WORKING PAPERS

N° 1467

August 2023

“One-step nonparametric instrumental regression using smoothing splines”

Jad Beyhum, Elia Lapenta and Pascal Lavergne

Toulouse School of Economics

One-step nonparametric instrumental regression using smoothing splines

Jad Beyhum
Department of Economics, KU Leuven

Elia Lapenta
CREST and ENSAE Paris

Pascal Lavergne *
Toulouse School of Economics, Université Toulouse Capitole

August 31, 2023

Abstract

We extend nonparametric regression smoothing splines to a context where there is endogeneity and instrumental variables are available. Unlike popular existing estimators, the resulting estimator is one-step and relies on a unique regularization parameter. We derive uniform rates of the convergence for the estimator and its first derivative. We also address the issue of imposing monotonicity in estimation. Simulations confirm the good performances of our estimator compared to some popular two-step procedures. Our method yields economically sensible results when used to estimate Engel curves.

Keywords: Instrumental variables, Nonparametric regression, Smoothing splines.

*Pascal Lavergne acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future program (Investissements d'Avenir, grant ANR-17-EURE-0010).
The authors thank Ben Deaner, Frank Windmeijer, Jasmin Fliegner, Daniel Wilhelm, Aureo de Paula, Kaspar Wuthrich, Christoph Breunig, Joachim Freyberger, seminar and conference participants at the University of Bonn, the 2023 Encounters in Econometric Theory at Nuffield College, Oxford University, and the 2023 Bristol Econometrics Study Group conference.

1 Introduction

The estimation of causal effects on outcomes is often complicated by omitted confounding variables or nonrandom selection. The problem is well known to affect observational studies, but it can also affect randomized controlled trials with issues like participant noncompliance. In economics, the problem is commonly framed in terms of a regression model from which important regressors have been omitted and so become part of the model's error. In this context, a variable is termed "exogenous" if it is not associated with the error term after conditioning on confounding variables, and "endogenous" otherwise. Instrumental variables (IVs) are widely used to solve the problems posed by endogeneity. An IV is associated with the outcome only through its association with the endogenous variable. Examples of IV models include estimation of Engel curves linking expenditures to the consumption share of a particular good, or estimation of the effects of class size on students' performances (Horowitz, 2011).

We consider the prototypical model

$$Y = g_0(Z) + \varepsilon \qquad \mathrm{E}\left(\varepsilon|W\right) = 0\,, \tag{1}$$

where $Y \in \mathbb{R}$ is the dependent variable, $Z \in \mathbb{R}$ is the endogenous continuous independent variable, and $W \in \mathbb{R}^p$ are the IVs. The goal is to estimate nonparametrically $g_0(\cdot)$, the causal effect of the variable Z on Y, using W to account for endogeneity. If we assumed linear relationships, we could use the two-stage least squares estimator: in a first stage, one obtains the linear projection of Z on W, then in a second stage one linearly regresses Y on the previously estimated linear projection. Considering a nonparametric function $g_0(\cdot)$ allows estimating the causal relationship of Y and Z in a more flexible manner. As detailed below, most existing nonparametric methods are also two-stage.

We here develop a nonparametric instrumental regression estimator for g_0 that bypasses first stage nonparametric estimation, is easy to implement, and leverages the excellent approximation properties of splines. To do so, we exhibit a global quantity that accounts for all the information contained in Model (1) and that is minimized by the true function $g_0(\cdot)$. We then consider an empirical equivalent, and we set up a minimization problem penalized by a roughness measure of the function to regularize the solution. Our estimator can be seen as a generalization to the instrumental variable regression of regression smoothing splines (Wahba, 1990; Green and Silverman, 1993). As a leading case, we consider the integral of the squared second derivative as a roughness measure, and we show that the solution is a natural cubic spline. We derive a closed-form expression of the solution. A characteristic feature of our estimator is that it depends upon a single regularization parameter, which can then be selected by some usual methods such as cross-validation. As an additional advantage of the smoothing spline nature of our estimator, straightforward estimators of derivatives of the function $g_0(\cdot)$ obtain. We further show how to impose monotonicity constraints by relying on a method proposed by Hall and Huang (2001). The constrained estimator is simple to implement in practice.

We derive uniform rates of convergences for the function itself and its derivative, with or without monotonicity constraints. Our estimator exhibits excellent finite sample performance in simulations compared to some existing two-step methods. We applied our method to the estimation of Engel curves. Our smoothing splines estimator and its constrained version yield comparable results and are reasonable from an economic viewpoint.

Existing procedures for estimating g_0 are typically two-step and do not make use of smoothing splines. Newey and Powell (2003) develop a nonparametric equivalent to the two-stage least squares estimator: they use linear-in-parameter series expansions of $\mathrm{E}(Y|W)$ and $\mathrm{E}\{g(Z)|W\}$ in a generalized method of moments framework, see also Ai and Chen (2003); Hall and Horowitz (2005); Blundell et al. (2007); Johannes et al. (2011); Horowitz (2014) for alternative series-based methods. Alternatively, Hall and Horowitz (2005), Darolles et al. (2011), and Gagliardini and Scaillet (2012) rely on kernel methods to estimate the unknown conditional expectations. In both cases, backing up a nonparametric estimate of $g_0(\cdot)$ is an ill-posed inverse problem. Hence, one needs some kind of regularization, such as hard thresholding (Horowitz, 2011; Chen and Pouzo, 2012), Tikhonov or ridge-type regularization (Newey and Powell, 2003; Darolles et al., 2011; Florens et al., 2011; Gagliardini and Scaillet, 2012; Singh et al., 2019), or a Landweber-type iterative method (Dunker et al., 2014). A general exposition of some of these methods is given by Carrasco et al. (2007). A recent machine learning literature considers solving a saddle point problem that is dual to a generalized method of moments criterion. Here one first maximizes an objective function with respect to a function of the instruments W, then one minimizes with respect to a function of Z to obtain $g_0(\cdot)$ (Bennett et al., 2019; Dikkala et al., 2020; Liao et al., 2020). Muandet et al. (2020) consider a related but different saddle point problem. If the set of functions upon which one optimizes is large, then one has in addition to introduce some penalization in the optimization problem (Dikkala et al., 2020; Liao et al., 2020).

By contrast to previous estimators based on series or kernel estimation, ours is a natural generalization of the popular regression smoothing splines estimator. The appeal of smoothing splines lies in their simplicity together with their excellent approximation properties of smooth functions and their derivatives, which have been extensively studied

(Schumaker, 2007). Our estimator cannot be obtained as a generalization or a special case of existing estimators. Our approach is also attractive because it is one-step. Two-stage procedures typically lead to theoretical and practical difficulties. First, one may need to estimate in a first-stage an object that may be more complex than the final object of interest, while imposing smoothness assumptions on this first-stage object. Second, from a statistical perspective, first-stage estimation typically affects the second-stage small sample and asymptotic properties. Third, each stage relies on a particular choice of a smoothing or regularization parameter, whose fine tuning may be difficult in practice while affecting the final results.

The paper is organized as follows. In Section 2, we detail the main steps that lead to the formulation of our estimator, and we show that it extends regression smoothing splines to the instrumental regression context. The asymptotic properties of our estimator are analyzed in Section 3. Section 4 deals with estimation under monotonicity constraints. Section 5 reports simulation results and an empirical application to Engel curves estimation.

2 Our estimator

2.1 General formulation

We assume that $g_0(\cdot)$ belong to some space of functions $\mathcal{G}$ on which identification holds, that is,

$$\mathrm{E}\left\{Y - g(Z)|W\right\} = 0 \quad \text{a.s.} \Rightarrow g(\cdot) = g_0(\cdot) \quad \text{a.s.} \tag{2}$$

For a discussion of this condition called *completeness*, see e.g. D'Haultfoeuille (2011) and Freyberger (2017). When Z is continuous, as we assume here, W should typically have at least one continuous component for completeness to hold. Some of the instruments, however, could be discrete, and this will not affect further our exposition and reasoning.

Instead of dealing directly with (2), as done by most previous work, we consider an equivalent formulation that does not require estimating a conditional expectation given the instruments W. By the results of Bierens (1982),

$$\mathrm{E}\left\{Y - g(Z)|W\right\} = 0 \Leftrightarrow \mathrm{E}\left\{(Y - g(Z))\exp(\mathbf{i}W^\top t)\right\} = 0 \quad \forall t \in \mathbb{R}^p\,, \tag{3}$$

where $\mathbf{i}$ is the unit imaginary number. Consider now

$$g_0 = \arg\min_{g\in\mathcal{G}} M(g), \qquad M(g) = \int \left|\mathrm{E}\left\{(Y - g(Z))\exp(\mathbf{i}W^\top t)\right\}\right|^2 d\mu(t)\,, \tag{4}$$

where μ is a symmetric probability measure with support $\mathbb{R}^p$. Then it is straightforward to see that $M(g) \geq 0$ for all $g \in \mathcal{G}$, and that under (2)

$$M(g) = 0 \Leftrightarrow g = g_0 \,.$$

With a random i.i.d. sample $\{(Y_i, Z_i, W_i), i = 1, \ldots n\}$ at hand, a natural estimator of $M(g)$ is

$$\begin{aligned} M_n(g) &= \int \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(Z_i)) \exp(\mathbf{i} W_i^\top t) \right|^2 d\mu(t) \\ &= \frac{1}{n^2} \sum_{1 \leq i,j \leq n} (Y_i - g(Z_i)) (Y_j - g(Z_j)) \, \omega(W_i - W_j) \,, \end{aligned} \tag{5}$$

where

$$\omega(z) = \int_{\mathbb{R}^p} \exp(it'w) \, d\mu(t) = \int_{\mathbb{R}^p} \cos(t'w) \, d\mu(t) \,,$$

due to the symmetry of μ. The function $\omega(\cdot)$ is (up to a constant) the Fourier transform of the density of μ. The above V-statistic formulation is used in practice for computational purposes. This statistic accounts for an infinity of moment conditions as stated in (3). It is different in nature from generalized method of moments criteria used in previous work, that account for an increasing but finite number of moment conditions.

The condition for μ to have support $\mathbb{R}^p$ translates into the restriction that $\omega(\cdot)$ should have a strictly positive Fourier transform almost everywhere. Examples include products of triangular, normal, logistic, see Johnson et al. (1995, Section 23.3), Student, including Cauchy, see Dreier and Kotz (2002), or Laplace densities. To achieve scale invariance, we recommend, as in Bierens (1982), to scale the exogenous instruments by a measure of dispersion, such as their empirical standard deviation. If W has bounded support, results from Bierens (1982) yield that the equivalence (3) holds when restricted to any t in a (arbitrary) neighborhood of 0 in $\mathbb{R}^p$. Hence, μ can be taken as any symmetric probability measure that contains 0 in the interior of its support. As noted by Bierens (1982), there is no loss of generality assuming a bounded support, as the equivalence result equally applies to a one-to-one transformation of W, which can be chosen with bounded image.

Minimizing $M_n(g)$ would lead to interpolation. We regularize the problem by assuming some smoothness for the function g. We assume that Z has compact support, say $[0, 1]$ without loss of generality, and that $\mathcal{G}$ is the space of twice differentiable functions on $[0, 1]$ defined as

$$\mathcal{G} = \left\{ g : [0, 1] \to \mathbb{R}, \int_0^1 |g''(t)|^2 \, dt < \infty \right\} \,.$$

We then estimate g_0 as a minimizer of a penalized version of $M_n(g)$ on $\mathcal{G}$. Specifically,

$$\widehat{g} \in \arg\min_{g\in\mathcal{G}} M_n(g) + \lambda \int_0^1 |g''(z)|^2\, dz\,, \tag{6}$$

where $\lambda > 0$ is a regularization parameter. Our estimator $\widehat{g}$ is a natural cubic spline with knots at points Z_i. This is implied by the following result of Green and Silverman (1993, Th. 2.3).

Theorem 2.1. *Suppose $n \geq 2$ and $0 \leq Z_1 < \cdots < Z_n \leq 1$. Let $\widetilde{g}$ be any function in $\mathcal{G}$ for which $\widetilde{g}(Z_i) = g_i$, $i = 1, \ldots, n$, and g a natural cubic spline such that $g(Z_i) = g_i$, $i = 1, \ldots, n$. Then*

$$\int_0^1 |\widetilde{g}''(t)|^2\, dt \geq \int_0^1 |g''(t)|^2\, dt\,.$$

with equality only if $\widetilde{g} = g$.

It thus follows that if $\widehat{g}$ is a solution to (6), it should be a natural cubic spline. The uniqueness of the above natural cubic spline interpolant is proven in Green and Silverman (1993, Th. 2.2).

A recent approach we became aware of when preparing this paper is the "kernel maximum moment loss" approach proposed by Zhang et al. (2023). While it does not smooth on the instruments, it assumes that the regression of interest belongs to a Reproducing Kernel Hilbert Space (RKHS) and solves a minimization problem by penalizing by the norm on such a space. The estimator thus depends on the chosen reproducing kernel. Differently, we assume that the regression of interest belongs to a space of smooth functions, and we penalize by the integral of the squared second derivative of the regression, which is a very intuitive measure of roughness, but not a RKHS norm.

2.2 Closed-form solution

We now show the existence and uniqueness of the solution to (6) and we characterize it. A natural cubic spline can be written as

$$g(z) = a_0 + a_1 z + \frac{1}{12}\sum_{i=1}^n \delta_i |z - Z_i|^3\,, \quad \sum_{i=1}^n \delta_i = \sum_{i=1}^n \delta_i Z_i = 0\,. \tag{7}$$

The function g is thus uniquely defined by the coefficients a_0, a_1, and $\delta_i, i = 1, \ldots n$, or equivalently by its value at the knots (Green and Silverman, 1993, Section 7.3). It will be useful for what follows to use matrix notations. Let

$$\boldsymbol{Z} = \begin{pmatrix} 1 & Z_1 \\ \vdots & \vdots \\ 1 & Z_n \end{pmatrix}\,,$$

$\boldsymbol{E} = \left[\frac{1}{12}|Z_i - Z_j|^3, i,j = 1,\dots n\right]$, and $\boldsymbol{g} = (g(Z_1),\dots g(Z_n))^T$. Then $\boldsymbol{g} = \boldsymbol{Za} + \boldsymbol{E\delta}$ with constraints $\boldsymbol{Z^T\delta} = 0$. Also, one can check that

$$\int g''(z)^2\,dz = \boldsymbol{\delta^T E\delta}\,,$$

see Green and Silverman (1993, Section 7.3). Let $\boldsymbol{Y}$ be the vector $(Y_1,\dots Y_n)^T$, then

$$M_n(g) + \lambda\int \left(g''(z)\right)^2\,dz = \left(\boldsymbol{Y} - \boldsymbol{Za} - \boldsymbol{E\delta}\right)^T \boldsymbol{\Omega}\left(\boldsymbol{Y} - \boldsymbol{Za} - \boldsymbol{E\delta}\right) + \lambda\boldsymbol{\delta^T E\delta}\,, \tag{8}$$

where $\boldsymbol{\Omega}$ is the matrix with generic element $n^{-2}\omega(W_i - W_j)$. Hence, we want to minimize a quadratic function in parameters under the constraints $\boldsymbol{Z^T\delta} = 0$. This yields a unique solution under the usual requirements. The following proposition gives a more convenient formula that characterizes our estimator.

Proposition 2.1. *For any $\lambda > 0$, if $\boldsymbol{Z}$ is full rank and all W_i's are different, the solution to (6) exists, is unique, and is a natural cubic spline $\widehat{g}$ characterized by*

$$\begin{bmatrix} \widetilde{\boldsymbol{E}} & \boldsymbol{Z} \\ \boldsymbol{Z^T} & \boldsymbol{0} \end{bmatrix} \begin{pmatrix} \widehat{\boldsymbol{\delta}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}, \qquad \widetilde{\boldsymbol{E}} = \boldsymbol{E} + \lambda\boldsymbol{\Omega}^{-1}\,, \tag{9}$$

and $\widehat{\boldsymbol{g}} = \left[\boldsymbol{P} + \boldsymbol{E}\widetilde{\boldsymbol{E}}^{-1}(\boldsymbol{I} - \boldsymbol{P})\right]\boldsymbol{Y}$, where $\boldsymbol{P} = \boldsymbol{Z}\left(\boldsymbol{Z}^T\widetilde{\boldsymbol{E}}^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\widetilde{\boldsymbol{E}}^{-1}$.

Our estimator is obtained directly by solving the linear system of equations (9). It does not necessitate estimation of other nonparametric quantities, and relies on only one regularization parameter λ. It also directly provides an estimator of the first derivative of $g(\cdot)$ as

$$\widehat{g}'(z) = \widehat{a}_1 + \frac{1}{4}\sum_{i=1}^n \widehat{\delta}_i\,\mathrm{sign}(z - Z_i)(z - Z_i)^2\,, \qquad \mathrm{sign}(u) = \mathbf{1}(u \geq 0) - \mathbf{1}(u < 0)\,. \tag{10}$$

There are alternative ways to (7) for expressing a natural cubic spline. We focus on this formulation as it does not rely on a particular support of Z, nor on the fact that the knots Z_i are arranged in increasing order. In particular, the closed-form expression in Proposition 2.1 is valid regardless of the support of Z and therefore it can be used without first transforming Z into $[0,1]$. We also found this formulation to be convenient for practical implementation. For large samples, where the above formula may not be computationally efficient, one can adapt to our context the Reinsch algorithm (Green and Silverman, 1993).

3 Asymptotic analysis

The formal study of our estimator is based on a reformulation of $M(g)$ in (4). Consider

$$\mathcal{H} = \{h \in \mathcal{G} : h(0) = h'(0) = 0\},$$

and the inner product $\langle h_1, h_2\rangle_{\mathcal{H}} = \int_0^1 h_1''(z)\, h_2''(z)\, dz$ on $\mathcal{H}$. Then each $g \in \mathcal{G}$ can be uniquely written as $g(z) = (1, z)\beta + h(z)$, where $\beta = (g(0), g'(0))^T \in \mathbb{R}^2$, $h(z) = g(z) - g(0) - g'(0)z$, $h \in \mathcal{H}$. Denote by L^2_μ the space of complex functions $l(\cdot)$ from $\mathbb{R}^p$ into $\mathbb{C}$ such that

$$\|l\|^2_\mu = \int |l(t)|^2\, d\mu(t) < \infty\,.$$

Consider the operators $\mathrm{A} : \mathcal{H} \mapsto L^2_\mu$ and $B : \mathbb{R}^2 \mapsto L^2_\mu$ such that

$$\mathrm{A}\, h = \mathrm{E}\{h(Z) \exp(\mathbf{i}W^T\cdot)\} \text{ and } \mathrm{B}\,\beta = \mathrm{E}\{(1, Z)\beta \exp(\mathbf{i}W^T\cdot)\}\,. \tag{11}$$

The minimization problem (4) identifying g_0 can be expressed as

$$\min_{(\beta,h)\in\mathbb{R}^2\times\mathcal{H}} \|\, \mathrm{E}\{Y \exp(\mathbf{i}W^T\cdot)\} - \mathrm{B}\,\beta - \mathrm{A}\, h\|^2_\mu\,. \tag{12}$$

The above quantity reaches its minimum zero at (β_0, h_0), with $g_0(z) = (1, z)\beta_0 + h_0(z)$. A key advantage of this formulation for theoretical analysis is that using orthogonal projection, we can profile (12) to first determine h_0, then β_0 as a function of h_0. In our proofs, we will also consider the penalized empirical counterpart of (12) and use a similar profiling method to obtain $(\widehat{\beta}, \widehat{h})$, and then $\widehat{g}(z) = (1, z)\widehat{\beta} + \widehat{h}(z)$.

We now detail our formal assumptions.

Assumption 3.1. *(i)* $\mathrm{E}\{Y^2\} < \infty$*; (ii)* Z *has a density* f_Z *on* $[0, 1]$*; (iii)* μ *is a symmetric probability measure with support* $\mathbb{R}^p$*. (iv)* $\int\int \mathrm{E}\{\exp(\mathbf{i}W^\top t) f_Z(z)\}^2 \mu(t) dz\, dt < \infty$*; (v)* W *has at least one continuous component.*

This assumption ensures that $\mathrm{E}\{Y \exp(\mathbf{i}W^T\cdot)\} \in L^2_\mu$, and that A and B are into L^2_μ. Our assumption on the support of Z is without much loss of generality, since we can always use a one-to-one transformation that maps Z into $[0, 1]$. We then formalize the completeness assumption, under which the problem (12) admits a unique solution (β_0, h_0).

Assumption 3.2. g_0 *belongs to* $\mathcal{G}$ *and the mapping* $g \in \mathcal{G} \mapsto \mathrm{E}\{g(Z)|W = \cdot\}$ *is injective.*

We introduce now a *source condition*, which is common in the literature on inverse problems (Carrasco et al., 2007). While it is not needed to establish the consistency of $\widehat{g}$ and its first derivative, it is necessary to obtain their uniform convergence rates.

Assumption 3.3. *Let M be the orthogonal projection onto the orthogonal of the span of B, and let $\mathrm{T} = \mathrm{M}\,\mathrm{A}$. Let $(\sigma_j, \varphi_j, \psi_j)_j$ be the singular system of T, where $(\varphi_j)_j$ is a sequence of orthonormal elements in $\mathcal{H}$, $(\psi_j)_j$ is a sequence of orthonormal elements in L^2_μ, and $(\sigma_j)_j$ is a sequence of strictly positive values in $\mathbb{R}$. Then there exists $\gamma > 0$ such that*

$$\sum_j \sigma_j^{-2\gamma} | \langle h_0, \varphi_j \rangle_{\mathcal{H}} |^2 < \infty \, .$$

Theorem 3.1. *Under Assumptions 3.1, 3.2, if $\lambda \to 0$ and $n\lambda \to \infty$, then*

$$\sup_{z\in[0,1]} |\widehat{g}(z) - g_0(z)| = o_P(1) \;\; \textit{and} \;\; \sup_{z\in[0,1]} |\widehat{g}'(z) - g_0'(z)| = o_P(1) \, .$$

If moreover Assumption 3.3 holds, then

$$\sup_{z\in[0,1]} |\widehat{g}(z) - g_0(z)| = O_P\left(\frac{1}{\sqrt{n\lambda}} + \lambda^{\frac{\gamma\wedge 2}{2}}\right) \;\; \textit{and} \;\; \sup_{z\in[0,1]} |\widehat{g}'(z) - g_0'(z)| = O_P\left(\frac{1}{\sqrt{n\lambda}} + \lambda^{\frac{\gamma\wedge 2}{2}}\right) .$$

We obtain consistency of our estimator and its derivative under mild assumptions, that only involve a standard condition on the regularization parameter λ. By contrast, in two-step estimation methods that smooth over the instruments, obtaining consistency of the first and second steps estimators typically necessitates smoothness assumptions on the distribution of Z given W and conditions that relate the different smoothing parameters, see e.g. Ai and Chen (2003); Chen and Pouzo (2012). One partial solution adopted by Horowitz (2011) is to choose the latter parameters equal, but this does not change the two-step nature of the underlying estimator. In some instances, consistency may further necessitate regularization parameters (Chen and Pouzo, 2012) and a source condition (Gagliardini and Scaillet, 2012). A general discussion can be found in Carrasco et al. (2007).

Turning now to our consistency rates, we do not claim that these are sharp or optimal. However, by contrast to previous results in this literature, they depend upon only one regularization parameter. If λ is optimally chosen, then we obtain convergence rates that are $n^{-\frac{\gamma\wedge 2}{2(1+\gamma\wedge 2)}}$. For $\gamma = 2$ or 1, this respectively yields $n^{-1/3}$ and $n^{-1/4}$. It is unclear how to compare the above rates to existing results on optimal convergence rates (Hall and Horowitz, 2005; Chen and Christensen, 2018) because these authors make assumptions on conditional expectation operators such as $\mathrm{E}\{g(Z)|W\}$, while Assumption 3.3 concerns an unconditional expectation operator. Our assumption, however, assumes that the problem is mildly ill-posed, while some previous work also considers the case of a severely ill-posed inverse problem.

4 Estimation under monotonicity

In many applications, we may expect the function of interest g_0 to be monotonic. For instance, if g_0 is the demand for a good as a function of price, we expect it to be decreasing. If g_0 is the Engel curve, that relates the proportion of consumer expenditure on a good to total expenditure, we typically expect this function to be increasing for a "normal" good and decreasing for an "inferior" good. Accounting for monotonicity in estimation is expected to improve accuracy in small and moderate samples (Chetverikov and Wilhelm, 2017).

First, note that since our spline smoothing estimator is linear in parameters, the derivative estimator(10) is linear as well. Let us express it in matrix form. Since $\boldsymbol{g} = \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{E}\boldsymbol{\delta}$, we can write $\boldsymbol{g'} = \boldsymbol{O}\boldsymbol{a} + \boldsymbol{D}\boldsymbol{\delta}$, where $\boldsymbol{g'} = (g'(Z_1), \ldots g'(Z_n))^T$,

$$\boldsymbol{D} = \left[\frac{1}{4}\,\mathrm{sign}(Z_i - Z_j)|Z_i - Z_j|^2\right]_{i,j=1,\ldots n}, \quad \boldsymbol{O} = \begin{pmatrix} 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}.$$

From Proposition 2.1,

$$\begin{pmatrix} \widehat{\boldsymbol{\delta}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = \boldsymbol{S}\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} \widetilde{\boldsymbol{E}} & \boldsymbol{Z} \\ \boldsymbol{Z^T} & \boldsymbol{0} \end{bmatrix}^{-1} \quad \Rightarrow \quad \boldsymbol{g'} = (\boldsymbol{D}, \boldsymbol{O})\,\boldsymbol{S}\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}. \tag{13}$$

To obtain a monotone estimator, we rely on a method proposed by Hall and Huang (2001), that is based on the same linear estimator but reweights the observations Y_i to impose monotonicity at observations points. It adjusts the unconstrained estimator by tilting the empirical distribution to make the least possible change, in the sense of a distance measure, subject to imposing the constraint of monotonicity at observation points. Specifically, if g_0 is assumed to be monotonically increasing, we consider the constrained optimization program

$$\min_{p_1,\ldots,p_n} n - \sum_{i=1}^{n}(np_i)^{1/2} \tag{14}$$

$$\text{subject to} \quad \sum_{i=1}^{n} p_i = 1\,, \quad p_i \geq 0 \text{ for all } i = 1,\ldots,n \quad , \quad (\boldsymbol{D},\boldsymbol{O})\,\boldsymbol{S}\begin{pmatrix} \boldsymbol{p}\circ\boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix} \geq 0\,,$$

where $\boldsymbol{p}\circ\boldsymbol{Y} = (p_1Y_1,\ldots,p_nY_n)^T$ is the Hadamard product between vectors. If g_0 was assumed to be monotonically decreasing, we would modify the last inequalities. Hall and Huang (2001) considered more general optimization problems based on a family of Cressie-Read divergences, but we focus on the above program for convenience. It is strictly convex, so it admits a unique solution $\boldsymbol{p}^*$, and it is computationally fast to solve. The final estimator is $\widehat{g}^*$, defined as a smoothing splines estimator where $\widehat{\boldsymbol{a}}^*$ and $\widehat{\boldsymbol{\delta}}^*$ are as in (13) where $\boldsymbol{p}^* \circ \boldsymbol{Y}$ replaces $\boldsymbol{Y}$.

We now state the asymptotic properties of our constrained smoothing splines estimator.

Assumption 4.1. *There exists $\eta > 0$ such that $g_0'(z) \geq \eta$ for all $z \in [0,1]$.*

Theorem 4.1. *Under Assumptions 3.1, 3.2, and 4.1 if $\lambda \to 0$ and $n\lambda \to \infty$, then*

$$\sup_{z\in[0,1]} |\widehat{g}^*(z) - g_0(z)| = o_P(1) \quad \textit{and} \quad \sup_{z\in[0,1]} |\widehat{g}^{*'}(z) - g_0'(z)| = o_P(1)\,.$$

If moreover Assumption 3.3 holds, then

$$\sup_{z\in[0,1]} |\widehat{g}^*(z) - g_0(z)| = O_P\left(\frac{1}{\sqrt{n\lambda}} + \lambda^{\frac{\gamma\wedge 2}{2}}\right) \quad \textit{and} \quad \sup_{z\in[0,1]} |\widehat{g}^{*'}(z) - g_0'(z)| = O_P\left(\frac{1}{\sqrt{n\lambda}} + \lambda^{\frac{\gamma\wedge 2}{2}}\right)\,.$$

The above result directly follows from Theorem 3.1. Indeed, as $\widehat{g}'$ is uniformly consistent, the constraint in the optimization problem (14) becomes asymptotically irrelevant under Assumption 4.1. Accordingly, $\widehat{g}^* = \widehat{g}$ with probability approaching one, and our results readily follow. While the monotonicity constraints become asymptotically irrelevant, they can matter in finite samples as shown by Chetverikov and Wilhelm (2017) and illustrated by our numerical results below.

5 Empirical results

5.1 Simulations

We used a DGP in line with Equation (1), where

$$\varepsilon = \frac{aV + \eta}{\sqrt{1+a^2}}\,, a = \sqrt{\frac{\rho_{\varepsilon V}^2}{1-\rho_{\varepsilon V}^2}}\,,$$
$$Z = \frac{\beta W + V}{\sqrt{1+\beta^2}}\,, \beta = \sqrt{\frac{\rho_{WZ}^2}{1-\rho_{WZ}^2}}\,,$$

and (W, V, η) are independent standard Gaussian. This yields standard Gaussian marginal distributions for ε and Z whatever the values of the parameters. We chose this setup so as to be able to simply tune the level of endogeneity of Z, which is here controlled by the correlation parameter $\rho_{\varepsilon V}$, and the strength of the instrument W, controlled by the correlation parameter ρ_{WZ}.

We implemented our smoothing splines estimator with $\omega(\cdot)$ equal to the density of a Laplace distribution with mean zero and variance 1. The choice of the penalty parameter

λ was based on 2-fold cross-validation. Namely, we split the data at random into two equally sized folds, we computed $\widehat{g}_{k,\lambda}$ for each fold $k = 1, 2$, and we created the cross-validated vector $\tilde{\boldsymbol{g}}_\lambda$, with typical element $\widehat{g}_{k,\lambda}(Z_i)$, where k is the fold that does not include Z_i. We then chose the value of λ that minimizes $M_n(\tilde{\boldsymbol{g}}_\lambda)$ within the grid $\{p/(1-p), p = 10^{-5} + k * (0.7 - 10^{-5})/399, k = 0, \ldots, 399\}$.

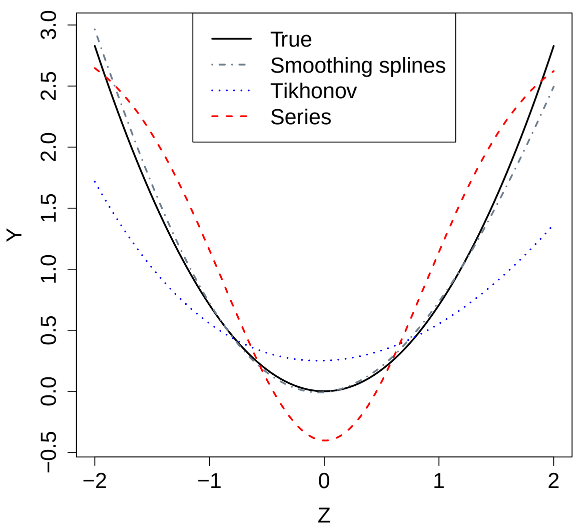
We compared our estimator to two existing methods, for which a data-driven procedure has been proposed for the choices of smoothing and regularization parameters. We considered first the kernel-based Tikhonov estimator of Darolles et al. (2011), hereafter referred as Tikhonov. We used Gaussian kernels of order 2, while the bandwidths were set using Silverman's rule of thumb, i.e., equal to $n^{-1/5}$ times the empirical standard deviation of the variable on which smoothing is performed (either Z or W). To select the regularization parameter, we used the pseudo-cross-validation procedure of Centorrino et al. (2017) searching for the minimum of the criterion on a grid between 0.0001 and 0.999. We also considered a series estimator based on a basis of Legendre polynomials. The number of polynomials in first and second stage estimation are chosen to be equal, and this number is selected using the adaptive procedure of Horowitz (2014). As the method is designed for variables belonging to $[0, 1]$, we transformed observations of Z and W by their respective empirical cumulative distribution functions (cdf). This implies in particular that even if the relation between Z and W is linear, the first-stage equation is not linear anymore in the transformed variables.

We first considered two functional forms for g_0, each normalized to have unit variance: a quadratic function $g_{0,1}(z) = z^2/\sqrt{2}$, and a non-polynomial function $g_{0,2}(z) = \sqrt{3\sqrt{3}}\, z \exp(-z^2/2)$. We ran 2000 Monte Carlo simulations with sample sizes $n = 200$ and 400. We considered three couples of values for $(\rho_{\varepsilon V}, \rho_{WZ})$: (i) $(0.5, 0.9)$, a setting with low endogeneity and a strong instrument, (ii) $(0.8, 0.9)$, corresponding to high endogeneity and a strong instrument, (iii) $(0.8, 0.7)$, a more complex setting with high endogeneity level but a weaker instrument. Table 1 reports the squared bias, variance, and mean squared error (MSE) of each estimator, averaged on a grid of 100 equidistant points on $[-2, 2]$. We note that the Tikhonov estimator is severely biased in all cases, while our estimator is close to be unbiased. The series estimator mostly lies in between, but with large differences depending on the setup. For instance, Figures 1a and 1b graph the pointwise average of each estimator for $n = 200$ and $(\rho_{\varepsilon V}, \rho_{WZ}) = (0.8, 0.7)$. Here, the series estimator is much steeper than the true quadratic curve $g_{0,1}$, while it fails to fit the sign changes in the first derivative for $g_{0,2}$. In both cases, the Tikhonov estimator is much less rough than the true curves, while the smoothing splines estimator is almost unbiased.

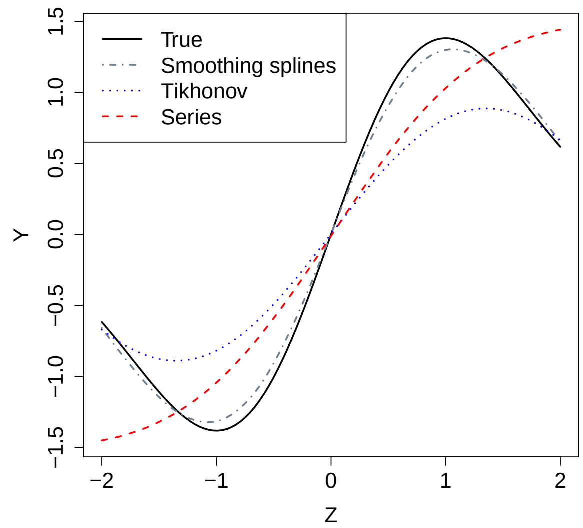
		$g_{0,1}$			$g_{0,2}$			$g_{0,3}$			
n		Sm.	Tikh.	Ser.	Sm.	Tikh.	Ser.	Cons.	Sm.	Tikh.	Ser.
		$\rho_{ZW} = 0.9\,,\ \rho_{\varepsilon V} = 0.5$									
200	Bias^2	0.000	0.148	0.092	0.001	0.065	0.005	0.000	0.000	0.088	0.104
200	Var	0.069	0.030	0.060	0.074	0.028	0.073	0.041	0.076	0.022	0.116
200	MSE	0.069	0.177	0.152	0.075	0.092	0.078	0.044	0.076	0.111	0.219
400	Bias^2	0.000	0.098	0.094	0.001	0.044	0.005	0.000	0.000	0.054	0.140
400	Var	0.052	0.020	0.031	0.053	0.018	0.025	0.026	0.056	0.015	0.044
400	MSE	0.052	0.118	0.125	0.054	0.063	0.030	0.027	0.056	0.069	0.184
		$\rho_{ZW} = 0.9\,,\ \rho_{\varepsilon V} = 0.8$									
200	Bias^2	0.001	0.148	0.092	0.001	0.064	0.005	0.000	0.000	0.093	0.107
200	Var	0.066	0.028	0.060	0.072	0.026	0.074	0.039	0.072	0.019	0.107
200	MSE	0.067	0.176	0.152	0.072	0.090	0.079	0.042	0.073	0.112	0.214
400	Bias^2	0.000	0.098	0.094	0.000	0.043	0.005	0.000	0.000	0.055	0.141
400	Var	0.049	0.019	0.030	0.051	0.018	0.026	0.025	0.054	0.013	0.042
400	MSE	0.050	0.117	0.125	0.052	0.061	0.030	0.026	0.054	0.068	0.183
		$\rho_{ZW} = 0.7\,,\ \rho_{\varepsilon V} = 0.8$									
200	Bias^2	0.009	0.307	0.133	0.004	0.158	0.158	0.012	0.012	0.184	0.189
200	Var	0.091	0.032	0.130	0.120	0.028	0.053	0.056	0.101	0.019	0.050
200	MSE	0.099	0.338	0.262	0.124	0.186	0.211	0.080	0.113	0.203	0.238
400	Bias^2	0.003	0.218	0.139	0.004	0.110	0.157	0.002	0.002	0.123	0.190
400	Var	0.069	0.024	0.053	0.087	0.021	0.024	0.041	0.086	0.014	0.021
400	MSE	0.073	0.242	0.192	0.090	0.131	0.180	0.051	0.089	0.138	0.212

Table 1: Average over a grid of 100 equidistant points on $[-2, 2]$ and 2000 Monte Carlo replications of the squared bias (Bias^2), the variance (Var), and the Mean Squared Error (MSE) for the constrained smoothing splines estimator (Cons.), the smoothing splines estimator (Sm.), the Tikhonov estimator (Tikh.), and the series estimator (Ser.).

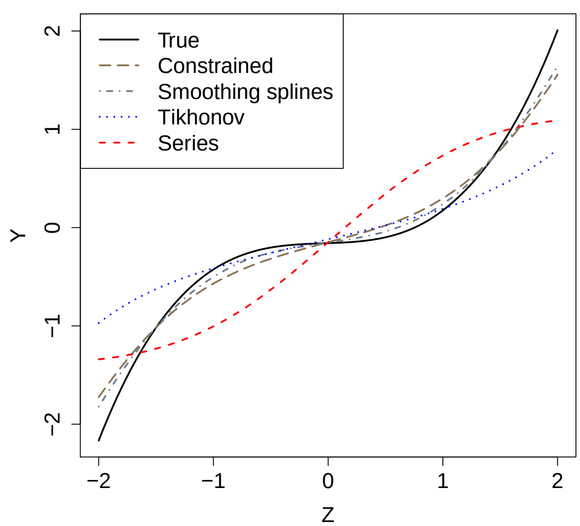
In terms of variance, Tikhonov does better than smoothing splines, that itself does better than series, in all but one instance. Smoothing splines performs best in terms of MSE in almost all cases. Exceptions are cases corresponding to the second function with $n = 400$ and strong instruments, where the series estimator is close to be unbiased. Overall, the severity of endogeneity does not affect much the estimators' performances. A decrease in the strength of the instrument has important detrimental effects for all estimators, but our smoothing splines estimator is affected to a much lesser degree than its competitors.



(a) $g_{0,1}$



(b) $g_{0,2}$



(c) $g_{0,3}$

Figure 1: True regression function (solid black line) and averages over all replications of the constrained (long dashed brown line), smoothing splines (dashed-dotted grey line), Tikhonov (dotted blue line) and series (dashed red line) estimators. The sample size is $n = 200$, $\rho_{\varepsilon V} = 0.8$, and $\rho_{ZW} = 0.7$.

To evaluate the gains of imposing monotonicity, we then considered a third function $g_{0,3}(z) = (\sqrt{(}10/3) \log(|z-1|+1) \operatorname{sign}(z-1) - 0.6z + 2z^3)/8$. We implemented the smoothing splines estimator as above, in particular the regularization parameter λ is chosen before the monotonizing step, and we used the the CVXR solver (Fu et al., 2020) to solve (14). As can be seen from Figure 1c and our detailed results in Table 1, imposing monotonicity does not affect much the average smoothing splines estimate, nor its average squared bias, but yields a substantial decrease in variance. Depending on the particular setup, it can be more than halved.

5.2 Application

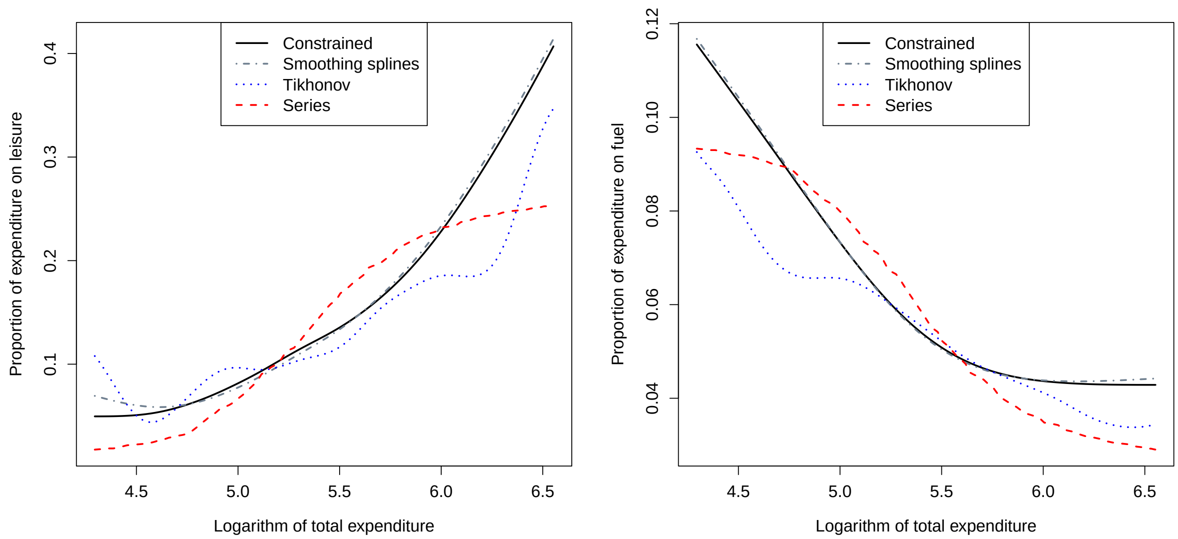
We applied our smoothing splines estimator to the estimation of Engel curves, which relate the proportion of spending on a given good to total expenditures. We used the "Engel95" data from the `R` package `np`, which is a random sample from the 1995 British Family Expenditure Survey and contains data for 1655 households of married couples for which the head-of-household is employed and between 25 and 55 years old. We focused on the subsample of 628 households with no kids. We report results for two Engel curves, pertaining to the expenditure shares on leisure and fuel. Economic theory suggests that the Engel curve for leisure is increasing and the one for fuel is decreasing. Following Blundell et al. (2007), we instrumented the endogenous variable *logarithm of total household's expenditure* by the *logarithm of total earnings before tax*. We consider the four estimators used in our simulations, and implementation details remain the same.

The resulting estimated nonparametric functions are reported in Figure 2. The Tikhonov estimate exhibits a non-monotonic and quite irregular behavior, while the series estimate is mainly monotonic and very regular. Since our smoothing splines estimates are monotonic, but at the boundaries of the data, our constrained and unconstrained estimates are very close. Both are in line with the findings of Blundell et al. (2007).

6 Conclusion

This paper generalizes regression smoothing splines to the context where there is endogeneity. We have considered the special case of cubic splines, but our approach easily generalizes to other spline orders depending upon the roughness measure chosen in the penalization. Though we have focused on a simple univariate nonparametric model, our work opens the path for considering more general models. A first example is the partly linear model (Heckman, 1986)

$$Y = X^T\beta + g_0(Z) + \varepsilon \qquad E(\varepsilon|X,W) = 0\,.$$



(a) Leisure (b) Fuel

Figure 2: Engel curves estimated with our constrained smoothing splines estimator (solid black line), our smoothing splines (dashed-dotted grey line), the Tikhonov (dotted blue line) and the Series estimator (dashed red line).

Our method of proof based on profiling should easily extend to this model. A second example is the multivariate model

$$Y = g_0(X,Z) + \varepsilon \qquad E(\varepsilon|X,W) = 0\,.$$

Consider for instance univariate variables X and Z, and the roughness penalty

$$J(g) = \int \left\{ \frac{\partial^2 g}{\partial x^2} + 2\frac{\partial^2 g}{\partial x \partial z} + \frac{\partial^2 g}{\partial z^2} \right\} dx\, dz\,.$$

A function $g(\cdot)$ is a natural thin plate spline with knots $(t_1,\ldots t_n)$, $t_i = (X_i, Z_i)$, if

$$g(t) = a_0 + t'b + \sum_{i=1}^{n} \delta_i \eta\left(\|t - t_i\|\right), \qquad \eta(r) = \frac{1}{16\pi} r^2 \log r^2\,,$$

where $\sum_{i=1}^n \delta_i = \sum_{i=1}^n X_i\delta_i = \sum_{i=1}^n Z_i\delta_i = 0$. It is known that the natural thin plate spline uniquely minimizes $J(g)$ among functions such that $g(t_i) = g_i, i = 1,\ldots n$ (Green and Silverman, 1993).

Since multivariate nonparametric estimators are subject to the curse of dimensionality, it may be wiser to consider an additive model such as

$$Y = \sum_{j=1}^{J} g_{0,j}(Z_j) + \varepsilon \qquad E(\varepsilon|W) = 0\,,$$

where W contain instruments, which may include some of the variables Z_j if these are deemed exogenous. All these extensions are left for further research.

References

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Bennett, A., Kallus, N., and Schnabel, T. (2019). Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, pages 3564–3574.

Bierens, H. J. (1982). Consistent model specification tests. *J Econometrics*, 20(1):105–134.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669.

Busby, R. C., Schochetman, I., and Smith, H. A. (1972). Integral operators and the compactness of induced representations. *Transactions of the American Mathematical Society*, 164:461–477.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6, chapter 77, pages 5633–5751. Elsevier.

Centorrino, S., Feve, F., and Florens, J.-P. (2017). Additive nonparametric instrumental regressions: A guide to implementation. *Journal of Econometric Methods*, 6(1).

Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84.

Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321.

Chetverikov, D. and Wilhelm, D. (2017). Nonparametric instrumental variable estimation under monotonicity. *Econometrica*, 85(4):1303–1320.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. (2020). Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262.

Dreier, I. and Kotz, S. (2002). A note on the characteristic function of the t-distribution. *Statistics & Probability Letters*, 57(3):221 – 224.

Dunker, F., Florens, J.-P., Hohage, T., Johannes, J., and Mammen, E. (2014). Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal of Econometrics*, 178:444–455.

D'Haultfoeuille, X. (2011). On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471.

Florens, J.-P., Johannes, J., and Van Bellegem, S. (2011). Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*, 27(3):472–496.

Freyberger, J. (2017). On completeness and consistency in nonparametric instrumental variable models. *Econometrica*, 85(5):1629–1644.

Fu, A., Narasimhan, B., and Boyd, S. (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34.

Gagliardini, P. and Scaillet, O. (2012). Tikhonov regularization for nonparametric instrumental variable estimators. *Journal of Econometrics*, 167(1):61–75.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Crc Press.

Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929.

Hall, P. and Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, 29(3):624–647.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(2):244–248.

Horowitz, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394.

Horowitz, J. L. (2014). Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics*, 180(2):158–173.

Johannes, J., Van Bellegem, S., and Vanhems, A. (2011). Convergence rates for ill-posed inverse problems with an unknown operator. *Econometric Theory*, 27(3):522–545.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, volume 2 of *Wiley series in probability and mathematical statistics: Applied probability and statistics*. Wiley & Sons.

Kress, R. (1999). *Linear Integral Equations*. Springer.

Kreyszig, E. (1978). *Introductory Functional Analysis with Applications*. John Wiley & Sons.

Liao, L., Chen, Y.-L., Yang, Z., Dai, B., Kolar, M., and Wang, Z. (2020). Provably efficient neural estimation of structural equation models: An adversarial approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 8947–8958.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. (2020). Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge Mathematical Library. Cambridge University Press, 3 edition.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Zhang, R., Imaizumi, M., Schölkopf, B., and Muandet, K. (2023). Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1):20220073.

Appendix

A Proof of Proposition 2.1

We begin by studying $\boldsymbol{E}$ and $\boldsymbol{\Omega}$. First, for any natural cubic spline (Green and Silverman, 1993, Section 7.3),

$$\int g''(z)^2\, dz = \boldsymbol{\delta^T E \delta} \geq 0\,,$$

so $\boldsymbol{E}$ is positive semi-definite. Second, let $\boldsymbol{b} = (b_1, \ldots, b_n)^T \in \mathbb{R}^n$, then

$$\boldsymbol{b^T \Omega b} = \int \left| \frac{1}{n} \sum_{i=1}^{n} b_i \exp(\mathbf{i} W_i^T t) \right|^2 \mu(dt) \geq 0\,.$$

Hence $\boldsymbol{b^T \Omega b} = 0$ iff $\frac{1}{n}\sum_{i=1}^n b_i \exp(\mathbf{i} W_i^T t) = 0$ for all $t \in \mathbb{R}^q$. Define the random vector $(\widetilde{b}, \widetilde{\boldsymbol{W}})$ that equals (b_i, W_i) with probability $1/n$, and $\widetilde{\mathrm{E}}$ the corresponding expectation. Then, $\widetilde{\mathrm{E}}\{\widetilde{b} \exp(\mathbf{i}\widetilde{\boldsymbol{W}}^T t)\} = 0$ for all $t \in \mathbb{R}^q$. From Bierens (1982), this implies that $\widetilde{\mathrm{E}}\{\widetilde{b} \,|\, \widetilde{\boldsymbol{W}} = \boldsymbol{W}_i\} = 0$. Since $\widetilde{\mathrm{E}}\{\widetilde{b} \,|\, \widetilde{\boldsymbol{W}} = \boldsymbol{W}_i\} = b_i$ if all W_is are different, $b_i = 0$ for all $i = 1, \ldots, n$. Hence, $\boldsymbol{\Omega}$ is positive definite.

Let us then write (8) as

$$\begin{pmatrix} \boldsymbol{\delta} \\ \boldsymbol{a} \end{pmatrix}^\top \begin{bmatrix} \boldsymbol{E^T \Omega E} + \lambda \boldsymbol{E} & \boldsymbol{E^T \Omega Z} \\ \boldsymbol{Z^T \Omega E} & \boldsymbol{Z^T \Omega Z} \end{bmatrix} \begin{pmatrix} \boldsymbol{\delta} \\ \boldsymbol{a} \end{pmatrix} - 2 \begin{pmatrix} \boldsymbol{\delta} \\ \boldsymbol{a} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{E^T \Omega Y} \\ \boldsymbol{Z^T \Omega Y} \end{pmatrix} + \boldsymbol{Y^T \Omega Y}\,.$$

This is a convex problem in $(\boldsymbol{\delta^T}, \boldsymbol{a^T})^T$. Differentiating with respect to $\left(\boldsymbol{\delta}^T, \boldsymbol{a}^T\right)^T$ yields

$$\begin{bmatrix} \boldsymbol{E^T \Omega} & \boldsymbol{0} \\ \boldsymbol{Z^T \Omega} & -\lambda \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{E}} & \boldsymbol{Z} \\ \boldsymbol{Z^T} & \boldsymbol{0} \end{bmatrix} \begin{pmatrix} \widehat{\boldsymbol{\delta}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = \begin{bmatrix} \boldsymbol{E^T \Omega} & \boldsymbol{0} \\ \boldsymbol{Z^T \Omega} & -\lambda \boldsymbol{I} \end{bmatrix} \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}.$$

This yields (9), and the solution satisfies the constraint $\boldsymbol{Z^T} \widehat{\boldsymbol{\delta}} = \boldsymbol{0}$. The matrix in (9) is full rank if $\boldsymbol{Z}$ is full rank and $\lambda > 0$. Indeed, assume

$$\begin{bmatrix} \boldsymbol{E} + \lambda \boldsymbol{\Omega}^{-1} & \boldsymbol{Z} \\ \boldsymbol{Z^T} & \boldsymbol{0} \end{bmatrix} \begin{pmatrix} \boldsymbol{\delta} \\ \boldsymbol{a} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix},$$

this implies $\boldsymbol{Z^T \delta} = \boldsymbol{0}$ and $0 = \boldsymbol{\delta^T}\left[(\boldsymbol{E} + \lambda \boldsymbol{\Omega}^{-1})\,\boldsymbol{\delta} + \boldsymbol{Za}\right] = \boldsymbol{\delta^T}\left(\boldsymbol{E} + \lambda \boldsymbol{\Omega}^{-1}\right)\boldsymbol{\delta}$. Since $\boldsymbol{E} + \lambda \boldsymbol{\Omega}^{-1}$ is positive definite, this implies that $\boldsymbol{\delta} = \boldsymbol{0}$, and in turn that $\boldsymbol{a} = \boldsymbol{0}$ as $\boldsymbol{Z}$ is full rank.

To obtain the values at the knots $\widehat{\boldsymbol{g}}$, note that the inverse of the matrix in (9) is

$$\begin{bmatrix} \widetilde{\boldsymbol{E}}^{-1}(\boldsymbol{I} - \boldsymbol{P}) & \widetilde{\boldsymbol{E}}^{-1}\boldsymbol{Z}\left(\boldsymbol{Z}^T \widetilde{\boldsymbol{E}}^{-1} \boldsymbol{Z}\right)^{-1} \\ \left(\boldsymbol{Z}^T \widetilde{\boldsymbol{E}}^{-1} \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T \widetilde{\boldsymbol{E}}^{-1} & -\left(\boldsymbol{Z}^T \widetilde{\boldsymbol{E}}^{-1} \boldsymbol{Z}\right)^{-1} \end{bmatrix}, \tag{A.15}$$

where $\boldsymbol{P} = \boldsymbol{Z}\left(\boldsymbol{Z}^T\widetilde{\boldsymbol{E}}^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\widetilde{\boldsymbol{E}}^{-1}$ is the oblique projection on the span of $\boldsymbol{Z}$ along the direction spanned by vectors $\boldsymbol{h}$ such that $\boldsymbol{Z}^T\widetilde{\boldsymbol{E}}^{-1}\boldsymbol{h} = 0$. Hence,

$$\begin{pmatrix} \boldsymbol{E}\widehat{\boldsymbol{\delta}} \\ \boldsymbol{Z}\widehat{\boldsymbol{a}} \end{pmatrix} = \begin{bmatrix} \boldsymbol{E}\widetilde{\boldsymbol{E}}^{-1}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} \\ \boldsymbol{P}\boldsymbol{Y} \end{bmatrix}.$$

Use $\widehat{\boldsymbol{g}} = \boldsymbol{Z}\widehat{\boldsymbol{a}} + \boldsymbol{E}\widehat{\boldsymbol{\delta}}$ to obtain the desired result.

B Proof of Theorem 3.1

We start by introducing some useful notations and results. Let $\mathcal{X}$ and $\mathcal{Y}$ be Hilbert spaces with corresponding inner products $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$, and consider a linear operator $\mathrm{D} : \mathcal{X} \mapsto \mathcal{Y}$. The norm of D is $\|\,\mathrm{D}\,\|_{op} = \sup_{f\in\mathcal{X},\|f\|_{\mathcal{X}}=1}\|\,\mathrm{D}\,f\|_{\mathcal{Y}}$. When $\|\,\mathrm{D}\,\|_{op} < \infty$, D is said to be bounded (or continuous), see Kress (1999, Chapter 2). Let D^* be the adjoint of D, defined as $\mathrm{D}^* : \mathcal{Y} \mapsto \mathcal{X}$ such that $\langle \mathrm{D}\,f, \psi\rangle_{\mathcal{Y}} = \langle f, \mathrm{D}^*\,\psi,\rangle_{\mathcal{X}}$ for any $(f,\psi) \in \mathcal{X}\times\mathcal{Y}$. When D is bounded, D^* always exists and $\|\,\mathrm{D}\,\|_{op} = \|\,\mathrm{D}^*\,\|_{op}$, see Kress (1999, Theorem 4.9). In what follows, we will repeatedly use the following properties: (i) $\|\,\mathrm{D}\,f\|_{\mathcal{Y}} \leq \|\,\mathrm{D}\,\|_{op}\|f\|_{\mathcal{X}}$ for any $f \in \mathcal{X}$, and (ii) if C is another linear operator, then $\|\,\mathrm{C}\,\mathrm{D}\,\|_{op} \leq \|\,\mathrm{C}\,\|_{op}\|\,\mathrm{D}\,\|_{op}$, whenever the composition $\mathrm{C}\,\mathrm{D}$ is well defined.

B.1 Main proof

We divide the proof into several steps. In Step 1, we analyze the minimization problem at the population level. In Step 2, we analyze the problem at the sample level. In Step 3, we bound the norm of $\widehat{h}-h_0$. In Step 4 and 5, we combine the results to first establish uniform consistency of $\widehat{g}$ and its first derivative, second to obtain uniform rates of convergence.

Step 1. From Assumption 3.2 and Bierens (1982),

$$g = 0 \Leftrightarrow \mathrm{E}\left\{g(Z)|W\right\} = 0 \Leftrightarrow \mathrm{E}\left\{g(Z)\exp(\mathbf{i}W^\top t)\right\} = 0 \;\forall t \in \mathbb{R}^p\,.$$

Hence, the null space of the linear mapping $g \mapsto \mathrm{E}\{g(Z)\exp(\mathbf{i}W^T\cdot)\}$ only contains the zero element, and such a mapping is injective (one-to-one). This implies that $\mathrm{A}\,h = \mathrm{E}\{h(Z)\exp(\mathbf{i}W^T\cdot)\}$ and $\mathrm{B}\,\beta = \mathrm{E}\{(1,Z)\beta\exp(\mathbf{i}W^T\cdot)\}$ are also injective.

Each $g \in \mathcal{G}$ can be uniquely written as $g(z) = (1,z)\beta + h(z)$, where $\beta = (g(0), g'(0))$, $h(z) = g(z) - g(0) - g'(0)z$, $h(\cdot) \in \mathcal{H}$. Hence, the intersection of the ranges of the operators A and B is the null function, since $\mathrm{A}\,h = \mathrm{B}\,\beta$ iff $(1,z)\beta - h(z) = 0$.

Consider the problem

$$\min_{\beta,h}\|r - \mathrm{B}\,\beta - \mathrm{A}\,h\|^2_\mu\,, \qquad r = \mathrm{E}\{Y\exp(\mathbf{i}W^T\cdot)\}\,, \tag{A.16}$$

where $\|\cdot\|_\mu$ is the L^2_μ norm. If $g_0(z) = (1,z)\beta_0 + h_0(z)$, then (β_0, h_0) is the unique solution. We will obtain an explicit expression of (β_0, h_0) solving (A.16). Let P be the orthogonal projection operator of functions in L^2_μ onto $\mathcal{R}(\mathrm{B})$ the range of B. Since B is defined on $\mathbb{R}^2$, its range $\mathcal{R}(\mathrm{B})$ is a linear finite dimensional space. As linear finite dimensional spaces are complete, see Kreyszig (1978, Theorem 2.4-2), $\mathcal{R}(\mathrm{B})$ is also linear and complete. By Kress (1999, Theorem 1.26), projection operators onto linear and complete spaces are well defined, and so is P. Let us show that P writes as $\mathrm{B}(\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^*$, where B^* is the adjoint of B. As previously noted, B is injective and its null space is $\mathcal{N}(\mathrm{B}) = \{0\}$. Then $\mathcal{N}(\mathrm{B}^*\mathrm{B}) = \mathcal{N}(\mathrm{B}) = \{0\}$, $\mathrm{B}^*\mathrm{B}$ is injective, and $(\mathrm{B}^*\mathrm{B})^{-1}$ exists. As linear operators mapping $\mathbb{R}^2$ into $\mathbb{R}^2$ are uniquely characterized by second order matrices, see Kreyszig (1978, Section 2.9), $\mathrm{B}^*\mathrm{B}$ is a second order matrix, as well as its inverse. Hence, the operator $\mathrm{B}(\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^* : L^2_\mu \mapsto L^2_\mu$ is well defined. For any $f \in L^2_\mu$ and $\beta \in \mathbb{R}^2$,

$$\left\langle f - \mathrm{B}(\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^* f, \mathrm{B}\,\beta\right\rangle_\mu = \langle \mathrm{B}^* f - \mathrm{B}^* f, \beta\rangle = 0\,.$$

Hence, $f - \mathrm{B}(\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^* f \perp \mathcal{R}(\mathrm{B})$, and $\mathrm{B}(\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^* f$ indeed represents the projection of f onto $\mathcal{R}(\mathrm{B})$, see Kress (1999). Therefore, $\mathrm{P} = \mathrm{B}(\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^*$.

Let $\mathrm{M} = I - \mathrm{P}$ be the orthogonal projection onto the orthogonal complement of $\mathcal{R}(\mathrm{B})$. Then

$$r = \mathrm{B}\,\beta_0 + \mathrm{A}\,h_0 \Rightarrow \mathrm{M}\,r = \mathrm{M}\,\mathrm{A}\,h_0 = \mathrm{T}\,h_0\,.$$

The operator $\mathrm{T} = \mathrm{M}\,\mathrm{A}$ is injective since the intersection of the ranges of A and B is the null function and A is injective. This yields

$$h_0 = \mathrm{T}^{-1}\,\mathrm{M}\,r\,, \qquad \beta_0 = (\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^*(r - Ah_0)\,.$$

Consider now the penalized problem

$$\min_{(\beta,h)\in\mathbb{R}^2\times\mathcal{H}} \|r - \mathrm{A}h - \mathrm{B}\beta\|^2_\mu + \lambda\|h\|^2_{\mathcal{H}}\,. \tag{A.17}$$

Let us profile with respect to β. For any fixed h,

$$\min_{\beta\in\mathbb{R}^2} \|r - \mathrm{A}h - \mathrm{B}\beta\|^2_\mu = \|r - \mathrm{A}h - \mathrm{P}(r - Ah)\|^2_\mu = \|\,\mathrm{M}\,r - \mathrm{T}h\|^2_\mu\,.$$

We thus need to solve first

$$\min_{h\in\mathcal{H}} \|\,\mathrm{M}\,r - \mathrm{T}h\|^2_\mu + \lambda\|h\|^2_{\mathcal{H}}\,.$$

From Lemma B.2(i) below, T is compact, and thus bounded. A direct application of Kress (1999, Theorem 16.1) ensures that the unique solution h_λ to the above minimization problem satisfies $(\mathrm{T}^*\mathrm{T} + \lambda I)h_\lambda = \mathrm{T}^*\mathrm{M}r$. Now, for any h,

$$\lambda\|h\|^2_{\mathcal{H}} \le \lambda\|h\|^2_{\mathcal{H}} + \|\,\mathrm{T}\,h\|^2_\mu = \lambda\,\langle h, h\rangle_{\mathcal{H}} + \langle h, \mathrm{T}^*\,\mathrm{T}\,h\rangle_{\mathcal{H}} = \langle h, (\mathrm{T}^*\mathrm{T} + \lambda I)h\rangle\,.$$

Hence, $(\mathrm{T}^*\mathrm{T}+\lambda I)$ is strictly coercive and has a bounded inverse by the Lax-Milgram Theorem, see Kress (1999, Theorem 13.26). Therefore,

$$h_\lambda = (\mathrm{T}^*\mathrm{T}+\lambda I)^{-1}\mathrm{T}^*\mathrm{M}r \tag{A.18}$$

Step 2. We study the minimization problem at the sample level and we obtain sample counterparts of the population objects of Step 1. Recall that $\widehat{g}$ solves

$$\min_{g\in\mathcal{G}} \int \left| \frac{1}{n}\sum_{i=1}^n [Y_i - g((Z_i))]\exp(\mathbf{i}W_i^T t)\right|^2 \mu(dt) + \lambda \int_0^1 |g''(z)|^2 dz\,. \tag{A.19}$$

By Proposition 2.1, under Assumption 3.1, the solution $\widehat{g}$ is unique with probability 1, and since each $g\in\mathcal{G}$ writes uniquely as $g(z)=(1,z)\beta+h(z)$, there is a unique $(\widehat{\beta},\widehat{h})$ such that $\widehat{g}(z)=(1,z)\widehat{\beta}+\widehat{h}(z)$. Define

$$\widehat{\mathrm{A}}:\mathcal{H}\mapsto L^2_\mu\,,\qquad \widehat{\mathrm{A}}h=\frac{1}{n}\sum_{i=1}^n h(Z_i)\exp(\mathbf{i}W_i^T\cdot)\,, \tag{A.20}$$

$$\widehat{\mathrm{B}}:\mathbb{R}^2\mapsto L^2_\mu\,,\qquad \widehat{\mathrm{B}}\beta=\frac{1}{n}\sum_{i=1}^n (1,Z_i)\beta\exp(\mathbf{i}W_i^T\cdot)\,, \tag{A.21}$$

and $\widehat{r}=(1/n)\sum_{i=1}^n Y_i\exp(\mathbf{i}W_i^T\cdot)$. The optimization problem (A.19) is equivalent to

$$\min_{(\beta,h)\in\mathbb{R}^2\times\mathcal{H}} \|\widehat{r}-\widehat{\mathrm{A}}h-\widehat{\mathrm{B}}\beta\|^2_\mu+\lambda\|h\|^2_{\mathcal{H}}\,. \tag{A.22}$$

We will profile with respect to β, and to do so require dealing with the orthogonal projection onto the range of $\widehat{\mathrm{B}}$. Let us proceed as in Step 1. First,

$$\|\widehat{\mathrm{B}}\beta\|^2_\mu=0 \Leftrightarrow \int\left|\frac{1}{n}\sum_{i=1}^n(1,Z_i)\beta\exp(\mathbf{i}W_i^Tt)\right|^2\mu(dt)=0\Leftrightarrow \beta^T\boldsymbol{Z}^T\boldsymbol{\Omega}\boldsymbol{Z}\beta=0\,.$$

From Assumption 3.1(v), W has at least one continuous component, so that all W_i's are different with probability 1, and thus $\boldsymbol{\Omega}>0$ with probability 1 from the proof of Proposition 2.1. Hence, $\|\widehat{\mathrm{B}}\beta\|^2_\mu=0$ iff $\boldsymbol{Z}\beta=0$. As $\boldsymbol{Z}$ has full column rank with probability 1 from Assumption 3.1, $\|\widehat{\mathrm{B}}\beta\|^2_\mu=0$ iff $\beta=0$, and B is injective. Let $\widehat{\mathrm{P}}$ be the orthogonal projection onto the range of $\widehat{\mathrm{B}}$, which is well defined and can be expressed as $\widehat{\mathrm{P}}=\widehat{\mathrm{B}}(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}\widehat{\mathrm{B}}^*$. Then,

$$\min_{\beta\in\mathbb{R}^2}\|\widehat{r}-\widehat{\mathrm{A}}h-\widehat{\mathrm{B}}\beta\|^2_\mu=\|\widehat{r}-\widehat{A}h-\widehat{P}(\widehat{r}-\widehat{A}h)\|_\mu=\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h\|^2_\mu\,,$$

where $\widehat{\mathrm{M}}=I-\widehat{\mathrm{P}}$ and $\widehat{\mathrm{T}}=\widehat{\mathrm{M}}\widehat{\mathrm{A}}$. We thus need to solve

$$\min_{h\in\mathcal{H}}\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h\|^2_\mu+\lambda\|h\|^2_{\mathcal{H}}\,. \tag{A.23}$$

From Lemma B.2(v) below, $\widehat{\mathrm{T}}$ is compact, and thus bounded. Thus, using a similar reasoning as in Step 1, the unique solution is

$$\widehat{h} = (\widehat{\mathrm{T}}^*\widehat{\mathrm{T}} + \lambda I)^{-1}\widehat{\mathrm{T}}^*\widehat{\mathrm{M}}\widehat{r}\,, \tag{A.24}$$

which in turn yields

$$\widehat{\beta} = (\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}\widehat{\mathrm{B}}^*(\widehat{r} - \widehat{\mathrm{A}}\widehat{h})\,. \tag{A.25}$$

Step 3. We now prove that

$$\|\widehat{h} - h_0\|_{\mathcal{H}} = O_P\left(\frac{1}{\sqrt{n\lambda}} + \|h_\lambda - h_0\|_{\mathcal{H}}\right)\,. \tag{A.26}$$

We will rely on the following results from Florens et al. (2011, Lemma A.1).

Lemma B.1. *Consider two Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ and a linear compact operator $\mathrm{K} : \mathcal{X} \mapsto \mathcal{Y}$. Then there are universal constants c and c' such that*

(i) $||\lambda(\lambda\,\mathrm{I} + \mathrm{K}^*\,\mathrm{K})^{-1}||_{op} \leq c$,

(ii) $||(\lambda\,\mathrm{I} + \mathrm{K}^*\,\mathrm{K})^{-1}\,\mathrm{K}^*\,||_{op} \leq \frac{c'}{\sqrt{\lambda}}$.

We will also use the following results, whose proofs are postponed to the next section.

Lemma B.2. *Under Assumptions 3.2 and 3.1, the following holds*

(i) *The operators* A *and* T *are compact;*

(ii) $\|\widehat{\mathrm{B}} - \mathrm{B}\,\|_{op} = O_P(n^{-1/2})$;

(iii) $\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1} - (\mathrm{B}^*\,\mathrm{B})^{-1}\|_{op} = O_P(n^{-1/2})$ *and* $\|\widehat{\mathrm{M}} - \mathrm{M}\,\|_{op} = O_P(n^{-1/2})$;

(iv) $\|\widehat{\mathrm{A}} - \mathrm{A}\,\|_{op} = O_P(n^{-1/2})$;

(v) *The operator* $\widehat{\mathrm{T}}$ *is compact;*

(vi) $\|\widehat{\mathrm{T}} - \mathrm{T}\,\|_{op} = O_P(n^{-1/2})$;

(vii) $\|\widehat{r} - r\|_{\mu} = O_P(n^{-1/2})$;

(viii) $\|\widehat{\mathrm{M}}\widehat{r} - \widehat{\mathrm{T}}h_0\|_{\mu} = O_P(n^{-1/2})$.

We consider the decomposition $\widehat{h} - h_0 = S_1 + S_2 + S_3 + S_4 + h_\lambda - h_0$, where

$$S_1 = (\mathrm{T}^*\,\mathrm{T} + \lambda I)^{-1}\,\mathrm{T}^*(\widehat{\mathrm{M}}\widehat{r} - \widehat{\mathrm{T}}h_0)\,, \quad S_2 = (\mathrm{T}^*\,\mathrm{T} + \lambda I)^{-1}(\widehat{\mathrm{T}}^* - \mathrm{T}^*)(\widehat{\mathrm{M}}\widehat{r} - \widehat{\mathrm{T}}h_0)\,,$$
$$S_3 = \left[(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}} + \lambda I)^{-1} - (\mathrm{T}^*\,\mathrm{T} + \lambda I)^{-1}\right]\widehat{\mathrm{T}}^*(\widehat{\mathrm{M}}\widehat{r} - \widehat{\mathrm{T}}h_0)\,, \quad S_4 = (\widehat{\mathrm{T}}^*\widehat{\mathrm{T}} + \lambda I)^{-1}\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}h_0 - h_\lambda\,.$$

We have

$$\|S_1\|_{\mathcal{H}} \leq \|(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\,\mathrm{T}^*\,\|_{op}\,\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h_0\|_\mu = O_P\left(\frac{1}{\sqrt{n\lambda}}\right).$$

Indeed, T is a compact operator from Lemma B.2(i), $\|(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\,\mathrm{T}^*\,\|_{op} \leq c'/\sqrt{\lambda}$ from Lemma B.1(ii), and $\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h_0\|_\mu = O_P(1/\sqrt{n})$ from Lemma B.2(viii). Next,

$$\|S_2\|_{\mathcal{H}} \leq \|(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\|_{op}\,\|\widehat{\mathrm{T}}^* - \mathrm{T}^*\,\|_{op}\,\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h_0\|_\mu = O_P\left(\frac{1}{n\lambda}\right),$$

as $\|(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\|_{op} \leq c/\lambda$ from Lemma B.1(i), $\|\widehat{\mathrm{T}}-\mathrm{T}\,\|_{op} = O_P(1/\sqrt{n})$ and $\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h_0\|_\mu = O_P(1/\sqrt{n})$ from Lemma B.2(vi) and (viii). Next,

$$\begin{aligned}
\|S_3\|_{\mathcal{H}} \leq& \|(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\widehat{\mathrm{T}}^* - (\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\widehat{\mathrm{T}}^*\|_{op}\,\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h_0\|_\mu \\
\leq& \|(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\widehat{\mathrm{T}}^* - (\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\,\mathrm{T}^* - (\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}(\widehat{\mathrm{T}}^* - \mathrm{T}^*)\|_{op}\,\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h_0\|_\mu \\
\leq& \|(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\widehat{\mathrm{T}}^* - (\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\,\mathrm{T}^*\,\|_{op}\,\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h_0\|_\mu \\
&+ \|(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\|_{op}\,\|\widehat{\mathrm{T}}^* - \mathrm{T}^*\,\|_{op}\,\|\widehat{\mathrm{M}}\widehat{r}-\widehat{\mathrm{T}}h_0\|_\mu = O_P\left(\frac{1}{\sqrt{n\lambda}}\right).
\end{aligned}$$

Indeed, $\widehat{\mathrm{T}}$ and T are compact operators from Lemma B.2(i) and (v), so $\|(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\widehat{\mathrm{T}}^* - (\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\,\mathrm{T}^*\,\|_{op} \leq 2c'/\sqrt{\lambda}$ and $\|(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\|_{op} \leq c/\lambda$ by Lemma B.1. Moreover, $\|\widehat{\mathrm{T}}^* - \mathrm{T}^*\,\|_{op} = O_P(1/\sqrt{n})$ from Lemma B.2(vi). Finally,

$$\begin{aligned}
\|S_4\|_{\mathcal{H}} =& \|(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}h_0 - (\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\,\mathrm{T}^*\,\mathrm{T}\,h_0\|_{\mathcal{H}} \\
=& \|(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I-\lambda I)h_0 - (\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}(\mathrm{T}^*\,\mathrm{T}+\lambda I-\lambda I)h_0\|_{\mathcal{H}} \\
=& \|\lambda[(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1} - (\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}]h_0\|_{\mathcal{H}} \\
=& \|\lambda(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}[\widehat{\mathrm{T}}^*\widehat{\mathrm{T}} - \mathrm{T}^*\,\mathrm{T}](\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}h_0\|_{\mathcal{H}} \\
=& \|\lambda(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}[\widehat{\mathrm{T}}^*(\widehat{\mathrm{T}}-\mathrm{T}) + (\widehat{\mathrm{T}}^* - \mathrm{T}^*)\,\mathrm{T}](\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}h_0\|_{\mathcal{H}} \\
\leq& \|(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\widehat{\mathrm{T}}^*\|_{op}\,\|\widehat{\mathrm{T}}-\mathrm{T}\,\|_{op}\,\|\lambda(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\|_{op}\,\|h_0\|_{\mathcal{H}} \\
&+ \|\lambda(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\|_{op}\,\|\widehat{\mathrm{T}}^* - \mathrm{T}^*\,\|_{op}\,\|\,\mathrm{T}(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\|_{op}\,\|h_0\|_{\mathcal{H}} = O_P\left(\frac{1}{\sqrt{n\lambda}}\right).
\end{aligned}$$

Here we use that $\|\widehat{\mathrm{T}}-\mathrm{T}\,\|_{op} = \|\widehat{\mathrm{T}}^* - \mathrm{T}^*\,\|_{op} = O_P(1/\sqrt{n})$ from Lemma B.2(vi), and that $\|\lambda(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\|_{op} \leq c$, $\|\lambda(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\|_{op} \leq c$, and $\|(\widehat{\mathrm{T}}^*\widehat{\mathrm{T}}+\lambda I)^{-1}\widehat{\mathrm{T}}^*\|_{op} \leq c'/\sqrt{\lambda}$ from Lemma B.1. Gathering results gives (A.26).

Step 4. We here show convergence of our estimators. Since T is injective from Step 1 and compact from Lemma B.2(i), $\|(\mathrm{T}^*\,\mathrm{T}+\lambda I)^{-1}\mathrm{T}^*\mathrm{T}h - h\|_{\mathcal{H}} = o(1)$ for all h whenever $\lambda \to 0$, see Kress (1999, Definition 15.5 and Theorem 15.23). Hence $\|h_\lambda - h_0\|_{\mathcal{H}} = o(1)$. This and (A.26) yields $\|\widehat{h} - h_0\|_{\mathcal{H}} = o_P(1)$ if in addition $n\lambda \to \infty$.

We now show that $\|\widehat{\beta}-\beta_0\| = O_P\left(\frac{1}{\sqrt{n}} + \|\widehat{h}-h_0\|_{\mathcal{H}}\right)$. From (A.25),

$$
\begin{aligned}
\widehat{\beta}-\beta_0 =& [(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1} - (\mathrm{B}^*\mathrm{B})^{-1}]\widehat{\mathrm{B}}^*(\widehat{r}-\widehat{\mathrm{A}}\widehat{h}) + (\mathrm{B}^*\mathrm{B})^{-1}[\widehat{\mathrm{B}}^* - \mathrm{B}^*](\widehat{r}-\widehat{\mathrm{A}}\widehat{h}) \\
&+ (\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^*(\widehat{r}-r) + (\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^*(\mathrm{A}-\widehat{\mathrm{A}})\widehat{h} - (\mathrm{B}^*\mathrm{B})^{-1}\mathrm{B}^*\mathrm{A}(\widehat{h}-h_0) \\
\Rightarrow \|\widehat{\beta}-\beta_0\| \leq& \|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1} - (\mathrm{B}^*\mathrm{B})^{-1}\|_{op}\|\widehat{\mathrm{B}}^*\|_{op}\left(\|\widehat{r}\|_\mu + \|\widehat{A}\|_{op}\|\widehat{h}\|_{\mathcal{H}}\right) \\
&+ \|(\mathrm{B}^*\mathrm{B})^{-1}\|_{op}\|\widehat{\mathrm{B}}^* - B^*\|_{op}\left(\|\widehat{r}\|_\mu + \|\widehat{A}\|_{op}\|\widehat{h}\|_{\mathcal{H}}\right) \\
&+ \|(\mathrm{B}^*\mathrm{B})^{-1}\|_{op}\|B^*\|_{op}\left(\|\widehat{r}-r\|_\mu + \|\widehat{A}-A\|_{op}\|\widehat{h}\|_{\mathcal{H}} + \|A\|_{op}\|\widehat{h}-h_0\|_{\mathcal{H}}\right).
\end{aligned}
$$

Lemma B.1 ensures that $\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1} - (\mathrm{B}^*\mathrm{B})^{-1}\|_{op}$, $\|\widehat{\mathrm{B}}^* - B^*\|_{op} = \|\widehat{\mathrm{B}} - B\|_{op}$, $\|\widehat{r}-r\|_\mu$, and $\|\widehat{A}-A\|_{op}$ all are $O_P(n^{-1/2})$. We have $\|B^*\|_{op} = \|\mathrm{B}\|_{op} < \infty$ as B is a linear operator with finite dimensional domain, see Kreyszig (1978, Theorem 2.7-8), and $\|\mathrm{B}^*\mathrm{B}\|_{op} = \|\mathrm{B}\|_{op}^2$. Similarly, $\|(\mathrm{B}^*\mathrm{B})^{-1}\|_{op} < \infty$ as B is injective. From Lemma B.2(i), A is compact and hence bounded, and from Lemma B.2(iv) $\|\widehat{\mathrm{A}}\|_{op} = O_P(1)$. From a similar reasoning, $\|\widehat{\mathrm{B}}^*\|_{op} = O_P(1)$. Also $\|\widehat{h}-h_0\|_{\mathcal{H}} = o_P(1)$ implies $\|\widehat{h}\|_{\mathcal{H}} = O_P(1)$. Combine these results to obtain that

$$
\|\widehat{\beta}-\beta_0\| = O_P\left(n^{-1/2} + \|\widehat{h}-h_0\|_{\mathcal{H}}\right) = o_P(1)\,.
$$

Since $\widehat{g}(z) = (1,z)\widehat{\beta} + \widehat{h}(z)$, to show uniform consistency of $\widehat{g}$ and $\widehat{g}'$, it now suffices to show that $\sup_{z\in[0,1]}|\widehat{h}(z) - h_0(z)|$ and $\sup_{z\in[0,1]}|\widehat{h}'(z) - h_0'(z)|$ are bounded by $\|\widehat{h}-h_0\|_{\mathcal{H}}$. As for any $h \in \mathcal{H}$, $h'(z) = \int_0^z h''(t)dt$,

$$
\sup_{z\in[0,1]}|\widehat{h}'(z) - h_0'(z)| \leq \sup_{z\in[0,1]}\int_0^z |\widehat{h}''(t) - h_0''(t)|dt \leq \int_0^1 |\widehat{h}''(t) - h_0''(t)|dt \leq \|\widehat{h}-h_0\|_{\mathcal{H}}\,,
$$

from Cauchy-Schwartz inequality. Since $h(z) = \int_0^z h'(t)dt$, a similar reasoning yields

$$
\sup_{z\in[0,1]}|\widehat{h}(z) - h_0(z)| \leq \sup_{z\in[0,1]}\int_0^z |\widehat{h}'(t) - h_0'(t)|dt \leq \sup_{z\in[0,1]}|\widehat{h}'(z) - h_0'(z)|\,.
$$

Step 5. We now obtain uniform convergence rates. Assumption 3.3 allows to apply Proposition 3.11 in Carrasco et al. (2007) to the operator T and yields $\|h_\lambda - h_0\|_{\mathcal{H}} = O\left(\lambda^{\frac{\gamma\wedge 2}{2}}\right)$. Combining with the results of Step 3 gives

$$
\|\widehat{h}-h_0\|_{\mathcal{H}} = O_P\left(\frac{1}{\sqrt{n\lambda}} + \lambda^{\frac{\gamma\wedge 2}{2}}\right) \text{ and } \|\widehat{\beta}-\beta_0\| = O_P\left(\frac{1}{\sqrt{n\lambda}} + \lambda^{\frac{\gamma\wedge 2}{2}}\right).
$$

We can now use the same arguments as in Step 4 to obtain uniform convergence rates.

B.2 Proof of Lemma B.2

(i). Let us show A is compact by compact embedding. Define $\widetilde{\mathrm{A}}$ as the extension of A to $L^2([0,1])$, where $L^2([0,1])$ is the space of real-valued square-integrable functions on $[0,1]$, i.e. $\widetilde{\mathrm{A}}h = \mathrm{E}\{h(Z)e^{\mathbf{i}W^\top \cdot}\}$ for any $h \in L^2([0,1])$. For all $h \in \mathcal{H}$, we have $h(z) = \int_0^z \int_0^x h''(t)dtdx$, so that

$$||h||^2_{L^2[0,1]} = \int_0^1 |h(z)|^2 dz \leq \sup_{z\in[0,1]} |h(z)|^2 \leq \sup_{z\in[0,1]} \left| \int_0^z h'(t)dt \right|^2 \leq \sup_{t\in[0,1]} |h'(t)|^2$$
$$\leq \sup_{t\in[0,1]} \left| \int_0^t h''(u)du \right|^2 \leq \left(\int_0^1 |h''(u)|du \right)^2 \leq \|h\|^2_{\mathcal{H}} ,$$

where we used the inequality of Cauchy-Schwartz to obtain the last inequality. Therefore, every bounded set on $(\mathcal{H}, ||\cdot||_{\mathcal{H}})$ is also a bounded set on $(L^2([0,1]), ||\cdot||)$. Hence, compactness of $\widetilde{\mathrm{A}}$ implies compactness of A. Now for any $h \in L^2[0,1]$,

$$(\widetilde{A}h)(t) = \mathrm{E}\{h(Z)\ \mathrm{E}\{\exp(\mathbf{i}W^T t)|Z\}\} = \int h(z)\ \mathrm{E}\{\exp(\mathbf{i}W^T t)|Z=z\}\, f_Z(z)\, dz\, ,$$

where

$$\int \left|\mathbb{E}\{\exp(\mathbf{i}W^T t)|Z=z\}\right|^2 \mu(t) f_Z(z)\, dt\, dz \leq 1\, ,$$

as $|\exp(\mathbf{i}\cdot)| \leq 1$. Since $\widetilde{A}$ is an integral operator whose kernel is Hilbert-Schmidt, i.e. squared integrable, we can apply Busby et al. (1972, Proposition 2.1) to conclude that $\widetilde{A}$ is compact.

Let us now show that T is compact. The range of B, $\mathcal{R}(\mathrm{B})$, is finite dimensional, linear, and closed. P is the orthogonal projection onto $\mathcal{R}(\mathrm{B})$, and is thus bounded by Kress (1999, Theorem 13.3). Hence, $\mathrm{M} = I - \mathrm{P}$ is bounded as well. Since $\mathrm{T} = \mathrm{M}\,\mathrm{A}$ is the composition of a bounded and a compact operator, it is compact by Kress (1999, Theorems 2.14 and 2.16).

(ii). For $\beta \in \mathbb{R}^2$, we have

$$\|(\widehat{\mathrm{B}} - \mathrm{B})\beta\|^2_\mu = \int |(\mathrm{E}_n - \mathrm{E})\{\exp(\mathbf{i}W^\top t)(1, Z)\}\beta|^2 \mu(dt)\, ,$$

where E_n denotes the empirical expectation. By the Cauchy-Schwarz inequality,

$$\mathrm{E}\,\|\widehat{\mathrm{B}} - \mathrm{B}\,\|^2_{op} \leq \mathrm{E}\left\{ \int (|(\mathrm{E}_n - \mathrm{E})\{\exp(\mathbf{i}W^\top t)\}|^2 + |(\mathrm{E}_n - \mathrm{E})\{Z\exp(\mathbf{i}W^\top t)\}|^2)\mu(dt) \right\}.$$

Since data are i.i.d.,

$$
\begin{aligned}
E\left\{|(\mathrm{E}_n-\mathrm{E})\{Z\exp(\mathbf{i}W^\top t)\}|^2\right\} &= \mathrm{E}\left\{\left|n^{-1}\sum_{i=1}^n Z_i\exp(\mathbf{i}W_i^\top t)-\mathrm{E}\{Z\exp(\mathbf{i}W^\top t)\}\right|^2\right\}\\
&= \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^n Z_i\exp(\mathbf{i}W_i^T t)\right\}\\
&= n^{-1}(\mathrm{E}\{|Z\exp(\mathbf{i}W^\top t)|^2\}-|\,\mathrm{E}\{Z\exp(\mathbf{i}W^\top t)\}|^2)=O(n^{-1})\,,
\end{aligned}
$$

as $|Z\exp(\mathbf{i}W^\top t)|\le 1$ for all $t\in\mathcal{T}$. Similarly, $\mathrm{E}\left\{|(\mathrm{E}_n-\mathrm{E})\{\exp(\mathbf{i}W^\top t)\}|^2\right\}=O(n^{-1})$. This implies $\mathrm{E}\,\|\widehat{B}-B\|_{op}^2=O(n^{-1})$, and by Markov's inequality, $\|\widehat{\mathrm{B}}-\mathrm{B}\,\|_{op}=O_P(n^{-1})$.

(iii). From Kreyszig (1978, Theorem 2.7-8), as B is a linear operator with a finite dimensional domain, it is bounded, and $\|\,\mathrm{B}\,\|_{op}<\infty$. Also $\|\,\mathrm{B}^*\,\|_{op}=\|\,\mathrm{B}\,\|_{op}$ and $\|\,\mathrm{B}^*\,\mathrm{B}\,\|_{op}=\|\,\mathrm{B}\,\|_{op}^2$. The operator $\mathrm{B}^*\,\mathrm{B}$ maps $\mathbb{R}^2$ into $\mathbb{R}^2$, and is thus a matrix. From (ii), $\|\widehat{\mathrm{B}}\|_{op}$ and $\|\widehat{\mathrm{B}}^*\|_{op}$ are $O_P(1)$, and

$$
\begin{aligned}
\|\,\mathrm{B}^*\,\mathrm{B}-\widehat{\mathrm{B}}^*\widehat{\mathrm{B}}\|_{op} &= \|(\mathrm{B}^*-\widehat{\mathrm{B}}^*)B+\widehat{\mathrm{B}}^*(\mathrm{B}-\widehat{\mathrm{B}})\|_{op}\\
&\le \|\,\mathrm{B}^*-\widehat{\mathrm{B}}^*\|_{op}\,\|\,\mathrm{B}\,\|_{op}+\|\widehat{\mathrm{B}}^*\|_{op}\,\|\,\mathrm{B}-\widehat{\mathrm{B}}\|_{op}=O_P(n^{-1/2})\,.
\end{aligned}
$$

Since B is injective, $\mathrm{B}^*\,\mathrm{B}$ is invertible, $(\mathrm{B}^*\,\mathrm{B})^{-1}$ exists and is bounded. By the continuous mapping theorem, $\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}-(\mathrm{B}^*\,\mathrm{B})^{-1}\|_{op}=o_P(1)$. Hence $\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}\|_{op}\le\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}-(\mathrm{B}^*\,\mathrm{B})^{-1}\|_{op}+\|(\mathrm{B}^*\,\mathrm{B})^{-1}\|_{op}=O_P(1)$. Moreover,

$$
\begin{aligned}
\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}-(\mathrm{B}^*\,\mathrm{B})^{-1}\|_{op} =&\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}(\mathrm{B}^*\,\mathrm{B}-\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})(\mathrm{B}^*\,\mathrm{B})^{-1}\|_{op}\\
\le&\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}\|_{op}\,\|\,\mathrm{B}^*\,\mathrm{B}-\widehat{\mathrm{B}}^*\widehat{\mathrm{B}}\|_{op}\,\|(\mathrm{B}^*\,\mathrm{B})^{-1}\|_{op}=O_P(n^{-1/2})\,.
\end{aligned}
$$

For the difference between $\mathrm{M}=I-\mathrm{B}(\mathrm{B}^*\,\mathrm{B})^{-1}\,\mathrm{B}^*$ and $\widehat{\mathrm{M}}=I-\widehat{\mathrm{B}}(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}\widehat{\mathrm{B}}^*$, we have

$$
\begin{aligned}
\|\widehat{\mathrm{M}}-\mathrm{M}\,\|_{op} =&\|(\widehat{\mathrm{B}}-\mathrm{B})(\widehat{\mathrm{B}^*\mathrm{B}})^{-1}\widehat{\mathrm{B}}^*+\mathrm{B}[(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}-(\mathrm{B}^*\,\mathrm{B})^{-1}]\widehat{\mathrm{B}}^*+\mathrm{B}(\mathrm{B}^*\,\mathrm{B})^{-1}[\widehat{\mathrm{B}}^*-\mathrm{B}^*]\|_{op}\\
\le&\|\widehat{\mathrm{B}}-\mathrm{B}\,\|_{op}\,\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}\|_{op}\,\|\widehat{\mathrm{B}}^*\|_{op}+\|\,\mathrm{B}\,\|_{op}\,\|(\widehat{\mathrm{B}}^*\widehat{\mathrm{B}})^{-1}-(\mathrm{B}^*\,\mathrm{B})^{-1}\|\,\|\widehat{\mathrm{B}}^*\|_{op}\\
&+\|\,\mathrm{B}\,\|_{op}\,\|(\mathrm{B}^*\,\mathrm{B})^{-1}\|_{op}\,\|\widehat{\mathrm{B}}^*-\mathrm{B}^*\,\|_{op}=O_P(n^{-1/2})\,.
\end{aligned}
$$

(iv). Recall that for any $h\in\mathcal{H}$, $h(z)=\int_0^z\int_0^x h''(u)dudx$. Thus,

$$
\begin{aligned}
(\widehat{A}h)(t) =&\frac{1}{n}\sum_{i=1}^n h(Z_i)\exp(\mathbf{i}W_i^T t)=\frac{1}{n}\sum_{i=1}^n\int_0^{Z_i}\int_0^x h''(u)\,du\,dx\,\exp(\mathbf{i}W_i^T t)\\
=&\int_{[0,1]^2}h''(u)\left[\frac{1}{n}\sum_{i=1}^n\mathbf{1}(0<u<x)\,\mathbf{1}(0<x<Z_i)\,\exp(\mathbf{i}W_i^T t)\right]du\,dx\\
=&\int_0^1 h''(u)\left[\int_0^1\frac{1}{n}\sum_{i=1}^n\mathbf{1}(u<x<Z_i)\,\exp(\mathbf{i}W_i^T t)\,dx\right]du\\
=&\int_0^1 h''(u)\,\widehat{k}(u,t)\,du\,,
\end{aligned}
\tag{A.27}
$$

where $\widehat{k}(u,t)$ is defined implicitly above. Exchanging the empirical measure with the population probability and using the same steps as above yield

$$(Ah)(t) = \int_0^1 h''(u) \left[\int_0^1 \mathrm{E}\{\mathbf{1}(u < x < Z) \exp(\mathbf{i}W^T t)\}\, dx\right] = \int_0^1 h''(u)\, k(u,t)\, du\,. \quad (A.28)$$

where $k(u,t) = \mathrm{E}\,\widehat{k}(u,t)$ is defined implicitly above. Next,

$$\begin{aligned}
\|\widehat{A} - A\|_{op}^2 &= \sup_{h\in\mathcal{H}\, \|h\|_{\mathcal{H}}=1} \|\widehat{A}h - Ah\|_\mu^2 = \sup_{h\in\mathcal{H}\, \|h\|_{\mathcal{H}}=1} \int \left|(\widehat{A}h)(t) - (Ah)(t)\right|^2 \mu(dt) \\
&= \sup_{h\in\mathcal{H}\, \|h\|_{\mathcal{H}}=1} \int \left| h''(u) \left[\widehat{k}(u,t) - k(u,t)\right] du \right|^2 \mu(dt) \\
&\leq \sup_{h\in\mathcal{H}\, \|h\|_{\mathcal{H}}=1} \int \left\{ \int_0^1 |h''(u)|^2 du \int_0^1 \left|\widehat{k}(u,t) - k(u,t)\right|^2 du \right\} \mu(dt) \\
&= \sup_{h\in\mathcal{H}\, \|h\|_{\mathcal{H}}=1} \|h\|_{\mathcal{H}} \int \left\{ \int_0^1 \left|\widehat{k}(u,t) - k(u,t)\right|^2 du \right\} \mu(dt) \\
&= \int_{[0,1]\times\mathbb{R}^q} \left|\widehat{k}(u,t) - k(u,t)\right|^2 du \otimes \mu(dt) \\
\Rightarrow \mathrm{E}\,\|\widehat{A} - A\|_{op}^2 &\leq \int_{[0,1]\times\mathbb{R}^q} \mathrm{E}\left|\widehat{k}(u,t) - k(u,t)\right|^2 du \otimes \mu(dt)\,.
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{E}\left|\widehat{k}(u,t) - k(u,t)\right|^2 &= \mathrm{E}\left|\int_0^1 (\mathrm{E}_n - \mathrm{E})\mathbf{1}(u < x < Z) \exp(\mathbf{i}W^T t)\, dx\right|^2 \\
&\leq \int_0^1 \mathrm{E}\left|(\mathrm{E}_n - \mathrm{E})\mathbf{1}(u < x < Z) \exp(\mathbf{i}W^T t)\right|^2 dx \\
&= \int_0^1 \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^n \mathbf{1}(u < x < Z_i) \exp(\mathbf{i}W_i^T t)\right\} dx \\
&= \frac{1}{n}\int_0^1 \mathrm{Var}\left\{\mathbf{1}(u < x < Z) \exp(\mathbf{i}W^T t)\right\} dx = O(n^{-1})\,.
\end{aligned}$$

Use Markov's inequality to obtain the desired result.

(v). By reasoning as in the proof of (i), compactness of $\widehat{\mathrm{T}} = \widehat{\mathrm{M}}\widehat{\mathrm{A}}$ follows if $\widehat{\mathrm{M}}$ is bounded and $\widehat{\mathrm{A}}$ is compact. The first claim is shown following similar arguments as in (i). To obtain compactness of $\widehat{\mathrm{A}}$, we will use Theorem 8.1-4 in Kreyszig (1978) stating that a bounded operator with a finite dimensional range is compact. As

$$\widehat{A}h = \sum_{i=1}^n h(Z_i)\, \frac{1}{n} \exp(W_i^T \cdot) \in \mathrm{Span}\left(\frac{1}{n}\exp(W_1^T \cdot), \ldots, \frac{1}{n}\exp(W_n^T \cdot)\right)\,,$$

the range of $\widehat{\mathrm{A}}$ is finite dimensional for all n. Moreover, using (A.27)

$$\|\widehat{\mathrm{A}}h\|_\mu^2 = \int \left|\int_0^1 h''(u)\, \widehat{k}(u,t)\, du\right|^2 \mu(dt) \leq \|h\|_{\mathcal{H}}^2 \sup_{u,t} |\widehat{k}(u,t)|^2\, du \leq \|h\|_{\mathcal{H}}^2\,,$$

as $|\widehat{k}(u,t)| \leq 1$. Hence, $\|\widehat{\mathrm{A}}\|_{op} \leq 1$, and $\widehat{\mathrm{A}}$ is compact.

(vi). Since $\widehat{\mathrm{T}} - \mathrm{T} = (\widehat{\mathrm{M}} - \mathrm{M})\widehat{\mathrm{A}} + \mathrm{M}(\widehat{\mathrm{A}} - \mathrm{A})$, the result follows from (ii), (iv), and the fact that M is bounded.

(vii). The proof is analogous to the proof of (ii).

(viii). Write $\widehat{\mathrm{M}}\widehat{r} - \mathrm{M}\, r = (\widehat{\mathrm{M}} - \mathrm{M})\widehat{r} + \mathrm{M}(\widehat{r} - r)$, and use $\|\widehat{\mathrm{M}} - \mathrm{M}\,\|_{op} = O_P(n^{-1/2})$, $\|\widehat{r} - r\|_{\mu} = O_P(n^{-1/2})$, and $\|\,\mathrm{M}\,\|_{op} < \infty$ from previous items to obtain $\|\widehat{\mathrm{M}}\widehat{r} - \mathrm{M}\, r\|_{\mu} = O_P(n^{-1/2})$. Use (vi) above to get $\|(\widehat{\mathrm{M}}\widehat{r} - \widehat{T}h_0) - (\mathrm{M}\, r - \mathrm{T}\, h_0)\|_{\mu} = O_P(n^{-1/2})$, and note that $\mathrm{M}_Z\, r - \mathrm{T}\, h_0 = 0$.