ESTIMATING TREATMENT EFFECTS WITH A NONPARAMETRIC RANDOM COEFFICIENTS SELECTION EQUATION

ERIC GAUTIER AND STEFAN HODERLEIN

Abstract. In this paper we study a binary treatment model where the outcome equation is of unrestricted form, and the selection equation contains multiple unobservables that enter through a nonparametric random coefficients specification. This specification is flexible because it allows for complex unobserved heterogeneity of economic agents and non-monotone selection into treatment. Employing continuous instruments, we establish that both the marginal conditional distributions of Y_0 and Y_1, the outcome for the untreated, respectively treated, given first stage random coefficients, are identified. We can thus identify an average treatment effect, conditional on first stage unobservables called U-CATE, which yields most treatment effects parameters that depend on averages, like ATE and ATT. Moreover, we provide sharp bounds on the variance, the joint distribution of Y_0, Y_1 and the distribution of treatment effects. When the outcomes are continuously distributed, we provide novel and weak conditions that allow to point identify the conditional distribution of Y_0, Y_1, given the unobservables. This allows to derive every treatment effect parameter, *e.g.* the distribution of treatment effects and the proportion of individuals who benefit from treatment. Finally, we present estimators together with their rates of convergence and we evaluate them in a simulation study.

Date: First Version: September 2010 ; This version: September 27, 2012.
Keywords: Treatment Effects, Endogeneity, Random Coefficients, Nonparametric Identification, Partial Identification, Roy Model, Ill-Posed Inverse Problems, Deconvolution, Radon Transform.

We are grateful to seminar participants at Boston College, Chicago, CREST, Harvard-MIT, Kyoto, Nanterre, Northwestern, Oxford, Princeton, UCL, 2011 CIRM New Trends in Mathematical Statistics, 2012 Bates White, CLAPEM, SCSE, and the World Congress in Probability and Statistics conferences and Arnaud Debussche for helpful comments. The authors are very grateful to Helen Broome, who is coauthor of a companion paper considering the evaluation of the returns to college education, for her many useful remarks and assistance in the simulation study.

1. Introduction

In this paper we consider estimation of treatment effect parameters in the presence of multiple sources of unobserved heterogeneity in the selection equation. We consider the following treatment effect model

$$Y = Y_0 + \Delta D, \quad \text{where } \Delta = Y_1 - Y_0, \tag{1.1}$$

$$D = \mathbf{1}\left\{V - Z'\Gamma - \Theta > 0\right\}. \tag{1.2}$$

To understand structure and limitations of this model, consider first the outcome equation (1.1): Here, Y_0 and Δ denote scalar random coefficients in a model with a binary endogenous regressor D. Note that this outcome equation allows for unrestricted heterogeneity, since it is equivalent to a nonseparable model $Y = \psi(D, U)$, with U being a (potentially infinite) vector of unobserved variables. The random coefficients have an interpretation: Y_0 is the outcome in the control group or base state, Y_1 the outcome in the treated group, and $\Delta = Y_1 - Y_0$ is the net gain from an ideal exogenous move of an individual from state 0 to state 1, called the effect of treatment. In our model, both Y_0 and Y_1 may be continuous or discrete, and individuals are observed in either of the two states 0 or 1.

In this model, participation in the treatment is endogenous: Individuals select themselves into treatment, if the net (expected) utility of participating in the treatment is positive, as formalized in equation (1.2), where $\mathbf{1}$ denotes the indicator function. This net utility depends on a vector of continuous instrumental variables $(V, Z') \in \mathbb{R}^L$ which are observable to the econometrician. It also depends, in a nonseparable fashion, on unobserved parameters $(\Gamma', \Theta) \in \mathbb{R}^L$ which are allowed to vary across the population. Because the scale of the net utility cannot be identified, we set the coefficient of V to 1. This can be done if, in the original scale, the coefficient of V is positive[1].

To fix ideas, one may think of the instruments as cost factors or elements of information about the net utility of treatment, and of the random slope coefficients as reflecting the heterogeneous impact of these factors on net utility. The random intercept can be interpreted as contributions to net utility that are unobserved to the econometrician such as the anticipated gains of being treated plus possibly some random term (*e.g.*, the random intercept of the cost function). While we allow for a rich structure in terms of unobservables that goes significantly beyond the common "scalar unobservable threshold crossing" model, we remark that at the same time we place potentially restrictive structure by requiring that this model be linear in parameters, and that the unobservables in the selection

[1] We can change V in $-V$ if it is negative.

equation have the same dimension as the vector of exogenous variables. While we do not literally believe in an exact linear structure on individual level, we think of a linear index structure as a good first order approximation. Rather than aiming to capture higher order terms in observable variables, in this paper we rather want to place emphasis on the dependence of the participation decision on an unobserved structure in a way that is more in line with structural economics.

The main results in this paper establish that under very general conditions,

$$f_{\Gamma',\Theta},\ f_{Y_j|\Gamma',\Theta}{}^{2},\ j=0,1,\quad \text{and } \mathbb{E}[\Delta \mid \Gamma=\gamma,\Theta=\theta], \tag{1.3}$$

are point identified. Moreover, we provide sharp bounds on $Var[\Delta]$ and F_Δ. Finally when the outcome is continuous, under additional conditions we show that $Var[\Delta|\Gamma=\gamma,\Theta=\theta]$, $f_{\Delta|\Gamma,\Theta}$ and $f_{Y_0,Y_1|\Gamma,\Theta}$ are point identified. Let us now elaborate on the individual objects.

The average treatment effect, conditional on first stage preference parameters, $\mathbb{E}[\Delta|\Gamma=\gamma,\Theta=\theta]$, abbreviated U-CATE, is similar in spirit to the marginal treatment effect (MTE, see Björklund and Moffitt (1987), and Heckman and Vytlacil (2005)) and shares many of its appealing properties (policy invariance, interpretation in terms of willingness to pay for people at the margin of indifference, averages like the average treatment effect (ATE) are straightforwardly derived, etc.).

It differs in as far as instead of depending on a single first stage unobservable, it allows to condition on the entire vector Γ,Θ of heterogenous first stage parameters. Unlike the scalar unobservable threshold crossing model, the selection equation neither imply monotonicity nor uniformity (see Imbens and Angrist (1994) and Heckman and Vytlacil (2005)), as soon as $L\geq 2$. It thus allows for more general heterogeneity patterns in the selection equation. In particular, there may be both compliers and defiers in the population. Because of this generality, the model (1.2) is suggested in Heckman and Vytlacil (2005) as a benchmark nonseparable, nonmonotonic model.

The marginals, $f_{Y_j|\Gamma',\Theta}f_{\Gamma',\Theta},\ j=0,1$, are identified under the same conditions as U-CATE without appealing to randomized experiments or selection on unobservables. From the marginals we can recover the unconditional marginals, as well as many other parameters, like the quantile treatment effect (QTE), see Abadie, Angrist and Imbens (2000) and Chernozhukov and Hansen (2005). We also derive sharp bounds for (1) the variance of treatment effect (VaTE), (2) the CDF of the two potential outcomes, and (3) the CDF of treatment effects.[3]

[2]Throughout this paper, we will refer to the density and cumulative distribution function (CDF for short) of a vector A as f_A, respectively F_A, we will write $f_{A|B}(\cdot|b)$ and $F_{A|B}(\cdot|b)$ the conditional densities and CDF of A given $B=b$.

[3]The bounds for (2) stem from the classical bounds of Hoeffding (1940) and Frechet (1951) and are used in Heckman, Smith and Clements (1997), Manski (1997) and Heckman and Smith (1998). Fan and Park (2010) and Firpo and Ridder

These bounds are potentially wide, and obtaining point identification under plausible conditions is thus desirable. To point identify the conditional variance of treatment effects (U-CVaTE) $Var[\Delta|\Gamma=\gamma,\Theta=\theta]$, and thus also the unconditional variance, we impose the additional assumption that conditional on (Γ',Θ), Y_0 and Δ are uncorrelated. For the Unobservables Conditioned Distribution of Treatment Effects (U-CDiTE) $f_{\Delta|\Gamma',\Theta}$ we impose the stronger assumption that Y_0 and Δ are independent conditional on Γ,Θ (which we denote as $Y_0 \perp \Delta|\Gamma,\Theta$)[4]. Conditional independence assumptions between the gain Δ and the base state Y_0 are made in Heckman, Smith and Clements (1997), Heckman and Clements (1998). However, in these references the independence assumption is conditional on D, or on observable variables X. In contrast, one attractive feature of the approach put forward in this paper is that the independence is conditional on the unobservables entering the selection equation (as well as control variables X). This makes this assumption more likely to hold, as we argue in detail in Section 3.4 using extensions of the Roy model. The unobservables in the selection equation contain information on ex-ante forecast of the gains and cost factors. At this point, we would only like to point out that this assumption is satisfied, if there exist otherwise unrestricted mappings ψ_0,ψ_Δ such that $Y_0=\psi_0(\Gamma',\Theta,U_0)$, and $\Delta=\psi_\Delta(\Gamma',\Theta,U_\Delta)$, with U_0,U_Δ possibly infinite dimensional, such that $U_0 \perp U_\Delta \perp (\Gamma',\Theta)$.[5] This paper therefore shows that allowing for several sources of unobserved heterogeneity in the selection equation is important, not just in its own right, but also to allow for more dependence between the gain Δ and the base state Y_0, and still point identify distributional treatment effect parameters.

Another material assumption is a support condition. When L, the dimension of instruments is 2 or larger, we impose a condition relating the support of the instruments and that of the unobserved heterogeneity parameters. Though we can deal with instruments with bounded support, it is a type of "large" support assumption which is required to fully recover the entire distribution of random

(2008) apply the Makarov (1981) bounds to infer bounds on the distribution of treatment effects, Fan and Zhu (2009) obtain bounds for functionals of the joint distribution of potential outcomes, like inequality measures based on the distribution of treatment effects. Unlike these references the bounds are obtained in the case of possible selection on observables and unobservables. As we shall see, the bounds on the distribution of treatment effects are also sharper than what a direct application of the Makarov bounds would yield because they are averages of Makarov bounds on CDFs of the marginals conditional on the observables and unobservables of equation (1.2).

[4] We indeed assume $Y_0 \perp \Delta|\Gamma,\Theta,X$ with some observables X that are used as control variables. This can make the independence assumption more likely in the same spirit as the missing at random assumption in the missing data literature. However, in total, we are able to condition both on X and the unobserved heterogeneity Γ,Θ.

[5] This can be interpreted as the fact that the selection equation reveals information about the common endogenous factors; there is (potentially complicated) remaining heterogeneity in Y_0 and Δ, but it is independent of everything else.

coefficients in the selection equation. When $L = 1$, we establish that when the variation of our instruments is small relative to that of the unobserved heterogeneity parameters, we can identify the average, variance or distribution of effects for the subpopulation defined by the range of the instrument. This subpopulation is related to the one considered in Angrist, Graddy and Imbens (2000), the LATE of Imbens and Angrist (1994) being a special case. We suspect that something similar is feasible for $L \geq 2$, but leave it for future research.

Based on these identification results, we provide sample counterparts estimators and obtain their rates of convergence. It is known from Beran, Feuerverger and Hall (1996), Hoderlein, Klemelä and Mammen (2010) and Gautier and Kitamura (2009) that the estimation of the distribution of random coefficients in the single equation case with exogenous regressors is an ill-posed inverse problem. We extend these papers by allowing dependence between the random coefficients and the regressors and by relaxing the full support assumption on the regressors[6]. Similar to Imbens and Newey (2009), we deal with an endogenous regressor D in equation (1.1) through a triangular system with an equation for the endogenous regressor but we allow for multiple sources of unobserved heterogeneity entering in a nonseparable fashion in (1.2).

In contrast to all of these references, in our approach the first stage (selection) equation allows for multiple sources of unobserved heterogeneity, and monotonicity is not imposed. Estimation of the marginals $(Y_0 + \Delta, \Gamma', \Theta)$, (Y_0, Γ', Θ) and (Γ', Θ), U-CATE, VaTE or extensions of the QTE, then relies on solving ill-posed inverse problems that involve the Radon transform. The estimation of the distribution of treatment effects is a deconvolution problem and features thus in addition another ill-posed inverse problem (see also Heckman, Smith and Clements (1997) and Heckman and Smith (1998)). More precisely, in our setup, it consists of a conditional deconvolution problem with unknown but estimable distributions of (1) the signal plus error and of (2) the error. Evdokimov (2010) considers conditional deconvolution in nonparametric panel data models with unobserved heterogeneity. In classical deconvolution, the density of Y_0 is known and the characteristic function of $Y_0+\Delta$ is estimated via the empirical characteristic function which estimates the true characteristic function at rate $1/\sqrt{N}$. An extension studied in the statistics literature considers the case where the density of Y_0 is estimable at rate $1/\sqrt{N}$ using a preliminary sample (see, *e.g.* Neumann (1997), Johannes (2009) and Comte and Lacour (2011)). In this paper the Fourier transforms of the densities of $(Y_0 + \Delta, \Gamma', \Theta)$ and (Y_0, Γ', Θ) with respect to the first argument are estimated solving inverse problems using the same

[6] In (1.1) D and (Y_0, Δ) are dependent and D has a limited support, in (1.2) we will allow V to have limited support and (V, Z') and (Γ', Θ) to be dependent but independent given control variables.

sample. Getting unconditional parameters requires to compute an integral over the conditional effects weighted by the joint density of the random coefficients.

More generally, this paper touches upon two related sets of literatures. The first is the treatment effect literature, in particular the part that is related to distributional treatment effects, the second is the random coefficients literature. First, we obtain results for treatment effect parameters that depend on averages. This is related to the important contributions of LATE (Imbens and Angrist(1994)) and MTE (Björklund and Moffitt (1987), Heckman and Vytlacil (1999, 2005, 2007)). But we also obtain results on the marginals which are related to the quantile treatment effects of Abadie, Angrist and Imbens (2002), Chernozhukov and Hansen (2005), and to results by Heckman, Smith and Clements (1997), Heckman and Smith (1998), Carneiro, Hansen and Heckman (2003), Aakvik, Heckman and Vytlacil (2005), Abbring and Heckman (2007) and Fan and Zhu (2009), Fan and Park (2010) and Firpo and Ridder (2008). Note that the first two results on quantile treatment effect essentially require a rank invariance assumption, *i.e.*, the individuals retain their ordering both in the treatment and the control group, an assumption which may only be weakened slightly. This assumption is restrictive, and has been criticized, see Heckman, Smith and Clements (1997). Carneiro, Hansen and Heckman (2003) and Aakvik, Heckman and Vytlacil (2005) consider a factor model approach that allows to identify the distribution of treatment effects. Because we implicitly recover the distribution of unobservables in the selection equation, our identification strategy has some similarities with that of Lewbel (2000) and Lewbel (2007) for the semi-parametric binary choice and endogenous selection model, but allows to handle multiple sources of unobserved heterogeneity that enter the index in the selection equation in a nonseparable, linear random coefficients fashion. Moffitt (2008) considers a model with multiple sources of unobserved heterogeneity in the selection equation, which enter in a more general form than the linear index structure of equation (1.2), but he imposes monotonicity. Klein (2010) considers a different specification where a second source of unobserved heterogeneity enters the scalar unobservable threshold crossing in a nonseparable fashion. The two sources of unobserved heterogeneity are independent and it is observed that recovering the distribution of unobserved heterogeneity is a nonlinear inverse problem.

The second related line of work is the literature on nonparametric random coefficients models. Random coefficients models allow the preference or production parameters to vary across the population. In this paper, we specifically allow for different individuals to have different costs and benefits of treatment. We emphasize the nonparametric aspect of our analysis, which allows to be flexible about the form of unobserved heterogeneity. References in econometrics include Elbers and Ridder (1982), Heckman and Singer (1984), Beran and Hall (1992), Beran, Feuerverger and Hall (1996), Ichimura and Thompson (1998), Fox and Gandhi (2009), Hoderlein, Klemelä and Mammen (2010), Gautier and Kitamura (2009) and Gautier and Le Pennec (2011). Fox and Gandhi (2009) study identification of the distribution of unobserved heterogeneity in Roy models, however they focus on the case of discretely supported random coefficients, and do not allow for a random intercept in the selection equation. The last three references focus on continuous random coefficients, and recognize that the estimation of the density of the latent random coefficients vector is a statistical inverse problem in this scenario. The literature on the treatment of these problems is extensive in statistics and econometrics, and we refer to Carrasco, Florens and Renault (2007) for a survey of applications in economics.

This paper is structured as followed. In section 2, we introduce the model formally, and discuss the basic assumptions. In section 3 we establish the main identification results. We show that under the baseline assumptions the marginals of each potential outcome are identified, conditional on random coefficients. This allows to obtain the U-CATE, many other treatment effect parameters that only depend on averages, and other parameters that solely depend on marginals, like the QTE. Moreover, we obtain bounds on the variance of treatment effects, on the joint distribution of the two potential outcomes, and consequently also on the distribution of treatment effects. Finally, in section 3.4, we introduce two nested assumptions that allow to point identify the variance and the distribution of treatment effects. We discuss the interpretation of these assumptions and their relevance in the context of extensions of the Roy model, discuss how they aid in identification, and present again sample counterparts estimators. In section 4, we present estimators and obtain general rates of convergence for our estimators. In section 5, we analyze the finite sample behavior through a simulation study, before an outlook concludes.

2. The Random Coefficients Model and Assumptions

This section introduces the formal setup in which we analyze the effects of treatment. We distinguish between two cases. The first one is the case where we have a vector of unobservables, and is the core innovation in this paper. The second one is the "traditional" case, which features a scalar unobservable and is displayed largely for comparison. Before we discuss these scenarios in detail, we start, however, by introducing some crucial pieces of notation and basic probabilistic assumptions.

Throughout this paper, we use uppercase letters for random variables and lowercase for their realizations. In addition to the observable variables (Y, D, V, Z) which have already been introduced above, we assume that there might be another observable random vector X on which we may want to condition upon when doing inference. Examples include household characteristics like age, gender, race etc. We do not impose any restriction on the dependence of Y_0, Y_1, V, Z, Γ, Θ on X, besides regularity conditions which we detail below. As with Y, we denote by X_0 and X_1 the random variable X when D is 0, respectively 1.

The data consists of the realizations of N independent and identically distributed copies of the population random variables, which we denote as $(y_i, d_i, v_i, z_i', x_i')_{i=1,\dots,N}$, where N is the sample size. We denote by $supp\,(A)$ or $supp\,(f_A)$, the support of the random vector A and by $Int(A)$ the interior of a set A.

For mathematical convenience, in the case where $L \geq 2$, it is useful to renormalize the index in (1.2). This could be done in several ways. In this paper, we assume that the (random) coefficient of V in the original net utility scale has a sign. A more general sufficient condition is presented in Section 6. Other than being more interpretable, the approach put forward in this paper allows us to handle the case where V has bounded support, too. As a next step, we divide by $\|(Z',1)\|$, and use the notations $\widetilde{S} = (Z',1)'/\|(Z',1)\|$, $\widetilde{V} = V/\|(Z',1)\|$ and $\widetilde{\Gamma} = (\Gamma',\Theta)'$. Then, (1.2) becomes

$$D = \mathbf{1}\{\widetilde{S}'\widetilde{\Gamma} < \widetilde{V}\}. \tag{2.1}$$

We invoke the following assumptions (when $L = 1$ simply drop Z and Γ below).

Assumption 2.1. (A-1) The conditional distribution of (V, Z, Θ, Γ') given $X = x$ is absolutely continuous with respect to the Lebesgue measure for almost every x in $supp\,(X)$;

(A-2) $(V,Z)\perp(Y_0,\Gamma',\Theta)|X$ and $(V,Z)\perp(Y_1,\Gamma',\Theta)|X$;

(A-3) $0 < \mathbb{P}(D=1|X) < 1 \quad a.s.$;

(A-4) $X_0 = X_1 \quad a.s.$.

Assumption (A-1) defines the setup of this paper as one with continuous instruments. In Sections 3.4.3, 4.3.3 and 4.3.4, we strengthen (A-1) to

(A-1') The conditional distribution of $(Y_0, Y_1, V, Z, \Theta, \Gamma')$ given $X = x$ is absolutely continuous with respect to the Lebesgue measure for almost every x in $supp\,(X)$.

It is only in these sections that we consider the particular case where the outcomes are continuous. The responses Y_0 and Y_1 are allowed to be heterogeneous in a general way and of the form $Y_0 = \psi_0(X,\Gamma',\Theta,U_0)$ and $Y_1 = \psi_1(X,\Gamma',\Theta,U_1)$ where U_0 and U_1 account for unobserved heterogeneity and can be infinite dimensional. Assumption (A-1) also implies an exclusion restriction: conditional on

$X = x$, Z is still continuous and has variation. In practice, this is achieved when Z contains variables that are not in the list of regressors X.

Assumption (A-2) requires the instruments to be independent of the random parameters given some variables X which are either exogenous or act as control variables. We allow (Y_0, Γ', Θ) and (Y_1, Γ', Θ) to depend on (V, Z) (unconditional endogeneity of the instruments) as long as we have at hand control variables X which yield independence. Randomization or pseudorandomization is a classical tool to generate instrumental variables satisfying this assumption unconditional on X. When $L \geq 2$, Assumption (A-2) can be written in terms of the renormalized instruments as

$$(\widetilde{V}, \widetilde{S}) \perp (Y_0, \widetilde{\Gamma}')|X \quad \text{and} \quad (\widetilde{V}, \widetilde{S}) \perp (Y_1, \widetilde{\Gamma}')|X.$$

Assumption (A-3) states that for any $x \in X$, there is always a fraction of the population that participates in treatment, and one that does not. Finally, Assumption (A-4) states that X is not caused by the treatment. As in Heckman and Vytlacil (2005), this last assumption is not strictly required for econometric analysis. However, it makes the inferred quantities more interpretable, and allows to still capture the total effects of D on Y, after conditioning on X.

It is possible, when considering only one unobservable, to consider a model where (1.2) is replaced by

$$D = \mathbf{1}\{\mu(V, Z) > \Theta\} \tag{2.2}$$

where μ is a CDF and $\Theta|X \sim \mathcal{U}(0, 1)$. This is a well established model studied, among others, in Heckman and Vytlacil (1998, 2005, 2007). We do not present the extension of their results in terms of variance and distribution of treatment effects in this text in order to make the presentation more concise. The arguments are closely related to the single unobservable case used for purpose of comparison below. This one unobservable framework extends the IV framework where general heterogeneity in response but not in choices is allowed. Additive separability has a strong implication called "monotonicity" in Imbens and Angrist (1994) and "uniformity" in Heckman and Vytlacil (2005): conditional on $X = x$ where x belongs to $supp(X)$, for arbitrary (v, z) and (v', z') in $supp(V, Z')$, if the instruments are moved for everyone from (v, z) to (v', z') then either for every $\theta \in [0, 1]$, $\mathbf{1}\{\mu(v, z) > \theta\} \geq \mathbf{1}\{\mu(v', z') > \theta\}$ or for every $\theta \in [0, 1]$, $\mathbf{1}\{\mu(v, z) > \theta\} < \mathbf{1}\{\mu(v', z') > \theta\}$, *i.e.*, there are either no compliers, or no defiers.

This is a substantial restriction regarding heterogeneity in choices of treatment. The case with more than two sources of unobserved heterogeneity in D that we advocate in this paper allows for more general heterogeneity in treatment choices in which uniformity breaks down. To see this, fix $v \in supp(\widetilde{V})$ and take s and s' in $supp\left(f_{\widetilde{S}|\widetilde{V}}(\cdot|v)\right)$ and denote by $D_s(\gamma) = \mathbf{1}\{s'\gamma < v\}$. In Figure 1 we consider the case where $L = 2$, $v = 0$ and thus the origin is where the two lines (defined through their normal vectors s and s') intersect. For an unobserved heterogeneity parameter γ in zone 2, $D_s = 0$ and $D_{s'} = 1$ while, for γ in zone 4, $D_s = 1$ and $D_{s'} = 0$, *i.e.*, parts of the population may be compliers, parts defiers.

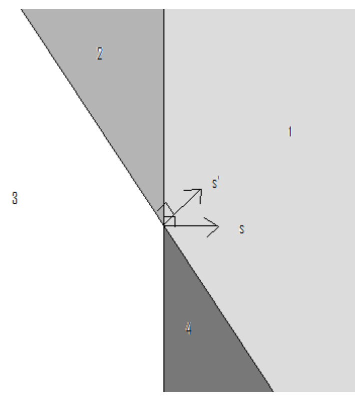

FIGURE 1. Non uniformity with more that 2 unobservables

Assumptions (A-1) and (A-2) together yield that the score function satisfies

$$\pi(v, x) = \int_{-\infty}^{v} f_{\Theta|X}(t|x)dt \tag{2.3}$$

when $L = 1$. In the case when $L \geq 2$,

$$\pi(s, v, x) = \mathbb{P}(D = 1|\widetilde{S} = s, \widetilde{V} = v, X = x) \tag{2.4}$$

$$= \int_{\mathbb{R}^L} \mathbf{1}\{\gamma' s < v\} f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma \tag{2.5}$$

$$= \int_{-\infty}^{v} \int_{P_{s,u}} f_{\widetilde{\Gamma}|X}(\gamma|x)d_{P_{s,u}}(\gamma)du \tag{2.6}$$

$$= \int_{-\infty}^{v} R[f_{\widetilde{\Gamma}|X}(\cdot|x)](s, u)du. \tag{2.7}$$

Here, R is called the Radon transform (see, *e.g.*, Helgason (1999)), and $P_{s,u} = \{\gamma : \ \gamma' s = u\}$ is the affine hyperplane of dimension $L - 1$ defined through the direction s and distance u to the origin, where s is in an hemisphere H of $\mathbb{S}^{L-1}$ (here $H^+ = \{x \in \mathbb{S}^{L-1} : \ x_L \geq 0\}$), and $u \in \mathbb{R}$. The Radon transform is a bounded operator. When applied to a function $f \in \mathrm{L}^1(\mathbb{R}^L)$, it yields the integral of that function on $P_{s,u}$

$$R[f](s,u)=\int_{P_{s,u}} f(\gamma)d_{P_{s,u}}(\gamma)$$

where $d_{P_{s,u}}$ is the Lebesgue measure on $P_{s,u}$. Mathematical results regarding this integral transformation (including injectivity and an inversion formula involving the adjoint of the Radon transform) can be found in Natterer (1996) and Helgason (1999). Statistical inverse problems involving this type of operator on the whole space appear in several problems in tomography (see, *e.g.*, Korostelev and Tsybakov (1993) and Cavalier (2000, 2001)), but also when one wishes to estimate the distribution of random coefficients in the linear model with random coefficients (see Beran, Feuerverger and Hall (1996) and Hoderlein, Klemelä and Mammen (2010)).

Assumption 2.2. When $L=1$, for every $x \in supp\,(X)$,

$$supp\left(f_{V|X}(\cdot|x)\right) \supset supp\left(f_{\Theta|X}(\cdot|x)\right).$$

When $L \geq 2$, for every $x \in supp\,(X)\,,\ supp\left(f_{\widetilde{S}|X}(\cdot|x)\right)=H^{+}$ and for every $s \in Int(H^{+})$,

$$supp\left(f_{\widetilde{V}|\widetilde{S},X}(\cdot|s,x)\right) \supset \left[\inf_{\gamma \in supp\left(f_{\widetilde{\Gamma}|X}(\cdot|x)\right)} s'\gamma,\ \sup_{\gamma \in supp\left(f_{\widetilde{\Gamma}|X}(\cdot|x)\right)} s'\gamma\right].$$

Assumption 2.2 is a large support assumption. It implies that the instruments have a large enough support to apprehend the whole distribution of the unobserved heterogeneity vector. This is crucial in our setup, as we want to recover the entire multivariate distribution of heterogeneity factors $\widetilde{\Gamma}$, and implicitly use this information to infer the average treatment effect from the conditional treatment effects. Assumption 2.2 is not required when $L=1$ to obtain treatment effect parameters conditional on the unobserved heterogeneity Θ, however, it is required to obtain population averages. When it is not satisfied, we can only make statements about a particular subpopulation related to the variation of the instrument, this is similar to the population apprehended by LATE of Imbens and Angrist (1994), see also Angrist, Grady and Imbens (2000). A similar assumption to Assumption 2.2 is made in Lewbel (2007). Unlike Gautier and Kitamura (2009), Assumption 2.2 still allows one of the instruments, for instance V, to have bounded support.

Finally, Heckman and Smith (1998) consider the case where in a welfare analysis one would like to consider treatment effects in terms of some social welfare criterion U which could be more general than simply the potential outcomes (*e.g.* consider a more general utility function than income if Y is income). Within our framework, it is easy to consider the case where the gains are expressed in terms $U(Y_1,X)-U(Y_0,X)$, for a known utility function U, by simply replacing everywhere, including

(1.1), Y, Y_0, Y_1 by $U(Y,X)$, $U(Y_0,X)$, $U(Y_1,X)$ and transform the variables y_i into $U(y_i,x_i)$ for $i=1,\dots,N$.

3. Identification of Structural Parameters

This section discusses identification and estimation of the distribution of random coefficients, $f_{\Gamma',\Theta}$, the marginal distribution of Y_0 and Y_1, respectively, given random coefficients, $f_{Y_j|\Gamma',\Theta}$, $j=0,1$, and several implied parameters, like U-CATE. Moreover, we show that the joint distribution and the variance of treatment effects are partially identified, and provide sharp bounds. Finally, we propose an additional assumption that allows us to point identify the variance and distribution of treatment effects, and argue that it is likely to be satisfied in many economically relevant cases.

3.1. **A Central Result for Identification.** We start again by clarifying the notation. In what follows we denote by $\overline{g}$ the extension of a function g as 0 outside its domain of definition, *e.g.* a regression function where regressors have bounded support outside of this support. Moreover, denote by R the Radon transform, and by R^{-1} its inverse. We use these objects in the following argument, which is at the core of our identification strategy.

First, note that Assumptions (A-1) and (A-2) yield that for (v,s',x') in $supp\left(\widetilde{V},\widetilde{S}',X'\right)$,

$$\begin{aligned}\mathbb{E}\left[\phi(Y)D|\widetilde{S}=s,\widetilde{V}=v,X=x\right] &= \int_{supp(\widetilde{\Gamma})}\mathbb{E}[\phi(Y_1)\mathbf{1}\{\gamma's<v\}|\widetilde{S}=s,\widetilde{V}=v,\widetilde{\Gamma}=\gamma,X=x]f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma\\ &= \int_{supp(\widetilde{\Gamma})}\mathbf{1}\{\gamma's<v\}\mathbb{E}[\phi(Y_1)|\widetilde{S}=s,\widetilde{V}=v,\widetilde{\Gamma}=\gamma,X=x]f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma\\ &= \int_{supp(\widetilde{\Gamma})}\mathbf{1}\{\gamma's<v\}\mathbb{E}[\phi(Y_1)|\widetilde{\Gamma}=\gamma,X=x]f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma.\end{aligned}$$

for any function ϕ such that $\mathbb{E}[|\phi(Y_1)|]<\infty$. Thus, by arguments from the previous section,

$$\mathbb{E}[\phi(Y)D|\widetilde{S}=s,\widetilde{V}=v,X=x]=\int_{-\infty}^{v}R\left[\overline{\mathbb{E}[\phi(Y_1)|\widetilde{\Gamma}=\cdot,X=x]}f_{\widetilde{\Gamma}|X}(\cdot|x)\right](s,u)du. \tag{3.1}$$

which yields, almost everywhere (a.e. for short) for u in $supp\left(f_{\widetilde{V}|\widetilde{S},X}(\cdot,s,x)\right)$,

$$\partial_v\mathbb{E}[\phi(Y)D|\widetilde{S}=s,\widetilde{V}=\cdot,X=x](u)=R\left[\overline{\mathbb{E}[\phi(Y_1)|\widetilde{\Gamma}=\cdot,X=x]}f_{\widetilde{\Gamma}|X}(\cdot|x)\right](s,u). \tag{3.2}$$

The right hand side of (3.2) is 0 for u outside $\left[\inf_{\gamma\in supp\left(f_{\widetilde{\Gamma}|X}(\cdot|x)\right)}s'\gamma,\sup_{\gamma\in supp\left(f_{\widetilde{\Gamma}|X}(\cdot|x)\right)}s'\gamma\right]$ (see Figure 2). Under Assumption 2.2 we hence know that extending the left hand side of (3.2) as 0 is innocuous.

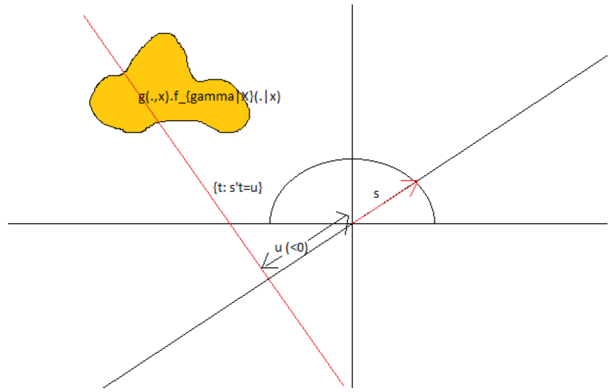

FIGURE 2. Radon transform and extensions

These arguments motivate our first theorem. To avoid tedious repetitions, we mostly display results for $L \geq 2$. The case of a scalar random coefficient is analogous. It can be obtained by leaving out the inverse of the Radon transform, and adapting the conditioning set accordingly replacing $\widetilde{V}$ and $\widetilde{\Gamma}$ by V and Θ.

Theorem 3.1. Consider an arbitrary function ϕ such that $\mathbb{E}\left[|\phi(Y_0)| + |\phi(Y_1)|\right] < \infty$. Let $L \geq 2$, and assume that Assumptions 2.1 and 2.2 hold. Then, almost surely in x in *supp*(X), the following statements are true:

$$f_{\widetilde{\Gamma}|X}(\cdot|x) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[D \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x\right]}\right] \tag{3.3}$$

$$\overline{\mathbb{E}\left[\phi(Y_1) \middle| \widetilde{\Gamma} = \cdot, X = x\right]} f_{\widetilde{\Gamma}|X}(\cdot|x) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\phi(Y) D \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x\right]}\right] \tag{3.4}$$

$$\overline{\mathbb{E}\left[\phi(Y_0) \middle| \widetilde{\Gamma} = \cdot, X = x\right]} f_{\widetilde{\Gamma}|X}(\cdot|x) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\phi(Y)(D-1) \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x\right]}\right]. \tag{3.5}$$

Discussion of Theorem 3.1: 1. This set of results always equate a structural parameter of interest on the left hand side to an object that can be estimated from data on the right hand side. It is instructive for this first result to compare them with the corresponding results that would be obtained in the scalar unobservable case ($L = 1$),

$$f_{\Theta|X}(\cdot|x) = \overline{\partial_v \mathbb{E}[D|V = \cdot, X = x]} \tag{3.6}$$

$$\overline{\mathbb{E}\left[\phi(Y_1)|\Theta = \cdot, X = x\right]} f_{\Theta|X}(\cdot|x) = \overline{\partial_v \mathbb{E}\left[\phi(Y) D|V = \cdot, X = x\right]} \tag{3.7}$$

$$\overline{\mathbb{E}\left[\phi(Y_0)|\Theta = \cdot, X = x\right]} f_{\Theta|X}(\cdot|x) = \overline{\partial_v \mathbb{E}[\phi(Y)(D-1)|V = \cdot, X = x]}. \tag{3.8}$$

All the results in the scalar unobservable case involve only one unbounded operator: a partial derivative with respect to v. In contrast, the results in the multiple unobservable case involve in addition a second unbounded operator: the inverse of the Radon transform, thus showing the more complex ill-posed inverse nature of estimation in this setup. More specifically, in the single unobservable case, the density of the scalar random coefficient Θ, conditional on exogenous factors X, is identified by the derivative of the propensity score with respect to v, see equation (3.6). In contrast, in order to recover the (conditional) density of $\widetilde{\Gamma}$ in the case of multiple unobservables, one has to also apply the inverse Radon transform to recover a similar object, see equation (3.3). The same remark applies to the comparison between equations (3.7) and (3.8) in the single unobservable case, and their counterparts (equations (3.4) and (3.5) respectively) in the multiple unobservable case: it is always the inverse Radon transform we have to apply. By trivial manipulations, using the identity for ϕ, we can recover a Heckman and Vytlacil (1998, 2005, 2007) type result,

$$\overline{\mathbb{E}\left[\Delta|\Theta=\cdot,X=x\right]}=\frac{\overline{\partial_v\mathbb{E}\left[Y|V=\cdot,X=x\right]}}{\overline{\partial_v\mathbb{E}[D|V=\cdot,X=x]}},$$

and can obviously provide an analog in the case of several unobservables. Because of its paramount importance, we focus on the discussion of this conditional average treatment effect in a separate section below.

2. Another important point to notice in Theorem 3.1 is the wide range of functions ϕ that can be used. Using for example $\phi(y)=\mathbf{1}\left\{(-\infty,y]\right\}$ allows to obtain (partial) CDFs, while using $\phi(t)=\exp(ity)$ allows to obtain partial Fourier transforms, which can be employed to recover densities. This is illustrated in the following corollary.

Corollary 3.1. Let Assumptions 2.1 and 2.2 be true, the marginal distributions of $(Y_0,\widetilde{\Gamma})$ and $(Y_1,\widetilde{\Gamma})$ given $X=x$ are identified. Integrating out $\widetilde{\Gamma}$, we obtain

$$F_{Y_1|X}(y|x)=\int_{\mathbb{R}^L}R^{-1}\left[\overline{\partial_v\mathbb{E}\left[\mathbf{1}\left\{Y\le y\right\}D\,\middle|\left(\widetilde{S},\widetilde{V}\right)=\cdot,X=x\right]}\right](\gamma)d\gamma,$$

and analogously for $F_{Y_0|X}(y|x)$.

From these CDFs we can obtain any inequality measure (*e.g.* the Gini index) for the outcome in the treated and control group. It is important to notice that these quantities are obtained without an ideal randomized experiment. As one example, we may obtain the quantile treatment effect (QTE) of Abadie, Angrist and Imbens (2002), and Chernozhukov and Hansen (2005), which is defined as,

$$QTE(x,\tau)=q(1,x,\tau)-q(0,x,\tau)$$

where $q(1,x,\tau)$ and $q(0,x,\tau)$ are the quantiles of $F_{Y_1|X}(y|x)$ and $F_{Y_0|X}(y|x)$, as well as the related average effect $\int_0^1 QTE(x,\tau)d\tau$, by simply inverting the CDFs obtained from Corollary 3.1.

3. When $L=1$, because we do not have R^{-1} in the formulas, Assumption 2.2 is not necessary and it is possible to consider cases where the support of the instrument V is not rich enough to provide identification for every value of the unobservable.

Proposition 3.1. Let Assumption 2.1 hold. For almost every x in $supp(X)$ such that $\mathbb{P}\left(\Theta \in supp\left(f_{V|X}(\cdot|x)\right)|X=x\right)>0$, we obtain that for every function ϕ s. th. $\mathbb{E}\left[|\phi(Y_0)|+|\phi(Y_1)|\right]<\infty$,

$$f_{\Theta|X,\Theta\in supp\left(f_{V|X}(\cdot|x)\right)}(\cdot|x)=\frac{\overline{\partial_v\mathbb{E}[D|(V=\cdot,X=x)]}}{\int_{\mathbb{R}}\overline{\partial_v\mathbb{E}[D|(V=\cdot,X=x)]}(t)dt} \tag{3.9}$$

$$\overline{\mathbb{E}\left[\phi(Y_1)|\Theta=\cdot,X=x\right]}f_{\Theta|X,\Theta\in supp\left(f_{V|X}(\cdot|x)\right)}(\cdot|x)=\frac{\overline{\partial_v\mathbb{E}\left[\phi(Y)D|V=\cdot,X=x\right]}(t)}{\int_{\mathbb{R}}\overline{\partial_v\mathbb{E}\left[D|V=\cdot,X=x\right]}(t)dt}, \tag{3.10}$$

and analogously for $\overline{\mathbb{E}\left[\phi(Y_0)|\Theta=\cdot,X=x\right]}$. This result allows to identify a large variety of parameters of interest, starting with marginals of potential outcomes

$$F_{Y_0|\Theta\in supp\left(f_{V|X}(\cdot|x)\right),X}(y|x)=\frac{\int_{supp\left(f_{V|X}(\cdot|x)\right)}\partial_v\mathbb{E}[\mathbf{1}\left\{Y\le y\right\}(D-1)|V=\cdot,X=x](t)dt}{\int_{supp\left(f_{V|X}(\cdot|x)\right)}\partial_v\mathbb{E}[D|V=\cdot,X=x](t)dt}.$$

and including average or quantile treatment effects, variance and distribution of treatment effects (under the respective assumptions to point identify theses quantities that we make below). The subpopulation for which we can make inference is the same as in Angrist, Grady and Imbens (2000). We conjecture that an analog result holds in the case of a multivariate unobservable.

To close this discussion, we remark that the equalities in Theorem 3.1 hold a.e., for convenience we will no longer mention this in the remainder of the paper. In the following, we discuss a key implication of Theorem 3.1. Due to its great importance, we state it in a separate subsection.

3.2. **The Average Treatment Effect Conditional on Unobservables (U-CATE).** In this section we extend the notion of the MTE in Björklund and Moffitt (1987) and Heckman and Vytlacil (1998, 2005, 2007) to our setup with potentially several unobservables. In the presence of only one unobservable in the selection equation, we call the parameter

$$U-CATE(\theta,x):=\mathbb{E}[\Delta|\Theta=\theta,X=x],$$

the Unobservables Conditioned Average Treatment Effect (U-CATE, for short). It is the average effect of treatment for the subpopulation with unobserved heterogeneity parameter equal to θ and observable $X=x$. The parameter $U-CATE(\theta,x)$ has the same interpretation as the MTE. It corresponds to the average effect for a subpopulation with $X=x$ who would be indifferent between participation and non-participation in the treatment, if they were exogenously assigned a value v of V such that $v = \theta$. The U-CATE parameter has the advantage that it can be generalized easily to the case where $L \geq 2$ as

$$U - CATE(\gamma, x) := \mathbb{E}[\Delta|\widetilde{\Gamma} = \gamma, X = x].$$

It is the average effect of treatment for the subpopulation with first stage unobserved heterogeneity vector equal to γ and observables $X = x$. It is also the average effect for people with $X = x$ who would be indifferent between participation and non-participation in the treatment, if they were exogenously assigned a value (s, v) of $(\widetilde{S}, \widetilde{V})$, such that $s'\gamma = v$. Because Assumption (A-2) yields that

$$\begin{aligned}
&\mathbb{E}\left[\Delta|\widetilde{\Gamma} = \gamma, X = x, \left(\widetilde{S}, \widetilde{V}\right) = (s, v)\right] \\
&= \mathbb{E}\left[Y_1|\widetilde{\Gamma} = \gamma, X = x, \left(\widetilde{S}, \widetilde{V}\right) = (s, v)\right] - \mathbb{E}\left[Y_0|\widetilde{\Gamma} = \gamma, X = x, \left(\widetilde{S}, \widetilde{V}\right) = (s, v)\right] \\
&= \mathbb{E}[Y_1|\widetilde{\Gamma} = \gamma, X = x] - \mathbb{E}[Y_0|\widetilde{\Gamma} = \gamma, X = x] \\
&= \mathbb{E}[\Delta|\widetilde{\Gamma} = \gamma, X = x],
\end{aligned}$$

$U - CATE(\gamma, x)$ does not depend on the values taken by the instruments $\left(\widetilde{S}, \widetilde{V}\right)$. It is thus policy invariant. All these properties are essential properties of the MTE. The extensions of MTE suggested in Heckman and Vytlacil (2005) for non additively separable models in the first stage selection equation do not satisfy these properties simultaneously so we believe that U-CATE is a natural parameter to extend the standard treatment effects analysis to more general models for the treatment choice. Moreover, like in Heckman and Vytlacil (2005), a large variety of treatment effect measures that depend on averages can be written as weighted averages of the U-CATE parameter. The ATE for example can be obtained from U-CATE under Assumption 2.2. To this end, we employ the general results in the previous subsection, with $\phi(y) = y$, $\forall y \in \mathbb{R}$ and make use of the following assumption.

Assumption 3.1. $\mathbb{E}[|Y_1| + |Y_0|] < \infty$.

In the following, we derive U-CATE formally, which is straightforward given the results above.

Theorem 3.2. Suppose assumptions 2.1 and 3.1 hold. When $L = 1$, we obtain for almost every θ,

$$\overline{U - CATE(\theta, x)} f_{\Theta|X}(\theta|x) = \partial_v \mathbb{E}\left[Y\,|V = \cdot, X = x\right](\theta). \tag{3.11}$$

Suppose now that $L \geq 2$ and invoke Assumption 2.2. Then, for almost every γ in $\mathbb{R}^L$,

$$\overline{U - CATE(\gamma, x)} f_{\widetilde{\Gamma}|X}(\gamma|x) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[Y\,\middle|\left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x\right]}\right](\gamma). \tag{3.12}$$

Obviously, from this equation it is trivial to solve for U-CATE by simply plugging in the expressions for $f_{\Theta|X}(\gamma|x)$ (respectively $f_{\widetilde{\Gamma}|X}(\gamma|x)$) from Theorem 3.1 into this equation. The right hand side of (3.12) is a natural extension of the local instrumental variable estimator (LIV for short). If $U-CATE(\gamma, x)$ is found to be not constant in one dimension of γ, it is an indication of heterogeneous "costs factors" in this specific direction. We then conclude that accounting for it through a random coefficients specification is essential for recovery of the classical treatment effect parameters.

U-CATE is a building block from which we may derive a large variety of treatment effect parameters that depend on averages, *e.g.*, the average treatment effect, or the treatment effect on the treated:

$$ATE(x) = \int_{\mathbb{R}^L} \overline{U - CATE(\gamma, x)} f_{\widetilde{\Gamma}|X}(\gamma|x) d\gamma = \int_{\mathbb{R}^L} R^{-1} \left[\overline{\partial_v \mathbb{E}\left[Y \left| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x \right. \right]} \right] (\gamma) d\gamma$$

$$ATT(x) = \int_{\mathbb{R}^L} h_{ATT}(\gamma, x) \overline{U - CATE(\gamma, x)} f_{\widetilde{\Gamma}|X}(\gamma|x) d\gamma = \int_{\mathbb{R}^L} h_{ATT}(\gamma, x) R^{-1} \left[\overline{\partial_v \mathbb{E}\left[Y \left| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x \right. \right]} \right] (\gamma) d\gamma$$

where $h_{ATT}(\gamma, x) = \mathbb{E}\left[\mathbf{1}\left\{ \widetilde{S}'\widetilde{\Gamma} < \widetilde{V} \right\} | X = x \right]^{-1} \mathbb{E}\left[\mathbf{1}\left\{ \widetilde{S}'\gamma < \widetilde{V} \right\} | X = x \right]$, and analogously for the treatment effect on the untreated. In Section 3.4, we discuss a natural extension of U-CATE to obtain more general treatment effect parameters which depend on the whole distribution of treatment effects, or on the full joint distribution of $(Y_0, Y_1, \widetilde{\Gamma})$.

This concludes our discussion of the point identified effects; however, we can use the results in Theorem 3.1 in additional ways, as the following subsection illustrates.

3.3. **Bounds on the Variance and the Distribution of Treatment Effects.** As discussed above, Theorem 3.1 reveals that under our assumptions, the marginals of Y_1 and Y_0, conditional on preference parameters, are identified. This information about the marginals, as well as moments of the marginals, can be used to bound moments of the difference $Y_1 - Y_0$, as well as the distribution of this difference, as this section illustrates.

For the formal statement of the bounds, we require the following notation. Let

$$F^L(y_0, y_1|x) = \max\left\{ \int_{\mathbb{R}^L} R^{-1} \left[\overline{\partial_v \mathbb{E}\left[\mathbf{1}\{Y \leq y\}(2D-1) \left| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x \right. \right]} \right] (\gamma) d\gamma - 1, 0 \right\}$$

$$F^U(y_0, y_1|x) = \min\left\{ \int_{\mathbb{R}^L} R^{-1} \left[\overline{\partial_v \mathbb{E}\left[\mathbf{1}\{Y \leq y\}(D-1) \left| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x \right. \right]} \right] (\gamma) d\gamma \right.,$$
$$\left. \int_{\mathbb{R}^L} R^{-1} \left[\overline{\partial_v \mathbb{E}\left[\mathbf{1}\{Y \leq y\} D \left| \left(\widetilde{S}, \widetilde{V}\right) = \cdot, X = x \right. \right]} \right] (\gamma) d\gamma \right\}$$

and

$$F^L(\delta|x) = \int_{\mathbb{R}^L} \sup_{y\in\mathbb{R}} \max\left\{R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\mathbf{1}\{Y\le y\}D\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot, X=x\right.\right]}\right](\gamma)\right.$$
$$\left. -R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\mathbf{1}\{Y\le y-\delta\}(D-1)\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot, X=x\right.\right]}\right](\gamma), 0\right\} d\gamma$$
$$F^U(\delta|x) = 1 + \int_{\mathbb{R}^L} \inf_{y\in\mathbb{R}} \min\left\{R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\mathbf{1}\{Y\le y\}D\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot, X=x\right.\right]}\right](\gamma)\right.$$
$$\left. -R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\mathbf{1}\{Y\le y-\delta\}(D-1)\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot, X=x\right.\right]}\right](\gamma), 0\right\} d\gamma.$$

Now we are in a position to characterize the bounds for the variance and the distribution of treatment effects in our baseline scenario.

Theorem 3.3. Suppose that Assumptions 2.1 and 2.2 hold. Then we obtain that, almost surely in x in $supp(X)$, for every $(y_0, y_1, \delta) \in \mathbb{R}^3$,

$$F^L(y_0, y_1|x) \le F_{Y_0,Y_1|X}(y_0, y_1|x) \le F^U(y_0, y_1|x) \tag{3.13}$$

and

$$F^L(\delta|x) \le F_{\Delta|X}(\delta|x) \le F^U(\delta|x)., \tag{3.14}$$

if in addition $\mathbb{E}\left[(Y_0)^2 + (Y_1)^2\right] < \infty$, then

$$\left(Var(\Delta|X=x) + \left(\int_{\mathbb{R}^L} R^{-1}\left[\overline{\partial_v \mathbb{E}\left[Y\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot, X=x\right.\right]}\right](\gamma)d\gamma\right)^2\right.$$
$$\left. + \int_{\mathbb{R}^L} R^{-1}\left[\overline{\partial_v \mathbb{E}\left[(1-2D)Y\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot, X=x\right.\right]}\right](\gamma)d\gamma\right)^2$$
$$\le 4\int_{\mathbb{R}^L} R^{-1}\left[\overline{\partial_v \mathbb{E}\left[(D-1)Y^2\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot, X=x\right.\right]}\right](\gamma)d\gamma \int_{\mathbb{R}^L} R^{-1}\left[\overline{\partial_v \mathbb{E}\left[DY^2\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot, X=x\right.\right]}\right](\gamma)d\gamma.$$

Note that this result implies that we can bound the variance and the distribution of treatment effects, $Var(\Delta|X=x)$ and $F_{\Delta|X}(\delta|x)$, respectively, entirely by observable quantities. The bounds on the variance are very easy to obtain and do not need to be obtained from the CDFs. The bound (3.13) is a direct application of the Fréchet-Hoeffding bounds, like Heckman, Smith and Clements (1997), Manski (1997) and Heckman and Smith (1998). It is obtained from the conditional CDFs $F_{Y_0|X}(y_0, x) = \mathbb{E}_{\widetilde{\Gamma}}\left[F_{Y_0|X,\widetilde{\Gamma}}(y_0|x,\widetilde{\Gamma})\right]$ and $F_{Y_1|X}(y_1|x) = \mathbb{E}_{\widetilde{\Gamma}}\left[F_{Y_1|X,\widetilde{\Gamma}}(y_1|x,\widetilde{\Gamma})\right]$. The bound (3.14) is, like Fan and Park (2010) and Firpo and Ridder (2008), a consequence of the Makarov bounds. Firpo and Ridder (2008) show that, unlike the Fréchet-Hoeffding bounds which are uniformly sharp, the Makarov bounds are only pointwise sharp (unless the outcomes are binary). We show that the average

Makarov bounds are pointwise sharp and tighter than the Makarov bounds on the average distribution. (3.14) is obtained by taking the expectation of the Makarov bounds evaluated at $F_{\Delta|X,\widetilde{\Gamma}}(\delta|x,\widetilde{\Gamma})$. In a similar fashion, we derive sharper bounds on F_Δ from (3.14) by averaging over covariates than what we would obtain by calculating the Makarov bounds from the CDFs of Y_0 and Y_1, *i.e.*,

$$\mathbb{E}_X\left[F^L(\delta|X)\right] \le F_\Delta(\delta) \le \mathbb{E}_X\left[F^U(\delta|X)\right].$$

(3.13) and (3.14) are obtained integrating out the unobservables, they take the random coefficients structure in the selection equation into account. While the above mentioned references present bounds in the case of randomized experiments or selection on observables, we obtain for the first time bounds in the case where there could be endogenous selection into treatment.

Confidence bands and bounds on functionals of $F_{\Delta|X}$ can be obtained in a similar spirit as Fan and Park (2010) and Firpo and Ridder (2008). Bounds on functionals of $F_{Y_0,Y_1|X}$ could be obtained in similar spirit of Fan and Zhu (2009). Unlike these references, these bounds would hold in the case of endogenous selection. They would also be tighter because we exploit the specific random coefficients structure in the selection equation. Nevertheless, these bounds could be large in practice. Hence, we now discuss possible alternatives that provide point identification.

3.4. **Point Identification of Variance and Distribution of Treatment Effects.** In this section, as well as in Sections 4.3.3 and 4.3.4, we consider the particular situation where the random variables Y_0 and Y_1 are continuously distributed. We thus replace (A-1) by (A-(1)') in Assumption 2.1.

As we have just seen, in order to point identify the variance and the distribution of treatment effects, it is imperative to invoke further assumptions. In our endogenous setup, these assumptions involve the unobservables. For the identification of the distribution of treatment effects, we make the following key assumption.

Assumption 3.2. $Y_0 \perp \Delta \;|\widetilde{\Gamma},\; X$

A slightly weaker form is sufficient for point identification of the variance of treatment effects. Heckman, Smith and Clements (1997) considers independence of the base state and the gains given D. We argue below that controlling for a vector of random coefficients instead of D retains the same spirit as in Heckman, Smith and Clements (1997), but makes the assumption more plausible. Heckman and Smith (1998) consider independence given a vector of observable characteristics X. As we argue subsequently in various extensions of the Roy model, it is important to control for the unobserved heterogeneity that enter in the selection equation as well. To keep the notation minimal, we suppress henceforth the dependence on X.

Assumption 3.2 can be readily interpreted in terms of control functions. It is useful to think of Y as being generated by a nonseparable model; in this case $Y = \psi(D, U)$, and $Y_0 = \psi(0, U)$, as well as $Y_1 = \psi(1, U) = Y_0 + \Delta$. If we identify U with a high dimensional unobservable, *e.g.*, preferences, it is interesting to note that this implies that our random coefficients Y_0 and Δ are two different functions of these unobservables, *i.e.*, $Y_0 = a(U) = \psi(0, U)$ and $\Delta = b(U) = \psi(1, U) - \psi(0, U)$. Without loss of generality, one could further partition the set of unobservable in vectors U_0, U_1, and U_2, and write $Y_0 = a(U_0, U_2)$ and $\Delta = b(U_1, U_2)$.

Assumption 3.2 restricts the heterogeneity appearing in this model. It is best understood in terms of the reformulation introduced above, namely $Y_0 = a(U_0, U_2)$ and $\Delta = b(U_1, U_2)$. A sufficient condition for Assumption (3.2) is that $(\Gamma', \Theta) = U_2$, and $U_0 \perp U_1 | (\Gamma', \Theta)$ (for the latter it would in turn be sufficient that $U_0 \perp U_1 \perp (\Gamma', \Theta)$). In words, there is a common driving factor that causes the selection bias and the dependence between Y_0 and Δ. This factor is given by (Γ', Θ), which, even though it is not recovered for every individual, implicitly serves as a control function. There is remaining randomness in Y_0 and Δ, however, once the driving factor for endogeneity in this system, *i.e.*, (Γ', Θ), is accounted for, there is no leftover dependence.

Note that it does **not** mean that $Y_0 \perp \Delta$. In fact, unless there is no endogenous selection there will generally be dependence between Y_0, and $Y_1 - Y_0$. In summary, there is endogenous selection into treatment, but as far as it is endogenous, it can be summarized by (Γ', Θ). Note that the assumption is more likely to hold in the model with several unobservables in the selection equation, in the sense that there is not just a single factor that we can employ to control for endogeneity, but a full vector of such variables.

The next subsection details that this assumption is sensible in several extensions of the Roy model.

3.4.1. *The Example of The Roy Model.* The aim of this section is to show that Assumption 3.2 is satisfied for several extensions of the popular Roy model. Consider a model where the individuals have at their disposal an information set $\mathcal{I}$ at the time of their decision to participate in the program. They select themselves into treatment if and only if their expected net utility exceeds expected costs.

Formally, the individuals choose $D = 1$ if and only if

$$\mathbb{E}[Y_1 - Y_0 - C_1 | \mathcal{I}] > 0,$$

where C_1 denotes the costs associated with participating in treatment. For simplicity, throughout this subsection, we do not assume to condition on observed factors X. However, we note that they could be used to make Assumption 3.2 more plausible. We use the notation $\Delta = \mathbb{E}[\Delta|\mathcal{I}] + \Xi$. We also assume that the expected costs $\mathbb{E}[C_1|\mathcal{I}]$ can be approximated on individual level by a linear function, *i.e.*, $\mathbb{E}[C_1|\mathcal{I}] \cong \overline{\Gamma}_0 - \overline{\Gamma}_1 V + \overline{\Gamma}' Z$, where $\overline{\Gamma}_1 > 0$ almost surely (the original V can be changed to $-V$). Since the population is heterogeneous, these coefficients vary across the population. Dividing the expected net utility of treatment by $\overline{\Gamma}_1$, we obtain the selection equation (1.2) with $\Gamma = \overline{\Gamma}/\overline{\Gamma}_1$ and $\Theta = (\overline{\Gamma}_0 - \mathbb{E}[\Delta|\mathcal{I}])/\overline{\Gamma}_1$.

Let us consider a first setup where Assumption 3.2 is satisfied. Suppose that $\overline{\Gamma}_0 - \overline{\Gamma}_1 V + \overline{\Gamma}' Z = h_0(\Psi) + h_1(\Psi)V + h_2(\Psi)'Z$ where Ψ denotes some deep economic parameters. When $L \geq 2$ we have

$$\Gamma = \frac{h_2(\Psi)}{h_1(\Psi)} = \left(\frac{h_{2\ 1}(\Psi)}{h_1(\Psi)}, \dots, \frac{h_{2\ L-1}(\Psi)}{h_1(\Psi)}\right)'.$$

It is reasonable to believe that when L is relatively large and the instruments are well chosen Γ captures a lot of features of the deep parameters Ψ. The following assumption considers an ideal situation.

Assumption 3.3 (Invertibility)**.** There is a bijective mapping from Ψ into $\Gamma = h_2(\Psi)/h_1(\Psi)$.

It implies that Ψ and hence also $\overline{\Gamma}_0$ and $\overline{\Gamma}_1$ are $\sigma(\Gamma)$-measurable, where $\sigma(A)$ denotes the sigma algebra generated by a random vector A. Hence $\mathbb{E}[\Delta|\mathcal{I}] = -\overline{\Gamma}_1\Theta + \overline{\Gamma}_0$ is $\sigma(\Gamma, \Theta)$-measurable. Note that we do not have to know or estimate this mapping.
Assume as well

Assumption 3.4. Ξ is $\sigma(\Gamma, \Theta)$-measurable.

When $\mathcal{I} \supset \sigma(\Gamma, \Theta)$ Assumption 3.4 can be rewritten in the form $\Xi = 0$, *i.e.* the agents have perfect foresight.

Proposition 3.2. Assumptions 3.3 and 3.4 imply Assumption 3.2. Also, Assumption 3.2 is satisfied even if we switch the labels between 0 and 1.

Proposition 3.2 is a direct consequence of the decomposition $\Delta = \mathbb{E}[\Delta|\mathcal{I}] + \Xi$. Indeed the assumptions yield that Δ is $\sigma(\Gamma, \Theta)$-measurable and $Y_0 \perp \Delta|\Gamma, \Theta$ is trivially satisfied as conditional on the unobservables the treatment effect is constant. Having an assumption that is independent of the labeling is a desirable property when there is no specific treatment but simply two different states, *e.g.*, two different employment sectors.

In the second setup we assume more generally that the forecast error on the gains is independent of the outcome in base state, given the rescaled sources of unobserved heterogeneity in the selection equation.

Assumption 3.5. $\Xi \perp Y_0|\Gamma, \Theta$.

The following proposition simply relies on the decomposition $\Delta = \mathbb{E}[\Delta|\mathcal{I}] + \Xi$.

Proposition 3.3. Assumptions 3.3 and 3.5 imply Assumption 3.2.

Note, however, that if Γ and Θ are known to the agents at the time the decision is made, $\mathbb{E}[\Delta|\mathcal{I}] = U - CATE$ is identified and $f_{\Xi|\widetilde{\Gamma}}$ is identified if $f_{\Delta|\widetilde{\Gamma}}$ is identified, see Section 3.4.3. Invertibility has thus a strong implication regarding the structure of the ex-ante information set.

In the third setup we relax Assumption 3.3 from setups 1 and 2 to allow $\overline{\Gamma}_0, \overline{\Gamma}_1$ to not be $\sigma(\Gamma)$-measurable. The following proposition gives a more general sufficient condition for Assumption 3.2 to hold that is satisfied under Assumptions 3.3 and 3.5.

Proposition 3.4. Assume that $\overline{\Gamma}_1 = \lambda(\Gamma, \Xi_1)$ for some measurable function λ, and that $(\overline{\Gamma}_0, \Xi, \Xi_1) \perp Y_0|\Gamma, \Theta$, then Assumption 3.2 is satisfied.

The proposition follows from the fact that

$$\begin{aligned}\Delta &= \mathbb{E}[\Delta|\mathcal{I}] + \Xi \\ &= -\lambda(\Gamma, \Xi_1)\Theta + \overline{\Gamma}_0 + \Xi.\end{aligned}$$

Note that, when in the original scale one coordinate of $\overline{\Gamma}$ is non random, then $\overline{\Gamma}_1 \in \sigma(\Gamma)$ and thus there exists a measurable function λ such that $\overline{\Gamma}_1 = \lambda(\Gamma, \Xi_1)$. Proposition (3.4) is also satisfied when $\overline{\Gamma}_1$ is non random or when $\overline{\Gamma}_1 = \Xi_1$ and $(\overline{\Gamma}_0, \Xi, \Xi_1) \perp Y_0|\Gamma, \Theta$. Moreover, it is worth mentioning that there could be arbitrary dependence between $(\overline{\Gamma}_0, \Xi, \Xi_1)$ holding fixed Γ, Θ and that we do not assume that $(\overline{\Gamma}_0, \Xi, \Xi_1) \perp Y_0$ but rather that they can only depend on each other through Γ, Θ.

The assumptions of Proposition 3.4 allow for more complex structure of the ex-ante information set than Assumptions 3.3 and 3.5. Indeed $\mathbb{E}[\Delta|\mathcal{I}] = -\lambda_1(\Gamma, \Xi_1)\Theta + \overline{\Gamma}_0$ can be different from U-CATE if λ is non constant in its last argument and/or constant in others (*e.g.* the agent does not fully know some of his Γ's).

Proposition 3.4 is the most general sufficient condition for Assumption 3.2 to hold that we present. It could certainly be further generalized. Thus there is a wide class of structural models that encompass generalizations of the Roy model for which Assumption 3.2 can hold.

3.4.2. *Variance of the Treatment Effect.* In the following we show that the Unobservables Conditioned Variance of the treatment effects, called U-CVaTE and defined as $Var\left(\Delta\left|\Theta=\theta,X=x\right.\right)$ and $Var\left(\Delta\left|\widetilde{\Gamma}=\gamma,X=x\right.\right)$, in the cases of $L=1$, and $L\geq 2$, respectively, is point identified under an assumption that is similar in spirit, but weaker than the full independence assumption specified above. It is the following conditional uncorrelatedness.

Assumption 3.6. $\mathbb{E}\left[Y_0^2+Y_1^2\right]<\infty$, $\mathbb{E}\left[Y_0\Delta\left|\Gamma,\Theta,X\right.\right]=\mathbb{E}\left[Y_0\left|\Gamma,\Theta,X\right.\right]\mathbb{E}\left[\Delta\left|\Gamma,\Theta,X\right.\right]$. When $L=1$, there are no Γ's in the conditioning set.

For the sake of completeness, we retain the case of a single unobservable in the selection equation.

Theorem 3.4. Let Assumptions 2.1 and 3.6 hold, and $L=1$. Then, almost surely in x in $supp(X)$, for almost every $\theta\in supp(\Theta)$

$$Var\left(\Delta\left|\Theta=\theta,X=x\right.\right)f_{\Theta|X}(t|x)=\overline{\partial_v\mathbb{E}\left[Y^2\left|V=\cdot,X=x\right.\right]}(t)+\frac{\overline{\partial_v\mathbb{E}\left[Y\left|V=\cdot,X=x\right.\right]}(t)\overline{\partial_v\mathbb{E}\left[(1-2D)Y\left|V=\cdot,X=x\right.\right]}(t)}{f_{\Theta|X}(t|x)}.$$

In the case where $L\geq 2$, if we also invoke Assumption 2.2, we obtain that for almost every $\gamma\in supp(\widetilde{\Gamma})$,

$$Var\left(\Delta\left|\widetilde{\Gamma}=\gamma,X=x\right.\right)f_{\widetilde{\Gamma}|X}(\gamma|x)=R^{-1}\left[\overline{\partial_v\mathbb{E}\left[Y^2\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot,X=x\right.\right]}\right](\gamma)$$
$$+\frac{R^{-1}\left[\overline{\partial_v\mathbb{E}\left[Y\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot,X=x\right.\right]}\right](\gamma)R^{-1}\left[\overline{\partial_v\mathbb{E}\left[(1-2D)Y\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot,X=x\right.\right]}\right](\gamma)}{f_{\widetilde{\Gamma}|X}(\gamma|x)}.$$

Finally, under the same additional Assumption 2.2, it holds that

$$Var(\Delta|X=x)=\int_{\mathbb{R}}Var(\Delta|\Theta=\theta,X=x)f_{\Theta|X}(t|x)dt+\int_{\mathbb{R}}\left(\mathbb{E}[\Delta|\Theta=\theta,X=x]-ATE(x)\right)^2f_{\Theta|X}(t|x)d\gamma.$$

when $L=1$, and

$$Var(\Delta|X=x)=\int_{\mathbb{R}^L}Var(\Delta|\widetilde{\Gamma}=\gamma,X=x)f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma+\int_{\mathbb{R}^L}\left(\mathbb{E}[\Delta|\widetilde{\Gamma}=\gamma,X=x]-ATE(x)\right)^2f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma.$$

when $L\geq 2$.

These rather involved formulae provide nevertheless a succinct description of the conditional variance, which does not require material assumptions beyond instrument independence, no correlation between base state and treatment effect, conditional on all observable and unobservable variables (Assumption 3.6), and a specific relation between the variation of the instruments and the support of the unobserved heterogeneity in the model for the selection into treatment (Assumption 2.2). When $L=1$, but the latter assumption does not hold, we may again identify the variance of the treatment effect for the population such that $\Theta \in supp\left(f_{V|X}(\cdot|x)\right)$ given $X=x$.

3.4.3. *U-CDiTE and Treatment Effect Parameters that Depend on the Distribution of Treatment Effects or the Joint Distribution of Potential Outcomes.* This section extends the analysis of the previous subsections to distributions of treatment effects. We define the Conditional Distribution of Treatment Effects as

$$U-CDiTE(\delta,\theta,x) := f_{\Delta|\Theta,X}(\delta|\theta,x)$$

when $L=1$ and

$$U-CDiTE(\delta,\gamma,x) := f_{\Delta|\widetilde{\Gamma},X}(\delta|\gamma,x)$$

when $L\geq 2$. This quantity is the distribution of treatment effects for the subpopulation with unobserved heterogeneity parameter (respectively vector) from the first stage selection equation equal to θ (respectively γ) and observables $X=x$. It is also the distribution of the gains in terms of Y_1-Y_0 for people with $X=x$ who would be indifferent between participation and non-participation in the treatment if they were exogenously assigned the value v of V (respectively (s,v) of $(\widetilde{S},\widetilde{V})$), such that $v=\theta$ (respectively $s'\gamma=v$). It is straightforward to check that, akin to U-CATE, Assumption (A-2) yields that $U-CDiTE(\theta,x)$ (respectively $U-CDiTE(\gamma,x)$) does not depend on the values taken by the instrument V (respectively $\left(\widetilde{S},\widetilde{V}\right)$), and is hence policy invariant. This quantity is at the heart of any calculation of more general treatment effects parameters that go beyond averages and depend on the distribution of either the treatment effects Y_1-Y_0, or the joint of potential outcomes (Y_1,Y_0).

To calculate some of these distributional treatment effects we sometimes have to replace Assumption 2.1 (A-2) by the following stronger assumption.

Assumption 3.7.

$$(V,Z)\perp(Y_0,Y_1,\Gamma',\Theta)|X.$$

More specifically, out of the list of the following parameters, which can be deduced from U-CDiTE, Assumption 3.7 is required to obtain the last three,

$$f_{\Delta|X}(\delta|x) = \int_{\mathbb{R}^L} \overline{U-CDiTE(\delta,\gamma,x)} f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma$$

$$\mathbb{P}\left(\Delta>0|X=x\right) = \int_{\mathbb{R}} \mathbf{1}\left\{\delta>0\right\} \int_{\mathbb{R}^L} \overline{U-CDiTE(\delta,\gamma,x)} f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma d\delta$$

$$f_{Y_0,Y_1|X}(y_0,y_1|x) = \int_{\mathbb{R}^L} \overline{U-CDiTE(y_1-y_0,\gamma,x) f_{Y_0|\widetilde{\Gamma},X}(y_0|\gamma,x)} f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma$$

$$f_{\Delta|D=1,X}(\delta|x) = \int_{\mathbb{R}^L} h_{ATT}(\gamma,x)\overline{U-CDiTE(\delta,\gamma,x)}f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma$$
$$f_{\Delta|D=1,Y_0,X}(\delta|y_0,x) = \int_{\mathbb{R}^L} h_{ATT}(\gamma,x)\overline{U-CDiTE(\delta,\gamma,x)f_{Y_0|\widetilde{\Gamma},X}(y_0|\gamma,x)}f_{\widetilde{\Gamma}|X}(\gamma|x)d\gamma.$$

At this stage, it is important to recall that $f_{Y_0|\widetilde{\Gamma},X}(y_0|\gamma,x)f_{\widetilde{\Gamma}|X}(\gamma|x)$ may be obtained from Theorem 3.1[7]. Setting $\phi(y)=e^{ity}$ in Theorem 3.1 yields, almost surely in x in $supp(X)$,

$$\mathcal{F}_1\left[f_{Y_0+\Delta,\Theta}\right](t,\cdot) = \overline{\partial_v\mathbb{E}\left[e^{itY}D|V=\cdot,X=x\right]}$$
$$\mathcal{F}_1\left[f_{Y_0,\Theta}\right](t,\cdot) = \overline{\partial_v\mathbb{E}[e^{itY}(D-1)|V=\cdot,X=x]},$$

while, when $L\geq 2$, under Assumption 2.2, we obtain

$$\mathcal{F}_1\left[f_{Y_0+\Delta,\widetilde{\Gamma}|X}(\cdot|x)\right](t,\cdot) = R^{-1}\left[\overline{\partial_v\mathbb{E}\left[e^{itY}D\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot,X=x\right.\right]}\right]$$
$$\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}|X}(\cdot|x)\right](t,\cdot) = R^{-1}\left[\overline{\partial_v\mathbb{E}\left[e^{itY}(D-1)\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot,X=x\right.\right]}\right].$$

where $\mathcal{F}_1$ denotes the Fourier transform of the joint density seen as a function of its first variable, holding the other arguments fixed. This object is called a partial Fourier transform.

The importance of Assumption 3.2 is that it allows factorization, *i.e.*,

$$\mathcal{F}_1\left[f_{Y_0+\Delta,\widetilde{\Gamma}|X}(\cdot|x)\right] = \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}|X}(\cdot|x)\right]\mathcal{F}_1\left[f_{\Delta,\widetilde{\Gamma}|X}(\cdot|x)\right],$$

when $L\geq 2$. We make moreover use of the following integrability assumption.

Assumption 3.8. For almost every x in $supp(X)$, for every θ in $supp\left(f_{\theta|X}(\cdot|x)\right)$, respectively for every γ in $supp\left(f_{\widetilde{\Gamma}|X}(\cdot|x)\right)$, $U-CDiTE(\cdot,\theta,x)$, respectively $U-CDiTE(\cdot,\gamma,x)$, belong to $\mathrm{L}^1(\mathbb{R})\cap\mathrm{L}^2(\mathbb{R})$.

Finally, we require a last technical assumption, which is common in the deconvolution literature.

Assumption 3.9. When $L=1$,

$$\forall x\in supp\left(X\right),\ \forall(t,\theta)\in\mathbb{R}\times supp\left(f_{\Theta|X}(\cdot|x)\right),\ \mathcal{F}\left[f_{Y_0|\Theta,X}(\cdot|\theta,x)\right](t)\neq 0$$

while when $L\geq 2$,

$$\forall x\in supp\left(X\right),\ \forall(t,\gamma)\in\mathbb{R}\times supp\left(f_{\widetilde{\Gamma}|X}(\cdot|x)\right),\ \mathcal{F}\left[f_{Y_0|\widetilde{\Gamma},X}(\cdot|\gamma,x)\right](t)\neq 0$$

where $\mathcal{F}$ is the Fourier transform.

[7]Similar expressions can be obtained when $L=1$, but are omitted for brevity of exposition. Moreover, in that case, when Assumption 2.2 does not hold, we may again identify the above parameters for the population such that $\Theta\in supp\left(f_{V|X}(\cdot|x)\right)$ and $X=x$.

We believe that it is possible to weaken this assumption and allow for isolated zeros in the spirit of Devroye (1989) , Carrasco and Florens (2011) and Evdokimov and White (2011) among others but prefer not to elaborate on this in this article.

Theorem 3.5. Let Assumptions 2.1 (with (1)), 3.2, 3.8 and 3.9 hold. In case $L=1$, for every δ in $\mathbb{R}$, almost surely in x in $supp(X)$, for every θ in $supp\left(f_{\Theta|X}(\cdot|x)\right)$, we obtain

$$U-CDiTE(\delta,\theta,x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-it\delta}\frac{\overline{\partial_v\mathbb{E}\left[e^{itY}D\,|V=\cdot,X=x\right](\theta)}}{\overline{\partial_v\mathbb{E}\left[e^{itY}(D-1)\,|V=\cdot,X=x\right](\theta)}}dt; \tag{3.15}$$

while, in case $L\geq 2$, for every δ in $\mathbb{R}$, almost surely in x in $supp(X)$, for every γ in $supp\left(f_{\widetilde{\Gamma}|X}(\cdot|x)\right)$, we obtain:

$$U-CDiTE(\delta,\gamma,x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-it\delta}\frac{R^{-1}\left[\overline{\partial_v\mathbb{E}\left[e^{itY}D\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot,X=x\right.\right]}\right](\gamma)}{R^{-1}\left[\overline{\partial_v\mathbb{E}\left[e^{itY}(D-1)\left|\left(\widetilde{S},\widetilde{V}\right)=\cdot,X=x\right.\right]}\right](\gamma)}dt. \tag{3.16}$$

Note that there is a slight abuse of notations in (3.15) and (3.16) because the numerators and denominators are only defined almost surely in θ, respectively γ. Implicitly, as in Heckman, Smith and Clements (1997), this result can be used to derive a test for the validity of Assumption 3.2: if for some x and θ (respectively γ) $U-CDiTE(\delta,\theta,x)$ (respectively $U-CDiTE(\delta,\gamma,x)$) fails to be a density, it is an indication that Assumption 3.2 is incorrect.

4. Estimation of Structural Parameters

In this section we focus on the case where $L\geq 2$, the case where $L=1$ requires minor modifications.

4.1. **Estimation of Quantities Related to Marginals.**

We consider the following regularized inverse[8] of the Radon transform

$$A_T[f](\gamma):=\int_{H^+}\int_{-\infty}^{\infty}K_T(s'\gamma-u)f(s,u)dud\sigma(s) \tag{4.1}$$

where $d\sigma$ denotes the classical spherical measure[9] on the sphere $\mathbb{S}^{L-1}$ and

$$\forall u\in\mathbb{R},\ K_T(u):=2(2\pi)^{-L}\int_0^{\infty}\cos(tu)t^{L-1}\psi\left(\frac{t}{T}\right)dt \tag{4.2}$$

[8] Up to our knowledge this modification of the classical Radon inverse has not been studied earlier in the literature.

[9] Its mass is the area of the sphere.

where T is a smoothing parameter and ψ is a symmetric rapidly decaying function in the Schwartz class

$$\mathcal{S}(\mathbb{R}) = \left\{ f \in C^\infty(\mathbb{R}) : \ \forall \alpha, \beta \in \mathbb{N}, \ |x|^\alpha \left| \partial^\beta f(x) \right| \underset{|x| \to \infty}{\longrightarrow} 0 \right\},$$

such that $\psi(0) = 1$. For simplicity of exposition we will take $\psi = \psi_0$ where $\psi_0 : \ x \mapsto \exp\left(1 - \max\left\{\frac{1}{1-x^2}, 0\right\}\right)$ which has also support in $[-1, 1]$.

The previous identification sections suggest using this regularized inverse as a building block for a sample counterparts estimator of the inverse Radon transform of, say, a derivative of a regression function, *i.e.*,

$$A_{T_N} \left[\overline{\partial_v \widehat{\mathbb{E}[\phi(Y)\varsigma(D)|(\widetilde{S}, \widetilde{V}) = \cdot, X = x]}} \right] (\gamma)$$

where $\varsigma(D)$ is either D or $D - 1$ and $\overline{\partial_v \widehat{\mathbb{E}[\phi(Y)\varsigma(D)|(\widetilde{S}, \widetilde{V}) = (\cdot), X = x]}}$ is the extension as 0 outside *supp* $\left(\widetilde{S}, \widetilde{V}\right)$ of an estimator of the derivative of the regression function (*e.g.*, using local polynomials). Moreover, T_N is chosen adequately, and tends to infinity as N goes to infinity. Replacing the various inverse Radon transforms by these regularized sample counterparts yields the following set of estimators:

$$\overline{\widehat{U - CATE(\gamma, x)}} f_{\widetilde{\Gamma}}(\gamma) = A_{T_N} \left[\overline{\partial_v \widehat{\mathbb{E}[Y|(\widetilde{S}, \widetilde{V}) = \cdot, X = x]}} \right] (\gamma), \tag{4.3}$$

$$\mathcal{F}_1 \left[\widehat{f_{Y_0+\Delta, \widetilde{\Gamma}|X}(\cdot|x)} \right] (t, \gamma) = A_{T_N} \left[\overline{\partial_v \widehat{\mathbb{E}[e^{itY} D|(\widetilde{S}, \widetilde{V}) = \cdot, X = x]}} \right] (\gamma), \tag{4.4}$$

$$\mathcal{F}_1 \left[\widehat{f_{Y_0, \widetilde{\Gamma}|X}(\cdot|x)} \right] (t, \gamma) = A_{T_N} \left[\overline{\partial_v \widehat{\mathbb{E}[e^{itY}(D - 1)|(\widetilde{S}, \widetilde{V}) = \cdot, X = x]}} \right] (\gamma), \tag{4.5}$$

$$\widehat{f_{\widetilde{\Gamma}|X}}(\gamma|x) = \max\left\{ A_{T_N} \left[\overline{\partial_v \widehat{\mathbb{E}[D|(\widetilde{S}, \widetilde{V}) = \cdot, X = x]}} \right] (\gamma), 0 \right\} \tag{4.6}$$ [10].

These individual elements can now be used to estimate many of the previously discussed quantities. From the first estimator, we may, for example, construct an estimator of the ATE.

$$\widehat{ATE}(x) = \int_{\mathbb{R}^L} \overline{\widehat{U - CATE(\gamma, x)}} f_{\widetilde{\Gamma}|X}(\gamma|x) d\gamma = \int_{\mathbb{R}^L} A_{T_N} \left[\overline{\partial_v \widehat{\mathbb{E}[Y|(\widetilde{S}, \widetilde{V}) = \cdot, X = x]}} \right] (\gamma) \mathbf{1}\{\gamma \in \mathcal{B}_N\} d\gamma,$$

where $\mathcal{B}_N$ is a large enough closed set in $\mathbb{R}^L$ which diameter depends on the sample size.

Alternative estimators that circumvent the numerical integration in (4.1) can be obtained as follows. Start out by defining

$$\forall u \in \mathbb{R}, \ \widetilde{K}_T(u) := -2(2\pi)^{-L} \int_0^T \sin(tu) t^L \psi\left(\frac{t}{T}\right) dt.$$

where $K_T'(u) = \widetilde{K}_T(u)$, with K_T defined by (4.2) and

$$B_T[f](\gamma) := \int_{H^+} \int_{-\infty}^{\infty} \widetilde{K}_T(s'\gamma - u) f(s,u) du d\sigma(s). \tag{4.7}$$

Because K_T is the Fourier transform of a function in $\mathcal{S}(\mathbb{R})$, K_T and $\widetilde{K}_T$ belong to $\mathcal{S}(\mathbb{R})$. Thus, they decay to zero faster than any polynomial and belong to any $\mathrm{L}^p(\mathbb{R})$[11].

In addition, for the alternative estimators, we require the following assumption.

Assumption 4.1. (i) For almost every $s \in H^+$ and almost surely in $x \in supp(X)$, $supp\left(f_{\widetilde{V}|\widetilde{S},X}(\cdot|s,x)\right) = \mathbb{R}$.

(ii) For the function ϕ considered, for almost every $s \in H^+$ and almost surely $x \in supp(X)$, $v \mapsto \mathbb{E}[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V}) = (s,v), X = x]$ is continuous and $v \mapsto \mathbb{E}[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V}) = (s,v), X = x]$ and $v \mapsto \partial_v \mathbb{E}[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V}) = (s,v), X = x]$ are bounded by a polynomial in v.

This assumption allows an integration by parts argument for the regularized inverse, which produces a structure that is easier to implement. It is based on the following proposition.

Proposition 4.1. Under Assumption 4.1, for every $T \in \mathbb{R}$ and $\gamma \in \mathbb{R}^L$,

$$A_T\left[\partial_v \mathbb{E}\left[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V}) = (\cdot), X = x\right]\right](\gamma) = B_T\left[\mathbb{E}\left[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V}) = (\cdot), X = x\right]\right](\gamma) \tag{4.8}$$

$$= \mathbb{E}\left[\left.\frac{\widetilde{K}_T(\widetilde{S}'\gamma - \widetilde{V})\phi(Y)\varsigma(D)}{f_{\widetilde{S},\widetilde{V}|X}(\widetilde{S},\widetilde{V}|x)}\right| X = x\right]. \tag{4.9}$$

Equation (4.8) suggests that one can take as an estimator B_T applied to an estimator of the regression function. (4.9) suggests the following trimmed sample counterpart estimator

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\widetilde{K}_{T_N}(\widetilde{s}_i'\gamma - \widetilde{v}_i) T_{\tau_N}(\phi(y_i))\varsigma(d_i)}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V},X}}(\widetilde{s}_i,\widetilde{v}_i,x), m_N\right)} \mathcal{K}_{\eta_N}(x_i - x) \tag{4.10}$$

where $\widehat{f_{\widetilde{S},\widetilde{V},X}}$ is a plug-in estimator for $f_{\widetilde{S},\widetilde{V},X}$, m_N is a trimming factor and $\mathcal{K}_\eta$ is a standard multivariate kernel with bandwidth vector η_N, T_τ is defined for τ positive and x in $\mathbb{R}$ by

$$T_\tau(x) = -\tau_N \mathbf{1}\{x < -\tau_N\} + x\mathbf{1}\{|x| \le \tau_N\} + \tau_N \mathbf{1}\{x > \tau_N\}$$

T_N, τ_N, m_N^{-1} and η_N^{-1} go to infinity as N goes to infinity. Trimming is introduced to avoid dividing by quantities that are too close to zero. Recall that this estimator can only be computed when $f_{\widetilde{S},\widetilde{V}|X}(\cdot|x)$ has full support, almost surely in x in $supp(X)$, implying that the density decays to zero both for large

[11]This is a very nice feature that is not shared for the classical Radon inverse where $\psi(x) = \mathbf{1}\{x \in [-1,1]\}$, also these yield good approximation results for the target quantities in every $\mathrm{L}^p(\mathbb{R}^L)$.

v and commonly when s approaches the boundary of H^+[12]. In particular, the true density $f_{\widetilde{S},\widetilde{V},X}$ is not bounded away from zero. Moreover we are replacing $f_{\widetilde{S},\widetilde{V},X}$ by an estimator, implying that the denominator could also be small due to estimation error. The truncation parameter τ_N is useful when ϕ is unbounded and $\phi(Y_0)$ (if $\varsigma(D)=D-1$) or $\phi(Y_1)$ (if $\varsigma(D)=D$) have fat tails. In the absence of conditioning on covariates X, the estimator simplifies to

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\widetilde{K}_{T_N}(\widetilde{s}_i'\gamma-\widetilde{v}_i)T_{\tau_N}(\phi(y_i))\varsigma(d_i)}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}. \tag{4.11}$$

4.2. **Estimation of the U-CDiTE and of the Distribution of Treatment Effects.** An estimator for U-CDiTE is obtained by applying the same principles. First, one may compute the following integral

(4.12)

$$\overline{\widehat{U-CDiTE}(\delta,\gamma,x)}=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-it\delta}K(th_{N,\gamma})\frac{\mathcal{F}_1\left[\widehat{f_{Y_0+\Delta,\widetilde{\Gamma}|X}(\cdot|x)}\right](t,\gamma)}{\mathcal{F}_1\left[\widehat{f_{Y_0,\widetilde{\Gamma}|X}(\cdot|x)}\right](t,\gamma)}\mathbf{1}\left\{\left|\mathcal{F}_1\left[\widehat{f_{Y_0,\widetilde{\Gamma}|X}(\cdot|x)}\right](t,\gamma)\right|>t_{N,t,\gamma}\right\}dt$$

where N is the sample size, $h_{N,\gamma}$ is a bandwidth going to zero with N, K denotes a kernel, $t_{N,t,\gamma}$ a proper trimming factor.

A typical example of a kernel is $K(t)=\mathbf{1}\{|t|\le 1\}$, with $h_{N,\gamma}=1/R_{N,\gamma}^{\Delta}$ it amounts to truncation of high frequencies. Devroye (1989) uses $\max(1-|t|,0)$ for the estimation with $L^1(\mathbb{R})$ loss. It is also possible to take $K=\psi$ where ψ belongs to $\mathcal{S}(\mathbb{R})$ and is such that $\psi(0)=1$ like in Section 4.1. In Section 5 we take $K(t)=\psi_0(t)$ with support in $[-1,1]$ (recall that $\psi_0(t)=\exp\left(1-\max\left\{\frac{1}{1-t^2},0\right\}\right)$). For every γ, the quantity $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}|X}(\cdot|x)\right](t,\gamma)\right|$ in the denominator of (4.12), decays to zero when t goes to infinity[13]. A smoothing kernel K should put less weight (possibly zero weight) on high frequencies[14] because this is where the denominator is small in modulus and the variance of the estimator can blow-up. In theory the bandwidth $h_{N,\gamma}$ may depend on γ[15] because of the possible different rates of decay of $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}|X}(\cdot|x)\right](t,\gamma)\right|$ for different values of γ.

Compared to usual deconvolution estimators, (4.12) also has an indicator function with a trimming factor. It is used because we are considering a case where the denominator is unknown and

[12] This is because $\widetilde{S}$ is a rescaled vector of original instruments and assuming that the density does not decay to zero when s approaches the boundary of H^+ would be a very strong assumption on the tails of the distributions of the original instruments (see, *e.g.*, Beran, Feuerverger and Hall (1996) and Hoderlein, Klemelä and Mammen (2011)).

[13] This is a consequence of the Riemann-Lebesgue lemma.

[14] The parameter t is the frequency.

[15] Every smoothing parameter can also depend on x but we omit it ease of notations.

has to be estimated. For that reason the denominator can also be small due to estimation error. This is the same structure as the estimator proposed in Neumann (1997)[16].

In the simulation study in Section 5, we take $h_{N,\gamma}$ and $t_{N,t,\gamma}$ that do not depend on t and γ for the smoothing parameters in the estimators of the numerator and denominator in (4.12). Using $K = \psi_0$ we also obtain exactly the same graphs for (4.12) regardless of whether or not we trim, so we present results with $t_{N,t,\gamma} = 0$. This is an important feature in practice, because in Proposition 4.5 we impose to tune $t_{N,t,\gamma} = r_{Y_0,N}$. But $r_{Y_0,N}$ is unknown since it depends on the smoothness of $f_{Y_0,\widetilde{\Gamma}|X}$ which is impossible to estimate. We believe that this is just a technical issue and that in practice no trimming works perfectly fine at least for smooth $f_{Y_0,\widetilde{\Gamma}|X}$ (super smooth in our simulation example because it is Gaussian).

An estimator of the unconditional distribution of treatment effects uses as plug-ins (4.12) and an estimator of the mixing density $f_{\widetilde{\Gamma}|X}$; it is given by

$$\widehat{f_{\Delta|X}}(\delta|x) = \int_{\mathbb{R}^L} \overline{U - \widehat{CDiTE}(\delta,\gamma,x)}\widehat{f_{\widetilde{\Gamma}|X}}(\gamma|x)\mathbf{1}\{\gamma \in \mathcal{B}_N\}d\gamma. \tag{4.13}$$

4.3. **Rates of Convergence.** To analyze the rate of convergence of our estimators, we have to introduce additional notation. We denote by $\|\cdot\|_p$ for $p \in [1,\infty)$ the classical L^p norms and by $\|\cdot\|_\infty$ the essential supremum norm, also called sup-norm. Moreover, for ease of notation we again omit the conditioning on X in this section.

4.3.1. *Estimation of Generic Quantities.* In this section we consider the estimation of one of the plug-in terms of the form

$$g(\gamma) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[\phi(Y)\varsigma(D)\middle|\left(\widetilde{S},\widetilde{V}\right) = \cdot, X = x\right]}\right](\gamma),$$

for some function ϕ and $\varsigma(D)$ is either D or $1-D$, by an estimator of the form (4.11)

$$\hat{g}(\gamma) = \frac{1}{N}\sum_{i=1}^N \frac{\widetilde{K}_{T_N}(\widetilde{s}_i'\gamma - \widetilde{v}_i)T_{\tau_N}(\phi(y_i))\varsigma(d_i)}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i), m_N\right)}.$$

where T_N, τ_N and m_N^{-1} increase with the sample size.

We restrict our attention to the sup-norm loss because this is what is later required in Assumption 4.3 and Proposition 4.5 for plug-in estimation. We specify the result to the case of ordinary smooth

[16]In Neumann (1997) and Comte and Lacour (2011) because the characteristic function of Y_0 is estimated at rate $1/\sqrt{N}$, $t_{N,t,\gamma}$ could be taken equal to $N^{-1/2}$, independent of t and γ.

functions (Sobolev classes). We will work with the following Sobolev spaces of all locally integrable functions that have weak derivatives up to order s in $\mathbb{N}\setminus\{0\}$ (see, *e.g.*, Evans (1998))

$$W^{s,\infty}(\mathbb{R}^L) := \left\{f \in \mathrm{L}^\infty(\mathbb{R}^L) :\ \forall|\alpha| \le s,\ \partial^\alpha f \in \mathrm{L}^\infty(\mathbb{R}^L)\right\}$$

where $\alpha \in \mathbb{N}^L$, $|\alpha| := \sum_{l=1}^L \alpha_l$ and $\partial^\alpha f := \prod_{l=1}^L \partial_l^{\alpha_l} f$ is the α^{th}-weak partial derivative of f. It is equipped with the norm

$$\|f\|_{s,\infty} := \sum_{\alpha:\ |\alpha|\le s} \|\partial^\alpha f\|_\infty.$$

We will consider the following Sobolev ellipsoids defined for M positive by

$$W^{s,\infty}(M) := \left\{f \in W^{s,\infty}(\mathbb{R}^L) :\ \|f\|_{s,\infty} \le M\right\}.$$

In what follows, $\mathcal{B}_N$ is a closed set in $\mathbb{R}^L$ and we denote by $d(\mathcal{B}_N)$ its diameter for the Euclidian norm.

Proposition 4.2. Let Assumption 4.1 hold, and assume moreover that

(i) g belongs to $\mathrm{W}^s_\infty(M)$ for some s in $\mathbb{N}\setminus\{0\}$ and M positive ;

(ii) there exists α positive such that $\log(T_N^3/m_N) + L\log(d(\mathcal{B}_N)) \le \alpha$;

(iii) there exists a sequence $r_{IV,N}$ going to 0 as N goes to infinity and M_{IV} positive such that with probability one

$$\overline{\lim}_{N\to\infty} r_{IV,N}^{-1} \max_{i=1,\dots,N} \left| f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i) - \widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i)\right| \le M_{IV} \ ; \tag{4.14}$$

then for some constants $M(\alpha)$ (which only depends on α and L) and $C(s)$ (which only depends on s and ψ), with probability one, for every ϵ positive there exists N large enough such that

$$\begin{aligned}
\|(\hat{g} - g)\,\mathbf{1}\{\mathcal{B}_N\}\|_\infty \le{}& (M_{IV}+\epsilon)\min(\tau_N, \|\phi\|_\infty)\, r_{IV,N} m_N^{-1} \left\| \mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma - \widetilde{V}\right)\right|}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}\left(\widetilde{S},\widetilde{V}\right), m_N\right)}\right]\right\|_\infty \\
&+ (M_{IV}+\epsilon)\min(\tau_N, \|\phi\|_\infty)\, r_{IV,N} m_N^{-3/2}(M(\alpha)+\epsilon)\left(\frac{\log N}{N}\right)^{1/2} T_N^{L+1/2} \\
&+ m_N^{-1/2}(M(\alpha)+\epsilon)\min(\tau_N, \|\phi\|_\infty)\left(\frac{\log N}{N}\right)^{1/2} T_N^{L+1/2} \\
&+ \min(\tau_N, \|\phi\|_\infty) \sup_{\gamma\in\mathcal{B}_N} \int_{\left\{(s,v):\ f_{\widetilde{S},\widetilde{V}}(s,v) < m_N\right\}} \left|\widetilde{K}_{T_N}(s'\gamma - v)\right| d\sigma(s)dv \\
&+ MC(s)T_N^{-s} \\
&+ \frac{1}{(2\pi)^L} T_N^{L+2} \||t|^L\psi\|_1 \mathbb{E}[|\phi(Y_j)|\mathbf{1}\{|\phi(Y_j)| > \tau_N\}]
\end{aligned}$$

where $j = 1$ if $\zeta(D) = D$ and $j = 0$ if $\zeta(D) = 1 - D$.

Let us make a few comments on this result.

(1) When $supp(\widetilde{\Gamma})$ is bounded then we can take $\mathcal{B}_N = supp(\widetilde{\Gamma}) = supp(g)$.

(2) Condition (4.15) can be relaxed to "bounded in probability" if we simply want to prove convergence in probability. This is actually the only thing that we need for the properties of the estimation of U-CDiTE that will follow.

(3) There are various ways to bound from above

$$\mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma - \widetilde{V}\right)\right|}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}\left(\widetilde{S},\widetilde{V}\right), m_N\right)}\right].$$

A first uniform upper bound uses

$$\mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma - \widetilde{V}\right)\right|}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}\left(\widetilde{S},\widetilde{V}\right), m_N\right)}\right] \leq \frac{|\mathbb{S}^{L-1}|}{2}\|\widetilde{K}_{T_N}\|_1$$

where $|\mathbb{S}^{L-1}|$ is the surface of the sphere $\mathbb{S}^{L-1}$. A second uniform upper bound, that has an analytic form[17] is given by

$$\mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma - \widetilde{V}\right)\right|}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}\left(\widetilde{S},\widetilde{V}\right), m_N\right)}\right] \leq \mathbb{E}\left[\left(\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma - \widetilde{V}\right)\right|}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}\left(\widetilde{S},\widetilde{V}\right), m_N\right)}\right)^2\right]^{1/2} \leq m_N^{-1}T_N^{2L+1}.$$

The last inequality uses (7.5) in the appendix.

(4) Because $\widetilde{K}_{T_N}$ is integrable[18], $\int_{\left\{(s,v):\ f_{\widetilde{S},\widetilde{V}}(s,v)<m_N\right\}}\left|\widetilde{K}_{T_N}(s'\gamma - v)\right| d\sigma(s)dv$ goes to zero as m_N goes to zero and the rate of convergence to zero depends on the tails of $f_{\widetilde{S},\widetilde{V}}$. Fat tails imply that this term is small. It could be made equal to zero if $f_{\widetilde{S},\widetilde{V}}$ were bounded away from zero.

(5) The contribution $MC(s)T_N^{-s}$ is an upper bound on the approximation error for functions in the ellipsoid $W^{s,\infty}(\mathbb{R}^L)$.

(6) We use truncation, because we deal with the fluctuation terms using the basic Bernstein inequality. It is possible to use the Bernstein inequality for random variables with bounded Orlicz norms (see, *e.g.*, Lemma 2.2.11 of Van der Vaart and Wellner (1996)), or other concentration results and avoid truncation in certain cases. When ϕ is bounded, we can take $\tau_N = \|\phi\|_\infty$ and the last term in the above upper bound disappears. In Proposition 4.2, we considered the property of estimators involving truncation for the estimation of the numerator of U-CATE

[17] We expect that it is not as sharp as the previous upper bound where, unfortunately, we do not have an upper bound for $\|\widetilde{K}_{T_N}\|_1$ in terms of its dependence in T_N.

[18] It belongs to $\mathcal{S}(\mathbb{R})$.

and for the estimation of the conditional variance of treatment effect when $|Y_0|$ and $|Y_1|$ can take arbitrarily large values. Because $\mathbb{E}\left[|\phi(Y_0)|+|\phi(Y_1)|\right]<\infty$, $\mathbb{E}\left[|\phi(Y_1)|\mathbf{1}\{|\phi(Y_1)|>\tau_N\}\right]$ and $\mathbb{E}\left[|\phi(Y_0)|\mathbf{1}\{|\phi(Y_0)|>\tau_N\}\right]$ go to zero when τ_N goes to infinity.

In the ideal case where: (1) $f_{\widetilde{S},\widetilde{V}}$ is bounded away from zero, (2) its density is smooth enough for the first term to be negligible and (3) the bias due to truncation is negligible (*e.g.* when ϕ is bounded), we obtain, for some M_I, with probability 1,

$$\overline{\lim}_{N\to\infty}\left(\frac{\log N}{N}\right)^{-\frac{s}{2s+2L+1}}\|\hat{g}-g\|_\infty\le M_I$$

by taking T_N of the order of $(N/\log(N))^{1/(2s+2L+1)}$. Recall that the rate of direct density estimation would be $(N/\log(N))^{s/(2s+L)}$. This is important, because it says that the degree of ill-posedness due to the presence of the unbounded operator is $\frac{L+1}{2}$. Recall that in positron emission tomography, the classical statistical inverse problem involving the Radon transform, the degree of ill-posedness is $\frac{L-1}{2}$ (see, *e.g.*, Korostelev and Tsybakov (1993)). Here we pay an additional price of 1 due to the extra differentiation. However, this degree of ill-posedness does not properly account for the difficulty of the problem. Equation (3.1) states that a regression function r is of the form $r=Qf$ where Q is an operator which has an unbounded inverse. The quantity $\frac{L-1}{2}$ only accounts for the smoothing properties of the operator K. But for identification we assumed that, in the original scale, all regressors but possibly V have full support. This implies that in most cases $f_{\widetilde{S}}(s)$ is not bounded away from zero and the rate of estimation of the regression function with L^∞ loss is slower than when the regressors have support on a compact set and their density is bounded from below. Estimation of a regression function when the density of the regressor can be 0 on its support (degeneracy) has been studied by several authors (see, *e.g.*, Hall, Marron, Neumann, Tetterington (1997), Guerre (1999) and Gaiffas (2005) and (2009))[19]. In our inverse problem setup this translates in the fact that in many cases $\int_{\left\{(s,v):\ f_{\widetilde{S},\widetilde{V}}(s,v)<m_N\right\}}\left|\widetilde{K}_{T_N}(s'\gamma-v)\right|d\sigma(s)dv$ cannot be made negligible which implies that a second degree of ill-posedness also has to be taken into account. Finally, a third degree of ill-posedness can appear when the conditional distributions of $\phi(Y_0)$ and/or $\phi(Y_1)$ given $\widetilde{\Gamma}$ have heavy tails.

An analogue of Proposition 4.2 has already been established when $\phi=1$ in Gautier and Kitamura (2009) with the scaling of Section 6 and an estimator based on smoothed projection kernels in the Fourier domain. In this paper, we use a function ψ in $\mathcal{S}(\mathbb{R})$ for the same reason for which we used smoothed projection kernels in Gautier and Kitamura (2009): to obtain rates of convergence for

[19] Upper bounds in an inverse problem setting for specific estimators are given in Hoderlein, Klemelä and Mammen (2011) and Gautier and Kitamura (2009).

all L^p risks for $1 \leq p \leq \infty$[20]. In Gautier and Le Pennec (2011) this smoothed projection kernel is used together with the Littlewood-Paley decomposition and a quadrature formula to obtain a needlet estimator. Gautier and Le Pennec (2011) provide minimax lower bounds for the estimation when $\phi = 1$ and show that their data-driven estimator is adaptive.

We have only considered the estimation of smooth functions for simplicity. If we consider "super smooth" functions, we expect that for certain functions ψ we could replace $MC(s)T_N^{-s}$ by an exponentially small term. In that case, like in statistical deconvolution (see, *e.g.*, Butucea (2004) and Butucea and Tsybakov (2007)), for a nicely behaved density of the instruments and for ϕ bounded, we could obtain parametric rates of convergence up to a logarithmic factor. Cavalier (2000) considers the estimation at a point of super smooth functions for the positron emission tomography problem in $\mathbb{R}^L$. The setup we consider is more involved because the inverse problem involves an extra derivative and we are in a regression framework with (1) random regressors whose density could be arbitrarily close to zero on its support and (2) possibly fat tails of the variables of interest. We do however consider super smooth functions in the more classical deconvolution framework in Section 4.3.4.

4.3.2. *Estimation of the Plug-in Terms for $f_{\Delta|\widetilde{\Gamma}}$.* In this section we consider the estimation of the partial Fourier transforms which are used as plug-ins in Section 4.3.4. Denote by

$$g(\gamma,t) = R^{-1}\left[\overline{\partial_v \mathbb{E}\left[e^{itY}\varsigma(D)\left|\left(\widetilde{S},\widetilde{V}\right) = \cdot, X = x\right.\right]}\right](\gamma)$$

and consider an estimator of the form (4.11)

$$\hat{g}(\gamma,t) = \frac{1}{N}\sum_{i=1}^{N}\frac{\widetilde{K}_{T_N}(\widetilde{s}_i'\gamma - \widetilde{v}_i)e^{ity_i}\varsigma(d_i)}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i), m_N\right)}$$

For technical reasons we introduce a maximum value $R_N^{\max}$ for the inverse of the smoothing parameter $h_{N,\gamma}^{-1}$[21] in the estimator (4.12). In section 4.3.4 we will only be able to adjust $h_{N,\gamma}^{-1}$ in the range $[0, R_N^{\max}]$.

Proposition 4.3. Make Assumption 4.1 and assume

(i) there exists s in $\mathbb{N}\setminus\{0\}$ and M positive such that for every t in $\mathbb{R}$, $\mathfrak{Re}\,[g(\gamma,t)]$ and $\mathfrak{Im}\,[g(\gamma,t)]$ belong to $\mathrm{W}_\infty^s(M)$;

(ii) there exists α positive such that $\log(T_N^3/m_N) + \log(R_N^{\max}) + L\log(d(\mathcal{B}_N)) \leq \alpha$;

[20]This is important to handle those plug-in terms to allow to use the whole range of the Hölder and Young inequalities.

[21]This is a classical feature of wavelet thresholding estimators and is called a maximal resolution level.

(iii) there exists a sequence $r_{IV,N}$ going to 0 as N goes to infinity and M_{IV} positive such that with probability one

$$\overline{\lim}_{N\to\infty} r_{IV,N}^{-1} \max_{i=1,\dots,N} \left| f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i) - \widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i)\right| \le M_{IV} \ ; \tag{4.15}$$

then for some constants $M(\alpha)$ (which only depends on α and L) and $C(s)$ (which only depends on s and ψ), with probability one, for every ϵ positive there exists N large enough such that

$$\begin{aligned}
&\sup_{t\in[-R_N^{\max},R_N^{\max}],\ \gamma\in\mathcal{B}_N} \left|\left(\hat{g}-g\right)(t,\gamma)\right| \\
&\le 2(M_{IV}+\epsilon) r_{IV,N} m_N^{-1}\left\{\left\|\mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma-\widetilde{V}\right)\right|}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}\left(\widetilde{S},\widetilde{V}\right),m_N\right)}\right]\right\|_\infty + m_N^{-1/2}(M(\alpha)+\epsilon)\left(\frac{\log N}{N}\right)^{1/2} T_N^{L+1/2}\right\} \\
&+ 2m_N^{-1/2}(M(\alpha)+\epsilon)\left(\frac{\log N}{N}\right)^{1/2} T_N^{L+1/2} \\
&+ 2\sup_{\gamma\in\mathcal{B}_N}\int_{\left\{(s,v):\ f_{\widetilde{S},\widetilde{V}}(s,v)<m_N\right\}} \left|\widetilde{K}_{T_N}(s'\gamma-v)\right| d\sigma(s)dv \\
&+ 2MC(s)T_N^{-s}
\end{aligned}$$

This is the same upper bound as in Proposition 4.2 up to a factor 2 (we separate the real and imaginary part of e^{ity}) and to a larger constant $M(\alpha)$), and the same remarks apply.

4.3.3. *Estimation of f_Δ.* We start with a proposition that relates the estimation of f_Δ to the estimation of $f_{\widetilde{\Gamma}}$ and of $f_{\Delta|\widetilde{\Gamma}}$. To this end, we make the following assumptions.

Assumption 4.2. $f_{\widetilde{\Gamma}}\in \mathrm{L}^\infty\left(\mathbb{R}^L\right)$ and there exists M_Δ positive such that $\sup_{\delta\in\mathbb{R},\ \gamma\in supp(\widetilde{\Gamma})} f_{\Delta|\widetilde{\gamma}}(\delta|\gamma)\le M_\Delta$.

In the next proposition we give an upper bound on the error in estimating f_Δ when we use the estimator (4.13).

Proposition 4.4. Let Assumptions (1) and 4.2 hold, then for every measurable set $\mathcal{B}_N$ in $\mathbb{R}^L$,

$$\begin{aligned}
\left\|\widehat{f_\Delta}-f_\Delta\right\|_2^2 &\le 3M_\Delta\left(\int_{\mathcal{B}_N^c} f_{\widetilde{\Gamma}}d\gamma + M_\Delta\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty^2\right) \\
&+3\|f_{\widetilde{\Gamma}}\|_\infty\left(1+\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty\right)^2\left\|\left(\widehat{f_{\Delta|\widetilde{\Gamma}}}(\cdot|\star)-f_{\Delta|\widetilde{\Gamma}}(\cdot|\star)\right)\mathbf{1}\{\star\in\mathcal{B}_N\}\right\|_2^2.
\end{aligned} \tag{4.16}$$

When $supp(\widetilde{\Gamma})$ is bounded, and $f_{\widetilde{\Gamma}}$ is bounded away from zero on its support, then we can take $\mathcal{B}_N = supp(\widetilde{\Gamma})$. Otherwise $\mathcal{B}_N$ should be (1) small enough so that $f_{\widetilde{\Gamma}}$ is bounded away from zero on $\mathcal{B}_N$ (recall as well that its diameter should not grow faster than polynomially in N to be able to apply the result from Section 4.3.1), and (2) large enough so that $\mathbb{P}(\widetilde{\Gamma} \in \mathcal{B}_N^c) = \int_{\mathcal{B}_N^c} f_{\widetilde{\Gamma}} d\gamma$ is small.

The next section studies the convergence to zero of the term

$$\left\| \left(\widehat{\overline{f_{\Delta|\widetilde{\Gamma}}}}(\cdot|\star) - \overline{f_{\Delta|\widetilde{\Gamma}}}(\cdot|\star) \right) \mathbf{1}\{\star \in \mathcal{B}_N\} \right\|_2 .$$

4.3.4. *Estimation of* $f_{\Delta|\widetilde{\Gamma}}$. In order to work with smoothing and trimming factors in (4.12) that are independent of t and γ, we work with sup-norm consistency of the estimators of the partial Fourier transforms.

Assumption 4.3.

$$\sup_{t\in[-R_N^{\max}, R_N^{\max}],\ \gamma\in\mathcal{B}_N} \left| \mathcal{F}_1 \left[\widehat{f_{Y_0+\Delta,\widetilde{\Gamma}}} \right] - \mathcal{F}_1 \left[f_{Y_0+\Delta,\widetilde{\Gamma}} \right] \right| = O_p\left(r_{Y_0+\Delta,N} \right)$$

$$\sup_{t\in[-R_N^{\max}, R_N^{\max}],\ \gamma\in\mathcal{B}_N} \left| \mathcal{F}_1 \widehat{\left[f_{Y_0,\widetilde{\Gamma}} \right]} - \mathcal{F}_1 \left[f_{Y_0,\widetilde{\Gamma}} \right] \right| = O_p\left(r_{Y_0,N} \right).$$

Recall that $R_N^{\max}$ is a maximal resolution level and we assume that $h_{N,\gamma}^{-1} \le R_N^{\max}$ and that $\mathcal{B}_N$ is a domain in $\mathbb{R}^L$ that could grow as N goes to infinity if $supp(\widetilde{\Gamma})$ is unbounded. Rates of estimation $r_{Y_0+\Delta,N}$ and $r_{Y_0,N}$ are given in Section 4.3.2.

Unlike deconvolution with noise observed on a preliminary sample, in this setup each rate is nonparametric; it is the rate of estimation in the respective inverse problems. Rates in sup norm are given in Proposition 4.2.

Proposition 4.5. Let Assumptions 2.1 (with (1)), 2.2, 3.2, 3.8, 3.9, 4.2 hold. Assume that K has support in $[-1,1]$ and that $h_{N,\gamma}^{-1} \le R_N^{\max}$. Take $t_{N,t,\gamma} = r_{Y_0,N}$. The following upper bound holds

$$\left\| \left(\widehat{\overline{f_{\Delta|\widetilde{\Gamma}}}}(\cdot|\star) - \overline{f_{\Delta|\widetilde{\Gamma}}}(\cdot|\star) \right) \mathbf{1}\{\star \in \mathcal{B}_N\} \right\|_2^2 \tag{4.17}$$

$$= O_p\left(\int_{\mathcal{B}_N \cap supp(\widetilde{\Gamma})} \int_{-\infty}^{\infty} \left[(1 - K(t\ h_{N,\gamma}))^2 \left| \mathcal{F}_1 \left[f_{\Delta|\widetilde{\Gamma}} \right](t|\gamma) \right|^2 \right.\right. \tag{4.18}$$

$$\left.\left. + \frac{K(t\ h_{N,\gamma})^2}{\left| \mathcal{F}_1 \left[f_{Y_0,\widetilde{\Gamma}} \right](t,\gamma) \right|^2} \left(r_{Y_0+\Delta,N}^2 + \left| \mathcal{F}_1 \left[f_{\Delta,\widetilde{\Gamma}} \right](t,\gamma) \right|^2 r_{Y_0,N}^2 \right) \right] dt d\gamma \right).$$

The first term in the upper bound is the square of the approximation bias. Consider now the following classes of ellipsoids for $f_{\Delta|\widetilde{\Gamma}}$

$$\mathcal{A}_{\delta,r,a}(L)=\left\{f\in \mathrm{L}^2(\mathbb{R}):\int_{-\infty}^{\infty}|\mathcal{F}[f](t)|^2\left(1+t^2\right)^{\delta}\exp\left(2a|t|^r\right)dtd\gamma\le \mathrm{L}^2\right\}$$

where $r\geq 0$, $a>0$, $\delta\in\ \mathbb{R}$ and $\delta>1/2$ if $r=0$, $l>0$. The case $r>0$ corresponds to an extension of the case of super smooth functions, otherwise the functions are extensions of ordinary smooth functions (in the Sobolev class). When $K(t)=\mathbf{1}\{|t|\le 1\}$, $h_{N,\gamma}$ is of the form $1/R_N^{\Delta}$, and f belongs to $\mathcal{A}_{\delta,r,a}(L)$ then we have

$$\int_{-\infty}^{\infty}(1-K(t\ h_{N,\gamma}))^2\,|\mathcal{F}[f](t)|^2\,dt\le \mathrm{L}^2\left(\left(R_N^{\Delta}\right)^2+1\right)^{-\delta}\exp\left(-2a\left(R_N^{\Delta}\right)^r\right).$$

The next proposition considers the case when $K(t)=\mathbf{1}\{|t|\le 1\}$ and $h_{N,\gamma}$ is of the form $1/R_N^{\Delta}$. We make the following assumption on the decay rate of $\left|\mathcal{F}_1\left[f_{Y_0|\widetilde{\Gamma}}\right](t|\gamma)\right|$ that strengthens Assumption 3.9.

Assumption 4.4. There exits $s\geq 0$, $b>0$, $\eta\in\mathbb{R}$ ($\eta>0$ if $s=0$) and $k_0,k_1>0$ such that for every γ in $supp(\widetilde{\Gamma})$,

$$k_0(1+t^2)^{-\eta/2}\exp\left(-b|t|^s\right)\le\left|\mathcal{F}_1\left[f_{Y_0|\widetilde{\Gamma}}\right](t|\gamma)\right|\le k_1(1+t^2)^{-\eta/2}\exp\left(-b|t|^s\right)$$

In the proposition below we use the short hand notation $\lambda_N=\lambda\left(\mathcal{B}_N\cap supp(\widetilde{\Gamma})\right)$, where $\lambda(B)$ is the Lebesgue measure of a set B.

Proposition 4.6. Let Assumptions 2.1 (with (1)), 2.2, 3.2, 3.8, 3.9, 4.2 and 4.3 and 4.4. Assume for every $\gamma\in supp(\widetilde{\Gamma})$, $f_{\Delta|\widetilde{\Gamma}}(\cdot|\gamma)$ belongs to $\mathcal{A}_{\delta,r,a}(L)$. The following upper bounds hold for every $R_N^{\Delta}\le R_N^{\max}$ and $\mathcal{B}_N$ measurable set in $\mathbb{R}^L$,

(A-1) if $s=r=0$, then

$$\begin{aligned}&\lambda_N^{-1}\left\|\left(\widehat{\overline{f_{\Delta|\widetilde{\Gamma}}}}(\cdot|\star)-\overline{f_{\Delta|\widetilde{\Gamma}}}(\cdot|\star)\right)\mathbf{1}\{\star\in\mathcal{B}_N\}\right\|_2^2\\&=O_p\left(\left(R_N^{\Delta}\right)^{-2\delta}+r_{Y_0+\Delta,N}^2\left(R_N^{\Delta}\right)^{2\eta+1}+r_{Y_0,N}^2\left(R_N^{\Delta}\right)^{2\max(\eta-\delta,0)+1}\right);\end{aligned}$$

(A-2) if $s>0$ and $r=0$,

$$\begin{aligned}&\lambda_N^{-1}\left\|\left(\widehat{\overline{f_{\Delta|\widetilde{\Gamma}}}}(\cdot|\star)-\overline{f_{\Delta|\widetilde{\Gamma}}}(\cdot|\star)\right)\mathbf{1}\{\star\in\mathcal{B}_N\}\right\|_2^2\\&=O_p\left(\left(R_N^{\Delta}\right)^{-2\delta}+e^{2b(R_N^{\Delta})^s}\left(r_{Y_0+\Delta,N}^2\left(R_N^{\Delta}\right)^{2\eta+1-s}+r_{Y_0,N}^2\left(R_N^{\Delta}\right)^{\min(1+2\eta-s,2(\eta-\delta))}\right)\right);\end{aligned}$$

(A-3) if $s=0$ and $r>0$, then

$$\lambda_N^{-1}\left\|\left(\widehat{\overline{f_{\Delta|\widetilde{\Gamma}}}}(\cdot|\star)-\overline{f_{\Delta|\widetilde{\Gamma}}}(\cdot|\star)\right)\mathbf{1}\{\star\in\mathcal{B}_N\}\right\|_2^2=O_p\left(\left(R_N^{\Delta}\right)^{-2\delta}e^{-2a(R_N^{\Delta})^r}+r_{Y_0+\Delta,N}^2\left(R_N^{\Delta}\right)^{2\eta+1}+r_{Y_0,N}^2\right);$$

(A-4) if $s>0$ and $r>0$, then

$$\lambda_N^{-1}\left\|\left(\widehat{\overline{f_{\Delta|\widetilde{\Gamma}}}}(\cdot|\star)-\overline{f_{\Delta|\widetilde{\Gamma}}}(\cdot|\star)\right)\mathbf{1}\{\star\in\mathcal{B}_N\}\right\|_2^2$$
$$=O_p\left(\left(R_N^{\Delta}\right)^{-2\delta}e^{-2a(R_N^{\Delta})^r}+r_{Y_0+\Delta,N}^2\left(R_N^{\Delta}\right)^{2\eta+1-s}e^{2b(R_N^{\Delta})^s}+r_{Y_0,N}^2\Delta\left(R_M^{\Delta}\right)\right)$$

where

$$\Delta\left(R_M^{\Delta}\right)=\left(R_N^{\Delta}\right)^{\min(1+2\eta-s,2(\eta-\delta))}e^{2b(R_N^{\Delta})^s}\mathbf{1}\{s>r\}+\left(R_N^{\Delta}\right)^{\max(2(\eta-\delta),0)}e^{2(b-a)(R_N^{\Delta})^s}\mathbf{1}\{r=s,b\geq a\}$$
$$+\mathbf{1}\{\{r>s\}\cup\{r=s,b<a\}\}.$$

When $supp(\widetilde{\Gamma})$ is bounded then λ_N is a constant.

We did not want to include results in other norms than the L^2 norm for brevity of exposition. However it is possible to obtain rates of convergence in sup-norm for an estimator with $K=\psi$, where ψ belongs to $\mathcal{S}(\mathbb{R})$ and $\psi(0)=1$. It would also be possible to use the sup-norm adaptive[22] estimator of Lounici and Nickl (2011).

5. Simulation study

We consider the model (1.1)-(1.2) with the added specification

$$Y_0=1+1.5\Gamma+\Theta+\varepsilon_0$$
$$Y_1=3+2.5\Gamma-\Theta+\varepsilon_1$$

where $(\Gamma,\Theta,\varepsilon_0,\varepsilon_1)=(1,-0.5,0,0)+W$, W is a centered Gaussian random vector with covariance matrix

$$\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&2&0\\0&0&0&1\end{pmatrix},$$

$\tilde{S}=(\cos N_t,\sin N_t)$ where N_t is a truncated Gaussian random variable with mean $\pi/2$ and variance $\pi^2/16$ on the interval $[0,\pi]$, V is a Gaussian random variable with mean -0.2 and variance 4 and V, N_t

[22] Specific to the case where the distribution of the error is known.

and $(\Gamma, \Theta, \varepsilon_0, \varepsilon_1)$ are independent. The sample size considered in the simulation study is $N = 10\ 000$ and we have performed $S = 100$ Monte Carlo repetitions.

We present the results with the estimators (4.3)-(4.6), using $\psi = \psi_0$. They outperformed the easier to calculate estimator (4.11) slightly. All numerical integrations were carried out by quadrature methods[23]. The choice of the smoothing parameters for the estimation of $f_{\Gamma,\Theta}$ were $T_N = 6$ for the regularized Radon inverse and $h_N = 1$ for the bandwidth of the local polynomial estimator of $\partial_v \mathbb{E}\left[D \middle| \left(\widetilde{S}, \widetilde{V}\right) = \cdot\right]$, while we took $T_N = 10$ and $h_N = 1$ for the estimation of $U - \widehat{CATE} \times f_{\Gamma,\Theta}$ and did not use truncation.

In Figures 3 and 4 we compare the truth, an empirical average of estimators over $S = 100$ simulations, and one typical simulation.

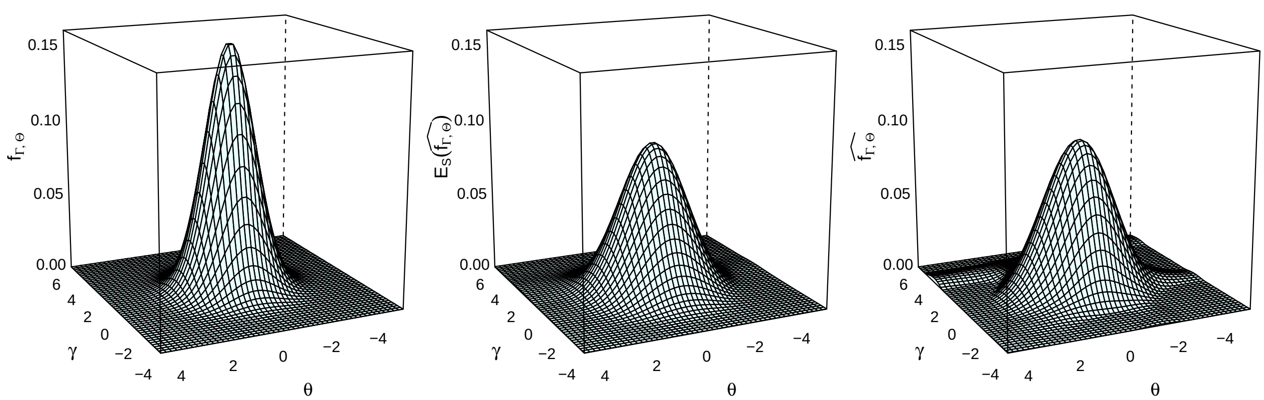

FIGURE 3. True $f_{\Gamma,\Theta}$ (left), average over S replications of the estimator (middle) and the estimator calculated for one data set (right).

Figure 3, based on the rectangle $[-4, 6] \times [-5.5, 4.5]$, shows that most of the mass of $\widehat{f_{\Gamma,\Theta}}$ is concentrated in a box around the mode at $(1, -0.5)$. Because $U - CATE$ is a conditional expectation given (Γ, Θ), it will only be well estimated at points where the density $f_{\Gamma,\Theta}$ is not too low. Based on $\widehat{f_{\Gamma,\Theta}}$ we experimented with several rectangular domains. Table 1 presents the empirical distribution

[23]We tried to apply an importance sampling Monte-Carlo method to calculate the multiple integrals using as proposals a Cauchy distribution for the integral with respect to v and a uniform distribution for the integral with respect to ϕ ($s = (\cos\phi, \sin\phi)$). Due to the presence of the $\mathcal{S}(\mathbb{R})$ function ψ_0 in K_T this Monte-Carlo approximation do have finite variance but the variance was too big to have a sufficiently good precision even using the highest possible sample size that we could generate in R. The variance of this Importance Sampling Monte-Carlo method is infinite if we replace ψ_0 by an indicator function. It is possible that another choice of rapidly decreasing function ψ can make it feasible.

over $S = 100$ replications of

$$\widehat{ATE}_j = \int_{\mathcal{B}_j} U - C\widehat{ATE} \times f_{\Gamma,\Theta}(\gamma,\theta)d\gamma d\theta$$
$$\widehat{ATT}_j = \int_{\mathcal{B}_j} h_{ATT}(\gamma)U - C\widehat{ATE} \times f_{\Gamma,\Theta}(\gamma,\theta)d\gamma d\theta$$

for estimators calculated on three such domains. The index $j = 1$ corresponds to $\mathcal{B}_1 = [-1.5, 3.5] \times [-3, 2]$ (2.5 standard errors in each direction), the index $j = 2$ corresponds to $\mathcal{B}_2 = [-1.75, 3.75] \times [-3.25, 2.25]$ (2.75 standard errors in each direction), while the index $j = 3$ corresponds to $\mathcal{B}_3 = [-2, 4] \times [-3.5, 2.5]$ (3 standard errors in each direction). For reference one should note that the true ATE is 4 while the true ATT calculated via Monte-Carlo is 4.507 (0.0016).

	Mean	**P5**	**P10**	**Median**	**P90**	**P95**
$\widehat{ATE}_1$	3.91	2.98	3.28	3.88	4.67	4.81
$\widehat{ATT}_1$	4.24	3.31	3.54	4.22	5.03	5.13
$\widehat{ATE}_2$	4.09	3.04	3.31	4.05	5.02	5.21
$\widehat{ATT}_2$	4.46	3.37	3.61	4.42	5.36	5.61
$\widehat{ATE}_3$	4.22	2.89	3.21	4.22	5.42	5.69
$\widehat{ATT}_3$	4.62	3.13	3.60	4.58	5.77	6.09

TABLE 1. The estimators indexed by 1 (resp. 2 and 3) correspond to the integration of $U - C\widehat{ATE} \times f_{\Gamma,\Theta}$ on $\mathcal{B}_1$ (resp. $\mathcal{B}_2$ and $\mathcal{B}_3$).

The plots in Figure 4 illustrate how our estimator performs for: the estimation of $U - CATE \times f_{\Gamma,\Theta}$ (top), and the estimation of $U - CATE$ only (bottom). For the former, we have used the domain $\mathcal{B}_1$, while for the latter we have employed the smaller domain $\mathcal{B}_4 = [-0.5, 2.5] \times [-2, 1]$. Note that $U - CATE \times f_{\Gamma,\Theta}$ should always be much more difficult to estimate than $f_{\Gamma,\Theta}$ because it involves a regression function with a conditional expectation with respect to (Γ, Θ) and the density of (Γ, Θ) is not bounded away from zero. In the same spirit, $U - CATE$ should be even more difficult to estimate in this simulation setup as the tails of the numerator are fatter than that of the denominator. However, this simulation study shows how estimators of the denominator and numerator of $U - CATE$ that do not perform extremely well when the risk is defined in terms of the sup-norm (see, *e.g.*, the heights of the peaks), can perform reasonably well when estimating $U - CATE$.

On the left panel of Figure 5 we compare the true $U - CDiTE(\delta, (1, -0.5))$ to an estimator with $K(t) = \exp\left(1 - \max\left\{\frac{1}{1-t^2}, 0\right\}\right)$ and $R^{\Delta}_{N,\gamma} = 0.85$. On the right panel we compare the true f_Δ

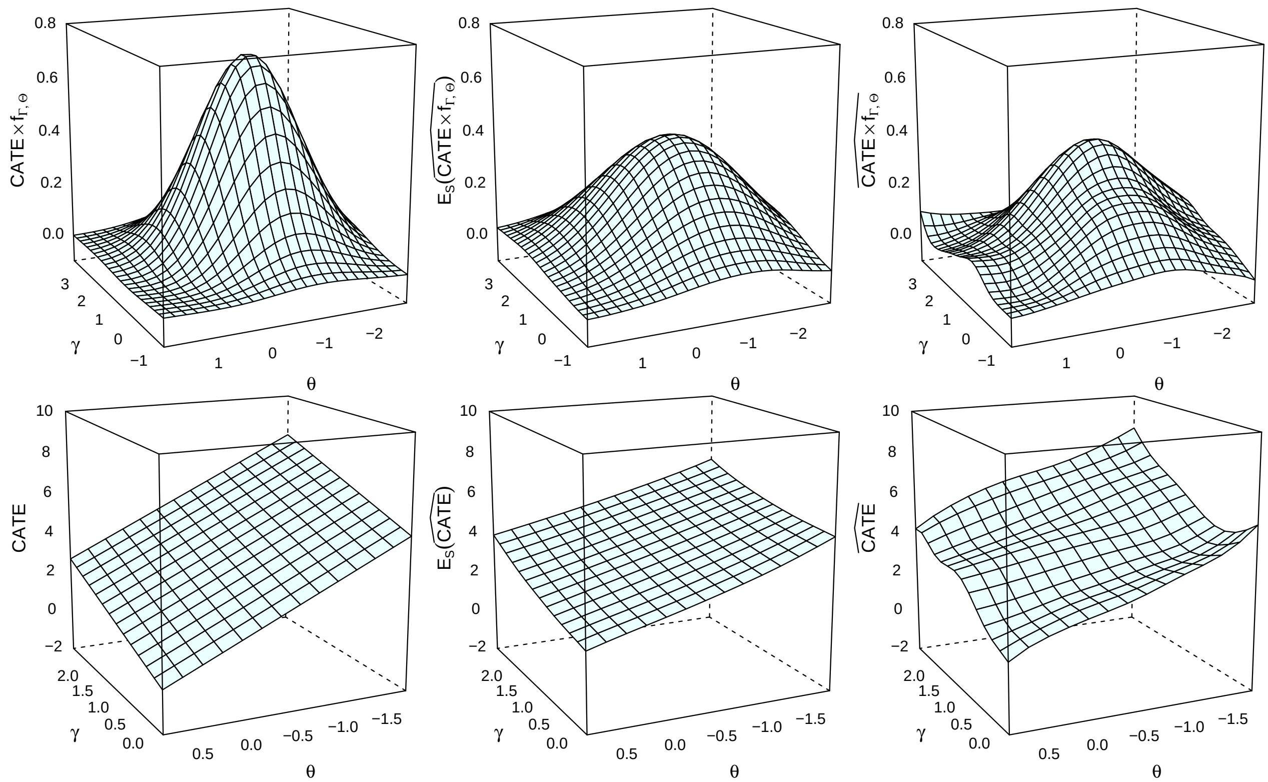

FIGURE 4. $CATE \times f_{\Gamma,\Theta}$ (up) and $CATE$ (down), truth (left), average over S replications of the estimator (middle) and the estimator calculated for one data set (right).

to an estimator obtained via a numerical integration on the box $\mathcal{B}_1$, with the same choice of $R^{\Delta}_{N,\gamma}$ and without trimming. Indeed, we tried several values of a trimming parameter and all estimates were virtually indistinguishable. Both estimators are calculated on the sample that we present on the right panel in figures 3 and 4[24].

In this simulation study we observe that the estimation of U-CATE is paradoxically more difficult than the estimation of U-CDiTE. Indeed, in our setup the denominator of U-CATE has thinner tails than the numerator. It is the opposite for U-CDiTE where the denominator has fatter tails than the numerator. Also, for U-CDiTE, the target density is super smooth and it is known that we can obtain very good rates of convergence for deconvolution estimators in that case (see the rates (A-4) in Proposition 4.6). Most quantities of interest are conditional expectations given (Γ, Θ). For these quantities it is vain to attempt to obtain an estimate at points where the density of (Γ, Θ) is too

[24]Averaging estimators calculated on our 100 replications would have taken too long given that we carry each numerical integration by quadrature methods and have to calculate many multiple integrals.

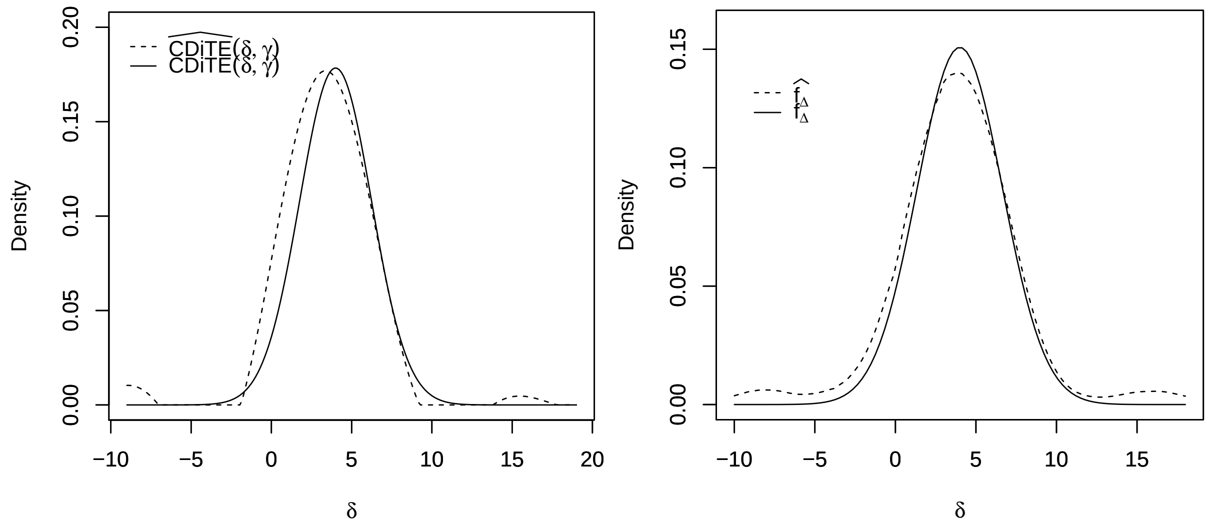

FIGURE 5. Comparison between (1) an estimator of U-CDiTE and the truth calculated at $(1,-0.5)$, the mode of (Γ,Θ), (left) and (2) an estimator of f_{Δ} and the truth.

low. We recommend to start by drawing plots of the density of (Γ,Θ), and ruling out such areas[25]. It is also more trustworthy to plot U-CATE on a small domain. A graphical representation of U-CATE allows to study the influence of the unobservables on the expected gains. With the data generating process of this simulation study it is clear both unobservables have an effect. It is thus essential to account for these two different sources of unobserved heterogeneity to get unbiased estimators of treatment effects parameters such as ATE or ATT. Obviously, the choice of a domain of integration is important when integration with respect to (γ,θ) has to be carried out. We recommend defining the domain of integration as the set $\left\{(\gamma,\theta):\ \widehat{f_{\Gamma,\Theta}}(\gamma,\theta)>\tau\right\}$ or equivalently to calculate an integral against $\widehat{f_{\Gamma,\Theta}}(\gamma,\theta)\mathbf{1}\left\{\widehat{f_{\Gamma,\Theta}}(\gamma,\theta)>\tau\right\}$. In this simulation study, the box $\mathcal{B}_1=[-1.5,3.5]\times[-3,2]$ corresponds to the choice $\tau\approx 0.021\left\|\widehat{f_{\Gamma,\Theta}}\right\|_{\infty}$ while the box $\mathcal{B}_2=[-1.75,3.75]\times[-3.25,2.25]$ (which seems to perform better to estimate the ATE and ATT) corresponds to the choice $\tau\approx 0.001\left\|\widehat{f_{\Gamma,\Theta}}\right\|_{\infty}$. The domain $\mathcal{B}_1$ contains 90% of the total mass while the domain $\mathcal{B}_2$ contains 94% of the total mass[26].

[25] Recall that the estimator of Gautier and Le Pennec (2011) is adaptive and thus achieves the minimax rate of convergence without having to chose a smoothing parameter.

[26] Indeed we have $\int_{-4}^{6}\int_{-5.5}^{4.5}\mathbb{E}_S\left[\widehat{f_{\Gamma,\Theta}}\right](\gamma,\theta)d\gamma d\theta \approx 1.014$ while $\int_{\mathcal{B}_1}\mathbb{E}_S\left[\widehat{f_{\Gamma,\Theta}}\right](\gamma,\theta)d\gamma d\theta \approx 0.916$ and $\int_{\mathcal{B}_2}\mathbb{E}_S\left[\widehat{f_{\Gamma,\Theta}}\right](\gamma,\theta)d\gamma d\theta \approx 0.952$.

6. An Alternative Scaling of the Random Coefficients Binary Choice Model

In this section, we present a different estimation approach in the case of the partial Fourier transforms. It is based on the scaling in Ichimura and Thomson (1998), Gautier and Kitamura (2009) and Gautier and Le Pennec (2011). Equation (1.2) is of the form $D = \mathbf{1}\{(V, Z', 1)(1, -\Gamma', -\Theta)' > 0\}$. Because the scale of $(1, -\Gamma', -\Theta)$ is not identified, instead of normalizing the first coordinate (the coefficient of V) to be one, we can work with the vector $\overline{\Gamma} = (1, -\Gamma', -\Theta)/\|(1, -\Gamma', -\Theta)\|$[27] which is of norm 1. This yields slightly more flexibility because for identification we have just have to assume that the support of $\overline{\Gamma}$ belongs to an hemisphere (see Gautier and Kitamura (2009)) and it is not required that in the original scale one specific coefficient is positive. We also rescale the instruments so that $S = (V, Z', 1)'/\|(V, Z', 1)\|$. Both S and $\overline{\Gamma}$ belong to the sphere $\mathbb{S}^L$ of the Euclidian space $\mathbb{R}^{L+1}$. We advertise this alternative approach here, because rescaling by one coefficient is potentially instable numerically, if one chooses as a "special" regressor V one for which the coefficient in the original scale of utilities can be close to zero for some individuals.

Under the above scaling (1.2) becomes $D = \mathbf{1}\{S'\overline{\Gamma} > 0\}$. Consider for example the case of $\mathcal{F}_1\left[f_{Y_0,\overline{\Gamma}}\right]$.

$$\begin{aligned}
\mathbb{E}[(1-D)e^{itY}|S=s] &= \mathbb{E}[(1-D)e^{itY_0}|S=s] \\
&= \mathbb{E}[e^{itY_0}] - \mathbb{E}[\mathbf{1}\{s'\overline{\Gamma} > 0\}e^{itY_0}] \quad (\text{using } (A-2)) \\
&= \mathbb{E}[e^{itY_0}] - \int_{\mathbb{S}^L} \mathbf{1}\left\{s'\gamma > 0\right\}\left(\mathcal{F}_1\left[f_{Y_0,\overline{\Gamma}}\right](t,\gamma)\right) d\sigma(\gamma) \\
&= \mathbb{E}[e^{itY_0}] - \mathcal{H}\left(\mathcal{F}_1\left[f_{Y_0,\overline{\Gamma}}\right](t,\cdot)\right)(s) \\
&= \frac{1}{2}\mathbb{E}[e^{itY_0}] - \mathcal{H}\left(\left(\mathcal{F}_1\left[f_{Y_0,\overline{\Gamma}}\right](t,\cdot)\right)^-\right)(s) \qquad (6.1)
\end{aligned}$$

where σ is the spherical measure on $\mathbb{S}^L$, $\mathcal{H}$ is the hemispherical transform (see, *e.g.*, Gautier and Kitamura (2009)) and f^- is the odd part[28] of a function f.

Assumption 6.1. (A-1) The rescaled vector of instruments S has a density with respect to σ and its support is the whole hemisphere $H^+ = \{s \in \mathbb{S}^L : s_{L+1} \geq 0\}$.

[27] Though the notation $\overline{\Gamma}$ has already been used in Section 3.4.1, here it is denoting something different.

[28] Odd, respectively even, functions are the closure in $L^2(\mathbb{S}^L)$ of continuous functions such that $\forall s \in \mathbb{S}^L,\ f(-s) = -f(s)$, respectively $\forall s \in \mathbb{S}^L,\ f(-s) = f(s)$. Each function in $L^2(\mathbb{S}^L)$ is the sum of its odd and even part. We denote by $L^2_{odd}(\mathbb{S}^L)$ the subspace of $L^2(\mathbb{S}^L)$ of odd functions.

(A-2) Γ has a density f_Γ with respect to σ which is defined point-wise and has support included in some hemisphere $H = \{s \in \mathbb{S}^L : \ s'\,\mathbf{n} \geq 0\}$, where $\mathbf{n}$ is a vector of norm 1 that does not need to be known.

Assumption 6.1 (1) corresponds to full support of the instruments. It is stronger than Assumption 2.2. Assumption 6.1 (2) is satisfied under the specification (1.1) where in the original scale the coefficient of V has a sign which is known, but it is slightly more general. Equation (6.1) yield that, under Assumption 6.1 (2), $\mathbb{E}[(1-D)e^{itY}|S=s] - \frac{1}{2}\mathbb{E}[e^{itY_0}]$ can be extended in a unique way as an odd function defined on the whole $\mathbb{S}^L$ (it is initially only defined on H^+ according to Assumption 6.1 (1)) through

$$\forall s \in H^+, \ R_{Y_0}(t,s) = \mathbb{E}[(1-D)e^{itY}|S=s] - \frac{1}{2}\mathbb{E}[e^{itY_0}]$$
$$\forall s \in -H^+, \ R_{Y_0}(t,s) = -R_{Y_0}(t,-s).$$

It is remarkable that $\mathbb{E}[e^{itY_0}]$ is also identified in this model. This is due to the smoothing properties of $\mathcal{H}$. Indeed (see, *e.g.*, Gautier and Kitamura (2009)), because $R_{Y_0}(t,\cdot)$ belong to $\mathcal{H}(\mathrm{L}^2_{odd}(\mathbb{S}^L))$, it is continuous and odd. Thus for any point $\tilde{s}$ on the boundary of H^+, $R_{Y_0}(t,\tilde{s}) = -R_{Y_0}(t,-\tilde{s})$. This yields

$$\lim_{s\to\tilde{s},\ s\in H^+} \mathbb{E}[(1-D)e^{itY}|S=s] + \lim_{s\to-\tilde{s},\ s\in H^+} \mathbb{E}[(1-D)e^{itY}|S=s] = \mathbb{E}[e^{itY_0}]. \tag{6.2}$$

Because the right hand side does not depend on $\tilde{s}$, an efficient estimator would take into account all these relations for all $\tilde{s}$ on the boundary of H^+. Given an estimator $\widehat{\mathcal{F}[f_{Y_0}]}(t)$ of $\mathbb{E}[e^{itY_0}]$, we can get an estimator of $\mathcal{F}_1[f_{Y_0,\Gamma}]$ with the same formulas as in Gautier and Kitamura (2009) or Gautier and Le Pennec (2011) replacing $2y_i - 1$ by $2(x_i-1)e^{ity_i} + \widehat{\mathcal{F}_1[f_{Y_0}]}(t)$. In the case of the estimator of Gautier and Kitamura (2009) (see the reference for more details) and delayed means smoothing kernels we get

$$\widehat{\mathcal{F}_1\left[f_{Y_0,\overline{\Gamma}}\right]}(t,\gamma) = \max\left(\frac{2}{|\mathbb{S}^L|}\sum_{p=0}^{T_N-1}\frac{\chi(2p+1,2T_N)h(2p+1,L)}{\lambda(2p+1,L)C_{2p+1}^{\nu(L)}(1)}\left(\frac{1}{N}\sum_{i=1}^N\frac{\left(2(x_i-1)e^{ity_i}+\widehat{\mathcal{F}[f_{Y_0}]}(t)\right)C_{2p+1}^{\nu(L)}(s'\gamma)}{\max\left(\hat{f}_S(s_i),m_N\right)}\right),0\right), \tag{6.3}$$

where $|\mathbb{S}^L| = \frac{2\pi^{(L+1)/2}}{\Gamma((L+1)/2)}$ is the surface measure of $\mathbb{S}^L$, $h(n,L) = \frac{(2n+L-1)(n+L-1)!}{n!(L-1)!(n+L-1)}$, $\nu(L) = (L-1)/2$, $\lambda(2p+1,L) = \frac{(-1)^p|\mathbb{S}^{L-1}|1\cdot3\cdots(2p-1)}{(L)(L+2)\cdots(L+2p)}$, $\chi(n,T) = \psi(n/T)$ where $\psi:\ [0,\infty)\to[0,\infty)$ is infinitely differentiable, nonincreasing, such that $\psi(x)=1$ if $x\in[0,1]$, $0\leq\psi(x)\leq1$ if $x\in[1,2]$, $\psi(x)=0$ if

$x \geq 2$, and $C_n^{\nu}(\cdot)$ are the Gegenbauer polynomials [29]. T_N is the smoothing parameter, m_N a trimming factor and $\hat{f}_S$ an estimator of the density of S.

7. Appendix

7.0.5. *Proof of Proposition 4.1.* Because K_T and $\widetilde{K}_T$ belong to $\mathcal{S}(\mathbb{R})$, due to Assumption 4.1 (ii), for the function ϕ considered, for almost every s in H^+, $v \mapsto \mathbb{E}[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V})=(s,v)]\widetilde{K}_T(v)$ and $v \mapsto \partial_v\mathbb{E}[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V})=(s,v)]K_T(v)$ are in $\mathrm{L}^1(\mathbb{R})$ and $\lim_{|v|\mapsto\infty}\mathbb{E}[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V})=(s,v)]K_T(v)=0$. This yields

$$\begin{aligned}
&A_T\left[\overline{\partial_v\mathbb{E}\left[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V})=(\cdot),X=x\right]}\right](\gamma)\\
&=\int_{supp\left(f_{\widetilde{S},\widetilde{V}|X}(\cdot|x)\right)}\widetilde{K}_T(s'\gamma-u)\mathbb{E}\left[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V})=(s,u),X=x\right]dud\sigma(s)\quad\text{(by integration by parts)}\\
&=\int_{supp\left(f_{\widetilde{S},\widetilde{V}|X}(\cdot|x)\right)}\widetilde{K}_T(s'\gamma-u)\mathbb{E}\left[\phi(Y)\varsigma(D)|(\widetilde{S},\widetilde{V})=(s,u),X=x\right]\frac{f_{\widetilde{S},\widetilde{V}|X}(s,u|x)}{f_{\widetilde{S},\widetilde{V}|X}(s,u|x)}dud\sigma(s)\\
&=\int_{supp\left(f_{\widetilde{S},\widetilde{V}|X}(\cdot|x)\right)}\mathbb{E}\left[\left.\frac{\widetilde{K}_T(s'\gamma-u)\phi(Y)\varsigma(D)}{f_{\widetilde{S},\widetilde{V}|X}(s,u|x)}\right|(\widetilde{S},\widetilde{V})=(s,u),X=x\right]f_{\widetilde{S},\widetilde{V}|X}(s,u|x)dud\sigma(s)\\
&=\mathbb{E}\left[\left.\frac{\widetilde{K}_T(\widetilde{S}^T\gamma-U)\phi(Y)\varsigma(D)}{f_{\widetilde{S},\widetilde{V}|X}(\widetilde{S},\widetilde{V}|x)}\right|X=x\right]\quad\text{(by the law of iterated conditional expectations)}
\end{aligned}$$

Q.E.D.

7.0.6. *Proof of Proposition 4.2.* We use the notations

$$g^I_{m,\tau}(\gamma)=\frac{1}{N}\sum_{i=1}^N\frac{\widetilde{K}_{T_N}(\widetilde{s}_i'\gamma-\widetilde{v}_i)T_{\tau_N}(\phi(y_i))\varsigma(d_i)}{\max\left(f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)},$$

$$g^I_{\tau}(\gamma)=\frac{1}{N}\sum_{i=1}^N\frac{\widetilde{K}_{T_N}(\widetilde{s}_i'\gamma-\widetilde{v}_i)T_{\tau_N}(\phi(y_i))\varsigma(d_i)}{f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i)},$$

$$g^I(\gamma)=\frac{1}{N}\sum_{i=1}^N\frac{\widetilde{K}_{T_N}(\widetilde{s}_i'\gamma-\widetilde{v}_i)\phi(y_i)\varsigma(d_i)}{f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i)},$$

[29] The Gegenbauer polynomials are given by

$$C_n^{\nu}(t)=\sum_{l=0}^{[n/2]}\frac{(-1)^l(\nu)_{n-l}}{l!(n-2l)!}(2t)^{n-2l},\quad \nu>-1/2,n\in\mathbb{N}$$

where $(a)_0=1$ and for n in $\mathbb{N}\setminus\{0\}$, $(a)_n=a(a+1)\cdots(a+n-1)=\Gamma(a+n)/\Gamma(a)$.

where the superscript I stands for ideal (this is because we replace the estimator of the density in the denominator by the true density). For two sequences of positive numbers $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n \lesssim b_n$ when there exists M positive such that $a_n \leq M b_n$ and $a_n \asymp b_n$ when $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Let us start by stating a few results on $\widetilde{K}_T$. Recall that

$$\widetilde{K}_T(u) = \frac{2}{(2\pi)^L}\int_0^{\infty} \sin(-tu)t|t|^{L-1}\psi\left(\frac{t}{T}\right)dt$$

therefore by the change of variables

$$\widetilde{K}_T(u) = \frac{2T^{L+1}}{(2\pi)^L}\int_0^{\infty} \sin(-Ttu)t|t|^{L-1}\psi(t)dt$$

thus

$$\left|\widetilde{K}_T(u)\right| \leq \frac{2T^{L+1}}{(2\pi)^L}\int_0^{\infty} t^L\psi(t)dt$$

and $\int_0^\infty t^L\psi(t)dt$ is a constant independent of T because $\psi \in \mathcal{S}(\mathbb{R}^L)$, therefore

$$\left|\widetilde{K}_T\right|_\infty \lesssim T^{L+1}. \tag{7.1}$$

Similarly we can show that

$$\left|\widetilde{K}_T'\right|_\infty \lesssim T^{L+2} \tag{7.2}$$

which implies that $\forall (u,v)\in\mathbb{R}^2$,

$$\left|\widetilde{K}_T(u) - \widetilde{K}_T(v)\right|_\infty \lesssim T^{L+2}|u-v|$$

which in turns yields

$$\left||\widetilde{K}_T(u)| - |\widetilde{K}_T(v)|\right|_\infty \lesssim T^{L+2}|u-v|$$

and $\forall (s,v)\in\mathbb{S}^{L-1}\times\mathbb{R}$,

$$\left|\widetilde{K}_T(s'\gamma - v) - \widetilde{K}_T(s'\overline{\gamma} - v)\right| \lesssim T^{L+2}|\gamma - \overline{\gamma}| \tag{7.3}$$

$$\left||\widetilde{K}_T(s'\gamma - v)| - |\widetilde{K}_T(s'\overline{\gamma} - v)|\right| \lesssim T^{L+2}|\gamma - \overline{\gamma}| \tag{7.4}$$

where on the right hand side of (7.3) and (7.4) $|\cdot|$ is the Euclidian norm in $\mathbb{R}^L$.
Also, because

$$\widetilde{K}_T(u) = \frac{1}{(2\pi)^L}\int_{-\infty}^{\infty} e^{-iut}t|t|^{L-1}\psi\left(\frac{t}{T}\right)dt,$$

$$\left|\widetilde{K}_T\right|_2 = \left\{\int_{-\infty}^{\infty} t^{2L}\psi^2\left(\frac{t}{T}\right)dt\right\}^{1/2}$$

$$= T^{(2L+1)/2}\left\{\int_{-\infty}^{\infty} t^{2L}\psi^2(t)dt\right\}^{1/2}$$

$$\lesssim T^{(2L+1)/2}. \tag{7.5}$$

We now rely on the decomposition

$$\hat{g}-g=\left(\hat{g}-g^I_{m,\tau}\right)+\left(g^I_{m,\tau}-\mathbb{E}\left[g^I_{m,\tau}\right]\right)+\left(\mathbb{E}\left[g^I_{m,\tau}\right]-\mathbb{E}\left[g^I_{\tau}\right]\right)+\left(\mathbb{E}\left[g^I_{\tau}\right]-\mathbb{E}\left[g^I\right]\right)+\left(\mathbb{E}\left[g^I\right]-g\right)$$
$$:= S_{\rm p}+S_{\rm e}+B_{\rm t}+B_{\rm trunc}+B_{\rm a}$$

where the expectation is with respect to $(\widetilde{S}_i,\widetilde{V}_i)_{i=1}^N$[30]. The contribution $S_{\rm p}$ corresponds to the stochastic component due to plug-in, $S_{\rm e}$ to the stochastic component of the infeasible estimator $g^I_{m,\tau}$, $B_{\rm t}$ to the trimming bias, $B_{\rm trunc}$ to the bias due to truncation and $B_{\rm a}$ to the approximation bias.

Let us study first the contribution of the term $S_{\rm p}$. The following upper bounds hold

$$\|S_{\rm p}\|_\infty=\left\|\frac{1}{N}\sum_{i=1}^N\frac{\widetilde{K}_{T_N}(\widetilde{s}_i'\cdot-\widetilde{v}_i)T_{\tau_N}(\phi(y_i))\varsigma(d_i)}{\max\left(f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}\left(\frac{\max\left(f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}-1\right)\mathbf{1}\{\cdot\in\mathcal{B}_N\}\right\|_\infty$$
$$\le\min\left(\tau_N,\|\phi\|_\infty\right)\left\|\frac{1}{N}\sum_{i=1}^N\frac{\left|\widetilde{K}_{T_N}(\widetilde{s}_i'\cdot-\widetilde{v}_i)\right|\mathbf{1}\{\cdot\in\mathcal{B}_N\}}{\max\left(f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}\right\|_\infty\max_{i=1,\dots,N}\left|\frac{\max\left(f_{\widetilde{S},\widetilde{V})}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}{\max\left(\widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}-1\right|$$
$$\le m_N^{-1}\min\left(\tau_N,\|\phi\|_\infty\right)\left\|\frac{1}{N}\sum_{i=1}^N\frac{\left|\widetilde{K}_{T_N}(\widetilde{s}_i'\cdot-\widetilde{v}_i)\right|\mathbf{1}\{\cdot\in\mathcal{B}_N\}}{\max\left(f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}\right\|_\infty\max_{i=1,\dots,N}\left|f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i)-\widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i)\right|$$
$$\le m_N^{-1}\min\left(\tau_N,\|\phi\|_\infty\right)\left(\|T_1\|_\infty+\|T_2\mathbf{1}\{\mathcal{B}_N\}\|_\infty\right)\max_{i=1,\dots,N}\left|f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i)-\widehat{f_{\widetilde{S},\widetilde{V}}}(\widetilde{s}_i,\widetilde{v}_i)\right|$$

where

$$T_1(\gamma)=\mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma-\widetilde{V}\right)\right|}{\max\left(f_{\widetilde{S},\widetilde{V}}\left(\widetilde{S},\widetilde{V}\right),m_N\right)}\right]$$
$$T_2(\gamma)=\frac{1}{N}\sum_{i=1}^N\left\{\frac{\left|\widetilde{K}_{T_N}(\widetilde{s}_i'\gamma-\widetilde{v}_i)\right|}{\max\left(f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i),m_N\right)}-\mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma-\widetilde{V}\right)\right|}{\max\left(f_{\widetilde{S},\widetilde{V}}\left(\widetilde{S},\widetilde{V}\right),m_N\right)}\right]\right\}$$

We just have to consider the term $\|T_2\mathbf{1}\{\mathcal{B}_N\}\|_\infty$. We cover $\mathcal{B}_N$ by $\mathfrak{N}(N,L)$ Euclidian balls $(B_i)_{i=1}^{\mathfrak{N}(N,L)}$ of centers $(\overline{\gamma}_i)_{i=1}^{\mathfrak{N}(N,L)}$ and radius $R(N,L)$. Because $\mathcal{B}_N$ is compact we have $\mathfrak{N}(N,L)\asymp d(\mathcal{B}_N)^L R(N,L)^{-L}$.

[30]Thus we do not integrate against the distribution of $\widetilde{\Gamma}$.

For $M(\alpha)$ positive and an appropriately chosen sequence (v_N) to be defined later

$$\mathbb{P}\left(v_N\|T_2\mathbf{1}\{\mathcal{B}_N\}\|_\infty \geq M(\alpha)\right) \tag{7.6}$$

$$\leq \mathbb{P}\left(\bigcup_{i=1,\dots,\mathfrak{N}(N,L)} \{v_N\left|T_2(\overline{\gamma}_i)\right| \geq M(\alpha)/2\}\right)$$

$$+\,\mathbb{P}\left(\exists i \in \{1,\dots,\mathfrak{N}(N,L)\}:\ v_N \sup_{\gamma\in B_i}\left|T_2(\gamma)-T_2(\overline{\gamma}_i)\right| \geq M(\alpha)/2\right). \tag{7.7}$$

By taking $R(N,L) \asymp m_N v_N^{-1} T_N^{-(L+2)} M(\alpha)$ for a well chosen constant, the first term on the right hand side is equal to zero. This follows from the fact that T_2 is Lipschitz with a constant proportional to $m_N^{-1}T_N^{-(L+2)}$. This is a consequence of (7.4). For such a choice of $R(N,L)$,

$$\mathbb{P}\left(v_N\|T_2\|_\infty \geq M(\alpha)\right) \leq \mathfrak{N}(N,L) \sup_{i=1,\dots,\mathfrak{N}(N,L)} \mathbb{P}\left(v_N\left|T_2(\overline{\gamma}_i)\right| \geq M(\alpha)/2\right). \tag{7.8}$$

Now

$$\mathbb{P}\left(v_N\left|T_2(\overline{\gamma}_i)\right| \geq M(\alpha)/2\right)$$

$$= \mathbb{P}\left(\left|\sum_{j=1}^N \frac{\left|\widetilde{K}_{T_N}\left(\widetilde{s}_j'\overline{\gamma}_i - \widetilde{v}_j\right)\right|}{\max\left(f_{\widetilde{S},\widetilde{V}}(\widetilde{s}_i,\widetilde{v}_i), m_N\right) m_N^{-1}T_N^{L+1}} - \mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\overline{\gamma}_i - \widetilde{V}\right)\right|}{\max\left(f_{\widetilde{S},\widetilde{V}}\left(\widetilde{S},\widetilde{V}\right), m_N\right) m_N^{-1}T_N^{L+1}}\right]\right| \geq t\right)$$

$$\leq 2\exp\left\{-\frac{1}{2}\left(\frac{t^2}{\omega + Lt/3}\right)\right\} \quad \text{(Bernstein inequality)} \tag{7.9}$$

where

$$t = T_N^{-(L+1)} v_N^{-1} m_N N M(\alpha)/2$$

$$\omega \geq \sum_{j=1}^N \mathrm{var}\left(\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{s}_j'\overline{\gamma}_i - \widetilde{v}_j\right)\right|}{m_N^{-1}T_N^{L+1}}\right)$$

$$L \geq \sup_{(s,v)\in supp(\widetilde{S},\widetilde{V})}\left|\frac{\widetilde{K}_{T_N}\left(s'\overline{\gamma}_i - v\right)}{\max\left(f_{\widetilde{S},\widetilde{V}}\left(\widetilde{S},\widetilde{V}\right), m_N\right) m_N^{-1}T_N^{L+1}}\right| \asymp 1 \quad \text{(using (7.1))}.$$

As

$$\sum_{j=1}^N \mathrm{var}\left(\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}_j'\overline{\gamma}_i - \widetilde{V}_j\right)\right|}{\max\left(f_{\widetilde{S},\widetilde{V}}\left(\widetilde{S},\widetilde{V}\right), m_N\right) m_N^{-1}T_N^{L+1}}\right) \leq \frac{m_N^2}{T_N^{2(L+1)}} \sum_{j=1}^N \mathbb{E}\left[\left(\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\overline{\gamma}_i - \widetilde{V}\right)\right|}{\max\left(f_{\widetilde{S},\widetilde{V}}\left(\widetilde{S},\widetilde{V}\right), m_N\right)}\right)^2\right]$$

$$\leq \frac{m_N N T_N^{2L+1}}{T_N^{2(L+1)}} \quad \text{(Due to (7.5))}$$

we shall take $\omega = m_N N T_N^{2L+1} T_N^{-2(L+1)}$.

Now choose v_N such that $t \asymp M(\alpha)\sqrt{\omega \log(N)}$. Thus ω is the leading term in the denominator of the exponent in (7.9). The corresponding v_N is

$$v_N \asymp (\log N)^{-1/2} \omega^{-1/2} T_N^{-(L+1)} m_N N$$
$$\asymp \left(\frac{N}{\log N}\right)^{1/2} m_N^{1/2} T_N^{-(L+1/2)}.$$

For these choices of the parameters

$$\frac{t^2}{\omega + Lt/3} \asymp (\log N) M(\alpha)^2 \tag{7.10}$$

and

$$R(N,L) \asymp \left(\frac{N}{\log N}\right)^{-1/2} m_N^{1/2} T_N^{-3/2} M(\alpha).$$

Due to (7.5) and because by assumption $\log(T_N^3/m_N) + L\log(d(\mathcal{B}_N)) \leq \alpha$, we obtain

$$\mathfrak{N}(N,L) \asymp d(\mathcal{B}_N)^L R(N,L)^{-L} = \exp\left((\alpha + L/2)\log N + o(\log N)\right). \tag{7.11}$$

Equations (7.8), (7.9), (7.10) and (7.11) imply that, for a positive constants C and C_2,

(7.12)

$$\mathbb{P}\left(\left(\frac{N}{\log N}\right)^{1/2} T_N^{-(L+1/2)} m_N^{1/2} \|T_2 \mathbf{1}\{\mathcal{B}_N\}\|_\infty \geq M(\alpha)\right) \leq C \exp\left\{(\log N)((\alpha + L/2) - C_2 M(\alpha)^2)\right\}$$

holds. For a large enough $M(\alpha)$, $(\alpha + L/2) - C_2 M(\alpha)^2 < -1$ which implies summability of the left hand side in (7.12), hence by the first Borel-Cantelli lemma for $M(\alpha)$ large enough with probability one

$$\overline{\lim}_{N\to\infty} \left(\frac{N}{\log N}\right)^{1/2} T_N^{-(L+1/2)} m_N^{1/2} \|T_2\|_\infty < M(\alpha).$$

In summary, we have obtained that for some constant M_{IV} and $M(\alpha)$, with probability one, for every ϵ positive, there exists N large enough such that

$$\|S_{\mathrm{p}} \mathbf{1}\{\mathcal{B}_N\}\|_\infty \leq (M_{IV} + \epsilon) \min(\tau_N, \|\phi\|_\infty) r_{IV,N} m_N^{-1} \left\{ \left\| \mathbb{E}\left[\frac{\left|\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma - \widetilde{V}\right)\right|}{\max\left(f_{\widetilde{S},\widetilde{V}}\left(\widetilde{S},\widetilde{V}\right), m_N\right)}\right]\right\|_\infty \right.$$
$$\left. + m_N^{-1/2}(M(\alpha) + \epsilon)\left(\frac{N}{\log N}\right)^{-1/2} T_N^{L+1/2} \right\}.$$

For the same reason, on the same event of probability 1, for every ϵ positive, there exists N large enough such that

$$\|S_{\rm e}\mathbf{1}\{\mathcal{B}_N\}\|_\infty \le m_N^{-1/2}(M(\alpha)+\epsilon)\min\left(\tau_N, \|\phi\|_\infty\right)\left(\frac{N}{\log N}\right)^{-1/2} T_N^{L+1/2}.$$

Consider now the bias term induced by trimming, evaluated at a point γ,

$$\begin{aligned} B_{\rm t}(\gamma) &= \mathbb{E}\left[\frac{\widetilde{K}_{T_N}\left(\widetilde{S}'\gamma-\widetilde{V}\right)\min\left(\phi(Y),\tau_N\right)\varsigma(D)}{f_{\widetilde{S},\widetilde{V}}(\widetilde{S},\widetilde{V})}\left(\frac{f_{\widetilde{S},\widetilde{V}}(\widetilde{S},\widetilde{V})}{\max\left(f_{\widetilde{S},\widetilde{V}}(\widetilde{S},\widetilde{V}),m_N\right)}-1\right)\right] \\ &= \int_{\{(s,v):\ f_{\widetilde{S},\widetilde{V}}(s,v)<m_N\}} \mathbb{E}\left[\min\left(\phi(Y),\tau_N\right)\varsigma(D)|\widetilde{S}=s,\widetilde{V}=v\right]\widetilde{K}_{T_N}\left(s'\gamma-v\right)\left(f_{\widetilde{S},\widetilde{V}}(s,v)m_N^{-1}-1\right)d\sigma(s)dv. \end{aligned}$$

This yields the following upper bound

$$|B_{\rm t}(\gamma)| \le \min\left(\tau_N, \|\phi\|_\infty\right)\int_{\left\{(s,v):\ f_{\widetilde{S},\widetilde{V}}(s,v)<m_N\right\}}\left|\widetilde{K}_{T_N}(s'\gamma-v)\right|d\sigma(s)dv.$$

Consider now the truncation bias $B_{\rm trunc}$. We obtain

$$|B_{\rm trunc}| \le \mathbb{E}\left[\left|\widetilde{K}_{T_N}(\widetilde{S}^T\gamma-\widetilde{V})T_{\tau_N}(\phi(Y))\zeta(D)\right|\right]$$

which allows to conclude using (7.1) with an explicit constant.

The upper bound for the approximation bias $B_{\rm a}$ is obtained as follows. Note that, for x in $\mathbb{R}^L$,

$$\begin{aligned} \mathbb{E}[g^I](x) &= \mathcal{F}^{-1}\left[\psi\left(\frac{\cdot}{T}\right)\mathcal{F}[g](\cdot)\right](x) \\ &= \rho_T * g(x) \end{aligned}$$

where $*$ is the usual convolution and

$$\begin{aligned} \rho_T(x) &= \frac{1}{(2\pi)^L}\int_{-\infty}^{\infty} e^{-i\xi' x}\psi\left(\frac{\xi}{T}\right)d\xi \\ &= T^L\rho_1\left(\frac{x}{T}\right). \end{aligned} \tag{7.13}$$

The collection $(\rho_T)_{T>0}$ is an approximate identity because of (7.13) and $\int_{-\infty}^{\infty}\rho_T(x)dx = 1 (= \psi(0))$. The rest of the argument is classical and is based on

$$g - \mathbb{E}[g^I](x) = \int_{\mathbb{R}^L}\left(g(x)-g(x-y)\right)\rho_T(y)dy.$$

Let us do the argument for $s=1$ and $s=2$ only for simplicity of the notations.

Case where $s=1$.

The inequalities

$$\|g - \mathbb{E}[g^I]\|_\infty \le L_g\int_{\mathbb{R}^L}|y||\rho_T(y)|dy$$

$$\leq \|g\|_{1,\infty} T^{-1} \int_{\mathbb{R}^L} |y||\rho_1(y)|dy$$

hold with L_g the Lipschitz constant of g which is itself upper bounded by $\|g\|_{1,\infty}$. The last integral is finite because ρ_1 is in $\mathcal{S}(\mathbb{R}^L)$. Indeed ρ_1 is the Fourier transform of a function in $\mathcal{S}(\mathbb{R}^L)$.

Case where $s = 2$.

Denoting by $Dg(x).y$ the differential of g at x applied to y, because ψ and thus ρ_1 is symmetric,

$$\|g - \mathbb{E}[g^I]\|_\infty = \int_{\mathbb{R}^L} (g(x) - g(x-y) - Dg(x).y)\rho_T(y)dy.$$

This yields

$$\begin{aligned}
\|g - \mathbb{E}[g^I]\|_\infty &= \int_{\mathbb{R}^L} \int_0^1 (Dg(x-\lambda y) - Dg(x)).y d\lambda \rho_T(y)dy \\
&\leq \|g\|_{2,\infty} \int_{\mathbb{R}^L} |y|^2 |\rho_T(y)|dy \\
&\leq \|g\|_{2,\infty} T^{-2} \int_{\mathbb{R}^L} |y|^2 |\rho_1(y)|dy
\end{aligned}$$

where again $\int_{\mathbb{R}^L} |y|^2|\rho_1(y)|dy < \infty$ because ρ_1 is in $\mathcal{S}(\mathbb{R}^L)$.

The upper bounds on the bias due to truncation follow from the expression of the difference between the two expectations when Assumption 2.1 holds.

Q.E.D.

7.0.7. *Proof of Proposition 4.3.* The proof of this result is almost the same as the proof of Proposition 4.2. We will thus only stress the differences. We will use the notation

$$\|f - g\|_\infty := \sup_{t \in [-R_N^{\max}, R_N^{\max}],\ \gamma \in \mathcal{B}_N} |(f-g)(t,\gamma)|.$$

We start by observing that

$$\|f - g\|_\infty \leq \|\mathfrak{Re}(f) - \mathfrak{Re}(g)\|_\infty + \|\mathfrak{Im}(f) - \mathfrak{Im}(g)\|_\infty$$

this yields the factor 2 in the upper bound of the proposition. Then it is easy to check that we obtain the same upper bounds for both the error on the estimation of the real part and the error on the estimation of the imaginary part. For both, all the terms can be bounded like in the proof of Proposition 4.2 (taking $\tau_N = 1$ and noting that $\|\phi\|_\infty = 1$ as here ϕ is either cos or sin) besides the term $S_{\rm a}$, the stochastic component of the infeasible estimator $g^I_{m,\tau}$.

We shall cover $\mathcal{B}_N$ by $\overline{\mathfrak{N}}(N,L)$ balls $(B_i)_{i=1}^{\overline{\mathfrak{N}}(N,L)}$ of centers $\left(\overline{\gamma}_i,\overline{t}_i\right)_{i=1}^{\overline{\mathfrak{N}}(N,L)}$ and radius $R(N,L)$ (will be the same as in the proof of Proposition 4.2) where balls are defined as

$$B_i=\left\{(\gamma,t):\ |\gamma-\overline{\gamma}_i|+|t-\overline{t}_i|\le R(N,L)\right\}$$

and again the norm $|\gamma-\overline{\gamma}_i|$ is the Euclidian norm in $\mathbb{R}^L$ while $|t-\overline{t}_i|$ is the absolute value. Because $B_d(0,1)\times[-1,1]$ is compact[31] it can be covered by a number of balls of the order of $R(N,L)^{-(L+1)}$ (the extra dimension due to t) and thus $\overline{\mathfrak{N}}(N,L)\asymp R_N^{\max}d(\mathcal{B}_N)^L R(N,L)^{-(L+1)}$.

The choice of $R(N,L)$ is based on the same reasoning as before and the fact that the functions

$$(\gamma,t)\to\frac{\widetilde{K}_{T_N}(s'\gamma-v)\cos(ty)}{\max\left(f_{\widetilde{S},\widetilde{V}}(s,v),m_N\right)}$$

and

$$(\gamma,t)\to\mathbb{E}\left[\frac{\widetilde{K}_{T_N}(\widetilde{S}'\gamma-\widetilde{V})\cos(tY)}{\max\left(f_{\widetilde{S},\widetilde{V}}\left(\widetilde{S},\widetilde{V}\right),m_N\right)}\right]$$

are Lipschitz with constant $m_N^{-1}T_N^{L+2}$. The same is true if we replace sin by cos. We can thus take the same t, ω and L and thus v_N as in the proof of Proposition 4.2. However due to the different covering number the constant C_1 changes which yields a different constant $M(\alpha)$.
Q.E.D.

7.0.8. *Proof of Proposition 4.4.* First note that

$$\begin{aligned}\left(\widehat{f_\Delta}-f_\Delta\right)(\delta)=&-\int_{\mathbb{R}^L}f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N^c\}d\gamma\\&+\int_{\mathbb{R}^L}\left(\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)-f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)\right)f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N\}d\gamma\\&+\int_{\mathbb{R}^L}\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)\left(\widehat{f_{\widetilde{\Gamma}}}(\gamma)-f_{\widetilde{\Gamma}}(\gamma)\right)\mathbf{1}\{\gamma\in\mathcal{B}_N\}d\gamma\end{aligned}$$

which yields

$$\begin{aligned}\left|\left(\widehat{f_\Delta}-f_\Delta\right)(\delta)\right|\le&\int_{\mathbb{R}^L}f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N^c\}d\gamma\\&+\int_{\mathbb{R}^L}\left|\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)-f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)\right|f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N\}d\gamma\\&+\int_{\mathbb{R}^L}\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)\left|\widehat{f_{\widetilde{\Gamma}}}(\gamma)-f_{\widetilde{\Gamma}}(\gamma)\right|\mathbf{1}\{\gamma\in\mathcal{B}_N\}d\gamma\\\le&\int_{\mathbb{R}^L}f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N^c\}d\gamma\end{aligned}$$

[31] $B_d(0,1)$ is a Euclidian ball centered at 0 or radius 1.

$$
\begin{aligned}
&+\int_{\mathbb{R}^L}\left|\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)-f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)\right| f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N\}d\gamma\\
&+\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty\int_{\mathbb{R}^L}\left|\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)\right| f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N\}d\gamma\\
&\le\int_{\mathbb{R}^L} f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N^c\}d\gamma\\
&+\left(1+\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty\right)\int_{\mathbb{R}^L}\left|\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)-f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)\right| f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in B_N\}d\gamma\\
&+\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty\int_{\mathbb{R}^L} f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N\}d\gamma\\
&\le\int_{\mathbb{R}^L} f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N^c\}d\gamma\\
&+\left(1+\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty\right)\int_{\mathbb{R}^L}\left|\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)-f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)\right| f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in B_N\}d\gamma\\
&+M_\Delta\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty
\end{aligned}
$$

thus, using the Cauchy-Schwartz inequality

$$
\begin{aligned}
\left(\widehat{f_\Delta}-f_\Delta\right)^2(\delta)\le\;&3\int_{\mathbb{R}^L} f^2_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N^c\}d\gamma\\
&+3\left(1+\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty\right)^2\int_{\mathbb{R}^L}\left(\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)-f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)\right)^2 f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N\}d\gamma\\
&+3M_\Delta^2\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty^2\\
\le\;&3M_\Delta\int_{\mathbb{R}^L} f_{\Delta,\widetilde{\Gamma}}(\delta,\gamma)\mathbf{1}\{\gamma\in\mathcal{B}_N^c\}d\gamma\\
&+3\|f_{\widetilde{\Gamma}}\|_\infty\left(1+\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty\right)^2\int_{\mathbb{R}^L}\left(\widehat{f_{\Delta|\widetilde{\Gamma}}}(\delta|\gamma)-f_{\Delta|\widetilde{\Gamma}}(\delta|\gamma)\right)^2\mathbf{1}\{\gamma\in B_N\}d\gamma\\
&+3M_\Delta^2\left\|\frac{\widehat{f_{\widetilde{\Gamma}}}-f_{\widetilde{\Gamma}}}{f_{\widetilde{\Gamma}}}\mathbf{1}\{\mathcal{B}_N\}\right\|_\infty^2.
\end{aligned}
$$

The inequality is now obtained by integration over δ.

Q.E.D.

7.0.9. *Proof of Proposition 4.5.* We introduce the notations

$$\overline{f}_{\Delta|\widetilde{\Gamma}}(\delta|\gamma) := \frac{1}{2\pi}\int_{-\infty}^{\infty} K(t\ h_{N,\gamma})e^{-i\delta t}\mathcal{F}_1\left[f_{\Delta|\widetilde{\Gamma}}\right](t|\gamma)dt,$$

$$R(t,\gamma) := \frac{\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| > r_{Y_0,N}\right\}}{\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)} - \frac{1}{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)}.$$

The following decomposition holds at a fixed γ by means of the Plancherel identity

$$\begin{aligned}\left\|\left(\widehat{f_{\Delta|\widetilde{\Gamma}}} - f_{\Delta|\widetilde{\Gamma}}\right)(\cdot|\gamma)\right\|_2^2 \le 4\left\|\left(\overline{f}_{\Delta|\widetilde{\Gamma}} - f_{\Delta|\widetilde{\Gamma}}\right)(\cdot|\gamma)\right\|_2^2 \\ + \frac{2}{\pi}\int_{-\infty}^{\infty}\frac{K(t\ h_{N,\gamma})^2}{\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^2}\left|\left(\widehat{\mathcal{F}_1\left[f_{Y_0+\Delta,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{Y_0+\Delta,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right|^2 dt \\ + \frac{2}{\pi}\int_{-\infty}^{\infty} K(t\ h_{N,\gamma})^2|R(t,\gamma)|^2\left|\left(\widehat{\mathcal{F}_1\left[f_{Y_0+\Delta,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{Y_0+\Delta,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right|^2 dt \\ + \frac{2}{\pi}\int_{-\infty}^{\infty} K(t\ h_{N,\gamma})^2\left|\mathcal{F}_1\left[f_{Y_0+\Delta,\widetilde{\Gamma}}\right](t,\gamma)\right|^2|R(t,\gamma)|^2 dt.\end{aligned}$$

We conclude using Lemma 7.1 below and the fact that by conditional independence

$$\frac{\left|\mathcal{F}_1\left[f_{Y_0+\Delta,\widetilde{\Gamma}}\right](t,\gamma)\right|^2}{\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^4} = \frac{\left|\mathcal{F}_1\left[f_{\Delta,\widetilde{\Gamma}}\right](t,\gamma)\right|^2}{\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^2}.$$

Q.E.D.

Lemma 7.1 below is an adaptation of the lemma of Neumann (1997). Denote by

$$\psi(t,\gamma) := \frac{1}{\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|}\min\left(1, \frac{r_{Y_0,N}}{\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|}\right).$$

Lemma 7.1.

$$\sup_{t\in[-R_N^{\max},R_N^{\max}],\ \gamma\in\mathcal{B}_N}\left\{\psi(t,\gamma)^{-1}|R(t,\gamma)|\right\} = O_p(1).$$

7.0.10. *Proof of Lemma 7.1.* We distinguish between two cases.

Case 1: Let t and γ be such that $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right| < 2r_{Y_0,N}$. Then, $\psi(t,\gamma)^{-1} \le 2\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|$ and it suffices to upper bound in probability $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right||R(t,\gamma)|$. By definition of $R(t,\gamma)$, $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right||R(t,\gamma)| \le 1$ on the event $\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| \le r_{Y_0,N}\right\}$, while $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right||R(t,\gamma)| \le (r_{Y_0,N})^{-1}\left|\left(\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right|$ on the complementary event

$\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| > r_{Y_0,N}\right\}$. This yields

$$\sup_{(t,\gamma):\ \left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|<2r_{Y_0,N}} \left\{\psi(t,\gamma)^{-1}\left|R(t,\gamma)\right|\right\} = O_p(1).$$

Case 2: Let now t and γ be such that $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right| \geq 2r_{Y_0,N}$. Then, $\psi(t,\gamma)^{-1} \leq 2\left(r_{Y_0,N}\right)^{-1}\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^2$ and it suffices to upper bound in probability $\left(r_{Y_0,N}\right)^{-1}\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^2\left|R(t,\gamma)\right|$.

By definition of $R(t,\gamma)$,

$$\begin{aligned}
&\left(r_{Y_0,N}\right)^{-1}\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^2\left|R(t,\gamma)\right| \\
&\leq \left(r_{Y_0,N}\right)^{-1}\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|\left(\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| \leq r_{Y_0,N}\right\}\right. \\
&\qquad\left.+\frac{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma) - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|}{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right|}\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| > r_{Y_0,N}\right\}\right)
\end{aligned}$$

Using

$$\frac{1}{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right|} \leq \frac{1}{\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|} + \frac{\left|\left(\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right|}{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right|\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|},$$

we obtain

$$\begin{aligned}
&\left(r_{Y_0,N}\right)^{-1}\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^2\left|R(t,\gamma)\right| \\
&\leq \left(r_{Y_0,N}\right)^{-1}\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| \leq r_{Y_0,N}\right\} \\
&+\left(r_{Y_0,N}\right)^{-1}\left(\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma) - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right| + \frac{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma) - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^2}{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right|}\right)\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| > r_{Y_0,N}\right\} \\
&\leq \left(r_{Y_0,N}\right)^{-1}\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| \leq r_{Y_0,N}\right\} \\
&+\left(\left(r_{Y_0,N}\right)^{-1}\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma) - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|\right. \\
&\quad\left.+\left(r_{Y_0,N}\right)^{-2}\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma) - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|^2\right)\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| > r_{Y_0,N}\right\}.
\end{aligned}$$

From the definition of the upper bound on the rate $r_{Y_0,N}$, the last term in the sum is, uniformly in t and γ such that $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right| \geq 2r_{Y_0,N}$, bounded in probability.

Moreover, because $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right| \geq 2r_{Y_0,N}$,

$$\begin{aligned}
\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| \leq r_{Y_0,N}\right\} &\leq \mathbf{1}\left\{\left|\left(\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right| \geq \left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right| - r_{Y_0,N}\right\} \\
&\leq \mathbf{1}\left\{\left|\left(\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right| \geq \left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|/2\right\} \\
&\leq 2\frac{\left|\left(\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right|}{\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|}
\end{aligned}$$

which yields

$$(r_{Y_0,N})^{-1}\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right|\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]}(t,\gamma)\right| \leq r_{Y_0,N}\right\} \leq (r_{Y_0,N})^{-1}\left|\left(\widehat{\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right|,$$

thus the first term is also, uniformly in t and γ such that $\left|\mathcal{F}_1\left[f_{Y_0,\widetilde{\Gamma}}\right](t,\gamma)\right| \geq 2r_{Y_0,N}$, bounded in probability.

Q.E.D.

7.0.11. *Proof of Proposition 4.6.* The proposition follows from adapting the upper bounds in Comte and Lacour (2011), (4.17) and the assumptions made.

Q.E.D.

References

[1] Aakvik, A., J. J. Heckman, and E. J. Vytlacil (2005): "Estimating Treatment Effects for Discrete Outcomes when Responses to Treatment Vary: an Application to Norwegian Vocational Rehabilitation Programs". *Journal of Econometrics*, **125**, 15–51.

[2] Abbring, J. H., and J. J. Heckman (2007): "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation". *Handbook of Econometrics*, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 72.

[3] Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings". *Econometrica*, **70**, 91–117.

[4] Angrist, J., K. Grady and G. Imbens G. (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish". *Review of Economic Studies*, **67**, 499–527.

[5] Beran, R., and P. Hall (1992): "Estimating Coefficient Distributions in Random Coefficient Regressions". *Annals of Statistics*, **20**, 1970–1984.

[6] Beran, R., A. Feuerverger, and P. Hall (1996): "On Nonparametric Estimation of Intercept and Slope in Random Coefficients Regression". *Annals of Statistics*, **24**, 2569–2592.

[7] Björklund, A., and R. Moffitt (1987): "The Estimation of Wage and Welfare Gains in Self-Selection Models", *Review of Economics and Statistics*, **69**, 42–49.

[8] Butucea, C. (2004): "Deconvolution of Supersmooth Densities with Smooth Noise". *Canadian Journal of Statistics*, **32**, 181–192.

[9] Butucea, C., and A. B. Tsybakov (2007): "Sharp Optimality in Density Deconvolution with Dominating Bias. I". *Rossiiskaya Akademiya Nauk. Teoriya Veroyatnostei i ee Primeneniya*, **52**, 111–128.

[10] Carneiro, P., K. T. Hansen, and J. Heckman (2003): "Estimating Distributions of Treatment Effects With an Application to the Return to Schooling and Measurement of the Effect of Uncertainty on College Choice". *International Economic Review*, **44**, 361–422.

[11] Carrasco, M., J. P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization". *Handbook of Econometrics* , J. J. Heckman and E. E. Leamer (eds.), vol. 6B, North Holland, chapter 77, 5633–5751.

[12] Carrasco, M., and J. P. Florens (2011): "A Spectral Method for Deconvolving a Density". *Econometric Theory*, **27**, 546—581.

[13] Cavalier, L. (2000): "Efficient Estimation of a Density in a Problem of Tomography". *Annals of Statistics*, **28** , 630–647.

[14] Cavalier, L. (2001): "On the Problem of Local Adaptive Estimation in Tomography". *Bernoulli*, **7**, 63–78.

[15] Chernozhukov, V., and C. Hansen. (2005): "An IV Model of Quantile Treatment Effects". *Econometrica*, **73**, 245–261.

[16] Comte, F., and C. Lacour (2011): "Data-driven Density Estimation in the Presence of Additive Noise with Unknown Distribution". *Journal of the Royal Statistical Society. Series B*, **73**, 601–627.

[17] Devroye, L. (1989): "Consistent Deconvolution in Density Estimation". *Canadian Journal of Statistics*, **17**, 235—239.

[18] Elbers, C., and G. Ridder (1982): "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Models". *Review of Economics Studies*, **49**, 403–410.

[19] Evans, L. C. (1998): *Partial Differential Equations*. Graduate Studies in Mathematics, American Mathematical Society.

[20] Evdokimov, K. (2010): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity" . Working paper.

[21] Evdokimov, K., and H. White (2011): "An Extension of a Lemma of Kotlarski". Working paper.

[22] Fan, Y., and S. S. Park (2010): "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference". *Econometric Theory*, **26**, 931–951.

[23] Fan, Y., and D. Zhu (2009): "Partial Identification and Confidence Sets for Functional of the Joint Distribution of the Potential Outcomes". Working paper.

[24] Firpo, S., and G. Ridder (2008): "Bounds on Functionals of the Distribution of Treatment Effects". Working paper.

[25] Fréchet, M. (1951): "Sur Les Tableaux de Corrélation Dont les Marges Sont Données", *Annales de l'Université de Lyon, Série 3*, **14**, 53–77.

[26] Fox, J., and A. Gandhi (2009): "Identifying Heterogenity in Economic Choice and Selection Models Using Mixtures". Working Paper.

[27] Gaïffas, S. (2005): "Convergence Rates for Pointwise Curve Estimation with a Degenerate Design". *Mathematical Methods of Statistics*, 14, 1-27.

[28] Gaïffas, S. (2009): "Uniform Estimation of a Signal Based on Inhomogeneous Data". *Statistica Sinica*, 19, 427-447.

[29] Gautier, E., and Y. Kitamura (2009): "Nonparametric Estimation in Random Coefficients Binary Choice Models". Preprint arXiv:0907.2451, forthcoming in *Econometrica*.

[30] Gautier, E., and E. Le Pennec (2011): "Adaptive Estimation in Random Coefficients Binary Choice Models Using Needlet Thresholding". Preprint arXiv:1106.3503.

[31] Guerre, E. (1999): "Efficient Random Rates for Nonparametric Regression Under Arbitrary Designs". Working Paper.

[32] Hall, P., Marron, J. S., Neumann, M. H., and Tetterington, D. M. (1997): "Curve Estimation When the Design Density is Low". *Annals of Statististics*, 25, 756770.

[33] Heckman, J. J., and B. Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data". *Econometrica*, **52**, 271–320.

[34] Heckman, J. J., and J. Smith (1998): "Evaluating The Welfare State". In: Strom, S. (Ed.), Econometrics and Economic Theory of the Twentieth Century: The Ragnar Frisch Centennial Symposium. Cambridge University Press, New York, pp. 241–318.

[35] Heckman, J. J., J. Smith, and N. Clements (1997): "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts". *Review of Economic Studies*, **64**, 487–635.

[36] Heckman, J. J., and E. Vytlacil (1999) "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects". Proceedings of the National Academy of Science, USA, **96**, 4730–4734.

[37] Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation". *Econometrica*, **73**, 669–738.

[38] Heckman, J. J., and E. Vytlacil (2007a) "Econometric Evaluation of Social Programs, Part I: Causal Models, Strucural Models and Econometric Evaluation of Public Policies". *Handbook of Econometrics*, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 70.

[39] Heckman, J. J., and E. Vytlacil (2007b) "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments". *Handbook of Econometrics*, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 71.

[40] Helgason, S. (1999): *The Radon Transform*. 2nd edition. Birkhauser Boston.

[41] Hoderlein, S., J. Klemelä, and E. Mammen (2010): "Analyzing the Random Coefficient Model Nonparametrically". *Econometric Theory*, **26**, 804–837.

[42] Hoeffding, W. (1940): "Masstabinvariate Korrelationstheorie". *Schriften des Mathematischen Instituts und Institutes Für Angewandte Mathematik der Universität Berlin*, **5**, 179–233.

[43] Ichimura, H., and T. S. Thompson (1998): "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution". *Journal of Econometrics*, **86**, 269–295.

[44] Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects". *Econometrica*, **62**, 467–475.

[45] Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity Corresponding". *Econometrica*, **77**, 1481–1512.

[46] Johannes, J. (2009): "Deconvolution with Unknown Error Distribution". *Annals of Statistics*, **37**, 2301–2323.

[47] Klein, T. (2010): "Heterogeneous Treatment Effects: Instrumental Variables Without Monotonicity?". *Journal of Econometrics*, **155**, 99–116.

[48] Korostelev, A. P., and A. B. Tsybakov (1993): *Minimax Theory of Image Reconstruction*. Springer, New-York, *Lecture Notes in Statistics* **82**.

[49] Lewbel, A. (2000): "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables". *Journal of Econometrics*, **97**, 145–177.

[50] Lewbel, A. (2007): "Endogenous Selection or Treatment Model Estimation". *Journal of Econometrics*, **141**, 777–806.

[51] Lounici, K., and R. Nickl (2011): "Global Uniform Risk Bounds for Wavelet Deconvolution Estimators". *Annals of Statistics*, **39**, 201–231.

[52] Makarov, G. D. (1981): "Estimates of the Distribution Function of a Sum of Two Random Variables when the Marginal Distributions are Fixed". *Theory of Probability and its Applicatons*, **26**, 803–806.

[53] Manski, C. F. (1997): "Monotone Treatment Effect". *Econometrica*, **65**, 1311–1334.

[54] Moffitt, R. (2008): "Estimating Marginal Treatment Effects in Heterogeneous Populations". *Annals of Economics and Statistics*, **91/92**, 239–261.

[55] Natterer, F. (1986): *The Mathematics of Computerized Tomography*. Wiley, Chichester.

[56] Neumann, M. H. (1997): "On the Effect of Estimating the Error Density in Nonparametric Deconvolution". *Journal of Nonparametric Statistics*, **7**, 307–330.

[57] van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes*. New York: Springer.

CREST (ENSAE), 3 avenue Pierre Larousse, 92 245 Malakoff Cedex, France.

E-mail address: Eric.Gautier@ensae-paristech.fr

Boston College, Chestnut Hill, MA 02467, USA.

E-mail address: Stefan.Hoderlein@bc.edu