ESTIMATING THE DISTRIBUTION OF TREATMENT EFFECTS

ERIC GAUTIER AND STEFAN HODERLEIN

ABSTRACT. In this paper we show that the distribution of treatment effects is point identified in a model where the outcome equation is of unrestricted form and the selection equation contains more than one unobservable. This is different and economically better motivated than the treatment effect on the distribution, usually the quantiles, which is commonly analyzed in the literature. Our key identifying assumption in the selection equation is a linear random coefficients structure and the assumption that the instruments are continuously distributed. This allows point identification of the entire distribution of treatment effects under conditions on unobserved heterogeneity that unlike the case of additively separable/monotonic scalar unobservables have a clear economic interpretation in terms of unobserved heterogeneity. Also, we obtain results on the distribution of treatment effects without invoking any scalar monotonicity assumption in the outcome equation. Moreover, the identification is constructive and suggests estimators of various quantities of interest by sample counterparts.

1. INTRODUCTION

Motivation In this paper we consider estimating the distribution of treatment effects in the following structural Roy type treatment effects model

$$(1.1) \qquad Y \;=\; A + BX$$

$$(1.2) \qquad X \;=\; \mathbf{1}\left\{Z^T\Gamma > 0\right\}$$

$$(1.3) \qquad Z \quad \text{is independent of } (A, B, \Gamma^T).$$

In the outcome equation (1.1), A and B are scalar random coefficients and $\mathbf{1}$ denotes the indicator function. In the treatment effect literature A is usually denoted as Y_0 and $B = Y_1 - Y_0$. i.e., A denotes the distribution of outcomes in the control group, and B the treatment effects. In the binary treatment case, the linearity in the outcome equation is unrestrictive.

The object of interest in this paper is to obtain the distribution of treatment effects f_B in the case of endogenous selection into treatment, and moments of the distribution like the mean (average)

and the variance of treatment effects. We model the binary choice to participate in the treatment in equation (1.2). The participation decision involves Z, a L-vector of instruments, and Γ, a random vector that accounts for heterogeneous preferences and information which in turn governs the selection into treatment. We allow for the first component of Z to be unity and for the first component of Γ to absorb the usual stochastic shock term.

As is evident from the equations (1.1)-(1.3), we do place some structure on the endogenous selection into treatment. First, we assume that we have instruments that are fully independent from the unobservables in the system, and, as we shall see below, also continuously distributed. The key identifying restriction is the linear random coefficients structure in the selection equation. While the linearity is clearly restrictive, it allows us to model several unobservables in a way that is more reminiscent of structural economics. Note that even if interest centers only on the average treatment effects, introducing some type of structure is necessary as there is not even point identification of the average of the distribution of treatment effect (ATE) in general[1].

Since the aim of this paper is to propose a minimal structure that point identifies the mean, the variance, as well as the entire distribution of treatment effects, we thus have to restrict the structure of the model at some point. We feel that our linear random coefficients structure is well motivated by heterogeneity in a population of economic agents: it corresponds to the notion that on individual level linearity of the selection equation is (at least approximately) a valid description of behavior. Depending on the feature of the distribution, we will require further assumptions. If the focus is on the mean of the distribution, we do not require any additional assumptions (apart from regularity conditions). For the variance of treatment effects, we provide sharp bounds, and show point identification under a covariance restriction. Finally, we establish identification of the distribution of treatment effects under the assumption that $A \perp B | \Gamma$. We will discuss this assumption in detail below. At this point we would like to point out that this assumption is satisfied, if there exist otherwise unrestricted mappings ψ_A, ψ_B such that $A = \psi_A(\Gamma, U_A)$, and $B = \psi_B(\Gamma, U_B)$, with U_A, U_B possibly infinite dimensional, such that $U_A \perp U_B \perp \Gamma$. This corresponds to a notion that the selection equation reveals information about the common endogenous factors; there is (potentially complicated) remaining heterogeneity in A and B, but it is independent of everything else.

---

[1] Indeed, Shaikh and Vytlacil (2009) provide sharp bounds for this effect which characterize the identified set, and Imbens and Angrist (1994) show that even under a monotonicity assumption on instruments, only the average effect for "compliers" (local average treatment effect, LATE), a specific subpopulation defined by instruments is point identified.

The main results in this paper establish that under the respective conditions, for $B = Y_1 - Y_0$,

$$\mathbb{E}[B \mid \Gamma = \gamma], \ Var[Y_1 - Y_0 \mid \Gamma = \gamma], \ \text{ and } \ f_{B|\Gamma}(b; \gamma)$$

are (point) identified. These objects correspond to generalizations of the Heckman and Vytlacil (2005) MTE to several sources of heterogeneity. From these results, we show that the unconditional average, variance and distribution treatment effect may be identified, but it requires an identification at infinity type of assumption. However, even in the absence of identification at infinity, these results can also be employed as building blocks to obtain policy relevant treatment effects, as in Carneiro, Heckman and Vytlacil (2010). Finally, and equally importantly, we establish that in the general case with limited support for our instruments, we can only identify the average, variance or distribution of effects for a subpopulation defined by the range of the instrument, a population that is related to the one considered in Angrist, Graddy and Imbens (2000).

Based on these identification results, we provide sample counterparts estimators. It is known from Hoderlein, Klemelä and Mammen (2010) and Gautier and Kitamura (2009) that the estimation of the distribution of random coefficients in the exogenous single equation cases are ill-posed inverse problems. It is clear from our identification argument that the estimation of the distribution of the treatment effect B is a different ill-posed inverse problem, akin to conditional deconvolution with noise and mixing distributions that are unknown but estimable by solving themselves inverse problems. We provide such estimators, and analyze their large sample behavior.

Literature Naturally, 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. Key references for the former are the quantile treatment effect of Abadie, Angrist and Imbens (2002), Chernozhukov and Hansen (2005) and Heckman, Smith and Clements (1996). Note that the first two results 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 slightly weakened. This assumption is restrictive, and has rightfully been criticized, see Heckman, Smith and Clements (1996), who point out that what is of interest is the distribution of treatment effects, and not the treatment effect on the distribution, and who provide (Frechet-) bounds for this quantity. In contrast, we provide point identification of the distribution of effects under different assumptions. As an implication, we can also obtain results for the average treatment effect, which is the expected value of the distribution of B. This is related to the seminal contribution

of LATE (Imbens and Angrist(1994)) and MTE (Heckman and Vytlacil (1999, 2005, 2007), henceforth HV).

The second line of work which is popular to model unobserved heterogeneity and related is, as mentioned, random coefficient models. Random coefficient models allow the preference or production parameters to vary across the population. In this paper, we allow for different individuals to have different preferences for treatment and that effect of treatment might differ for each individual. 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), Ichimura and Thompson (1998), Fox and Gandhi (2009), Hoderlein, Klemelä and Mammen (2010), Gautier and Kitamura (2009) and Gautier and Le Pennec (2011). The last two references recognize that the estimation of the density of the latent random coefficients vector is a statistical inverse problem. The literature on the treatment of these problems is extensive in statistics and econometrics (see, *e.g.*, Carrasco, Florens and Renault (2007) for a survey of applications in economics). Fox and Gandhi (2009) are the first to study the identification of the distribution of unobserved heterogeneity in treatment effects models, however they do not allow for an intercept in the binary choice model.

The two equations model in this paper combines the random coefficients linear model studied by Beran and Hall (1992), Hoderlein, Klemelä and Mammen (2010) and the random coefficients binary choice model studied by Ichimura and Thompson (1998), Gautier and Kitamura (2009), and Gautier and Le Pennec (2011). The first equation though has a regressor which is a dummy variable whose effect varies by a random coefficient. This is not handled in Beran and Hall (1992) or Hoderlein, Klemelä and Mammen (2010).

The problem is doubly related to inverse problems: the conditional densities of $A + B$ and A given Γ as well as that of Γ are obtained solving inverse problems. We present two approaches when $L \geq 3$. The one presented in Section 3 involves the inversion of the Radon transform (see, *e.g.*, Helgason (1999) and Korostelev and Tsybakov (1993), Cavalier (2000) and Hoderlein, Klemelä and Mammen (2010) for statistical inverse problems involving the Radon transform on $\mathbb{R}^d$) applied to the derivative of a regression function. The one of Section 5.1 relies on the inversion of the Hemispherical transform (see, *e.g.*, Funk (1916), Groemer (1996), Rubin (1999) and Gautier and Kitamura (2009) and Gautier and Le Pennec (2011) for a statistical inverse problem involving the Hemispherical transform). A second inverse problem appears when we obtain the conditional density of B given Γ assuming that $A \perp B | \Gamma$. This corresponds to conditional deconvolution. Evdokimov (2010) considers conditional

deconvolution in a different problem. In the classical deconvolution problem, the density of A is known and the characteristic function of $A + B$ is estimated via the empirical characteristic function which estimates the true characteristic function at rate $1/\sqrt{N}$. An extension studied in Diggle and Hall (1993), Neumann (1997), Comte and Lacour (2009) and Johannes (2009) considers the case where the density of A is estimable at rate $1/\sqrt{N}$ using a preliminary sample. In this paper $A + B$ and B are unobserved and both conditional characteristic functions given Γ are estimated solving inverse problems using the same sample.

Throughout this paper, we will analyze the distribution of random coefficients under the assumption that we observe an independent and identically distributed sample $(y_i, x_i, z_i^T)_{i=1,\dots,N}$ where N is the sample size, and assume that the independent and identically distributed realizations $(a_i, b_i, \gamma_i^T)_{i=1,\dots,N}$ are unobserved. Because it allows to relate our results closely to the literature, in Section 2 we consider in detail the important case of a single continuous instrument. In section 3 we present the more general case that there is more than one continuous instrument and we keep the index structure of model defined by equations (1.1)-(1.3). In the fourth section, we analyze the large sample behavior of our estimator.

2. The Single Unobservable Case

In this section, we start out with the standard specification for the selection equation in the treatment effects literature. We review the main result of Heckman and Vytlacil (2005, henceforth HV) in this setup, and show how to obtain the variance or the distribution of the treatment effect including a discussion of the respective assumptions.

2.1. **MTE and ATE with a Single Unobservable in the Selection Equation.**

2.1.1. *Model and Assumptions.* The setup we shall consider can be formalized as follows in a random coefficients setup:

$$(2.1) \qquad \begin{cases} Y = A + BX \\ X = \mathbf{1}\{P > V\} \quad V|Z \sim \mathcal{U}(0,1) \end{cases}$$

where X is a binary treatment indicator, $P = \pi(Z) = Pr(X = 1 \mid Z)$ is the selection probability (sometimes also referred to as "propensity score"). This model could arise in several ways, and it is neither more or less general than the pure random coefficients model defined by equations (1.1)-(1.3). However, when we have one instrument the latter is nested in the former. To see this, recall that the

selection equation (1.2) is then defined as

$$(2.2) \qquad\qquad X = \mathbf{1}\{\Gamma_1 + \Gamma_2 Z_2 > 0\}.$$

For identification of the density of $\Gamma/\|\Gamma\|$, we know from Gautier and Kitamura (2009) that it is sufficient that the support be included in some half sphere. This occurs if the coefficient of Z_2 has a fixed sign. Changing Z_2 in $-Z_2$ we can assume that the coefficient is nonnegative. Assuming it is as well nonzero we can write the event inside the indicator of equation (2.2) as $Z_2 \geq -\Gamma_1/\Gamma_2$. Moreover, this is observationally equivalent to

$$(2.3) \qquad\qquad X = \mathbf{1}\{F_{-\Gamma_1/\Gamma_2}(Z_2) > V\},$$

where $F_{-\Gamma_1/\Gamma_2}$ is the cdf of $-\Gamma_1/\Gamma_2$, with $V \backsim \mathcal{U}(0,1)$. This obviously leads to model (2.1) with $\pi(Z_2) = F_{-\Gamma_1/\Gamma_2}(Z_2)$. However, in the case of d instruments, the two models are generally different, hence we treat the two approaches separately. Observe that at this stage, we impose additive separability inside the indicator of the first stage (i.e., the selection) equation, which is obviously restrictive in terms of the unobservables that are allowed for, but at least allows for an unrestricted function π of the instruments. We start out by treating this model for two reasons. First, we want to provide a useful addition to the standard HV type of framework. Secondly, we want to understand similarities and differences with the case when we have more than one unobservables[2] in the selection equation.

The formulation is somewhat different from the standard treatment effects formulation. To see the parallels, $A = Y_0$, and $B = Y_1 - Y_0$. With this notation, $Y = A + BX = Y_0 + (Y_1 - Y_0)X$ which is more standard. It is also useful to think of Y as being generated by a nonseparable model; in this case $Y = \psi(X, U)$, and $Y_0 = \psi(0, U) = A$, as well as $Y_1 = \psi(1, U) = A + B$. As is obvious from these equalities, in the binary case the linearity in X in the outcome equation is without loss of generality. However, if we identify U with a high dimensional unobservable, e.g., preferences, it is interesting to note that we are thinking of our random coefficients A and B as two different functions of the unobservables, i.e., $A = a(U) = \psi(0, U)$ and $B = b(U) = \psi(1, U) - \psi(0, U)$. Without loss of generality, one could further partition the set of unobservable in vectors U_0, U_1, and U_2. Here, U_2 are unobservables that are common to both the treatment and the control group [3]

---

[2] With the scaling of (2.3), there is only one unobservable which is V, though the size L parameter introduced later which corresponds to vectors in the coordinate system before rescaling is 2.

[3] We prefer the term subpopulation, but use this common name in the literature. treatment, U_0 are specific to the control group but these unobservables do not affect the effect of treatment, and U_1 impacts only the effect of treatment. This means that $A = a(U_0, U_2)$ and $B = b(U_1, U_2)$.

We will analyze the model under materially the same assumptions as HV, but add one crucial condition to deal with distributions. The effect we analyze is the marginal treatment effect (MTE), which is, for $p \in (0,1)$:

$$\mathbb{E}[B \mid V = p] = \mathbb{E}[Y_1 - Y_0 \mid V = p],$$

which is the average effect of treatment for the subpopulation, for which V , the first stage preference parameter, takes on the value p. For p, we can think of this population as being indifferent between participation and non-participation in the treatment, see HV (2005, 2007) for a more extensive discussion of this parameter.

The assumptions we employ to first state the HV result, and then proceed to the distribution of effects, are as follows:

Assumption 2.1. *Let* (Ω, F, P) *be a complete probability space on which are defined the random vectors* $(Y, X, Z, A, B, V) : \Omega \to \mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \times \mathcal{A} \times \mathcal{B} \times \mathcal{V}$, $\mathcal{Y} \subseteq \mathbb{R}, \mathcal{X} = \{0,1\}, \mathcal{Z} \subseteq \mathbb{R}^2, \mathcal{A} \subseteq \mathbb{R}, \mathcal{B} \subseteq \mathbb{R}$ *and* $\mathcal{V} \subseteq \mathbb{R}$. *The causal model is defined by equation* (2.1) *where* $\pi : \mathcal{Z} \to [0,1]$ *is Borel measurable function, and realizations of* (Y, X, Z) *observable whereas those of* (A, B, V) *are not.*

Assumption 2.2. *All of the defined probability distributions (joint, marginal, and conditional) involving* (Y, X, Z, A, B, V) *, but* X *only appearing in the conditioning set, are absolutely continuous with respect to Lebesgue measure.*

Assumption 2.3. *(i)* (A, B, V) *are independent of* Z *. (ii) The distribution of* $A = Y_0$ *and the distribution of* $B = Y_1 - Y_0$ *given* $V = p$ *have moments of order one and* $\mathbb{E}[B|V = \cdot] \in L^1(\mathbb{R})$. *(iii) The conditional density of* $B = Y_1 - Y_0$ *given* $V = p$ *is in* $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ *and* $\mathbb{E}[e^{itA}|V = \cdot]$ *and* $\mathbb{E}[e^{it(A+B)}|V = \cdot]$, *where* $i = \sqrt{-1}$ *are in* $L^1(\mathbb{R})$. *(iv) The conditional distributions of* $A = Y_0$ *and* $B = Y_1 - Y_0$ *given* $V = p$ *have second moments and* $\mathbb{E}[A^2|V = \cdot], \mathbb{E}[B^2|V = \cdot]$ *and* $\mathbb{E}[AB|V = \cdot]$ *are in* $L^1(\mathbb{R})$.

As will be clear from the proofs, the model and the respective assumptions in Assumption 2.3 imply that the conditional expectations $\mathbb{E}[Y \mid \pi(Z) = p]$, $\mathbb{E}[Xe^{itY} \mid \pi(Z) = p]$ and $\mathbb{E}[(1 - X)e^{itY} \mid \pi(Z) = p]$, exist and are differentiable. This does not need to be assumed.

2.1.2. *Main Result.* These assumptions allow us to characterize the MTE

Theorem 2.1. Suppose that Assumptions 2.1, 2.2 (i), and 2.3 (i) and (ii) hold in the model defined through equation (2.1). Then,

$$\mathbb{E}[B \mid V = p] = \partial_p \mathbb{E}[Y \mid \pi(Z) = p]$$

holds.

Proof. See Heckman and Vytlacil (2005).

For an in depth discussion, see HV (2005, 2007). For our purpose, note that the MTE on the left-hand side is identified by the local instrumental variable (LIV) on the right-hand side. The only thing we want to point out is that, in order to obtain the average treatment effect (ATE), we would need to integrate over p from 0 to 1, or,

$$ATE = \mathbb{E}[B] = \int_0^1 \mathbb{E}[Y_1 - Y_0 \mid V = p]dp,$$

which requires "identification at infinity" in the sense that the instrument has to be informative enough to shift $\pi(z)$ to both zero and unity.

2.2. The Distribution of Treatment Effects with a Single Unobservable Case in the Selection Equation.

2.2.1. *Parameter of Interest.* Having laid out the setup and related it to previous work, we introduce now the object of interest. In exactly the same setup as defined above in equation (2.1), we will be concerned with the distribution of treatment effects for the same subpopulation as the one considered by HV. More formally, we are interested in recovering

$$f_{B|V}(b; p) = f_{Y_1 - Y_0|V}(b; p),$$

for any $b \in supp(Y_1 - Y_0)$, where $supp(Q)$ denotes the support of a random variable Q. In analogy to HV, we call this the "Distribution of Treatment Effects at the Margin", and abbreviate it DITEM. The interpretation is also quite similar to HV: For the subpopulation who is indifferent between participation and non participation, it provides us with a measure for the effect of treatment. However, this measure is now the distribution of effects. As mentioned in the introduction, the distribution of effects is different from the effect of treatment on the distribution.

We will study this object under an additional identifying assumption

Assumption 2.4. $A \perp B|V$

Under our maintained set of assumptions, this assumption is sufficient for point identification of $f_{B|V}(b; p)$. This assumption restricts heterogeneity appearing in this model. It is best understood in terms of the reformulation introduced above, namely $A = a(U_0, U_2)$ and $B = b(U_1, U_2)$. A sufficient condition for assumption (2.4) is that $V = U_2$, and $U_0 \perp U_1|V$ (for the latter it would in turn be

sufficient that $U_0 \perp U_1 \perp V$). In words, there is a common driving factor that causes endogeneity in the selection model, and it is given by V, which, even though it is not recovered, may serve as a control function. There is remaining randomness in A and B, however, once the driving factor for endogeneity in this system, i.e., V, is accounted for, there is no leftover endogeneity.

Note that it does **not** mean that $A \perp B$. In fact, unless there is no endogenous selection there will generally be dependence between Y_0 , and $Y_1 - Y_0$. In other words, there is endogenous selection into treatment, but as far as it is endogenous, it can be summarized by V.

This assumption will be weakened in the model with several unobservables in the selection in the sense that there is not just one factor that we can employ to control for endogeneity, but we have a full vector of such variables, meaning that if we have a richer set of unobservables $V_1, \dots, V_K$, a condition like $A \perp B | V_1, \dots, V_K$ is more likely to hold.

The proof of this results relies crucially on conditional characteristic functions (ccf's) of the scalar random unobservables H (A, $A + B$ and B), conditional on $V = p$, defined as

$$\mathbb{E}\left[e^{itH} | V = p\right].$$

To recover the ccf of B at $V = p$, we require the following condition:

Assumption 2.5.

$$\forall t \in \mathbb{R}, \forall p \in (0,1) : \ \partial_p \mathbb{E}\left[(1 - X)e^{itY} | P = p\right] \neq 0.$$

As we shall see in the proof of Theorem 2.2, the quantity in Assumption 2.5 is indeed $\mathbb{E}\left[e^{iA} | V = p\right]$: the ccf of the distribution of the untreated subpopulation conditional on the unobservables in the selection equation (the source of endogeneity). This assumption is technical and it is classical in deconvolution problems where A is the noise corrupting the signal B. Characteristic functions of most standard distributions (normal, log-normal, Cauchy, Laplace, χ^2, Student-t, etc.) do not vanish.

2.2.2. *Distribution of Treatment Effects.* These assumptions allow us to characterize the DITEM:

Theorem 2.2. Suppose that Assumptions 2.1, 2.2, 2.3 (i) and (iii)[4], 2.4 and 2.5 hold in the model defined through equation (2.1). Then,

$$\mathbb{E}\left[e^{itB} | V = p\right] = -\frac{\partial_p \mathbb{E}\left[X e^{itY} | P = p\right]}{\partial_p \mathbb{E}\left[(1 - X)e^{itY} | P = p\right]}$$

---

[4] The square integrability is indeed only useful for the analysis of the rates of convergence of the estimator.

and

$$f_{B|V}(b;p) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itb}\frac{\partial_p\mathbb{E}\left[Xe^{itY}|P=p\right]}{\partial_p\mathbb{E}\left[(1-X)e^{itY}|P=p\right]}dt$$

holds.

Remarks: Since $B = Y_1 - Y_0$, this gives the distribution of treatment effects. From there on, we can get many quantities of interest. If the score varies in the whole range $(0,1)$, which amounts to identification at infinity, we can obtain the unconditional characteristic function

$$\mathbb{E}\left[e^{itB}\right] = \int_0^1 -\frac{\partial_p\mathbb{E}\left[Xe^{itY}|P=p\right]}{\partial_p\mathbb{E}\left[(1-X)e^{itY}|P=p\right]}dp,$$

or the unconditional density of the treatment effect

$$f_B(b) = -\frac{1}{2\pi}\int_0^1\int_{-\infty}^{\infty} e^{-itb}\frac{\partial_p\mathbb{E}\left[Xe^{itY}|P=p\right]}{\partial_p\mathbb{E}\left[(1-X)e^{itY}|P=p\right]}dtdp.$$

Alternatively, akin to Carneiro, Heckman and Vytlacil (2010) we can look at how weighted averages change with θ, i.e., $\partial_\theta\chi(t,\theta)$ or $\partial_\theta\zeta(b,\theta)$ (Marginal policy relevant treatment effect) where

$$\chi(t,\theta) = \int_0^1 -\frac{\partial_p\mathbb{E}\left[Xe^{itY}|P=p\right]}{\partial_p\mathbb{E}\left[(1-X)e^{itY}|P=p\right]}\omega(p,\theta)dp$$

or

$$\zeta(b,\theta) = -\frac{1}{2\pi}\int_0^1\int_{-\infty}^{\infty} e^{-itb}\frac{\partial_p\mathbb{E}\left[Xe^{itY}|P=p\right]}{\partial_p\mathbb{E}\left[(1-X)e^{itY}|P=p\right]}dt\omega(p,\theta)dp.$$

For such quantities it is not necessary that the score varies in the whole range $(0,1)$.

2.3. **Variance of Treatment Effects.** Similar computations as those used in the proof of Theorem 2.2 allow to obtain various moments of the distribution of treatment effects under milder assumptions than in Section 2.2. For example, we state Theorem 2.3 which allows to get the variance. It is obtained by simple algebra.

Theorem 2.3. If Assumptions 2.1, 2.2, 2.3 (i) and (iv) and $\mathbb{E}[AB|V] = \mathbb{E}[A|V]\mathbb{E}[B|V]$ hold, then

$$Var(Y_1 - Y_0|V = p) = Var(B|V = p) = \gamma(p)$$

where

$$\gamma(p) = \partial_p\mathbb{E}\left[XY^2|P=p\right] - (2\partial_p\mathbb{E}\left[(1-X)Y|P=p\right] + \partial_p\mathbb{E}\left[Y|P=p\right])\partial_p\mathbb{E}\left[Y|P=p\right]$$
$$- \partial_p\mathbb{E}\left[(1-X)Y^2|P=p\right].$$

Strikingly, even if we do not assume $\mathbb{E}[AB|V] = \mathbb{E}[A|V]\mathbb{E}[B|V]$ we can get the following bound on this quantity.

Theorem 2.4. If Assumptions 2.1, 2.2, 2.3 (i) and (iv) hold, then the following sharp bounds hold

$$(Var(B|V=p) - \tilde{\gamma}(p))^2 \le 4\partial_p \mathbb{E}\left[(1-X)Y^2|P=p\right]\partial_p \mathbb{E}\left[XY^2|P=p\right]$$

where

$$\tilde{\gamma}(p) = \partial_p \mathbb{E}\left[XY^2|P=p\right] - (\partial_p \mathbb{E}\left[Y|P=p\right])^2 + \partial_p \mathbb{E}\left[(1-X)Y^2|P=p\right].$$

2.4. **Instrument with Limited Support.** In the scalar instrument case and one unobservable case, if we consider the initial model (1.1)-(1.3) and assume that the coefficient on Z_2 is positive (recall that $Z_1 = 1$), we obtain a selection equation of the form:

$$X = \mathbf{1}\left\{Z_2 > \widetilde{\Gamma}\right\}$$

for some scalar unobservable $\widetilde{\Gamma}$ that is independent of Z_2 but not necessarily uniformly distributed. Starting from here we use the notation $\widetilde{\Gamma}$ to denote a rescaled version of Γ, we denote by γ without the tildes the arguments of functions when it is not subject to confusion. Assume that $supp(Z_2) = I_Z$ is an interval and denote by $\text{int}(I_Z)$ its interior. Consider for example the MTE, because from Assumption 2.3 Z_2 and $\widetilde{\Gamma}$ are independent, setting

$$\forall \gamma \in I_Z, \ r_Y(\gamma) \triangleq \mathbb{E}[Y|Z_2 = \gamma] = \mathbb{E}[A] + \int_{supp(\widetilde{\Gamma})} \mathbb{E}[B|\widetilde{\Gamma} = \tilde{\gamma}]f_{\widetilde{\Gamma}}(\tilde{\gamma})\mathbf{1}\left\{\gamma \ge \tilde{\gamma}\right\}d\tilde{\gamma}.$$

This yields

(2.4) $$\forall \gamma \in \text{int}(I_Z), \ r'_Y(\gamma) = \mathbb{E}[B|\widetilde{\Gamma} = \gamma]f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\left\{\gamma \in supp\left(\widetilde{\Gamma}\right)\right\}.$$

Similarly, setting

$$\forall \gamma \in I_Z, \ r_X(\gamma) \triangleq \mathbb{E}[X|Z_2 = \gamma] = \int_{supp(\widetilde{\Gamma})} f_{\widetilde{\Gamma}}(\tilde{\gamma})\mathbf{1}\left\{\gamma \ge \tilde{\gamma}\right\}d\tilde{\gamma},$$

we get

(2.5) $$\forall \gamma \in \text{int}(I_Z), \ r'_X(\gamma) = f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\left\{\gamma \in supp\left(\widetilde{\Gamma}\right)\right\}.$$

We extend both derivatives r'_Y and r'_X to be 0 on the complement of $\text{int}(I_Z)$. The extension of the derivatives are the natural extensions when $supp\left(\widetilde{\Gamma}\right) \subset I_Z$ since, for example, for any z_2 in I_Z such that $z_2 \ge \gamma$ for every γ in $supp\left(\widetilde{\Gamma}\right)$, $\mathbb{E}[Y|Z_2 = z_2] = \mathbb{E}[A] + \mathbb{E}[B]$ and when $z_2 \le \gamma$ for every γ in $supp\left(\widetilde{\Gamma}\right)$, $\mathbb{E}[Y|Z_2 = z_2] = \mathbb{E}[A]$ and the derivatives are 0. Note that

$$\int_{-\infty}^{\infty} r'_X(\gamma)d\gamma = \mathbb{P}(\widetilde{\Gamma} \in I_Z).$$

This yields

$$\mathbb{E}[B|\widetilde{\Gamma} \in I_Z] = \frac{\int_{-\infty}^{\infty} r'_Y(\gamma)d\gamma}{\int_{-\infty}^{\infty} r'_X(\gamma)d\gamma}.$$

It corresponds to the average treatment effect for the sub-population whose unobservables vary in a range that can be apprehended by the variation of the instrument. Under the supplementary assumption that $supp\left(\widetilde{\Gamma}\right) \subset I_Z$ this is the MTE for the whole population. Similarly, it is easy to revisit the derivations of Section 2.2 and get for example that

$$(2.6) \qquad f_{B|\widetilde{\Gamma}\in I_Z}(b) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-itb} \frac{\mathcal{F}_1\left[f_{A+B,\widetilde{\Gamma}}\right](t,\gamma)}{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)} dt f_{\widetilde{\Gamma}|\widetilde{\Gamma}\in I_Z}(\gamma)d\gamma$$

$$f_{\widetilde{\Gamma}|\widetilde{\Gamma}\in I_Z}(\gamma) = \frac{r'_X(\gamma)}{\int_{-\infty}^{\infty} r'_X(\gamma)d\gamma}$$

where $\forall \gamma \in \text{int}(I_Z)$,

$$(2.7) \qquad \mathcal{F}_1\left[f_{A+B,\widetilde{\Gamma}}\right](t,\gamma) = \partial_\gamma \mathbb{E}[Xe^{itY}|Z_2 = \gamma]$$

$$(2.8) \qquad \mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma) = -\partial_\gamma \mathbb{E}[(1-X)e^{itY}|Z_2 = \gamma],$$

$\mathcal{F}_1$ stands for the partial Fourier transform with respect to the first argument and we use the notations $f_{A+B,\widetilde{\Gamma}}$, $f_{A,\widetilde{\Gamma}}$, and $f_{B,\widetilde{\Gamma}}$ for the joint densities of respectively $(A+B,\widetilde{\Gamma})$, $(A,\widetilde{\Gamma})$ and $(B,\widetilde{\Gamma})$. We will also use the notation $\mathcal{F}_2$ to denote the partial Fourier transform with respect to the second argument which will be γ in the sequel. Recall the definition of the partial Fourier transform:

$$\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma) = \int_{-\infty}^{\infty} e^{itx} f_{A,\widetilde{\Gamma}}(x,\gamma)dx,$$

$$\mathcal{F}_2\left[f_{A,\widetilde{\Gamma}}\right](x,\omega) = \int_{\mathbb{R}^{L-1}} e^{i\omega^T\gamma} f_{A,\widetilde{\Gamma}}(x,\gamma)d\gamma.$$

We denote by $\mathcal{F}$ the usual Fourier transform, for example,

$$\mathcal{F}\left[f_{A,\widetilde{\Gamma}}\right](t,\omega) = \int_{\mathbb{R}^L} e^{i(tx+\omega^T\gamma)} f_{A,\widetilde{\Gamma}}(x,\gamma)dxd\gamma.$$

The conditional density of the treatment effect given by (2.6) coincides with the density of the treatment effect for the whole population if and only if $supp\left(\widetilde{\Gamma}\right) \subset I_Z$, *i.e.* the instrument has enough variation to capture all the heterogeneity in the population. The assumption $supp\left(\widetilde{\Gamma}\right) \subset I_Z$ could be easily tested as it is enough to check that $\int_{I_Z} r'_X(\gamma)d\gamma = 1$.

2.5. **Principles of Estimation.** To focus on the central innovation in this paper, we consider now the case of one instrument and limited support, as in the previous subsection (in the case of many instruments, we advocate the use of several unobservables, see below).

The integral over t in (2.6) is the Fourier transform inversion formula. The inverse Fourier transform of a ratio of characteristic functions is a classical structure for an estimator in deconvolution problems where one observes $A + B$ where A and B are independent and one knows the characteristic function of the error A. Optimal rates for deconvolution problems under various smoothness assumptions on the densities of A and B are given in Fan (1991), Butucea (2004) and Butucea and Tsybakov (2007).

Here we are reasoning conditional on the unobservable $\widetilde{\Gamma}$, because A and B are independent given $\widetilde{\Gamma}$ only. Conditional deconvolution also appears in Evdokimov (2010). In this article the density of the second equation unobserved heterogeneity parameter $\widetilde{\Gamma}$ acts as a mixing distribution. So we are dealing with conditional deconvolution in a nonparametric mixture context. The degree of ill-posedness should hence related to the decay to zero of the partial Fourier transform $\left| \mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right] (t, \gamma) \right|$ in t in a certain sense.

A second difference to standard deconvolution is that the distribution of A is unknown but estimable. This is a particular case of an inverse problem with an unknown but estimable operator, also encountered in econometrics in the case of nonparametric instrumental variables (see, *e.g.* Carrasco, Renault and Florens (2007)). This situation in a deconvolution problem has been studied in Diggle and Hall (1993), Neumann (1997) and later by Johannes (2009) and Comte and Lacour (2009). They obtain mathematical results when a preliminary sample allows to estimate the characteristic function of A via the empirical characteristic function[5] at rate $1/\sqrt{N}$. In there setting $A + B$ is observed on a second sample and the characteristic function of $A + B$ is again estimated via the empirical characteristic function. Neumann (1997) and Comte and Lacour (2009) obtain lower bounds, under various smoothness assumptions on the densities of A and B, that account for the extra difficulty of estimating the characteristic function of A. The estimator of Comte and Lacour (2009), which is built on that of Neumann (1997), is adaptive.

---

[5]

$$\widehat{\phi_A}(t) = \frac{1}{N} \sum_{i=1}^{N} \exp\left(it a_i\right).$$

A third difference is that we neither observe $A+B$ but can estimate the partial Fourier transform $\mathcal{F}_1\left[f_{A+B,\widetilde{\Gamma}}\right](t,\gamma)$ solving an inverse problem (see next paragraph). Up to our knowledge, this case has never been studied before. In Evdokimov (2010) $A+B$ is observed and the empirical characteristic function of $A+B$ in classical deconvolution problems is replaced by the Nadaraya-Watson type estimator of the characteristic function of $A+B$ given $\widetilde{\Gamma}$. This is not possible here.

In this article we use (2.7) and (2.8) to estimate the partial Fourier transform $\mathcal{F}_1\left[f_{A+B,\widetilde{\Gamma}}\right](t,\gamma)$ and $\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]$ that replace the characteristic functions in the classical deconvolution problem. The mixing distribution $f_{\widetilde{\Gamma}}$ is estimated estimating the two quantities in the ratio in (2.6). Estimation of the numerator is also the estimation of the derivative of a regression function. Estimation of derivatives of regression functions is typically an ill-posed inverse problem. Note also that in the case of degenerate design (points where the density is zero) the rates of estimation of regression functions can be degraded as for inverse problems, see for example Gaïffas (2009) and the references therein. Gaïffas (2009) proposes an adaptive procedure for the sup-norm using local polynomials and a selection rule similar to Lepski's method. However, up to our knowledge, there are no results for the estimation of derivatives in the presence of random and possibly degenerate design or extensions to an inverse problem setting (the case of Section 3). The case of degenerate random design is a situation encountered in an inverse problem setting in both Hoderlein, Klemelä and Mammen (2010) and Gautier and Kitamura (2009) where regressors are unbounded and degeneracy occurs at "infinity" for many designs of interest.

To summarize, our proposed estimator is computed in the following way:

(1) Step 1: $\mathcal{F}_1\left[\widehat{f_{A+B,\widetilde{\Gamma}}}\right]$ and $\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right](t,\gamma)$ are obtained by estimating the derivatives of the two regression functions. There are many estimators of derivatives of regression functions.[6]

---

[6] For example, local polynomials estimators (see, e.g., Fan and Gijbels (1996) and Tsybakov (2009)) can be used. In the case of (2.7) the estimator is obtained as follows

$$\mathcal{F}_1\left[\widehat{f_{A+B,\widetilde{\Gamma}}}\right](t,\gamma) \triangleq \frac{1}{Nh_N^2}\sum_{i=1}^N x_i e^{ity_i}W_{Ni}(\gamma) \text{ where } \qquad W_{Ni}(\gamma) \triangleq e_1 \mathcal{B}_{N\gamma}^{-1} U\left(\frac{z_{2i}-\gamma}{h_N}\right)K\left(\frac{z_{2i}-\gamma}{h_N}\right),$$

$$U(u) \triangleq \left(1,u,u^2/2!,\ldots,u^l/l!\right)^T \qquad \text{and} \qquad \mathcal{B}_{N\gamma} \triangleq \frac{1}{Nh_N}\sum_{i=1}^N U\left(\frac{z_{2i}-\gamma}{h_N}\right)U\left(\frac{z_{2i}-\gamma}{h_N}\right)^T K\left(\frac{z_{2i}-\gamma}{h_N}\right).$$

K is a kernel, h_N a bandwidth, e_1 a row of size $l+1$ with all coordinates being 0 but the second one which is equal to 1 and $l>1$, in practice $l=2$ is used for the estimation of one derivative.

(2) Step 2: Compute the following integral (in practice this is carried out numerically)[7]

$$(2.11) \qquad \widehat{f_{B|\widetilde{\Gamma}}}(b;\gamma) \triangleq \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ibt} K\left(t\ h_{N,\gamma}\right) \frac{\mathcal{F}_1\left[\widehat{f_{A+B,\widetilde{\Gamma}}}\right]}{\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right]}(t,\gamma) \mathbf{1}\left\{\left|\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right](t,\gamma)\right| > t_{N,t,\gamma}\right\} dt$$

where N is the sample size, $h_{N,\gamma}$ is a bandwidth going to zero with N, K a kernel (a typical kernel that we will use in Section 4 is $K(t) = \mathbf{1}\{|t| \leq 1\}$ with $h_{N,\gamma} = 1/R_{N,\gamma}^B$, it amounts to truncation of high frequencies), $t_{N,t,\gamma}$ a proper trimming factor. It has the same structure as that of Neumann (1997) [8]. $h_{N,\gamma}$ depends on γ because of the possible different decay rates of $\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)$ for different values of γ. When $\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right]$ is known, K avoids dividing by small values of $\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right]$ and $\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right]$ is small for large values of the frequency t.

(3) Step 3: Rely on plug-in estimators of the numerators and denominators of the estimable mixing density $r_X'(\gamma)/\int_{-\infty}^{\infty} r_X'(\gamma)d\gamma$. The numerator could be obtained by estimating the derivative of the choice probability and the denominator by integrating it numerically.[9] If we denote the corresponding estimator $\widehat{f_{\widetilde{\Gamma}|\widetilde{\Gamma}\in I_Z}}$, the global estimator is given by

$$(2.12) \qquad \widehat{f_{B|\widetilde{\Gamma}\in I_Z}}(b) = \int_{-\infty}^{\infty} \widehat{f_{B|\widetilde{\Gamma}}}(b;\gamma) \widehat{f_{\widetilde{\Gamma}|\widetilde{\Gamma}\in I_Z}}(\gamma) d\gamma.$$

Because the last integral with respect to γ in (2.12) should in practice be carried out numerically using a quadrature method, (1) and (2) need only to be carried a finite number of times. Section 4 considers the asymptotic behavior of this estimator.

---

[7] Alternatively the two following more simple estimators are possible

$$(2.9) \qquad \widehat{f_{B|\widetilde{\Gamma}}}(b;\gamma) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ibt} \frac{\mathcal{F}_1\left[\widehat{f_{A+B,\widetilde{\Gamma}}}\right]}{\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right]}(t,\gamma) \mathbf{1}\left\{\left|\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right](t,\gamma)\right| > t_{N,t,\gamma}\right\} dt$$

$$(2.10) \qquad \widehat{f_{B|\widetilde{\Gamma}}}(b;\gamma) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ibt} K\left(t\ h_{N,\gamma}\right) \frac{\mathcal{F}_1\left[\widehat{f_{A+B,\widetilde{\Gamma}}}\right]}{\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right]}(t,\gamma) dt.$$

(2.10) is in the spirit of Diggle and Hall (1993) and the asymptotic analysis and tuning of the bandwidth is known to be tricky, see, e.g. Neumann (1997).

[8] In Neumann (1997) and Comte and Lacour (2009) because the characteristic function of A is estimated at rate $1/\sqrt{N}$, $t_{N,t,\gamma}$ could be taken equal to $N^{-1/2}$, independent of t and γ.

[9] Under the assumption that $supp\left(\widetilde{\Gamma}\right) \subset I_Z$ the denominator is equal to 1 and we get the distribution of the treatment effect for the whole population. In that case the denominator does not need to be estimated. Recall that this could be tested.

To obtain marginal relevant treatment effect simply replace $d\gamma$ by $\omega(\gamma, \theta)d\gamma$ above. The asymptotic analysis in that case is the same as without the weight $\omega(\gamma, \theta)$ when the weight is bounded.

3. The Multivariate Unobservables Case

In this section, we extend the previous framework to cover models with a higher dimension of unobservables at the expense of being linear in the instruments. We show how to obtain the MTE, which is now a function of several unobservables, and extend the approach to obtain the distribution of treatment effects, including a discussion of the respective assumptions.

3.1. **MTE and ATE in the Case of Multivariate Unobservables and Instruments in the Selection Equation.**

3.1.1. *Model and Assumptions.* Compared to the univariate unobservable in the selection equation, which results in a model that has more of the flavor of a reduced form model, we formalize our (more structural) model with several unobservables and as many instruments as follows:

$$(3.1) \qquad \begin{cases} Y = A + BX \\ X = \mathbf{1}\{D^T Z_2 + V Z_1 < Z_3\} \end{cases}$$

where $Z_1 = 1$, and D is now a vector of unobservables of dimension $L - 2$. As discussed previously and in Gautier and Kitamura (2009), imposing that one coefficient in the selection equation has a sign is sufficient for identification of the distribution of the (scaled) random coefficients vector in this equation. To account for scale invariance, we divide the latent equation in (1.2) by this coefficient. We get the second equation of (3.1) when the coefficient of Z_3 is negative, otherwise change Z_3 in $-Z_3$. We use the notation $\widetilde{\Gamma} = (D^T, V)^T$ for our random vector of scaled first stage (selection equation, abbreviated FS) unobservables. Observe that this only makes a difference if $L = \dim(Z) > 2$, or else we are generally back in the previous model. We introduce the notation $S = (Z_2, 1)/\|(Z_2, 1)\|$ and $U = Z_3/\|(Z_2, 1)\|$, so that the FS becomes $X = \mathbf{1}\{\widetilde{\Gamma}^T S < U\}$. Note that the support of S is necessarily included in a hemisphere H, while U is a scalar which can be positive and negative.

It is possible to do everything below assuming working with a vector Γ of norm 1 in (1.2) as in Gautier and Kitamura (2009). We decide to present this different scaling in order to present a new approach to deal with problems involving a random coefficients binary choice. This alternative approach shares many similarities with HV. With this normalization the natural operator is no longer the Hemispherical transform but the Radon transform (see, *e.g.*, Helgason (1999)).

Theorem 3.1 below involves the Radon transform which is defined for $f \in L^1(\mathbb{R}^{L-1})$, s in a hemisphere of the Euclidian space $\mathbb{R}^{L-1}$ and u in $\mathbb{R}$ through

$$Rf(s,u) = \int_{P_{s,u}} f(\gamma) d_{P_{s,u}}(\gamma)$$

where $P_{s,u} = \{\gamma : \gamma^T s = u\}$ is an affine hyperplane of dimension $L-2$ in $\mathbb{R}^{L-1}$ and $d_{P_{s,u}}$ is the Lebesgue measure on $P_{s,u}$. Mathematical results regarding this integral transformation (among which its injectivity, an inversion formula involving the adjoint of the Radon transform and the projection theorem) can be found in Natterer (1996) and Helgason (1999). Statistical inverse problems involving this operator on the whole space appear in several problems from tomography (see for example Korostelev and Tsybakov (1993) and Cavalier (2000)) but also when one wishes to estimate distribution of random coefficients in the linear model with random coefficients (see Hoderlein, Klemelä and Mammen (2010)).

Assumption 3.1. *Let (Ω, F, P) be a complete probability space on which are defined the random vectors $(Y, X, S, U, A, B, \widetilde{\Gamma}) : \Omega \to \mathcal{Y} \times \mathcal{X} \times \mathcal{S} \times \mathcal{U} \times \mathcal{A} \times \mathcal{B} \times \mathcal{G}, \mathcal{Y} \subseteq \mathbb{R}, \mathcal{X} = \{0,1\}, \mathcal{S} \subseteq \mathbb{S}^{L-2}, \mathcal{U} \subset \mathbb{R}, \mathcal{A} \subseteq \mathbb{R}, \mathcal{B} \subseteq \mathbb{R}, \mathcal{G} \subseteq \mathbb{R}^{L-1}$ where L is an integer. The causal model is defined by equations (3.1) where the realizations of (Y, X, S, U) are observable whereas those of $(A, B, \widetilde{\Gamma})$ are not.*

Assumption 3.2. *(i) $(A, B, \widetilde{\Gamma})$ are independent of Z. (ii) The distribution of $A = Y_0$ and the distribution of $B = Y_1 - Y_0$ given $\widetilde{\Gamma} = \gamma$ have moments of order one and $\mathbb{E}[B|\widetilde{\Gamma} = \cdot]f_{\widetilde{\Gamma}}(\cdot) \in L^1(\mathbb{R}^{L-1})$. (iii) The conditional density of B given $\widetilde{\Gamma} = \gamma$ is in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ for almost every γ and $\mathbb{E}[e^{itA}|\widetilde{\Gamma} = \cdot]f_{\widetilde{\Gamma}}(\cdot)$ and $\mathbb{E}[e^{it(A+B)}|\widetilde{\Gamma} = \cdot]f_{\widetilde{\Gamma}}(\cdot)$ are in $L^1(\mathbb{R}^{L-1})$.*

Assumption 3.3. *There exists a ball B centered at 0 such that $supp\left(\widetilde{\Gamma}\right) \subset B \subset supp\left(SU\right)$.*

This assumption allows to deal with instruments with limited support as long as the unobserved heterogeneity varies within a limited range.

Theorem 3.1. Make assumptions 3.1, 3.2 (i) and (ii) and 3.3. Then, defining the arguments of R^{-1} below as 0 outside the support of (S, U), the following formula holds

$$\mathbb{E}[B|\widetilde{\Gamma} = \gamma] = \frac{R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}[Y|(S,U) = (\cdot)]\right](\gamma)}{R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}[X|(S,U) = (\cdot)]\right](\gamma)}$$

where R^{-1} is the inverse of the Radon transform (see, *e.g.* page 15 of Helgason (1999)).

Remark 3.1. Suppose that in the case of Section 2, instead of regressing on P, we regress on Z (Z_2 since $Z_1 = 1$), then we would obtain that

$$\mathbb{E}[B|V = p] = \frac{\partial_{z_2}\mathbb{E}\left[Y|Z_2 = \cdot\right]}{\partial_{z_2}\mathbb{E}\left[X|Z_2 = \cdot\right]}(\pi^{-1}(p)).$$

3.2. The Distribution of Treatment Effects in the Multivariate Unobservables and Instruments Case.

3.2.1. *Parameter of Interest.* In exactly the same setup as defined above in equation (3.1), we will be concerned with the distribution of treatment effect for the same subpopulation as the one considered by HV. More formally, we are interested in recovering

$$f_{B|\widetilde{\Gamma}}(b; \gamma) = f_{Y_1 - Y_0|\widetilde{\Gamma}}(b; \gamma),$$

for any $b \in supp(Y_1 - Y_0)$, and any $\gamma \in \mathcal{G}$. To emphasize the parallels to above, we call this the Distribution of Treatment Effects at the Margin", and abbreviate it DITEM. The interpretation is also quite similar to HV: For the subpopulation defined by $\widetilde{\Gamma} = \gamma$, it provides us with a measure for the effect of treatment. However, this measure is now the distribution of effects. As mentioned in the introduction, the distribution of effects is different from the effect of treatment on the distribution.

We will study this object under an additional identifying assumption

Assumption 3.4. $A \perp B|\widetilde{\Gamma}$

Recall that if $A = a(U_0, U_2)$ and $B = b(U_1, U_2)$, a sufficient condition for assumption (3.4) is that $\widetilde{\Gamma} = U_2$, and $U_0 \perp U_1|\widetilde{\Gamma}$. In words, there is a common driving factor that causes endogeneity in the selection models, and it is given by $\widetilde{\Gamma}$. In contrast to before, this is now an entire vector of unobservables and it is more realistic that this vector accounts for a potentially complicated structure of heterogeneity and correlation. Recall again that it does **not** mean that $A \perp B$. In fact, unless there is no endogenous selection, in general there will be dependence between Y_0, and $Y_1 - Y_0$. In other words, there is endogenous selection into treatment, but as far as it is endogenous, it is captured by the vector $\widetilde{\Gamma}$.

To recover the ccf of B at $\widetilde{\Gamma} = \gamma$, we require the following condition:

Assumption 3.5. Extending as zero outside the support of (S, U) the conditional expectation in the argument of R^{-1} below,

$$\forall (t, \gamma) \in \mathbb{R} \times supp\left(\widetilde{\Gamma}\right), \ R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}[e^{itY}(1 - X)|(S, U) = (s, u)]\right](\gamma) \neq 0.$$

This assumption is the analogue of the technical assumption 2.5 and is the classical assumption in deconvolution problems. Indeed we will see in the proofs that this corresponds to $\mathbb{E}[e^{itA}|\widetilde{\Gamma} = \gamma]f_{\widetilde{\Gamma}=\gamma}$.

3.2.2. *Main Result.* These assumptions allow us to characterize the DITEM.

Theorem 3.2. Let assumptions 3.1, 3.2 (i) and (iii), 3.3-3.5 be true. Defining the arguments of R^{-1} below to be zero outside the support of (S, U), the following formulas hold

$$(3.2) \qquad f_{B|\widetilde{\Gamma}}(b; \gamma) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itb} \frac{R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}\left[e^{itY}X|(S,U) = (\cdot)\right]\right](\gamma)}{-R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}[e^{itY}(1-X)|(S,U) = (\cdot)]\right](\gamma)} dt$$

$$(3.3) \qquad f_{\widetilde{\Gamma}}(\gamma) = R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}[X|(S,U) = (\cdot)]\right](\gamma).$$

Note the parallel to above: if we do not take the conditioning and derivative wrt p, but wrt z, we obtain:

$$f_{B|V}(b; \pi(z)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itb} \frac{\partial_z \pi^{-1} \partial_z \mathbb{E}\left[X e^{itY}|Z = z\right]}{-\partial_z \pi^{-1} \partial_z \mathbb{E}\left[(1-X)e^{itY}|Z = z\right]} dt.$$

3.3. **Estimation.** The overall estimator is the same as in Section 2.5 and just requires an adaptation of Step 1. We thus only present two possible ways to estimate the inverse of the derivatives of regression functions which appears in the numerator and denominator of (3.2) and in (3.3). We simply present the estimation of $f_{\widetilde{\Gamma}}$, the estimation of the numerator and denominator of (3.2) are the same.

A classical regularized inverse of the Radon transform is given by:

$$(3.4) \qquad R_R^{-1}[f](\gamma) \triangleq \int_H \int_{-\infty}^{\infty} K_R(s^T\gamma - u)f(s,u)dud\sigma(s)$$

where R is a smoothing parameter and

$$(3.5) \qquad \forall u \in \mathbb{R}, \ K_R(u) \triangleq 2(2\pi)^{-(L-1)} \int_0^R \cos(tu)t^{L-2}dt.$$

This suggests using as an estimator of the inverse Radon transform of a derivative of a regression function

$$\widehat{f_{\widetilde{\Gamma}}}(\gamma) = R_R^{-1}\left[\widehat{\frac{\partial}{\partial u}\mathbb{E}[X|(S,U)} = (\cdot)]\right](\gamma)$$

in the case of the estimation of $f_{\widetilde{\Gamma}}$ (for example), where $\widehat{\frac{\partial}{\partial u}\mathbb{E}[X|(S,U)} = (\cdot)]$ is an estimator of the derivative of the regression function and R is chosen adequately.

The following second estimator is easy to compute but only works in specific situations. We introduce

$$\forall u \in \mathbb{R}, \ \widetilde{K}_R(u) \triangleq 2(2\pi)^{-(L-1)} \int_0^R \sin(tu)t^{L-1}dt.$$

it is such that $K'_R(u) = \widetilde{K}_R(u)$ where K_R is defined by (3.5). Note that $\lim_{|u| \mapsto \infty} K'_R(u) = \lim_{|u| \mapsto \infty} K_R(u) = 0$ by the Riemann-Lebesgue Theorem. We make the following assumption

Assumption 3.6. *(i)* $\forall s \in H$, $supp\left(f_{U|S}(\cdot; s)\right) = \mathbb{R}$. *(ii) For almost every s in H, $u \mapsto \mathbb{E}[X|(S,U) = (s,u)]\widetilde{K}_R(u)$ and $u \mapsto \partial_u \mathbb{E}[X|(S,U) = (s,u)]K_R(u)$ are in $\mathrm{L}^1(\mathbb{R})$ and $u \mapsto \mathbb{E}[X|(S,U) = (s,u)]$ is continuous and $\lim_{|u| \mapsto \infty} \mathbb{E}[X|(S,U) = (s,u)]K_R(u) = 0$.*

This Assumption allows to justify an integration by parts argument for the regularized inverse counterparts of (3.3).

Proposition 3.1. Under Assumption 3.6,

$$\forall R \in \mathbb{R}, \ \forall \gamma \in \mathbb{R}^{L-1}, \ R_R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}\left[X|(S,U) = (\cdot)\right]\right](\gamma) = \mathbb{E}\left[\frac{\widetilde{K}_R(S^T\gamma - U)X}{f_{(S,U)}(S,U)}\right].$$

The trimmed sample counterpart estimator is given by

$$(3.6) \qquad \widehat{f_{\widetilde{\Gamma}}}(\gamma) \triangleq \frac{1}{N}\sum_{i=1}^{N} \frac{\widetilde{K}_{R_N^{\widetilde{\Gamma}}}(s_i^T\gamma - u_i)x_i}{\widehat{f_{(S,U)}}(s_i, u_i)}\mathbf{1}\left\{\left|\widehat{f_{(S,U)}}(s_i, u_i)\right| > t_N^{\widetilde{\Gamma}}\right\}$$

where $\widehat{f_{(S,U)}}$ is a plug-in estimator of $f_{(S,U)}$. We classically introduce trimming to avoid dividing by denominators too close to 0[10]. $t_N^{\widetilde{\Gamma}}$, $R_N^{\widetilde{\Gamma}}$ are trimming and smoothing parameters that should be adequately chosen.

For the estimation of the partial Fourier transforms in the numerator and denominator of (3.2) simply replace above x_i by respectively $e^{ity_i}x_i$ and $-e^{ity_i}(1-x_i)$. Because of possible different smoothness, the trimming and smoothing parameters should be adjusted as well for both the numerator and denominator.

Note that estimators already[11] exist if we normalize $\widetilde{\Gamma}$ to be of norm 1 using Gautier and Kitamura (2009) and Gautier and Le Pennec (2011).

---

[10] The trimming could be suppressed to obtain convergence in probability results when $f_{(S,U)}$ is bounded from below. This is not compatible with Assumption 3.6 (i). Note the similarity between the introduction of trimming here and in the deconvolution problem (via the kernel and trimming factor that accounts for estimation error on the denominator). When $f_{(S,U)}$ is unbounded from below the rates are degraded.

[11] There is an easy analogue for the estimation of the partial Fourier transforms.

4. Asymptotic Analysis

We denote by $\|f\|_p = \left(\int_{-\infty}^{\infty} f(b)^p db\right)^{1/p}$ for $p \in [1, \infty)$ the classical L^p norms and by $\|f\|_\infty$ the essential supremum norm, also called sup-norm for simplicity. We consider in this paper an upper bound on the squared L^2 risk for simplicity. C is a constant whose value can change from line to line.

4.1. **Estimation of f_B.** We will start with a proposition that relates the estimation of f_B with the estimation of $f_{\widetilde{\Gamma}}$ and of $f_{B|\widetilde{\Gamma}}$. Take $w: \ supp\left(\widetilde{\Gamma}\right) \to \mathbb{R}$ a weight function. We will make the following assumptions.

Assumption 4.1. $w^{-1} f_{\widetilde{\Gamma}} \in \mathrm{L}^\infty\left(supp\left(\widetilde{\Gamma}\right)\right).$

Assumption 4.2. $w f_{B|\widetilde{\Gamma}} \in \mathrm{L}^2\left(\mathbb{R} \times supp\left(\widetilde{\Gamma}\right)\right).$

Proposition 4.1. Let assumptions 3.1, 3.2 (i) and (iii), 3.3-3.5, 4.1 and 4.2 be true, then

$$\left\|\widehat{f_B} - f_B\right\|_2^2 \leq 3 \left\|\left(\widehat{f_{\widetilde{\Gamma}}} - f_{\widetilde{\Gamma}}\right) w^{-1}\right\|_\infty^2 \left\|\left(\widehat{f_{B|\widetilde{\Gamma}}}(\cdot;\star) - f_{B|\widetilde{\Gamma}}(\cdot;\star)\right) w(\star)\right\|_2^2$$

$$(4.1) \qquad + 3 \left\|f_{\widetilde{\Gamma}} w^{-1}\right\|_\infty^2 \left\|\left(\widehat{f_{B|\widetilde{\Gamma}}}(\cdot;\star) - f_{B|\widetilde{\Gamma}}(\cdot;\star)\right) w(\star)\right\|_2^2 + 3 \left\|\left(\widehat{f_{\widetilde{\Gamma}}} - f_{\widetilde{\Gamma}}\right) w^{-1}\right\|_\infty^2 \left\|f_{B|\widetilde{\Gamma}}(\cdot;\star) w(\star)\right\|_2^2$$

We now consider convergence in probability in order to easily handle the various plug-in terms, especially of $\widehat{f_{(S,U)}}$ in the second estimator in Section 3.3, and multiplications.

4.2. **Estimation of $f_{B|\widetilde{\Gamma}}$.** In order to work with smoothing and trimming factors in (2.11) that are independent of t and γ, we work with sup-norm consistency of the estimators of the partial Fourier transforms.

Assumption 4.3.

$$\sup_{t\in\mathbb{R}, \ \gamma\in supp(\widetilde{\Gamma})} \left|\mathcal{F}_1\left[\widehat{f_{A+B,\widetilde{\Gamma}}}\right] - \mathcal{F}_1\left[f_{A+B,\widetilde{\Gamma}}\right]\right| = O_p\left(r_{A+B,N}\right)$$

$$\sup_{t\in\mathbb{R}, \ \gamma\in supp(\widetilde{\Gamma})} \left|\mathcal{F}_1\left[\widehat{f_{A,\widetilde{\Gamma}}}\right] - \mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]\right| = O_p\left(r_{A,N}\right)$$

Unlike deconvolution situations with noise observed on a preliminary sample, each rate is nonparametric and it is the rate of estimation in an inverse problem. As mentioned in Section 3.1.1, we have used a scaling giving rise to the Radon transform to present an approach to the estimation of $f_{\widetilde{\Gamma}}$ different from that of Gautier and Kitamura (2009). However, we could rescale Γ to be on the sphere and inverse the hemispherical transform to obtain $f_{\widetilde{\Gamma}}$ and the partial Fourier transforms in (2.6), this is done in Section 5.1 . The minimax rates of estimation as well as an adaptive estimator

are given in given in Gautier and Le Pennec (2011) in the case of an estimator of f_Γ. In the proposition below we use the notation $\phi_{B|\widetilde{\Gamma}}(t;\gamma)$ for the ccf of B given $\widetilde{\Gamma}$ for $\widetilde{\Gamma} = \gamma$ evaluated at t.

Proposition 4.2. Let assumptions 3.1, 3.2 (i) and (iii), 3.3, 3.5 and 4.3 hold. Take $t_{N,t,\gamma} = r_{A,N}$. The following upper bound holds

(4.2)

$$\left\| \left(\widehat{f_{B|\widetilde{\Gamma}}}(\cdot;\star) - f_{B|\widetilde{\Gamma}}(\cdot;\star) \right) w(\star) \right\|_2^2 = O_p \left(\int_{supp(\widetilde{\Gamma})} \int_{-\infty}^{\infty} \left[(1 - K(t\, h_{N,\gamma}))^2 \left| \phi_{B|\widetilde{\Gamma}}(t;\gamma) \right|^2 \right. \right.$$

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

Besides the integral in γ, the upper bound is the same as in Comte and Lacour (2009) where $r_{A+B,N}^2$ and $r_{A,N}^2$ are respectively $1/\sqrt{N}$ and $1/\sqrt{M}$, with M is the sample size of the preliminary sample used to estimate the characteristic function of A. Here, these parametric rates are replaced by ill-posed inverse problems nonparametric rates. The first term in the upper bound is the square of the approximation bias.

Because $\left| \phi_{B|\widetilde{\Gamma}}(t;\gamma) \right| \leq 1$, we obtain as a corollary of (4.2)

(4.3) $$\left\| \left(\widehat{f_{B|\widetilde{\Gamma}}}(\cdot;\star) - f_{B|\widetilde{\Gamma}}(\cdot;\star) \right) w(\star) \right\|_2^2 = O_p \left(\int_{supp(\widetilde{\Gamma})} \int_{-\infty}^{\infty} \left[(1 - K(t\, h_{N,\gamma}))^2 \left| \phi_{B|\widetilde{\Gamma}}(t;\gamma) \right|^2 \right. \right.$$

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

When $supp\left(\widetilde{\Gamma}\right)$ is bounded, we can take $w \equiv 1$ above. Another sensible choice is to take $w = \left(f_{\widetilde{\Gamma}} \right)^\alpha$ for some $\alpha \in (0,1]$ ensuring integrability. Note that when $\alpha = 1$, the following term $\left| \phi_{B|\widetilde{\Gamma}}(t;\gamma) f_{\widetilde{\Gamma}}(\gamma) \right|^2 = \left| \mathcal{F}_1 \left[f_{B,\widetilde{\Gamma}} \right](t,\gamma) \right|^2$ appears.

More precise rates could be obtained making smoothness assumptions implying specific rates $r_{A+B,N}^2$ and $r_{A,N}^2$ as well as a smoothness assumption on $f_{B|\widetilde{\Gamma}}$ and an assumption on the decay rate to zero of $\left| \mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right](t;\gamma) \right|$. An adaptation of classical ellipsoids for $f_{B|\widetilde{\Gamma}}$ is

$$\mathcal{A}_{\delta,r,a,w}(L) = \left\{ f \text{ conditional density on } \mathbb{R} \text{ given } \gamma \in \mathbb{R}^{L-1} : \right.$$

$$\left. \int_{supp(\widetilde{\Gamma})} \int_{-\infty}^{\infty} |\mathcal{F}[f](t;\gamma)|^2 \left(1 + t^2\right)^\delta \exp\left(2a|t|^r\right) dt w^2(\gamma) d\gamma \leq L^2 \right\}$$

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). The case where $w \equiv 1$ and $w = f_{\widetilde{\Gamma}}$ are the more natural ones.

When $K(t) = \mathbf{1}\{|t| \leq 1\}$ and we take $h_{N,\gamma}$ of the form $1/R_N^B$, the square of the approximation bias can be bounded in the following way

$$\int_{supp(\widetilde{\Gamma})} \int_{-\infty}^{\infty} \left| \phi_{B|\widetilde{\Gamma}}(t;\gamma) w(\gamma) \right|^2 dt d\gamma \leq L^2 \left(\left(R_N^B \right)^2 + 1 \right)^{-\delta} \exp\left(-2a \left(R_N^B \right)^r \right).$$

The assumption on the decay rate of $\left| \mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right](t;\gamma) \right|$ strengthens the assumption 3.5.

Assumption 4.4. *There exits $s(\gamma) \geq 0$, $b(\gamma) > 0$, $\eta(\gamma) \in \mathbb{R}$ ($\eta(\gamma) > 0$ if $s(\gamma) = 0$) and $k_0(\gamma), k_1(\gamma) > 0$ such that*

(1) [(i)]

(2)

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

(3) *or*

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

Proposition 4.3. Let assumptions 3.1, 3.2 (i) and (iii), 3.3, 3.5 and 4.3 hold. Assume either (1): $supp\left(\widetilde{\Gamma}\right)$ is bounded, $w \equiv 1$, and $f_{B|\widetilde{\Gamma}}$ belongs to $\mathcal{A}_{\delta,r,a,1}(L)$ and Assumption 4.4 (2) with constants independent of γ, or (2): $w = f_{\widetilde{\Gamma}}$, $f_{B|\widetilde{\Gamma}}$ belongs to $\mathcal{A}_{\delta,r,a,f_{\widetilde{\Gamma}}}(L)$ and Assumption 4.4 (3) with constants independent of γ. Take $t_{N,t,\gamma} = r_{A,N}$. The following upper bounds hold:

(1) [(i)]

(2) if $s = r = 0$, then

(4.4)
$$\left\| \left(\widehat{f_{B|\widetilde{\Gamma}}}(\cdot;\star) - f_{B|\widetilde{\Gamma}}(\cdot;\star) \right) w(\star) \right\|_2^2 = O_p\left(\left(R_N^B \right)^{-2\delta} + r_{A+B,N}^2 \left(R_N^B \right)^{2\eta+1} + r_{A,N}^2 \left(R_N^B \right)^{2\max(\eta-\delta,0)+1} \right);$$

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

(4.5)
$$\left\| \left(\widehat{f_{B|\widetilde{\Gamma}}}(\cdot;\star) - f_{B|\widetilde{\Gamma}}(\cdot;\star) \right) w(\star) \right\|_2^2 = O_p\left(\left(R_N^B \right)^{-2\delta} + e^{2b(R_N^B)^s} \left(r_{A+B,N}^2 \left(R_N^B \right)^{2\eta+1-s} + r_{A,N}^2 \left(R_N^B \right)^{\min(1+2\eta-s,2(\eta-\delta))} \right) \right);$$

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

(4.6) $\quad \left\| \left(\widehat{f_{B|\widetilde{\Gamma}}}(\cdot;\star) - f_{B|\widetilde{\Gamma}}(\cdot;\star) \right) w(\star) \right\|_2^2 = O_p\left(\left(R_N^B \right)^{-2\delta} e^{-2a(R_N^B)^r} + r_{A+B,N}^2 \left(R_N^B \right)^{2\eta+1} + r_{A,N}^2 \right);$

(5) if $s > 0$ and $r > 0$, then

(4.7)
$$\left\| \left(\widehat{f_{B|\widetilde{\Gamma}}}(\cdot;\star) - f_{B|\widetilde{\Gamma}}(\cdot;\star) \right) w(\star) \right\|_2^2 = O_p \left(\left(R_N^B \right)^{-2\delta} e^{-2a(R_N^B)^r} + r_{A+B,N}^2 \left(R_N^B \right)^{2\eta+1-s} e^{2b(R_N^B)^s} + r_{A,N}^2 \Delta \left(R_M^B \right) \right)$$

where

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

$$+ \mathbf{1}\{\{r > s\} \cup \{r = s, b < a\}\}.$$

5. APPENDIX

5.1. An alternative approach to the identification and estimation of the partial Fourier transforms.

In this section we assume that $\mathbb{P}(\Gamma = 0) = 0$, we rescale Γ in (1.2) to be of norm 1 and denote by $S = X/\|X\|$. Consider for example the case of $\mathcal{F}_1 [f_{A,\Gamma}]$.

$$\mathbb{E}[(1 - X)e^{itY} | S = s] = \mathbb{E}[(1 - X)e^{itA} | S = s]$$

$$= \mathbb{E}[e^{itA}] - E[\mathbf{1}\{s^T\gamma > 0\}e^{itA}] \quad \text{(using (1.3))}$$

$$= \mathbb{E}[e^{itA}] - \int_{\mathbb{S}^{L-1}} \mathbf{1}\left\{ s^T\gamma > 0 \right\} \left(\mathcal{F}_1 [f_{A,\Gamma}] (t,\gamma) \right) d\sigma(\gamma)$$

$$= \mathbb{E}[e^{itA}] - \mathcal{H} \left(\mathcal{F}_1 [f_{A,\Gamma}] (t,\cdot) \right) (s)$$

(5.1)
$$= \frac{1}{2}\mathbb{E}[e^{itA}] - \mathcal{H} \left(\left(\mathcal{F}_1 [f_{A,\Gamma}] (t,\cdot) \right)^- \right) (s)$$

where σ is the spherical measure on the sphere $\mathbb{S}^{L-1}$ of the Euclidian space $\mathbb{R}^L$, $\mathcal{H}$ is the hemispherical transform (see, e.g., Gautier and Kitamura (2009)) and f^- is the odd part[12] of a function f. If we assume full support of the regressors then the (i) of the following assumption holds.

Assumption 5.1. (1) [(i)]

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

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

---

[12] Odd, respectively even, functions are the closure in $L^2(\mathbb{S}^{L-1})$ of continuous functions such that $\forall s \in \mathbb{S}^{L-1}$, $f(-s) = -f(s)$, respectively $\forall s \in \mathbb{S}^{L-1}$, $f(-s) = f(s)$. Each function in $L^2(\mathbb{S}^{L-1})$ is the sum of its odd and even part. We denote by $L^2_{odd}(\mathbb{S}^{d-1})$ the subspace of $L^2(\mathbb{S}^{d-1})$ of odd functions.

Assumption 5.1 (3) is slightly weaker than the assumption we made in Section 3.1.1. Equation (5.1) yield that, under Assumption 5.1 (3), $\mathbb{E}[(1-X)e^{itY}|S = s] - \frac{1}{2}\mathbb{E}[e^{itA}]$ can be extended in a unique way as an odd function defined on the whole $\mathbb{S}^{L-1}$ (it is initially only defined on H^+ according to Assumption 5.1 (2)) through

$$\forall s \in H^+, \ R_A(t, s) = \mathbb{E}[(1-X)e^{itY}|S = s] - \frac{1}{2}\mathbb{E}[e^{itA}]$$

$$\forall s \in -H^+, \ R_A(t, s) = -R_A(t, -s).$$

It is remarkable that $\mathbb{E}[e^{itA}]$ 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 $R_A(t, \cdot)$ belong to $\mathcal{H}(\mathrm{L}^2_{odd}(\mathbb{S}^{d-1}))$, it is continuous and odd. Thus for any point $\tilde{s}$ on the boundary of H^+, $R_A(t, \tilde{s}) = -R_A(t, -\tilde{s})$. This yields

(5.2) $$\lim_{s \to \tilde{s}, \ s \in H^+} \mathbb{E}[(1-X)e^{itY}|S = s] + \lim_{s \to -\tilde{s}, \ s \in H^+} \mathbb{E}[(1-X)e^{itY}|S = s] = \mathbb{E}[e^{itA}].$$

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

(5.3)
$$\widehat{\mathcal{F}_1[f_{A,\Gamma}]}(t, \gamma) = \max\left(\frac{2}{|\mathbb{S}^{L-1}|}\sum_{p=0}^{T_N-1}\frac{\chi(2p+1, 2T_N)h(2p+1, L)}{\lambda(2p+1, L)C_{2p+1}^{\nu(L)}(1)}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\left(2(x_i-1)e^{ity_i} + \widehat{\phi_A}(t)\right)C_{2p+1}^{\nu(L)}(s_i^T\gamma)}{\max\left(\hat{f}_S(s_i), m_N\right)}\right), 0\right),$$

where $|\mathbb{S}^{L-1}| = \frac{2\pi^{L/2}}{\Gamma(L/2)}$ is the surface measure of $\mathbb{S}^{L-1}$, $h(n, L) = \frac{(2n+L-2)(n+L-2)!}{n!(L-2)!(n+L-2)}$, $\nu(L) = (L-2)/2$, $\lambda(2p+1, L) = \frac{(-1)^p|\mathbb{S}^{L-2}|1\cdot3\cdots(2p-1)}{(L-1)(L+1)\cdots(L+2p-1)}$, $\chi(n, T) = \psi(n/T)$ where $\psi : [0, \infty) \to [0, \infty)$ is infinitely differentiable, nonincreasing, such that $\psi(x) = 1$ if $x \in [0, 1]$, $0 \leq \Psi(x) \leq 1$ if $x \in [1, 2]$, $\psi(x) = 0$ if $x \geq 2$, and $C_n^\nu(\cdot)$ are the Gegenbauer polynomials[13]. T_N is the smoothing parameter, m_N a trimming factor and $\hat{f}_S$ an estimator of the density of S.

5.2. Proofs.

---

[13] 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 $\mathbb{N}\setminus\{0\}$, $(a)_n = a(a+1)\cdots(a+n-1) = \Gamma(a+n)/\Gamma(a)$.

5.2.1. *Proof of Theorem 2.2.* Consider the conditional expectations:

$$\mathbb{E}\left[(1-X)e^{itY}|P=p\right] = \mathbb{E}\left[(1-X)e^{itA}e^{itBX}|P=p\right]$$

Note that $(1-X)e^{itA}e^{itBX} = (1-X)e^{itA}$. Hence

$$
\begin{aligned}
-\mathbb{E}\left[(1-X)e^{itY}|P=p\right] &= -\mathbb{E}\left[(1-X)e^{itA}|P=p\right] \\
&= -\mathbb{E}\left[(1-X)\mathbb{E}\left[e^{itA}|V,P\right]|P=p\right] \\
&= -\mathbb{E}\left[(1-X)\mathbb{E}\left[e^{itA}|V\right]|P=p\right] \quad \text{(from Assumption (2.3) (i))} \\
&= -\int_p^1 \mathbb{E}\left[e^{itA}|V=v\right]dv \quad \text{(because } V|Z \sim \mathcal{U}(0,1)) \\
&= -\int_0^1 \mathbb{E}\left[e^{itA}|V=v\right]dv + \int_0^p \mathbb{E}\left[e^{itA}|V=v\right]dv
\end{aligned}
$$

Differentiating with respect to p produces

$$-\partial_p\mathbb{E}\left[(1-X)e^{itY}|P=p\right] = \mathbb{E}\left[e^{itA}|V=p\right]$$

Similarly,

$$
\begin{aligned}
\mathbb{E}\left[Xe^{itY}|P=p\right] &= \int_0^p \mathbb{E}\left[e^{it(A+B)}|V=v\right]dv \\
&= \int_0^p \mathbb{E}\left[e^{itA}e^{itB}|V=v\right]dv \\
&= \int_0^p \mathbb{E}\left[e^{itA}|V=v\right]\mathbb{E}\left[e^{itB}|V=v\right]dv \quad \text{(from Assumption (2.4))}
\end{aligned}
$$

Differentiating wrt p produces

$$\partial_p\mathbb{E}\left[Xe^{itY}|P=p\right] = \mathbb{E}\left[e^{itA}|V=p\right]\mathbb{E}\left[e^{itB}|V=p\right].$$

As a consequence,

$$-\frac{\partial_p\mathbb{E}\left[Xe^{itY}|P=p\right]}{\partial_p\mathbb{E}\left[(1-X)e^{itY}|P=p\right]} = \mathbb{E}\left[e^{itB}|V=p\right]$$

Q.E.D.

5.2.2. *Proof of Theorem 2.3 and Theorem 2.4.* Like in the proof of Theorem 2.2 we can check that

$$\partial_p\mathbb{E}\left[XY^2|P=p\right] = \mathbb{E}\left[B^2|V=p\right] + 2\mathbb{E}\left[AB|V=p\right] + \mathbb{E}\left[A^2|V=p\right]$$

and

$$
\begin{aligned}
&\partial_p\mathbb{E}\left[XY^2|P=p\right] - \partial_p\mathbb{E}\left[(1-X)Y^2|P=p\right] \\
&= \mathbb{E}\left[B^2|V=p\right] + 2\mathbb{E}\left[AB|V=p\right] + \mathbb{E}\left[A^2|V=p\right] - \mathbb{E}\left[A^2|V=p\right]
\end{aligned}
$$

$$= \mathbb{E}\left[B^2|V=p\right] + 2\mathbb{E}\left[AB|V=p\right].$$

Since

(5.4) $$\left|\mathbb{E}\left[AB|V=p\right]\right| \le \mathbb{E}\left[|AB|\,|V=p\right] \le \sqrt{\mathbb{E}\left[B^2|V=p\right]\mathbb{E}\left[A^2|V=p\right]},$$

we get

$$\frac{\left\{\partial_p\mathbb{E}\left[XY^2|P=p\right] - \partial_p\mathbb{E}\left[(1-X)Y^2|P=p\right] - \mathbb{E}\left[B^2|V=p\right]\right\}^2}{4\mathbb{E}\left[B^2|V=p\right]} \le \mathbb{E}\left[A^2|V=p\right]$$
$$= \partial_p\mathbb{E}\left[(1-X)Y^2|P=p\right].$$

Theorem 2.1 yields that

$$\left(Var(B|V=p) - \partial_p\mathbb{E}\left[XY^2|P=p\right] + (\partial_p\mathbb{E}\left[Y|P=p\right])^2 + \partial_p\ \mathbb{E}\left[(1-X)Y^2|P=p\right]\right)^2$$
$$\le 4\partial_p\mathbb{E}\left[(1-X)Y^2|P=p\right]\left(Var(B|V=p) + (\partial_p\mathbb{E}\left[Y|P=p\right])^2\right),$$

thus

$$\left(Var(B|V=p) - \partial_p\mathbb{E}\left[XY^2|P=p\right] + (\partial_p\mathbb{E}\left[Y|P=p\right])^2 - \partial_p\ \mathbb{E}\left[(1-X)Y^2|P=p\right]\right)^2$$
$$\le 4\partial_p\mathbb{E}\left[(1-X)Y^2|P=p\right]\partial_p\mathbb{E}\left[XY^2|P=p\right].$$

The bounds obtained are sharp because the only inequality comes from (5.4) and there could equality in the inequality.

Q.E.D.

5.2.3. *Proof of Theorem 3.1.* Similar computations as before yield

$$\mathbb{E}[Y|(S,U)=(s,u)] = \mathbb{E}[A] + \int_{\mathbb{R}^{L-1}} \mathbb{E}[B|\widetilde{\Gamma}=\gamma]f_{\widetilde{\Gamma}}(\gamma)\mathbf{1}\{\gamma^T s < u\}d\gamma$$
$$= \mathbb{E}[A] + \int_{-\infty}^{u}\int_{P_{s,v}} \mathbb{E}[B|\widetilde{\Gamma}=\gamma]f_{\widetilde{\Gamma}}(\gamma)dP_{s,v}(\gamma)dv,$$

where the last identity holds because the Lebesgue measure on $\mathbb{R}^{L-1}$ is the product of the Lebesgue measure on $\mathbb{R}^{L-2}$ and on $\mathbb{R}$. As a consequence, for (u,s) in $supp\left((U,S)\right)$,

$$\frac{\partial}{\partial u}\mathbb{E}[Y|(S,U)=(s,u)] = R\left[\mathbb{E}[B|\widetilde{\Gamma}=\cdot]f_{\widetilde{\Gamma}}(\cdot)\right](s,u)$$

and

$$\frac{\partial}{\partial u}\mathbb{E}[X|(S,U)=(s,u)] = R\left[f_{\widetilde{\Gamma}}(\cdot)\right](s,u).$$

The equations in integral form imply both that, on the support of (S,U), the derivatives of the left hand-side regression functions exists and that they are images via the Radon transform of well defined

$L^1(\mathbb{R}^{L-1})$ functions. Assumption 3.3 implies that $R\left[\mathbb{E}[B|\widetilde{\Gamma}=\cdot]f_{\widetilde{\Gamma}}(\cdot)\right](s,u)$ and $R\left[f_{\widetilde{\Gamma}}(\cdot)\right](s,u)$ are 0 outside the support of (S,U) so 0 is the natural extension for the left hand side expressions outside the support of (S,U). Based on these extensions, it is now possible to apply the inverse operator R^{-1}. This yields that

$$R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}[Y|(S,U)=(\cdot)]\right](\gamma)=\mathbb{E}[B|\widetilde{\Gamma}=\gamma]f_{\widetilde{\Gamma}}(\gamma)$$

and

$$R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}[X|(S,U)=(\cdot)]\right](\gamma)=f_{\widetilde{\Gamma}}(\gamma).$$

Q.E.D.

5.2.4. *Proof of Theorem 3.2.* The proof is very similar to that of Theorem 2.2 and we first show that

$$\begin{aligned}\frac{\partial}{\partial u}\mathbb{E}\left[e^{itY}X|(S,U)=(s,u)\right]&=&R\left[\mathbb{E}[e^{itA+B}|\widetilde{\Gamma}=\cdot]f_{\widetilde{\Gamma}}(\cdot)\right](s,u)\\&=&R\left[\mathbb{E}[e^{itA}|\widetilde{\Gamma}=\cdot]\mathbb{E}[e^{itB}|\widetilde{\Gamma}=\cdot]f_{\widetilde{\Gamma}}(\cdot)\right](s,u),\end{aligned}$$

where the second equality follows from the conditional independence assumption (3.4).
Moreover

$$\frac{\partial}{\partial u}\mathbb{E}[e^{itY}(1-X)|(S,U)=(s,u)]=-R\left[\mathbb{E}[e^{itA}|\widetilde{\Gamma}=\cdot]f_{\widetilde{\Gamma}}(\cdot)\right](s,u).$$

Equation (3.3) has been proved in the proof of Theorem 3.1.

Q.E.D

5.2.5. *Proof of Proposition 3.1.* Consider the case of the numerator in (3.2). The result is based on the following computations.

$$\begin{aligned}&R_R^{-1}\left[\frac{\partial}{\partial u}\mathbb{E}\left[e^{itY}X|(S,U)=(\cdot)\right]\right](\gamma)\\&=\int_{supp(S)}\widetilde{K}_R(s^T\gamma-u)\mathbb{E}\left[e^{itY}X|(S,U)=(s,u)\right]dud\sigma(s)\\&=\int_{supp((S,U))}\widetilde{K}_R(s^T\gamma-u)\mathbb{E}\left[e^{itY}X|(S,U)=(s,u)\right]\frac{f_{(S,U)}(s,u)}{f_{(S,U)}(s,u)}dud\sigma(s)\\&=\int_{supp((S,U))}\mathbb{E}\left[\left.\frac{\widetilde{K}_R(s^T\gamma-u)e^{itY}X}{f_{(S,U)}(s,u)}\right|(S,U)=(s,u)\right]f_{(S,U)}(s,u)dud\sigma(s)\\&(5.5)\quad=\mathbb{E}\left[\frac{\widetilde{K}_R(S^T\gamma-U)e^{itY}X}{f_{(S,U)}(S,U)}\right]\quad\text{(using the law of iterated conditional expectations)}\end{aligned}$$

where for the first equality we use the integration by parts formula and Assumption 3.6.

Q.E.D.

5.2.6. *Proof of Proposition 4.1.* The proposition is a direct consequence of the relations $\hat{a}\hat{b} - ab = (\hat{a} - a)(\hat{b} - b) + a(\hat{b} - b) + (\hat{a} - a)b$, $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ and the Hölder inequality. Q.E.D.

5.2.7. *Proof of Proposition 4.2.* We introduce the notations

$$\overline{f}_{B|\widetilde{\Gamma}}(b; \gamma) \triangleq \frac{1}{2\pi} \int_{-\infty}^{\infty} K(t\, h_{N,\gamma}) e^{-ibt} \phi_{B|\widetilde{\Gamma}}(t; \gamma) dt,$$

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

The following decomposition holds by means of the Plancherel identity:

$$\left\|\left(\widehat{f_{B|\widetilde{\Gamma}}} - f_{B|\widetilde{\Gamma}}\right)(\cdot; \gamma)\right\|_2^2 \leq 4\left\|\left(\overline{f}_{B|\widetilde{\Gamma}} - f_{B|\widetilde{\Gamma}}\right)(\cdot; \gamma)\right\|_2^2$$

$$+ \frac{2}{\pi} \int_{-\infty}^{\infty} \frac{K(t\, h_{N,\gamma})^2}{\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t, \gamma)\right|^2} \left|\left(\mathcal{F}_1\left[\widehat{f_{A+B,\widetilde{\Gamma}}}\right] - \mathcal{F}_1\left[f_{A+B,\widetilde{\Gamma}}\right]\right)(t, \gamma)\right|^2 dt$$

$$+ \frac{2}{\pi} \int_{-\infty}^{\infty} K(t\, h_{N,\gamma})^2 |R(t, \gamma)|^2 \left|\left(\mathcal{F}_1\left[\widehat{f_{A+B,\widetilde{\Gamma}}}\right] - \mathcal{F}_1\left[f_{A+B,\widetilde{\Gamma}}\right]\right)(t, \gamma)\right|^2 dt$$

$$+ \frac{2}{\pi} \int_{-\infty}^{\infty} K(t\, h_{N,\gamma})^2 \left|\mathcal{F}_1\left[f_{A+B,\widetilde{\Gamma}}\right](t, \gamma)\right|^2 |R(t, \gamma)|^2 dt.$$

We conclude using Lemma 5.1 below and the fact that by conditional independence

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

Q.E.D.

Lemma 5.1 below is an adaptation of the lemma of Neumann (1997). Denote by

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

Lemma 5.1.

$$\sup_{t \in \mathbb{R},\ \gamma \in supp(\widetilde{\Gamma})} \left\{\psi(t, \gamma)^{-1} |R(t, \gamma)|\right\} = O_p(1).$$

5.2.8. *Proof of Lemma 5.1.* We distinguish between two cases.

Case 1: Let t and γ be such that $\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right| < 2r_{A,N}$. Then, $\psi(t,\gamma)^{-1} \le 2\left|\widehat{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]}(t,\gamma)\right|$ and

it suffices to upper bound in probability $\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right||R(t,\gamma)|$. By definition of $R(t,\gamma)$,

$\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right||R(t,\gamma)| \le 1$ on the event $\left\{\left|\widehat{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]}(t,\gamma)\right| \le r_{A,N}\right\}$, while

$\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right||R(t,\gamma)| \le (r_{A,N})^{-1}\left|\left(\widehat{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]} - \mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]\right)(t,\gamma)\right|$ on the complementary event

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

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

Case 2: Let now t and γ be such that $\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right| \ge 2r_{A,N}$. Then, $\psi(t,\gamma)^{-1} \le 2(r_{A,N})^{-1}\left|\widehat{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]}(t,\gamma)\right|^2$

and it suffices to upper bound in probability $(r_{A,N})^{-1}\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right|^2|R(t,\gamma)|$.

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

$$(r_{A,N})^{-1}\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right|^2|R(t,\gamma)|$$

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

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

Using

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

we obtain

$$(r_{A,N})^{-1}\left|\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right|^2|R(t,\gamma)|$$

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

$$+ (r_{A,N})^{-1}\left(\left|\widehat{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]}(t,\gamma) - \mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right| + \frac{\left|\widehat{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]}(t,\gamma) - \mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right](t,\gamma)\right|^2}{\left|\widehat{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]}(t,\gamma)\right|^2}\right)\mathbf{1}\left\{\left|\widehat{\mathcal{F}_1\left[f_{A,\widetilde{\Gamma}}\right]}(t,\gamma)\right| > r_{A,N}\right\}$$

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

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

$$+ \left. (r_{A,N})^{-2} \left| \widehat{\mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right]}(t,\gamma) - \mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right](t,\gamma) \right|^2 \right) \mathbf{1} \left\{ \left| \widehat{\mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right]}(t,\gamma) \right| > r_{A,N} \right\}.$$

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

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

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

$$\leq \mathbf{1} \left\{ \left| \left(\widehat{\mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right]} - \mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right] \right)(t,\gamma) \right| \geq \left| \mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right](t,\gamma) \right| / 2 \right\}$$

$$\leq 2 \frac{\left| \left(\widehat{\mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right]} - \mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right] \right)(t,\gamma) \right|}{\left| \mathcal{F}_1 \left[f_{A,\widetilde{\Gamma}} \right](t,\gamma) \right|}$$

which yields

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

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

Q.E.D.

5.2.9. *Proof of Proposition 4.3.* The proposition follows from adapting the upper bounds in Comte and Lacour (2009), (4.2) and the assumptions made.

Q.E.D.

REFERENCES

[1] 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.

[2] BERAN, R., AND P. HALL (1992): "Estimating Coefficient Distributions in Random Coefficient Regression". *Annals of Statistics*, **20**, 1970–1984.

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

[4] 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.

[5] CARNEIRO, P., J. HECKMAN, AND E. VYTLACIL (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin". *Econometrica*, **78**, 377–394.

[6] CARNEIRO, P., AND S. LEE (2009): "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality". *Journal of Econometrics*, **149**, 191–208.

[7] 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.

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

[9] CHERNOZHUKOV, V., AND C. HANSEN. (2005): "An IV Model of Quantile Treatment Effects". *Econometrica*, **73**, 245–261.

[10] CHERNOZHUKOV, V., AND C. HANSEN (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models". *Journal of Econometrics*, **132**, 491–525.

[11] COMTE, F., AND C. LACOUR (2011): "Deconvolution with Estimated Characteristic Function of the Errors". *Journal of the Royal Statistical Society: Series B*, **73**, 601–627.

[12] DIGGLE, P. J., AND P. HALL (1993): "Fourier Approach to Nonparametric Deconvolution of a Density Estimate". *Journal of the Royal Statistical Society. Series B*, **55**, 523–531.

[13] 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.

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

[15] FAN, J. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems". *Annals of Statistics*, **19**, 1257–1272.

[16] FAN, J., AND I. GIJBELS (1996): *Local Polynomial Modelling and Its Applications*. Chapman Hall, London.

[17] FOX, J., AND A. GANDHI (2009): "Identifying Heterogenity in Economic Choice and Selection Models Using Mixtures". Working Paper.

[18] FUNK, P. (1916): "Uber Eine Geometrische Anwendung der Abelschen Integralgleichung". *Mathematische Annalen*, **77**, 129–135.

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

[20] GAUTIER, E., AND Y. KITAMURA (2009): "Nonparametric Estimation in Random Coefficients Binary Choice Models". Preprint arXiv:0907.2451.

[21] GAUTIER, E., AND E. LE PENNEC (2011): "Adaptive Estimation in Random Coefficients Binary Choice Models Using Needlet Thresholding". Preprint arXiv:1106.3503.

[22] GROEMER, H. (1996): *Geometric Applications of Fourier Series and Spherical Harmonics*. Cambridge University Press: Cambridge, *Encyclopedia of Mathematics and its Applications*.

[23] 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.

[24] 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.

[25] HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation". *Econometrica*, **73**, 669–738.

[26] HECKMAN, J. J., AND E. VYTLACIL (2007) "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.

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

[28] HODERLEIN, S., J. KLEMELÄ, AND E. MAMMEN (2010): "Analyzing the Random Coefficient Model Nonparametrically". *Econometric Theory*, **26**, 804–837.

[29] 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.

[30] IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects". *Econometrica*, **62**, 467–475.

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

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

[33] KOROSTELEV, A. P., AND A. B. TSYBAKOV (1993): *Minimax Theory of Image Reconstruction*. Springer, New-York, *Lecture Notes in Statistics* **82**.

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

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

[36] RUBIN, B. (1999): "Inversion and Characterization of the Hemispherical Transform". *Journal d'Analyse Mathématique*, **77**, 105–128.

[37] TSYBAKOV, A. B. (2009): *Introduction to Nonparametric Estimation*. Springer, New-York.

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`