, B.1-B.3 and B.6, we have uniformly in* $x_1 \in \mathcal{X}_1$*:*
(i) $\left\|\hat{\psi}_{x_1}\right\|_{L^2_{x_1}(\mathcal{Z}_1)} = O_p\left(\sqrt{\frac{\log T}{T h_{x_1,T}^{d_{X_1}} h_T^{d_{Z_1}}}}\right)$; (ii) $\left\|A^*_{x_1}\hat{\psi}_{x_1}\right\|_{H^l(\mathcal{X}_2)} = O_p\left(\sqrt{\frac{\log T}{T h_{x_1,T}^{d_{X_1}}}}\right)$;
(iii) $\left\|\zeta_{x_1}\right\|_{L^2_{x_1}(\mathcal{Z}_1)} = O\left(h_T^m + h_{x_1,T}^m\right)$.

Let $\|.\|_{\mathcal{L}(H_1,H_2)}$ denote the operator norm for operators from Banach space H_1 into Banach space H_2, with $\|.\|_{\mathcal{L}(H_1)} := \|.\|_{\mathcal{L}(H_1,H_1)}$ when $H_1 = H_2$. By using $\left\|\left(\lambda_{x_1,T} + A^*_{x_1}A_{x_1}\right)^{-1}\right\|_{\mathcal{L}\left(H^l(\mathcal{X}_2)\right)}$

$\leq 1/\lambda_{x_1,T}$ and Lemma A.4 (ii), we have $\|\mathcal{V}_{x_1,T}\|_{H^l(\mathcal{X}_2)} = O_p\left(\frac{1}{\lambda_{x_1,T}}\sqrt{\frac{\log T}{Th_{x_1,T}^{d_{X_1}}}}\right)$, uniformly in $x_1 \in \mathcal{X}_1$. By using $\left\|\left(\lambda_{x_1,T} + A_{x_1}^* A_{x_1}\right)^{-1} A_{x_1}^*\right\|_{\mathcal{L}\left(L^2_{x_1}(\mathcal{Z}_1),H^l(\mathcal{X}_2)\right)} \leq 1/\sqrt{\lambda_{x_1,T}}$ (see CGS) and Lemma A.4 (i), we have $\|\mathcal{V}_{x_1,T}\|_{H^l(\mathcal{X}_2)} = O_p\left(\sqrt{\frac{\log T}{\lambda_{x_1,T}Th_{x_1,T}^{d_{X_1}}h_T^{d_{Z_1}}}}\right)$, uniformly in $x_1 \in \mathcal{X}_1$. Thus, we get $\|\mathcal{V}_{x_1,T}\|_{H^l(\mathcal{X}_2)} = O_p\left(\sqrt{\frac{\log T}{Th_{x_1,T}^{d_{X_1}}\lambda_{x_1,T}\left(\lambda_{x_1,T}\vee h_T^{d_{Z_1}}\right)}}\right)$, uniformly in $x_1 \in \mathcal{X}_1$. Moreover, from Lemma A.4 (iii) we get $\left\|\mathcal{B}^e_{x_1,T}\right\|_{H^l(\mathcal{X}_2)} = O\left(\frac{h_T^m + h_{x_1,T}^m}{\sqrt{\lambda_{x_1,T}}}\right)$, uniformly in $x_1 \in \mathcal{X}_1$. From (27) and (15)-(19) we get:

$$\begin{aligned}\left\|\hat{\varphi}_{x_1} - \varphi_{x_1,0}\right\|_{H^l(\mathcal{X}_2)} &= O_p\left(\sqrt{\frac{\log T}{Th_{x_1,T}^{d_{X_1}}\lambda_{x_1,T}\left(\lambda_{x_1,T}\vee h_T^{d_{Z_1}}\right)}} + \frac{h_T^m + h_{x_1,T}^m}{\sqrt{\lambda_{x_1,T}}} + \lambda_{x_1,T}^{\delta_{x_1}}\right)\\ &\quad + O_p\left(\frac{1}{\lambda_{x_1,T}}\frac{\log T}{Th_{x_1,T}^{d_{X_1}}h_T^{d_{Z_1}+d_{X_2}}} + \frac{h_T^{2m} + h_{x_1,T}^{2m}}{\lambda_{x_1,T}}\right),\end{aligned}$$

uniformly in $x_1 \in \mathcal{X}_1$. Then, Propositions 1 and 2 follow.

A.3.4 MISE (proof of Proposition 3)

The next Lemma A.5 shows that the L^2-norm of the remainder term $\mathcal{R}_{x_1,T}$ is asymptotically negligible in mean square sense.

Lemma A.5: *Under Assumptions 5, B.1-B.6, B.7 (iii), B.8 and if* $\frac{1}{Th_{x_1,T}^{d_{X_1}}h_T^{d_{X_2}\vee d_{Z_1}}} + h_T^{2m} + h_{x_1,T}^{2m} = O\left(\lambda_{x_1,T}^{2+\varepsilon}\right)$, $\varepsilon > 0$, $\frac{(\log T)^2}{Th_{x_1,T}^{d_{X_1}}h_T^{d_{Z_1}+d_{X_2}}} = O(1)$, $\frac{1}{Th_{x_1,T}^{d_{X_1}}h_T^{d_{Z_1}+d_{X_2}}} + h_{x_1,T}^{2m} + h_T^{2m} = o\left(\lambda_{x_1,T}b\left(\lambda_{x_1,T},h_{x_1,T}\right)\right)$, *we have:* $E\left[\|\mathcal{R}_{x_1,T}\|^2_{L^2(\mathcal{X}_2)}\right] = o\left(E\left[\|\mathcal{V}_{x_1,T}\|^2_{L^2(\mathcal{X}_2)}\right] + \left\|\mathcal{B}^r_{x_1,T} + \mathcal{B}^e_{x_1,T}\right\|^2_{L^2(\mathcal{X}_2)}\right)$.

From Equation (13) and Lemma A.5 we have:

$$E\left[\left\|\hat{\varphi}_{x_1}-\varphi_{x_1,0}\right\|^2_{L^2(\mathcal{X}_2)}\right] = E\left[\left\|\mathcal{V}_{x_1,T}+\mathcal{B}^r_{x_1,T}+\mathcal{B}^e_{x_1,T}\right\|^2_{L^2(\mathcal{X}_2)}\right](1+o(1))$$
$$= \left(E\left[\left\|\mathcal{V}_{x_1,T}\right\|^2_{L^2(\mathcal{X}_2)}\right]+\left\|\mathcal{B}^r_{x_1,T}+\mathcal{B}^e_{x_1,T}\right\|^2_{L^2(\mathcal{X}_2)}\right)(1+o(1)), \quad (28)$$

for any $x_1 \in \mathcal{X}_1$. To derive the asymptotic expansion of the MISE, we need sharp bounds for the variance contribution $E\left[\left\|\mathcal{V}_{x_1,T}\right\|^2_{L^2(\mathcal{X}_2)}\right]$ and the bias contribution $\left\|\mathcal{B}^r_{x_1,T}+\mathcal{B}^e_{x_1,T}\right\|^2_{L^2(\mathcal{X}_2)}$. They are given in next Lemmas A.6 and A.7.

Lemma A.6: *Under Assumptions B.1-B.4, B.8 and 5, we have* $E\left[\left\|\mathcal{V}_{x_1,T}\right\|^2_{L^2(\mathcal{X}_2)}\right] = \left(\frac{\omega^2 f_{X_1}(x_1)}{T h^{d_{X_1}}_{x_1,T}}\sum_{j=1}^{\infty}\frac{\nu_{x_1,j}}{\left(\lambda_{x_1,T}+\nu_{x_1,j}\right)^2}\left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)}\right)(1+o(1))$, *for any* $x_1 \in \mathcal{X}_1$.

Lemma A.7: *Under Assumptions 5, B.1-B.4 and B.6, and if* $\frac{h_T h^{m-1}_{x_1,T}+h^m_T}{\sqrt{\lambda_{x_1,T}}} = o\left(b(\lambda_{x_1,T},h_{x_1,T})\right)$, *we have* $\left\|\mathcal{B}^r_{x_1,T}+\mathcal{B}^e_{x_1,T}\right\|_{L^2(\mathcal{X}_2)} = \left\|\mathcal{B}^r_{x_1,T}+h^m_{x_1,T}\left(\lambda_{x_1,T}+A^*_{x_1}A_{x_1}\right)^{-1}A^*_{x_1}\Xi_{x_1}\right\|_{L^2(\mathcal{X}_2)}(1+o(1))$, *for any* $x_1 \in \mathcal{X}_1$.

From (28) and Lemmas A.6 and A.7, Proposition 3 follows.

Appendix 4: Proof of Proposition 4

(i) Since $A^*_{x_1}A_{x_1} = \mathcal{D}^{-1}\widetilde{A}_{x_1}A_{x_1}$ (see Lemma A.1 (iv) with $l=1$) and $\mathcal{D}^{-1}\tilde{\phi}_{x_1,j} = \tau^{-1}_{x_1,j}\tilde{\phi}_{x_1,j}$, we have $A^*_{x_1}A_{x_1}\tilde{\phi}_{x_1,j} = \frac{\nu_{x_1,j}}{\tau_{x_1,j}}\tilde{\phi}_{x_1,j}$. The normalization of the eigenfunctions in $H^1(\mathcal{X}_2)$ is obtained from $\left\|\tilde{\phi}_{x_1,j}\right\|^2_{H^1(\mathcal{X}_2)} = \langle\tilde{\phi}_{x_1,j},\mathcal{D}\tilde{\phi}_{x_1,j}\rangle_{L^2(\mathcal{X}_2)} = \tau_{x_1,j}$ (see Lemma A.1 (iii) with $l=1$).

(ii) To compute the asymptotic behaviour of $\sigma^2_{x_1}(\lambda_{x_1,T})$ and $b_{x_1}(\lambda_{x_1,T}, h_{x_1,T})^2$ in Proposition 3, we use the next lemma.

Lemma A.8: *Let* $\nu_j \asymp j^{-\alpha_1}e^{-\alpha_2 j}$, $a_j \asymp j^{-\alpha_3}e^{-\alpha_4 j}$ *for* $\alpha_1, \alpha_3 \geq 0$, $\alpha_2, \alpha_4 > 0$. *Let* $n_\lambda \in \mathbb{N}$ *be such that* $\nu_{n_\lambda} \asymp \lambda$ *as* $\lambda \to 0$. *Then, as* $\lambda \to 0$*:*

$$\sum_{j=1}^{\infty} \frac{a_j}{(\lambda+\nu_j)^2} \asymp \begin{cases} \lambda^{-2+\alpha_4/\alpha_2} n_\lambda^{\frac{\alpha_1\alpha_4}{\alpha_2}-\alpha_3} & , \text{ if } \alpha_4 < 2\alpha_2 \\ 1 & , \text{ if } \alpha_4 > 2\alpha_2 \end{cases}.$$

From Lemma A.8 and Condition (a), we get $\sigma^2_{x_1}(\lambda_{x_1,T}) \asymp \dfrac{1}{\lambda_{x_1,T} n^{\beta}_{\lambda_{x_1,T}}}$. Now, by using that the functions $\phi_{x_1,j}$ are orthogonal w.r.t. $\langle ., . \rangle_{L^2(\mathcal{X}_2)}$ (see Part (i)), the squared bias function is given by $b_{x_1}(\lambda_{x_1,T}, h_{x_1,T})^2 = \displaystyle\sum_{j=1}^{\infty} \frac{\left(\lambda_{x_1,T} d_{x_1,j} - h^m_{x_1,T}\sqrt{\nu_{x_1,j}}\xi_{x_1,j}\right)^2}{(\lambda_{x_1,T}+\nu_{x_1,j})^2} \left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)}$. We develop the parentheses, and show by using Lemma A.8 and Condition (a) that $\lambda^2_{x_1,T} \displaystyle\sum_{j=1}^{\infty} \frac{d^2_{x_1,j}}{(\lambda_{x_1,T}+\nu_{x_1,j})^2} \left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)} \asymp \lambda^{2\delta}_{x_1,T} n^{(2\delta-1)\beta}_{\lambda_{x_1,T}}$ and $h^{2m}_{x_1,T} \displaystyle\sum_{j=1}^{\infty} \frac{\nu_{x_1,j}\xi^2_{x_1,j}}{(\lambda_{x_1,T}+\nu_{x_1,j})^2} \left\|\phi_{x_1,j}\right\|^2_{L^2(\mathcal{X}_2)} \asymp h^{2m}_{x_1,T}\lambda^{2\rho-1}_{x_1,T} n^{(2\rho-1)\beta}_{\lambda_{x_1,T}}$. In the Technical Report, we use Condition (b) to control the cross term in $b_{x_1}(\lambda_{x_1,T}, h_{x_1,T})^2$ and get $M_{x_1,T}(\lambda_{x_1,T}, h_{x_1,T}) \asymp \dfrac{1}{T h^{d_{X_1}}_{x_1,T}\lambda_{x_1,T} n^{\beta}_{\lambda_{x_1,T}}} + \lambda^{2\delta}_{x_1,T} n^{(2\delta-1)\beta}_{\lambda_{x_1,T}} + h^{2m}_{x_1,T}\lambda^{2\rho-1}_{x_1,T} n^{(2\rho-1)\beta}_{\lambda_{x_1,T}}$. Moreover, $n_{\lambda_{x_1,T}} = O(\log(1/\lambda_T))$. Thus, for $\lambda_{x_1,T}$ such that $\lambda_{x_1,T} \asymp T^{-\gamma}$, powers of $n_{\lambda_{x_1,T}}$ contribute multiplicative terms of logarithmic order, and Part (ii) follows.

(iii) The bandwidth and regularization parameter sequences that optimize the convergence rate of the MISE up to logarithmic terms are the minima of the function $\Psi_T(\lambda_{x_1,T}, h_{x_1,T}) = \dfrac{1}{T h^{d_{X_1}}_{x_1,T}\lambda_{x_1,T}} + \lambda^{2\delta}_{x_1,T} + h^{2m}_{x_1,T}\lambda^{2\rho-1}_{x_1,T}$. The partial derivatives are given by $\dfrac{\partial \Psi_T}{\partial h_{x_1,T}} = -\dfrac{d_{X_1}}{T h^{d_{X_1}+1}_{x_1,T}\lambda_{x_1,T}} + 2m h^{2m-1}_{x_1,T}\lambda^{2\rho-1}_{x_1,T}$ and $\dfrac{\partial \Psi_T}{\partial \lambda_{x_1,T}} = -\dfrac{1}{T h^{d_{X_1}}_{x_1,T}\lambda^2_{x_1,T}} + 2\delta\lambda^{2\delta-1}_{x_1,T} + (2\rho-1) h^{2m}_{x_1,T}\lambda^{2\rho-2}_{x_1,T}$. By setting these

partial derivatives equal to zero, we get:

$$h_{x_1,T}^{2m} = \frac{2\delta}{\frac{2m}{d_{X_1}} + 1 - 2\rho} \lambda_{x_1,T}^{2(\delta-\rho)+1}. \tag{29}$$

By plugging this into function Ψ_T, we get the concentrated function $\Psi_T(\lambda_{x_1,T}) = \frac{c_1}{T\lambda_{x_1,T}^{1+\frac{d_{X_1}}{2m}(2(\delta-\rho)+1)}} + c_2\lambda_{x_1,T}^{2\delta}$, for some constants $c_1, c_2 > 0$. By minimizing this function w.r.t. $\lambda_{x_1,T}$, the optimal rate γ for the regularization parameter follows. Then, the optimal rate η for the bandwidth is deduced from (29). Finally, by plugging the optimal $\lambda_{x_1,T}$ and $h_{x_1,T}$ into $\Psi_T(\lambda_{x_1,T}, h_{x_1,T})$, the optimal rate of the MISE follows.

References

Abramowitz, M. and I. Stegun, 1970, Handbook of mathematical functions, Dover Publications, New York.

Adams, R. and J. Fournier (2003): *Sobolev Spaces*, Academic Press, Boston.

Ai, C. and X. Chen, 2003, Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, 1795-1843.

Andrews, D., 1994, Empirical process methods in econometrics, in: R. Engle and D. McFadden, (Eds.), Handbook of Econometrics, Vol. 4, North Holland, Amsterdam, 2247-2294.

Banks, J., Blundell, R. and A. Lewbel, 1997, Quadratic Engel curves and consumer demand. Review of Economics and Statistics 79, 527-539.

Blundell, R., X. Chen and D. Kristensen, 2007, Semi-nonparametric IV estimation of shape invariant Engel curves. Econometrica 75, 1613-1669.

Blundell, R. and J. Horowitz, 2007, A non-parametric test of exogeneity. Review of Economic Studies 74, 1035-1058.

Blundell, R. and J. Powell, 2003, Endogeneity in semiparametric and nonparametric regression models, in: M. Dewatripont, L. Hansen, and S. Turnovsky (Eds.), Advances in economics and econometrics: theory and applications, Cambridge University Press, Cambridge, pp. 312-357.

Carrasco, M. and J.-P. Florens, 2011, Spectral method for deconvolving a density. Econometric Theory 27 (3), 546-581.

Carrasco, M., Florens, J.-P. and E. Renault, 2007, Linear inverse problems in structural econometrics: estimation based on spectral decomposition and regularization, in: J. Heckman and E. Leamer (Eds.), Handbook of Econometrics, Vol. 6, Part 2, North Holland, Amsterdam, pp. 5633-5751.

Chen, X., 2007, Large sample Sieve estimation of semi-nonparametric models, in: J. Heckman and E. Leamer (Eds.), Handbook of Econometrics, Vol. 6, Part 2, North Holland, Amsterdam, pp. 5549-5632.

Chen, X. and S. Ludvigson, 2009, Land of addicts? An empirical investigation of habit-based asset pricing models. Journal of Applied Econometrics 24, 1057-1093.

Chen, X. and D. Pouzo, 2009, Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. Journal of Econometrics 152, 46-60.

Chen, X. and D. Pouzo, 2011, Estimation of nonparametric conditional moment models with possibly nonsmooth moments. Forthcoming in Econometrica.

Chernozhukov, V., Gagliardini, P. and O. Scaillet, 2006, Nonparametric instrumental variable estimation of structural quantile effects. Working Paper.

Chernozhukov, V. and C. Hansen, 2005, An IV model of quantile treatment effects. Econometrica 73, 245-271.

Chernozhukov, V., Imbens, G. and W. Newey, 2007, Instrumental variable estimation of nonseparable models. Journal of Econometrics 139, 4-14.

Darolles, S., Fan, Y., Florens, J.-P. and E. Renault, 2011, Nonparametric instrumental regression. Forthcoming in Econometrica.

Florens, J.-P., 2003, Inverse problems and structural econometrics: the example of instrumental variables, in: M. Dewatripont, L. Hansen, and S. Turnovsky (Eds.), Advances in economics and econometrics: theory and applications, Cambridge University Press, Cambridge, 284-311.

Florens, J.-P., Johannes, J. and S. Van Bellegem, 2005, Instrumental regression in partially linear models. Working Paper.

Gagliardini, P. and C. Gouriéroux, 2007, An efficient nonparametric estimator for models with nonlinear dependence. Journal of Econometrics 137, 189-229.

Gagliardini, P. and O. Scaillet, 2006, Tikhonov regularization for functional minimum distance estimators. Working Paper.

Gagliardini, P. and O. Scaillet, 2007, A specification test for nonparametric instrumental variable regression. Working Paper.

Groetsch, C. W., 1984, The theory of Tikhonov regularization for Fredholm equations of the first kind, Pitman Advanced Publishing Program, Boston.

Hall, P. and J. Horowitz, 2005, Nonparametric methods for inference in the presence of instrumental variables. Annals of Statistics 33, 2904-2929.

Hansen, B., 2008, Uniform convergence rates for kernel estimation with dependent data. Econometric Theory 24, 726-748.

Hoderlein, S. and H. Holzmann, 2011, Demand analysis as an ill-posed inverse problem with semiparametric specification. Econometric Theory, 27 (3), 609-638.

Horowitz, J., 2006, Testing a parametric model against a nonparametric alternative with identification through instrumental variables. Econometrica 74, 521-538.

Horowitz, J., 2007, Asymptotic normality of a nonparametric instrumental variables estimator. International Economic Review 48, 1329-1349.

Horowitz, J. and S. Lee, 2007, Nonparametric instrumental variables estimation of a quantile regression model. Econometrica 75, 1191-1208.

Hu, Y. and S. Schennach, 2008, Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions using instruments. Econometrica 76, 195-216.

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, 2008, Nonparametric transformation to white noise. Journal of Econometrics 142, 241-264.

Loubes, J.-M. and A. Vanhems, 2004, Estimation of the solution of a differential equation with endogenous effect. Working Paper.

Mammen, E., Linton, O. and J. Nielsen, 1999, The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Annals of Statistics 27, 1443-1490.

Newey, W. and D. McFadden, 1994, Large sample estimation and hypothesis testing, in: R. Engle and D. McFadden (Eds.), Handbook of Econometrics, Vol. 4, pp. 2111-2245.

Newey, W. and J. Powell, 2003, Instrumental variable estimation of nonparametric models. Econometrica 71, 1565-1578.

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., 1963a, On the solution of incorrectly formulated problems and the regularization method. Soviet Mathematics Doklady 4, 1035-1038 (English translation).

Tikhonov, A., 1963b, Regularization of incorrectly posed problems. Soviet Mathematics Doklady 4, 1624-1627 (English translation).

White, H. and J. Wooldridge, 1991, Some results on Sieve estimation with dependent observations, in: W. Barnett, J. Powell and G. Tauchen (Eds.), Nonparametric and semiparametric methods in econometrics and statistics, Cambridge University Press, Cambridge, pp. 459-493.

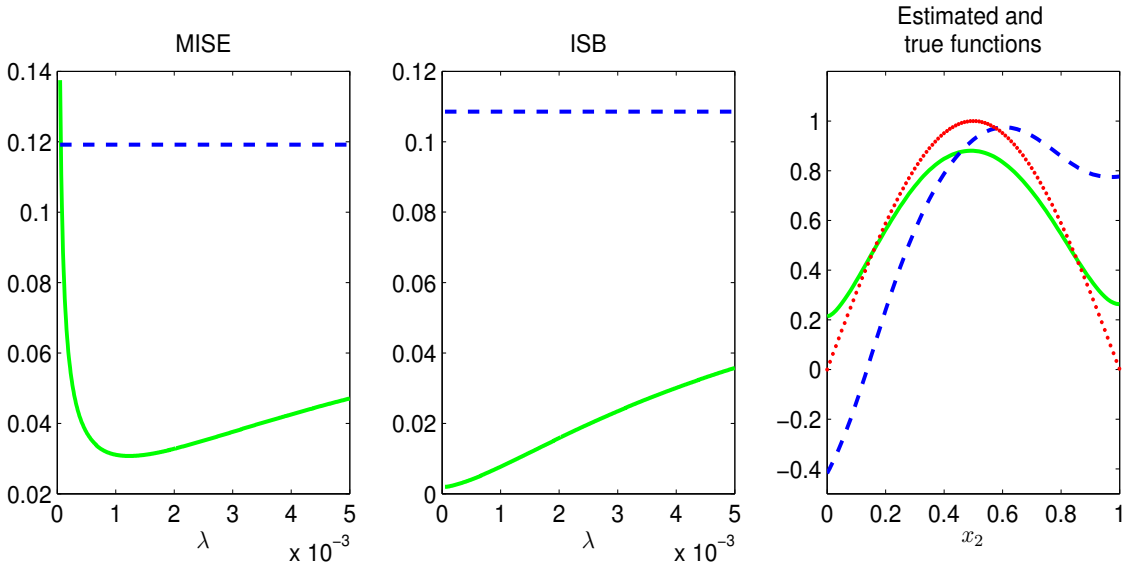

Figure 1: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for the kernel regression estimator (dashed line). The true function is the dotted line in the right panel. Correlation parameter is $\rho = 0.5$, value of the exogenous variable X_1 is fixed at $x_1 = \Phi(0)$, and sample size is $T = 1000$.

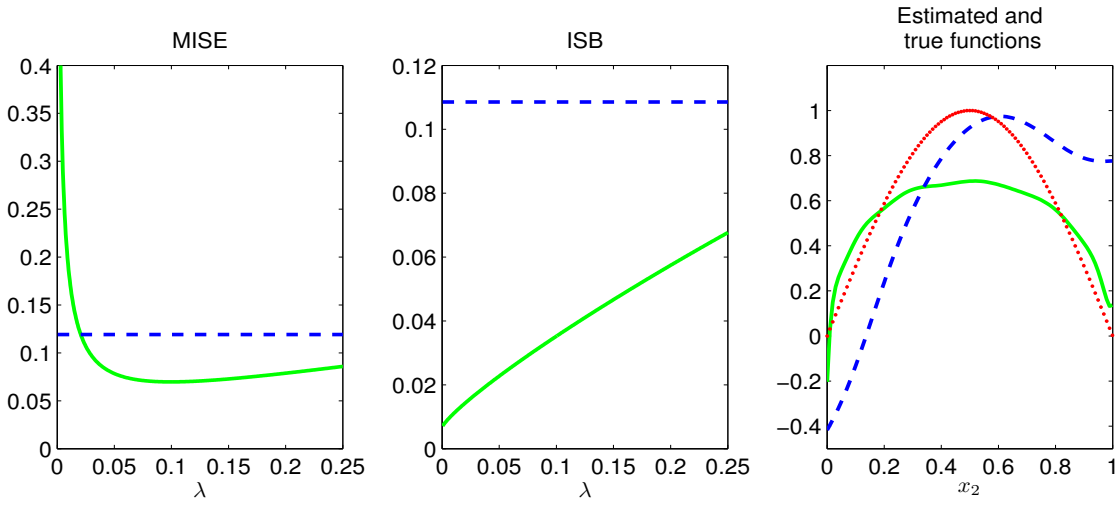

Figure 2: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the regularised estimator using L^2 norm (solid line) and for the kernel regression estimator (dashed line). The true function is the dotted line in the right panel. Correlation parameter is $\rho = 0.5$, value of the exogenous variable X_1 is fixed at $x_1 = \Phi(0)$, and sample size is $T = 1000$.

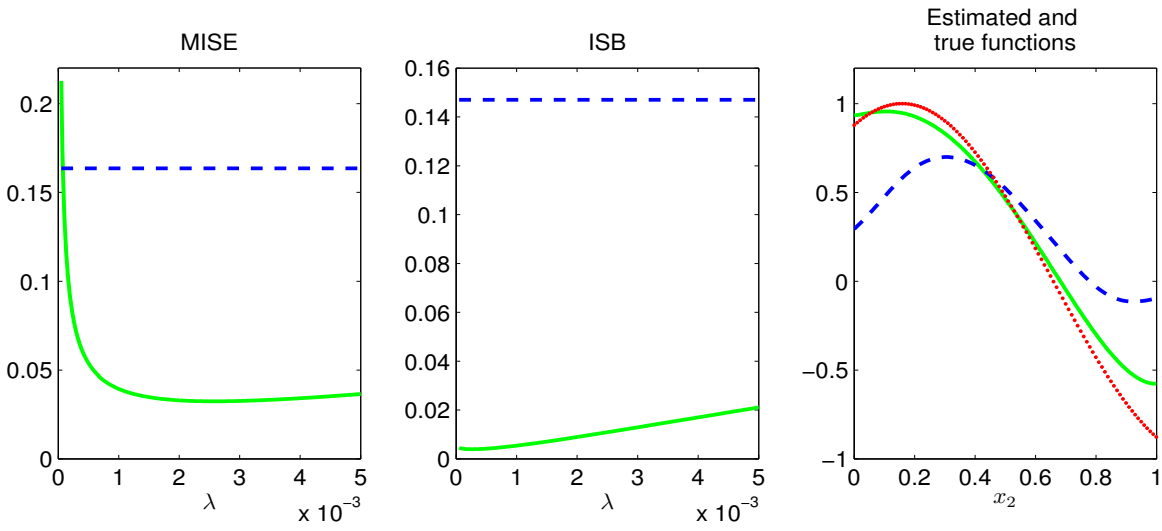

Figure 3: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for the kernel regression estimator (dashed line). The true function is the dotted line in the right panel. Correlation parameter is $\rho = 0.5$, value of the exogenous variable is X_1 is fixed at $x_1 = \Phi(1)$, and sample size is $T = 1000$.

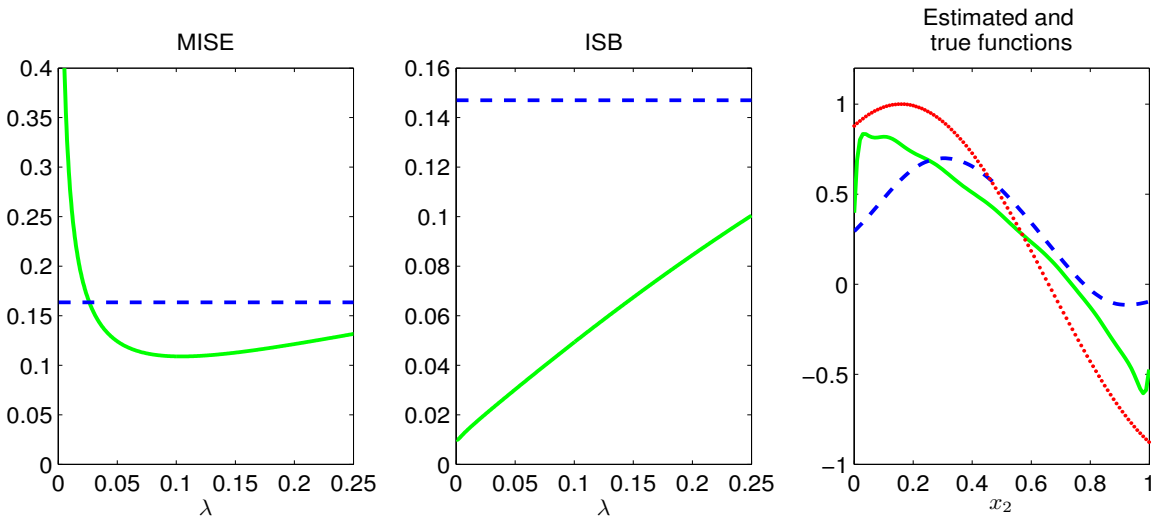

Figure 4: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the regularised estimator using L^2 norm (solid line) and for the kernel regression estimator (dashed line). The true function is the dotted line in the right panel. Correlation parameter is $\rho = 0.5$, value of the exogenous variable is X_1 is fixed at $x_1 = \Phi(1)$, and sample size is $T = 1000$.

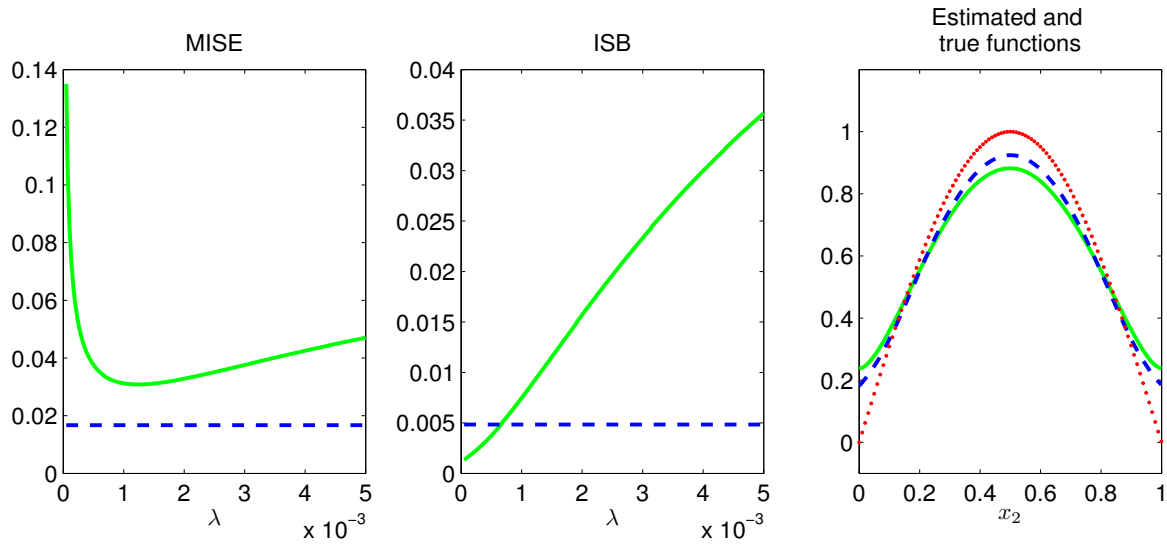

Figure 5: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for the kernel regression estimator (dashed line). The true function is the dotted line in the right panel. Correlation parameter is $\rho = 0$, value of the exogenous variable is X_1 is fixed at $x_1 = \Phi(0)$, and sample size is $T = 1000$.

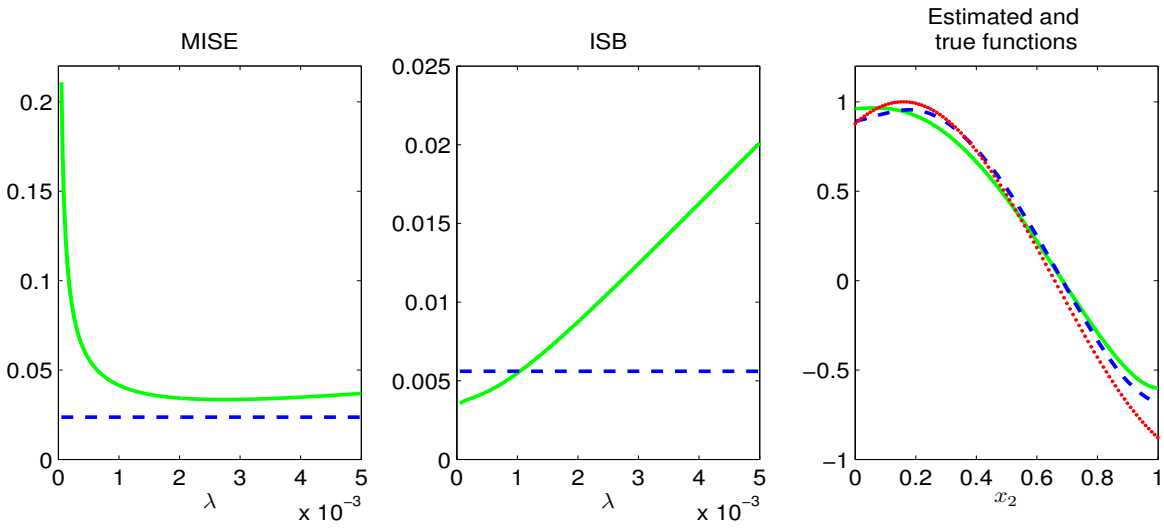

Figure 6: MISE (left panel), ISB (central panel) and mean estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for the kernel regression estimator (dashed line). The true function is the dotted line in the right panel. Correlation parameter is $\rho = 0$, value of the exogenous variable is X_1 is fixed at $x_1 = \Phi(1)$, and sample size is $T = 1000$.

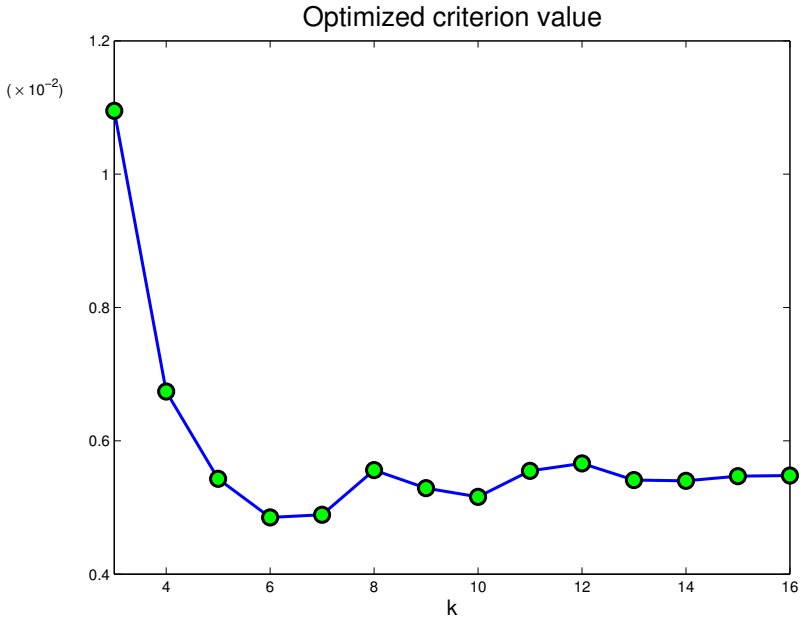

Figure 7: Value of the optimized objective function as a function of the number k of polynomials. The regularization parameter is selected with a data-driven approach.

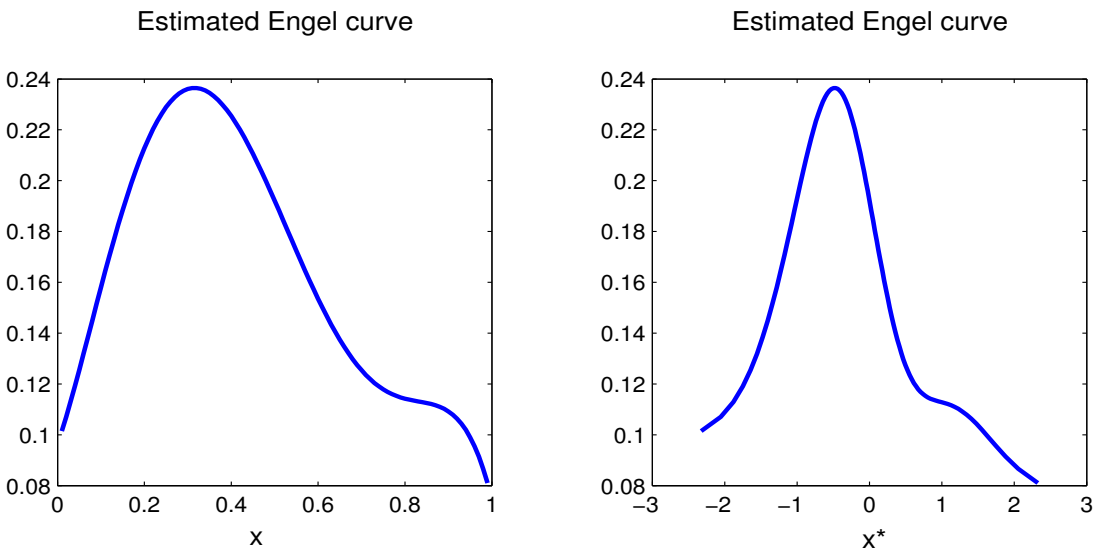

Figure 8: Estimated Engel curves for 785 household-level observations from the 1996 US Consumer Expenditure Survey. Food expenditure share Y is plotted as a function of the standardized logarithm X_2^* of total expenditures (right panel), and of transformed variable $X_2 = \Phi(X_2^*)$ with support $[0, 1]$, where Φ is the cdf of the standard normal distribution (left panel). Instrument Z_1 is the standardized logarithm of annual income from wages and salaries.