Electronic Journal of Statistics
Vol. 15 (2021) 5333–5379
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1921

Nonparametric estimation of accelerated failure-time models with unobservable confounders and random censoring*

Samuele Centorrino†

Economics Department
Stony Brook University
Stony Brook, NY USA
e-mail: samuele.centorrino@stonybrook.edu

Jean-Pierre Florens

Toulouse School of Economics
University of Toulouse Capitole
Toulouse, France
e-mail: jean-pierre.florens@tse-fr.eu

Abstract: We consider nonparametric estimation of an accelerated failure-time model when the response variable is randomly censored on the right, and regressors are not mean independent of the error component. This dependence can arise, for instance, because of measurement error. We achieve identification and conduct estimation using a vector of instrumental variables. Censoring is independent of the response variable given the instruments. We consider settings in which regressors are continuously distributed. However, the instruments may or may not be continuous, and we show how various independence restrictions allow us to identify and estimate the unknown function of interest depending on the nature of instruments. We provide rates of convergence of our estimator and showcase its finite sample properties in simulations.

MSC2020 subject classifications: Primary 62N01,62N02, 62G08; secondary 45A05,45G05.
Keywords and phrases: Accelerated failure-time models, censoring, instrumental variables, nonparametric, regularization, Landweber-Fridman.

Received February 2021.

Contents

*Jean-Pierre Florens acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future program (Investissements d'Avenir, grant ANR-17-EURE-0010).
†Corresponding author.

1. Introduction

We consider identification and estimation of the following nonparametric accelerated failure-time (AFT) model in log form

$$T = \varphi(Z) + U, \tag{1.1}$$

when the conventional assumption that $E(U \mid Z) = 0$ no longer holds due to potential dependence between U and Z. This dependence can arise for several reasons: measurement error, omitted variables, or simultaneity.

For instance, when analyzing the effect of systolic blood pressure as a possible risk factor for developing cardiovascular diseases, measurement error is often an issue due to time constraints and other unobservable factors in routine care. This measurement error can bias, in an unknown direction, statistical evaluations of this effect [10].

Similarly, recent studies have tried to uncover the relation between unemployment status and body mass index (BMI) to explain the elevated morbidity and mortality among job seekers [see 43, among others]. In particular, one may be interested in understanding how BMI affects spells of unemployment duration. However, it is plausible that there are individual characteristics, unobserved to the statistician, that may determine both the length of unemployment spells and the subject's physical well-being. This simultaneity issue may render estimators based on the standard assumption that $E[U \mid Z] = 0$ inconsistent.

Moreover, T is often not fully observed, and we assume it is subject to random right censoring, C. In particular, we consider a setting in which one observes $Y = T \wedge C$ and $\delta = \mathbb{1}(T \leq C)$.

Our analysis focuses on the regression function φ when the regressor Z is restricted to have a continuous distribution with respect to the Lebesgue measure. To achieve identification and carry on estimation, we rely on a vector of instrumental variables, W, which are taken to satisfy some independence restrictions with respect to the error term. In the examples provided above, plausible instrumental variables are given by the same measurement run on parents or relatives.

For instance, parents' BMI is often used as an instrument for an individual's own BMI [see, e.g., 51, among others]. Similarly, genetic markers can be used as a source of exogenous variation to help the identification of causal effects in these settings [69].

In this paper, we consider that the censoring variable C is independent of T given the exogenous instruments, W. We thus exclude unobservable factors that affect both T and C simultaneously. Upon additional exclusion restrictions that we discuss below, we can recover the unknown function φ.

We consider two settings, depending on the properties of W. In the first setting, we take W to be mean independent of the error term, i.e., $E[U \mid W] = 0$. We recover the regression function φ by solving the following integral equation

$$E(V \mid W) = E(\varphi(Z) \mid W), \tag{1.2}$$

where V is an appropriate transformation of the censored response Y [see, e.g., 48, 66, among others]. Identification in this setting requires, among other things, W to be continuously distributed. One can then recover an estimator of φ by replacing population objects in equation (1.2) with sample counterparts.

In a second setting, we consider the stronger assumption that $U \perp\!\!\!\perp W$, i.e., W is independent of U, and $E(U) = 0$. Independence is equivalent to

$$S_{U|W}(u \mid w) = S_U(u) \quad \forall u, w, \tag{1.3}$$

where $S_{U|W}(u \mid w)$ and $S_U(u)$ are the conditional and unconditional survivor functions of U, respectively [14, 27, 26]. Both survivor functions can be consistently estimated using the conditional and unconditional Kaplan-Meier estimators to take censoring into account [22, 46, 68]. Upon additional conditions that guarantee the consistency of the product-limit estimator, censoring can be very easily handled in this setting, and it does not require any major modification of the estimation procedure.

In parametric models, the assumption of independence is often justified by efficiency considerations. One can decrease the variance of a parametric estimator by taking one step towards the Maximum Likelihood Estimator [63, 65]. In this nonparametric setting, the stronger assumption of independence simply allows us to relax the relevance condition and accommodate settings in which W may only be binary or discrete.

In both cases, an additional technical difficulty for implementation and for the derivation of the asymptotic properties is that the resulting estimators are the solution, respectively, of a linear and a nonlinear ill-posed inverse problem. Hence, besides the smoothing step, which is common in nonparametric regressions, we have a further regularization step [see 28, 45].

Popular regularization approaches in the literature include Tikhonov regularization [24, 38, 67]; Landweber-Fridman regularization [34, 36, 50]; and a more recent method based on finite-dimensional approximations of the function space, which has become popular in econometrics [so-called sieve regularization, see, e.g., 19, 41, 62].

In this paper, we consider regularization through the Landweber-Fridman (LF) approach. In finite samples, the relative advantage of LF regularization is that it iteratively approximates the inverse of a conditional expectation operator. Thus, it avoids exact inversion of large matrices necessary, for instance, in Tikhonov regularization. Compared to sieve regularization, it does not require that the unknown function is well approximated by just a few terms of its series expansion [see, e.g., 19].

We provide a detailed explanation of the implementation of the Nonparametric IV estimator with LF regularization, and we derive an upper bound on the L^2 loss. We show that, under our identification assumptions, the random right-censoring does not affect the properties of the estimator. That is, the rate of convergence is the same as when the dependent variable is fully observed. When the instrument is mean independent, this result holds upon an appropriate choice of the bandwidth parameter.

The estimator based upon mean-independence is a linear estimator and thus relatively straightforward to implement. The L^2 rates are minimax under weak conditions on the smoothing and regularization parameters [20]. By contrast, the nonlinearity of the estimation procedure based on independence introduces some significant theoretical and practical difficulties. The LF procedure relies upon a first-order linear approximation of the nonlinear problem. Hence, regularity conditions only hold in the vicinity of the true solution. Moreover, the loss function can be decomposed into two parts: one which is due to estimation and can be handled similarly as in the case with mean-independence; and another one due to the nonlinearity of the inverse problem. The latter determines the convergence of our estimator and, in certain instances, can be controlled to reach the same rates as in the linear case. Whether these rates are minimax remains an open question, to the best of our knowledge.

Related work has considered the estimation of duration models with endogeneity and (possibly random) right-censoring. Frandsen [35] discusses nonparametric identification and estimation of a model with a binary endogenous treatment variable and a binary instrument, independent of the error term [see also 70]. More recently, Beyhum et al. [6] analyze a nonparametric duration model with endogenous treatment. They provide identification and estimation based on an instrumental variable assumption when the outcome is randomly censored on the right. Their estimator is also a solution to a nonlinear inverse problem, although they avoid ill-posedness by restricting the endogenous treatment to be discrete. They also discuss partial identification of the treatment effect when censoring is fixed. Sant'Anna [64] provides a nonparametric test of treatment effect heterogeneity for a binary treatment variable in cases where the treatment is assigned independently of the potential outcome conditional on observables. He also assumes that outcome and censoring are independent conditional on the treatment. He also allows for the treatment to be endogenous. Our work contributes to this literature by allowing the treatment variable to be continuous and potentially endogenous. However, all these papers assume that the effect of the treatment is heterogeneous, which is ruled by the additive separability of our model. This would be an interesting contribution, that we defer to further

research.

2. Identification

2.1. Framework

We consider a random element (T, Z, W) with $T \in \mathbb{R}$, $Z \in \mathbb{R}^p$, and W is a q-dimensional random vector, with $q \geq p$. We let F denote the joint distribution of the random vector (T, Z, W); and L^2_Z or L^2_W, the spaces of functions of Z or W, respectively, that are square-integrable with respect to F. Depending on the setting, we will impose additional restrictions on the distribution of (Z, W).

We analyze the model

$$Y = T \wedge C = (\varphi(Z) + U) \wedge C, \quad \varphi \in L^2_Z, \tag{2.1}$$

with $\delta = \mathbb{1}(T \leq C)$, and $C \in \mathbb{R}$. We maintain the following assumption.

Assumption 2.1. $T \perp\!\!\!\perp C \mid W$.

This assumption allows any relations between the unobserved response and the censoring variable to happen through observable components. For instance, the restriction in Assumption 2.1 holds when the censoring variable can be written as $C = \psi(W, \nu)$, with $\nu \perp\!\!\!\perp (T, Z, W)$. When Z is exogenous, that is $W = Z$, Assumption 2.1 reduces to the standard exclusion restriction commonly imposed in AFT regressions with random censoring [see 48].

In the following, we let $S_{\cdot|W}(\cdot \mid w)$ be the survivor function conditional on W. Our identification strategy is based on the following assumption about $S_{C|W}(C \mid w)$.

Assumption 2.2. *Let $\mathcal{T}$ be the support of T, such that $\sup_{t\in\mathcal{T}} |t| = T_0 < \infty$. For every w, we have that $S_{C|W}(T_0 \mid w) > \epsilon$, for a constant $\epsilon > 0$.*

Assumption 2.2 implies that the supremum of $\mathcal{T}$ is not censored with positive probability. This condition is relatively standard in this literature, and point identification of the parameters of interest is not possible without a similar restriction, to the best of our knowledge. For instance, when T represents unemployment spells and a randomly assigned interview determines censoring, Assumption 2.2 implies that the longest duration of a spell is finite and that interviews can be conducted late enough to guarantee that at least some of the individuals with the longest spell are interviewed after they found employment. If Assumption 2.2 does not hold, one can only hope to identify φ for those t which are in the interior of the support of C [6]. Assumption 2.2 is violated, in particular, when censoring is fixed (see Remark 1 below).

2.2. Case 1: U is mean independent of W

We first treat the case in which $E(U \mid W) = 0$ and the joint distribution of (T, Z, W) is absolutely continuous with respect to the Lebesgue measure.

Define the following random variable

$$V = \frac{\delta Y}{S_{C|W}(Y \mid W)}, \tag{2.2}$$

where $S_{C|W}$ is the survivor function of C conditional on W. We have $Y = T$, whenever $\delta = 1$.

To achieve identification of the function φ, we consider the following assumption.

Assumption 2.3. *$E(\varphi(Z) \mid W) = 0$ implies $\varphi = 0$, almost surely.*

This *completeness condition* is an unsettled assumption for identification in nonparametric instrumental regressions. The terminology used is in analogy with the notion of complete statistic [see, e.g., 52], and it is sometimes referred to as a strong identification condition [see, e.g., 31]. When the pair (Z, W) is continuously distributed, Andrews [3] has derived a class of distributions for which completeness holds *generically*, in a sense defined within that paper. Some additional results about completeness that rely on stronger restrictions on the DGP are provided in D'Haultfoeuille [25]. When completeness fails, Babii and Florens [4] and Florens et al. [32] show that the estimator may still converge to the minimal norm solution.

Under the conditions above, we have the following proposition.

Proposition 2.1. *Under Assumptions 2.1-2.3, the regression function, φ, is identified.*

Proof. From the definition of V, and when Assumptions 2.1 and 2.2 hold, we directly have that

$$E(V \mid W) = E\left(\frac{\delta Y}{S_{C|W}(Y \mid W)} \mid W\right) = E(T \mid W) = E(\varphi(Z) \mid W).$$

Let φ_1 and φ_2, two possible solutions to the integral equation $E(T - \varphi(Z) \mid W) = 0$. Then we must have that $E(\varphi_1(Z) - \varphi_2(Z) \mid W) = 0$. By Assumption 2.3, this is only true if and only if $\varphi_1 = \varphi_2$, almost surely. This concludes the proof. □

Remark 1 (Identification with fixed censoring)**.** *When censoring is fixed, one could achieve point identification of the function φ (up to location) as follows. Let $\varepsilon = T - E[T|W]$, with $\varepsilon \perp\!\!\!\perp W$. Then we can identify $E[T|W]$ using the approach in Lewbel and Linton [53]. Finally, φ can be identified by solving the linear integral equation $E[T|W] = E[\varphi(Z)|W]$, if Assumption 2.3 holds. The additional assumption that $\varepsilon \perp\!\!\!\perp W$ is strong and might be justified only in specific settings [15].*

2.3. Case 2: U is independent of W

We now turn to the case when $U \perp\!\!\!\perp W$ with $E(U) = 0$.

The joint distribution of (T, Z) is still restricted to be absolutely continuous with respect to the Lebesgue measure. However, we do not impose any condition on the distribution of W. Therefore, we can identify the function φ with purely discrete instruments. Our presentation is largely based on Centorrino et al. [14], Dunker [26], and Dunker et al. [27], who consider identification and estimation in a similar model without random censoring.

We rewrite the independence condition as follows:

$$F(t, z \mid w) = \frac{\partial}{\partial z_1} \cdots \frac{\partial}{\partial z_p} P(T \geq t, Z \geq z \mid W = w), \tag{2.3}$$

and

$$F(t, z) = \frac{\partial}{\partial z_1} \cdots \frac{\partial}{\partial z_p} P(T \geq t, Z \geq z), \tag{2.4}$$

where, roughly speaking F is a survivor function in terms of t, and the negative of the density as a function of z.

The independence restriction, therefore, implies that

$$\int F(\varphi(z) + u, z \mid w) dz = \int F(\varphi(z) + u, z) dz. \tag{2.5}$$

We notice that, conceptually, nothing changes compared to the case when T is fully observed, at least for identification purposes. The error term U still has a well-defined (conditional and unconditional) survivor function. The main difference with the existing approach will be tackled in estimation, where the standard nonparametric estimators are replaced with Kaplan-Meier estimators.

Equation (2.5) may be written as

$$A(\varphi_\dagger) = \int \left[F(\varphi_\dagger(z) + u, z \mid w) - F(\varphi_\dagger(z) + u, z) \right] dz = 0, \tag{2.6}$$

which defines a nonlinear integral equation of the first kind, where $\varphi_\dagger$ is its true solution.

Identification of $\varphi_\dagger$ is more complex in this context. In particular, given the nonlinear nature of the integral equation, we have to consider both conditions for global and local identification. We focus here on the latter that are easier to derive and are more easily interpretable. Interested readers can refer to Centorrino et al. [14], Chernozhukov and Hansen [21], and Fève et al. [30], for a discussion of global identification conditions in this context.

Our discussion of local identification is based on the linearization of the operator $A(\cdot)$. We provide mild sufficient conditions such that its Fréchet derivative exists and it is well-behaved. This discussion of local identification will lead us to impose, among others, a condition that is similar to the one in Assumption 2.3.

We start by assuming the following.

Assumption 2.4. *$F(t, z \mid w)$ and $F(t, z)$ are differentiable with respect to t. Their first partial derivatives with respect to t are the conditional density and the*

density, denoted $f_{T,Z|W}(t,z \mid w)$ *and* $f_{T,Z}(t,z)$*. These densities are continuous and have continuous and uniformly bounded first partial derivatives with respect to their first argument.*

Under the conditions in Assumption 2.4, the nonlinear operator A is Fréchet differentiable, and its Fréchet derivative $A'_\varphi(\tilde{\varphi})$ satisfies

$$A'_\varphi(\tilde{\varphi}) = -\int \left(f_{T,Z|W}(\varphi(z)+u,z \mid w) - f_{T,Z}(\varphi(z)+u,z)\right)\tilde{\varphi}(z)dz, \tag{2.7}$$

where $A'_\varphi(\tilde{\varphi})$ denotes the derivative of A at φ as a linear function of $\tilde{\varphi}$.

The nonlinear operator A is defined on the centered functions in L^2_Z and valued in $L^2_{U\times W}$. The mean of φ can be identified by noticing that $E(T) = E(\varphi(Z))$, as long as $E(U) = 0$. Therefore, we restrict our attention to the space of L^2 centered functions of Z, without loss of generality.

The Fréchet derivative A'_φ operates between L^2_Z and $L^2_{U\times W}$, and, under the conditions in Assumption 2.4 is a continuous linear operator for any φ. Under additional minor regularity conditions, it is also a Hilbert Schmidt operator, and thus compact and bounded [see 11].

We let $\mathcal{D}(A)$ be the domain of A, and for some finite constant $R > 0$, let

$$\mathcal{B}_R(\varphi_\dagger) \equiv \{\varphi \in L^2_Z : \|\varphi - \varphi_\dagger\| < R\}. \tag{2.8}$$

We have the following definition.

Definition 2.1 (17)**.** *The model* $T = \varphi(Z) + U$ *is locally identified on* $\mathcal{B}_R(\varphi_\dagger)$ *when* $U \perp\!\!\!\perp W$ *and* $E(U) = 0$*, if*

(i) The operator A *is Fréchet differentiable.*
(ii) A'_φ *is a one-to-one linear operator.*
(iii) There exist a finite constant $M' > 0$*, which depends on* R*, such that, for all* $\varphi \in \mathcal{B}_R(\varphi_\dagger)$*,*

$$\|A(\varphi) - A(\varphi_\dagger) - A'_{\varphi_\dagger}(\varphi - \varphi_\dagger)\| \leq M'\|\varphi - \varphi_\dagger\|^2.$$

Fréchet differentiability of the operator helps control the behavior of our nonlinear problem in the vicinity of the true solution. We further take the Fréchet derivative to be injective, which is tantamount to a rank condition on $A'_{\varphi_\dagger}$ [see 17, Assumption 1, p. 788]. Finally, we need to restrict the amount of nonlinearity that is allowed for the ill-posed inverse problem at hand [see 17, Assumption 2, p. 789]. The last condition can be proven, using uniform boundedness of the first derivative of the conditional and marginal pdfs with respect to their first argument. We omit the proof for brevity. The last statement also follows from a Lipschitz continuity condition on $A'_{\varphi_\dagger}$ [26].

Let

$$\mathcal{E} = \{\varphi \in L^2_Z : E(\varphi) = 0\}.$$

We consider the following Assumptions.

Assumption 2.5 (Conditional completeness)**.** *Let $\varphi \in \mathcal{E}$. Then*

$$E(\varphi(Z) \mid U = u, W = w) \overset{a.s.}{=} E(\varphi(Z) \mid U = u) \Rightarrow \varphi \overset{a.s.}{=} 0.$$

Assumption 2.6 (Scalability)**.** *For every $\varphi \in \mathcal{B}_R(\varphi_\dagger)$, we can write,*

$$A'_\varphi = G_{\varphi_\dagger,\varphi} A'_{\varphi_\dagger},$$

where $\{G_{\varphi_\dagger,\varphi}, \varphi \in \mathcal{B}_R(\varphi_\dagger)\}$ is a family of bounded operators, such that

$$\|G_{\varphi_\dagger,\varphi} - I\| \leq M\|\varphi - \varphi_\dagger\|,$$

with $0 < M < \infty$.

Assumption 2.5 states that the projection of φ under $L^2_{U\times W}$ differs from the projection of φ under L^2_U except if φ is constant. Moreover, this constant is equal to 0, under the additional assumption $E(U) = 0$.

This condition, under independence, is immediately implied by the completeness condition in Assumption 2.3. This is because, under the independence condition $E(E(\varphi(Z) \mid U) \mid W) = E(\varphi(Z)) = 0$, for all $\varphi \in \mathcal{E}$. Therefore

$$\begin{aligned} &E(\varphi(Z) \mid W) = E(\varphi(Z) - E(\varphi(Z)) \mid W) \\ =&E(E(\varphi(Z) \mid W, U) - E(\varphi(Z) \mid U) \mid W) = 0, \end{aligned}$$

which finally implies $E(\varphi(Z) \mid W, U) = E(\varphi(Z) \mid U)$, where equalities are intended almost surely. However, the inverse is not true in general: conditional completeness does not imply completeness.

Assumption 2.5 implies condition (ii) in Definition 2.1. To further clarify its role, we consider the following example from Dunker et al. [27].

Example 1. *Let us assume that* $\begin{pmatrix} U \\ Z \end{pmatrix} \mid W \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho(W) \\ \rho(W) & 1 \end{pmatrix}\right)$ *and $W \in \{0, 1\}$. If $\rho(0) \neq \rho(1)$, Assumption 2.5 is verified and the model $Y = \varphi(Z) + U$ ($U \perp\!\!\!\perp W$ and $E(\varphi(Z)) = 0$) is locally identified [see 27, for a formal proof].*

Assumption 2.6 of scalability is proven in Centorrino et al. [14], under more primitive conditions on the conditional expectation operator. This assumption implies condition (iii) in Definition 2.1 (see also 27, 39).

We thus revisit the conditions in Definition 2.1 to obtain the following.

Proposition 2.2. *If the operator A is Fréchet differentiable and Assumptions 2.1, 2.2, 2.5 and 2.6 hold, then the separable NPIV model is locally identified on $\mathcal{B}_R(\varphi_\dagger)$ if $U \perp\!\!\!\perp W$.*

3. Estimation

3.1. Framework

We observe an IID sample $\{(Y_i, \delta_i, Z_i, W_i), i = 1, \ldots, n\}$ from the joint distribution of the random vector (Y, δ, Z, W). We take the supports of Z and W, when

continuous, to be compact. In particular, we restrict the support of (Z, W) to be the unit hypercube of dimension $p+q$, without loss of generality. One could allow for the support of the data to be unbounded. However, this substantially complicates the proof [see 9, for convergence results in sieve Nonparametric IV when the data have unbounded support].

In the following, we also let $K(v)$ be a standard univariate kernel function, such as Gaussian or Epanechnikov. We also let $K_h(\cdot) = K(h^{-1}\cdot)$, and $\mathbf{K}_h(v) = \prod_j K_h(v_j)$, the standard product kernel, for a scalar bandwidth h. Below we focus mainly on the practical implementation of our estimation procedure. As in any other nonparametric framework, we face the issue of selecting smoothing parameters. However, in nonparametric regressions with instrumental variables, we come across two kinds of 'smoothing parameters'. Namely, the bandwidth used for kernel estimates; and N, the number of iterations, used to regularize the ill-posed inverse problem. Separately, these two problems are standard, and several adaptive rules have been proposed. In the nonparametric instrumental regression setting, bandwidths and the regularization parameter compensate for one another. There likely exists a set of jointly 'optimal' choices for these two elements. However, this is a topic we do not tackle in this paper [see 12, for additional results on the selection of the tuning parameters in linear ill-posed inverse problems]. Below we thus consider data-driven procedures for the choice of tuning parameters that, although not optimal in the sense of oracle minimization of a given risk function, behave reasonably well in practice.

3.2. Case 1: U is mean independent of W

The estimation procedure is based on equation (2.2), where the conditional survivor function $S_{C|W}$ is replaced by a generalization of a Kaplan-Meier type estimator [48, 66]; and on equation (1.2). To estimate the former, we follow the approach of Beran [5] [see also 22, 37, 68]. The latter can be cast as a linear integral equation of the first kind [49].

Beran's (1981) estimator of the conditional survivor function can be written as follows

$$\hat{S}_{C|W}(y \mid w) = \prod_{i=1}^{n} \left\{ 1 - \frac{\mathbf{K}_{h_S}(W_i - w)}{\sum_{l=1}^{n} \mathbb{1}(Y_l \leq Y_i)\mathbf{K}_{h_S}(W_l - w)} \right\}^{\mathbb{1}(Y_i \leq y, \delta_i = 0)}, \quad (3.1)$$

where h_S is a bandwidth parameter. This estimator reduces to the standard Kaplan and Meier [46] estimator when the weights are all equal to n^{-1}. We provide conditions for the strong uniform consistency of this estimator in Section 4.

Further, let A be the following conditional expectation operator

$$(A\varphi)(w) = E(\varphi(Z) \mid W = w),$$

such that $A : L^2_Z \to L^2_W$, and $r(w) = E(V \mid W = w)$. Similarly, let

$$(A^*\psi)(z) = E(\psi(W) \mid Z = z),$$

such that $A^* : L^2_W \to L^2_Z$ is the adjoint of the operator A [see 24, among others]. This notation allows us to express equation (1.2) as follows

$$A\varphi = r. \tag{3.2}$$

Assumption 2.3 implies that the operator A is injective and therefore invertible. Under this condition, a unique solution to this problem exists, as shown in Proposition 2.1. However, the solution obtained by inverting the conditional expectation operator directly is not stable, and therefore we are faced with a linear ill-posed inverse problem. Heuristically, one could interpret the problem stated in equation (3.2) as a system of equations, in which the (infinite dimensional) matrix A is singular [13]. As discussed in the introduction, we explore the properties of our estimators using a Landweber-Fridman regularization approach.

The intuition underlying this regularization method is as follows. Equation (3.2) can be equivalently written as

$$A^*A\varphi = A^*r.$$

With simple algebra, one can show that the last identity also implies $cA^*r = [I - (I - cA^*A)]\varphi$, where $c \in (0, 1)$ is an arbitrary constant, which satisfies $\|cA^*A\| < 1$, with $\|\cdot\|$ being the operator norm. The solution φ thus needs to satisfy the following recursive identity

$$\varphi = cA^*r + (I - cA^*A)\varphi.$$

An exact solution for φ would be given by the infinite sum

$$\varphi = c\sum_{k=0}^{\infty}(I - cA^*A)^k A^*r. \tag{3.3}$$

A *regularized* solution is obtained by stopping this infinite sum after N terms:

$$\varphi_N = c\sum_{k=0}^{N-1}(I - cA^*A)^k A^*r. \tag{3.4}$$

Similarly, equation (3.4) can be expressed recursively as

$$\varphi_k = \varphi_{k-1} + cA^*(r - A\varphi_{k-1}), \text{ for } k = 1, \ldots, N, \tag{3.5}$$

with $\varphi_0 = 0$.

The regularized estimator of φ is obtained by replacing r, A, and A^* in equation (3.5) by consistent nonparametric estimators and using a stopping rule to determine the total number of iterations, N. Equation (3.4) is a solution of a linear optimization problem. We start iterating from $N = 1$, with $\varphi_1 = cA^*r$. For $k = 1, 2, 3, \ldots$, the iterative scheme converges towards the true solution as long as

$$\|\varphi_k - \varphi_{k-1}\| \leq c\|A^*(r - A\varphi_{k-1})\| = c\|A^*A\,(\varphi - \varphi_{k-1})\,\|$$

$$\leq c\|A^*A\|\|\varphi - \varphi_{k-1}\| < \|\varphi - \varphi_{k-1}\|.$$

Hence, we need to select c in such a way that $c\|A^*A\| < 1$. This condition implies that our iterative scheme is a contraction. Notice that $\|A^*A\| = \|A\|^2 = 1$, as A is a conditional expectation operator, and its norm is equal to 1. Therefore, any $c < 1$ would guarantee convergence of our iteration scheme. Besides this restriction, the specific choice of c does not matter for our purposes, and the solution is insensitive to it. As in Engl et al. [28, p. 155], we can rewrite equation (3.2) in a way that $A_c\varphi = \sqrt{c}r$, with $A_c = \sqrt{c}A$, and

$$\varphi_k = \varphi_{k-1} + A_c^*\left(\sqrt{c}r - A_c\varphi_{k-1}\right),$$

with $A_c^* = \sqrt{c}A^*$. This converges if $A_c^*A_c$ is a contraction. That is, as long as $\|A_c^*A_c\| < 1$, which is the same condition as above. Values of c closer to the upper bound result in larger steps and fewer iterations for convergence. By contrast, if c is close to 0, the number of iterations can be extraordinarily large and, albeit precise, reaching the solution would require greater computational time. In our numerical experiment and empirical application, we use $c = 0.5$.

Below we outline the practical implementation of our estimator.

1. We compute the kernel weighted estimator of $S_{C|W}(y \mid w)$, $\hat{S}_{C|Z}(y \mid w)$, as in equation (3.1), using local constant weights and bandwidth parameter h_S. We then construct the dependent variable

$$\hat{V}_i = \frac{\delta_i Y_i}{\hat{S}_{C|W}(Y_i \mid W_i)}.$$

2. For the estimation of $r(w) = E(V \mid W = w)$ and all the other population objects hereafter, we advocate using local polynomial regressions. While our asymptotic properties are developed using generalized kernels [see 61] to control the behavior of the estimator at the boundaries of the support, these are seldom used in practice. Local polynomial regressions are simpler to implement and do not have any boundary effects [29]. To simplify our exposition and without loss of generality, we consider local linear fitting. Let $\hat{\mathbf{V}}$ to be the $n \times 1$ vector of the generated dependent variable; $\bar{\mathbf{K}}_{W,h_W}(w)$, the $n \times n$ diagonal matrix of kernel weights at the point w,

$$\bar{\mathbf{K}}_{W,h_W}(w) = diag\left(\mathbf{K}_{h_W}\left(W_1 - w\right), \ldots, \ldots, \mathbf{K}_{h_W}\left(W_n - w\right)\right),$$

where h_W is a bandwidth parameter; and $\mathbf{W}(w)$ an $n \times 2$ matrix with i-th row equal to $(1, W_i - w)$. We write

$$\hat{r}(w) = e_1'\left(\mathbf{W}(w)'\bar{\mathbf{K}}_{W,h_W}(w)\mathbf{W}(w)\right)^{-1}\left(\mathbf{W}(w)'\bar{\mathbf{K}}_{W,h_W}(w)\hat{\mathbf{V}}\right) = \mathbf{M}(w)\hat{\mathbf{V}},$$

with $e_1' = (1, 0)$ and $\mathbf{M}(w)$ a $1 \times n$ vector.

3. Next, we estimate the two conditional expectation operators, A and A^*. Both operators are linear and can therefore be approximated by linear smoothers. To construct an estimator of A using local linear regressions,

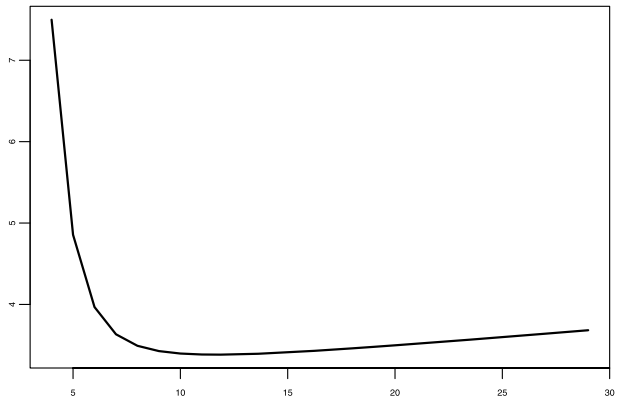


FIG 1. *Stopping function* $||\hat{r} - \hat{A}_{-1}\hat{\varphi}_{k,-1}||^2$*, where* k *is the iteration index, for one draw from the simulated DGP in Section 5,* $n = 500$*, optimal stopping iteration* $N = 12$*.*

we stack in a matrix of dimension $n \times n$, the vectors $\mathbf{M}(W_i)$, for all $i = 1, \ldots, n$, in a way that

$$\hat{A} = [\mathbf{M}(W_1)', \ldots, \mathbf{M}(W_n)']'.$$

As above, we let

$$\bar{\mathbf{K}}_{Z,h_Z}(z) = diag\left(\mathbf{K}_{h_Z}(Z_1 - z), \ldots, \ldots, \mathbf{K}_{h_Z}(Z_n - z)\right),$$

and $\mathbf{Z}(z)$ a matrix with i-th row equal to $(1, Z_i - z)$. We finally have

$$\begin{aligned}\mathbf{M}(z) =& e_1'\left(\mathbf{Z}(z)'\bar{\mathbf{K}}_{Z,h_Z}(z)\mathbf{Z}(z)\right)^{-1}\mathbf{Z}(z)'\bar{\mathbf{K}}_{Z,h_Z}(z)\\ \hat{A}^* =& [\mathbf{M}(Z_1)', \ldots, \mathbf{M}(Z_n)']'.\end{aligned}$$

4. Given estimators of r, A and A^*, we start our iteration scheme from $\hat{\varphi}_1 = c\hat{A}^*\hat{r}$. We compute each subsequent iteration as

$$\hat{\varphi}_{k+1} = \hat{\varphi}_k + c\hat{A}^*(\hat{r} - \hat{A}\hat{\varphi}_k), \text{ for } k = 1, \ldots, N-1.$$

5. To determine when to stop iterating, we adopt the cross-validation criterion developed in Centorrino [12]. We compute the leave-one-out version of $\hat{\varphi}_k$, denoted $\hat{\varphi}_{k,-1}$. Then we let

$$CV(\hat{\varphi}_{k,-1}) = ||\hat{r} - \hat{A}_{-1}\hat{\varphi}_{k,-1}||^2,$$

for $k = 1, 2, \ldots$. This function's typical shape can be observed in Figure 1 (this is the stopping function for one draw from the simulated DGP in Section 5, $n = 500$).

Equation (3.5) involves unknown density, distribution, and conditional mean functions, which are consistently estimated using locally weighted kernel approaches. We employ Gaussian kernels and select the bandwidth parameters,

$\{h_S, h_W, h_Z\}$, by Silverman's rule-of-thumb. This procedure delivers a consistent estimator of the unknown function φ.

Remark 2 (Additional Confounders)**.** *In many empirical settings, it is common to have additional observable confounders, $X \in \mathbb{R}^d$, which may be continuous or discrete and should be included in the regression model. The statistical model is*

$$T_i = \varphi(Z_i, X_i) + U_i, \text{ for } i = 1, \dots, n,$$

with $E[U_i|W_i, X_i] = 0$, and Assumptions 2.1 and 2.2 must now hold conditional on (W, X). Because of theoretical considerations that will be explained in more detail below, it is not possible to modify the definition of the operators to include the additional exogenous variables. However, as explained in Hall and Horowitz [38], we can obtain an estimator of the function for every fixed value of X_i and then smooth it with respect to it. That is,

$$\hat{\varphi}_k(z, x) = \hat{\varphi}_{k-1}(z, x) + c \sum_{i=1}^{n} \left[A^*(r - A\varphi_{k-1})\right](z, X_i)\mathbf{M}_i(x),$$

where $\mathbf{M}_i(\cdot)$ is the i-th elements of a mixed kernel projection vector [54], which depends on an additional bandwidth, h_X, and is defined as above.

This estimation strategy suffers from the well-known curse of dimensionality. As an alternative, one could consider a partially linear specification, $\varphi(Z, X) = \varphi_0(Z) + X\beta$ [1, 33]; or a varying coefficient specification $\varphi(Z, X) = \varphi_0(Z) + X\varphi_1(Z)$ [16]. The latter is desirable and naturally arises when X is purely discrete.

3.3. Case 2: U is independent of W

Estimation in the independent case proceeds similarly as above.

An estimator of the conditional cdf of the error term can be obtained using Beran's (1981) approach as in equation (3.1). An estimator of the unconditional survivor function can instead be obtained using a smoothed version of the simple Kaplan-Meier estimator.

The Landweber-Fridman estimator of $\varphi_\dagger$ is based on a recursive definition as above

$$\hat{\varphi}_{k+1} = \hat{\varphi}_k - c\hat{A}'^*_{\hat{\varphi}_k}(\hat{A}(\hat{\varphi}_k)), \tag{3.6}$$

where $k = 0, 1, 2, \dots$ is an integer, and $N > 0$ is the total number of iterations; $\hat{A}(\hat{\varphi}_k)$ is an estimator of $A(\varphi)$ computed at the point $\hat{\varphi}_k$; $\hat{A}'^*_{\hat{\varphi}_k}$ is an estimator of the adjoint operator of the Fréchet derivative, and $c < 1$ is a strictly positive constant that determines the size of the step between consecutive iterations.

An additional step for implementation is to derive a closed-form expression for A'^*_φ. Recall that A'_φ is a linear operator from L^2_Z into $L^2_{U\times W}$. Therefore, A'^*_φ is a linear operator from $L^2_{U\times W}$ into L^2_Z which ought to satisfy the following relation

$$\int\int \left[A'_\varphi(\tilde\varphi)\right](u,w)\ \psi(u,w) f_{U,W}(u,w)dudw$$
$$= \int \tilde\varphi(z)\left[A'^*_\varphi(\psi)\right](z) f_Z(z)dz \quad \forall \tilde\varphi \in L^2_Z, \quad \psi \in L^2_{U\times W},$$

with

$$\int\int \left[A'_\varphi(\tilde\varphi)\right](u,w)\ \psi(u,w) f_{U,W}(u,w)dudw$$
$$= -\int\int\int \tilde\varphi(z)\psi(u,w)\left[f_{T,Z|W}(\varphi(z)+u,z\mid w)\right.$$
$$\left. -f_{T,Z}(\varphi(z)+u,z)\right] f_U(u) f_W(w)dzdwdu.$$

From some elementary computations, we get

$$\left(A'^*_\varphi\psi\right)(z) = -\int\int \psi(u,w)\left[f_{T,Z,W}(\varphi(z)+u,z,w)\right.$$
$$\left. -f_{T,Z}(\varphi(z)+u,z) f_W(w)\right]\frac{f_U(u)}{f_Z(z)}dudw,$$

which reduces to

$$\left(A'^*_\varphi\psi\right)(z) = -E\left[\left(\psi(u,w) - E_W\psi(u,W)\right) f_U(u)\mid Z=z\right], \tag{3.7}$$

where E_W denotes the expectation taken with respect to the marginal distribution of W.

Let us now describe the practical implementation of this algorithm. In the following, we let $\bar T$ to be the estimator of the mean of T obtained by integrating the uncensored observations with respect to the empirical Kaplan-Meier distribution.

- We select an initial value φ_0. Different choices of the initial conditions are possible. We may take φ_0 equal to the nonparametric estimation of the conditional expectation of T given Z, obtained as in Dabrowska [22]. This is not a consistent estimator if Z is endogenous but in many cases, the endogeneity bias is not too strong, and $E(Y\mid Z)$ may be a reasonable starting value. Another possible choice is to solve the linear problem $E(V\mid W) = E(\varphi(Z)\mid W)$ as detailed above. If $\varphi_\dagger$ is identified under the mean independence condition, this solution is a consistent estimator, and we conjecture that imposing the independence restriction should improve the properties of this estimator. If $\varphi_\dagger$ is under-identified this estimation gives an approximation [see 4, 32]. Finally, one could use a linear or nonlinear parametric instrumental variable estimator.
- At each iteration $k \geq 0$, we compute the estimated centered residuals

$$\hat U_{ki} = Y_i - \hat\varphi_k(Z_i) - \bar T + \frac{1}{n}\sum_{i=1}^n \hat\varphi_k(Z_i), \tag{3.8}$$

where $\bar{T}$ is the sample mean of the random variable T estimated from the censored observations $Y = T \wedge C$ [66]. Notice that this location normalization of the residuals correctly identifies the location of the function φ, under the assumption that $E(U) = 0$. $\hat{A}(\hat{\varphi}_k)$ can be taken to be the difference between the conditional product-limit estimator of the distribution of U given W, and the unconditional product-limit estimator of the distribution of U. That is, we let

$$\hat{S}_{U|W}(u \mid w) = \prod_{i=1}^{n} \left\{ 1 - \frac{\mathbf{K}_{h_S}(W_i - w)}{\sum_{l=1}^{n} \mathbb{1}(\hat{U}_{kl} \leq \hat{U}_{ki}) \mathbf{K}_{h_S}(W_l - w)} \right\}^{\mathbb{1}(\hat{U}_{ki} \leq u, \delta_i = 1)}, \tag{3.9}$$

and

$$\hat{S}_U(u) = \prod_{i=1}^{n} \left\{ 1 - \frac{1}{n - i + 1} \right\}^{\mathbb{1}(\hat{U}_{ki} \leq u, \delta_i = 1)}, \tag{3.10}$$

so that finally

$$\hat{A}(\hat{\varphi}_k)(u, w) = \hat{S}_{U|W}(u \mid w) - \hat{S}_U(u). \tag{3.11}$$

If W is discrete, the conditional cdf may be computed by sorting with respect to the different (finite) values of W, allowing us to reach faster convergence rates. We provide a more detailed description of the latter case in Section 5. Finally,

$$\hat{A}'^{*}_{\hat{\varphi}_k}(\hat{A}(\hat{\varphi}_k)) = -\frac{\sum_{i=1}^{n} r(\hat{U}_{ki}, W_i) \hat{f}_{\hat{U}_k}(\hat{U}_{ki}) \mathbf{K}_{h_Z}(Z_i - z, z)}{\sum_{i=1}^{n} \mathbf{K}_{h_Z}(Z_i - z, z)}, \tag{3.12}$$

where

$$r(\hat{U}_{ki}, W_i) = \hat{A}(\hat{\varphi}_k)(\hat{U}_{ki}, W_i) - \frac{1}{n} \sum_{i=1}^{n} \hat{A}(\hat{\varphi}_k)(\hat{U}_{ji}, W_i),$$

with tuning parameter h_Z, and with $\hat{f}_{\hat{U}_k}$ being a nonparametric density of the residuals at iteration k, whose construction is discussed in more detail below. Bandwidth parameters are chosen by Silverman's rule-of-thumb. Finally, we take $c = 0.5$ as discussed above.

- An important component in the construction of the estimator of the adjoint operator is $\hat{f}_{\hat{U}_k}(\hat{U}_{ki})$, the nonparametric estimator of the density of the error term. As our observations are right-censored, we follow the approach in Marron and Padgett [57], and Mielniczuk [59], and use the following estimator

$$\hat{f}_{\hat{U}_k}(\hat{U}_{ki}) = \frac{1}{n h_U} \sum_{i=1}^{n} \mathbf{K}_{h_U}(U_i - u, u) \, \Delta \hat{F}_{\hat{U}_k}(U_i),$$

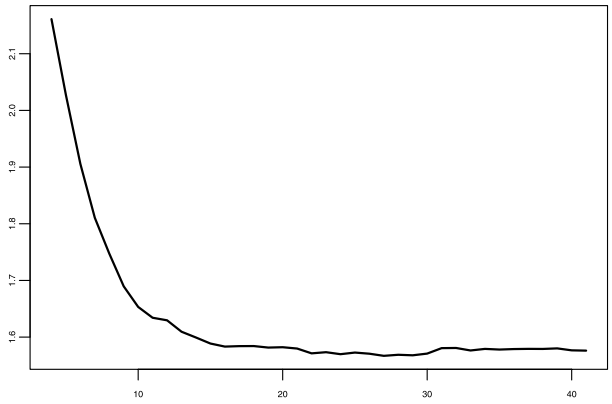


FIG 2. *Stopping function* $||\hat{A}(\hat{\varphi}_k)||^2$*, where* k *is the iteration index, for one draw from the simulated DGP in Section 5,* $n = 500$*, optimal stopping iteration* $N = 27$*.*

where $\hat{F}_{\hat{U}_k}$ is the Kaplan-Meier estimator of the distribution of U in iteration k, and $\Delta\hat{F}_{\hat{U}_k}(U_i)$ are its finite differences. The rate of convergence of this estimator is the same as the usual nonparametric density estimator under standard assumptions.

- The last point is the choice of the stopping rule. This choice is crucial, as the regularization of the ill-posed inverse problem is provided by the stopping rule. It is common in the mathematical literature to adopt the so-called Morozov's discrepancy principle [see 8, 45, 60]. This principle leads to iterate up to $N_0 > 0$, such that

$$\|\hat{A}(\hat{\varphi}_{N_0-1})\|^2 > \tau\delta \geq \|\hat{A}(\hat{\varphi}_{N_0})\|^2, \tag{3.13}$$

where δ is a noise that is usually known, and τ is a positive constant, which depends on the properties of the known operator A. In this problem, however, we have an additional estimation error because of a nonparametrically generated regressor [56], which may blow our variance further as $N \to \infty$. We, therefore, proceed as follows. We fix a maximum number of iterations, $N_{\max}$, based on the asymptotic theory derived below. We then check the norm of $\hat{A}(\hat{\varphi}_k)$, at each iteration $j = 0, 1, 2, \ldots, N_{\max}$, and take N_0 as the iteration where the norm reaches its minimum. Otherwise, we take $N_0 = N_{\max}$. The typical shape of this function can be seen in Figure 2 (this is the stopping function for one draw from the simulated DGP in Section 5, $n = 500$).

Remark 3 (Additional Confounders)**.** *Adding additional confounders,* $X \in \mathbb{R}^d$*, to the model with independent instruments requires more careful consideration of the underlying independence assumptions. One could potentially assume that* $U \perp\!\!\!\perp W|X$*, with* $E[U|X] = 0$*, in a way that still allows for arbitrary het-*

eroskedasticity of U wrt X. The identifying restriction thus becomes

$$[A(\varphi_\dagger)](u,w,x) = \int [F(\varphi_\dagger(z,x)+u,z \mid x,w) - F(\varphi_\dagger(z,x)+u,z \mid x)]\, dz = 0,$$

with the adjoint operator of the Fréchet derivative written as

$$\left(A'^{*}_{\varphi}\psi\right)(z,x) = -E\left[\left(\psi(u,w,x) - E_W\psi(u,W,x)\right) f_{U|X}(u|x) \mid Z=z, X=x\right].$$

The final estimator can be written as

$$\hat{\varphi}_{k+1}(z,x) = \hat{\varphi}_k(z,x) - c\sum_{i=1}^{n}\left[\hat{A}'^{*}_{\hat{\varphi}_k}(\hat{A}(\hat{\varphi}_k))\right](z,X_i)\mathbf{M}_i(x),$$

where $\mathbf{M}_i(\cdot)$ is defined as in Remark 2. This approach requires an estimator of the conditional density of U given X at each iteration, which suffers from the curse of dimensionality and may result in very slow rates of convergence. A potential alternative is to use the more restrictive assumption that $U \perp\!\!\!\perp (X,W)$ and $E(U)=0$, together with a flexible semi-parametric structure.

4. Rates of convergence

4.1. Framework

We briefly give the main result about the rate of convergence of our estimators. Our proofs are based on results by Engl et al. [28, Section 6.1], Carrasco et al. [11, Section 3], Johannes et al. [44], Florens et al. [34], Dunker et al. [27], Dunker [26], and Centorrino et al. [14].

We start by collecting assumptions that are common across the various frameworks. To clarify our notations, we let

$$K_h(hu,t) = \begin{cases} K_+(u,1) & \text{if } h \le t \le 1-h \\ K_+\left(u,\frac{t}{h}\right) & \text{if } 0 \le t \le h \\ K_-\left(u,\frac{1-t}{h}\right) & \text{if } 1-h \le t \le 1 \\ 0 & \text{otherwise} \end{cases},$$

to be a generalized kernel with correction at the endpoints as defined in Müller [61], where $K_+(\cdot,t)$ and $K_-(\cdot,t)$ are functions supported on $[-1,t]\times[0,1]$ and $[-t,1]\times[0,1]$, respectively, and $K_+(u,1) = K_-(u,1) = K(u)$, where K is a standard kernel function.

Assumption 4.1. *The univariate generalized kernel function $K_h(\cdot,\cdot)$ satisfies the following properties:*

(i) It has order $\ell \ge 2$.

(ii) For each $t\in[0,1]$, $K_h(h\cdot,t)$ is supported on $[(t-1)/h, t/h]\cap\mathcal{K}$, where $\mathcal{K}$ is a compact interval that does not depend on t and:

$$\sup_{h>0,t\in[0,1],u\in\mathcal{K}} \mid K_h(hu,t) \mid < \infty$$

(iii) $K_+(\cdot,t)$ *and* $K_-(\cdot,t)$ *are Lipschitz continuous.*
(iv) $K_h(h\cdot,1) = K(\cdot)$ *is a Lipschitz continuous and symmetric kernel function with compact support.*

4.2. Case 1: U mean independent of W

For this section, we let

$$\hat{r}(w) = \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)\,\hat{V}_i}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)},$$
$$\hat{A} = \left(\frac{\mathbf{K}_{h_W}(W_i - W_l, W_l)}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - W_l, W_l)}\right)_{i,l=1}^n$$
$$\hat{A}^* = \left(\frac{\mathbf{K}_{h_Z}(Z_i - Z_l, Z_l)}{\sum_{i=1}^n \mathbf{K}_{h_Z}(Z_i - Z_l, Z_l)}\right)_{i,l=1}^n$$

where $\mathbf{K}_{h_W}(\cdot,\cdot)$ is a multivariate generalized kernel function as defined above, and $\hat{V}_i = \delta_i Y_i / \hat{S}_{C|W}(Y_i \mid W_i)$. Estimators of the operators A and A^* are constructed as $n \times n$ matrices of kernel weights.

We also need the following additional assumptions.

Assumption 4.2. *The random vector* (T, C, Z, W) *is characterized by its joint distribution* F*, which is absolutely continuous with respect to the Lebesgue measure.*

Assumption 4.3.

(i) The joint density $f_{ZW}(z,w)$ *is* $\lambda \geq 2$ *times differentiable and uniformly bounded away from* 0 *and* ∞*.*
(ii) The joint and the marginal densities of (Z,W) *satisfy*

$$\int\int \left[\frac{f_{ZW}(z,w)}{f_Z(z) f_W(w)}\right]^2 f_Z(z) f_W(w) < \infty.$$

Assumption 4.4. *The conditional mean* $E(V \mid W)$ *is at least* $\rho \geq 2$ *times differentiable with respect to both its arguments and the conditional variance of* V *given* W *is uniformly bounded on* $[0,1]^q$*.*

Assumption 4.5. *The smoothing parameters satisfy* $h_S, h_W, h_Z \to 0$*,* $(nh_S^q)^{-1}\log n \to 0$ *and* $(nh_Z^p h_W^q)^{-1}\log n \to 0$*.*

Assumption 4.2 requires the random vector (T, C, Z, W) to have continuous density. Our identification results do not hold if the censoring variable does not have sufficient variation. This implies that the joint distribution of (Y, Z, W) is also continuous, which is a standard condition invoked in the literature on nonparametric instrumental regressions [see 24, 41]. Assumption 4.3(i) imposes smoothness restrictions on the joint density of the random vector (Z, W). The conditions in Assumption 4.3(ii)-(iii) imply that A and A^* are compact and

injective. In particular, Condition (ii) entails that A and A^* admit a singular value decomposition, with their singular values having zero as a limit point. This property generates the *ill-posedness* of the inverse problem defined by equation (3.2). We distinguish two cases: when the singular values of A^*A converge to zero at a polynomial rate, we say that the inverse problem is *mildly ill-posed*, while if they have an exponential rate of convergence, we say that the problem is *severely ill-posed*. The degree of ill-posedness is related to the smoothness of the joint density of (Z, W). In practice, the smoother the joint density is, the more the function φ is *blurred* when integrated with respect to it, and the more difficult the estimation problem becomes. The following example shows how the decay of the singular values of A and A^* is related to the joint distribution of (Z, W).

Example 2 (The Normal Case)**.** *Suppose that $(Z, W) \in \mathbb{R}^2$ is jointly normal with mean zero and covariance matrix given by:*

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

with $\mid \rho \mid < 1$. This implies that the conditional distribution of Z given $W = w$ is normal with a mean equal to ρw and a variance equal to $1 - \rho^2$. Therefore, the eigenvectors associated to the operator A are Hermite polynomials, and its singular values are given by ρ^j, for $j = 0, 1, 2, \ldots$. As $j \to \infty$, the eigenvalues are converging to zero at an exponential rate. In this jointly normal case, the inverse problem is therefore severely ill-posed.

Remark 4. *The conditions in Assumptions 4.2 and 4.3(ii) do not hold when Z and W have elements in common, i.e., there are other observed confounders X included in the regression model. In this case, one can proceed as discussed in Remarks 2 and 3 above. The conditions in Assumptions 4.2 and 4.3(ii) will then have to hold for the densities conditional on $X = x$.*

Assumption 4.4 is a smoothness condition on the conditional expectation of V given W, and the second part is tantamount to the requirement that V is square-integrable.

Finally, Assumptions 4.5, along with the conditions on the kernel function in Assumption 4.1 and the differentiability conditions in Assumption 4.3(i) are used to show the uniform consistency of the nonparametric estimators of the joint and marginal densities of (Z, W), and of the conditional survival function $S_{C|W}$ [22, 68].[1]

We obtain the following.

Proposition 4.1. *Under Assumptions 4.1-4.5, when $\ell \geq \rho$, and $\ell \geq \lambda$, the following holds*

[1] Gonzalez-Manteiga and Cadarso-Suarez [37] also provide uniform consistency results for Beran's generalization of the Kaplan-Meier estimator. However, their results rely on the additional condition that $nh_n^{2\rho+q} \to 0$, which, as explained in 68, excludes the optimal choice of bandwidth.

(i) Estimation of r.

$$||\hat{r} - \hat{A}\varphi||^2 = O_P\left(\frac{1}{nh_W^q} + \frac{1}{nh_S^q} + h_W^{2\rho} + h_S^{2\rho}\right),$$

which implies $||\hat{r} - \hat{A}\varphi||^2 = O_P(n^{-2\rho/(2\rho+q)})$, provided $h_S = O(h_W)$, and $h_W = O(n^{-1/(2\rho+q)})$;

(ii) Estimation of the operator A:

$$||\hat{A} - A||^2 \sim ||\hat{A}^* - A^*||^2 = O_P\left(n^{-\frac{2\lambda}{2\lambda+p+q}}\right).$$

□

The first part of the proposition gives the rate of convergence for the nonparametric estimator $\hat{r}$. The dependent variable is estimated using standard kernel regressions, so that a projection argument, which is standard in this literature, makes us conclude that the first step estimation of the conditional survivor function is negligible, provided the bandwidth h_S is chosen accordingly. Notice that this choice of bandwidth allows us to achieve the same rate for the estimation of $E(T|W)$ as if T was fully observed, and therefore map the results that follow into the class of *standard* nonparametric IV estimators. We argue that such a requirement is easily satisfied by taking $h_S = h_W$. That is, we use the same bandwidth to estimate the conditional survivor function and the conditional expectation. A proof of the first part of the Proposition is given in Appendix, under additional assumptions on the asymptotic representation of $\hat{S}_{C|W}$.

The second part of the Proposition follows from the results in Darolles et al. [24]. The rate of convergence for the estimation of the operators is standard in nonparametric econometrics. As the operators are effectively estimators of conditional densities, their rates of convergence are those of the nonparametric estimator of the joint density of (Z, W).

Denote by $\mathcal{R}$ the range of an operator. We present the main convergence rate in the following Theorem.

Theorem 4.1. *Under the assumptions and result of Proposition 4.1, we have*

(i) Either the following strong source condition holds

$$\exists, \beta > 0 \text{ such that } \varphi \in \mathcal{R}(A^*A)^{\frac{\beta}{2}}, \tag{4.1}$$

and

$$||\hat{\varphi}^N - \varphi||^2 = O_P\left(Nn^{-\frac{2\rho}{2\rho+q}} + n^{-\frac{2\lambda}{2\lambda+p+q}}N^{1-\beta} + N^{-\beta}\right).$$

If $\frac{\lambda}{2\lambda+p+q} \geq \frac{\rho}{(2\rho+q)(\beta+1)}$, this value is minimized for $N_0 \asymp \left(n^{\frac{2\rho}{2\rho+q}}\right)^{\frac{1}{\beta+1}}$ and

$$||\hat{\varphi}^{N_0} - \varphi||^2 = O_P\left(n^{-\frac{2\rho\beta}{(2\rho+q)(\beta+1)}}\right).$$

(ii) or the following weak source condition holds

$$\exists\, \beta > 0 \text{ such that } \varphi \in \mathcal{R}(-\log(A^*A))^{-\frac{\beta}{2}} \tag{4.2}$$

and

$$||\hat{\varphi}^N - \varphi||^2 = O_P\left(Nn^{-\frac{2\rho}{2\rho+q}} + n^{-\frac{2\lambda}{2\lambda+p+q}}N(\log(N)^{-\beta}) + (\log(N))^{-\beta}\right).$$

For a small constant $\epsilon > 0$ *such that,* $\epsilon < \frac{2\rho}{2\rho+q}$ *and* $\frac{\lambda}{2\lambda+p+q} \geq \frac{\rho}{2\rho+q} - \frac{\epsilon}{2}$, *this value is minimized for* $N_0 \asymp n^{\frac{2\rho}{2\rho+q}-\epsilon}$ *and*

$$||\hat{\varphi}^{N_0} - \varphi||^2 = O_P\left((\log(n))^{-\beta}\right). \quad \square$$

The strong and weak source conditions link the smoothness properties of the conditional expectation operators, A and A^* with the smoothness properties of the function φ. The first part of the theorem establishes the convergence rates for the mildly ill-posed problem, which are polynomial in the sample size. In the second part, we instead provide the rates for the severely ill-posed case, which are instead polynomial in the logarithm of the sample size. This is a standard result in the literature, as the super-smoothness of the joint density implies that the data contain very little information about the function φ, and a large sample is required to obtain a precise estimate [see 20].

For the mildly ill-posed problem, we note that our rate is the rate of estimation of $E(V \mid W)$ at a power $\frac{\beta}{\beta+1}$ smaller than 1 which we may view as the cost of the resolution of the inverse problem. Note that one advantage of the Landweber-Fridman method is that β is not constrained by the qualification of the method, such as in the Tikhonov regularization where β is limited by 2 (see the Appendix for a formal definition).

Remark 5 (Minimax rates)**.** *When the inverse problem is mildly ill-posed, the eigenvalues of the operator* A^*A *decay geometrically at a speed equal to* $2a$, *with* $a > 0$. *If* $s > 0$ *is the smoothness of the function* φ, *then* $\rho = s + a$, *and* $\beta = s/a$. *Therefore, our rates of convergence would be equal to*

$$||\hat{\varphi}^{N_0} - \varphi||^2 = O_P\left(n^{-\frac{2s}{2s+2a+q}}\right),$$

where q *is the dimension of* W. *If* $p = q$, *that is, we have as many instruments as endogenous variables, then this rate is minimax [18, 20].*

When the inverse problem is severely ill-posed, the rate of convergence is dominated by the bias term, which converges at a logarithmic rate. The rate is minimax for the given choice of the tuning parameters.

4.3. Case2: U independent of W

For nonlinear inverse problems, iteration methods like the one used here would in general not converge globally. We prove local convergence by appropriately restricting the initial condition and controlling the behavior of the Fréchet derivative of the operator $\hat{A}$. We assume the following.

Assumption 4.6. *Let* $\mathcal{B}_R(\varphi_0)$ *to be a ball of radius* $R < \infty$ *around the initial condition, such that* $\mathcal{B}_R(\varphi_0) \subset \mathcal{D}(A)$. *We have*

$$\varphi_\dagger \in \mathcal{B}_R(\varphi_0).$$

Assumption 4.7. *A and $\hat{A}$ are Fréchet differentiable, with A' and $\hat{A}'$ bounded linear operators.*

We also impose the following additional Assumptions.

Assumption 4.8.

(i) The conditional probability density function $f_{T,Z|W}(t,z \mid w)$ and the density function $f_{T,Z}(t,z)$ are $\lambda \geq 2$ times continuously differentiable and uniformly bounded away from ∞.

(ii) The densities $f_{ZW}(z,w)$, $f_Z(z)$ and $f_W(w)$ are uniformly bounded away from 0 and ∞.

Assumption 4.9. *The density of the error term $f_U(u)$ is absolutely continuous with respect to the Lebesgue measure, and $\lambda \geq 2$ times continuously differentiable.*

Assumption 4.10. *Let ϖ_n a real sequence that is either bounded or diverges slowly to ∞ with n. The density of the error term, U, satisfies*

$$\kappa_n = \inf_{|u| \leq \varpi_n} f_U(u) > 0,$$

with $\kappa_n \to 0$, as $n \to \infty$.

Assumption 4.11. *The smoothing parameters satisfy $h_U, h_W, h_Z \to 0$, and $(nh_Z^p h_W^q h_U)^{-1} \ln n \to 0$.*

Assumption 4.8 is a standard regularity condition of conditional and unconditional densities. Part (ii) is not restrictive as long as we maintain that the joint support of (Z,W) is compact. Assumption 4.9 restricts the density of the error term to be continuous and differentiable. Finally, Assumptions 4.10 and 4.11 are used for the uniform consistency of the nonparametric density estimators. One crucial difference of this estimator compared to the one we have previously described is that it involves estimating the density of the error term at each iteration. While it is plausible to assume that the support of the independent variables is bounded, such an assumption would be too restrictive for the error component U. Assumption 4.10 helps us accommodate possibly unbounded support of the error, following the approach of Hansen [40]. We choose the points u in expanding sets of the form $\{u :| u |\leq \varpi_n\}$.

We let the following hold.

Assumption 4.12. *We have that*

$$\begin{aligned} E\|\hat{A}(\varphi) - A(\varphi)\|^2 =& O(\delta_n^2(\ell,\lambda,q)) \\ E\|\hat{A}'^*_\varphi - A'^*_\varphi\|^2 =& O(\gamma_n^2(\ell,\lambda,p,q)), \end{aligned}$$

with $\delta_n(\ell,\lambda,q), \gamma_n(\ell,\lambda,p,q) \to 0$, as $n \to \infty$.

In the following, to simplify notations, we shall remove the dependence of δ_n and γ_n from the parameters $\{\ell,\lambda,p,q\}$. We leave the values of δ_n and γ_n unspecified as they depend on the nature of the instrumental variable. For instance, if

the instruments are binary, and their dimension q is relatively small compared to the sample size, one can sort the sample in a way to obtain $\delta_n \asymp n^{-1/2}$. On the contrary, if $W \in \mathbb{R}^q$ is continuous, $\ell \geq \lambda$ and one uses a standard nonparametric estimator for conditional distribution as in Li and Racine [55], then we have that $\delta_n \asymp n^{-\lambda/(2\lambda+q)}$. This high-level assumption holds under Assumptions 2.1 and 2.2, and further regularity conditions similar to the ones provided in Assumptions 4.1 and 4.8-4.11.

Finally, we need to further restrict the local behavior of the Fréchet derivative, its adjoint, and their estimators. In practice, this is done by extending Assumption 2.6 to the estimators of the Fréchet derivative and its adjoint. This is presented in more detail in Appendix.

We also make two additional assumptions.

Assumption 4.13 (Strong source condition)**.**

$$\exists, \beta > 0 \text{ such that } \varphi_0 - \varphi_\dagger \in \mathcal{R}(A'^*_{\varphi_\dagger} A'_{\varphi_\dagger})^{\frac{\beta}{2}},$$

Assumption 4.14 (Tuning parameters)**.** *The tuning parameters satisfy the following restrictions*

(i)

$$\frac{(\delta_n \vee \gamma_n)}{h_U^2 \kappa_n \sqrt{N^{-1}\ln(N)}} = O(1).$$

(ii) For $\beta \leq 1/2$,

$$\begin{cases} (h_u^2\kappa_n)^{-1} N^{-\beta/2} = O(1) & \text{if } \beta < 1/2 \\ (h_u^2\kappa_n)^{-1} N^{1/4}\ln(N) = O(1) & \text{if } \beta = 1/2 \end{cases}.$$

(iii) There exists $\beta^* \in (1/2, 1)$*, such that*

$$\frac{N^{(\beta^*-1)/2}}{h_u^2\kappa_n} = O(1).$$

Assumption 4.13 is a source condition. Differently from the statement of Theorem 4.1, the source condition is not assumed on the function $\varphi_\dagger$ directly, but rather on the difference between our initial condition and the true solution. This is due to the local nature of our estimation procedure. Similarly, when the inverse problem is nonlinear, we cannot allow for a weak source condition. This is because the error accumulates across iterations at a polynomial rate. Therefore, when the regularization bias only decreases at a logarithmic rate, the Landweber-Fridman algorithm cannot converge. Assumption 4.14 imposes restrictions on the tuning parameters. All restrictions depend on the unknown regularity of the ill-posed inverse problem which is determined by β. Proposing a data-driven procedure for the choice of these parameters is an essential step to be pursued in future research.

The following Theorem contains the main result of this Section.

Theorem 4.2. *Let Assumptions 2.1-2.2, 2.4-2.6, 4.1, and 4.6-4.14 hold. Then*

$$\|\hat{\varphi}_N - \varphi_\dagger\|^2 = O_P\left(N\delta_n^2 + \gamma_n^2 N^{1-\beta} + D^2(N)\right),$$

where,

$$D(N) = \begin{cases} N^{-\beta/2} & \text{for } \beta \leq \beta^* \\ (h_u^2\kappa_n)^{-1}N^{-\beta/2} & \text{otherwise} \end{cases}.$$

Otherwise, if Assumptions 4.14(i) and 4.14(iii) do not hold, and we only have $(\delta_n \vee \gamma_n)\sqrt{N\ln(N)} = O(1)$*, then*

$$\|\hat{\varphi}_N - \varphi_\dagger\|^2 = O_P\left(\frac{1}{h_u^4\kappa_n^2}\left(N\delta_n^2 + \gamma_n^2 N^{1-\beta} + N^{-\beta}\right)\right),$$

The result of this Theorem gives an upper bound on the mean square error of our estimator.

For $\beta \leq \beta^*$, as defined in Assumption 4.14(iii), the upper bound is the same one we get in Theorem 4.1(i), under a strong source condition. However, for $\beta > \beta^*$, we cannot reach the same upper bound. Heuristically, we have an additional term, $(h_U^2\kappa_n)^{-1}$, due to the estimation of the density of the error term. When $\beta = 1$, the regularization bias that accumulates across iterations converges to zero exactly as $1/N$, and thus the term $1/(Nh_U^4\kappa_n^2)$ dominates. The same effect holds for any β close enough to 1, or, more precisely, for any $\beta > \beta^*$.

The same heuristic does not apply to the other terms in the decomposition when we can choose h_U large enough and N small enough so that the nonlinearity error does not dominate. The condition in Assumption 4.14(i) on the tuning parameters serves exactly this purpose.

The last statement in the Theorem applies if we cannot choose $N \to \infty$ slow enough to satisfy the conditions in Assumptions 4.14(i) and 4.14(iii). In this case, N satisfies

$$\sqrt{N}\delta_n \vee \gamma_n N^{-\frac{\beta-1}{2}} \asymp N^{-\beta/2},$$

which would be equivalent to the optimal choice of the regularization parameter for the linear ill-posed inverse problem in Theorem 4.1. However, in this case, the rate of convergence is slower because of the additional term $(h_u^4\kappa_n^2)^{-1}$.

A potential way to let h_U go to zero more slowly is to use higher-order kernels, which is what we advocate in practice.

Example 3. *Let us consider the case in which both Z and W are continuous and scalar. We further take kernels of order $\ell \geq \lambda \wedge 2$. In this case, $\hat{A}(\varphi)$ is an estimator of the conditional cdf of U given the instrument W, so that one could take*

$$\delta_n^2 \asymp (h_W n)^{-1} + h_W^{2\lambda}.$$

Let $h_W \asymp n^{-\frac{1}{2\lambda+1}}$, in a way that $\delta_n \asymp n^{-\frac{\lambda}{2\lambda+1}}$. Similarly, $\hat{A}'_\varphi$ is a conditional expectation operator, and

$$\gamma_n^2 \asymp (h_Z^2 n)^{-1} + h_Z^{2\lambda} \asymp n^{-\frac{2\lambda}{2\lambda+2}},$$

where the last equivalence follows by taking $h_Z \asymp n^{-\frac{1}{2\lambda+2}}$. *Thus,* $\delta_n \vee \gamma_n = \gamma_n$.

Let us take $h_U \asymp n^{-1/(2\lambda+1)}$. *The condition on the growth of the number of iterations becomes*

$$\kappa_n^{-1} n^{-\left(\frac{\lambda}{2\lambda+2} - \frac{2}{2\lambda+1}\right)} \sqrt{N \ln(N)} = O(1),$$

which requires $\lambda > 2$, *with* $N \to \infty$. *The result of Theorem 4.2 finally implies*

$$\|\hat{\varphi}_N - \varphi_\dagger\|^2 = O_P\left(N n^{-\frac{2\lambda}{2\lambda+1}} + n^{-\frac{2\lambda}{2\lambda+3}} N^{\frac{1-\beta}{2}} + D^2(N)\right),$$

where the unknown value of β *determines the optimal convergence of the regularization constant,* N.

Having established the rate of convergence of the proposed estimator, we now turn to an assessment of its finite-sample performance.

5. Finite-sample behavior

We consider a Monte Carlo simulation based on the framework of Darolles et al. [24] and Florens et al. [34]. The data generating process is as follows

$$T = \varphi(Z) + U,$$

where the function $\varphi(z)$ is taken to be equal to $\varphi_1(z) = -(2z-1)^2$ and $\varphi_2(z) = -1.75 \exp(- \mid 2z - 1 \mid)$, respectively.

We generate a bivariate instrumental variable $W = (W_1, W_2)$ from a truncated normal distribution in $[-1, 1]^2$, with covariance matrix equal to

$$\begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}.$$

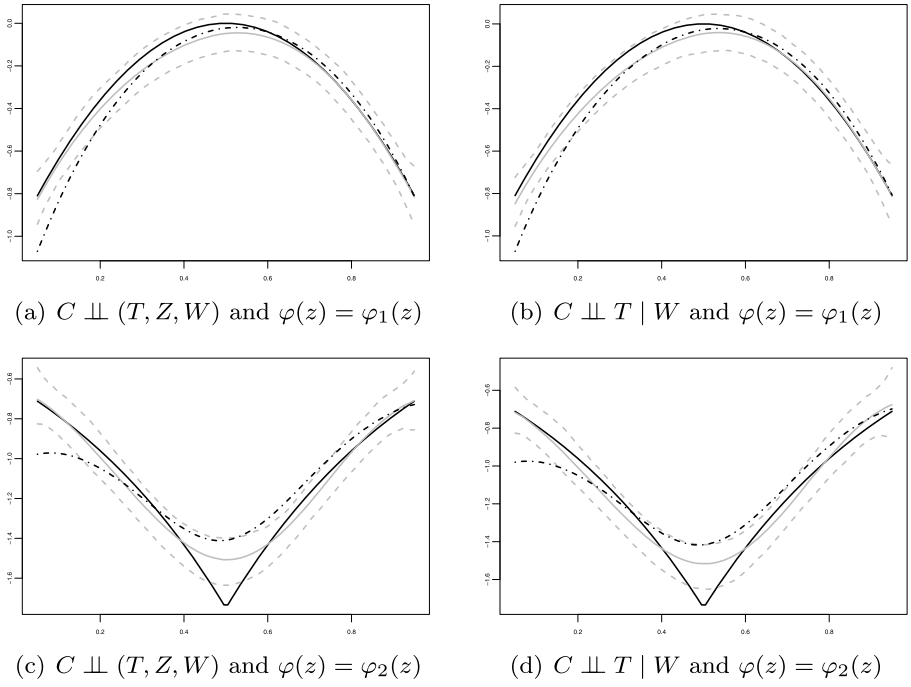
We then let

$$\begin{aligned} Z &= \frac{1}{1 + \exp\left(2(W_1 + W_2) + (W_1 + W_2) * \zeta + \zeta\right)}, \\ U &= -\left(\zeta - 0.1\right) + \varepsilon, \end{aligned}$$

where $\zeta \sim N(0.1, 0.4^2)$, and $\varepsilon \sim N(0, 0.25^2)$. This generates dependence between U and Z in such a way that $E(U \mid Z) \neq 0$, while obviously $E(U \mid W) = 0$, as W is taken to be independent of U in this example.

We consider two separate scenarios for the censoring variable C. In the first case, we take C to be independent of all other variables in the model. We generate C from a normal distribution with mean equal to the $90th$ percentile of T, and variance equal to the variance of T. In the second case, we simulate $\nu \sim N(\mu_\nu, 0.25^2)$, where μ_ν is twice the $90th$ percentile of T. We thus have that $\nu \perp\!\!\!\perp (T, Z, W)$, and we take $C = Z\nu$.

For each simulated DGP, about 20% of the observations is censored.



(a) $C \perp\!\!\!\perp (T, Z, W)$ and $\varphi(z) = \varphi_1(z)$ (b) $C \perp\!\!\!\perp T \mid W$ and $\varphi(z) = \varphi_1(z)$

(c) $C \perp\!\!\!\perp (T, Z, W)$ and $\varphi(z) = \varphi_2(z)$ (d) $C \perp\!\!\!\perp T \mid W$ and $\varphi(z) = \varphi_2(z)$

FIG 3. *Nonparametric IV with* $E[U|W] = 0$ *(solid gray line) vs nonparametric regression with* $E[U|Z] = 0$ *(dashed black line) for* $M = 1000$ *Monte Carlo replications and* $n = 500$ *observations. The solid black line indicates the true function and the dashed gray line are the* 95% *simulated confidence intervals for the Nonparametric IV estimator.*

We first look at estimation under the mean independence condition $E(U \mid W) = 0$. Figure 3 plots the median of the estimated function (solid gray line), and a simulated 95% confidence interval (dashed gray line) for $M = 1000$ simulated samples of size $n = 500$ drawn from these DGPs. The black dashed-dotted line is the nonparametric regression estimator under the assumption of mean independence [i.e., $E(U \mid Z) = 0$, see 22, 23]; and the solid black line is the true regression function. We can notice how the simple nonparametric regression estimator is never fully contained in the 95% confidence bands, and it can highly distort marginal effects at every point of the support of Z. Numerical comparison of the Mean Integrated Squared Error (MISE) of the simple nonparametric regression against our nonparametric instrumental variable estimator confirms the graphical results (see Table 1). We point out that, as the sample size increases, the relative improvement of the MISE becomes larger. This has to be expected, as the simple nonparametric regression is inconsistent in this setting.

Finally, we consider $M = 1000$ Monte Carlo replications for increasing sample sizes $n = \{250, 500, 1000\}$. For each replication, we use the nonparametric estimator outlined above to estimate $\varphi(z)$. The stopping rule used is the one described above. The constant for the Landweber-Fridman iteration was set at

TABLE 1
MISE of the nonparametric regression estimator with $E[U|Z] = 0$ relative to the instrumental variable estimator with $E[U|W] = 0$.

	$C \perp\!\!\!\perp (T, Z, W)$		$C \perp\!\!\!\perp T \mid W$	
	$\varphi_1(z)$	$\varphi_2(z)$	$\varphi_1(z)$	$\varphi_2(z)$
250	1.906	2.153	1.919	2.061
500	2.455	2.626	2.400	2.526
1000	3.408	3.172	3.217	3.125

0.5. All bandwidths for the conditional mean objects were selected via Silverman's rule-of-thumb. We consider local linear estimators of conditional means and operators, as described in Section 3, and we provide the MISE of $\hat{\varphi}(z)$ with respect to the unfeasible estimator $\tilde{\varphi}(z)$ for each replication. The unfeasible estimator is defined as the estimator of φ that would be obtained if T was observed without censoring. We report summary results in Table 2, along with the median number of iterations N required for convergence.

We can observe from Table 2 that the feasible estimator has reasonable properties compared to the unfeasible one. In most cases, the ratio between the median MISE tends to 1 as n increases, and as predicted by our theoretical results. As we take a larger proportion of censored observations, we can expect our estimator to approach more slowly the unfeasible one.

TABLE 2
MISE relative to the unfeasible estimator with $E[U|W] = 0$ and median number of iterations, N.

	$C \perp\!\!\!\perp (T, Z, W)$				$C \perp\!\!\!\perp T \mid W$			
	$\varphi_1(z)$		$\varphi_2(z)$		$\varphi_1(z)$		$\varphi_2(z)$	
n	$MISE$	N	$MISE$	N	$MISE$	N	$MISE$	N
250	1.53	8	1.09	11	1.67	8	1.13	11
500	1.64	10	1.06	15	1.90	10	1.08	14
1000	1.64	12	1.05	21	2.01	12	1.05	20

We also consider the same DGP when estimation is carried using an independence restriction. We keep the same data generating process as above. As we have noticed, our continuous instruments satisfy the restriction of independence, so that our estimator can be implemented using this stronger restriction as described in Section 3.

We can also directly compare the performance of the estimator under mean independence and independence. The latter restriction carries more information about the data generating process. However, rates of convergence can be slower due to the nonlinear ill-posed inverse problem. Moreover, as instruments are continuous, the linear estimator under mean independence is consistent. Finally, we conjecture that taking only one iteration from the mean independence estimator towards independence should be sufficient to achieve the smallest MISE. This is in parallel with the scoring method in Maximum likelihood estimation [63, 65]. In that approach, any consistent estimator of the unknown parameter reaches the efficiency bound by taking a one-step deviation towards the Maximum Like-

lihood estimator.

We can therefore assess a) how sensitive the performance of the estimation procedure is to various choices of the initial condition; b) test if our conjecture holds, at least in a limited simulation setting.

The results of this comparison are reported in Table 3 for various sample sizes.

The table is divided into four sub-tables. The first one refers to the estimator under mean independence. The second refers to the estimator under independence when the initial value is taken to be the local linear estimator of the conditional expectation of Y given Z, $\hat{\varphi}_{LL}$. The third sub-table considers the performance when the initial condition is the estimator under mean independence, $\hat{\varphi}_{MI}$. Finally, the last sub-table considers the performance of the independence estimator when we simply take a one-step deviation from $\hat{\varphi}_{MI}$. For each sample size and type of simulation, we report the MISE and the median number of iterations N performed. For the latter estimator, the median number of iterations is always equal to 1, and it is therefore not reported. The median mean square error is multiplied by a factor of 100, for the convenience of the reader.

As expected, the performances of all estimators improve as the sample size increases. The properties of the estimators under mean independence are better than those of the estimator under independence. This may also be due to a choice of tuning parameters that is not optimal in the latter case. This would require further research that is beyond the scope of this work.

There does not appear to be a substantial difference in the properties of the estimator under independence when we take different initial conditions. The median mean square error does not change dramatically, and neither does the median number of iterations.

Finally, we find some evidence in favor of our conjecture, at least in our simulation study. That is, taking a one-step iteration using the independence restriction has, in some cases, better performance than taking multiple iterations.

Finally, we consider a Monte-Carlo simulation in which we replace the two continuous instruments with a single binary instrument, W, generated from a Bernoulli distribution with parameter equal to 0.5.

We then independently generate a normal random variable, $\varepsilon \sim N\left(0, 0.25^2\right)$, and a uniform random variable, ω. We then let

$$\zeta = \log\left(\frac{\omega}{1-\omega}\right),$$

in a way that ζ follows a standard logistic distribution. Furthermore

$$\begin{aligned} Z =& \frac{1}{1+\exp\left(-\frac{(-0.5W+\zeta+\zeta W)}{2}\right)}, \\ U =& -0.4\zeta + \varepsilon, \end{aligned}$$

TABLE 3
MISE and median number of iterations, N, with $U \perp\!\!\!\perp W$ and W continuous.

		$C \perp\!\!\!\perp (T, Z, W)$				$C \perp\!\!\!\perp T \mid W$			
		$\varphi_1(z)$		$\varphi_2(z)$		$\varphi_1(z)$		$\varphi_2(z)$	
	n	$MISE$	N	$MISE$	N	$MISE$	N	$MISE$	N
$E(U \mid W) = 0$	250	0.53	8	1.01	11	0.58	8	1.05	11
	500	0.32	10	0.66	15	0.38	10	0.67	14
	1000	0.20	12	0.45	21	0.24	12	0.46	20
$U \perp\!\!\!\perp W$	250	0.62	30	0.76	27	0.61	30	0.78	29
$\varphi_0 = \hat{\varphi}_{LL}$	500	0.34	39	0.48	34	0.33	39	0.49	33
	1000	0.19	50	0.33	43	0.19	49	0.32	34
$U \perp\!\!\!\perp W$	250	0.65	28	0.89	27	0.63	28	0.91	29
$\varphi_0 = \hat{\varphi}_{MI}$	500	0.35	36	0.52	35	0.34	38	0.54	39
	1000	0.20	50	0.33	46	0.19	51	0.33	52
$U \perp\!\!\!\perp W$	250	0.59		1.07		0.61		1.10	
One-step from $\hat{\varphi}_{MI}$	500	0.32		0.69		0.34		0.70	
	1000	0.19		0.47		0.21		0.47	

Otherwise, we keep the same specifications of the regression function φ and the censoring variable C.

In this example, the estimation of the operator A is obtained by sorting the sample according to the values of the instrument W and obtaining an estimator for both survivor functions. Our estimators need to satisfy Assumption 4.7. Hence, we do not directly employ the Kaplan-Meier estimator, but its kernel smoothed version [see 47, 58, among others]. We let $\hat{S}_{U|W}(u \mid W = 1)$ and $\hat{S}_{U|W}(u \mid W = 0)$ be the estimators of the survivor function of U conditional on $W = 1$ and $W = 0$, respectively.

Let $\psi(u, w) = \hat{S}_{U|1}(u) - \hat{S}_{U|0}(u)$. One can write the estimator of the adjoint operator $A^{*\prime}_{\varphi}$ in the following form

$$\left(\hat{A}'^{*}_{\varphi_\dagger}\psi\right)(z) = -\frac{\sum_{i=1}^{n}\psi(u,w)\hat{f}_U(u)K_{h_Z}(Z_i - z, z)}{\sum_{i=1}^{n}K_{h_Z}(Z_i - z, z)},$$

with $\psi \in L^2_{U\times W}$, and $\hat{f}_U(u)$ a nonparametric estimator of the density of U as explained in Section 3.

In this case, the model is not identified under the mean independence restriction, as the completeness condition in Assumption 2.3 fails. As a matter of fact, the restriction $E(\varphi(Z) \mid W) = 0$ reduces to

$$\int \varphi(z) f_{Z|W}(z \mid w = 0)dz = \int \varphi(z) f_{Z|W}(z \mid w = 1)dz = 0$$

which cannot imply $\varphi = 0$, except when Z is also binary, or when φ is a two-parameter function in Z.

Results for this simulation exercise are reported in Table 4. As above, the MISE is multiplied by a factor of 100.

The performance of our estimator worsens compared to the case where we have two continuous instruments, which may be expected. Nonetheless, we can

TABLE 4
MISE and median number of iterations, N, with $U \perp\!\!\!\perp W$ and W binary.

	$C \perp\!\!\!\perp (T, Z, W)$				$C \perp\!\!\!\perp T \mid W$			
	$\varphi_1(z)$		$\varphi_2(z)$		$\varphi_1(z)$		$\varphi_2(z)$	
n	$MISE$	N	$MISE$	N	$MISE$	N	$MISE$	N
250	4.00	21	9.48	52	10.94	38	9.21	94
500	1.27	12	6.70	47	3.92	24	7.89	149
1000	0.48	8	6.00	64	1.50	17	7.32	213

appreciate how the MISE decreases as the number of observations increases. Moreover, the median number of iterations taken is often larger than above, which may be related to the fact that the information contained in each single iteration step is much smaller in this context.

Appendix A

Proof of Proposition 4.1. We only prove the first part of the Proposition, which is specific to this paper. The proof of the second part is identical to Darolles et al. [24] and Florens et al. [34], and it is omitted here for brevity.

We introduce the following additional notations

$$\begin{aligned} H_{C|W}(y \mid w) =& P(C \leq y \mid W = w) \\ H^{\delta}_{C|W}(y \mid w) =& P(C \leq y, \delta = 0 \mid W = w), \end{aligned}$$

where the definition of these objects should be apparent. Under Assumption 2.2, we immediately obtain that $H_{C|W}(y \mid w) < 1$ and $H^{\delta}_{C|W}(y \mid w) < 1$. We make use of the following Lemma.

Lemma A.1 (68)**.** *Let Assumptions 2.2, 4.1(iii) and 4.5 hold. Further, suppose that the functions $H_{C|W}$ and $H^{\delta}_{C|W}$ are twice continuously differentiable on $\mathcal{C} \times [0,1]^q$. Then*

$$\begin{aligned} \hat{F}_{C|W}(y \mid w) - F_{C|W}(y \mid w) &= S_{C|W}(y \mid w) - \hat{S}_{C|W}(y \mid w) \\ &= \sum_{i=1}^{n} \frac{\mathbf{K}_{h_S}(W_i - w, w)}{\sum_{i=1}^{n} \mathbf{K}_{h_S}(W_i - w, w)} \xi_{y,w}(Y_i, \delta_i) + O_P\left(\left(\frac{\log n}{n h_S^q}\right)^{3/4}\right), \end{aligned}$$

on $\mathcal{C} \times [0,1]^q$, as $n \to \infty$, with

$$\begin{aligned} \xi_{y,w}(Y_i, \delta_i) =& S_{C|W}(y \mid w) \left\{ \int_{\mathcal{C}} \frac{\mathbb{1}(Y_i \leq s) - H_{C|W}(s \mid w)}{(1 - H_{C|W}(s \mid w))^2} dH^{\delta}_{C|W}(s \mid w) \right. \\ &+ \frac{\mathbb{1}(Y_i \leq y, \delta_i = 0) - H^{\delta}_{C|W}(y \mid w)}{1 - H_{C|W}(y \mid w)} \\ &\left. - \int_{\mathcal{C}} \frac{\mathbb{1}(Y_i \leq s, \delta_i = 0) - H^{\delta}_{C|W}(s \mid w)}{(1 - H_{C|W}(s \mid w))^2} dH_{C|W}(s \mid w) \right\}, \end{aligned}$$

with

$$\sup_{(y,\tilde{y},w,\tilde{w}):y,\tilde{y}<\infty;w,\tilde{w}\in[0,1]^q} \mid \xi_{y,w}(\tilde{y},\tilde{w}) \mid <\infty.$$

We let

$$\tilde{r}(w) = \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)\, V_i}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)},$$

with

$$V_i = \frac{\delta_i Y_i}{S_{C|W}(Y_i \mid W_i)}.$$

We therefore decompose

$$\left(\hat{r} - \hat{A}\varphi\right)(w) = (\hat{r} - \tilde{r})(w) + \left(\tilde{r} - \hat{A}\varphi\right)(w),$$

where

$$||\tilde{r} - \hat{A}\varphi||^2 = O_P\left(n^{-\frac{2\rho}{2\rho+q}}\right),$$

directly, under Assumption 4.4 [see 24, 34]. We are therefore left with the term $\hat{r} - \tilde{r}$. After simple computations, this difference can be rewritten as

$$\begin{aligned}
&(\hat{r} - \tilde{r})(w) \\
&= \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)\, \delta_i Y_i \left(\frac{1}{\hat{S}_{C|W}(Y_i|W_i)} - \frac{1}{S_{C|W}(Y_i|W_i)}\right)}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)} \\
&= \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)\, V_i \left(\frac{S_{C|W}(Y_i|W_i) - \hat{S}_{C|W}(Y_i|W_i)}{S_{C|W}(Y_i|W_i)}\right)}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)} \\
&\quad + \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)\, V_i \left(\frac{\left[S_{C|W}(Y_i|W_i) - \hat{S}_{C|W}(Y_i|W_i)\right]^2}{\hat{S}_{C|W}(Y_i|W_i) S_{C|W}(Y_i|W_i)}\right)}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)} \\
&= \sum_{i=1}^n \left[\frac{\mathbf{K}_{h_W}(W_i - w, w)}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)} V_i \times \right. \\
&\quad \left. \left(\frac{\sum_{j=1}^n \frac{\mathbf{K}_{h_S}(W_j - W_i, W_i)}{\sum_{j=1}^n \mathbf{K}_{h_S}(W_j - W_i, W_i)} \xi_{Y_i,W_i}(Y_j, \delta_j) + O_P\left(\left(\frac{\log n}{n h_S^q}\right)^{3/4}\right)}{\hat{S}_{C|W}(Y_i \mid W_i)} (1 + o_P(1)) \right) \right],
\end{aligned}$$

where the last step follows from Lemma A.1 and Assumption 4.5.

Therefore, we have that

$$\|\hat{r} - \tilde{r}\|^2 = \left\| \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)\, V_i \left(\frac{\sum_{j=1}^n \frac{\mathbf{K}_{h_S}(W_j - W_i, W_i)}{\sum_{j=1}^n \mathbf{K}_{h_S}(W_j - W_i, W_i)} \xi_{Y_i,W_i}(Y_j, \delta_j)}{\hat{S}_{C|W}(W_i|W_i)} \right)}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)} \right\|^2$$

$$= \left(\frac{1}{\inf_{(y,w):y<\infty;w\in[0,1]^q} \mid \hat{S}_{C|W}(y \mid w) \mid} \right)^2 \times$$
$$\left\| \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) V_i \sum_{j=1}^n \frac{\mathbf{K}_{h_S}(W_j - W_i, W_i)}{\sum_{j=1}^n \mathbf{K}_{h_S}(W_j - W_i, W_i)} \xi_{Y_i,W_i}(Y_j, \delta_j)}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)} \right\|^2$$
$$= O_P(1) \left\| \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) \sum_{j=1}^n \frac{\mathbf{K}_{h_S}(W_j - W_i, W_i)}{\sum_{j=1}^n \mathbf{K}_{h_S}(W_j - W_i, W_i)} \xi_{Y_i,W_i}(Y_j, \delta_j) V_i}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)} \right\|^2,$$

where the last step follows from the conditions in Assumptions 4.3(i), 4.4(iii) and 4.5, which imply the uniform convergence of the conditional Kaplan-Meier estimator [22]; and Assumption 2.2, which implies that the conditional survivor function is almost surely bounded away from 0.

Directly from the results in Darolles et al. [24], we obtain that

$$\left\| \frac{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) \sum_{j=1}^n \frac{\mathbf{K}_{h_S}(W_j - W_i, W_i)}{\sum_{j=1}^n \mathbf{K}_{h_S}(W_j - W_i, W_i)} \xi_{Y_i,W_i}(Y_j, \delta_j) V_i}{\sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w)} \right\|^2$$
$$\leq \left(\frac{1}{\inf_{w:w\in[0,1]^q} \mid \hat{f}_W(w) \mid} \right)^2 \times$$
$$\left\| \frac{1}{nh_W^q} \sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) \sum_{j=1}^n \frac{\frac{1}{n}\mathbf{K}_{h_S}(W_j - W_i, W_i)}{\frac{1}{n}\sum_{j=1}^n \mathbf{K}_{h_S}(W_j - W_i, W_i)} \xi_{Y_i,W_i}(Y_j, \delta_j) V_i \right\|^2$$
$$= O_P(1) \left\| \frac{1}{nh_W^q} \sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) \sum_{j=1}^n \frac{\frac{1}{nh_S^q}\mathbf{K}_{h_S}(W_j - W_i, W_i)}{\hat{f}_W(W_i)} \xi_{Y_i,W_i}(Y_j, \delta_j) V_i \right\|^2$$
$$\leq O_P(1) \left(\frac{1}{\inf_{w:w\in[0,1]^q} \mid \hat{f}_W(w) \mid} \right)^2 \times$$
$$\left\| \frac{1}{n^2 h_W^q h_S^q} \sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) \sum_{j=1}^n \mathbf{K}_{h_S}(W_j - W_i) \xi_{Y_i,W_i}(Y_j, \delta_j) V_i \right\|^2$$
$$= O_P(1) \left\| \frac{1}{nh_W^q} \sum_{j=1}^n \left(\frac{1}{nh_S^q} \sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) \mathbf{K}_{h_S}(W_j - W_i, W_i) V_i \right) \xi_{Y_i,W_i}(Y_j, \delta_j) \right\|^2,$$

where $\hat{f}_W(w)$ is the Nadaraya-Watson estimator of the density of W using generalized kernels as defined in Assumption 4.1. We now consider

$$\frac{1}{n^2 h_W^q h_S^q} \sum_{j=1}^n \sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) \mathbf{K}_{h_S}(W_j - W_i, W_i) V_i \xi_{Y_i,W_i}(Y_j, \delta_j)$$
$$= \frac{1}{n^2 h_W^q h_S^q} \sum_{i=1}^n \sum_{j=1,j\neq i}^n \mathbf{K}_{h_W}(W_i - w, w) \mathbf{K}_{h_S}(W_j - W_i, W_i) V_i \xi_{Y_i,W_i}(Y_j, \delta_j)$$
$$+ \frac{1}{n^2 h_W^q h_S^q} \sum_{i=1}^n \mathbf{K}_{h_W}(W_i - w, w) \mathbf{K}_{h_S}(0, W_i) V_i \xi_{Y_i,W_i}(Y_i, \delta_i)$$
$$= \frac{1}{n^2 h_W^q h_S^q} \sum_{i=1}^n \sum_{j=1,j\neq i}^n \mathbf{K}_{h_W}(W_i - w, w) V_i \mathbf{K}_{h_S}(W_j - W_i, W_i) E\left[\xi_{Y_i,W_i}(Y_j, \delta_j) | W_j\right]$$

$$
\begin{aligned}
&+\frac{1}{n^2 h_W^q h_S^q}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbf{K}_{h_W}\left(W_i-w,w\right)V_i\mathbf{K}_{h_S}\left(W_j-W_i,W_i\right)\times\\
&\qquad\qquad\left(\xi_{Y_i,W_i}(Y_j,\delta_j)-E\left[\xi_{Y_i,W_i}(Y_j,\delta_j)|W_j\right]\right)\\
&+\frac{1}{n^2 h_W^q h_S^q}\sum_{i=1}^{n}\mathbf{K}_{h_W}\left(W_i-w,w\right)\mathbf{K}_{h_S}\left(0,W_i\right)V_i\xi_{Y_i,W_i}(Y_i,\delta_i)\\
=&I+II+III.
\end{aligned}
$$

Let

$$
\begin{aligned}
E\left[III\right]=&\frac{1}{nh_W^q h_S^q}E\left[\mathbf{K}_{h_W}\left(W_i-w,w\right)\mathbf{K}_{h_S}\left(0,W_i\right)E\left[V_i\xi_{Y_i,W_i}(Y_i,\delta_i)|W_i\right]\right]\\
=&\frac{1}{nh_S^q h_W^q}\int\mathbf{K}_{h_W}\left(W_i-w,w\right)\mathbf{K}_{h_S}\left(0,W_i\right)E\left[V_i\xi_{Y_i,W_i}(Y_i,\delta_i)|W_i\right]f_W(W_i)dW_i\\
=&\frac{1}{nh_S^q}\int\mathbf{K}\left(h_W u,w\right)\mathbf{K}_{h_S}\left(0,w+h_W u\right)\times\\
&\qquad\qquad E\left[V_i\xi_{Y_i,W_i}(Y_i,\delta_i)|W_i=w+h_W u\right]f_W(w+h_W u)du\\
=&O\left(\frac{1}{nh_S^q}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
Var\left(III\right)=&\frac{1}{n^3 h_W^{2q}h_S^{2q}}Var\left(\mathbf{K}_{h_W}\left(W_i-w,w\right)\mathbf{K}_{h_S}\left(0,W_i\right)V_i\xi_{Y_i,W_i}(Y_i,\delta_i)\right)\\
\leq&\frac{1}{n^3 h_W^{2q}h_S^{2q}}E\left(\mathbf{K}_{h_W}^2\left(W_i-w,w\right)\mathbf{K}_{h_S}^2\left(0,W_i\right)E\left[V_i^2\xi_{Y_i,W_i}^2(Y_i,\delta_i)|W_i\right]\right)\\
=&O\left(\frac{1}{(nh_S^q)^2}\frac{1}{nh_W^q}\right),
\end{aligned}
$$

where the conclusion follows by the usual change of variable, the uniform boundedness of the kernel function, Lemma A.1 and Assumption 4.4. Using a similar argument, we have that

$$
\frac{1}{h_S^q}E\left[\mathbf{K}_{h_S}\left(W_j-W_i,W_i\right)E\left[\xi_{Y_i,W_i}(Y_j,\delta_j)|W_j\right]|Y_i,W_i\right]=O\left(h_S^{\rho}\right),
$$

which directly implies

$$
\|I\|^2=O_P\left(h_S^{2\rho}\right).
$$

Finally,

$$
\begin{aligned}
Var\left(II\right)\leq&\frac{1}{nh_W^{2q}h_S^{2q}}E\left[\mathbf{K}_{h_W}\left(W_i-w,w\right)\mathbf{K}_{h_W}\left(W_{i'}-w,w\right)V_iV_{i'}\times\right.\\
&\mathbf{K}_{h_S}\left(W_j-W_i,W_i\right)\mathbf{K}_{h_S}\left(W_j-W_{i'},W_{i'}\right)\left(\xi_{Y_i,W_i}(Y_j,\delta_j)-E\left[\xi_{Y_i,W_i}(Y_j,\delta_j)|W_j\right]\right)\times\\
&\left.\left(\xi_{Y_{i'},W_{i'}}(Y_j,\delta_j)-E\left[\xi_{Y_{i'},W_{i'}}(Y_j,\delta_j)|W_j\right]\right)\right]\\
+&\frac{1}{n^2 h_W^{2q}h_S^{2q}}E\left[\mathbf{K}_{h_W}^2\left(W_i-w,w\right)V_i^2\mathbf{K}_{h_S}^2\left(W_j-W_i,W_i\right)\times\right.\\
&\left.\left(\xi_{Y_i,W_i}(Y_j,\delta_j)-E\left[\xi_{Y_i,W_i}(Y_j,\delta_j)|W_j\right]\right)^2\right],
\end{aligned}
$$

where the remaining terms are zero by the law of iterated expectations. By the usual change of variable, Assumption 4.4 and Lemma A.1, one can show that

$$Var\,(II) = O\left(\frac{1}{nh_S^q} + \frac{1}{nh_S^q}\frac{1}{nh_W^q}\right) = O\left(\frac{1}{nh_S^q}\right) + o\left(\frac{1}{nh_S^q}\right).$$

The result of the Proposition follows from Markov inequality, and the assumption $h_S = O(h_W)$. □

Proof of Theorem 4.1. We just give the main steps of the proof. More details can be found in Carrasco et al. [11] and Centorrino [12]. Let us denote by R a generic positive constant. To reduce the notational burden, we use the notation $\Phi_\beta(N)$, where $\Phi_\beta(N) = N^{-\beta}$, under the strong source condition, and $\Phi_\beta(N) = (\log(N))^{-\beta}$ under the weak source condition, respectively. As the operator A^*A is compact and thus admits a singular value decomposition, we also use the notation $\Phi_\beta(A^*A)$ to signify that the function Φ_β is applied to the singular values of A^*A. Finally, the source condition implies that we can write $\varphi = \Phi_{\beta/2}(A^*A)v$, with $v \in L^2_Z$, and $\|v\| \leq R$.

We first recall the following definition.

Definition A.1 (Qualification)**.** *A regularization procedure, g_N, is said to have qualification of order $\kappa > 0$, if:*

$$\sup_{0<a\leq\|A\|^2} \mid 1 - ag_N(a) \mid a^\eta \leq RN^{-\eta}, \tag{A.1}$$

for $0 < \eta \leq \kappa$. □

In particular, the Landweber-Fridman regularization has qualification equal to ∞, in the sense that for every $\eta > 0$, the inequality in equation (A.1) holds with

$$1 - ag_N(a) = (1 - ca)^N.$$

Moreover, we need the following.

Assumption A.1. *There exist two positive constants R and η such that:*

$$\sup_{N^{-1}\leq a\leq\|A\|^2} \frac{\Phi_{\beta/2}(a)}{a^\eta} \leq R\Phi_{\beta/2}(N)N^\eta. \tag{A.2}$$

□

Definition A.1 and Assumption A.1 together imply that

$$\begin{aligned}
&\sup_{0<a\leq\|A\|^2} \mid (1-ca)^N \mid \Phi_{\beta/2}(a) \\
&= \sup_{0<a\leq\|A\|^2} \mid (1-ca)^N \mid a^\eta \frac{\Phi_{\beta/2}(a)}{a^\eta} \\
&\leq \sup_{0<a\leq\|A\|^2} \mid (1-ca)^N \mid a^\eta \sup_{0<a\leq\|A\|^2} \frac{\Phi_{\beta/2}(a)}{a^\eta} \leq R\Phi_{\beta/2}(N).
\end{aligned}$$

This result is used repeatedly in the proof below. We have

$$\begin{aligned}\hat{\varphi}^N - \varphi &= c\sum_{j=0}^{N-1}(I - c\hat{A}^*\hat{A})^j\hat{A}^*(\hat{r} - \hat{A}\varphi) \\ &+ c\sum_{j=0}^{N-1}(I - c\hat{A}^*\hat{A})^j\hat{A}^*\hat{A}\varphi - c\sum_{j=0}^{N-1}(I - cA^*A)^jA^*A\varphi \\ &+ c\sum_{j=0}^{N-1}(I - cA^*A)^jA^*A\varphi - \varphi \\ &= I + II + III.\end{aligned}$$

Given the source condition and the qualification of Landweber-Fridman regularization, we directly have that $||III||^2 = O_P\left(\Phi_\beta(N)\right)$. Moreover

$$\begin{aligned}||I||^2 &= O_P\left(||c\sum_{j=0}^{N-1}(I - c\hat{A}^*\hat{A})^j\hat{A}^*||^2||\hat{r} - \hat{A}\varphi||^2\right) \\ &= O_P\left(Nn^{-\frac{2\rho}{2\rho+q}}\right),\end{aligned}$$

directly from the result in Proposition 4.1, and with $h_S = O_P(h_W)$. Finally, thanks to

$$\sum_{j=0}^{N-1}(I - cA^*A)^jA^*A = I - (I - cA^*A)^N$$

we have:

$$||II||^2 = ||\left[(I - c\hat{A}^*\hat{A})^N - (I - cA^*A)^N\right]\varphi||^2.$$

By using the Taylor theorem for integer powers of positive operators in Bhatia and Sinha [7], which applies provided $c < 1$, we obtain

$$\begin{aligned}[(I - c\hat{A}^*\hat{A})^N - (I - cA^*A)^N]\varphi &= N\left(\hat{A}^*\hat{A} - A^*A\right)(I - cA^*A)^{N-1}\varphi \\ &+ O\left(N(N-1)\|\left(\hat{A}^*\hat{A} - A^*A\right)(I - cA^*A)^{N-2}\varphi\|^2\right).\end{aligned}$$

We ignore for the moment the remainder of the Taylor expansion, which is shown to be negligible under identical conditions. We thus have

$$\begin{aligned}||II||^2 &= \|N\left(\hat{A}^*\hat{A} - A^*A\right)(I - cA^*A)^{N-1}\varphi\|^2 \\ &\leq N^2\|\hat{A}^* - A^*\|^2\|A(I - cA^*A)^{N-1}\varphi\|^2 \\ &\quad + N^2\|A^*\left(\hat{A} - A\right)(I - cA^*A)^{N-1}\varphi\|^2 \\ &= ||IIa||^2 + ||IIb||^2.\end{aligned}$$

Therefore,

$$||IIa||^2 = O_P\left(n^{-\frac{2\lambda}{2\lambda+p+q}} N\Phi_\beta(N)\right),$$

which follows directly from the result of Proposition 4.1 and Assumption A.1; and

$$\begin{aligned}
||IIb||^2 &= N^2\left\langle A^*\left(\hat{A}-A\right)(I-cA^*A)^{N-1}\varphi, A^*\left(\hat{A}-A\right)(I-cA^*A)^{N-1}\varphi\right\rangle \\
&= N^2\left\langle \left(\hat{A}-A\right)(I-cA^*A)^{N-1}\varphi, AA^*\left(\hat{A}-A\right)(I-cA^*A)^{N-1}\varphi\right\rangle \\
&\leq N^2\|\left(\hat{A}-A\right)(I-cA^*A)^{N-1}\varphi\|\|AA^*\left(\hat{A}-A\right)(I-cA^*A)^{N-1}\varphi\| \\
&\leq R^2N^2\|\left(\hat{A}-A\right)(I-cA^*A)^{N-1}\Phi_{\beta/2}(A^*A)\|\times \\
&\quad \|AA^*\left(\hat{A}-A\right)(I-cA^*A)^{N-1}\Phi_{\beta/2}(A^*A)\| \\
&\leq R^2N^2\|\left(\hat{A}-A\right)(I-cA^*A)^{N-1}\Phi_{\beta/2}(A^*A)\|\times \\
&\quad \|(I-cA^*A)^{N-1}\Phi_{\beta/2}(A^*A)A^*A\| \\
&\leq R^2N^2\|\hat{A}-A\|^2\|(I-cA^*A)^{N-1}\Phi_{\beta/2}(A^*A)(A^*A)^{1/2}\|^2 \\
&= O_P\left(n^{-\frac{2\lambda}{2\lambda+p+q}} N\Phi_\beta(N)\right),
\end{aligned}$$

where the last result follows from Proposition 4.1 and Assumption A.1. We finally notice that the reminder of the Taylor expansion can be treated in the same way. It can be therefore proven that the reminder is of the order $n^{-\frac{4\lambda}{2\lambda+p+q}} N^2\Phi_{2\beta}(N)$, and thus negligible under the conditions given in the statement of the Theorem. Finally,

$$||II||^2 = O_P\left(n^{-\frac{2\lambda}{2\lambda+p+q}} N\Phi_\beta(N)\right).$$

The result of the theorem follows. □

Proof of Theorem 4.2. To obtain uniform consistency of the nonparametric estimators, we must impose some additional assumptions. These are listed below. Without loss of generality, we use the word density irrespective of W being discrete or continuous.

The marginal density of W can be estimated by different methods depending on the nature and the dimension of the instrument. Therefore, we suppose that there is a function $d(\cdot)$, such that

$$\hat{f}_W(w) = \frac{1}{n}\sum_{i=1}^{n} d(W_i - w).$$

This function could be a kernel for continuous or discrete variables [see 2, 54]; or a product of indicator functions for purely discrete instruments. Then we define

$$\hat{f}_Z(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_Z}(Z_i - z, z)$$

$$\hat{f}_U(u) = \frac{1}{n}\sum_{i=1}^{n} K_{h_U}(U_i - u, u)\,\Delta\hat{F}_U(U_i)$$
$$\hat{f}_{T,Z}(t,z) = \frac{1}{n}\sum_{i=1}^{n} K_{h_U}(Y_i - t, t)\,\mathbf{K}_{h_Z}(Z_i - z, z)\,\Delta\hat{F}_T(Y_i)$$
$$\hat{f}_{T,Z,W}(t,z,w) = \frac{1}{n}\sum_{i=1}^{n} K_{h_U}(Y_i - t, t)\,K_{h_Z}(Z_i - z, z)\,d(W_i - w)\Delta\hat{F}_T(Y_i)$$
$$\hat{f}_{T,Z|W}(t,z \mid w) = \frac{\hat{f}_{T,Z,W}(t,z,w)}{\hat{f}_W(w)}.$$

We first list lemmas that are useful to prove the main result of the theorem. Some of the results of this section are taken from Centorrino et al. [14] and are given without proof.

It is convenient to recall that, for any functions $\varphi, \tilde{\varphi} \in L^2_Z$, and $\psi \in L^2_{U\times W}$, we can write

$$\left(\hat{A}'_\varphi\tilde{\varphi}\right)(u,w) = \int \hat{a}(\varphi(z)+u,z,w)\tilde{\varphi}(z)dz$$
$$\left(\hat{A}'^*_\varphi\psi\right)(z) = \int\int \hat{a}^*(\varphi(z)+u,z,w)\psi(u,w)dudw,$$

where we take

$$\hat{a}(\varphi(z)+u,z,w) = \hat{f}_{T,Z|W}(\varphi(z)+u,z \mid w) - \hat{f}_{T,Z}(\varphi(z)+u,z)$$
$$\hat{a}^*(\varphi(z)+u,z,w) = \frac{\hat{f}_{T,Z,W}(\varphi(z)+u,z,w)\hat{f}_U(u) - \hat{f}_{T,Z}(\varphi(z)+u,z)\hat{f}_W(w)\hat{f}_U(u)}{\hat{f}_Z(z)},$$

to be the operator kernels of $\hat{A}'_\varphi$ and $\hat{A}'^*_\varphi$, respectively. When W is a discrete variable, integrals can be replaced by sums, where appropriate, and we use the integral notation without loss of generality.

We state the following Lemma.

Lemma A.2 (Centorrino et al. [14])**.** *Let Assumptions 4.12 and 4.8-4.11 hold, $\tilde{\varphi} \in L^2_Z$ and $\psi \in L^2_{U\times W}$. There exist operators $G_{\varphi_\dagger,\hat{\varphi}_k}$ and $G^*_{\varphi_\dagger,\hat{\varphi}_k}$ such that*

$$\left(\hat{A}'_{\hat{\varphi}_k}\tilde{\varphi}\right)(u,w) = \left(G_{\varphi_\dagger,\hat{\varphi}_k}\hat{A}'_{\varphi_\dagger}\tilde{\varphi}\right)(u,w),$$
$$\left(\hat{A}'^*_{\hat{\varphi}_k}\psi\right)(z) = \left(\hat{A}'^*_{\varphi_\dagger}G^*_{\varphi_\dagger,\hat{\varphi}_k}\psi\right)(z),$$

and

$$\|G_{\varphi_\dagger,\hat{\varphi}_k} - I\| \leq \varpi_1\kappa_n^{-1}\|\hat{\varphi}_k - \varphi_\dagger\|,$$
$$\|G^*_{\varphi_\dagger,\hat{\varphi}_k} - I\| \leq \varpi_2(h_u^2\kappa_n)^{-1}\|\hat{\varphi}_k - \varphi_\dagger\|$$

where I is the identity operator, $\varpi_1, \varpi_2 < \infty$, positive constants, and κ_n is such that $(h_u^2\kappa_n)^{-1}(\delta_n \vee \gamma_n) \to 0$, as $n \to \infty$.

Among other things, this lemma implies that

$$\begin{aligned}&\|\hat{A}(\hat{\varphi}_k) - \hat{A}(\varphi_\dagger) - \hat{A}'_{\varphi_\dagger}(\hat{\varphi}_k - \varphi_\dagger)\| \leq \|\left(\hat{A}'_{\hat{\varphi}_k} - \hat{A}'_{\varphi_\dagger}\right)(\hat{\varphi}_k - \varphi_\dagger)\| \\ =&\|\left(G_{\varphi_\dagger,\hat{\varphi}_k} - I\right)\hat{A}'_{\varphi_\dagger}(\hat{\varphi}_k - \varphi_\dagger)\| \leq \varpi_1\kappa_n^{-1}\|\hat{\varphi}_k - \varphi_\dagger\|\|\hat{A}'_{\varphi_\dagger}(\hat{\varphi}_k - \varphi_\dagger)\|.\end{aligned}$$

This condition implies (but it is not implied by) a Lipschitz continuity condition on $\hat{A}'$ [see 26, 42, 45].

We will also use the following results below.

Lemma A.3 (Kaltenbacher et al. (2008, Lemma 2.9, p. 17)). *Let a and b be non-negative. Then there is a positive constant $M(a,b)$ independent of N so that*

$$\sum_{j=0}^{N-1}(N-j)^{-a}(j+1)^{-b} \leq M(a,b)N^{1-a-b}D(N),$$

with

$$D(N) = \begin{cases}1, & a \vee b < 1 \\ \ln(N), & a \vee b = 1 \\ N^{a\vee b-1}, & a \vee b > 1\end{cases}.$$

Lemma A.4 (Kaltenbacher et al. (2008, Lemma 2.10, p. 18)). *Let A be a compact operator such that $c\|A\|^2 \leq 1$, with A^* be its adjoint. Further let $s \in [0,1]$, and $N \geq 0$, an integer. Then the following estimates hold*

$$\begin{aligned}\|\left(I - cA^*A\right)^N(A^*A)^s\| &= O\left((N+1)^{-s}\right), \\ \left\|c\sum_{k=0}^{N-1}\left(I - cA^*A\right)^j(A^*A)^s\right\| &= O\left(N^{1-s}\right).\end{aligned}$$

We now turn to the proof of the main result of the Theorem. We have

$$\begin{aligned}\hat{\varphi}_N - \varphi_\dagger &= \hat{\varphi}_{N-1} - \varphi_\dagger - c\hat{A}'^*_{\hat{\varphi}_{N-1}}\left(\hat{A}(\hat{\varphi}_{N-1})\right) \\ &= \hat{\varphi}_{N-1} - \varphi_\dagger - c\hat{A}'^*_{\hat{\varphi}_{N-1}}\left(\hat{A}(\hat{\varphi}_{N-1}) - \hat{A}(\varphi_\dagger)\right) - c\hat{A}'^*_{\hat{\varphi}_{N-1}}\left(\hat{A}(\varphi_\dagger) - A(\varphi_\dagger)\right) \\ &= \hat{\varphi}_{N-1} - \varphi_\dagger - c\hat{A}'^*_{\varphi_\dagger}\left(\hat{A}(\hat{\varphi}_{N-1}) - \hat{A}(\varphi_\dagger)\right) \\ &\quad - c\left(\hat{A}'^*_{\hat{\varphi}_{N-1}} - \hat{A}'^*_{\varphi_\dagger}\right)\left(\hat{A}(\hat{\varphi}_{N-1}) - A(\varphi_\dagger)\right) \\ &\quad - c\hat{A}'^*_{\varphi_\dagger}\left(\hat{A}(\varphi_\dagger) - A(\varphi_\dagger)\right) \\ &= \hat{\varphi}_{N-1} - \varphi_\dagger - c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}(\hat{\varphi}_{N-1} - \varphi_\dagger) \\ &\quad - c\hat{A}'^*_{\varphi_\dagger}\left(\hat{A}(\hat{\varphi}_{N-1}) - \hat{A}(\varphi_\dagger) - \hat{A}'_{\varphi_\dagger}(\hat{\varphi}_{N-1} - \varphi_\dagger)\right) \\ &\quad - c\left(\hat{A}'^*_{\hat{\varphi}_{N-1}} - \hat{A}'^*_{\varphi_\dagger}\right)\left(\hat{A}(\hat{\varphi}_{N-1}) - A(\varphi_\dagger)\right) \\ &\quad - c\hat{A}'^*_{\varphi_\dagger}\left(\hat{A}(\varphi_\dagger) - A(\varphi_\dagger)\right),\end{aligned}$$

where the second line follows from $A(\varphi_\dagger) = 0$. By replacing iteratively $\hat{\varphi}_j$, for all $k = 0, \ldots, N-2$, and letting $\hat{e}_k = \hat{\varphi}_k - \varphi_\dagger$, for all $k = 0, 1, 2, \ldots$, we finally

obtain

$$\begin{aligned}\hat{e}_N = &\left(I - c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^N (\varphi^0 - \varphi_\dagger) \\ &- c\sum_{k=0}^{N-1}\left(I - c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^k \hat{A}'^*_{\varphi_\dagger}\left(\hat{A}(\varphi_\dagger) - A(\varphi_\dagger)\right) \\ &- c\sum_{j=0}^{N-1}\left(I - c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^{N-k-1} \hat{A}'^*_{\varphi_\dagger}\left(\hat{A}(\hat{\varphi}_k) - \hat{A}(\varphi_\dagger) - \hat{A}'_{\varphi_\dagger}\hat{e}_k\right) \\ &- c\sum_{k=0}^{N-1}\left(I - c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^{N-k-1}\left(\hat{A}'^*_{\hat{\varphi}_k} - \hat{A}'^*_{\varphi_\dagger}\right)\left(\hat{A}(\hat{\varphi}_k) - A(\varphi_\dagger)\right) \\ &= I + II + III + IV.\end{aligned}$$

The first two terms are similar as in the asymptotic expansion of Landweber-Fridman regularization for linear inverse problems (see the proof of Theorem 4.1). By contrast, the terms in III and IV come from the nonlinearity of the inverse problem in our framework. As a matter of fact, these latter terms are identically zero when the ill-posed inverse problem is linear. To control these terms, we use the main result provided in Lemma A.2.

Let δ_n and γ_n to be defined as in Assumption 4.12. We again use the letter R to denote a strictly positive constant, which may take different values in different instances. We start by considering the term in I. It follows from the strong source condition that

$$\begin{aligned}||I||^2 &\leq 2||\left[\left(I - c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^N - \left(I - cA'^*_{\varphi_\dagger}A'_{\varphi_\dagger}\right)^N\right](A'^*_{\varphi_\dagger}A'_{\varphi_\dagger})^{\beta/2}v||^2 \\ &\quad + 2||\left(I - cA'^*_{\varphi_\dagger}A'_{\varphi_\dagger}\right)^N (A'^*_{\varphi_\dagger}A'_{\varphi_\dagger})^{\beta/2}v||^2 \\ &= 2||I_a||^2 + 2||I_b||^2.\end{aligned}$$

Under the same conditions outlined earlier, $||I_b||^2 = O_P\left(N^{-\beta}\right)$, and

$$||I_a||^2 = O_P\left(\gamma_n N^{(1-\beta)/2}\right).$$

Similarly, for II, we have

$$||II||^2 \leq ||c\sum_{j=0}^{N-1}\left(I - c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^j \hat{A}'^*_{\varphi_\dagger}||^2||\hat{A}(\varphi_\dagger) - A(\varphi_\dagger)||^2 = O_P\left(N\delta_n^2\right).$$

We now control the nonlinear terms following the approach in Centorrino et al. [14]. Let

$$\begin{aligned}&\left(E||III||^2\right)^{1/2} \\ &\leq c\sum_{k=0}^{N-1}\left(E\left\|\left(I - c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^{N-k-1}\hat{A}'^*_{\varphi_\dagger}\right\|^2 ||\hat{A}(\hat{\varphi}_k) - \hat{A}(\varphi_\dagger) - \hat{A}'_{\varphi_\dagger}(\hat{\varphi}_k - \varphi_\dagger)||^2\right)^{1/2}\end{aligned}$$

$$\leq cR\kappa_n^{-1}\sum_{k=0}^{N-1}(N-k)^{-1/2}\left(E||\hat{e}_k||^4E||\hat{A}'_{\varphi_\dagger}\hat{e}_k||^4\right)^{1/4},$$

where the second inequality follows from Lemma A.4 with $s=0.5$. Similarly,

$$\begin{aligned}
&\left(E||IV||^2\right)^{1/2}\\
&=\left(E\left\|c\sum_{k=0}^{N-1}\left(I-c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^{N-k-1}\hat{A}'^*_{\varphi_\dagger}\left(G^*_{\hat{\varphi}_k,\varphi_\dagger}-I\right)\left(\hat{A}(\hat{\varphi}_k)-A(\varphi_\dagger)\right)\right\|^2\right)^{1/2}\\
&\leq c\sum_{k=0}^{N-1}\left(E\left\|\left(I-c\hat{A}'^*_{\varphi_\dagger}\hat{A}'_{\varphi_\dagger}\right)^{N-k-1}\hat{A}'^*_{\varphi_\dagger}\right\|^2\left\|G_{\hat{\varphi}_k,\varphi_\dagger}-I\right\|^2\left\|\hat{A}(\hat{\varphi}_k)-A(\varphi_\dagger)\right\|^2\right)^{1/2}\\
&\leq cR(h_u^2\kappa_n)^{-1}\sum_{k=0}^{N-1}(N-k)^{-1/2}\left(E\left\|\hat{e}_k\right\|^2\left\|\hat{A}(\hat{\varphi}_k)-A(\varphi_\dagger)\right\|^2\right)^{1/2}\\
&\leq cR(h_u^2\kappa_n)^{-1}\sum_{k=0}^{N-1}(N-k)^{-1/2}\left(E\left\|\hat{e}_k\right\|^4E\left\|\hat{A}'_{\varphi_\dagger}\hat{e}_k\right\|^4\right)^{1/4}\\
&\quad+cR(h_u^2\kappa_n)^{-1}\sum_{k=0}^{N-1}(N-k)^{-1/2}\left(E\left\|\hat{e}_k\right\|^4E\left\|\hat{A}(\varphi_\dagger)-A(\varphi_\dagger)\right\|^4\right)^{1/4}\\
&\leq cR(h_u^2\kappa_n)^{-1}\sum_{k=0}^{N-1}(N-k)^{-1/2}\left(E\left\|\hat{e}_k\right\|^4E\left\|\hat{A}'_{\varphi_\dagger}\hat{e}_k\right\|^4\right)^{1/4}\\
&\quad+cR\delta_n(h_u^2\kappa_n)^{-1}\sum_{k=0}^{N-1}(N-k)^{-1/2}\left(E\left\|\hat{e}_k\right\|^4\right)^{1/4}\\
&=O\left(h_u^{-2}\left(E||III||^2\right)^{1/2}\right)+cR\delta_n(h_u^2\kappa_n)^{-1}\sum_{k=0}^{N-1}(N-k)^{-1/2}\left(E\left\|\hat{e}_k\right\|^4\right)^{1/4}.
\end{aligned}$$

The last result implies

$$E||III||^2=o\left(E||IV||^2\right),$$

so that the convergence of the nonlinearity term is dominated by the terms in IV.

We prove the result of the Theorem by induction. We let

$$\begin{aligned}
E\left\|\hat{e}_k\right\|^2=&O\left((k+1)\delta_n^2+(k+1)^{1-\beta}\gamma_n^2+(k+1)^{-\beta}\right)\\
E\left\|\hat{A}'_{\varphi_\dagger}\hat{e}_k\right\|^2=&O\left(\delta_n^2+(k+1)^{-\beta}\gamma_n^2+(k+1)^{-\beta-1}\right),
\end{aligned}$$

and for an integer $l\geq 1$

$$\begin{aligned}
\left(E\left\|\hat{e}_k\right\|^{2l}\right)^{1/2l}=&O\left(E\left\|\hat{e}_k\right\|^2\right)^{1/2}\\
\left(E\left\|\hat{A}'_{\varphi_\dagger}\hat{e}_k\right\|^{2l}\right)^{1/2l}=&O\left(E\left\|\hat{A}'_{\varphi_\dagger}\hat{e}_k\right\|^2\right)^{1/2}.
\end{aligned}$$

The final result is established by controlling the remaining terms. We have

$$\left(E||III||^2\right)^{1/2}$$

$$\leq R\kappa_n^{-1}\left(\delta_n^2\sum_{k=0}^{N-1}(N-k)^{-1/2}(k+1)^{1/2}\right.$$
$$\left.+\gamma_n^2\sum_{k=0}^{N-1}(N-k)^{-1/2}(k+1)^{1/2-\beta}+\sum_{k=0}^{N-1}(N-k)^{-1/2}(k+1)^{-\beta-1/2}\right),$$

and we analyze these terms one by one.

First

$$\delta_n^2\sum_{k=0}^{N-1}(N-k)^{-1/2}(k+1)^{1/2}\leq R\delta_n^2\left(\sum_{k=0}^{N-1}(N-k)^{-1}\right)^{1/2}\left(\sum_{k=0}^{N-1}(k+1)\right)^{1/2}$$
$$\leq R\delta_n^2 N(\ln(N))^{1/2}.$$

Similarly,

$$\gamma_n^2\sum_{k=0}^{N-1}(N-k)^{-1/2}(k+1)^{1/2-\beta}\leq R\gamma_n^2\begin{cases}N^{1-\beta}(\ln(N))^{1/2} & \beta\in(0,1/2)\\ N^{1-\beta} & \beta\in[1/2,1]\end{cases}.$$

The condition in Assumption 4.14(iii) is enough to control the latter terms so that they are both negligible asymptotically. For the bias component, we have instead

$$\sum_{k=0}^{N-1}(N-k)^{-1/2}(k+1)^{-\beta-1/2}\leq R\begin{cases}N^{-\beta} & 0<\beta<1/2\\ N^{-1/2}\ln(N) & \beta=1/2\\ N^{-1/2} & 1/2<\beta\leq 1\end{cases}.$$

Thus

$$\left(E||III||^2\right)^{1/2}$$
$$=O\left(\kappa_n^{-1}\left(\delta_n^2 N(\ln(N))^{1/2}+\gamma_n^2 N^{1-\beta}(\ln(N))^{1/2}\right.\right.$$
$$\left.\left.+\ N^{-\beta/2}(N^{\beta/2}\mathbb{1}(\beta<1/2)+N^{-\beta/2}\ln(N)\mathbb{1}(\beta=1/2)+N^{(\beta-1)/2})\right)\right).$$

In an analogous fashion, one can prove that

$$\delta_n(h_u^2\kappa_n)^{-1}\sum_{k=0}^{N-1}(N-k)^{-1/2}\left(E\left\|\hat{e}_k\right\|^4\right)^{1/4}$$
$$\leq R\delta_n(h_u^2\kappa_n)^{-1}\left(\delta_n N(\ln(N))^{1/2}+\gamma_n N^{1-\beta/2}(\ln(N))^{1/2}+N^{(1-\beta)/2}\right).$$

So that finally,

$$\left(E||IV||^2\right)^{1/2}=O\left(h_u^{-2}\left(E||III||^2\right)^{1/2}\right.$$
$$\left.+\ \delta_n(h_u^2\kappa_n)^{-1}\left(\delta_n N(\ln(N))^{1/2}+\gamma_n N^{1-\beta/2}(\ln(N))^{1/2}+N^{(1-\beta)/2}\right)\right).$$

Because of the restriction imposed in Assumption 4.14(i), we have that

$$\begin{aligned} N\delta_n^2(h_u^2\kappa_n)^{-1}(\ln(N))^{1/2} &\asymp \sqrt{N}\delta_n \\ \delta_n\gamma_n(h_u^2\kappa_n)^{-1}N^{1-\beta/2}(\ln(N))^{1/2} &\asymp \gamma_n N^{(1-\beta)/2} \\ \delta_n(h_u^2\kappa_n)^{-1}N^{(1-\beta)/2} &= o(N^{-\beta/2}), \end{aligned}$$

Finally, we need to bound the term $h_u^{-2}\left(E||III||^2\right)^{1/2}$. The bias component can be controlled as follows: for $\beta < 1/2$, it is enough to have $(h_u^2\kappa_n)^{-1}N^{-\beta/2} = O(1)$, and, for $\beta = 1/2$, we need $(h_u^2\kappa_n)^{-1}N^{-1/4}\ln(N) = O(1)$. However, for $\beta > 1/2$, this requires

$$(h_u^2\kappa_n)^{-1}N^{(\beta-1)/2} = O(1),$$

which, for $\beta \leq 1$ and $h_u^2\kappa_n = o(1)$, cannot be satisfied for all β's. Therefore, we say there exists a $\beta^* < 1$, such that the condition above is satisfied. This is equivalent to the condition given in Assumption 4.14(iii). Finally, reasoning as above,

$$\begin{aligned} \delta_n^2(h_u^2\kappa_n)^{-1}N(\ln(N))^{1/2} &\asymp \sqrt{N}\delta_n \\ \gamma_n^2(h_u^2\kappa_n)^{-1}N^{1-\beta}(\ln(N))^{1/2} &= o\left(\gamma_n N^{(1-\beta)/2}\right). \end{aligned}$$

The result of the Theorem follows from Markov's inequality. □

Acknowledgments

The authors would like to thank the Editor, Domenico Marinucci, the Associate Editor, and two anonymous referees whose comments and suggestions helped improve the manuscript.

References

[1] Ai, C. and X. Chen (2003). Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica 71*(6), 1795–1843. MR2015420

[2] Aitchison, J. and C. G. G. Aitken (1976). Multivariate binary discrimination by the kernel method. *Biometrika 63*(3), 413–420. MR0443222

[3] Andrews, D. W. K. (2017). Examples of L^2-Complete and Boundedly-Complete Distributions. *Journal of Econometrics 199*(2), 213–220. MR3681027

[4] Babii, A. and J.-P. Florens (2017). Is Completeness Necessary? Estimation in Non-identified Linear Models. *Mimeo-UNC Chapel Hill*.

[5] Beran, R. (1981). Nonparametric Regression with Randomly Censored Survival Data. Technical report, University of California Berkeley.

[6] Beyhum, J., J.-P. Florens, and I. Van Keilegom (2021). Nonparametric Instrumental Regression With Right Censored Duration Outcomes. *Journal of Business & Economic Statistics Forthcoming*.

[7] Bhatia, R. and K. Sinha (1994). Variation of Real Powers of Positive Operators. *Indiana Univ. Math. J. 43*, 913–925. MR1305952

[8] Blanchard, G., M. Hoffmann, and M. Reiß (2018). Optimal Adaptation for Early Stopping in Statistical Inverse Problems. *SIAM/ASA Journal on Uncertainty Quantification 6*(3), 1043–1075. MR3829522

[9] Blundell, R., X. Chen, and D. Kristensen (2007). Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves. *Econometrica 75*(6), 1613–1669. MR2351452

[10] Brakenhoff, T. B., M. van Smeden, F. L. J. Visseren, and R. H. H. Groenwold (2018, 02). Random measurement error: Why worry? An example of cardiovascular risk factors. *PLOS ONE 13*(2), 1–8.

[11] Carrasco, M., J.-P. Florens, 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*, pp.5633–5751. Elsevier.

[12] Centorrino, S. (2016). Data-Driven Selection of the Regularization Parameter in Additive Nonparametric Instrumental Regressions. *Mimeo - Stony Brook University*.

[13] Centorrino, S., F. Fève, and J.-P. Florens (2017). Additive Nonparametric Instrumental Regressions: a Guide to Implementation. *Journal of Econometric Methods 6*(1). MR3602594

[14] Centorrino, S., F. Fève, and J.-P. Florens (2019). Nonparametric Instrumental Regressions with (Potentially Discrete) Instruments Independent of the Error Term. *Mimeo - Stony Brook University*.

[15] Centorrino, S. and J.-P. Florens (2021). Nonparametric Instrumental Variable Estimation of Binary Response Models with Continuous Endogenous Regressors. *Econometrics and Statistics 17*, 35–63. MR4214845

[16] Centorrino, S. and J. S. Racine (2017). Semiparametric Varying Coefficient Models with Endogenous Covariates. *Annals of Economics and Statistics* (128), 261–295.

[17] Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014). Local Identification of Nonparametric and Semiparametric Models. *Econometrica 82*(2), 785–809. MR3191719

[18] Chen, X. and T. Christensen (2018). Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression. *Quantitative Economics 9*(1), 39–84. MR3789729

[19] Chen, X. and D. Pouzo (2012). Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals. *Econometrica 80*(1), 277–321. MR2920758

[20] Chen, X. and M. Reiss (2011). On Rate Optimality for Ill-Posed Inverse Problems in Econometrics. *Econometric Theory 27*(3), 497–521. MR2806258

[21] Chernozhukov, V. and C. Hansen (2005). An IV Model of Quantile Treat-

ment Effects. *Econometrica 73*(1), 245–261. MR2115636
[22] Dabrowska, D. M. (1989). Uniform Consistency of the Kernel Conditional Kaplan-Meier Estimate. *Ann. Statist. 17*(3), 1157–1167. MR1015143
[23] Dabrowska, D. M. (1992). Variable Bandwidth Conditional Kaplan-Meier Estimate. *Scandinavian Journal of Statistics 19*(4), 351–361. MR1211789
[24] Darolles, S., Y. Fan, J. P. Florens, and E. Renault (2011). Nonparametric Instrumental Regression. *Econometrica 79*(5), 1541–1565. MR2883763
[25] D'Haultfoeuille, X. (2011, 5). On the Completeness Condition in Nonparametric Instrumental Problems. *Econometric Theory 27*, 460–471. MR2806256
[26] Dunker, F. (2018). Convergence of the risk for nonparametric IV quantile regression and nonparametric IV regression with full independence. Courant Research Centre: Poverty, Equity and Growth - Discussion Papers 192, Courant Research Centre PEG.
[27] Dunker, F., J.-P. Florens, T. Hohage, J. Johannes, and E. Mammen (2014). Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal of Econometrics 178*(Part 3), 444 – 455. MR3132443
[28] Engl, H. W., M. Hanke, and A. Neubauer (2000). *Regularization of Inverse Problems*, Volume 375 of *Mathematics and Its Applications*. Dordrecht: Kluwer Academic Publishers. MR1408680
[29] Fan, J. and I. Gijbels (1992, 12). Variable Bandwidth and Local Linear Regression Smoothers. *Ann. Statist. 20*(4), 2008–2036. MR1193323
[30] Fève, F., J.-P. Florens, and I. Van Keilegom (2018). Estimation of Conditional Ranks and Tests of Exogeneity in Nonparametric Nonseparable Models. *Journal of Business & Economic Statistics 36*(2), 334–345. MR3790218
[31] Florens, J., M. Mouchart, and J. Rolin (1990). *Elements of Bayesian Statistics*. Pure and Applied Mathematics. M. Dekker. MR1051656
[32] Florens, J.-P., J. Johannes, and S. Van Bellegem (2011). Identification and Estimation by Penalization in Nonparametric Instrumental Regression. *Econometric Theory 27*(3), 472–496. MR2806257
[33] Florens, J.-P., J. Johannes, and S. Van Bellegem (2012). Instrumental Regressions in Partially Linear Models. *The Econometrics Journal 15*(2), 304–324. MR2951059
[34] Florens, J.-P., J. Racine, and S. Centorrino (2018). Nonparametric Instrumental Variable Derivative Estimation. *Jounal of Nonparametric Statistics 30*(2), 368–391. MR3794398
[35] Frandsen, B. R. (2015). Treatment Effects With Censoring and Endogeneity. *Journal of the American Statistical Association 110*(512), 1745–1752. MR3449070
[36] Fridman, V. M. (1956). A method of successive approximations for Fredholm integral equations of the first kind. *Uspeskhi, Math. Nauk. 11*, 233–334. MR0076183
[37] Gonzalez-Manteiga, W. and C. Cadarso-Suarez (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Journal of Nonparametric Statistics 4*(1), 65–78. MR1366364

[38] Hall, P. and J. L. Horowitz (2005). Nonparametric Methods for Inference in the Presence of Instrumental Variables. *Annals of Statistics 33*(6), 2904–2929. MR2253107

[39] Hanke, M., A. Neubauer, and O. Scherzer (1995, Nov). A convergence analysis of the Landweber iteration fornonlinear ill-posed problems. *Numerische Mathematik 72*(1), 21–37. MR1359706

[40] Hansen, B. E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory 24*(03), 726–748. MR2409261

[41] Horowitz, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica 79*(2), 347–394. MR2809374

[42] Horowitz, J. L. and S. Lee (2007). Nonparametric Instrumental Variables Estimation of a Quantile Regression Model. *Econometrica 75*(4), 1191–1208. MR2333498

[43] Hughes, A. and M. Kumari (2017). Unemployment, underweight, and obesity: Findings from Understanding Society (UKHLS). *Preventive Medicine 97*, 19 – 25.

[44] Johannes, J., S. Van Bellegem, and A. Vanhems (2013, January). Iterative Regularization in Nonparametric Instrumental Regression. *Journal of Statistical Planning and Inference 143*(1), 24–39. MR2969008

[45] Kaltenbacher, B., A. Neubauer, and O. Scherzer (2008). *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. Berlin, Boston: De Gruyter. MR2459012

[46] Kaplan, E. L. and P. Meier (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association 53*(282), 457–481. MR0093867

[47] Kim, C., B. U. Park, W. Kim, and C. Lim (2003, Jun). Bezier curve smoothing of the Kaplan-Meier estimator. *Annals of the Institute of Statistical Mathematics 55*(2), 359–367. MR2001869

[48] Koul, H., V. Susarla, and J. V. Ryzin (1981, 11). Regression analysis with randomly right-censored data. *Ann. Statist. 9*(6), 1276–1288. MR0630110

[49] Kress, R. (1999). *Linear integral equations*. Applied mathematical sciences. Springer-Verlag. MR1723850

[50] Landweber, L. (1951). An iterative formula for Fredholm integral equations of the first kind. *American Journal of Mathematics 73*, 615–624. MR0043348

[51] Lee, C. Y., T. A. Ledoux, C. A. Johnston, G. X. Ayala, and D. P. O'Connor (2019). Association of parental body mass index (BMI) with child's health behaviors and child's BMI depend on child's age. *BMC Obesity 6*(1), 11.

[52] Lehmann, E. L. and H. Scheffe (1947). On the Problem of Similar Regions. *Proceedings of the National Academy of Sciences of the United States of America 33*(12), 382–386. MR0023504

[53] Lewbel, A. and O. Linton (2002). Nonparametric Censored and Truncated Regression. *Econometrica 70*(2), 765–779. MR1913830

[54] Li, Q. and J. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press. MR2283034

[55] Li, Q. and J. S. Racine (2008). Nonparametric Estimation of Conditional

CDF and Quantile Functions with Mixed Categorical and Continuous Data. *Journal of Business & Economic Statistics 26*(4), 423–434. MR2459343

[56] Mammen, E., C. Rothe, and M. Schienle (2012, 04). Nonparametric Regression with Nonparametrically Generated Covariates. *Annals of Statistics 40*(2), 1132–1170. MR2985946

[57] Marron, J. S. and W. J. Padgett (1987, 12). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *The Annals of Statistics 15*(4), 1520–1535. MR0913571

[58] McNichols, D. T. and W. J. Padgett (1986). Mean and Variance of a Kernel Density Estimator under the Koziol-Green Model of Random Censorship. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 48*(2), 150–168. MR0905456

[59] Mielniczuk, J. (1986). Some Asymptotic Properties of Kernel Estimators of a Density Function in Case of Censored Data. *The Annals of Statistics 14*(2), 766–773. MR0840530

[60] Mozorov, V. (1967). Choice of a Parameter for the Solution of Functional Equations by the Regularization Method. *Sov. Math. Doklady 8*, 1000–1003.

[61] Müller, H.-G. (1991). Smooth Optimum Kernel Estimators Near Endpoints. *Biometrika 78*(3), pp. 521–530. MR1130920

[62] Newey, W. K. and J. L. Powell (2003). Instrumental Variable Estimation of Nonparametric Models. *Econometrica 71*(5), 1565–1578. MR2000257

[63] Rao, C. (1973). *Linear Statistical Inference and its Applications*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley. MR0440638

[64] Sant'Anna, P. H. C. (2020). Nonparametric Tests for Treatment Effect Heterogeneity With Duration Outcomes. *Journal of Business & Economic Statistics Forthcoming*. MR4272938

[65] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley. MR0595165

[66] Susarla, V. and J. V. Ryzin (1980). Large Sample Theory for an Estimator of the Mean Survival Time from Censored Samples. *Ann. Statist. 8*(5), 1002–1016. MR0585699

[67] Tikhonov, A. N. and V. Y. Arsenin (1977). *Solutions of Ill-Posed Problems*. John Wiley & Sons, Ltd. MR0455365

[68] Van Keilegom, I. and N. Veraverbeke (1997). Estimation and Bootstrap with Censored Data in Fixed Design Nonparametric Regression. *Annals of the Institute of Statistical Mathematics 49*(3), 467–491. MR1482368

[69] von Hinke, S., G. D. Smith, D. A. Lawlor, C. Propper, and F. Windmeijer (2016). Genetic markers as instrumental variables. *Journal of Health Economics 45*, 131 – 148.

[70] Wei, B., L. Peng, M.-J. Zhang, and J. P. Fine (2021). Estimation of causal quantile effects with a binary instrumental variable and censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) Forthcoming*. MR4294544