Partially Linear Models under Data Combination*

Xavier D'Haultfoeuille† Christophe Gaillac‡ Arnaud Maurel§

First Version: April 11, 2022
This Version: March 3, 2023

Abstract

We study partially linear models when the outcome of interest and some of the covariates are observed in two different datasets that cannot be linked. This type of data combination problem arises very frequently in empirical microeconomics. Using recent tools from optimal transport theory, we derive a constructive characterization of the sharp identified set. We then build on this result and develop a novel inference method that exploits the specific geometric properties of the identified set. Our method exhibits good performances in finite samples, while remaining very tractable. We apply our approach to study intergenerational income mobility over the period 1850-1930 in the United States. Our method allows us to relax the exclusion restrictions used in earlier work, while delivering confidence regions that are informative.

Keywords: Partially Linear Model; Data combination; Partial Identification; Intergenerational Mobility.

*We thank the Editor, Francesca Molinari, three anonymous referees, Federico Bugni, Nathael Gozlan, Jinyong Hahn, Jim Heckman, Matt Masten, David Pacini, Adam Rosen, Andres Santos, Jörg Stoye, Martin Weidner and conference and seminar participants at Duke, Oxford, Tilburg, UCLA, the 2021 European Winter Meeting of the Econometric Society and the 2021 Bristol Econometric Study Group for useful comments and suggestions. We also thank Hongchang Guo, Zhangchi Ma and Frank Yan for capable research assistance.

†CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr. Xavier D'Haultfoeuille thanks the hospitality of PSE where part of this research was conducted.

‡Nuffield College and the University of Oxford, christophe.gaillac@economics.ox.ac.uk.

§Duke University, NBER and IZA, arnaud.maurel@duke.edu. Arnaud Maurel thanks the hospitality of the University of Pennsylvania where part of this research was conducted.

1 Introduction

In this paper, we derive partial identification and inference results for a partially linear model, in a context where the outcome of interest and some of the covariates are observed in two different datasets that cannot be merged. Relevant situations include cases where the researcher is interested in the effect of a particular variable that is not observed jointly with the outcome variable, as well as cases where the outcome and covariates of interest are jointly observed but some of the potential confounders are observed in a different dataset.

Our analysis focuses on a partially linear model of the following form:

$$E(Y|X) = f(X_c) + X'_{nc}\beta_0, \quad X = (X_{nc}, X_c), \tag{1}$$

in a data combination environment where F_{Y,X_c} and F_{X_{nc},X_c} are supposed to be identified, but the joint distribution $F_{Y,X}$ is not. The variable X_c is thus common to both datasets, whereas the variable $X_{nc} \in \mathbb{R}^p$ is only observed in one of the two datasets. In this setup, $\beta_0 = (\beta_{01}, ..., \beta_{0p})'$ is generally not point-identified, and as a result we focus on the identified set of either β_0 or β_{0k} for some $k \in \{1, ..., p\}$; the identified set of f can then be deduced from that of β_0.

We first derive a tractable characterization of the identified set of β_0. Unlike many other models considered in the partial identification literature, our setup does not deliver a tractable characterization of the identified set through the support function. However, using Strassen's theorem (Strassen, 1965), a recent result in optimal transport by Backhoff-Veraguas et al. (2019), and a convenient characterization of second-order stochastic dominance, we show that this set is convex, compact, includes the origin and can be simply constructed from its radial function.[1] The identified set of β_{0k}, then, can also be computed at low computational cost by solving an unconstrained convex minimization problem.

The characterization of the identified set also implies that point identification may be achieved if $\beta_0 = 0$, or under a restriction on the unobserved term $Y - f(X_c) - X'_{nc}\beta_0$. While the latter condition is not directly testable, we show how to assess its

[1] The radial function S of a closed, compact convex set $\mathcal{C}$ including the origin is defined, for any q on the unit sphere, by $S(q) = \max_{\lambda q \in \mathcal{C}} \lambda$.

plausibility when one has access to a validation sample in which the outcome and covariates are jointly observed.

In the partially identified case, the identification region may be reduced by adding restrictions on $f(\cdot)$. The two-sample two-stage least squares estimator (TSTSLS) relies on the assumption $f(X_c) = X'_{c,i}\gamma_0$ for some γ_0 and $X_c = (X'_{c,e}, X'_{c,i})'$. In this context, $X_{c,e}$ (resp. $X_{c,i}$) corresponds to the excluded (resp. included) instruments. This is a leading example that results in point identification. But the exclusion restriction that $E(Y|X)$ does not depend on $X_{c,e}$ may not be credible. We show that alternative restrictions, such as imposing a lower bound on the R^2 of the "long regression" of Y on X_{nc} and X_c (in a similar spirit as Oster, 2019) or shape restrictions such as monotonicity or convexity of f, may in practice dramatically reduce the identified set, and allow to, e.g., identify the sign of β_{0k}.

Our identification result is constructive, and readily leads to a simple, plug-in estimator of the identified sets for β_0 or β_{0k}. A difficulty arises, however, as the estimator of the radial function is generally not asymptotically normal. To construct asymptotically valid confidence regions on β_0 or confidence intervals on β_{0k}, we propose to use subsampling (Politis et al., 1999).

Our method is based on a specific characterization of the identified set, and one may wonder whether alternative characterizations would be more convenient. In particular, the identified set can also be expressed through an infinite collection of moment inequalities. Therefore, general approaches for such problems such as that developed by Andrews and Shi (2017) could in principle be used instead. We show through simulations the key computational advantage of relying on the method we propose. With a univariate X_{nc}, confidence regions are typically computed in seconds, whereas they take up to 30 seconds with a bivariate X_{nc}. Compared to the method of Andrews and Shi (2017), this corresponds to a dramatic reduction by a factor of more than 1,000 in computational time.

We apply our method to study intergenerational income mobility over the period 1850 to 1930 in the United States, revisiting the analysis of Olivetti and Paserman (2015). In this context where the main variable and outcome of interest are observed in two different datasets that cannot be linked, we show that the confidence sets obtained using our method are quite informative in practice, while allowing us to

relax the exclusion restrictions underlying the TSTSLS approach used in Olivetti and Paserman (2015). In the appendix, we consider another application where a key control variable is observed in a separate database. When incorporating sign constraints, our bounds are again very informative.

Related literatures

The method we develop in this paper can be used in a broad set of data combination environments. Two such contexts have attracted much attention in the empirical literature.

One can use our method to conduct inference on the relationship between a particular covariate and an outcome variable, in situations where both variables are not jointly observed. A large literature on intergenerational income mobility often faces the unavailability of linked income data across generations and relies on exclusion restrictions, as in the application we revisit (see Santavirta and Stuhler, 2022, for a recent survey). More generally, this type of data combination environment frequently arises in various subfields of empirical microeconomics, including in education and returns to skill estimation (Rothstein and Wozny, 2013; Piatek and Pinger, 2016; Garcia et al., 2020; Hanushek et al., 2021), health (Manski, 2018; Robbins et al., 2022) and labor (Athey et al., 2020). A leading example that has attracted much interest in the literature is one where the researcher seeks to combine experimental data with another observational dataset, in particular situations where data on long-term outcomes is not available in the experimental data.

Our approach can also be used to conduct inference on the causal effect of a variable of interest, in a setup where some of the confounders are observed in an auxiliary dataset. As such, our paper expands the range of data environments in which unconfoundedness is a credible assumption, complementing a literature that focuses on evaluating its reasonableness in the absence of data combination (see, e.g., Altonji et al., 2005; Oster, 2019; Diegert et al., 2022).

From a methodological standpoint, our paper is connected to the seminal article of Cross and Manski (2002) and subsequent work by Molinari and Peski (2006). They consider the issue of identifying the "long regression", in our context $E(Y|X_c, X_{nc})$,

in the same data combination set-up as here. Importantly though, these two papers focus on deriving the identification region for $E(Y|X_c, X_{nc})$, but do not address the issue of inference. They also consider a setup where the covariates X_{nc} have a discrete distribution with finite support, while we allow X_{nc} to be continuously distributed. On the other hand their setup is entirely nonparametric, whereas we focus on a model that is linear in the covariates X_{nc} and without interaction terms with X_c. The linearity assumption plays an important role in our ability to derive a tractable inference method. The absence of interaction further implies that in our set-up, and in contrast with these two papers, the identified set shrinks as one considers different values of X_c.

Our paper is also related to Pacini (2019) and Hwang (2022). Both papers construct bounds on the best linear predictor of Y on X in a similar data combination framework as here. We show that if one is ready to impose the usual assumption that the model is partially linear, large identification gains may be achieved, possibly up to point identification. Hwang (2022) also considers a set-up where some of the X's are only observed with Y but not with X_{nc}, a case we do not study in this paper.

More generally speaking, our paper relates to the broader literature on data combination problems in econometrics and statistics. We refer the reader to Ridder and Moffitt (2007) for a survey of this literature and to Fan et al. (2014), Fan et al. (2016), Buchinsky et al. (2022), and Athey et al. (2020) for recent contributions. Contrary to ours, most of these papers impose restrictions that entail point identification.

Within the data combination literature, our paper is technically closest to D'Haultfoeuille et al. (2021). Though that paper considered the entirely different context of rational expectation testing, we also relied therein on Strassen's theorem to obtain a characterization of the null hypothesis of rational expectations. Importantly, we extend here our previous main result in a highly non-trivial way, by relying in particular on recent results from Backhoff-Veraguas et al. (2019) to handle the case where X_{nc} is multivariate. Also, we previously based our inference on Andrews and Shi (2017). In contrast, a key contribution of our paper lies in the novel and tractable inference method that we derive.

Finally, by developing in this data combination context a feasible inference method that can be implemented at a very limited computational cost, our paper also adds to

the growing set of papers that propose tractable computational methods for partially identified models (see Bontemps and Magnac, 2017 and Molinari, 2020 for recent surveys). In particular, our paper fits into the strand of the literature that uses tools from optimal transport to devise computationally tractable identification and inference methods for partially identified models (Galichon and Henry, 2011; Galichon, 2016). By characterizing the sharp identified set based on the radial function, a novel approach in the partial identification literature, we show that it is possible to achieve very substantial tractability gains in this context, relative to a more standard characterization in terms of many moment inequalities.

Organization of the paper

The remainder of the paper is organized as follows. In Section 2 we present our main identification results for the two-sample partially linear model described above. Section 3 studies estimation and inference for this model. In Section 4, we apply our method to intergenerational income mobility in the United States. Section 5 concludes. The Appendix of the paper gathers additional results on robustness to measurement errors, identification in models with heterogeneous effects of X_{nc} on Y, and a test for point-identification. It also presents our second application to the black-white wage gap in the United States. Monte Carlo simulation results, additional material on the application, and the proofs are collected in the online Appendix. Some complements of the proofs appear in supplementary material available in our working paper version (see D'Haultfœuille et al., 2023). Finally, our inference method can be implemented using our companion R package, `RegCombin`, available at CRAN.R-project.org/package=RegCombin.

2 Identification

Before presenting our main identification results, we introduce some notation that will be used throughout the paper. We let $\|\cdot\|$, 0_p and $\mathcal{S}_p$ denote respectively the usual Euclidean norm in $\mathbb{R}^p$, the vector 0 and the unit sphere in $\mathbb{R}^p$; we may omit the index p in the absence of ambiguity. For any cumulative distribution function (cdf) F defined on $\mathbb{R}$, we let $F^{-1}(t) = \inf\{x : F(x) \geq t\}$ denote its generalized inverse

and $\overline{F} = 1 - F$ be the corresponding survival function. For any random variable A, we let $\text{Supp}(A)$ be its support, F_A denote its cdf. and $V(A)$ its variance, if defined. We also let $\succ_{\text{cv}}$ denote the convex ordering, namely, for two random variables A and B with $E[|A|] < \infty$ and $E[|B|] < \infty$, $A \succ_{\text{cv}} B$ if $E[\phi(A)] \geq E[\phi(B)]$ for all convex functions ϕ.[2] We write $A \not\succ_{\text{cv}} B$ when $A \succ_{\text{cv}} B$ does not hold. Finally, for any sets C and C', we denote by ∂C the boundary of C and by $d_H(C, C')$ the Hausdorff distance between C and C', defined by

$$d_H(C, C') = \max\left(\sup_{c' \in C'} \inf_{c \in C} ||c - c'||, \sup_{c \in C} \inf_{c' \in C'} ||c - c'||\right).$$

2.1 Identification without common regressors

2.1.1 A tractable characterization of the identified set

We first consider a linear model and derive the sharp identified set of β_0 in the absence of common regressors observed in both datasets. We suppose that we observe from two samples that can not be merged the distributions of the outcome, F_Y, and covariates, F_X. We maintain the following assumption:

Assumption 1. *We have $E(Y^2) < \infty$, $E(\|X\|^2) < \infty$, $V(Y) > 0$ and $V(X)$ is non-singular. Moreover, $E(Y|X) = \alpha_0 + X'\beta_0$ for some $(\alpha_0, \beta_0) \in \mathbb{R} \times \mathbb{R}^p$.*

We focus hereafter on the identified set $\mathcal{B}$ of β_0. Since $\mathcal{B}$ is the set of all vectors in $\mathbb{R}^p$ that are compatible with the model and the marginal distributions of Y and X, we have

$$\mathcal{B} = \left\{\beta \in \mathbb{R}^p : \exists \text{ r.v. } (\widetilde{X}, \widetilde{Y}) : \ E(\widetilde{Y}_0|\widetilde{X}_0) = \widetilde{X}_0'\beta, \ \widetilde{X} \overset{d}{=} X, \ \widetilde{Y} \overset{d}{=} Y\right\}, \tag{2}$$

where, for any random variable A with $E[|A|] < \infty$, we let $A_0 = A - E(A)$ and we have used that $E(Y|X) = \alpha_0 + X'\beta_0$ for some α_0 is equivalent to $E(Y_0|X_0) = X_0'\beta_0$. Now, our goal is to express $\mathcal{B}$ to make it amenable to (simple) estimation. To this end, we define, for any $\alpha \in (0, 1)$, F and G cdfs with expectation 0, the following functions:

$$R(\alpha, F, G) = \frac{\int_\alpha^1 F^{-1}(t)dt}{\int_\alpha^1 G^{-1}(t)dt}, \tag{3}$$

[2]Even though we may have $E[|\phi(A)|] = \infty$, $E[\phi(A)]$ is always well-defined because $E[\max(0, -\phi(A))] < \infty$, since there exists a, b such that for all x, $\phi(x) \geq a + bx$.

$$S(F,G) = \inf_{\alpha\in(0,1)} R(\alpha,F,G).$$

These two functions play an important role in our analysis. Remark that, since F and G are cdfs of mean zero distributions, $\int_\alpha^1 F^{-1}(t)dt$ and $\int_\alpha^1 G^{-1}(t)dt$ are both positive, so that the ratio of superquantiles $R(\alpha,F,G)$ is well-defined, with $R(\alpha,F,G)>0$ and $S(F,G)\geq 0$. Theorem 1 is our main identification result.

Theorem 1. *Suppose that Assumption 1 holds. Then*

$$\mathcal{B} = \left\{\lambda q :\ q\in\mathcal{S},\ 0\leq\lambda\leq S(F_{Y_0},F_{X_0'q})\right\}. \tag{4}$$

$\mathcal{B}$ includes 0_p and is a convex, compact subset of $\mathcal{B}^V=\{\beta\in\mathbb{R}^p :\ \beta'V(X)\beta\leq V(Y)\}$.

We now give a sketch of the proof of (4). Let $\mathcal{B}'$ denote the set on the right-hand side of (4). First, by definition of $S(F_{Y_0},F_{X_0'q})$,

$$\mathcal{B}' = \left\{\beta\in\mathbb{R}^p : \forall\alpha\in(0,1),\ \int_\alpha^1 F^{-1}_{X_0'\beta}(t)dt\leq\int_\alpha^1 F^{-1}_{Y_0}(t)dt\right\}.$$

This, in turn, is equivalent to $F_{X_0'\beta}$ dominating F_{Y_0} at the second order (see, e.g. De la Cal and Cárcamo, 2006), implying that

$$\mathcal{B}' = \left\{\beta\in\mathbb{R}^p : Y_0\succ_{\text{cv}} X_0'\beta\right\}.$$

The inclusion $\mathcal{B}\subset\mathcal{B}'$ then follows essentially from Jensen's inequality. As a side remark, note that we can also express $\mathcal{B}'$ through infinitely many moment inequality restrictions:

$$\mathcal{B}' = \left\{\beta\in\mathbb{R}^p :\ E\left[\max(0,Y_0-t)\right]\geq E\left[\max(0,X_0'\beta-t)\right]\ \forall t\in\mathbb{R}\right\}. \tag{5}$$

This equality directly follows from Fubini-Tonelli, applied to the standard characterization of the second-order stochastic dominance condition, namely $\int_{-\infty}^y F_{Y_0}(t)dt\geq\int_{-\infty}^y F_{X_0'\beta}(t)dt\ \forall y\in\mathbb{R}$. We return to this alternative characterization of the identified set in Subsections C.1 and C.4 of the online appendix, where we document the computational advantages of using our characterization instead.

The inclusion $\mathcal{B}'\subset\mathcal{B}$ is more intricate to prove. First, if $Y_0\succ_{\text{cv}} X_0'\beta$, we have, by Strassen's theorem (Theorem 8 in Strassen, 1965),

$$\inf_{(\widetilde{Y},\widetilde{X}^\beta):\widetilde{Y}\overset{d}{=}Y,\ \widetilde{X}^\beta\overset{d}{=}X'\beta} E\left[\left|\widetilde{X}_0^\beta - E[\widetilde{Y}_0|\widetilde{X}_0^\beta]\right|\right] = 0. \tag{6}$$

This result was already used in D'Haultfoeuille et al. (2021) to characterize the restrictions on F_Y and F_ψ entailed by the rational expectation hypothesis $E(Y|\psi) = \psi$, where ψ denotes the subjective expectations on an outcome Y. Importantly though, when X is multivariate, (6) is not sufficient to conclude that $\mathcal{B}' \subset \mathcal{B}$, as the σ-algebras generated by X and $X'\beta$ are not equal in general. Nonetheless, we prove,[3] using a recent result in optimal transport (Theorem 1.3 in Backhoff-Veraguas et al., 2019), that

$$\inf_{(\widetilde{Y},\widetilde{X}):\widetilde{Y}\overset{d}{=}Y,\,\widetilde{X}\overset{d}{=}X} E\left[\left|\widetilde{X}_0'\beta - E[\widetilde{Y}_0|\widetilde{X}_0]\right|\right] \leq \inf_{(\widetilde{Y},\widetilde{X}^\beta):\widetilde{Y}\overset{d}{=}Y,\,\widetilde{X}^\beta\overset{d}{=}X'\beta} E\left[\left|\widetilde{X}_0^\beta - E[\widetilde{Y}_0|\widetilde{X}_0^\beta]\right|\right]. \quad (7)$$

Together, (6), (7), and the existence of a minimizer on the left-hand side of (7) (Theorem 1.2 in Backhoff-Veraguas et al., 2019), imply that we can find random variables $\widetilde{Y}$ and $\widetilde{X}$ such that $E[\widetilde{Y}_0|\widetilde{X}_0] = \widetilde{X}_0'\beta$, $\widetilde{Y} \overset{d}{=} Y$ and $\widetilde{X} \overset{d}{=} X$. Thus, $\beta \in \mathcal{B}$.

Turning to the second part of the theorem, $0_p \in \mathcal{B}$ follows by noting that one can always rationalize, from the sole knowledge of their marginal distributions, that X and Y are independent. That $\mathcal{B} \subset \mathcal{B}^V$ comes from the inclusion $\mathcal{B} \subset \mathcal{B}'$, combined with the fact that $Y_0 \succ_{\text{cv}} X_0'\beta$ implies $V(Y) \geq V(X'\beta)$. Hence, $\mathcal{B}$ is included in a bounded ellipsoid. The equality $\mathcal{B} = \mathcal{B}^V$ occurs for instance when Y and X are normally distributed. Otherwise, $\mathcal{B}$ may be substantially smaller than $\mathcal{B}^V$, as we illustrate below. In such cases, $\mathcal{B}^V$ remains a natural benchmark as it is very simple to characterize using $V(Y)$ and $V(X)$ only, and straightforward to estimate.

Remark 2.1. *Using the exact same reasoning as above, one can prove that without any linear restriction on the conditional expectation, the identified set for* $E[Y_0|X_0]$ *is* $\{g : Y_0 \succ_{cv} g(X_0)\}$. *Similarly, if we only impose that* $m(x) := E[Y_0|X_0 = x]$ *belongs to a linear space* $\mathcal{Z}$ *of functions, then the identified set for* m *is* $\{\lambda q : q \in \mathcal{Z} : E[|q(X_0)|] = 1, E[q(X_0)] = 0,\, 0 \leq \lambda \leq S(F_{Y_0}, F_{q(X_0)})\}$.[4]

Radial vs. support function characterization of the identified set. A key takeaway from Equation (4) is that the identified set admits a very simple expression as a function of S, which is the inverse of the Minkowski gauge function of $\mathcal{B}$ (see, e.g.,

[3] We thank Nathael Gozlan for his help in obtaining (7).

[4] We thank a referee for pointing out this extension.

Definition 1.2.4 p.137 and Proposition 3.2.4 p.157 Hiriart-Urruty and Lemaréchal, 2012), also known as the radial function of $\mathcal{B}$. This function differs from the support function σ of $\mathcal{B}$, defined by $\sigma(q, F_{Y_0}, F_{X_0}) = \sup_{b \in \mathcal{B}} q'b$. The difference between these two functions is illustrated in Figure 1.

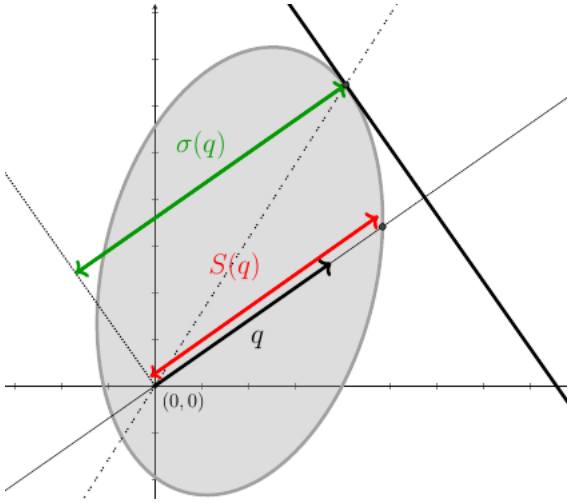


Figure 1: Two characterizations of a closed convex set including the origin, either through its support function σ (green), or through the radial function S (red).

The partial identification literature has largely relied on support functions, as these are powerful tools that uniquely characterize their convex sets. But the radial function also uniquely characterizes convex sets if, as is the case here, these sets include the origin. Importantly, this approach allows us to characterize the sharp identified set by minimizing a simple function over the interval $(0, 1)$. In contrast, the support function approach will generally be significantly less tractable in our context as it would require solving a high-dimensional constrained optimization problem. Namely, using the characterization of the identified set given in Equation (5) above, the support function can be obtained by solving the following program:

$$\sigma(q, F_{Y_0}, F_{X_0}) = \sup_{b \in \mathbb{R}^p} q'b \quad \text{s.t.} \inf_{t \in \mathbb{R}} E\left[\max(0, Y_0 - t)\right] - E\left[\max(0, X_0'b - t)\right] \geq 0, \quad (8)$$

where the constraint itself involves an optimization problem. Simulation results indicate that using the radial function rather than the support function approach does result in very large computational gains, see Online Appendix C.4 for details on this.

Partial identification of subcomponents of β_0. The support function still plays a key role in our context when one is interested in a component of $\beta_0 = (\beta_{0,1}, ..., \beta_{0,p})'$, say $\beta_{0,k}$. The following result shows that we can actually recover this function at a low computational cost once S is known. Hereafter, we let e_k denotes the k-th element of the canonical basis in $\mathbb{R}^p$ and use the convention $1/0 = \infty$ and $1/\infty = 0$.

Corollary 1. *Suppose that Assumption 1 holds. Then, the identified set $\mathcal{B}_k$ of $\beta_{0,k}$ satisfies $\mathcal{B}_k = [-\sigma(-e_k, F_{Y_0}, F_{X_0}), \sigma(e_k, F_{Y_0}, F_{X_0})]$. Moreover,*

$$\sigma(e_k, F_{Y_0}, F_{X_0}) = \frac{1}{\inf_{q \in \mathbb{R}^p : q_k = 1} 1/S(F_{Y_0}, F_{X_0'q})}. \tag{9}$$

The same holds with $\sigma(-e_k, F_{Y_0}, F_{X_0})$, after replacing $q_k = 1$ by $q_k = -1$.

We use the expression (9) of the support function, rather than the simpler expression $\sigma(e_k, F_{Y_0}, F_{X_0}) = \sup_{q \in \mathbb{R}^p : q_k = 1} S(F_{Y_0}, F_{X_0'q})$, because $q \mapsto 1/S(F_{Y_0}, F_{X_0'q})$ is convex (see the proof of Proposition 6, which also applies when $\varepsilon = 0$), whereas $q \mapsto S(F_{Y_0}, F_{X_0'q})$ may not be concave. It follows that one can recover the support function σ, and in turn the sharp bounds on $\beta_{0,k}$, by simply minimizing a convex function over $\mathbb{R}^{p-1}$.

2.1.2 Point identification

In some cases, our approach yields point identification of the parameters of interest, or subcomponents of it. Proposition 1 below presents two such cases under which the identified sets $\mathcal{B}$ and $\mathcal{B}_1$, respectively, boil down to a singleton.

Proposition 1. *Suppose that Assumption 1 holds and let ϕ be a convex function such that $E[\phi(Y)] < \infty$. Then:*

1. *If for all $\beta \neq 0_p$, $E[\phi(X'\beta)] = \infty$, then $\mathcal{B} = \{\beta_0\} = \{0_p\}$.*
2. *If $E[\phi(X_1\beta_1)] = \infty$ for all $\beta_1 \neq 0$ and $E[\phi(X_{-1}'\beta_{-1})] < \infty$ for all $\beta_{-1} \in \mathbb{R}^{p-1}$, then $\mathcal{B}_1 = \{\beta_{0,1}\} = \{0\}$.*

Recall from our main identification result above that the identified set $\mathcal{B}$ always includes the origin. The first point of Proposition 1 further establishes point identification of $\beta_0 = 0_p$ when, basically, Y has lighter tails than any linear index of X. The second point is similar but focuses on a subcomponent instead: if Y and $X_{-1}'\beta_{-1}$ have

lighter tails than X_1, then $\beta_{0,1} = 0$ is point identified. As an example of function ϕ for which Proposition 1 holds, one might consider for instance $\phi(x) = |x|^a$ for some $a > 2$ (in which case $X'\beta$ or X_1 have heavy tails), or $\phi(x) = \exp(a|x|^b)$ for some $a, b > 0$ (in which case $X'\beta$ or X_1 have exponential tails).

To illustrate Point 1 of Proposition 1, suppose that $p = 1$, X follows a Laplace distribution (with density $\exp(-|x|)/2$ on $\mathbb{R}$) and $Y \sim \mathcal{N}(0, 1)$. Then, by using $\phi(x) = \exp(|x|^{3/2})$, it follows from Point 1 of Proposition 1 that $\beta_0 = 0$ is point identified in this case. On the other hand, the variance restrictions only set identify β_0, with an identified set given by $\mathcal{B}^V = [-1/\sqrt{2}, 1/\sqrt{2}] \simeq [-0.707, 0.707]$. This example illustrates the (in this case point-) identifying power of higher-order moments of the distributions of X and Y.

2.2 Identification with common regressors

We now turn to the frequent situation where some regressors are observed in both datasets. Namely, suppose we observe regressors X_c that are common to both datasets, and assume that the partially linear model (1) holds:

$$E(Y|X) = f(X_c) + X'_{nc}\beta_0, \quad X = (X_{nc}, X_c),$$

The key here is to note, following Robinson (1988), that this case is equivalent to the previous setup without common regressors once we compute the following residuals, for all x in the support of X_c:

$$X^x = X_{nc} - E(X_{nc}|X_c = x),$$
$$Y^x = Y - E(Y|X_c = x).$$

It directly follows that β_0 satisfies $E(Y^x|X^x) = X^{x\prime}\beta_0$, which allows us to use the characterization of the identified set without common regressors obtained in Section 2.1.

Let $\mathcal{B}^c$ and $\mathcal{F}$ denote the identified sets of β_0 and f, respectively. We have the following characterization of $\mathcal{B}^c$ and $\mathcal{F}$:

Proposition 2. *Suppose that $E(Y^2) < \infty$, for all $x \in Supp(X_c)$, $E(X^x X^{x\prime}|X_c = x)$ is nonsingular and (1) holds. Then:*

$$\mathcal{B}^c = \left\{\lambda q : q \in \mathcal{S},\ 0 \leq \lambda \leq \overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c})\right\},$$

$$\mathcal{F} = \{x \mapsto E(Y|X_c = x) - E(X_{nc}|X_c = x)'\beta : \ \beta \in \mathcal{B}^c\},$$

where $\overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c}) = \inf_{x \in Supp(X_c)} S(F_{Y^x|X_c=x}, F_{X^{x\prime}q|X_c=x})$. $\mathcal{B}^c$ *includes* 0_p, *is compact and convex.*

It is possible to extend (1) by including interaction terms. Notably, such specification allows for heterogeneous effects of X_{nc} on Y, which can be important in practice. We consider this extension in Appendix A.2. Another interesting extension corresponds to cases where $E(Y|X) = f(X_c) + X'_{nc}\beta_0 + X'_a\delta_0$ and we observe in a first dataset (Y, X_a, X_c) and in a second dataset, (X_c, X_{nc}). This setup leads to qualitatively different results. For instance, if there is no common regressors and (Y, X_a) and X_{nc} are Gaussian, one can show that the sharp identified set of (β_0, δ_0) is not convex and does not include 0_{p+r} (with r the dimension of X_a). We refer the reader to Hwang (2022) for outer bounds on the best linear predictor in this setup and leave its study for future research.

2.3 Identifying power of additional restrictions

We now consider additional restrictions that may reduce the identified set, in some cases resulting in point identification of the parameters of interest.

2.3.1 Lower bound on the R^2 of the long regression

A first way to reduce the identified set is to use a lower bound on the predictive power of X_{nc} and X_c with respect to Y. To formalize this idea, we assume that R^2_ℓ, the coefficient of determination of the "long" regression of Y on X_{nc} and X_c is higher than a certain threshold. This threshold may be absolute (e.g., 0.1) or relative to $R^2_s := V(E(Y|X_c))/V(Y)$, the R^2 of the "short" regression of Y on X_c, which is directly identified from the data. This is in the same spirit as Oster (2019), who suggests fixing R^2_ℓ/R^2_s to 1.3. Note that

$$f(X_c) + X'_{nc}\beta = E(Y|X_c) + (X_{nc} - E(X_{nc}|X_c))'\beta,$$

and the two components on the right-hand side are uncorrelated. Thus,

$$R^2_\ell = \frac{V(E(Y|X_c)) + \beta' E(V(X_{nc}|X_c))\beta}{V(Y)} = R^2_s + \frac{\beta' E(V(X_{nc}|X_c))\beta}{V(Y)}.$$

Then, if one imposes a lower bound $\underline{R}^2$ on R^2_ℓ such that $\underline{R}^2 \geq R^2_s$, the identified set on β becomes

$$\left\{ \lambda q : q \in \mathcal{S},\ \left(\frac{(\underline{R}^2 - R^2_s)V(Y)}{q' E(V(X_{nc}|X_c))q} \right)^{1/2} \leq \lambda \leq \overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c}) \right\},$$

provided that $E(V(X_{nc}|X_c))$ is nonsingular. This restriction has three key attractive features. First, one can in practice motivate this restriction based on a "validation sample", namely a subset of the population or another population (e.g., a different country than that under investigation), for which we identify the joint distribution of the outcome and covariates, and thus the R^2 of the "long" regression. Second, imposing a lower bound such that $\underline{R}^2 > R^2_s$ allows one to exclude 0_p from the identified set. Third, the identified set still admits a very simple expression.

2.3.2 Shape restrictions

Another way to narrow the identified set $\mathcal{B}^c$ with common regressors is to impose some constraints on $f(\cdot)$. Shape restrictions such as monotonicity or convexity often follow from economic theory; see Matzkin (1994) and Chetverikov et al. (2018) for econometric reviews, and Tripathi (2000) and Abrevaya and Jiang (2005) for their use and testability with partially linear models. We characterize here the identified set when we impose such restrictions on f.

We model these restrictions by $[Rf](r) \geq \underline{c}(r)$ for all $r \in \mathcal{R}$, with R a known linear operator, $\underline{c}$ a known, real function and $\mathcal{R}$ the domain of $[Rf]$ and $\underline{c}$. For instance, if X_c is discrete such that $\text{Supp}(X_c) = \{x_{c,1}, ..., x_{c,K}\} \subset \mathbb{R}$, with $K > 1$ and $x_{c,1} < ... < x_{c,K}$, considering $[Rf](r) = f(x_{c,r+1}) - f(x_{c,r})$ for $r \in \mathcal{R} = \{1, ..., K-1\}$ (resp. $[Rf](r) = (f(x_{c,r+2}) - f(x_{c,r+1}))/(x_{c,r+2} - x_{c,r+1}) - (f(x_{c,r+1}) - f(x_{c,r}))/(x_{c,r+1} - x_{c,r})$ for $r \in \mathcal{R} = \{1, ..., K-2\}$ with $K > 2$) and $\underline{c}(r) = 0$ corresponds to imposing that f is non-decreasing (resp. convex). When X_c is continuous, the same two constraints can be imposed by considering $[Rf](r) = f'(r)$ and $[Rf](r) = f''(r)$, with $\mathcal{R} = \text{Supp}(X_c)$.

This framework also accommodates restrictions on the magnitude of the effect of X_c on Y. Namely, suppose for simplicity that X_c is binary and consider $[Rf](1) = -[Rf](2) = f(x_{c,2}) - f(x_{c,1})$ with $\mathcal{R} = \{1, 2\}$ and $\underline{c}(1) = \underline{c}(2) = \underline{c} \geq 0$. The extreme case $\underline{c} = 0$ corresponds to X_c having no effect on Y, as in the two-sample

two-stage least squares strategy (see the next subsection for a related, more general point identification result in this context). More generally, this corresponds to the constraint that the magnitude of the effect of X_c is bounded by the cutoff $\underline{c}$, $|f(x_{c,2}) - f(x_{c,1})| \leq \underline{c}$.[5] By increasing $\underline{c}$, one can therefore study how the identified set varies when relaxing the exclusion restriction, in a similar spirit to, e.g., Masten and Poirier (2018).

Hereafter, we denote by $m_Y(\cdot) = E[Y|X_c = \cdot]$, $m_{X_{nc}}(\cdot) = E[X_{nc}|X_c = \cdot]$ and

$$\underline{S}^c(m_Y, m_{X_{nc}}, q) = \sup_{\substack{r \in \mathcal{R}: \\ [Rm'_{X_{nc}} q](r) \leq 0}} \lim_{u \downarrow 0} \frac{[Rm_Y - \underline{c}](r) + u}{[Rm'_{X_{nc}} q](r) - u^2},$$

$$\overline{S}^c(m_Y, m_{X_{nc}}, q) = \inf_{\substack{r \in \mathcal{R}: \\ [Rm'_{X_{nc}} q](r) \geq 0}} \lim_{u \downarrow 0} \frac{[Rm_Y - \underline{c}](r) + u}{[Rm'_{X_{nc}} q](r) + u^2},$$

where we let $\sup \emptyset = -\inf \emptyset = -\infty$ and we note that the two functions above may be infinite. We introduce limits to deal with the cases where $[Rm'_{X_{nc}} q](r) = 0$. Proposition 3 characterizes the identified sets of β_0 and f under such shape restrictions.

Proposition 3. *Suppose that the conditions of Proposition 2 hold and* $[Rf](r) \geq \underline{c}(r)$ *for all* $r \in \mathcal{R}$. *Then, the identified sets* $\mathcal{B}^{con}$ *and* $\mathcal{F}^{con}$ *of* β_0 *and* f *satisfy*

$$\mathcal{B}^{con} = \left\{ \lambda q : q \in \mathcal{S}^+,\ \underline{S}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) \leq \lambda \leq \overline{S}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) \right\},$$

$$\mathcal{F}^{con} = \left\{ x \mapsto E(Y|X_c = x) - E(X_{nc}|X_c = x)'\beta :\ \beta \in \mathcal{B}^{con} \right\},$$

where $\mathcal{S}^+ = \mathcal{S} \cap \{(x_1, ..., x_p) \in \mathbb{R}^p : x_1 \geq 0\}$ *and*

$$\underline{S}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) = \max\left(-\overline{S}(F_{Y,X_c}, F_{-X'_{nc}q,X_c}), \underline{S}^c(m_Y, m_{X_nc}, q) \right),$$

$$\overline{S}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) = \min\left(\overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c}), \overline{S}^c(m_Y, m_{X_nc}, q) \right).$$

$\mathcal{B}^{con}$ *is compact, convex but does not include* 0_p *if for some* $r \in \mathcal{R}$, $[Rm_Y - \underline{c}](r) < 0$.

In contrast to our baseline identification results in the absence of additional restrictions, the resulting identified set may exclude the origin. This illustrates the practical

[5] If X_c has $K > 2$ points of support, the same idea can be generalized by imposing restrictions on $|f(x_{c,k}) - f(x_{c,j})|$ for specific pairs $(j, k) \in \{1, ..., K\}^2$, $j \neq k$.

importance of imposing these types of shape restrictions in contexts where these are likely to hold. Suppose for instance that $p = 1$, X_c is binary ($\text{Supp}(X_c) = \{0, 1\}$), $\mathcal{R} = \{1\}$ and $[Rf](1) = f(1) - f(0)$, namely we impose that f is non-decreasing. If $f(1) - f(0) < (m_{X_{nc}}(0) - m_{X_{nc}}(1))\beta_0$, then $m_Y(1) < m_Y(0)$. As a result, $0 \notin \mathcal{B}^{\text{con}}$. The condition $f(1) - f(0) < (m_{X_{nc}}(0) - m_{X_{nc}}(1))\beta_0$ holds for instance if $m_{X_{nc}}$ is decreasing and β_0 is positive and large enough.

Remark 2.2. *While we focus here on the identifying power of each type of restrictions considered separately, researchers may in some contexts want to jointly impose several of these restrictions and consider the intersection of the associated identified sets. In the particular cases of the shape restrictions and the restrictions on the R^2 considered above, the identified sets share the same structure. Thus, the identified set resulting from both types of constraints can be simply computed by replacing the lower bound on λ by the maximum of the lower bounds of the initial sets, and proceeding symmetrically for the upper bound.*

2.3.3 Functional form restrictions involving common regressors

One may alternatively be willing to impose functional form restrictions on f. The following proposition shows that this may yield point identification.

Proposition 4. *Suppose that $E(Y^2) < \infty$, $E[\|X\|^2] < \infty$ and f belongs to a vector space $\mathcal{G}$. Then, if for all $\gamma \neq 0$, $m'_{X_{nc}}\gamma \notin \mathcal{G}$, β_0 and f are point identified.*

This proposition encompasses several popular restrictions. We consider in particular three such restrictions, for which the key point-identifying condition $m'_{X_{nc}}\gamma \notin \mathcal{G}$ has a simple interpretation:

1. $f(X_c) = f_1(X_{i,c})$, with $X_c = (X'_{i,c}, X'_{e,c})'$. This restriction, which is implicit in, and central to the two-sample two-stage least squares strategy, states that conditional on $(X_{nc}, X_{i,c})$, Y is mean-independent of $X_{e,c}$. In such a case, $m'_{X_{nc}}\gamma \notin \mathcal{G}$ for all $\gamma \neq 0$ basically means that $m_{X_{nc}}(X_c)$ varies with $X_{e,c}$. To see this, consider the simple case where $m_{X_{nc}}(X_c) = m_i(X_{i,c}) + \Pi X_{e,c}$, for some function m_i and a $p \times q$ matrix Π. Then, $m'_{X_{nc}}\gamma \notin \mathcal{G}$ is equivalent to Π having rank p, which is the usual rank condition in linear instrumental variable models.

2. $f(X_c) = X_c'\gamma_0$. Under this linearity restriction on $f(\cdot)$, $m_{X_{nc}}'\gamma \notin \mathcal{G}$ for all $\gamma \neq 0$ basically means that $m_{X_{nc}}(X_c)$ is nonlinear in X_c (the two notions are actually equivalent if $X_{nc} \in \mathbb{R}$). Note that this point identification result fully relies on the linearity of $f(\cdot)$ combined with the nonlinearity of $E(X_{nc}|X_c)$, and is thus akin to, e.g., the identification of sample selection models without instruments exploiting the nonlinearity of the inverse Mill's ratio. Also, this result does not apply when X_c is binary, since in this case $m_{X_{nc}}(X_c)$ is necessarily linear in X_c.

3. $f(X_c) = \sum_{j=1}^{J} f_j(X_{j,c})$, with $X_c = (X_{1,c}, ..., X_{J,c})'$. Under this additivity restriction on $f(\cdot)$, $m_{X_{nc}}'\gamma \notin \mathcal{G}$ for all $\gamma \neq 0$ means that $m_{X_{nc}}(X_c)$ is not additive in X_c. If for instance $X_c = (X_{1,c}, X_{2,c})$ with $X_{1,c}, X_{2,c}$ both binary, $m_{X_{nc}}'\gamma \notin \mathcal{G}$ for all $\gamma \neq 0$ holds if in the regression of X_{nc} on $X_{1,c}, X_{2,c}$ and $X_{1,c} \times X_{2,c}$, the coefficient of $X_{1,c} \times X_{2,c}$ is not zero.

2.3.4 Tail conditions

Finally, if one is ready to impose a relative tail condition between the error term $U := Y_0 - X_0'\beta_0$ and $X_0'\beta_0$, the identified set is considerably reduced. For simplicity, we assume here that there are no common regressors but Proposition 5 readily extends to accomodate such regressors.

Proposition 5. *Suppose that Assumption 1 holds. Then:*

1. *If there exists a convex function ϕ such that $E[\phi(U\lambda)] < E[\phi(X_0'\beta_0\lambda)] = \infty$ for all $\lambda > 1$, the identified set of β_0 is included in $\partial\mathcal{B}$;*

2. *$X \in \mathbb{R}$, $E[\phi(U\lambda)] < E[\phi(X\lambda)] = \infty$ for all $\lambda > 0$ and it is known that $\beta_0 > 0$, β_0 is point identified.*

With $X \in \mathbb{R}$, the condition $E[\phi(U\lambda)] < E[\phi(X\lambda)] = \infty$ for all $\lambda > 0$ holds for instance if $E[|U|^a] < E[|X|^a] = \infty$ for some $a > 2$. More generally, the condition $E[\phi(U\lambda)] < E[\phi(X_0'\beta_0\lambda)] = \infty$ basically imposes that $X_0'\beta_0$ has fatter tails than U. In this sense, this condition is similar to those in Proposition 1 above.

Testability. Note that we cannot test the condition $E[\phi(U\lambda)] < E[\phi(X_0'\beta_0\lambda)] = \infty$ for some convex function ϕ and all $\lambda > 1$, simply because U is not identified. On the other hand, we can assess the plausibility of $\beta_0 \in \partial\mathcal{B}$ using a validation sample, as defined above. Denoting by (Y_v, X_v) the variables corresponding to this validation sample, it becomes possible to test whether the corresponding parameter $\beta_v = V(X_v)^{-1}\text{cov}(X_v, Y_v)$ is at the boundary of the identified set one would get from the sole knowledge of F_{Y_v} and F_{X_v}. Provided that $\beta_v \neq 0$, this condition is indeed equivalent to $\|\beta_v\| = S(F_{Y_{v0}}, F_{X_{v0}'\beta_v/\|\beta_v\|})$ or, in simpler terms,

$$S(F_{Y_{v0}}, F_{X_{v0}'\beta_v}) = 1.$$

We consider a statistical test of this condition in Appendix A.3, and apply it in Section 4 below.

2.4 Numerical illustration

We illustrate the previous results by considering the following model:

$$Y = \gamma_{0,0} + X_c^{1.3}\gamma_{0,1} + X_{nc,1}\beta_{nc,1} + X_{nc,2}\beta_{nc,2} + U,\ U|X \sim \mathcal{N}(0, 9).$$

We set the coefficients as follows: $\gamma_{0,0} = -0.1$, $\gamma_{0,1} = 0.3$, $\beta_{nc,1} = 1$ and $\beta_{nc,2} = 1$. The variables X are transformations of $(N_1, N_2, N_3)'$, which is supposed to follow a multivariate normal distribution with mean 0 and covariance matrix

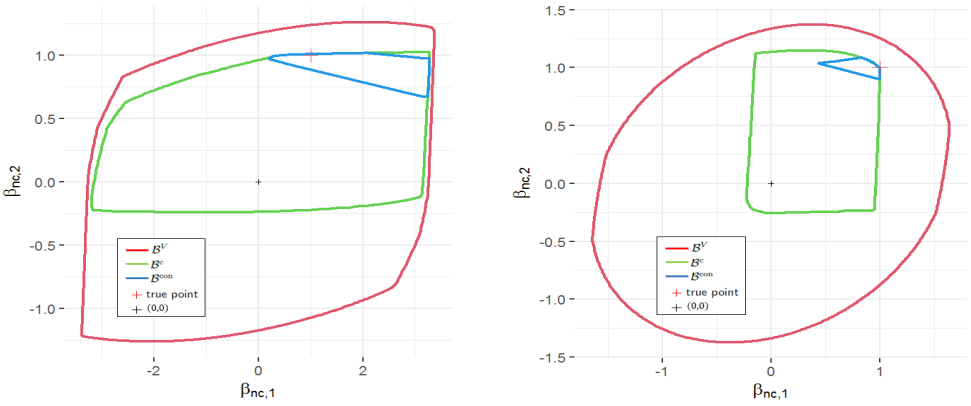
$$\Sigma = \begin{pmatrix} 1 & -0.3 & -0.8 \\ -0.3 & 1 & -0.1 \\ -0.8 & -0.1 & 1 \end{pmatrix}.$$

Specifically, the common regressor is given by $X_c = \sum_{k=1}^{K}(k-1)1\{c_{k-1} \leq N_1 \leq c_k\}$, $K = 4$, $c_0 = -\infty$, $c_1, \ldots, c_{K-1}$, are respectively the quantiles of order 0.1, 0.37, 0.67 and 0.9 of the standard normal, and $c_K = \infty$. We consider two cases for the regressors that are observed in one of the datasets only, X_{nc}. In the first case, $(X_{nc,1}, X_{nc,2}) = (N_2, \exp(N_3))$ and in the second, $(X_{nc,1}, X_{nc,2}) = (\exp(N_2), \exp(N_3))$.

Figure 2 displays several identified sets for each of the two data-generating processes (DGPs) described above, each of them being associated with particular restrictions. Namely, the set in red, denoted by $\mathcal{B}^V$, is obtained from the variance restrictions only:

$$\mathcal{B}^V = \left\{\beta : \beta' V(X^0)\beta \leq V(Y^0)\right\} \cap \left\{\beta : \beta' V(X^1)\beta \leq V(Y^1)\right\},$$

where X^x and Y^x are defined as in Section 2.2. Hence, $\mathcal{B}^V$ is the intersection of two ellipses. The set in green, $\mathcal{B}^c$, is obtained as in Proposition 2 and relies on the restrictions $E(Y^x|X_{nc}, X_c = x) = X^{x\prime}\beta_0$ for $x \in \{0, 1\}$. Finally, the set in blue, $\mathcal{B}^{\text{con}}$, is a subset of $\mathcal{B}^c$ that imposes both convexity and monotonicity constraints on X_c.



(a) $X_{nc,1}$ and $\ln(X_{nc,2})$ normal

(b) $X_{nc,1}$ and $X_{nc,2}$ lognormal

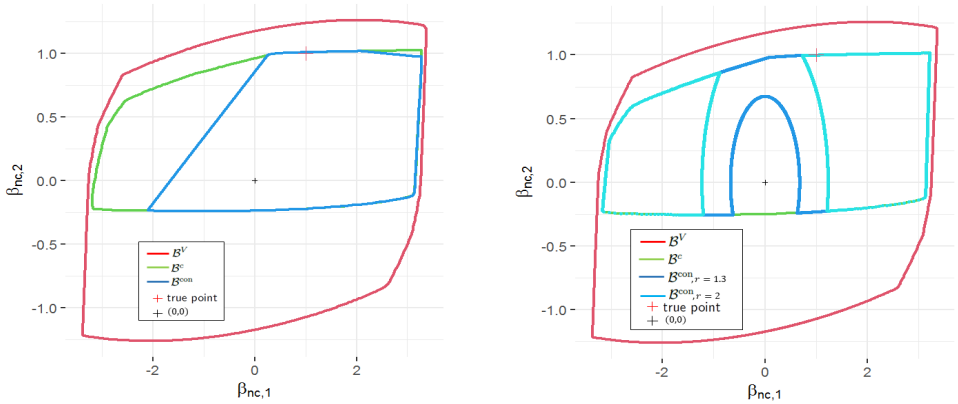
Note: the sets are obtained using a sample of size 100,000 and taking the convex hull of the set obtained from a uniform grid of 1,000 directions on the 2 dimensional sphere. $\mathcal{B}^{\text{con}}$ uses both convexity and monotonicity constraints on X_c.

Figure 2: Identification regions for different distributions of $(X_{nc,1}, X_{nc,2})$

A couple of comments are in order. In case (a) the restrictions implied by the model are much more informative than the variance restrictions, because of the non-normality of $X_{nc,2}$, and in particular the fact that it has fatter tails than the residuals U. The true point is at the boundary of $\mathcal{B}^c$, illustrating Proposition 1 applied conditional on $X_c = 0$ and $X_c = 1$. In this case, the shape restrictions are sufficient to imply that $0_2 \notin \mathcal{B}^{\text{con}}$ but also that $\beta_{nc,1} = 0$ and $\beta_{nc,2} = 0$. The identified set $\mathcal{B}^c$ is reduced further in case (b), as a result of the fatter tails of both $X_{nc,1}$ and $X_{nc,2}$. Like in case (a), the shape constraints on $f(X_c)$ allow to reduce dramatically the identified set.

Figure 3 presents convexity constraints and different constraints on the R^2 on the first DGP. While unlike case (a) of Figure 2, convexity constraints alone fail to reject $0_2 \notin \mathcal{B}^{\text{con}}$, imposing a constraint of the form $\underline{R}^2 \geq rR_s^2$ with $r > 1$ rejects it by

definition. In the latter case, the identified set is no longer convex, allowing to exclude some directions from the identified set and providing an informative lower bound on $|\beta_{nc,1}|$. Overall, that the sharp identified sets $\mathcal{B}^c$ are much more informative than the identified set $\mathcal{B}^V$ based on the variance restrictions highlights the importance of using all of the restrictions implied by the model. Another takeaway from these numerical illustrations is that sign constraints can be very informative in practice, resulting in significant shrinkage of the identified set.



(a) Convexity constraint on $f(\cdot)$ (b) $R_l^2 \geq rR_s^2$ constraints

Note: For Panel 3(a) and $\mathcal{B}^V$ and $\mathcal{B}^c$ in Panel 3(b), the sets are obtained as in Figure 2. For the constraints $R_l^2 \geq rR_s^2$ in Panel 3(b), we use 1,500 directions and no convexification. For this DGP, the true values of the R^2 of the long and short regressions are $R_l^2 = 0.307$ and $R_s^2 = 0.107$.

Figure 3: Identification regions for different shape restrictions

2.5 Regularization

An issue for estimation and inference on $\mathcal{B}$ is that when $\alpha \to 0$ or $\alpha \to 1$, $R(\alpha, F, G)$ is a ratio of two terms tending to 0. It follows that its plug-in estimator may become very unstable. To regularize the problem, we consider an outer set of $\mathcal{B}$ based on the removal of extreme values of α. We will focus on this outer set when we turn to estimation and inference in Section 3. Specifically, we define, for any $\varepsilon \in (0, 1/2)$,

$$S_\varepsilon(F, G) = \min_{\alpha \in [\varepsilon, 1-\varepsilon]} R(\alpha, F, G), \tag{10}$$

$$\mathcal{B}_\varepsilon = \left\{ \lambda q : q \in \mathcal{S}, 0 \leq \lambda \leq S_\varepsilon(F_{Y_0}, F_{X_0'q}) \right\}.$$

Note that for all F, G, $\alpha \mapsto R(\alpha, F, G)$ is continuous on $[\varepsilon, 1-\varepsilon]$. Thus, the minimum in (10) is well-defined. Proposition 6 below describes some properties of $\mathcal{B}_\varepsilon$ and relates it to the sharp identified set $\mathcal{B}$.

Proposition 6. *Suppose that Assumption 1 holds. Then:*

1. *For all $\varepsilon \in (0, 1/2)$, $\mathcal{B}_\varepsilon$ includes 0_p, is compact and convex;*
2. *For all $0 < \varepsilon < \varepsilon' < 1/2$, $\mathcal{B} \subset \mathcal{B}_\varepsilon \subset \mathcal{B}_{\varepsilon'}$ and $\cap_{\varepsilon \in (0,1/2)} \mathcal{B}_\varepsilon = \mathcal{B}$;*
3. *Suppose that F_Y is continuous and $U := Y_0 - X_0'\beta_0$ satisfies*

$$\forall \lambda > 0, \quad \lim_{t\to\infty} \sup_s \frac{\overline{F}_{\|X_0\|}(\lambda t)}{\overline{F}_{U|X_0'\beta_0 = s}(t)} = 0, \ \lim_{t\to\infty} \sup_s \frac{\overline{F}_{\|X_0\|}(\lambda t)}{\overline{F}_{-U|X_0'\beta_0 = s}(t)} = 0. \tag{11}$$

Then, there exists $\varepsilon_0 \in (0, 1/2)$ such that for all $\varepsilon \in (0, \varepsilon_0]$, $\mathcal{B} = \mathcal{B}_\varepsilon$.

The first part of Proposition 6 states that the regularized set $\mathcal{B}_\varepsilon$, for all $\varepsilon \in (0, 1/2)$, preserves the compactness and convexity of the sharp identified set $\mathcal{B}$. The second part states that $\mathcal{B}_\varepsilon$ is always a superset of $\mathcal{B}$, which is arbitrarily close to $\mathcal{B}$ as $\varepsilon \downarrow 0$. The third part states that if, basically, the tails of $\|X_0\|$ are thinner than those of U (Condition (11)), the set $\mathcal{B}_\varepsilon$ coincides with the sharp set $\mathcal{B}$ for ε small enough. Condition (11) holds in particular if X has a bounded support and $\text{Supp}(U) = \mathbb{R}$, or if U is symmetric and has a tail index larger than that of $\|X_0\|$. Note that $\mathcal{B} = \mathcal{B}_\varepsilon$ may hold even without (11). For instance, if both U and X are normally distributed, it is easy to check that $\mathcal{B}_\varepsilon = \mathcal{B}$ for all $\varepsilon \in (0, 1/2)$.

On the other hand, when U has thinner tails than $X'\beta_0$, $\mathcal{B}_\varepsilon$ will be a strict superset of $\mathcal{B}$ for ε large enough. In such cases, and under additional restrictions, we provide upper bounds on the Hausdorff distance between $\mathcal{B}$ and $\mathcal{B}_\varepsilon$ in Proposition 7. Intuitively, these bounds inform us about the maximal possible loss, in terms of identification, that is due to regularization.

Proposition 7. *Suppose that Assumption 1 holds and let $U := Y_0 - X_0'\beta_0$. Then:*

1. *Assume that X has an elliptical distribution with nonsingular variance matrix Σ, a density with respect to the Lebesgue measure and $\liminf_{|x|\to\infty} |x|^{1+c} f_{X_0'\beta_0}(x) > 0$ for some $c > 1$. Suppose also that $\limsup_{x\to\infty} x^d \overline{F}_{|U|}(x) < \infty$ for some $d > c$. Then, there exists $K_1 > 0$ such that for ε small enough,*

$$d_H(\mathcal{B}, \mathcal{B}_\varepsilon) \leq K_1 \varepsilon^{\frac{1/c - 1/d}{1+1/d}}.$$

2. *Assume that* $\beta_0 = 0_p$, $\liminf_{x\to\infty} \inf_{q\in\mathcal{S}} x^c \overline{F}_{X_0'q}(x) > 0$ *and* $\limsup_{x\to\infty} x^d \overline{F}_{|U|}(x) < \infty$ *for some* $d > c$. *Then, there exists* $K_2 > 0$ *such that for* ε *small enough,*

$$d_H(\mathcal{B}, \mathcal{B}_\varepsilon) \leq K_2 \varepsilon^{1/c-1/d}.$$

The tail conditions imposed in Proposition 7 are basically the opposite as in Point 3 of Proposition 6, as they imply that $\|X\|$ has fatter tails than U. The assumption that X has an elliptical distribution in Point 1 allows us to relate $S_\varepsilon(F_{Y_0}, F_{X_0'q}) - S(F_{Y_0}, F_{X_0'q})$, for any $q \in \mathcal{S}$, with $S_\varepsilon(F_{Y_0}, F_{X_0'\beta_0}) - S(F_{Y_0}, F_{X_0'\beta_0})$, but it is not necessary to obtain an upper bound on $S_\varepsilon(F_{Y_0}, F_{X_0'\beta_0}) - S(F_{Y_0}, F_{X_0'\beta_0})$.

In the two cases of Proposition 7, we produce upper bounds on the Hausdorff distance between $\mathcal{B}$ and $\mathcal{B}_\varepsilon$ that are, up to some constants, power of the regularization parameter ε. The upper bounds are close to 0 when ε is small, in line with Point 2 of Proposition 6. They are also closer to 0 the smaller c is, i.e. the fatter the tails of $X'\beta_0$ (or $X'q$) are, or the larger d is, i.e. the thinner the tails of U are.

3 Inference

We now consider the estimation of the identified set, and how to conduct inference on the parameters of interest β_0. As in the previous section, we first consider the case without common regressors before showing how to incorporate such regressors and combine them with additional constraints. We conclude this section by discussing some computational aspects of our procedure. We illustrate the finite sample performances of our inference method in Online Appendix C.

3.1 No common regressors

3.1.1 Estimation of the identification region and confidence region

We rely on random samples from the distributions of Y and X.

Assumption 2. *We observe* $(Y_1, ..., Y_{n_Y})$ *and* $(X_1, ..., X_{n_X})$, *two independent samples of i.i.d. variables with the same distribution as* Y *and* X, *respectively.*

For any $q \in \mathcal{S}$, let $\widehat{F}_Y$ and $\widehat{F}_{X'q}$ denote the empirical cdf of Y and $X'q$ and let $\widehat{F}_{Y_0}(t) = \widehat{F}_Y(t + \overline{Y})$ and $\widehat{F}_{X_0'q}(t) = \widehat{F}_{X'q}(t + \overline{X}'q)$. We simply estimate $R(\alpha, F_{Y_0}, F_{X_0'q})$ and $S_\varepsilon(F_{Y_0}, F_{X_0'q})$ by their empirical counterpart $R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ and $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$. It turns out that these functions can be computed quickly, as detailed in Section 3.3 below. We then also simply estimate the identified set $\mathcal{B}_\varepsilon$ by plug-in:

$$\widehat{\mathcal{B}}_\varepsilon := \left\{\lambda q : q \in \mathcal{S},\ 0 \leq \lambda \leq S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right)\right\}.$$

Next, we build confidence regions on β_0. The asymptotic distribution of $S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right)$ is not Gaussian in general, so we rely on subsampling (Politis et al., 1999). One could alternatively use the numerical bootstrap, see the discussion pp. 18-19 in the first version of D'Haultfœuille et al. (2023).

Let $n = (n_X n_Y)/(n_X + n_Y)$ and let b_n denote the size of the subsample. For any estimator $\widehat{\theta}$, let $\widehat{\theta}^*$ denotes its subsampling counterpart. For a nominal coverage of $1 - \alpha$, the confidence region on β_0 we consider is given by

$$\mathrm{CR}_{1-\alpha}(\beta_0) = \left\{\lambda q :\ q \in \mathcal{S},\ 0 \leq \lambda \leq S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right) - \widehat{c}_{\alpha,\varepsilon}(q) n^{-1/2}\right\},$$

where $\widehat{c}_{\alpha,\varepsilon}(q)$ is the quantile of order α of the distribution of $b_n^{1/2}[S_\varepsilon(\widehat{F}^*_{Y_0}, \widehat{F}^*_{X_0'q}) - S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})]$, conditional on the data.

Inference on subcomponents of β_0. In practice, one is often interested in conducting inference on subcomponents of β_0. In view of (9), the identified (outer) set $\mathcal{B}_{k,\varepsilon}$ of $\beta_{0,k}$ corresponding to $\mathcal{B}_\varepsilon$ satisfies

$$\mathcal{B}_{k,\varepsilon} = [-\sigma_\varepsilon(-e_k, F_{Y_0}, F_{X_0}),\ \sigma_\varepsilon(e_k, F_{Y_0}, F_{X_0})], \tag{12}$$

where $\sigma_\varepsilon(\cdot, F_{Y_0}, F_{X_0})$ denotes the support function associated to $q \mapsto S_\varepsilon(F_{Y_0}, F_{X_0'q})$ and e_k is the k-th element of the canonical basis of $\mathbb{R}^p$. To construct confidence intervals on β_{0k}, we first estimate $\sigma_\varepsilon(\cdot, F_{Y_0}, F_{X_0})$ by

$$\sigma_\varepsilon(e, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) = \frac{1}{\inf_{q \in \mathbb{R}^p : q'e = 1} 1/S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right)}, \tag{13}$$

see Corollary 1. Then, denoting by $\widetilde{c}_{\beta,\varepsilon}(e)$ the quantile of order $\beta \in (0,1)$ of the distribution of $b_n^{1/2}(\sigma_\varepsilon(e, \widehat{F}^*_{Y_0}, \widehat{F}^*_{X_0}) - \sigma_\varepsilon(e, \widehat{F}_{Y_0}, \widehat{F}_{X_0}))$, conditional on the data, the confidence

interval we consider for $\beta_{0,k}$ is

$$\mathrm{CI}_{1-\alpha}(\beta_{0,k}) = \left[\left(-\sigma_\varepsilon(-e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) + \frac{\widetilde{c}_{\alpha,\varepsilon}(-e_k)}{n^{1/2}}\right)^-, \left(\sigma_\varepsilon(e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) - \frac{\widetilde{c}_{\alpha,\varepsilon}(e_k)}{n^{1/2}}\right)^+\right],$$

where $x^- = \min(0, x)$ and $x^+ = \max(0, x)$. The rationale for using $(\cdot)^-$ and $(\cdot)^+$ is to ensure that $0 \in \mathrm{CI}_{1-\alpha}(\beta_{0,k})$: recall that without constraints, $0 \in \mathcal{B}_{k,\varepsilon}$. The advantage, then, is that we can still use the quantiles of order α while maintaining coverage even under point identification, as formally shown in Theorem 3 below.

Choice of the regularization parameter ε. Because $S_\varepsilon(F_{Y_0}, F_{X_0'q}) \geq S(F_{Y_0}, F_{X_0'q})$, the confidence regions and intervals above are conservative in general. To gain in efficiency, we suggest using several ε, and, basically, keep the one leading to the smallest confidence regions or intervals. We distinguish the cases $p = 1$, where we can adapt the choice to the direction $q \in \mathcal{S}$ while preserving the convexity of $\widehat{\mathcal{B}}_\varepsilon$, from the case $p > 1$. When $p = 1$, let us define, for $q \in \mathcal{S} = \{-1, 1\}$,

$$\varepsilon(q) = \underset{\varepsilon \in \mathcal{E}}{\operatorname{argmin}}\ S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) - \widehat{c}_{\alpha,\varepsilon}(q) n^{-1/2}, \tag{14}$$

where $\mathcal{E}$ is a finite grid in $(0, 1/2]$. Hence, $\varepsilon(q)$ simply minimizes the boundary value of the confidence region in the direction $q \in \mathcal{S}$. This idea is similar to that of Chernozhukov et al. (2013) in the context of intersection bounds.

Now consider the case $p > 1$. If one focuses on confidence intervals on β_{0k}, we need to choose the parameter ε that appears in $\sigma_\varepsilon(\pm e_k, F_{Y_0}, F_{X_0})$. To this end, we simply use $\varepsilon(q)$ as given above, with $q = \pm e_k$. If we are interested instead in the set $\mathcal{B}$ itself, we recommend using $\underline{\varepsilon} = \min_{q \in \mathcal{Q}} \varepsilon(q)$, where $\mathcal{Q}$ is a finite subset of $\mathcal{S}$.

3.1.2 Consistency and validity of the confidence region

The following theorem shows that $\widehat{\mathcal{B}}_\varepsilon$ is consistent for $\mathcal{B}_\varepsilon$, in the sense of the Hausdorff distance, under mild regularity conditions.

Theorem 2. *Suppose that Assumptions 1-2 hold. Then, as $n \to \infty$,*

$$d_H\left(\widehat{\mathcal{B}}_\varepsilon, \mathcal{B}_\varepsilon\right) \xrightarrow{\mathbb{P}} 0.$$

Next, we establish the asymptotic validity of $\mathrm{CR}_{1-\alpha}(\beta_0)$ and $\mathrm{CI}_{1-\alpha}(\beta_{0,k})$, under Assumptions 3 and 4 respectively. Assumption 5 (resp. 6) is used to establish the asymptotic validity of $\mathrm{CR}_{1-\alpha}(\beta_0)$ (resp. $\mathrm{CI}_{1-\alpha}(\beta_{0,k})$) using $\varepsilon(q)$ or $\underline{\varepsilon}$ (resp. $\varepsilon(\pm e_k)$), as defined above, instead of a fixed ε.

Assumption 3. *(Regularity conditions for $CR_{1-\alpha}(\beta_0)$) $E[\|X\|^2] < \infty$, $E[Y^2] < \infty$. Also, for all $q \in \mathcal{S}$, there exists $\varepsilon' \in (0, \varepsilon)$ such that $F_{X'q}$ and F_Y are continuous and strictly increasing on $[F^{-1}_{X'q}(\varepsilon'), F^{-1}_{X'q}(1-\varepsilon')]$ and $[F^{-1}_{Y}(\varepsilon'), F^{-1}_{Y}(1-\varepsilon')]$ respectively.*

Assumption 4. *(Regularity conditions for $CI_{1-\alpha}(\beta_{0,k})$) $E[\|X\|^2] < \infty$, $E[Y^2] < \infty$. Also, there exists $\varepsilon' \in (0, \varepsilon)$ such that for all $(\alpha, \alpha') \in [\varepsilon', 1-\varepsilon']^2$, there exists a strictly increasing and continuous function m such that $m(0) = 0$ and*

$$\begin{aligned} \sup_{q \in \mathcal{S}} \left| F^{-1}_{X'q}(\alpha') - F^{-1}_{X'q}(\alpha) \right| &< m(|\alpha' - \alpha|), \\ \left| F^{-1}_{Y}(\alpha') - F^{-1}_{Y}(\alpha) \right| &< m(|\alpha' - \alpha|). \end{aligned} \tag{15}$$

Finally, for all $e = \pm e_k$ $(k = 1, ..., p)$, either (i) $\sigma_\varepsilon(e, F_{Y_0}, F_{X_0}) > \sigma(e, F_{Y_0}, F_{X_0})$, (ii) $q \mapsto [qS_\varepsilon(F_{Y_0}, F_{X_0'q})]'e$ admits a unique maximizer on $\mathcal{S}$, or (iii) for all $q_m \in \arg\max_{q \in \mathcal{S}} [qS_\varepsilon(F_{Y_0}, F_{X_0'q})]'e$, $a \mapsto R(a, F_Y, F_{X'q_m})$ admits a unique minimizer on $[\varepsilon, 1-\varepsilon]$.

Assumption 5. *(Regularity conditions for the validity of $CR_{1-\alpha}(\beta_0)$ based on data-dependent ε) For all $q \in \mathcal{S}$, we either have (i) $S_\varepsilon(F_{Y_0}, F_{X_0'q}) > S(F_{Y_0}, F_{X_0'q})$ for all $\varepsilon \in \mathcal{E}$, or (ii) $a \mapsto R(a, F_{Y_0}, F_{X_0'q})$ admits a unique minimizer on $(0, 1)$.*

Assumption 6. *(Regularity conditions for the validity of $CI_{1-\alpha}(\beta_{0,k})$ based on data-dependent ε) For all $e = \pm e_k$ $(k = 1, ..., p)$, we either have (i) $\sigma_\varepsilon(e, F_{Y_0}, F_{X_0}) > \sigma(F_{Y_0}, F_{X_0})$ for all $\varepsilon \in \mathcal{E}$ or (ii) for all $q_m \in \arg\max_{q \in \mathcal{S}} [qS_{\varepsilon_{j_0}}(F_{Y_0}, F_{X_0'q})]'e$, $a \mapsto R(a, F_{Y_0}, F_{X_0'q_m})$ admits a unique minimizer $a(q_m)$ on $(0, 1)$, with $a(q_m) \in [\varepsilon_{j_0}, 1-\varepsilon_{j_0}]$ and $\varepsilon_{j_0} := \max\{\varepsilon \in \mathcal{E} : \sigma_\varepsilon(F_{Y_0}, F_{X_0'q}) = \sigma(F_{Y_0}, F_{X_0'q})\}$.*

The second part of Assumption 3 holds if for all $q \in \mathcal{S}$, the distributions of $X'q$ and Y are continuous with respect to the Lebesgue distribution and their support is a (possibly unbounded) interval. The first part of Assumption 4 is basically a reinforcement of Assumption 3 to ensure that some of our results hold uniformly over q. This is

needed when we consider the support function, as this function implies an optimization over q. A sufficient condition for (15) is that, for all $q \in \mathcal{S}$, $X'q$ admits a density $f_{X'q}$ with respect to the Lebesgue measure and $\inf_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} f_{X'q}(F^{-1}_{X'q}(\alpha)) > 0$. The conditions (ii) and (iii) in Assumption 4 are sufficient conditions for the continuity of the asymptotic distribution of $n^{1/2}\left(\sigma_\varepsilon(e,\widehat{F}_{Y_0},\widehat{F}_{X_0}) - \sigma_\varepsilon(e,F_{Y_0},F_{X_0})\right)$, which is necessary for the validity of subsampling.

In Assumption 5, we exclude cases where $S_{\varepsilon_0}(F_{Y_0},F_{X_0'q}) = S(F_{Y_0},F_{X_0'q})$ for some $\varepsilon_0 \in (0,1/2)$ and there are multiple minima for $a \mapsto R(a,F_{Y_0},F_{X_0'q})$. This occurs when both X and Y are Gaussian. Assumption 6 is basically similar to Assumption 5 but somewhat more complicated, as we consider therein the support function instead of the radial function.

Theorem 3. *Fix $(\varepsilon,\alpha) \in (0,1/2)^2$ and suppose that $n_X/(n_X+n_Y) \to \mu \in (0,1)$, $b_n \to \infty$, $b_n/n \to 0$ and Assumptions 1-2 hold. Then:*

1. *If Assumption 3 also holds,*

$$\inf_{\beta\in\mathcal{B}} \liminf_{n\to\infty} P\left(\beta \in CR_{1-\alpha}(\beta_0)\right) \geq 1-\alpha, \tag{16}$$

with equality if $\mathcal{B} = \mathcal{B}_\varepsilon$. Moreover, if Assumption 5 also holds, (16) *is still true if we use $\varepsilon(q)$ (when $p=1$) or $\underline{\varepsilon}$ (when $p>1$) instead of ε.*

2. *If Assumption 4 also holds,*

$$\liminf_{n\to\infty} \inf_{\beta_k\in\mathcal{B}_k} P\left(\beta_k \in CI_{1-\alpha}(\beta_{0,k})\right) \geq 1-\alpha, \tag{17}$$

with equality if $\mathcal{B}_k = \mathcal{B}_{k,\varepsilon}$. Moreover, if Assumption 6 also holds, (17) *is still true if we use $\varepsilon(e_k)$ and $\varepsilon(-e_k)$ instead of ε.*

To prove (16)-(17), we first show the weak convergence of

$$\sqrt{n}\left(R(\alpha,\widehat{F}_{Y_0},\widehat{F}_{X_0'q}) - R(\alpha,F_{Y_0},F_{X_0'q})\right),$$

seen as a process indexed by either α or (α,q). The convergence in distribution of $S_\varepsilon(\widehat{F}_{Y_0},\widehat{F}_{X_0'q})$ and $\sigma_\varepsilon(e,\widehat{F}_{Y_0},\widehat{F}_{X_0})$, and in turn (16)-(17), then essentially follows by the Hadamard directional differentiability of the minimum and maximin maps, shown respectively by Cárcamo et al. (2020) and Firpo et al. (2023).

Our results for a fixed $\varepsilon > 0$ extend to the data-dependent $\varepsilon(q)$ and $\underline{\varepsilon}$, under the additional conditions provided above. Note that one could avoid these conditions by using sample splitting, with one subsample used to choose $\varepsilon(q)$ or $\underline{\varepsilon}$ and the other to construct the confidence regions/intervals. One drawback of this alternative solution, though, is that it increases the size of confidence regions/intervals, to a point that we may lose the benefits of using a data-dependent rather than a fixed ε.

3.2 Common regressors and possible constraints

We now turn to inference on β_0 with common regressors X_c. Recall from Proposition 2 that the identified set on β_0 is

$$\mathcal{B}^c = \left\{\lambda q : q \in \mathcal{S},\ 0 \leq \lambda \leq \overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c})\right\},$$

with $\overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c}) = \inf_{x \in \text{Supp}(X_c)} S(F_{Y^x|X_c=x}, F_{X^{x\prime}q|X_c=x})$.

Let us first assume that X_c has a finite support. Let $\widehat{F}_{Y^x|X_c=x}$ and $\widehat{F}_{X^{x\prime}q|X_c=x}$ denote the empirical estimators of $F_{Y^x|X_c=x}$ and $F_{X^{x\prime}q|X_c=x}$, respectively. Following the same logic as above, we estimate $\overline{S}(F_{Y,X_c}, F_{X'_{nc}q,X_c})$ by

$$\widehat{\overline{S}}(q, F_{Y,X_c}, F_{X'_{nc}q,X_c}) = \min_{x \in \text{Supp}(X_c)} S_\varepsilon(\widehat{F}_{Y^x|X_c=x}, \widehat{F}_{X^{x\prime}q|X_c=x}).$$

Let $\widehat{c}^c_{\alpha,\varepsilon}(q)$ be the quantile of order $\alpha \in (0,1)$ of the distribution of $b_n^{1/2}(\widehat{\overline{S}}^*(q, F_{Y,X_c}, F_{X_{nc},X_c}) - \widehat{\overline{S}}(q, F_{Y,X_c}, F_{X_{nc},X_c}))$, conditional on the data. For a nominal coverage of $1-\alpha$, the confidence region on β_0 we consider is

$$\text{CR}^c_{1-\alpha}(\beta_0) = \left\{\lambda q : q \in \mathcal{S},\ 0 \leq \lambda \leq \widehat{\overline{S}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) - \widehat{c}^c_{\alpha,\varepsilon}(q) n^{-1/2}\right\}.$$

With continuous common regressors, one can adapt the earlier arguments using sieve estimation. Specifically, suppose that Model (1) holds and consider a linear sieve approximation of $f(\cdot)$ by a step function $x_c \mapsto \sum_{k=1}^{K_n} \mathbb{1}\{x_c \in I_{n,k}\}\gamma_k$ for some partition $(I_{n,k})_{k=1\ldots K_n}$ of the support of X_c and with K_n tending to infinity at an appropriate rate. Then, one can construct a confidence region on β_0 by following a similar logic as above.[6]

[6] Establishing the asymptotic validity of such a confidence region would require to handle both the bias stemming from the approximation of $f(\cdot)$ and the increasing complexity of the approximation. We leave this analysis for future research.

We now discuss how to conduct inference under constraints on the R^2 or shape restrictions, as considered in Subsections 2.3.1 and 2.3.2 respectively. The main difference with above is that for a given direction $q \in \mathcal{S}$, both the lower and upper bounds on the identified set need to be estimated. As before, we estimate them with plug-in estimators.[7] The only substantive difference is that in the confidence regions, we then account for the variability of both bounds. For instance, with shape restrictions, we consider the following confidence region:

$$\mathrm{CR}^{con}_{1-\alpha}(\beta_0) = \Big\{ \lambda q : q \in \mathcal{S},\ \underline{\widehat{S}}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) + \underline{\widehat{c}}^{con}_{1-\alpha/2,\varepsilon}(q) n^{-1/2} \leq \lambda \\ \leq \overline{\widehat{S}}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) - \overline{\widehat{c}}^{con}_{\alpha/2,\varepsilon}(q) n^{-1/2} \Big\},$$

where $\underline{\widehat{c}}^{con}_{\delta,\varepsilon}(q)$ is the quantile of order δ of $b_n^{1/2}(\underline{\widehat{S}}^{con*}(q, F_{Y,X_c}, F_{X_{nc},X_c}) - \underline{\widehat{S}}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}))$, conditional on the data and similarly for $\overline{\widehat{c}}^{con}_{\delta,\varepsilon}$. Although to conserve on space we do not establish any formal result, using quantiles of order $\alpha/2$ and $1-\alpha/2$, depending on whether $\underline{S}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c}) = \overline{S}^{con}(q, F_{Y,X_c}, F_{X_{nc},X_c})$, should ensure that the confidence region is asymptotically conservative.

3.3 Computational aspects

We first discuss how to efficiently compute $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$. Let $Y_{(1)} < ... < Y_{(m_y)}$ represent the $m_y \leq n_y$ distinct, ordered values of the $(Y_i)_{i=1,...,n_y}$ and let $W^Y_{(j)} = \#\{i : Y_i = Y_{(j)}\}/n_y$. Let us also define $I^Y = \{\sum_{j=1}^i W^Y_{(j)} : i = 1, ..., m_y - 1\}$. We define similarly $W^{X'q}_{(j)}$ and $I^{X'q}$. By construction, the numerator $\widehat{f}^Y(\alpha) := \int_\alpha^1 \widehat{F}^{-1}_{Y_0}(t)dt$ of $R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ is linear on all intervals $[\sum_{j=1}^i W^Y_{(j)}, \sum_{j=1}^{i+1} W^Y_{(j)}]$ $(i = 0, ..., m_y - 1)$. Moreover, for any $\alpha = \sum_{j=1}^i W^Y_{(j)} \in I^Y$,

$$\widehat{f}^Y(\alpha) = \sum_{j=i+1}^{m_y} W^Y_{(j)} \left(Y_{(j)} - \overline{Y} \right). \tag{18}$$

[7] With a finite number of constraints and X_c finitely supported, one can prove that the plug-in estimators $\underline{S}^c(m_Y, m_{X_{nc}}, q)$ and $\overline{S}^c(m_Y, m_{X_{nc}}, q)$ are consistent if there are no $r \in \mathcal{R}$ such that $[Rm'_{X_{nc}}q](r) = [Rm_y - \underline{c}](r) = 0$. In the latter case, one may need to replace the plug-in estimator by removing the limit and using u_n instead of u, with $u_n \to 0$ at an appropriate rate. We leave this issue for future research.

The same holds for the denominator $\widehat{f}^{X'q}(\alpha)$ of $R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$. As a result, $R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ is of the form $(a\alpha + b)/(c\alpha + d)$ on intervals between two consecutive values of $I^Y \cup I^{X'q}$. Now, observe that the minimum of such a function is reached at one of the endpoints of the interval. As a result, we can compute $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ using the following algorithm:

1. Compute $\widehat{f}^Y(\cdot)$ on I^Y using (18) and let $\widehat{f}^Y(0) = \widehat{f}^Y(1) = 0$. Proceed similarly with $\widehat{f}^{X'q}(\cdot)$;

2. Interpolate linearly $\widehat{f}^Y(\cdot)$ (resp. $\widehat{f}^{X'q}(\cdot)$) on $\{\varepsilon, 1-\varepsilon\} \cup I^{X'q}$ (resp. $\{\varepsilon, 1-\varepsilon\} \cup I^Y$).

3. Compute $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) = \min_{\alpha \in \{\varepsilon, 1-\varepsilon\} \cup I^Y \cup I^{X'q}} \widehat{f}^Y(\alpha)/\widehat{f}^{X'q}(\alpha)$.

To compute $\sigma_\varepsilon(\pm e_k, F_{Y_0}, F_{X_0})$, we solve (13), in which $q \mapsto 1/S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ is also convex. In practice, we use the BFGS quasi-Newton method implemented in the R package `optim`, using as a starting point the considered direction e.

Finally, the exact computation of $\widehat{\mathcal{B}}_\varepsilon$ and $\mathrm{CR}_{1-\alpha}(\beta_0)$ requires the computation of $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ and $\widehat{c}_{\alpha,\varepsilon}(q)$ for all $q \in \mathcal{S}$, which is in practice infeasible if $p > 1$ as $\mathcal{S}$ is infinite. Instead, we suggest to (i) fix a grid $\widetilde{\mathcal{S}} \subset \mathcal{S}$; (ii) compute $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ and $\widehat{c}_{\alpha,\varepsilon}(q)$ for each $q \in \widetilde{\mathcal{S}}$; (iii) construct an approximation of $\widehat{\mathcal{B}}_\varepsilon$ and $\mathrm{CI}_{1-\alpha}$ by computing the convex hulls of $\{S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})q : \ q \in \widetilde{\mathcal{S}}\}$ and $\left\{\left(S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) - \widehat{c}_{\alpha,\varepsilon}(q)n^{-1/2}\right) q : q \in \widetilde{\mathcal{S}}\right\}$, respectively.[8] The resulting sets, $\widetilde{\mathcal{B}}_\varepsilon$ and $\widetilde{\mathrm{CR}}_{1-\alpha}(\beta_0)$ say, are convex, inner approximations of $\widehat{\mathcal{B}}_\varepsilon$ and $\mathrm{CR}_{1-\alpha}(\beta_0)$, and satisfy, as $d_H(\mathcal{S}, \widetilde{\mathcal{S}}) \to 0$, $d_H(\widetilde{\mathcal{B}}_\varepsilon, \widehat{\mathcal{B}}_\varepsilon) \to 0$ and $d_H(\widetilde{\mathrm{CR}}_{1-\alpha}(\beta_0), \mathrm{CR}_{1-\alpha}(\beta_0)) \to 0$.

The computation of the estimated set, the confidence regions on β_0 and γ_0 in the specification $f(X_c) = X_c'\gamma_0$ (where X_c is the vector of all dummy variables associated with a finitely supported variable) and the confidence intervals on the corresponding subcomponents are implemented in our companion R package `RegCombin`. The package also handles shape restrictions and lower bound on the R^2 of the long regression, as well as combinations of these. The `RegCombin` vignette, available through the description of the package on CRAN, provides additional details about the implementation, including the choice of the tuning parameters $\mathcal{E}$ and b_n.

[8] The convex hull of n points in $\mathbb{R}^p$ can be computed efficiently by the quickhull algorithm (Barber et al., 1996), which requires around $n^{p/2}$ operations.

4 Application to intergenerational mobility in the United States

We now apply our method to conduct inference on the intergenerational income mobility over the period 1850 to 1930 in the United States, revisiting the influential analysis of Olivetti and Paserman (2015) on this question. We follow their paper and focus on the father-son and father-son-in-law intergenerational income elasticities. We conduct our analysis using 1 percent extracts from the decennial censuses of the United States, over the period 1850 to 1930 (1850-1930 IPUMS).[9]

An important feature of the historical Census data used in this analysis is that father's and son's (as well as son-in-law's) incomes are not jointly observed. Olivetti and Paserman (2015) address this measurement issue by predicting, for any given child (John, say) observed in one of the Census datasets, their father's log earnings using the mean log earnings of fathers whose children have the same first name (namely, John). Olivetti and Paserman then estimate in a second step the intergenerational elasticity by regressing son's log earnings on the predicted father's log earnings computed from the previous step. This procedure boils down to a two-sample two-stage least squares estimator (TSTSLS).[10] The corresponding exclusion restriction that the son's first name does not predict his log earnings, once we control for his father's log earnings, may nonetheless be problematic; see Santavirta and Stuhler (2022) for a critical review of the empirical literature using TSTSLS in this context of intergenerational mobility. For the periods 1860-1880 and 1880-1900 only, the IPUMS Linked Representative Samples link fathers and sons using information on first and last names, which allows us to estimate more directly the father-son elasticity using OLS.

[9]We refer the reader to Section 2 of Olivetti and Paserman (2015) for a detailed discussion of the data used in the analysis. Note that they estimate the evolution of the intergenerational income mobility over a longer time window (1850 to 1940) than we do. We confine our analysis to the period 1850-1930 as the 1940 portion of the data (1% extract of the IPUMS Restricted Complete Count Data) is not publicly available.

[10]Another limitation of the data used in Olivetti and Paserman (2015) and in this application is that it does not allow us to directly calculate the intergenerational elasticity in income. Instead, we follow the baseline specification of Olivetti and Paserman (2015) and proxy income using an index of occupational standing available from IPUMS (OCCSCORE), which is constructed as the median total income of the persons in each occupation in 1950.

Using our notation and consistent with Olivetti and Paserman (2015), the population parameter of interest here is given by

$$\theta_0 := \frac{\text{Cov}(Y, X_{nc})}{V(X_{nc})} = \beta_0 + \left(\frac{\text{Cov}(X_c, X_{nc})}{V(X_{nc})}\right)' \gamma_0,$$

where Y denotes the son's (or son-in-law's) log-income, X_{nc} the father's log-income and X_c the vector of indicators corresponding to the son's (or son-in-law's) first names observed in both datasets. The second equality follows from (1), since X_c is discrete and thus $f(X_c) = X_c'\gamma_0$ for some γ_0. In what follows, we report the upper bound of the estimated identified set and confidence interval on θ_0.

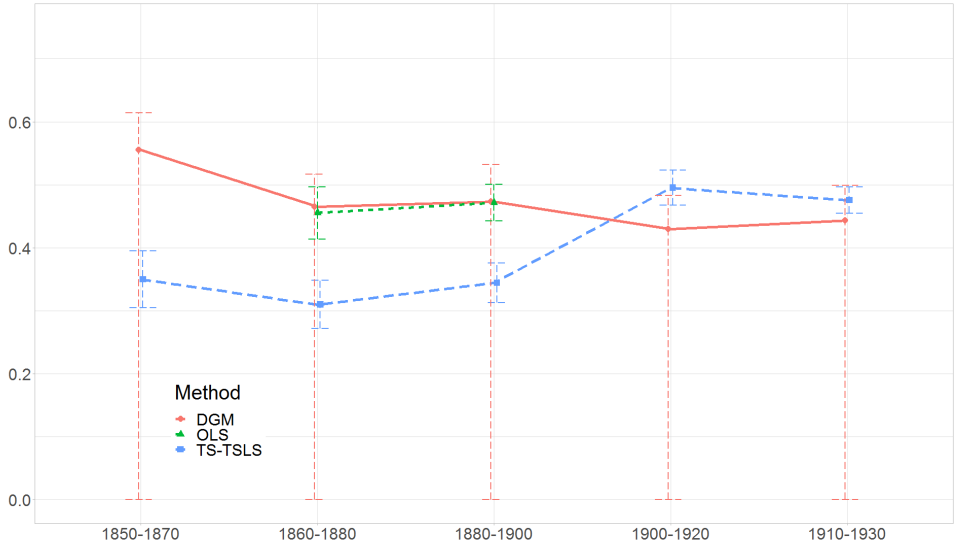
Even though the sample sizes as well as the number of common regressors X_c are quite large, our method can still be implemented at a very reasonable computational cost. For instance, for the sample of sons over the first period (1850-1870), the computation of the confidence intervals only takes less than 4 minutes with our R package. As expected, computational time is highest for the period 1910-1930 associated with the largest number of observations, with $n > 100,000$ for both samples of Y and X_{nc}. Nonetheless, our inference procedure remains tractable in this case too, with a computational time of about 11 minutes.[11] Overall, this illustrates the applicability of our method, which can be easily implemented even in this type of rich and high-dimensional data environment.

Figures 4(a)-4(b) and Table 1 below display the results, for the father-son as well as father-son-in-law elasticities, obtained using our approach, the TSTSLS and, for the sample of sons over the years 1860-1880 and 1880-1900, the OLS.[12] Specifically, we report in Figures 4(a)-4(b) the estimated upper bounds of the identified sets (in solid red) and the confidence intervals (dashed red) obtained with our method, the TSTSLS estimates and confidence intervals (solid and dashed blue, resp.) as well as,

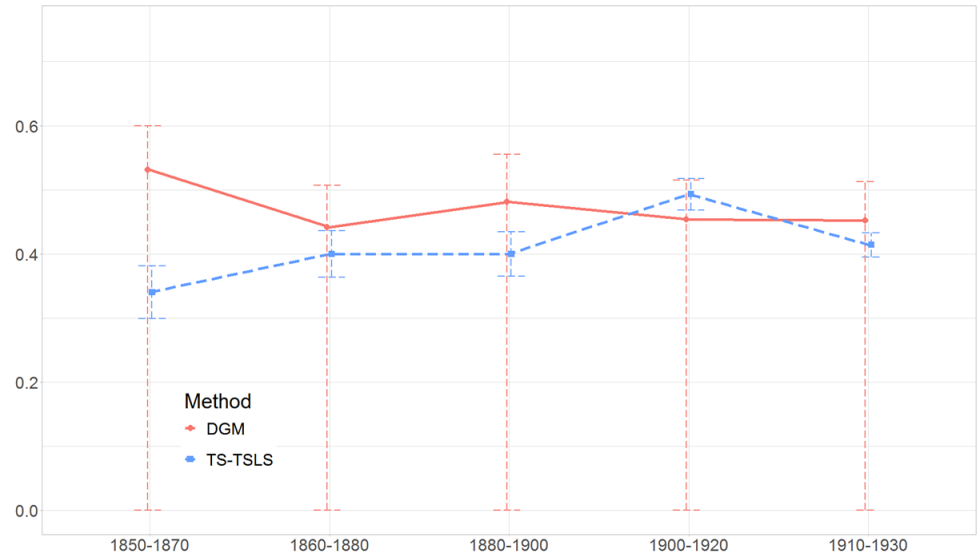
[11] These CPU times are obtained using our companion R package, parallelized on 20 CPUs on an Intel Xeon Gold 6130 CPU 2.10GHz with 382Gb of RAM.

[12] In practice we need to restrict the set of first names included in X_c to avoid very uncommon occurrences that are perfect predictors of the outcome variable Y. In our baseline specification, we implement this by restricting X_c to the set of first names that account for at least 0.01% of the observations in the pooled sample, and appear at least 10 times in either of the samples. We discuss in the following the robustness of our results to alternative cutoffs.

for 1860-1880 and 1880-1900 and the sample of sons only, the OLS estimates and confidence intervals (solid and dashed green, resp.).



(a) For sons



(b) For sons-in-law

Note: for readability and because 0 is a natural lower bound, the y-axis starts at 0, even though the lower bounds of our confidence intervals without restrictions are negative (see Table 1).

Figure 4: Intergenerational income correlation using different methods.

Sample: **Sons**	**1850-1870**	**1860-1880**	**1880-1900**	**1900-1920**	**1910-1930**
DGM, set	[-0.555,0.555]	[-0.465,0.465]	[-0.473,0.473]	[-0.430,0.430]	[-0.443,0.443]
DGM, CI.	[-0.614,0.614]	[-0.517,0.517]	[-0.532,0.532]	[-0.483,0.483]	[-0.499,0.499]
DGM, $\underline{R}^2 \geq 1.3R_s^2$, set	[0.081,0.555]	[0.075,0.465]	[0.075,0.473]	[0.076,0.430]	[0.071,0.443]
DGM, $\underline{R}^2 \geq 1.3R_s^2$, CI.	[0.033,0.617]	[0.034,0.519]	[0.039,0.527]	[0.044,0.477]	[0.047,0.491]
DGM, $\underline{R}^2 \geq 2R_s^2$, set	[0.163,0.555]	[0.151,0.465]	[0.153,0.473]	[0.164,0.430]	[0.159,0.443]
DGM, $\underline{R}^2 \geq 2R_s^2$, CI.	[0.095,0.601]	[0.093,0.506]	[0.102,0.513]	[0.127,0.459]	[0.132,0.477]
TSTSLS, pt.	0.350	0.310	0.344	0.495	0.476
TSTSLS, CI.	[0.305,0.395]	[0.272,0.348]	[0.313,0.376]	[0.468,0.523]	[0.454,0.497]
Test of equality, p-value	<0.001	0.001	0.001	0.999	0.014
(Stat.; critical val. 95%)	(28.52; 15.16)	(25.21; 12.37)	(25.92; 28.30)	(15.09; 17.69)	(8.18; 6.79)
OLS, pt.		0.455	0.472		
OLS, CI.		[0.414,0.497]	[0.443,0.501]		
Test pt identification, p-value		0.147	0.003		
(Stat.; critical val. 95%)		(9.21,17.03)	(23.06,6.33)		
Number of names X_c	225	261	382	514	598
Sample sizes Y and X_{nc}	(39,734; 34,603)	(55,728; 47,014)	(85,340; 73,999)	(116,986; 102,053)	(131,089; 116,328)
Sample: **Sons-in-law**	**1850-1870**	**1860-1880**	**1880-1900**	**1900-1920**	**1910-1930**
DGM, set	[-0.531,0.531]	[-0.442,0.442]	[-0.481,0.481]	[-0.454,0.454]	[-0.452,0.452]
DGM, CI.	[-0.601,0.600]	[-0.507,0.507]	[-0.554,0.555]	[-0.515,0.515]	[-0.513,0.513]
DGM, $\underline{R}^2 \geq 1.3R_s^2$, set	[0.089,0.531]	[0.085,0.442]	[0.075,0.481]	[0.073,0.454]	[0.062,0.452]
DGM, $\underline{R}^2 \geq 1.3R_s^2$, CI.	[0.030,0.605]	[0.036,0.505]	[0.029,0.552]	[0.037,0.513]	[0.033,0.509]
DGM, $\underline{R}^2 \geq 2R_s^2$, set	[0.186,0.531]	[0.178,0.442]	[0.159,0.481]	[0.164,0.454]	[0.146,0.452]
DGM, $\underline{R}^2 \geq 2R_s^2$, CI.	[0.115,0.596]	[0.114,0.490]	[0.105,0.534]	[0.122,0.499]	[0.113,0.496]
TSTSLS, pt.	0.340	0.400	0.400	0.493	0.414
TSTSLS, CI.	[0.299,0.381]	[0.364,0.436]	[0.365,0.434]	[0.469,0.518]	[0.395,0.433]
Test of equality, p-value	<0.001	0.998	0.012	1	1
(Stat.; critical val. 95%)	(23.12; 9.67)	(5.87; 18.53)	(13.08; 12.94)	(8.03; 13.28)	(8.33; 13.07)
Number of names X_c	155	212	323	468	545
Sample sizes Y and X_{nc}	(25,760; 33,256)	(32,970; 45,800)	(49,068; 71,141)	(73,425; 99,871)	(85,122; 112,763)

Notes: Dependent variable Y is son's (or son-in-law's) log income. Common regressors X_c are dummies for the first names appearing more than 0.01% in the pooled dataset and 10 times in both datasets. "DGM, set" and "DGM, CI." refer to the estimated identified set and 95% confidence interval, respectively, obtained with our method. "TSTSLS, pt." and "TSTSLS, CI." refer to the TSTSLS point estimate and 95% confidence interval, respectively. The test of equality between the TSTSLS (β_{TSTSLS}) estimates and DGM (β_{DGM}) upper bound estimates is performed using subsampling with 1,000 replications. The statistic ("Stat.") is $n^{1/2}\widehat{\theta}$, where $\widehat{\theta} = \widehat{\beta}_{TSTSLS} - \widehat{\beta}_{DGM}$ and $n = n_y n_x/(n_y + n_x)$, n_y and n_x being the respective sample sizes of Y and X_{nc}. The critical value corresponds to the $1-\alpha$ quantile of the distribution of $b_n^{1/2}|\widehat{\theta}^* - \widehat{\theta}|$, where $\widehat{\theta}^*$ is a subsampled version of $\widehat{\theta}$ and b_n is the subsample size. The sample sizes where the joint distribution is observed for both periods 1860-1880 and 1880-1900 are respectively 3,947 and 9,076. The R^2 on the short and long regressions are respectively 0.04 and 0.18 for 1860-1880, and 0.02 and 0.17 for 1880-1900. The test for point identification is performed with the ε selected in (14), however this choice appears conservative on simulations. Taking $\varepsilon/2$ yields p-values of 0.69 for the period 1860-1880 and 0.04 for 1880-1900.

Table 1: Intergenerational income correlation for sons using different methods.

A first conclusion from these results is that the upper bounds of the confidence intervals associated with our method range, depending on the periods, between 0.48 and 0.61 (0.51 and 0.6) for the sample of sons (sons-in-law). These values of the intergenerational correlation coefficient are all well below the natural upper bound of 1. Also, even though the estimates vary depending on the data and econometric specification being used, most of the existing point estimates of the father-son income elasticity range between 0.40 and 0.50 (Olivetti and Paserman, 2015). Overall, this clearly indicates that our method leads to informative inference on the parameter of interest.

Second, consider the two cases where the linked data is available (1860-1880 and 1880-1900 for the sample of sons). Results in Table 1 indicate that the corresponding OLS estimates of the intergenerational income elasticities are quantitatively very close to the estimated upper bound of our identified set. Recall that, from Proposition 5 in Section 2.3.4, the upper bound of our identified set ($\overline{\theta}_0$, say) plays a special role: under an additional restriction on the distributions of X_{nc} and the error term, θ_0 is actually point identified and equal to $\overline{\theta}_0$.[13] In other words, the results from these two periods support the hypothesis that the restriction on the distributions of X_{nc} and the error term guaranteeing point identification of θ_0 by $\overline{\theta}_0$ hold. Besides, the fact that we do not reject at standard levels the null hypothesis of point identification with our formal test described in Section A.3 for the period 1860-1880 (p-value of 0.147) provides suggestive evidence in this direction.[14] Under this assumption, our results are informative not only on the maximal father-son elasticity coefficient for a given period of time, but also on its evolution. It follows in particular that our estimates point to a mild decrease in this elasticity coefficient for sons between 1850 and 1930.

Third, the results from the test of equality reported in Table 1 indicate that the TSTSLS estimates are in several cases statistically distinguishable from the estimated upper bounds of our identified sets. This includes, for the sample of sons, all periods

[13] Proposition 5 is obtained without X_c. Yet, it can be combined with Proposition 2 to show that β_0, and in turn γ_0 (and thus θ_0 here) are point identified with such X_c.

[14] Simulation results available from the authors upon request indicate that our choice of ε tends to be conservative for the test of point identification. One would not reject either at the 1% level the null hypothesis for the period 1880-1900 with a less conservative choice of ε (e.g. we obtain a p-value of 0.04 using $\varepsilon/2$).

with the exception of 1900-1920, and the periods 1850-1870 and 1880-1900 for the sample of sons-in-law. Besides, for the sample of sons in particular, the TSTSLS estimates exhibit a sharp increase, while our estimated upper bound decreases between the periods 1880-1900 and 1900-1920. In that sense, our results offer suggestive evidence that the intergenerational income correlation might have been more stable at the beginning of the 20th century than what one would infer from the TSTSLS estimates.

Fourth, we also report in Table 1 the estimated identified set and confidence intervals associated with our method when we impose a lower bound on the R^2 of the long regression, namely $\underline{R}^2 \geq 1.3R_s^2$ or $\underline{R}^2 \geq 2R_s^2$. Imposing any of these restrictions, which are satisfied for the periods 1860-1880 and 1880-1900 for which the linked data is available, results in substantially tighter confidence intervals. In particular, for the sample of sons, the confidence intervals obtained under the restriction $\underline{R}^2 \geq 2R_s^2$ allow us to reject values of the intergenerational income elasticity coefficient smaller than 0.13 and larger than 0.48 for the years 1910-1930.

Finally, we consider in Tables 8 and 9, and Figure 5 in online Appendix D several robustness checks. They relate to the set of first names that we include as controls in our estimation procedure (Panel A), the choice of ε (Panel B and Figure 5), and restrictions of the sample to the set of individuals whose first name is included in the set of controls X_c (Panel C). Throughout the tables, we focus on the upper bound of the estimated identified set ("DGM, set") and of the confidence interval ("DGM, CI.").

The main takeaway from Table 8 and Figure 5 is that, for the sample of sons, the results from our inference procedure are qualitatively, and in most cases quantitatively, robust to these different sensitivity analyses. The one case that exhibits more sensitivity is the specification where we control for the first names that account for at least 0.02% of the sample, instead of 0.01% in our baseline specification. The upper bound of our confidence interval for the period 1900-1920 increases in this case from 0.48 to 0.58, the results remaining, however, stable for the other periods. The results for the sample of sons-in-law (Table 9) are also, for most periods at the exception of the same limit for 1900-1920, qualitatively, and in some cases quantitatively similar across specifications. The main difference with the sample of sons is that the choice

of ε does appear to matter more for the sons-in-law, a limitation that one should keep in mind when interpreting the findings for this subgroup. Nonetheless, to the extent that our baseline choice of ε (see Section 3.3) is motivated by the theory and is found to perform well in our Monte Carlo simulation exercises, we do not view this as particularly worrisome.

5 Conclusion

We study the identification of and inference on partially linear models, in an environment where the outcome of interest and some of the covariates are observed in two different datasets that can not be matched. This setup arises in particular when one is interested in the effect of a variable that is not observed jointly with the outcome variable, or in cases where potential confounders are observed in a different dataset from the one including the outcome and regressor of interest. In such situations, researchers often rely on strong assumptions to point identify their parameters of interest. Our approach offers a useful alternative when such assumptions are debatable. The application shows that in addition to its tractability, our method is able to deliver informative bounds. Finally, beyond the model considered in this paper, our analysis suggests that the radial function is an appealing tool in partial identification problems where the support function proves difficult to compute.

References

Abrevaya, J. and W. Jiang (2005). Unobservable selection and coefficient stability: theory and evidence. *Journal of Business and Economic Statistics 23*(1), 1–19.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy 113*(1), 151–184.

Andrews, D. W. and X. Shi (2017). Inference based on many conditional moment inequalities. *Journal of Econometrics 196*(2), 275–287.

Athey, S., R. Chetty, and G. W. Imbens (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. arXiv preprint arXiv:2006.09676v1.

Backhoff-Veraguas, J., M. Beiglböck, and G. Pammer (2019). Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations 58*(6), 1–28.

Barber, C. B., D. P. Dobkin, and H. Huhdanpaa (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS) 22*(4), 469–483.

Bontemps, C. and T. Magnac (2017). Set identification, moment restrictions, and inference. *Annual Review of Economics 9*, 103–129.

Buchinsky, M., F. Li, and Z. Liao (2022). Estimation and inference of semiparametric models using data from several sources. *Journal of Econometrics 226*(1), 80–103.

Cárcamo, J., A. Cuevas, and L.-A. Rodríguez (2020). Directional differentiability for supremum-type functionals: Statistical applications. *Bernoulli 26*(3), 2143–2175.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). Intersection bounds: estimation and inference. *Econometrica 81*(2), 667–737.

Chetverikov, D., A. Santos, and A. M. Shaikh (2018). The econometrics of shape restrictions. *Annual Review of Economics 10*(1), 31–63.

Cross, P. J. and C. F. Manski (2002). Regressions, short and long. *Econometrica 70*(1), 357–368.

Davydov, Y. A., M. A. Lifshits, and N. V. Smorodina (1998). *Local properties of distributions of stochastic functionals*. American Mathematical Society.

De la Cal, J. and J. Cárcamo (2006). Stochastic orders and majorization of mean order statistics. *Journal of Applied Probability 43*(3), 704–712.

Del Barrio, E., E. Giné, and C. Matrán (1999). Central limit theorems for the wasserstein distance between the empirical and the true distributions. *Annals of Probability 31*, 1009–1071.

D'Haultfoeuille, X., C. Gaillac, and A. Maurel (2021). Rationalizing rational expectations: Characterizations and tests. *Quantitative Economics 12*(3), 817–842.

D'Haultfœuille, X., C. Gaillac, and A. Maurel (2023). Partially linear models under data combination. arXiv preprint arXiv:2204.05175.

Diegert, P., M. A. Masten, and A. Poirier (2022). Assessing omitted variable bias when the controls are endogenous. arXiv preprint arXiv:2206.02303.

Embrechts, P. and R. Wang (2015). Seven proofs for the subadditivity of expected shortfall. *Dependence Modeling 3*(1), 126–140.

Fan, Y., R. Sherman, and M. Shum (2014). Identifying treatment effects under data combination. *Econometrica 82*(2), 811–822.

Fan, Y., R. Sherman, and M. Shum (2016). Estimation and inference in an ecological inference model. *Journal of Econometric Methods 5*(1), 17–48.

Firpo, S., A. F. Galvao, and T. Parker (2023). Uniform inference for value functions. *Journal of Econometrics Forthcoming*.

Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.

Galichon, A. and M. Henry (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies 78*(4), 1264–1298.

Garcia, J., J. Heckman, L. D.E., and M. Prados (2020). Quantifying the life-cycle benefits of an influential early-childhood program. *Journal of Political Economy 128*(7), 2502–2541.

Gozlan, N., C. Roberto, P.-M. Samson, Y. Shu, and P. Tetali (2018). Characterization of a class of weak transport-entropy inequalities on the line. *Annales de l'IHP 54*(3), 1667–1693.

Hanushek, E. A., L. Kinne, P. Lergetporer, and L. Woessmann (2021). Culture and student achievement: The intertwined roles of patience and risk-taking. *Economic Journal 132*(646), 2290–2307.

Hiriart-Urruty, J.-B. and C. Lemaréchal (2012). *Fundamentals of convex analysis*. Springer Science & Business Media.

Horowitz, J. L. and C. F. Manski (1995). Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society 63*(2), 281–302.

Hwang, Y. (2022). Bounding omitted variable bias using auxiliary data with an application to estimate neighborhood effects. SSRN 3866876.

Manski, C. F. (2018). Credible ecological inference for medical decisions with personalized risk assessment. *Quantitative Economics 9*(2), 541–569.

Masten, M. A. and A. Poirier (2018). Identification of treatment effects under conditional partial independence. *Econometrica 86*(1), 317–351.

Matzkin, R. (1994). Restrictions of economic theory in nonparametric methods. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics, Volume 4*, Volume 4 of *Handbook of Econometrics*, pp. 2523–58. Elsevier.

Milgrom, P. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *Econometrica 70*(2), 583–601.

Molinari, F. (2020). Microeconometrics with partial identification. In S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin (Eds.), *Handbook of Econometrics, Volume 7A*, Volume 7 of *Handbook of Econometrics*, pp. 355–486. Elsevier.

Molinari, F. and M. Peski (2006). Generalization of a result on "regressions, short and long". *Econometric Theory 22*(1), 159–163.

Neal, D. A. and W. R. Johnson (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy 104*(5), 869–895.

Olivetti, C. and M. D. Paserman (2015). In the name of the son (and the daughter): Intergenerational mobility in the united states, 1850-1940. *American Economic Review 105*(8), 2695–2724.

Oster, E. (2019). Unobservable selection and coefficient stability: theory and evidence. *Journal of Business and Economic Statistics 37*(2), 187–204.

Pacini, D. (2019). Two-sample least squares projection. *Econometric Reviews 38*(1), 95–123.

Piatek, R. and P. Pinger (2016). Maintaining (locus of) control? data combination for the identification and inference of factor structure models. *Journal of Applied Econometrics 31*, 734–755.

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer Science & Business Media.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*(2), 186–199.

Ridder, G. and R. Moffitt (2007). The econometrics of data combination. *Handbook of Econometrics 6*, 5469–5547.

Robbins, M. W., S. Bauhoff, and L. Burgette (2022). Data fusion for predicting long-term program impacts. arXiv preprint arXiv:2205.01904v1.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Rothstein, J. and N. Wozny (2013). Permanent income and the black-white test score gap. *Journal of Human Resources 48*(3), 510–544.

Santavirta, T. and J. Stuhler (2022). Name-based estimators of intergenerational mobility. Mimeo.

Strassen, V. (1965). The existence of probability measures with given marginals. *The Annals of Mathematical Statistics 36*(2), 423–439.

Sundaram, R. K. (1996). *A first course in optimization theory*. Cambridge university press.

Tripathi, G. (2000). Local semiparametric efficiency bounds under shape restrictions. *Econometric Theory 16*(5), 729–739.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer.

Wijsman, R. A. (1966). Convergence of sequences of convex sets, cones and functions. ii. *Transactions of the American Mathematical Society 123*(1), 32–45.

A Additional theoretical results

A.1 Measurement errors

We have assumed so far that the outcome and covariates are perfectly observed. However, measurement errors are pervasive in survey data. We now explore the robustness of the identified set proposed earlier to measurement errors on the outcome and covariates, which we denote by Y^* and X^*. Specifically, consider a situation where both the covariates and the outcome are measured with error, such that:

$$\begin{cases} X = X^* + \xi_X, \quad \xi_X \perp\!\!\!\perp X^*, \\ Y = Y^* + \xi_Y, \quad \xi_Y \perp\!\!\!\perp (X^*, Y^*). \end{cases} \tag{19}$$

We introduce a new set, $\mathcal{B}^*$, which is defined as the original identified set $\mathcal{B}$ after replacing the observed measurement error-ridden covariates and outcome (X, Y) by their latent counterparts (X^*, Y^*).

Proposition 8. *If Assumption 1 is satisfied with (X, Y) replaced by (X^*, Y^*), (19) holds and for all $\beta \in \mathcal{B}^*$, $\xi_{Y_0} \succ_{cv} \xi'_{X_0}\beta$, then $\mathcal{B}^* \subset \mathcal{B}$.*

The proof is in our supplementary material. This proposition establishes that the identified set is robust to measurement errors in the following sense: if (centered) measurement errors on the outcome Y^* second-order stochastically dominate those on the linear index $X_0^{*\prime}\beta$ for all $\beta \in \mathcal{B}^*$, the identified set $\mathcal{B}$ based on the observed covariates X and outcome Y always contains the true value of the parameter of interest.[15] To better understand the above domination condition, suppose that $p = 1$, $\xi_Y \sim \mathcal{N}(0, \sigma_Y^2)$ and $\xi_X \sim \mathcal{N}(0, \sigma_X^2)$. Then, recalling that any $\beta \in \mathcal{B}^*$ satisfies the variance restriction $\beta^2 V(X^*) \leq V(Y^*)$, a sufficient condition for the dominance condition $\xi_{Y_0} \succ_{\text{cv}} \xi_{X_0}\beta$ is $\sigma_Y^2 \geq [V(Y^*)/V(X^*)]\sigma_X^2$. In our application for instance, Y^* and X^* are the log earnings of fathers and sons (or sons-in-law), respectively, so $V(Y^*) \simeq V(X^*)$ and $\sigma_Y^2 \simeq \sigma_X^2$ seem credible. This suggests that the key domination condition from Proposition 8 is likely to hold in this context.

[15] This result and underlying assumptions are closely related to the robustness to measurement errors on the beliefs of the test of rational expectations proposed in D'Haultfoeuille et al. (2021) (Subsection 2.2.4).

A.2 Identification of a model with interaction terms

Let $X_c = (X_{1,c}, X_{-1,c})$ and $X_{nc} = (X_{1,nc}, X_{-1,nc})$. We consider here the following model

$$E(Y|X) = f(X_c) + X'_{nc}\beta_0 + X_{1,nc}X_{1,c}\delta_0.$$

First define, for $x \in \text{Supp}(X_{1c})$ and $q \in \mathbb{R}^p$ ($q \neq 0_p$),

$$\overline{S}_x(q, F_{Y,X_c}, F_X) = \inf_{x_{-1,c} \in \text{Supp}(X_{-1,c}|X_{1,c}=x)} S(F_{Y|X_{-1,c}=x_{-1,c},X_{1,c}=x}, F_{X'_{nc}q|X_{-1,c}=x_{-1,c},X_{1,c}=x})$$

$$\mathcal{B}_x = \left\{ \lambda q : \, q \in \mathcal{S}, 0 \leq \lambda \leq \overline{S}_x(q, F_{Y,X_c}, F_X) \right\}.$$

Proposition 2 applied to the subpopulation $X_{1,c} = x$ implies that for all $x \in \text{Supp}(X_{1,c})$, $\beta_0 + x\delta_0 e_1 \in \mathcal{B}_x$, where $e_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^p$. Because the converse also holds, the identified set $\mathcal{B}^{\delta\beta}$ of (δ_0, β_0) is

$$\mathcal{B}^{\delta\beta} = \{(\delta, \beta) : \forall\, x \in \text{Supp}(X_{1,c}),\ \beta + x\delta e_1 \in \mathcal{B}_x\}. \tag{20}$$

The sets $\mathcal{B}_x$ are convex and include 0_p. Hence, $\mathcal{B}^{\delta\beta}$ is convex too, and also includes 0_{p+1}. Moreover, because $\mathcal{B}_x$ are compact, any $(\delta, \beta) \in \mathcal{B}^{\delta\beta}$ satisfies, for any $(x, x') \in \text{Supp}(X_{1,c})^2$, $x \neq x'$,

$$|\delta||x - x'| \leq \|\beta + x\delta e_1\| + \|\beta + x'\delta e_1\| \leq M_x + M_{x'}, \tag{21}$$

for some $M_x, M_{x'} > 0$. Moreover,

$$\|\beta\| \leq \|\beta + x\delta e_1\| + |x||\delta| \leq M_x + \frac{|x|}{|x - x'|}(M_x + M_{x'}),$$

which implies that $\mathcal{B}^{\delta\beta}$ is also compact. Thus, $\mathcal{B}^{\delta\beta}$ can also be described by its radial function, which we denote by $S(q, F_{Y,X_c}, F_X)$. Moreover, it follows from (20) that

$$S(q, F_{Y,X_c}, F_X) = \inf_{x \in \text{Supp}(X_{1,c})} \overline{S}_x(q_{-1} + xq_1 e_1, F_{Y,X_c}, F_X).$$

A.3 Test for point-identification

We develop here a statistical test that can be used to check whether $\beta_0 \in \partial\mathcal{B}$. Following the discussion in Subsection 2.3.4, this boils down to testing for

$$H_0 : \ S(F_{Y_{v0}}, F_{X'_{v0}\beta_v}) = 1 \quad \text{against} \quad H_1 : S(F_{Y_{v0}}, F_{X'_{v0}\beta_v}) > 1, \tag{22}$$

where we recall that the joint distribution of the validation data (X_v, Y_v) is observed and $\beta_v = V(X_v)^{-1}\text{cov}(X_v, Y_v)$. We consider a statistical test based on i.i.d. data $(X_{vi}, Y_{vi})_{i=1,\dots,n}$. The test statistic is

$$T = b_n^{1/2}\left(S_\varepsilon(\widehat{F}_{Y_{v0}}, \widehat{F}_{X'_{v0}\widehat{\beta}_v}) - 1\right),$$

where $\widehat{\beta}_v$ is the OLS estimator of β_v. The critical value is then $q_{1-\alpha}(T^*)$, the quantile of order $1-\alpha$ (defined conditional on the data) of

$$T^* = n^{1/2}\left(S_\varepsilon(\widehat{F}^*_{Y_{v0}}, \widehat{F}^*_{X'_{v0}\widehat{\beta}^*_v}) - S_\varepsilon(\widehat{F}_{Y_{v0}}, \widehat{F}_{X'_{v0}\widehat{\beta}_v})\right),$$

where $\widehat{F}^*_{Y_{v0}}$, $\widehat{F}^*_{X'_{v0}q}$ and $\widehat{\beta}^*_v$ are the subsampling counterpart of $\widehat{F}_{Y_{v0}}$, $\widehat{F}_{X'_{v0}q}$ and $\widehat{\beta}_v$, respectively. We establish the asymptotic properties of the test under the following assumption.

Assumption 7. *$E[\|X_v\|^{2+\delta}] < \infty$ for some $\delta > 0$, $E[Y_v^2] < \infty$, $\beta_v \neq 0$ and $S(F_{Y_{v0}}, F_{X'_{v0}\beta_v}) = S_\varepsilon(F_{Y_{v0}}, F_{X'_{v0}\beta_v})$. Also, there exists $\mathcal{V} \subset \mathcal{S}$, compact and including a ball of positive radius centered at $\beta_v/\|\beta_v\|$, and $\varepsilon' \in (0, \varepsilon)$ such that for all $(\alpha, \alpha') \in [\varepsilon', 1-\varepsilon']^2$, there exists $c > 0$ and a strictly increasing and continuous function m such that $m(0) = 0$ and*

$$\begin{aligned}
\inf_{q\in\mathcal{V}} \left|F^{-1}_{X'q}(\alpha') - F^{-1}_{X'q}(\alpha)\right| &> c|\alpha' - \alpha|,\\
\sup_{q\in\mathcal{V}} \left|F^{-1}_{X'q}(\alpha') - F^{-1}_{X'q}(\alpha)\right| &< m(|\alpha' - \alpha|),\\
\left|F^{-1}_{Y}(\alpha') - F^{-1}_{Y}(\alpha)\right| &< m(|\alpha' - \alpha|).
\end{aligned}$$

Up to the condition $S(F_{Y_{v0}}, F_{X'_{v0}\beta_v}) = S_\varepsilon(F_{Y_{v0}}, F_{X'_{v0}\beta_v})$ on which we come back below, Assumption 7 is very close to the first part of Assumption 4, but it is weaker as we require that it holds over $\mathcal{V}$ instead of $\mathcal{S}$.

Proposition 9. *Suppose that $b_n \to \infty$, $b_n/n \to 0$ and Assumptions 1-2 and 7 hold. Then:*

1. *If H_0 in (22) holds, $\lim_{n\to\infty} P(T > q_{1-\alpha}(T^*)) = \alpha$.*
2. *If H_1 in (22) holds, $\lim_{n\to\infty} P(T > q_{1-\alpha}(T^*)) = 1$.*

The proof is in our supplementary material. Note that if H_0 holds but $S(F_{Y_{v0}}, F_{X'_{v0}\beta_v}) < S_\varepsilon(F_{Y_{v0}}, F_{X'_{v0}\beta_v})$, T will tend to infinity and H_0 will be rejected. Because we are testing here the validation of the tail condition described above, failing to reject H_0 under the alternative is more of an issue than wrongly rejecting H_0. Thus, potential over-rejection is arguably not as problematic as in other more standard contexts, such as testing the null of no effect of a treatment.

B Application to the black-white wage gap

We apply our method to estimate the black-white wage gap among young males in the United States using the 1979 panel of the National Longitudinal Survey of Youth (NLSY79), revisiting the seminal work of Neal and Johnson (1996) on this question. Considering the same restrictions as Neal and Johnson (1996) leads to a sample of size $n = 1,675$.[16] We focus on the following model :

$$Y = \gamma_{c,0} + X_{c,1}\gamma_{c,1} + X_{c,2}\gamma_{c,2} + X_{nc}\beta_{nc} + \epsilon, \quad E\left[\epsilon | X_c, X_{nc}\right] = 0,$$

where Y is the mean log wage in 1990-1991, X_{nc} denotes the AFQT and $X_{c,k}$, $k = 1, 2$ are dummy variables for being black or Hispanic. While (Y, X_c, X_{nc}) is jointly observed in the NLSY79 dataset, we proceed in the following as if AFQT, which is used in Neal and Johnson (1996) to control for pre-market factors, was not observed jointly with wages. This setup, which mimics the data environments in several other countries, allows us to directly compare the confidence intervals based on our partial identification approach with the ones obtained from the oracle OLS specification.

Results in Table 2 below show the effect on our bounds when we impose different sets of constraints, namely i) a negative sign constraint on the coefficient $\gamma_{c,1}$ associated with the black indicator as well as a positive sign constraint on the coefficient β_{nc} associated with the AFQT, ii) the latter constraints combined with the constraint $\underline{R}^2 \geq 1.3R_s^2$, and iii) the sign constraints i) combined with a less conservative bound $\underline{R}^2 \geq 2R_s^2$. Focusing on the main coefficient of interest $\gamma_{c,1}$, these results indicate that imposing these constraints on the $\underline{R}^2$ results in an identified set and confidence interval that are quite informative. Notably, the lower bound of the confidence interval

[16] We refer the reader to Neal and Johnson (1996) for a detailed discussion on the data.

is equal to $-.25$ and $-.2$ respectively in cases ii) and iii), against $-.17$ (i.e. a 17 log points wage penalty) for the OLS estimator. Taken together, these results show that our method is able to deliver confidence intervals that are very informative in practice.

	OLS	DGM			
Constraints		Without	With signs constraints		
			Only	And $\underline{R}^2 \geq 1.3R_s^2$	And $\underline{R}^2 \geq 2R_s^2$
	(1)	(2)	(3)	(4)	(5)
Omitted variable X_{nc}					
AFQT	0.150	[-0.437,0.437]	[0,0.154]	[0.045,0.154]	[0.082,0.154]
CI	[0.11,0.19]	[-0.522,0.522]	[0,0.215]	[0.004,0.211]	[0.010,0.207]
Common variables X_c					
Black	-0.076	[-0.664,0.318]	[-0.173,0]	[-0.123,0]	[-0.081,0]
CI	[-0.171,0.02]	[-0.847,0.507]	[-0.304,0]	[-0.247,0]	[-0.199,0]
Hispanic	0.016	[-0.334,0.266]	[-0.034,0.071]	[-0.003,0.071]	[0.022,0.071]
CI	[-0.083,0.116]	[-0.506,0.450]	[-0.197,0.226]	[-0.186,0.225]	[-0.174,0.223]

Notes: Y is average log wage in 1990 and 1991, X_{nc} is the AFQT, X_c are dummies for being Black or Hispanic. The sample size is $n = 1,675$, which is randomly split in two to artificially create a dataset where we observe (Y, X_c) and another one with (X_{nc}, X_c). The first column presents the OLS estimates on the full dataset, where the 95% CI have been multiplied by $\sqrt{2}$ to make it comparable with the DGM procedure using only half of it. The second column (2) presents the DGM estimates without constraints. Column (3) is the DGM estimates with a negative sign constraint on the coefficient of Black and a positive one on the coefficient of AFQT. Column (4) and (5) gather the DGM estimates with the latter constraints plus a lower bound constraint R^2 of the long regression: $\underline{R}^2 \geq rR_s^2$, with respectively $r = 1.3$ and $r = 2$. The R^2 on the short and long regressions are respectively 0.051 and 0.142.

Table 2: Bounds on the wage gap under different constraints for NLSY79

Online Appendix

C Monte Carlo simulations

In this section we study the finite sample performances of our inference method through Monte Carlo simulations. We first consider the baseline case where no common regressor is available, before evaluating the performance of our method in the presence of a common regressor. Finally, we discuss the computational time of our procedure compared to a many moment inequality-based alternative.

C.1 Univariate case without common regressors

We first explore the finite sample performances of our inference method with $p = 1$ and no X_c, considering the following DGP:

$$Y = X_{nc}\beta_0 + U, \quad \beta_0 = 1, \; X_{nc} \perp\!\!\!\perp U.$$

Then, we either assume that $X_{nc} \sim \mathcal{N}(0, 1.5)$ and $U \sim \mathcal{N}(0, 1)$, referred to in the following as the normal case, or $X_{nc} \sim \Gamma(1, 2)$ and $U \sim \Gamma(0.4, 2)$, which we refer to as the gamma case.

We compare the finite sample performances of our inference method with those based on Andrews and Shi, 2017, henceforth AS. Specifically, recall from (5) above that

$$\mathcal{B} = \{\beta \in \mathbb{R}^p : \; E\left[\max(0, Y_0 - t)\right] \geq E\left[\max(0, X_{nc0}'\beta - t)\right] \; \forall t \in \mathbb{R}\}.$$

Hence, $\mathcal{B}$ is characterized by infinitely many moment inequalities. We then construct confidence regions for β_0 by inverting tests that these moment inequalities hold.[17]

In Table 3 below, we report the average bounds, across all 500 simulations, of the estimated identified sets and the 95% confidence intervals associated with each of the

[17] These tests involve several tuning parameters. Following the recommendation of AS (and using their notation), we fix $\epsilon = 0.05$ and $\eta = 10^{-6}$. To fix b_0 and κ, we follow the same procedure as in D'Haultfoeuille et al. (2021), which yields $b_0 = 0.5$ and $\kappa = 10^{-4}$. To construct a confidence region on β_0, we first fix a few directions $(q_1, ..., q_n)$ in $\mathcal{S}$. Then, for $q = q_k$, we compute by a bisection method the maximal $\lambda \in \mathbb{R}^+$ such that the test of the moment inequalities at $\beta = \lambda q$ is not rejected.

five different sample sizes (Column "Bounds") obtained with our method ("DGM") and by applying Andrews and Shi (2017) ("AS"). In order to isolate sampling uncertainty, we report for each sample size and separately for our method and AS what we call the excess length ("Ex. length"), namely the mean difference between the length of the confidence sets and that of the identified set. We also report the coverage rates across simulations ("Coverage"). Finally, we report the average, across all simulations, of the estimates of the identified set $\mathcal{B}_{\varepsilon(q)}$, where $\varepsilon(q)$ is given by (14) and thus varies from one simulation to another.

	DGM				**AS**		
Sample size	Bounds	Ex. length	Coverage	$\widehat{\mathcal{B}}_{\varepsilon(q)}$	Bounds	Ex. length	Coverage
Normal							
Identified set	[-1.202,1.202]				[-1.202,1.202]		
400	[-1.305,1.307]	0.208	0.938	[-1.202,1.202]	[-1.374,1.367]	0.337	0.983
800	[-1.280,1.280]	0.156	0.942	[-1.202,1.202]	[-1.329,1.328]	0.253	0.985
1,200	[-1.266,1.267]	0.129	0.940	[-1.202,1.202]	[-1.301,1.301]	0.198	0.978
2,400	[-1.246,1.247]	0.089	0.948	[-1.202,1.202]	[-1.268,1.270]	0.134	0.975
4,800	[-1.234,1.235]	0.065	0.936	[-1.202,1.202]	[-1.251,1.250]	0.097	0.980
Gamma							
Identified set	[-0.025,1.046]				[-0.025,1.046]		
400	[-0.758,1.357]	1.043	1	[-0.464,1.287]	[-0.538,1.343]	0.809	1
800	[-0.603,1.302]	0.834	0.996	[-0.340,1.257]	[-0.466,1.313]	0.707	1
1,200	[-0.546,1.28]	0.754	1	[-0.293,1.247]	[-0.438,1.302]	0.668	1
2,400	[-0.458,1.243]	0.629	1	[-0.237,1.220]	[-0.391,1.277]	0.596	1
4,800	[-0.391,1.213]	0.532	1	[-0.199,1.197]	[-0.362,1.258]	0.548	1

Notes: results obtained with 500 simulations. Column "Bounds" reports either the identified set or the average of the bounds of the 95% confidence intervals over simulations. "Ex. length" is the excess length, i.e. the average length of the confidence region minus the length of the identified set. Column "Coverage" displays the minimum, over $\beta \in \mathcal{B}$, of the estimated probability that $\beta \in \mathrm{CR}_{1-\alpha}(\beta_0)$. Column "$\widehat{\mathcal{B}}_{\varepsilon(q)}$" displays the average, across all simulations, of the estimates of the identified set $\mathcal{B}_{\varepsilon(q)}$, where $\varepsilon(q)$ is given by (14). We use 1,000 subsampling (resp. bootstrap) replications to compute the confidence intervals for the DGM (resp. AS) method.

Table 3: Finite sample performances for $p = 1$

A couple of remarks are in order. First, as expected, the 95% confidence intervals shrink with the sample sizes n. For both DGPs and all sample sizes, comparing the

identified set with the confidence intervals indicates that identification uncertainty clearly dominates sampling uncertainty. This is especially striking for the normal case, which yields a substantially wider identified set, but also holds in the gamma case, where the regressor X_{nc} has thicker tails. In particular, considering the excess length in the normal case, the confidence set is only between 8.6% (for $n = 400$) and 2.7% (for $n = 4,800$) wider than the identified set. In the gamma case, the confidence set ranges between 20.7% and 14.8% larger than the (regularized) identified set ($\mathcal{B}_\varepsilon$).

Second, the coverage of our confidence intervals is good: coverage rates are always larger than 93.6%. Third, our inference method generally performs similarly or better than AS, delivering consistently tighter confidence sets for sufficiently large sample size. For example, in the normal case, the excess length of the confidence set is reduced by around 30% to 39% depending on the sample sizes. In the gamma case, the two methods are close, AS doing slightly better only for sample sizes smaller than $n = 4,800$. These results are consistent with our inference method exploiting the specific geometric structure of the identified set. This could also be due to the fact that we do not need to bear the cost, in terms of statistical power, of incorporating potentially many non-binding inequality constraints.

Finally, the good finite sample performances of our inference method offers supporting evidence that our choice of the regularization parameter $\varepsilon(q)$, given by (14) and motivated in Section 3.3 above, is appropriate. In the normal case where $\mathcal{B}_\varepsilon = \mathcal{B}$ for all ε, $\varepsilon(q)$ remains close to 0.5 for all sample sizes. In contrast, in the gamma case where the minimum of $R(\cdot, F_{Y_0}, F_{X_0'q})$ is reached at $\varepsilon = 0$ for both $q = 1$ and $q = -1$, $\varepsilon(q)$ tends to 0 as n tends to infinity. Overall, the results suggest that the chosen $\varepsilon(q)$ achieves a good balance between identification (a large ε leading to an increase in $\mathcal{B}_\varepsilon$) and statistical uncertainty (a small ε leading to more volatility when estimating S_ε and thus larger quantiles $\widehat{c}_{\alpha,\varepsilon}$).

C.2 Multivariate case without common regressor

We now consider the multivariate case ($p = 2$) with the following DGP:

$$Y \;=\; \gamma_0 + X_{nc}'\beta_0 + U,\; U|X_{nc} \sim \mathcal{N}(0,4). \tag{23}$$

We set the coefficients as follows: $\gamma_0 = -0.1$, $\beta_{0,1} = 1$, and $\beta_{0,2} = 1$. The variables X_{nc} follow a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}.$$

We report in Table 4 below the performances of our inference method, applied to the first component of β_0, for the same sample sizes as above, along with the identified set of the projection. These results were obtained using 500 simulations. We restrict to the first component of β_0 as the results are very similar for the second component. The main takeaway of this table is that our inference method exhibits similar finite-sample performances to the ones discussed in the univariate ($p = 1$) normal case. In particular, the excess length of the confidence sets relative to the identified set tends to be quite small, and declines as n gets larger.

Average	Bounds	Excess length	Coverage
Identified set	[-2.367, 2.367]		
Sample size			
400	[-2.599, 2.599]	0.465	0.94
800	[-2.555, 2.554]	0.376	0.962
1,200	[-2.523, 2.522]	0.312	0.96
2,400	[-2.496, 2.497]	0.26	0.982
4,800	[-2.475, 2.474]	0.217	0.986

Notes: results obtained with 500 simulations. Column "Bounds" reports either the identified set or the average of the bounds of the 95% confidence intervals over simulations. "Excess length" is the average length of the confidence region minus the length of the identified set. Column "Coverage" displays the minimum, over $\beta_1 \in \mathcal{B}_1$, of the estimated probability that $\beta_1 \in \mathrm{CI}_{1-\alpha}(\beta_{0,1})$. We use 200 subsampling replications to compute the confidence intervals.

Table 4: Finite sample performances for $\beta_{0,1}$ with $p = 2$

C.3 Case with a common regressor and possible constraints

We now examine the performances of our inference method in the presence of a common regressor. Namely, we consider the DGP:

$$Y = X_c\gamma_0 + X_{nc}\beta_0 + U,\ U|X \sim \mathcal{N}(0,4).$$

We set the coefficients as follows: $\gamma_0 = 0.3$ and $\beta_0 = 1$. The covariates are transformations of $(N_1, N_2)'$, which follows a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1.5 \end{pmatrix}.$$

Specifically, the common regressor is given by $X_c = \mathbb{1}\{N_1 \leq 0.3\}$, and the regressors observed in one of the datasets only are such that $X_{nc} = N_2$.

We report in Table 5 the performances of our inference method applied to the parameters β_0 and γ_0 along with the identified sets, with or without imposing the sign constraint $\gamma_0 \geq 0$. For β_0, coverage ranges between 95.4% and 97%. Similar to the baseline case without common regressors, the excess length of the confidence interval relative to the identified set declines as n grows, and becomes quite small for the largest sample sizes. For instance, for $n = 4,800$, our confidence interval is only 4% larger than the identified set, highlighting again the limited role of sampling uncertainty in this context. The sign constraint reduces considerably the confidence interval, allowing to reject that $\beta_0 = 0$ for all the considered sample sizes.

Similar comments apply to the results on γ_0. The coverage rate is always over 96.4%. The length of our confidence intervals is between 8% and 30.5% larger than the one of the unconstrained identified set. Note that the upper bounds of our confidence intervals on γ_0 are larger with the sign constraint than without it as in the former case we use critical values based on quantiles of order $\alpha/2$ and $1-\alpha/2$ to ensure that the confidence region $\mathrm{CR}^{con}_{1-\alpha}(\beta_0)$ is asymptotically conservative.

Table 6 illustrates the performances of our inference method on β_0, using the same DGP as above except for γ_0 which is set equal to zero, and compare them to the TSTSLS confidence intervals which are valid under this particular DGP. We implement our inference method without imposing the constraint that $\gamma_0 = 0$. A couple

of remarks are in order. First, the coverage with our method ranges between 94.8% and 99.4%, with the exception of one case ($n = 400$ and the constraint $\gamma_0 \geq 0$, where the coverage is 92.2%). Second, while the bounds obtained without imposing the sign constraint are substantially larger than the TSTSLS ones, which rely on the constraint $\gamma_0 = 0$, the non-negativity constraint $\gamma_0 \geq 0$ does result in significantly tighter confidence intervals. In particular, the lower bounds on β_0 become close to the TSTSLS ones, and exclude 0.

	Without sign constraint			With the constraint $\gamma_0 \geq 0$		
Average	Bounds	Excess length	Coverage	Bounds	Excess length	Coverage
	Parameter β_0					
Identified set	[-2.125, 2.125]			[0.768, 2.125]		
Sample size						
400	[-2.445, 2.445]	0.640	0.966	[0.376, 2.495]	0.761	0.944
800	[-2.339, 2.341]	0.430	0.970	[0.408, 2.376]	0.611	0.978
1,200	[-2.297, 2.300]	0.347	0.962	[0.460, 2.329]	0.512	0.994
2,400	[-2.247, 2.251]	0.248	0.966	[0.541, 2.273]	0.374	0.982
4,800	[-2.206, 2.213]	0.170	0.954	[0.603, 2.229]	0.268	0.976
	Parameter γ_0					
Identified set	[-3.738, 1.754]			[0, 1.754]		
Sample size						
400	[-4.578, 2.590]	1.676	0.98	[0, 2.729]	0.975	0.996
800	[-4.306, 2.348]	1.162	0.984	[0, 2.448]	0.694	0.998
1,200	[-4.197, 2.214]	0.919	0.990	[0, 2.296]	0.542	0.996
2,400	[-4.062, 2.076]	0.646	0.976	[0, 2.135]	0.381	0.988
4,800	[-3.967, 1.990]	0.465	0.976	[0, 2.032]	0.278	0.992

Notes: results obtained with 500 simulations. Column "Bounds" reports either the identified set or the average of the bounds of the 95% confidence intervals over simulations. "Excess length" is the average length of the confidence region minus the length of the identified set. Column "Coverage" displays the minimum of the estimated probability that $\gamma \in \mathrm{CR}_{1-\alpha}(\gamma_0)$. We use 1,000 subsampling replications to compute the confidence intervals.

Table 5: Finite sample performances for β_0 and γ_0 with and without sign constraints

	Without sign constraint			With the constraint $\gamma_0 \geq 0$			TSTSLS
Average	Bounds	Excess length	Coverage	Bounds	Excess length	Coverage	Bounds
Identified set	[-2.125, 2.125]			[1, 2.125]			[1, 1]
Sample size							
400	[-2.411, 2.411]	0.571	0.980	[0.642, 2.457]	0.689	0.922	[0.634, 1.426]
800	[-2.342, 2.343]	0.434	0.980	[0.67, 2.377]	0.582	0.964	[0.733, 1.285]
1,200	[-2.296, 2.296]	0.341	0.968	[0.709, 2.324]	0.49	0.982	[0.791, 1.241]
2,400	[-2.241, 2.246]	0.236	0.968	[0.779, 2.267]	0.362	0.994	[0.844, 1.159]
4,800	[-2.200, 2.207]	0.157	0.948	[0.836, 2.223]	0.261	0.970	[0.891, 1.114]

Notes: results obtained with 500 simulations. Column "Bounds" reports either the identified set or the average of the bounds of the 95% confidence intervals over simulations. "Excess length" is the average length of the confidence region minus the length of the identified set. Column "Coverage" displays the minimum of the estimated probability that $\beta \in \text{CR}_{1-\alpha}(\beta_0)$. We use 1,000 subsampling replications to compute the confidence intervals.

Table 6: Finite sample performances for β_0 with one common regressor $\gamma_0 = 0$

C.4 Computational time

First, and following the discussion around Eq. (8), we compare our approach based on the radial function and the direct computation of the support function based on (8). We consider the DGP $Y = X_{nc}'\beta + \epsilon$, where $\beta = (1, \ldots, 1)' \in \mathbb{R}^p$, $X_{nc} \in \mathbb{R}^p$ with independent marginals $\mathcal{N}(0, 2.25)$ and $\epsilon | X_{nc} \sim \mathcal{N}(0, 1)$. We then compute $\sigma(\pm e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0})$ for $k = 1, ..., p$ on 100 samples of size $n = 2,000$. Our approach turns out to be 100 times faster when $p = 1$, because it avoids the double optimization, and 11 (resp. 4) times faster when $p = 2$ (resp. $p = 3$).[18]

Next, we examine the computational time of our inference method and that of AS when p, the dimension of X_{nc}, is equal to either 1 or 2, for the DGPs considered in Sections C.1 and C.2, respectively, and for the five different sample sizes considered above. Table 7 below reports the computational time for $CR_{1-\alpha}(\beta_0)$ when $p = 1$, and for the two confidence intervals $CI_{1-\alpha}(\beta_{0,1})$ and $CI_{1-\alpha}(\beta_{0,2})$ when $p = 2$.

[18] All the computational times are obtained for a single simulation using our companion R package, on an Intel Xeon Gold 6130 CPU 2.10GHz with 382Gb of RAM and a single core. For the support function approach (8), we use the Constrained Optimization by Linear Approximations (COBYLA) algorithm for solving the linear optimization program under nonlinear constraints.

Sample size	400	800	1,200	2,400	4,800
$p = 1$					
AS (s)	241.8	349.2	458.4	823.2	1137.0
DGM (s)	0.70	0.73	0.77	0.86	0.93
$p = 2$					
AS fast (min)	18.3	29.5	40.0	71.7	150.3
AS recommended (min)	177.8	296.5	393.5	702.8	1500.2
DGM (s)	18.9	19.8	20.8	23.9	30.6

Notes: The CPU time for the DGM method when $p = 2$ corresponds to the computation of the 4 projections associated to $\pm e_k$, $k = 1, 2$. For $p = 2$, the "AS fast" approximation uses 25 directions to evaluate the computational time of the AS based method. The average over 50 replications of the excess length between the confidence intervals obtained with 250 directions and 25 directions over the length of the confidence intervals obtained with 250 directions ("AS recommended") is 3.2% for $n = 1,200$. As in Sections C.1-C.2, we use 1,000 subsampling (resp. bootstrap) replications when $p = 1$ and 200 replications when $p = 2$ for the DGM (resp. AS) method.

Table 7: CPU time as function of sample size and dimension p of X_{nc}.

In the univariate case ($p = 1$), the computational gains of our method range from a factor of 342 to 1,217 compared to AS, for $n = 400$ and $n = 4,800$, respectively. While the computational time associated with our method increases with the sample size, it remains very modest (less than 1 second) for $n = 4,800$.

In the multivariate case ($p = 2$), we compare our method with two alternative implementations of the AS method. "AS fast" corresponds to an approximation of the confidence intervals for both components of β_0 that uses 25 directions in $\mathcal{S}$ to implement the method, while "AS recommended" corresponds to the computational time associated with 250 directions. Since our method does not rely on any numerical approximation of this kind (as we exactly compute $1/\inf_{q\in\mathbb{R}^p:q_k=1} 1/S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$), it is arguably more relevant to compare the computational times of our method and the "AS recommended" implementation. While the computational time of our method increases with p, it does remain tractable even with fairly large sample sizes, taking for instance 30.6 seconds only to run for $n = 4,800$. In the multivariate case also our method outperforms both implementations of the AS method. For instance, for

$n = 2,400$, our method runs 1,768 times faster than the recommended implementation of AS. In this case, computing $\varepsilon(q)$ for one direction with our method takes the same time as in the univariate case ($p = 1$). The main difference and computational bottleneck with $p > 1$ lies in the subsampling of the convex optimization in (13).

To conclude, our approach can be implemented at a very limited computational cost, and achieves in our context considerable computational gains relative to the alternative method of AS.

D Additional results on the application

Sample:	**1850-1870**	**1860-1880**	**1880-1900**	**1900-1920**	**1910-1930**
Baseline specification					
DGM, set	0.555	0.465	0.473	0.43	0.443
DGM, CI	0.614	0.517	0.532	0.483	0.499
Number of names X_c	225	261	382	514	598
Panel A: Robustness to the set of first names					
Threshold 0.005%					
DGM, set	0.555	0.465	0.473	0.43	0.443
DGM, CI	0.611	0.521	0.529	0.484	0.499
Number of names X_c	225	261	382	515	626
Threshold 0.02%					
DGM, set	0.555	0.465	0.493	0.511	0.477
DGM, CI	0.609	0.522	0.554	0.578	0.54
Number of names X_c	225	261	332	378	415
Panel B: Robustness to the choice of ε					
$\varepsilon/2$					
DGM, set	0.555	0.442	0.473	0.419	0.415
DGM, CI	0.612	0.49	0.534	0.471	0.467
Number of names X_c	224	259	380	512	596
2ε					
DGM, set	0.555	0.465	0.473	0.43	0.443
DGM, CI	0.616	0.52	0.532	0.483	0.5
Number of names X_c	224	259	380	512	596
Panel C: Restricting the sample to the selected first names					
DGM, set	0.556	0.465	0.472	0.43	0.442
DGM, CI	0.616	0.517	0.531	0.48	0.499
Sample sizes Y	33,796	46,296	73,961	99,874	111,126
Sample sizes X_{nc}	29,209	40,431	62,567	85,202	99,270

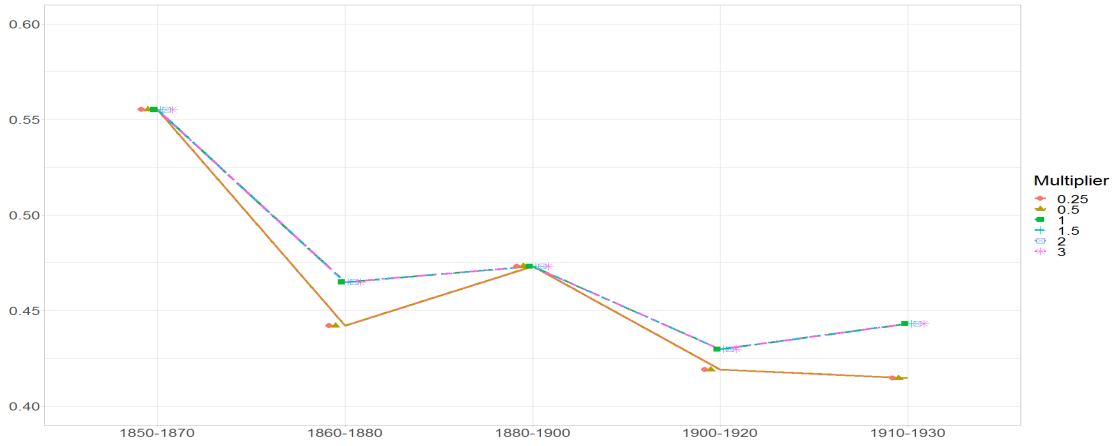
Notes: Y=son's log income. The baseline specification restricts X_c to be the dummies for the names appearing in the pooled dataset more than 0.01%, and 10 times in both datasets. Panel A presents the results when we consider names appearing more than 0.005% or 0.02% in the pooled dataset. In the baseline specification, the parameter ε is chosen according to the data-driven rule (14). Panel B presents the results when using 0.5 or 2 times this choice of ε. Panel C presents results when we restrict the samples to the selected names based on our rule in the baseline case. We report the corresponding modified sample sizes.

Table 8: Robustness checks for the upper bound on intergenerational income correlation for sons.

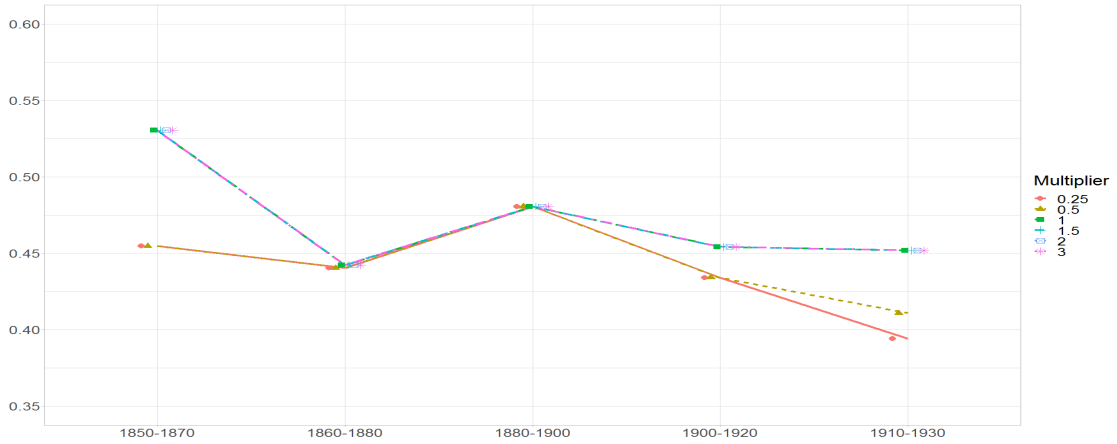
Sample:	**1850-1870**	**1860-1880**	**1880-1900**	**1900-1920**	**1910-1930**
Baseline specification					
DGM, set	0.531	0.442	0.481	0.454	0.452
DGM, CI	0.6	0.507	0.555	0.515	0.513
Number of names X_c	155	212	323	468	545
Panel A: Robustness to the set of first names					
Threshold 0.02%					
DGM, set	0.531	0.442	0.48	0.573	0.452
DGM, CI	0.604	0.503	0.554	0.658	0.514
Number of names X_c	155	212	316	397	430
Panel B: Robustness to the choice of ε					
$\varepsilon/2$					
DGM, set	0.455	0.44	0.481	0.434	0.411
DGM, CI	0.51	0.505	0.553	0.495	0.466
Number of names X_c	155	212	323	468	545
2ε					
DGM, set	0.531	0.442	0.481	0.454	0.452
DGM, CI	0.599	0.504	0.551	0.517	0.514
Number of names X_c	155	212	323	468	545
Panel C: Restricting the sample to the selected first names					
DGM, set	0.534	0.445	0.481	0.456	0.453
DGM, CI	0.61	0.509	0.554	0.52	0.514
Sample sizes Y	20,375	26,418	41,212	61,742	70,656
Sample sizes X_{nc}	27,096	37,231	57,474	81,551	94,706

Notes: Y=son-in-law's log income. The baseline specification restricts X_c to be the dummies for the names appearing in the pooled dataset more than 0.01%, and 10 times in both datasets. Panel A presents the results when we consider names appearing more than 0.02% in the pooled dataset. Results with considering names appearing more than 0.005% in the pooled dataset are identical to the baseline, hence note reported. In the baseline specification, the parameter ε is chosen according to the data-driven rule (14). Panel B presents the results when using 0.5 or 2 times this choice of ε. Panel C presents results when we restrict the samples to the selected names based on our rule in the baseline case. We report the corresponding modified sample sizes.

Table 9: Robustness checks for the upper bound on intergenerational income correlation for sons-in-law.



(a) For sons



(b) For sons-in-law

Note: the graphs display the value of $\overline{S}_{\varepsilon'}(q, \widehat{F}_{Y,X_c}, \widehat{F}_{X_{nc},X_c})$ for $\varepsilon' = c\varepsilon$, where ε is selected via (14) and $c \in \{0.25, 0.5, 1, 1.5, 2, 3\}$.

Figure 5: $\overline{S}_{\varepsilon}(q, \widehat{F}_{Y,X_c}, \widehat{F}_{X_{nc},X_c})$ for different ε.

E Proofs

E.1 Notation

We denote by $\mathcal{P}_q(\mathbb{R}^p)$ the set of Borel probability measures on $\mathbb{R}^p$ with q finite absolute moments. We assimilate herafter probability measures on $\mathbb{R}^p$ with their cdf, so we may write for instance $F \in \mathcal{P}_q(\mathbb{R}^p)$. We let W_1 denote the 1-Wasserstein distance and

recall that for $(F, G) \in \mathcal{P}_1(\mathbb{R})^2$,

$$W_1(F,G) = \inf_{U\sim F, V\sim G} E\left[|U-V|\right] = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt = \int_{-\infty}^{\infty} |F(t) - G(t)|dt. \tag{24}$$

We denote by $\ell^\infty(\mathcal{X})$ the space of bounded functions on $\mathcal{X}$ for the uniform metric. Finally, $g(x) \lesssim h(x)$ means that $g(x) \leq Ah(x)$ for some universal constant $A > 0$.

E.2 Theorem 1

Let $\mathcal{B}'$ denote the set on the right-hand side of (4). We first show that $\mathcal{B} \subset \mathcal{B}'$. Then, we show the other inclusion. Finally, we show the other properties of $\mathcal{B}$.

1. $\mathcal{B} \subset \mathcal{B}'$

Let F be such that $0 < \int x^2 dF(x) < \infty$ and $\int x dF(x) = 0$ and define $g(\alpha) = \int_\alpha^1 F^{-1}(t)dt$, for any $\alpha \in [0,1]$. Then, $g'(\alpha) = -F^{-1}(\alpha)$ is decreasing, which implies that g is concave. Moreover, $g(0) = g(1) = 0$. For some $\alpha \in (0,1)$, $F^{-1}(\alpha) \geq \int x dF(x) = 0$ so $g(\alpha) \geq (1-\alpha)F^{-1}(\alpha) \geq 0$. Assume that $g(\alpha) = 0$. Then, by concavity, $g(x) = 0$ for all $x \in [0,1]$. This implies that $F^{-1}(x) = 0$ for all $x \in (0,1)$, which contradicts $\int x^2 dF(x) > 0$. Thus, for all $\alpha \in (0,1)$, $g(\alpha) > 0$.

Then, because $E(X_0 X_0')$ is nonsingular, $\int_\alpha^1 F^{-1}_{X_0' q}(t)dt > 0$ for all $\alpha \in (0,1)$. This means that $0 \leq \lambda \leq S(F_{Y_0}, F_{X_0' q})$ is equivalent to

$$\int_\alpha^1 F^{-1}_{X_0'(\lambda q)}(t)dt \leq \int_\alpha^1 F^{-1}_{Y_0}(t)dt \quad \forall \alpha \in (0,1).$$

This, in turn, is equivalent to $F_{X_0'(\lambda q)}$ dominating F_{Y_0} at the second order (see, e.g. De la Cal and Cárcamo, 2006). Then, by definition of second-order stochastic dominance,

$$\mathcal{B}' = \{\beta \in \mathbb{R}^p : E[\phi(Y_0)] \geq E[\phi(X_0'\beta)] \quad \forall \phi \text{ convex}\}.$$

Now, for any $\beta \in \mathcal{B}$, there exists $(\widetilde{X}, \widetilde{Y})$ such that $E(\widetilde{Y}_0|\widetilde{X}_0) = \widetilde{X}_0'\beta$, $\widetilde{X} \overset{d}{=} X$ and $\widetilde{Y} \overset{d}{=} Y$. Then, for all convex function ϕ, we have, by Jensen's inequality,

$$E[\phi(\widetilde{Y}_0)|\widetilde{X}_0] \geq \phi(E[\widetilde{Y}_0|\widetilde{X}_0]) = \phi(\widetilde{X}_0'\beta).$$

As a result, $\beta \in \mathcal{B}'$.

2. $\mathcal{B}' \subset \mathcal{B}$

For any $(F, G) \in \mathcal{P}_1(\mathbb{R}) \times \mathcal{P}_1(\mathbb{R}^{p+1})$, let G_1 denote the first marginal of G and define

$$\begin{aligned} W_w(F, G_1) &:= \inf_{F_{U,V_1}: F_U = F, F_{V_1} = G_1} E\left[|V_1 - E[U|V_1]|\right], \\ W_c(F, G) &:= \inf_{F_{U,V_1,V_2}:\ F_U = F,\ F_{V_1,V_2} = G} E\left[|V_1 - E(U|V_1, V_2)|\right]. \end{aligned} \tag{25}$$

We first prove that $W_c(F, G) \leq W_w(F, G_1)$. To this end, let us define $c(x, H) = |x_1 - \int y dH(y)|$, for any $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}^p$ and $H \in \mathcal{P}_1(\mathbb{R})$. Because the function c satisfies the assumptions of Theorem 1.3. in Backhoff-Veraguas et al. (2019), we have

$$W_c(F, G) = \sup_{f \in \Phi_{\text{bel}}} \left\{ \int R_c(f)(x_1, x_2) dG(x_1, x_2) - \int f(y) dF(y) \right\},$$

where we define

$$\begin{aligned} \Phi_{\text{bel}} = &\Big\{ \psi : \mathbb{R} \to \mathbb{R} \text{ continuous s.t. } \exists (a, b, \ell, x_0) \in \mathbb{R}^4 : \\ &\forall x \in \mathbb{R},\ \ell \leq \psi(x) \leq a + b|x - x_0| \Big\}, \\ R_c(f)(x_1, x_2) = &\inf_{H \in \mathcal{P}_1(\mathbb{R})} \int f(y) dH(y) + \left| x_1 - \int y dH(y) \right|. \end{aligned}$$

Let $U \sim F$ and $V = (V_1, V_2) \sim G$. By definition of $R_c(f)$,

$$R_c(f)(x_1, x_2) \leq E[f(U)|V_1 = x_1] + |x_1 - E[U|V_1 = x_1]|.$$

As a result,

$$\begin{aligned} &\int R_c(f)(x_1, x_2) dG(x_1, x_2) - \int f(y) dF(y) \\ \leq &E[f(U)] + E\left[|V_1 - E(U|V_1)|\right] - \int f(y) dF(y) \\ = &E\left[|V_1 - E(U|V_1)|\right]. \end{aligned}$$

Since this holds for all (U, V_1) with $U \sim F$ and $V_1 \sim G_1$,

$$\int R_c(f)(x_1, x_2) dG(x_1, x_2) - \int f(y) dF(y) \leq W_w(F, G_1).$$

Taking the supremum over $f \in \Phi_{\text{bel}}$, we obtain $W_c(F, G) \leq W_w(F, G_1)$.

Now, let $\beta \in \mathcal{B}'$. By Strassen's theorem (Theorem 8 in Strassen, 1965; see also Theorem 3.1 in Gozlan et al., 2018), we have $W_w(F_{Y_0}, F_{X_0'\beta}) = 0$. As a result,

$W_c(F_{Y_0}, F_{X_0'\beta,X_0}) = 0$. Because the function c satisfies the assumptions of Theorem 1.2 in Backhoff-Veraguas et al. (2019), there exists a minimizer reaching the infimum in (25). This implies that there exist random variables $(\widetilde{Y}, \widetilde{X^\beta}, \widetilde{X})$ with $F_{\widetilde{Y}} = F_Y$, $F_{\widetilde{X^\beta},\widetilde{X}} = F_{X'\beta,X}$ and satisfying $\widetilde{X}_0^\beta = E[\widetilde{Y}_0|\widetilde{X}_0^\beta, \widetilde{X}_0]$. The equality $F_{\widetilde{X^\beta},\widetilde{X}} = F_{X'\beta,X}$ implies that $\widetilde{X^\beta} = \widetilde{X}'\beta$ almost surely. Then, $E[\widetilde{Y}_0|\widetilde{X}_0] = \widetilde{X}_0'\beta$ and in view of (2), $\beta \in \mathcal{B}$. The result follows.

3. Other properties of $\mathcal{B}$

Let $\widetilde{X}, \widetilde{Y}$ be independent variables such that $\widetilde{X} \overset{d}{=} X$ and $\widetilde{Y} \overset{d}{=} Y$. Then

$$E[\widetilde{Y}_0|\widetilde{X}_0] = E[\widetilde{Y}_0] = 0 = \widetilde{X}_0'0_p.$$

Hence, $0_p \in \mathcal{B}$. Now, let $(\beta_1, \beta_2) \in \mathcal{B}^2$ and $t \in [0,1]$. For any convex function ϕ, we have

$$\phi(X_0'(t\beta_1 + (1-t)\beta_2)) \leq t\phi(X_0'\beta_1) + (1-t)\phi(X_0'\beta_2).$$

Hence, because $(\beta_1, \beta_2) \in \mathcal{B}'^2$,

$$E\left[\phi(X_0'(t\beta_1 + (1-t)\beta_2))\right] \leq E\left[\phi(Y_0)\right],$$

which also implies that $t\beta_1 + (1-t)\beta_2 \in \mathcal{B}' \subset \mathcal{B}$. Thus, $\mathcal{B}$ is convex. The inclusion $\mathcal{B} \subset \mathcal{B}^V$ follows from $\mathcal{B} \subset \mathcal{B}'$ and the convexity of $x \mapsto x^2$ which implies $\mathcal{B}' \subset \mathcal{B}^V$.

This last point also implies that $\mathcal{B}$ is bounded, as a subset of $\mathcal{B}^V$. Thus, to prove that $\mathcal{B} = \mathcal{B}'$ is compact, it suffices to show that it is closed. First, remark that in the definition of $\mathcal{B}'$, we can replace "ϕ convex" by "ϕ continuous and convex" (in fact, we can focus on the functions $x \mapsto \max(0, x-t)$ for $t \in \mathbb{R}$). Let $(\beta_n)_{n\in\mathbb{N}}$ be such that $\beta_n \in \mathcal{B}'$ and $\beta_n \to \beta$. By Fatou's lemma,

$$E\left[\phi(X_0'\beta)\right] = E\left[\liminf_n \phi(X_0'\beta_n)\right] \leq \liminf_n E\left[\phi(X_0'\beta_n)\right] \leq E\left[\phi(Y_0))\right].$$

Thus, $\beta \in \mathcal{B}' = \mathcal{B}$, and $\mathcal{B}$ is closed.

E.3 Corollary 1

By definition, $\mathcal{B}_k = \{b_k : \exists \beta \in \mathcal{B} : \beta_k = b_k\}$. Because $\mathcal{B}$ is convex and compact, $\mathcal{B}_k$ is a compact interval $[\underline{b}_k, \overline{b}_k]$, with $\underline{b}_k = \inf_{\beta\in\mathcal{B}} e_k'\beta$ and $\overline{b}_k = \sup_{\beta\in\mathcal{B}} e_k'\beta$. Thus, $\overline{b}_k = \sigma(e_k, F_{Y_0}, F_{X_0})$ and, similarly, $\underline{b}_k = -\sigma(-e_k, F_{Y_0}, F_{X_0})$.

Next, remark that solutions β of $\sup_{\beta\in\mathcal{B}} e_k'\beta$ are at the boundary of $\mathcal{B}$ and are thus of the form $\beta = S(F_{Y_0}, F_{X_0'q})q$ for some $q \in \mathcal{S}$ such that $q_k := e_k'q > 0$. Thus,

$$\begin{aligned}
\sigma(e_k, F_{Y_0}, F_{X_0}) &= \sup_{q\in\mathcal{S}:q_k>0} q_k S(F_{Y_0}, F_{X_0'q}) \\
&= \sup_{q\in\mathcal{S}:q_k>0} S\left(F_{Y_0}, F_{X_0'q/q_k}\right) \\
&= \sup_{q\in\mathbb{R}^p:q_k>0} S\left(F_{Y_0}, F_{X_0'q/q_k}\right) \\
&= \sup_{q\in\mathbb{R}^p:q_k=1} S\left(F_{Y_0}, F_{X_0'q}\right) \\
&= \frac{1}{\inf_{q\in\mathbb{R}^p:q_k=1} 1/S\left(F_{Y_0}, F_{X_0'q}\right)},
\end{aligned}$$

where the second equality follows by definition of S. The same reasoning applies to $\sigma(-e_k, F_{Y_0}, F_{X_0})$.

E.4 Proposition 1

Point 1. Let $\psi(y) = \phi(y/2)$. By convexity of ϕ, $\psi(Y_0) \leq [\phi(Y) + \phi(-E(Y))]/2$. Thus, $E[\psi(Y_0)] < \infty$. Now, let $b \neq 0$. By convexity again, $\phi(X'b/4) \leq \{\phi(X_0'b/2) + \phi[E(X_0'b/2)]\}/2$. Since $E[\phi(X'b/4)] = \infty$, this implies $E[\psi(X_0'b)] = \infty$. Because $\mathcal{B} = \{\beta : Y_0 \succ_{\text{cv}} X_0'\beta\}$, $b \notin \mathcal{B}$. Thus $\mathcal{B} = \{0\}$. The result follows since $\beta_0 \in \mathcal{B}$.

Point 2. Let $\beta = (\beta_1, \beta_{-1}) \in \mathcal{B}$. Since $\mathcal{B} = \{\beta : Y_0 \succ_{\text{cv}} X_0'\beta\}$, we have, as above, $\infty > E[\psi(Y_0)] \geq E[\psi(X_0'\beta)]$. Moreover, by convexity of ψ,

$$\psi(X_1\beta_1/3) \leq \frac{1}{3}\left\{\psi(X_0'\beta) + \psi(-X_{-1}'\beta_{-1}) + \psi[E(X'\beta)]\right\}.$$

Moreover, by assumption, $E[\psi(-X_{-1}'\beta_{-1})] < \infty$. Thus,

$$E[\phi(X_1\beta_1/6)] = E[\psi(X_1\beta_1/3)] < \infty,$$

which, by assumption, implies $\beta_1 = 0$. The result follows since $\beta_{0,1} \in \mathcal{B}_1$.

E.5 Proposition 3

By Proposition 2 and linearity of R, we have

$$\mathcal{B}^{\text{con}} = \left\{\lambda q : q \in \mathcal{S}^+ : -\overline{S}(F_{Y,X_c}, F_{-X_{nc}'q,X_c}) \leq \lambda \leq \overline{S}(F_{Y,X_c}, F_{X_{nc}'q,X_c}),\right.$$

$$\forall r \in \mathcal{R} : [Rm_Y - \underline{c}](r) \geq \lambda [Rm'_{X_{nc}} q](r) \Big\}.$$

Remark that when $[Rm'_{X_{nc}} q](r) > 0$, $[Rm_Y - \underline{c}](r) \geq \lambda [Rm'_{X_{nc}} q](r)$ is equivalent to $\lambda \leq [Rm_Y - \underline{c}](r)/[Rm'_{X_{nc}} q](r)$. This implies that

$$\lambda \leq \inf_{\substack{r \in \mathcal{R}: \\ [Rm'_{X_{nc}} q](r) > 0}} \frac{[Rm_Y - \underline{c}](r)}{[Rm'_{X_{nc}} q](r)}.$$

When $[Rm'_{X_{nc}} q](r) = 0$, there are two cases: either $[Rm_Y - \underline{c}](r) \geq 0$, in which case we have no constraint on λ (equivalently, $\lambda \leq \infty$); or $[Rm_Y - \underline{c}](r) < 0$, in which case $\lambda q \notin \mathcal{B}^{\text{con}}$ for any $\lambda \in \mathbb{R}$ (equivalently, $\lambda \leq -\infty$). This can be summarized by

$$\lambda \leq \inf_{\substack{r \in \mathcal{R}: \\ [Rm'_{X_{nc}} q](r) \geq 0}} \lim_{u \downarrow 0} \frac{[Rm_Y - \underline{c}](r) + u}{[Rm'_{X_{nc}} q](r) + u^2}.$$

The reasoning is similar for the lower bound, yielding the final expression for $\mathcal{B}^{\text{con}}$. The expression of $\mathcal{F}^{\text{con}}$ follows as in Proposition 2.

Finally, $\mathcal{B}^{\text{con}}$ is closed and convex, as the intersection of $\mathcal{B}^c$ and $\{\beta \in \mathbb{R}^p : \forall r \in \mathcal{R}, [Rm_Y - \underline{c}](r) \geq [Rm'_{X_{nc}} \beta](r)\}$, which are both closed and convex. Since $\mathcal{B}^c$ is bounded, it is also bounded and thus compact. Finally, because $0_p \in \mathcal{B}^c$, $0_p \in \mathcal{B}^{\text{con}}$ if and only if $[Rm_Y - \underline{c}](r) \geq 0$ for all $r \in \mathcal{R}$.

E.6 Proposition 4

First, $E(Y|X_c) = f(X_c) + m(X_c)'\beta_0$. Assume that $(\tilde{f}, \tilde{\beta})$ also rationalizes the data and the model. Then

$$[f - \tilde{f}](X_c) = m(X_c)'[\tilde{\beta} - \beta_0].$$

Because $f - \tilde{f} \in \mathcal{G}$, we must have $\tilde{\beta} = \beta_0$ and in turn $\tilde{f} = f$.

E.7 Proposition 5

Point 1. Fix $c > 0$. For any $M > 0$, let

$$\phi_M(x) = \phi(x) \mathbb{1}\{|x| \leq M\} + \phi'_+(-M)(-M - x)^+ + \phi'_-(M)(x - M)^+,$$

where ϕ'_+ (resp. ϕ'_-) denotes the right (resp. left) derivative of ϕ. Because $\phi_M(x) \leq K_1 + K_2|x|$ for some $K_1, K_2 > 0$, we have $E[\phi_M(X'_0\beta_0(1+c))] < \infty$. Also, $\phi_M(x) \uparrow \phi(x)$

as $M \uparrow \infty$. Then, by the monotone convergence theorem,

$$\lim_{M\to\infty} E[\phi_M(X_0'\beta_0(1+c))] = E[\phi(X_0'\beta_0(1+c))] = \infty.$$

On the other hand, $E[\phi_M((1/c+1)U)] \leq E[\phi((1/c+1)U)] < \infty$. Thus, there exists M_c such that

$$E[\phi_{M_c}((1/c+1)U)] < E[\phi_{M_c}(X_0'\beta_0(1+c))].$$

Moreover, using $Y_0 = X_0'\beta_0 + U$ and convexity of ϕ_{M_c}, we obtain

$$\phi_{M_c}(Y_0) \leq \frac{1}{1+c}\phi_{M_c}(X_0'\beta_0(1+c)) + \frac{c}{1+c}\phi_{M_c}((1/c+1)U).$$

Combining the two inequalities, we obtain[19]

$$E[\phi_{M_c}(Y_0)] < E[\phi_{M_c}(X_0'\beta_0(1+c))].$$

Because ϕ_{M_c} is convex, this implies that $\beta_0(1+c) \notin \mathcal{B}$. Since $c > 0$ was arbitrary, $\beta_0 \in \partial\mathcal{B}$. The result follows.

Point 2. By convexity, $\phi(X\lambda/2) \leq [\phi(X_0\lambda)+\phi(E(X)\lambda)]/2$ for all $\lambda > 0$. Therefore, for such λ, $E[\phi(X_0\lambda)] = \infty$. Since $X \in \mathbb{R}$ and $\beta_0 > 0$, this implies $E[\phi((X_0'\beta_0)\lambda)] = \infty$ for all $\lambda > 1$. Thus, the condition of Point 1 holds and the identified set of β_0 is included in $\partial\mathcal{B}$, which is of the form $\{b, \beta_0\}$ for some $b \leq 0$ (since $0 \in \mathcal{B}$). Because it is known that $\beta_0 > 0$, the identified set is $\{\beta_0\}$.

E.8 Proposition 6

1. $\mathcal{B}_\varepsilon$ is compact and convex.

We showed in the proof of Theorem 1 that for all $\alpha \in (0,1)$, $\int_\alpha^1 F^{-1}_{X_0'q}(t)dt > 0$ and $\int_\alpha^1 F^{-1}_{Y_0}(t)dt > 0$. Then, by continuity of $\alpha \mapsto \int_\alpha^1 F^{-1}_{Y_0}(t)dt / \int_\alpha^1 F^{-1}_{X_0'q}(t)dt$,

$$S_\varepsilon(F_{Y_0}, F_{X_0'q}) = \min_{\alpha\in[\varepsilon,1-\varepsilon]} \frac{\int_\alpha^1 F^{-1}_{Y_0}(t)dt}{\int_\alpha^1 F^{-1}_{X_0'q}(t)dt} > 0.$$

Hence, $p_\varepsilon(q) := 1/S_\varepsilon(F_{Y_0}, F_{X_0'q})$ is well-defined and

$$p_\varepsilon(q) = \max_{\alpha\in[\varepsilon,1-\varepsilon]} \frac{\int_\alpha^1 F^{-1}_{X_0'q}(t)dt}{\int_\alpha^1 F^{-1}_{Y_0}(t)dt}.$$

[19] Using ϕ instead of ϕ_{M_c} would not work: $E[\phi(Y_0)] \geq E[\phi(X_0'\beta_0)]$, so we may have $E[\phi(Y_0)] = \infty$.

Besides, for any random variables U and V, and $\lambda \in [0,1]$,

$$\begin{aligned}\int_\alpha^1 F^{-1}_{\lambda U+(1-\lambda)V}(t)dt &\leq \int_\alpha^1 F^{-1}_{\lambda U}(t)dt + \int_\alpha^1 F^{-1}_{(1-\lambda)V}(t)dt \\ &= \lambda \int_\alpha^1 F_U^{-1}(t)dt + (1-\lambda)\int_\alpha^1 F_V^{-1}(t)dt,\end{aligned}$$

where the first inequality follows from Theorem 1.1 in Embrechts and Wang (2015). As a result, for any $\alpha \in (0,1)$, the function $q \mapsto \int_\alpha^1 F^{-1}_{X_0'q}(t)dt$ is convex. Because the maximum of convex functions is also convex, the function p_ε is convex on $\mathbb{R}^p$. As such, it is also continuous. This implies that $\mathcal{B}_\varepsilon = \{q \in \mathbb{R}^p : p_\varepsilon(q) \leq 1\}$ is convex and closed. Finally, by continuity of $q \mapsto S_\varepsilon(F_{Y_0}, F_{X_0'q})$,

$$\sup_{q\in\mathcal{S}} S_\varepsilon(F_{Y_0}, F_{X_0'q}) = \max_{q\in\mathcal{S}} S_\varepsilon(F_{Y_0}, F_{X_0'q}) < \infty,$$

which implies that $\mathcal{B}_\varepsilon$ is bounded, and thus compact.

2. For all $0 < \varepsilon < \varepsilon' < 1/2$, $\mathcal{B} \subset \mathcal{B}_\varepsilon \subset \mathcal{B}_{\varepsilon'}$ and $\cap_{\varepsilon\in(0,1/2)}\mathcal{B}_\varepsilon = \mathcal{B}$.

The first result follows since by definition, $S_\varepsilon(F,G) \leq S_{\varepsilon'}(F,G)$ for any $0 < \varepsilon < \varepsilon' < 1/2$. Now,

$$\cap_{\varepsilon\in(0,1/2)}\mathcal{B}_\varepsilon = \left\{\lambda q :\ q \in \mathcal{S},\ 0 \leq \lambda \leq \inf_{\varepsilon\in(0,1/2)} S_\varepsilon(F_{Y_0}, F_{X_0'q})\right\}.$$

Thus, to prove $\cap_{\varepsilon\in(0,1/2)}\mathcal{B}_\varepsilon = \mathcal{B}$, it suffices to show that $\inf_{\varepsilon\in(0,1/2)} S_\varepsilon(F,G) = S(F,G)$. First, $\inf_{\varepsilon\in(0,1/2)} S_\varepsilon(F,G) \geq S(F,G)$ since $S_\varepsilon(F,G) \geq S(F,G)$ for all $\varepsilon \in (0,1/2)$. Now, fix $\eta > 0$. By definition, there exists $\alpha_0 \in (0,1)$ such that

$$S(F,G) > R(\alpha_0, F, G) - \eta.$$

Hence, there exists $\varepsilon \in (0,1/2)$ such that

$$S(F,G) > S_\varepsilon(F,G) - \eta \geq \inf_{\varepsilon\in(0,1/2)} S_\varepsilon(F,G) - \eta.$$

Since η is arbitrary, we have $S(F,G) \geq \inf_{\varepsilon\in(0,1/2)} S_\varepsilon(F,G)$. The result follows.

3. Under the stated conditions, there exists $0 < \varepsilon_0 < 1/2$ such that $\mathcal{B}_{\varepsilon_0} = \mathcal{B}$.

We first show that for all q, as $\alpha \to 1$, $R(\alpha, F_{Y_0}, F_{X_0'q}) \to \infty$. First, $E(X_0'\beta_0) = 0$ implies that $P(X_0'\beta_0 \geq 0) > 0$. Next, for all λ and M, there exists t_0 such that for

all $t \geq t_0$ and all s, $\overline{F}_{U|X_0'\beta_0}(t|s) > (M/P(X_0'\beta_0 \geq 0))\overline{F}_{\|X_0\|}(\lambda t)$. Then, for all $t \geq t_0$,

$$\begin{aligned}
\overline{F}_{Y_0}(t) =& E[\overline{F}_{U|X_0'\beta_0}(t - X_0'\beta_0|X_0'\beta_0)] \\
\geq& E[\overline{F}_{U|X_0'\beta_0}(t - X_0'\beta_0|X_0'\beta_0)\mathbb{1}\{X_0'\beta_0 \geq 0\}] \\
\geq& E[\overline{F}_{U|X_0'\beta_0}(t|X_0'\beta_0)\mathbb{1}\{X_0'\beta_0 \geq 0\}] \\
\geq& M\overline{F}_{\|X_0\|}(\lambda t).
\end{aligned} \tag{26}$$

In other words,

$$\forall \lambda > 0,\ \lim_{t\to\infty} \frac{\overline{F}_{\|X_0\|}(\lambda t)}{\overline{F}_{Y_0}(t)} = 0. \tag{27}$$

If $\sup \mathrm{Supp}(X_0'q) < \infty$, (26) together with $\mathrm{Supp}(U) = \mathbb{R}$ implies that $\sup \mathrm{Supp}(Y_0) = \infty$ and thus $F^{-1}_{\|X_0\|}(\alpha) = o(F^{-1}_{Y_0}(\alpha))$. Now, if $\sup \mathrm{Supp}(\|X_0\|) = \infty$, $F^{-1}_{\|X_0\|}(\alpha) \to \infty$ as $\alpha \to 1$. Thus,

$$\forall \lambda > 0, \quad \lim_{\alpha\to 1} \frac{\overline{F}_{\|X_0\|}(F^{-1}_{\|X_0\|}(\alpha))}{\overline{F}_{Y_0}(\lambda F^{-1}_{\|X_0\|}(\alpha))} = 0.$$

Now, remark that by continuity of F_{Y_0}, $\overline{F}_{\|X_0\|}(F^{-1}_{\|X_0\|}(\alpha)) \leq 1 - \alpha = \overline{F}_{Y_0}(F^{-1}_{Y_0}(\alpha))$. Therefore,

$$\forall \lambda > 0, \quad \lim_{\alpha\to 1} \frac{\overline{F}_{Y_0}(F^{-1}_{Y_0}(\alpha))}{\overline{F}_{Y_0}(\lambda F^{-1}_{\|X_0\|}(\alpha))} = 0.$$

Since $\overline{F}_{Y_0}$ is decreasing, this implies that there exists $\alpha_0(\lambda)$ such that, for all $\alpha \geq \alpha_0(\lambda)$,

$$F^{-1}_{Y_0}(\alpha) > \lambda F^{-1}_{\|X_0\|}(\alpha).$$

Because λ was arbitrary, this proves $F^{-1}_{\|X_0\|}(\alpha) = o(F^{-1}_{Y_0}(\alpha))$. Then, by integration, we obtain, as $\alpha \to 1$,

$$R(\alpha, F_{Y_0}, F_{\|X_0\|}) \to \infty.$$

The exact same reasoning shows that as $\alpha \to 0$, $R(\alpha, F_{Y_0}, F_{\|X_0\|}) \to \infty$. Now, let us define $M := \sup_{q\in\mathcal{S}} S_{1/4}(F_{Y_0}, F_{X_0'q})$. We proved in Point 1 above that $q \mapsto S_{1/4}(F_{Y_0}, F_{X_0'q})$ is continuous, implying that $M < \infty$. Then, by what precedes, there exists $\varepsilon_0 \in (0, 1/4)$ such that

$$\inf_{\alpha\in(0,\varepsilon_0)\cup(1-\varepsilon_0,1)} R(\alpha, F_{Y_0}, F_{\|X_0\|}) > M.$$

Moreover, by the Cauchy-Schwarz inequality, $R(\alpha, F_{Y_0}, F_{X_0'q}) \geq R(\alpha, F_{Y_0}, F_{\|X_0\|})$. As a result,

$$\inf_{q\in\mathcal{S}} \inf_{\alpha\in(0,\varepsilon_0)\cup(1-\varepsilon_0,1)} R(\alpha, F_{Y_0}, F_{X_0'q}) > M.$$

By definition, for all $q \in \mathcal{S}$, $S_{\varepsilon_0}(F_{Y_0}, F_{X_0'q}) \leq S_{1/4}(F_{Y_0}, F_{X_0'q}) \leq M$. Then,

$$\begin{aligned} S(F_{Y_0}, F_{X_0'q}) &= \min\left(\inf_{\alpha\in(0,\varepsilon_0)\cup(1-\varepsilon_0,1)} R(\alpha, F_{Y_0}, F_{X_0'q}), S_{\varepsilon_0}(F_{Y_0}, F_{X_0'q})\right), \\ &= S_{\varepsilon_0}(F_{Y_0}, F_{X_0'q}). \end{aligned}$$

This proves that $\mathcal{B} = \mathcal{B}_{\varepsilon_0}$.

E.9 Proposition 7

In both cases, it suffices to prove the result for ε small enough.

Proof of Point 1

The proof proceeds in two steps. First, we obtain an upper bound $S_\varepsilon(F_{Y_0}, F_{X_0'\beta_0}) - S(F_{Y_0}, F_{X_0'\beta_0})$. Then, we obtain the bound on $d_H(\mathcal{B}, \mathcal{B}_\varepsilon)$.

Step 1: upper bound on $S_\varepsilon(F_{Y_0}, F_{X_0'\beta_0}) - S(F_{Y_0}, F_{X_0'\beta_0})$.

First, observe that $R(\cdot, F_{Y_0}, F_{X_0'\beta_0})$ is differentiable and

$$\begin{aligned} &S_\varepsilon(F_{Y_0}, F_{X_0'\beta_0}) - S(F_{Y_0}, F_{X_0'\beta_0}) \\ \leq &\max\Bigg(R(\varepsilon, F_{Y_0}, F_{X_0'\beta_0}) - \inf_{\alpha\in[0,\varepsilon)} R(\alpha, F_{Y_0}, F_{X_0'\beta_0}), \\ &\qquad R(1-\varepsilon, F_{Y_0}, F_{X_0'\beta_0}) - \inf_{\alpha\in[1-\varepsilon,1)} R(\alpha, F_{Y_0}, F_{X_0'\beta_0})\Bigg) \\ \leq &\int_{[0,\varepsilon]\cup[1-\varepsilon,1]} \left|\frac{\partial R}{\partial \alpha}(\alpha, F_{Y_0}, F_{X_0'\beta_0})\right| d\alpha. \end{aligned} \tag{28}$$

Now, $F^{-1}_{X_0'\beta_0}(\alpha_0) > 0$ for some $\alpha_0 < 1$. Then, for $\alpha \geq \alpha_0$,

$$\begin{aligned} \left|\frac{\partial R}{\partial \alpha}(\alpha, F_{Y_0}, F_{X_0'\beta_0})\right| =& \frac{1}{\int_\alpha^1 F^{-1}_{X_0'\beta_0}(t)dt}\left|-F^{-1}_{Y_0}(\alpha) + R(\alpha, F_{Y_0}, F_{X_0'\beta_0})F^{-1}_{X_0'\beta_0}(\alpha)\right| \\ =& \frac{|F^{-1}_{X_0'\beta_0}(\alpha)|}{\int_\alpha^1 F^{-1}_{X_0'\beta_0}(t)dt}\left|R(\alpha, F_{Y_0}, F_{X_0'\beta_0}) - \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)}\right| \\ \leq& \frac{1}{1-\alpha}\left|R(\alpha, F_{Y_0}, F_{X_0'\beta_0}) - \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)}\right|. \end{aligned} \tag{29}$$

Let $w_\alpha(t) = F^{-1}_{X_0'\beta_0}(t)/\int_\alpha^1 F^{-1}_{X_0'\beta_0}(u)du$. For $t \geq \alpha_0$, $w_\alpha(t) > 0$. Then, for $\alpha \geq \alpha_0$,

$$\begin{aligned}\left|R(\alpha, F_{Y_0}, F_{X_0'\beta_0}) - \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)}\right| &= \left|\int_\alpha^1 w_\alpha(t)\frac{F^{-1}_{Y_0}(t)}{F^{-1}_{X_0'\beta_0}(t)}dt - \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)}\right| \qquad (30)\\ &= \left|\int_\alpha^1 w_\alpha(t)\left(\frac{F^{-1}_{Y_0}(t)}{F^{-1}_{X_0'\beta_0}(t)} - 1\right)dt - \left(\frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)} - 1\right)\right| \\ &\leq 2\sup_{t\in[\alpha,1]}\left|\frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)} - 1\right| \\ &\lesssim (1-\alpha)^{\frac{1/c-1/d}{1+1/d}}, \qquad (31)\end{aligned}$$

where the last inequality follows from Lemma 2 in the supplementary material. Combining (29) and (31), we obtain, for $\varepsilon \leq 1 - \alpha_0$,

$$\int_{1-\varepsilon}^1 \left|\frac{\partial R}{\partial \alpha}(\alpha, F_{Y_0}, F_{X_0'\beta_0})\right| d\alpha \lesssim \varepsilon^{\frac{1/c-1/d}{1+1/d}}. \qquad (32)$$

Similarly, note that $\int_\alpha^1 F^{-1}_{X_0'\beta_0}(t)dt = -\int_0^\alpha F^{-1}_{X_0'\beta_0}(t)dt \geq -\alpha F^{-1}_{X_0'\beta_0}(\alpha)$ and $F^{-1}_{X_0'\beta_0}(\alpha_1) < 0$ for some α_1. Then, for $\alpha \leq \alpha_1$, we obtain, instead of (29),

$$\left|\frac{\partial R}{\partial \alpha}(\alpha, F_{Y_0}, F_{X_0'\beta_0})\right| \leq \frac{1}{\alpha}\left|R(\alpha, F_{Y_0}, F_{X_0'\beta_0}) - \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)}\right|. \qquad (33)$$

The same reasoning as to get (31) but using $R(\alpha, F_{Y_0}, F_{X_0'\beta_0}) = \int_0^\alpha F^{-1}_{Y_0}(t)dt/\int_0^\alpha F^{-1}_{X_0'\beta_0}(t)dt$, $w_\alpha(t) = F^{-1}_{X_0'\beta_0}(t)/\int_0^\alpha F^{-1}_{X_0'\beta_0}(u)du$ and, again, Lemma 2 yields, for $\alpha \leq \alpha_1$,

$$\left|R(\alpha, F_{Y_0}, F_{X_0'\beta_0}) - \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)}\right| \lesssim \alpha^{\frac{1/c-1/d}{1+1/d}}. \qquad (34)$$

Thus, for $\varepsilon \leq \alpha_1$,

$$\int_0^\varepsilon \left|\frac{\partial R}{\partial \alpha}(\alpha, F_{Y_0}, F_{X_0'\beta_0})\right| d\alpha \lesssim \varepsilon^{\frac{1/c-1/d}{1+1/d}}. \qquad (35)$$

Then, (28), (32) and (35) imply that for $\varepsilon \leq \min(1-\alpha_0, \alpha_1)$,

$$S_\varepsilon(F_{Y_0}, F_{X_0'\beta_0}) - S(F_{Y_0}, F_{X_0'\beta_0}) \lesssim \varepsilon^{\frac{1/c-1/d}{1+1/d}}. \qquad (36)$$

Step 2: upper bound on $d_H(\mathcal{B}, \mathcal{B}_\varepsilon)$.

X has an elliptical distribution with nonsingular variance matrix Σ. As a result, for all $q \in \mathcal{S}$, there exists $\sigma(q)$ such that $X_0'q \overset{d}{=} \sigma(q)X_0'\beta_0$, with

$$\sigma(q)^2 = \frac{q'\Sigma q}{\beta_0'\Sigma\beta_0} \geq \frac{\underline{\lambda}_\Sigma}{\beta_0'\Sigma\beta_0},$$

where $\underline{\lambda}_\Sigma > 0$ denotes the smallest eigenvalue of Σ and the inequality can be reached. Then,

$$\begin{aligned} d_H(\mathcal{B}, \mathcal{B}_\varepsilon) &\leq \sup_{q\in\mathcal{S}} S_\varepsilon(F_{Y_0}, F_{X_0'q}) - S(F_{Y_0}, F_{X_0'q}) \\ &= \sup_{q\in\mathcal{S}} S_\varepsilon(F_Y, F_{\sigma(q)X_0'\beta_0}) - S(F_Y, F_{\sigma(q)X_0'\beta_0}) \\ &= \left[S_\varepsilon(F_{Y_0}, F_{X_0'\beta_0}) - S(F_{Y_0}, F_{X_0'\beta_0})\right] \sup_{q\in\mathcal{S}}[1/\sigma(q)] \\ &= \left[\frac{\beta_0'\Sigma\beta_0}{\underline{\lambda}_\Sigma}\right]^{1/2} \left[S_\varepsilon(F_{Y_0}, F_{X_0'\beta_0}) - S(F_{Y_0}, F_{X_0'\beta_0})\right] \\ &\lesssim \varepsilon^{\frac{1/c-1/d}{1+1/d}}, \end{aligned}$$

where the first inequality uses the definition of the Hausdorff distance and $\mathcal{B} \subset \mathcal{B}_\varepsilon$, the first equality follows since $X_0'q \overset{d}{=} \sigma(q)X_0'\beta_0$, the second equality uses the definition of R, S and S_ε and the last inequality is due to (36).

Proof of Point 2

First, our assumptions imply that, for $\alpha \geq \alpha_0$ (resp. $\alpha \leq \alpha_1$) and all $q \in \mathcal{S}$, $F^{-1}_{X_0'q}(\alpha) \gtrsim (1-\alpha)^{-1/c}$ (resp. $F^{-1}_{X_0'q}(\alpha) \gtrsim \alpha^{-1/c}$) and $F_U^{-1}(\alpha) \lesssim (1-\alpha)^{-1/d}$ (resp. $F_U^{-1}(\alpha) \lesssim \alpha^{-1/d}$), hence using $\beta_0 = 0_p$,

$$\forall \alpha \geq \alpha_0, \quad \sup_{q\in\mathcal{S}} \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'q}(\alpha)} \lesssim (1-\alpha)^{1/c-1/d}, \quad \forall \alpha \leq \alpha_1, \quad \sup_{q\in\mathcal{S}} \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'q}(\alpha)} \lesssim \alpha^{1/c-1/d}.$$

Now, remark that (28), (29), and (33) still hold with $X_0'\beta_0$ replaced by $X_0'q$, $q \in \mathcal{S}$. Then, using (30), we obtain, instead of (31) and (34), for $\alpha \in (0, \alpha_0] \cup [\alpha_1, 1)$,

$$\begin{aligned} \sup_{q\in\mathcal{S}} \left| R(\alpha, F_{Y_0}, F_{X_0'q}) - \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'q}(\alpha)} \right| &\leq 2 \sup_{q\in\mathcal{S}} \sup_{t\in[\alpha,1]} \frac{F^{-1}_{Y_0}(t)}{F^{-1}_{X_0'q}(t)} \\ &\lesssim (1-\alpha)^{1/c-1/d} \mathbb{1}\{\alpha \geq \alpha_0\} + \alpha^{1/c-1/d} \mathbb{1}\{\alpha \leq \alpha_1\}. \end{aligned}$$

As a result, for $\varepsilon \leq \min(1-\alpha_0, \alpha_1)$,

$$d_H(\mathcal{B}, \mathcal{B}_\varepsilon) \leq \sup_{q\in\mathcal{S}} S_\varepsilon(F_{Y_0}, F_{X_0'q}) \lesssim \varepsilon^{1/c-1/d},$$

where in the first inequality we used $S_\varepsilon(F_{Y_0}, F_{X_0'q}) - S(F_{Y_0}, F_{X_0'q}) \leq S_\varepsilon(F_{Y_0}, F_{X_0'q})$.

E.10 Theorem 2

Recall that $\widehat{F}_{Y_0}(t) = \frac{1}{n_Y}\sum_{i=1}^{n_Y} \mathbb{1}\left\{Y_i - \overline{Y} \leq t\right\}$ and $\widehat{F}_{X_0'q}(t)$ is defined similarly. The proof proceeds in two steps. We first prove that for all $q \in \mathcal{S}$, $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) \xrightarrow{\mathbb{P}} S_\varepsilon(F_{Y_0}, F_{X_0'q})$. Then, we show that $d_H(\widehat{\mathcal{B}}_\varepsilon, \mathcal{B}_\varepsilon) \xrightarrow{\mathbb{P}} 0$.

Step 1: $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) \xrightarrow{\mathbb{P}} S_\varepsilon(F_{Y_0}, F_{X_0'q})$**, for all** $q \in \mathcal{S}$**.**

The idea is to apply the continuous mapping theorem, with the metric

$$d((F,G),(F',G')) = W_1(F,F') + W_1(G,G'),$$

where we recall that W_1 is the 1-Wasserstein distance. To this end, we first show that $(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ converges to $(F_{Y_0}, F_{X_0'q})$ for this metric. It suffices to prove that $W_1(\widehat{F}_{Y_0}, F_{Y_0}) \xrightarrow{\mathbb{P}} 0$, the proof being similar for $X_0'q$. Remark that $\widehat{F}_{Y_0}(t) = \widehat{F}_Y(t+\overline{Y})$ and $F_{Y_0}(y) = F_Y(y + E(Y))$. Then,

$$\begin{aligned} W_1(\widehat{F}_{Y_0}, F_{Y_0}) =& \int_{-\infty}^{\infty} |\widehat{F}_Y(t+\overline{Y}) - F_Y(t+\overline{Y}) + F_Y(t+\overline{Y}) - F_Y(t+E(Y))|dt \\ \leq& W_1(\widehat{F}_Y, F_Y) + \int_{-\infty}^{\infty} |F_Y(t+\overline{Y}) - F_Y(t+E(Y))|dt \\ =& W_1(\widehat{F}_Y, F_Y) + |\overline{Y} - E(Y)|, \end{aligned}$$

where the first equality follows by (24) and the last equality by Fubini's theorem. Because $E[|Y|] < \infty$, we have, by the law of large numbers $|\overline{Y} - E(Y)| \xrightarrow{\mathbb{P}} 0$ and also (see (1.3) in Del Barrio et al., 1999) $W_1(\widehat{F}_Y, F_Y) \xrightarrow{\mathbb{P}} 0$.

Thus, the first step follows if we prove that S_ε is continuous for the metric d. First, by Lemma 3, R is continuous with respect to the metric d' on $[\varepsilon, 1-\varepsilon] \times \mathcal{D}^2$, where $\mathcal{D}$ denote the set of cdfs with mean 0 and d' is defined by

$$d'((\alpha,F,G),(\alpha',F',G')) = |\alpha' - \alpha| + W_1(F,F') + W_1(G,G'). \tag{37}$$

Now, because the product topology is induced by d', R is continuous on the product $[\varepsilon, 1-\varepsilon] \times \mathcal{D}^2$. Since $[\varepsilon, 1-\varepsilon]$ is compact, it follows from Berge maximum theorem (see, e.g., Theorem 9.14 in Sundaram, 1996) that S_ε is also continuous with respect to the metric d. The result follows.

Step 2: Convergence of the set $\widehat{\mathcal{B}}_\varepsilon$.

We showed in the proof of Proposition 6 that $S_\varepsilon(F_{Y_0}, F_{X_0'q}) > 0$ for all $q \in \mathcal{S}$. Then, let $p_\varepsilon(q) = 1/S_\varepsilon(F_{Y_0}, F_{X_0'q})$ and $\widehat{p}_\varepsilon(q) = 1/S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$. By the continuous mapping theorem, for all $q \in \mathcal{S}$, $\widehat{p}_\varepsilon(q) \xrightarrow{\mathbb{P}} p_\varepsilon(q)$. Moreover,

$$\widehat{p}_\varepsilon(q) = \max_{\alpha \in [\varepsilon, 1-\varepsilon]} 1/R\left(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right). \tag{38}$$

Note that for any (F_Y, F_X) and $\alpha \in [\varepsilon, 1-\varepsilon]$, $q \mapsto 1/R(\alpha, F_{Y_0}, F_{X_0'q})$ is convex (see the proof of Point 1 in Proposition 6). Then, (38) implies that $\widehat{p}_\varepsilon$ is also convex. As a result, by the convexity lemma of Pollard (1991),

$$\sup_{q \in \mathcal{S}} |\widehat{p}_\varepsilon(q) - p_\varepsilon(q)| \xrightarrow{\mathbb{P}} 0. \tag{39}$$

By construction, $\widehat{p}_\varepsilon$ (resp. p_ε) is the gauge function of the set $\widehat{\mathcal{B}}_\varepsilon$ (resp. $\mathcal{B}_\varepsilon$). The gauge function of a nonempty, compact and convex set H containing the origin is defined as the support function of its polar set (see, e.g., Corollary 3.2.5 p.149 in Hiriart-Urruty and Lemaréchal, 2012). Thus, using Theorem 3.3.6 p.155 in Hiriart-Urruty and Lemaréchal (2012) and denoting respectively by $\widehat{\mathcal{B}}_\varepsilon^\circ$ and $\mathcal{B}_\varepsilon^\circ$ the polar sets of $\widehat{\mathcal{B}}_\varepsilon$ and $\mathcal{B}_\varepsilon$, we obtain

$$d_H\left(\widehat{\mathcal{B}}_\varepsilon^\circ, \mathcal{B}_\varepsilon^\circ\right) = \sup_{q \in \mathcal{S}} |\widehat{p}_\varepsilon(q) - p_\varepsilon(q)|.$$

Thus, by (39), $d_H\left(\widehat{\mathcal{B}}_\varepsilon^\circ, \mathcal{B}_\varepsilon^\circ\right) \xrightarrow{\mathbb{P}} 0$. The result follows because convergence of polar sets for the Hausdorff distance implies convergence of the sets themselves for the same distance, see Theorem 7.2 in Wijsman (1966).

E.11 Theorem 3

1. Asymptotic validity of the confidence region

Let us define $\iota(G) = \inf_{\alpha \in [\varepsilon, 1-\varepsilon]} G(\alpha)$. By definition,

$$S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) = \iota\left[R(\cdot, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q})\right].$$

Moreover, by Theorem 2.1 of Cárcamo et al. (2020), ι is Hadamard directionally differentiable. Then, by Lemma 4 in the supplementary material and the functional delta

method for Hadamard directionally differentiable functions (see, e.g., Proposition 2.1 in Cárcamo et al., 2020), we have

$$n^{1/2}\left(S_\varepsilon(\widehat{F}_{Y_0},\widehat{F}_{X_0'q})-S_\varepsilon(F_{Y_0},F_{X_0'q})\right)\xrightarrow{d}\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}(\mathbb{F}), \tag{40}$$

where, in view of Corollary 2.3 in Cárcamo et al. (2020), $\iota'_f(h)=\inf\{h(x):x\in \operatorname{argmin}_{\alpha\in[\varepsilon,1-\varepsilon]}f(\alpha)\}$ for any continuous functions f and h.

Now, let us show that $\widehat{c}_{\alpha,\varepsilon}\xrightarrow{\mathbb{P}}c_{\alpha,\varepsilon}$. Denote by H the cdf of $\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}(\mathbb{F})$. Note that $-\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}$ is convex. Then, by Theorem 11.1 in Davydov et al. (1998), its cdf H is continuous and strictly increasing in a neighborhood of every point of its support except perhaps at $\underline{r}:=\inf\{r\in\mathbb{R}:H(r)>0\}$. By Problem 11.3 in Davydov et al. (1998), we also have that $H(r)>0$ for any $r\in\mathbb{R}$. Thus, H is continuous and strictly increasing on $\mathbb{R}$. Since $-c_{\alpha,\varepsilon}$ is the quantile of order $1-\alpha$ of $-\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}(\mathbb{F})$ and using (40), it follows from Theorem 2.2.1 in Politis et al. (1999) that $\widehat{c}_{\alpha,\varepsilon}\xrightarrow{\mathbb{P}}c_{\alpha,\varepsilon}$.

Finally, fix $\beta\in\mathcal{B}_\varepsilon$, so that $\beta=\lambda q$ with $\lambda\in[0,S_\varepsilon(F_{Y_0},F_{X_0'q})]$. By definition, $\beta\in \mathrm{CR}_{1-\alpha}(\beta_0)$ if and only if

$$n^{1/2}\left(S_\varepsilon(\widehat{F}_{Y_0},\widehat{F}_{X_0'q})-\lambda\right)-\widehat{c}_{\alpha,\varepsilon}\geq 0. \tag{41}$$

Suppose first that $\lambda<S_\varepsilon(F_{Y_0},F_{X_0'q})$. Since $S_\varepsilon(\widehat{F}_{Y_0},\widehat{F}_{X_0'q})$ is consistent for $S_\varepsilon(F_{Y_0},F_{X_0'q})$ and $\widehat{c}_{\alpha,\varepsilon}=O_P(1)$, (41) holds with probability approaching one and $\liminf_{n\to\infty}P(\beta\in \mathrm{CR}_{1-\alpha}(\beta_0))=1$. Now, suppose that $\lambda=S_\varepsilon(F_{Y_0},F_{X_0'q})$. Then, by what precedes,

$$n^{1/2}\left(S_\varepsilon(\widehat{F}_{Y_0},\widehat{F}_{X_0'q})-\lambda\right)-\widehat{c}_{\alpha,\varepsilon}\xrightarrow{d}\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}(\mathbb{F})-c_{\alpha,\varepsilon}.$$

Moreover, by continuity of the cdf of $\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}(\mathbb{F})$ at $c_{\alpha,\varepsilon}$,

$$P(\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}(\mathbb{F})-c_{\alpha,\varepsilon}\geq 0)=1-\alpha.$$

Thus, $\liminf_{n\to\infty}P(\beta\in \mathrm{CR}_{1-\alpha}(\beta_0))=1-\alpha$. Equation (16) follows since $\beta\in\mathcal{B}_\varepsilon\subset\mathcal{B}$.

Now, suppose that Assumption 5 holds and let us prove that (16) is still true if ε is replaced by $\varepsilon(q)$ (if $p=1$) or $\underline{\varepsilon}$ (if $p>1$). We can focus on $\beta\in\partial\mathcal{B}_\varepsilon$, $\beta=S_\varepsilon(F_{Y_0},F_{X_0'q})q$. If $S_\varepsilon(F_{Y_0},F_{X_0'q})>S(F_{Y_0},F_{X_0'q})$ for all $\varepsilon\in\mathcal{E}$, we get, for $\varepsilon'=\varepsilon(q)$ or $\varepsilon'=\underline{\varepsilon}$,

$$S_{\varepsilon'}(\widehat{F}_{Y_0},\widehat{F}_{X_0'q})\geq\min_{\varepsilon\in\mathcal{E}}S_\varepsilon(\widehat{F}_{Y_0},\widehat{F}_{X_0'q})\xrightarrow{\mathbb{P}}\min_{\varepsilon\in\mathcal{E}}S_\varepsilon(F_{Y_0},F_{X_0'q})>S(F_{Y_0},F_{X_0'q}),$$

where the convergence holds by the convergence in probability of $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$ for any $\varepsilon \in \mathcal{E}$ and the continuous mapping theorem. Equation (16) follows. Suppose instead that $a \mapsto R(a, F_{Y_0}, F_{X_0'q})$ admits a unique minimizer a_0 on $(0,1)$. Up to replacing a_0 by $1-a_0$, we can suppose without loss of generality that $a_0 \leq 0.5$. Let $\mathcal{E} = \{\varepsilon_1, ..., \varepsilon_J\}$, with $\varepsilon_1 < ... < \varepsilon_J < 1/2$. Reasoning as above, we have

$$\sqrt{n}\begin{pmatrix} S_{\varepsilon_1}(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) - S_{\varepsilon_1}(F_{Y_0}, F_{X_0'q}) \\ \vdots \\ S_{\varepsilon_J}(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) - S_{\varepsilon_J}(F_{Y_0}, F_{X_0'q}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \iota'_{\varepsilon_1, R(\cdot, F_{Y_0}, F_{X_0'q})}(\mathbb{F}) \\ \vdots \\ \iota'_{\varepsilon_J, R(\cdot, F_{Y_0}, F_{X_0'q})}(\mathbb{F}) \end{pmatrix},$$

where, compared to (40), we let the dependence of ι' on ε explicit. If $\varepsilon_1 > a_0$, then for any $\varepsilon \in \mathcal{E}$, $S_\varepsilon(F_{Y_0}, F_{X_0'q}) > S(F_{Y_0}, F_{X_0'q})$, and the reasoning above applies. Otherwise, let $\varepsilon_{j_0} = \max\{\varepsilon \in \mathcal{E} : \varepsilon \leq a_0\}$ (where we simply let $\varepsilon_{J+1} = 1$ if $j_0 = J$). Then, with probability approaching one, $\varepsilon(q) \in \{\varepsilon_1, ..., \varepsilon_{j_0}\}$. Moreover, the expression of $\iota'_{\varepsilon, R(\cdot, F_{Y_0}, F_{X_0'q})}$ and that $a \mapsto R(a, F_{Y_0}, F_{X_0'q})$ admits a unique minimizer a_0 imply that $\iota'_{\varepsilon_1, R(\cdot, F_{Y_0}, F_{X_0'q})}(\mathbb{F}) = \cdots = \iota'_{\varepsilon_{j_0}, R(\cdot, F_{Y_0}, F_{X_0'q})}(\mathbb{F}) = \mathbb{F}(a_0)$. As a result,

$$\left(\widehat{c}_{\alpha,\varepsilon_1}, ..., \widehat{c}_{\alpha,\varepsilon_{j_0}}\right) \xrightarrow{\mathbb{P}} (c_\alpha, ..., c_\alpha),$$

where c_α is the quantile of order α of $\mathbb{F}(a_0)$. Combining these results yield, for all $j \in \{2, ..., j_0\}$,

$$S_{\varepsilon_j}(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) - \widehat{c}_{\alpha,\varepsilon_j} n^{-1/2} = S_{\varepsilon_1}(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) - \widehat{c}_{\alpha,\varepsilon_1} n^{-1/2} + o_P(n^{-1/2}).$$

In turn, this implies that

$$S_{\varepsilon(q)}(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) - \widehat{c}_{\alpha,\varepsilon(q)} n^{-1/2} = S_{\varepsilon_1}(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) - \widehat{c}_{\alpha,\varepsilon_1} n^{-1/2} + o_P(n^{-1/2}),$$

which ensures that using $\varepsilon(q)$ leads to asymptotically correct coverage in this case. Finally, remark that by definition of $\varepsilon(q)$ (and letting the dependence of the confidence region on ε explicit),

$$P\left(S(F_{Y_0}, F_{X_0'q})q \in \mathrm{CR}^{\underline{\varepsilon}}_{1-\alpha}(\beta_0)\right) \geq P\left(S(F_{Y_0}, F_{X_0'q})q \in \mathrm{CR}^{\varepsilon(q)}_{1-\alpha}(\beta_0)\right),$$

which ensures the validity of using $\underline{\varepsilon}$ instead of a fixed ε.

2. Asymptotic validity of the confidence interval

Let $\beta_k \in \mathcal{B}_k$. First assume that $\beta_k \leq 0$. Because $0 \in \mathrm{CI}_{1-\alpha}(\beta_{0,k})$, $\beta_k \notin \mathrm{CI}_{1-\alpha}(\beta_{0,k})$ only if

$$\beta_k < -\sigma_\varepsilon(-e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) + n^{-1/2}\widetilde{c}_{\alpha,\varepsilon}(-e_k).$$

In turn, this event implies that $\underline{E}_n$ holds, with

$$\underline{E}_n := \left\{ n^{1/2} \left(-\sigma_\varepsilon(-e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) + \sigma(-e_k, F_{Y_0}, F_{X_0}) \right) > -\widetilde{c}_{\alpha,\varepsilon}(-e_k) \right\}. \tag{42}$$

Hence, $\sup_{\beta_k \in \mathcal{B}_k \cap \mathbb{R}^-} P(\beta_k \notin \mathrm{CI}_{1-\alpha}(\beta_{0,k})) \leq P\left(\underline{E}_n\right)$. Reasoning similarly for $\beta_k \geq 0$, we obtain

$$\sup_{\beta_k \in \mathcal{B}_k} P(\beta_k \notin \mathrm{CI}_{1-\alpha}(\beta_{0,k})) \leq \max\left[P\left(\underline{E}_n\right), P\left(\overline{E}_n\right) \right],$$

where we let $\overline{E}_n := \left\{ n^{1/2} \left(\sigma_\varepsilon(e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) - \sigma(e_k, F_{Y_0}, F_{X_0}) \right) < \widetilde{c}_{\alpha,\varepsilon}(e_k) \right\}$. As the reasoning is similar for $\underline{E}_n$ and $\overline{E}_n$, it suffices to prove that $\limsup_{n\to\infty} P(\underline{E}_n) \leq \alpha$, with equality if $\sigma(-e_k, F_{Y_0}, F_{X_0}) = \sigma_\varepsilon(-e_k, F_{Y_0}, F_{X_0})$. To this end, first remark that

$$\sigma_\varepsilon(-e_k, F_{Y_0}, F_{X_0}) = \sup_{q \in \mathcal{S}} \inf_{\alpha \in [\varepsilon, 1-\varepsilon]} \left[R(\alpha, F_{Y_0}, F_{X_0'q})q \right]'(-e_k).$$

Let us define $\kappa(f) := \sup_{q\in\mathcal{S}} \inf_{\alpha\in[\varepsilon,1-\varepsilon]} f(q,\alpha)$ and $G(q,\alpha) := [R(\alpha, F_{Y_0}, F_{X_0'q})q]'(-e_k)$. By Lemma B.1 in Firpo et al. (2023), κ is Hadamard directionally differentiable. Moreover, by Lemma 4 in the supplementary material, the process $(q,\alpha) \mapsto [\mathbb{F}_n(q,\alpha)q]'(-e_k)$ converges weakly to a Gaussian process ($\widetilde{\mathbb{F}}$, say). Then, as above,

$$n^{1/2} \left(\sigma_\varepsilon(-e_k, \widehat{F}_{Y_0}, \widehat{F}_{X_0}) - \sigma_\varepsilon(-e_k, F_{Y_0}, F_{X_0}) \right) \xrightarrow{d} \kappa'_G(\widetilde{\mathbb{F}}), \tag{43}$$

where the expression of κ' is given by (3.10) in Firpo et al. (2023).

Now, suppose that (i) in Assumption 4 holds: $\sigma_\varepsilon(-e_k, F_{Y_0}, F_{X_0}) > \sigma(-e_k, F_{Y_0}, F_{X_0})$. By, e.g., Theorem 2.2.1 in Politis et al. (1999), the subsampling counterpart of (43) holds. This implies that $\widetilde{c}_{\alpha,\varepsilon}(-e_k) = O_P(1)$. Combined with (42) and (43), this implies that $P(\underline{E}_n) \to 0$.

Next, suppose that (ii) in Assumption 4 holds. Then, in view of (3.10) in Firpo et al. (2023) and since G and $\widetilde{\mathbb{F}}$ are continuous, $\kappa'_G(\widetilde{\mathbb{F}}) = \min_{\alpha\in[\varepsilon,1-\varepsilon]} \widetilde{\mathbb{F}}(\widetilde{q},\alpha)$, where $\widetilde{q}$ is the only $q \in \mathcal{S}$ such that $\inf_{\alpha\in[\varepsilon,1-\varepsilon]} G(\widetilde{q},\alpha) = \kappa(G)$. Because $\widetilde{\mathbb{F}}(\widetilde{q},\cdot)$ is Gaussian, the same reasoning as in Point 2 above applies, and $\widetilde{c}_{\alpha,\varepsilon}(-e_k) \xrightarrow{\mathbb{P}} c^s_{\alpha,\varepsilon}(-e_k)$, the quantile of order α of $\kappa'_G(\widetilde{\mathbb{F}})$. Then, $P(E_n) \to \alpha$.

Finally, suppose (iii) in Assumption 4 holds. We then obtain, still using (3.10) in Firpo et al. (2023),

$$\kappa'_G(\widetilde{\mathbb{F}}) = \max_{q_m \in \arg\max_{q\in\mathcal{S}}[qS_\varepsilon(F_{Y_0}, F_{X'_0 q})]'(-e_k)} \widetilde{\mathbb{F}}(q_m, a_\varepsilon(q_m)), \tag{44}$$

where for each $q_m \in \arg\max_{q\in\mathcal{S}}[qS_\varepsilon(F_{Y_0}, F_{X'_0 q})]'(-e_k)$, $a_\varepsilon(q_m)$ is the only $a \in (\varepsilon, 1-\varepsilon)$ such that $R(a_\varepsilon(q_m), F_Y, F_{X'q_m}) = \inf_{\alpha\in[\varepsilon,1-\varepsilon]} R(\alpha, F_Y, F_{X'q_m})$. Because $\widetilde{\mathbb{F}}(\cdot, a_\varepsilon(\cdot))$ is Gaussian, the same reasoning as above applies once more and again, $P(E_n) \to \alpha$.

To conclude the proof, we show the validity of using $\varepsilon(e)$ instead of a fixed ε under Assumption 6. We do this by proving that we still have $\limsup_{n\to\infty} P(\underline{E}_n^{\varepsilon(e)}) \leq \alpha$, now indexing $\underline{E}_n$ by ε to avoid any ambiguity. If (i) of Assumption 6 holds for all $\varepsilon \in \mathcal{E}$, we have, by what precedes,

$$P(\underline{E}_n^{\varepsilon(e)}) \leq P\left(\cup_{\varepsilon\in\mathcal{E}} \underline{E}_n^{\varepsilon}\right) \leq \sum_{\varepsilon\in\mathcal{E}} P(\underline{E}_n^{\varepsilon}) \to 0. \tag{45}$$

Otherwise, (ii) in Assumption 6 holds. Let ε_{j_0} be as in Assumption 6 and let us first show that for all $j \in \{1, \cdots, j_0\}$, $\mathcal{Q}_j = \mathcal{Q}_{j_0}$, with $\mathcal{Q}_j := \arg\max_{q\in\mathcal{S}}[S_{\varepsilon_j}(F_{Y_0}, F_{X'_0 q})q]'(-e_k)$. First, for all $q_m \in \mathcal{Q}_j$ and using that $q'_m(-e_k) \geq 0$,

$$\begin{aligned}
[S_{\varepsilon_{j_0}}(F_{Y_0}, F_{X'_0 q_m})q_m]'(-e_k) &\geq [S_{\varepsilon_j}(F_{Y_0}, F_{X'_0 q_m})q_m]'(-e_k) \\
&= \sigma_{\varepsilon_j}(-e_k, F_{Y_0}, F_{X_0}) \\
&= \sigma_{\varepsilon_{j_0}}(-e_k, F_{Y_0}, F_{X_0}).
\end{aligned}$$

Thus, $q_m \in \mathcal{Q}_{j_0}$ and $\mathcal{Q}_j \subset \mathcal{Q}_{j_0}$. Conversely, for any $q_m \in \mathcal{Q}_{j_0}$, by assumption,

$$\begin{aligned}
S_{\varepsilon_j}(F_{Y_0}, F_{X'_0 q_m}) &= \min_{a\in[\varepsilon_j, 1-\varepsilon_j]} R(a, F_{Y_0}, F_{X'_0 q_m}) \\
&= R(a(q_m), F_{Y_0}, F_{X'_0 q_m}) \\
&= S_{\varepsilon_{j_0}}(F_{Y_0}, F_{X'_0 q_m}).
\end{aligned}$$

As a result,

$$\begin{aligned}
[S_{\varepsilon_j}(F_{Y_0}, F_{X'_0 q_m})q_m]'(-e_k) &= [S_{\varepsilon_{j_0}}(F_{Y_0}, F_{X'_0 q_m})q_m]'(-e_k) \\
&= \sigma_{\varepsilon_{j_0}}(-e_k, F_{Y_0}, F_{X_0}) \\
&= \sigma_{\varepsilon_j}(-e_k, F_{Y_0}, F_{X_0}),
\end{aligned}$$

implying that $q_m \in \mathcal{Q}_j$. Thus, $\mathcal{Q}_{j_0} \subset \mathcal{Q}_j$ and then $\mathcal{Q}_{j_0} = \mathcal{Q}_j$. Now, by (44) but making the dependence on ε explicit, we have, for all $j \leq j_0$,

$$\kappa'_{\varepsilon_j,G}(\widetilde{\mathbb{F}}) = \max_{q_m \in \mathcal{Q}_j} \widetilde{\mathbb{F}}(q_m, a_{\varepsilon_j}(q_m)).$$

Hence, $\kappa'_{\varepsilon_1,G}(\widetilde{\mathbb{F}}) = \cdots = \kappa'_{\varepsilon_{j_0},G}(\widetilde{\mathbb{F}})$. Reasoning as in Point 2 above, we obtain $P(\underline{E}_n^{\varepsilon_j}) = P(\underline{E}_n^{\varepsilon_1}) + o(1)$. Then,

$$\begin{aligned} P(\underline{E}_n^{\varepsilon(e)}) &= \sum_{j=1}^{J} P\left(\underline{E}_n^{\varepsilon_j}, \varepsilon(e) = \varepsilon_j\right) \\ &= \sum_{j=1}^{j_0} P\left(\underline{E}_n^{\varepsilon_j}, \varepsilon(e) = \varepsilon_j\right) + o(1) \\ &= \sum_{j=1}^{j_0} P\left(\underline{E}_n^{\varepsilon_1}, \varepsilon(e) = \varepsilon_j\right) + o(1) \\ &\leq P\left(\underline{E}_n^{\varepsilon_1}\right) + o(1), \end{aligned}$$

where the second equality holds since when $\varepsilon(e) > \varepsilon_{j_0}$, $\sigma_{\varepsilon(e)}(-e_k, F_{Y_0}, F_{X_0}) > \sigma(-e_k, F_{Y_0}, F_{X_0})$ and we can apply the same reasoning leading to (45). The result follows since by what precedes, $\liminf_{n\to\infty} P\left(\underline{E}_n^{\varepsilon_1}\right) \leq \alpha$.

Supplementary material

(not for publication)

1. Complements on the proof of Proposition 7

Lemma 1. *For any random variables U_1 and U_2, $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, we have:*

$$F^{-1}_{U_1+U_2}(\alpha) \leq F^{-1}_{U_1}(\alpha+\beta) + F^{-1}_{U_2}(1-\beta). \tag{46}$$

Proof: Fix $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$. We have

$$\begin{aligned}
&P(U_1 + U_2 \leq F^{-1}_{U_1}(\alpha+\beta) + F^{-1}_{U_2}(1-\beta)) \\
&\geq P(U_1 \leq F^{-1}_{U_1}(\alpha+\beta),\, U_2 \leq F^{-1}_{U_2}(1-\beta)) \\
&\geq P(U_1 \leq F^{-1}_{U_1}(\alpha+\beta)) + P(U_2 \leq F^{-1}_{U_2}(1-\beta)) - 1 \\
&\geq \alpha.
\end{aligned}$$

Equation (46) follows by definition of quantiles.

We now establish an upper bound on $|F^{-1}_{Y_0}(\alpha)/F^{-1}_{X_0'\beta_0}(\alpha) - 1|$ for α or $1-\alpha$ small:

Lemma 2. *Under the assumptions of Proposition 7.1 and with α_0 and α_1 as in the proof of Proposition 7, we have:*

$$\forall \alpha \geq \alpha_0,\ \left| \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)} - 1 \right| \lesssim (1-\alpha)^{\frac{1/c-1/d}{1+1/d}}, \quad \forall \alpha \leq \alpha_1,\ \left| \frac{F^{-1}_{Y_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)} - 1 \right| \lesssim \alpha^{\frac{1/c-1/d}{1+1/d}}.$$

Proof: we focus hereafter on the case $\alpha \geq \alpha_0$; the other case can be treated similarly. Fix $\gamma > 1$ and note that by Lemma 1, we have

$$F^{-1}_{Y_0}(\alpha) \leq F^{-1}_{X_0'\beta_0}(\alpha + (1-\alpha)^\gamma) + F^{-1}_{U}(1-(1-\alpha)^\gamma).$$

Moreover, since $X_0'\beta_0 = Y_0 - U$,

$$F^{-1}_{X_0'\beta_0}(\alpha - (1-\alpha)^\gamma) \leq F^{-1}_{Y_0}(\alpha) + F^{-1}_{-U}(1-(1-\alpha)^\gamma).$$

Thus,

$$\frac{F^{-1}_{X_0'\beta_0}(\alpha - (1-\alpha)^\gamma) - F^{-1}_{X_0'\beta_0}(\alpha)}{F^{-1}_{X_0'\beta_0}(\alpha)} - \frac{F^{-1}_{-U}(1-(1-\alpha)^\gamma)}{F^{-1}_{X_0'\beta_0}(\alpha)}$$

$$\leq \frac{F_{Y_0}^{-1}(\alpha)}{F_{X_0'\beta_0}^{-1}(\alpha)} - 1 \leq \frac{F_{X_0'\beta_0}^{-1}(\alpha+(1-\alpha)^\gamma) - F_{X_0'\beta_0}^{-1}(\alpha)}{F_{X_0'\beta_0}^{-1}(\alpha)} + \frac{F_U^{-1}(1-(1-\alpha)^\gamma)}{F_{X_0'\beta_0}^{-1}(\alpha)}. \quad (47)$$

Now, the tail conditions imply that $F_{X_0'\beta_0}^{-1}(\alpha) \gtrsim (1-\alpha)^{-1/c}$ and $F_U^{-1}(\alpha) \lesssim (1-\alpha)^{-1/d}$. As a result, we get, for some $h(\alpha) \in (0, (1-\alpha)^\gamma)$,

$$\begin{aligned}
\frac{F_{Y_0}^{-1}(\alpha)}{F_{X_0'\beta_0}^{-1}(\alpha)} - 1 &\leq \frac{(1-\alpha)^\gamma}{f_{X'\beta_0}(F_{X_0'\beta_0}^{-1}(\alpha+h(\alpha)))\, F_{X_0'\beta_0}^{-1}(\alpha)} + \frac{F_U^{-1}(1-(1-\alpha)^\gamma)}{F_{X_0'\beta_0}^{-1}(\alpha)} \\
&\lesssim (1-\alpha)^\gamma (1-\alpha-h(\alpha))^{-1-1/c}(1-\alpha)^{1/c} + (1-\alpha)^{1/c}(1-\alpha)^{-\gamma/d} \\
&\lesssim (1-\alpha)^{\gamma-1} + (1-\alpha)^{1/c-\gamma/d}.
\end{aligned}$$

Choosing $\gamma = (1+1/c)/(1+1/d)$ yields

$$\frac{F_{Y_0}^{-1}(\alpha)}{F_{X_0'\beta_0}^{-1}(\alpha)} - 1 \lesssim (1-\alpha)^{\frac{1/c-1/d}{1+1/d}}.$$

The exact same reasoning with the lower bound in (47) finally yields the result.

2. Complements on the proof of Theorem 2

Lemma 3. *R is continuous for the metric d' defined by* (37).

Proof: First, remark that for all $a, a', b, b' > 0$, we have

$$\left|\frac{a'}{b'} - \frac{a}{b}\right| \leq \frac{1}{b}\left[|a'-a| + \left|\frac{a'}{b'} - \frac{a}{b}\right| |b'-b| + \frac{a}{b}|b'-b|\right]. \quad (48)$$

Therefore, if $|b'-b| < b$,

$$\left|\frac{a'}{b'} - \frac{a}{b}\right| \leq \frac{|a'-a| + a/b\,|b'-b|}{b - |b'-b|}.$$

Fix $\alpha \in [\varepsilon, 1-\varepsilon]$, F and G and let G' be such that $W_1(G,G') < (1/4)\int_\alpha^1 G^{-1}(t)dt$. Let also $\alpha' \in [\varepsilon, 1-\varepsilon]$ be such that

$$\left|\int_\alpha^{\alpha'} G^{-1}(t)dt\right| < \frac{1}{2}\int_\alpha^1 G^{-1}(t)dt.$$

Then,

$$\int_\alpha^1 G^{-1}(t)dt = \int_\alpha^{\alpha'} G^{-1}(t)dt + \int_{\alpha'}^1 G^{-1}(t)dt < \frac{1}{2}\int_\alpha^1 G^{-1}(t)dt + \int_{\alpha'}^1 G^{-1}(t)dt.$$

Thus, $\int_\alpha^1 G^{-1}(t)dt < 2\int_{\alpha'}^1 G^{-1}(t)dt$. Moreover, since $W_1(F,F') = \int_0^1 |F^{-1}(t) - F'^{-1}(t)|dt$, we have

$$\left|\int_{\alpha'}^1 G'^{-1}(t) - G^{-1}(t)dt\right| \leq W_1(G,G') < \frac{1}{2}\int_{\alpha'}^1 G^{-1}(t)dt.$$

Let $c_F = |F^{-1}(\varepsilon)| \vee |F^{-1}(1-\varepsilon)|$ and define c_G similarly. Then, using (48), we get

$$\begin{aligned}
|R(\alpha',F,G) - R(\alpha,F,G)| &\leq \frac{\left|\int_\alpha^{\alpha'} F^{-1}(t)dt\right| + R(\alpha,F,G)\left|\int_\alpha^{\alpha'} G^{-1}(t)dt\right|}{\int_\alpha^1 G^{-1}(t)dt - \left|\int_\alpha^{\alpha'} G^{-1}(t)dt\right|} \\
&\leq \frac{|\alpha'-\alpha|\left(|F^{-1}(\alpha)| \vee |F^{-1}(\alpha')| + R(\alpha,F,G)|G^{-1}(\alpha)| \vee |G^{-1}(\alpha')|\right)}{1/2\int_\alpha^1 G^{-1}(t)dt} \\
&\leq \frac{2|\alpha'-\alpha|\left(c_F + R(\alpha,F,G)c_G\right)}{\int_\alpha^1 G^{-1}(t)dt}. \qquad (49)
\end{aligned}$$

Next, for any F', using again (48),

$$\begin{aligned}
|R(\alpha',F',G') - R(\alpha',F,G)| &\leq \frac{\left|\int_{\alpha'}^1 F^{-1}(t) - F'^{-1}(t)dt\right| + R(\alpha',F,G)\left|\int_{\alpha'}^1 G^{-1}(t) - G'^{-1}(t)dt\right|}{\int_{\alpha'}^1 G^{-1}(t)dt - \left|\int_{\alpha'}^1 G'^{-1}(t) - G^{-1}(t)dt\right|} \\
&\leq \frac{W_1(F,F') + R(\alpha',F,G)W_1(G,G')}{1/4\int_\alpha^1 G^{-1}(t)dt} \\
&\leq \frac{4}{\int_\alpha^1 G^{-1}(t)dt}\left[W_1(F,F') + \left(\frac{2|\alpha'-\alpha|\left(c_F + R(\alpha,F,G)c_G\right)}{\int_\alpha^1 G^{-1}(t)dt}\right.\right. \\
&\qquad \left.\left. + R(\alpha,F,G)\right)W_1(G,G')\right]. \qquad (50)
\end{aligned}$$

The result follows by Inequalities (49) and (50) and the triangle inequality.

3. Complements on the proof of Theorem 3

Lemma 4. *Fix $\varepsilon \in (0,1/2)$ and suppose that $n_X/(n_X+n_Y) \to \mu \in (0,1)$ and Assumptions 1-2 and 4 hold. Then, $\mathbb{F}_n$, as a process indexed by $(q,\alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]$, converges weakly to a Gaussian process $\mathbb{F}$. The same holds but for $\mathbb{F}_n$ indexed by $\alpha \in [\varepsilon, 1-\varepsilon]$ only if Assumption 4 is replaced by Assumption 3.*

Proof: First, $R(\alpha, F_{Y_0}, F_{X_0'q}) = \theta_1(q,\alpha)/\theta_2(q,\alpha)$, where $\theta_1(q,\alpha) = \int_\alpha^1 F_{Y_0}^{-1}(t)dt$, $\theta_2(q,\alpha) = \int_\alpha^1 F_{X_0'q}^{-1}(t)dt$ and we suppress the dependence of θ_1 and θ_2 in F_{Y_0} and $F_{X_0'q}$ for simplicity. Moreover, $R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) = \widehat{\theta}_1(q,\alpha)/\widehat{\theta}_2(q,\alpha)$ with $\widehat{\theta}_1(q,\alpha) = \int_\alpha^1 \widehat{F}_{Y_0}^{-1}(t)dt$ and $\widehat{\theta}_2(q,\alpha) = \int_\alpha^1 \widehat{F}_{X_0'q}^{-1}(t)dt$. The map $(U,V) \mapsto U/V$, from $\ell^\infty(\mathcal{S} \times [\varepsilon,1-\varepsilon])^2$ to $\ell^\infty(\mathcal{S} \times [\varepsilon,1-\varepsilon])$, is Hadamard differentiable at any (U,V) such that

$\inf_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} V(q,\alpha) > 0$. Now, $\theta_2(\cdot,\alpha)$ is continuous (see the proof of Proposition 6). $\theta_2(q,\cdot)$ is also continuous. Thus,

$$\inf_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} \theta_2(q,\alpha) = \min_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} \theta_2(q,\alpha) > 0.$$

Hence, by the functional delta method, $\mathbb{F}_n$ converges weakly as long as

$$n^{1/2}\left(\widehat{\theta}_1(q,\alpha) - \theta_1(q,\alpha), \widehat{\theta}_2(q,\alpha) - \theta_2(q,\alpha)\right)$$

converges weakly. By independence of the two samples, it suffices to show the weak convergence of each component. We focus on the second hereafter, as the proof is similar (and actually simpler) for the first. Also, it suffices to show the weak convergence of $n_X^{1/2}\left(\widehat{\theta}_2(q,\alpha) - \theta_2(q,\alpha)\right)$, as $n/n_X \to 1-\mu$ by assumption.

Let us define

$$\widetilde{\theta}_2(q,\alpha) = \frac{1}{n_X}\sum_{i=1}^{n_X}\left(X_i'q - \overline{X'q}\right)\mathbb{1}\left\{\widehat{F}_{X'q}(X_i'q) > \alpha\right\}.$$

Because $F_{X'q}$ is continuous, almost surely there are no ties and $\widehat{\theta}_2(q,\alpha) = \widetilde{\theta}_2(q,\alpha)$ for all $\alpha \in \{0/n_X, ..., (n_X-1)/n_X\}$. Elsewhere, if $\alpha = [ti + (1-t)(i+1)]/n_X$, $t \in (0,1)$, we have $\widehat{\theta}_2(q,\alpha) = t\widetilde{\theta}_2(q,i/n_X) + (1-t)\widetilde{\theta}_2(q,(i+1)/n_X)$. As a result,

$$\begin{aligned} n_X^{1/2}\sup_{\alpha\in[\varepsilon,1-\varepsilon]}\left|\widehat{\theta}_2(q,\alpha) - \widetilde{\theta}_2(q,\alpha)\right| \leq & \frac{\sup_{i=\lfloor n\varepsilon\rfloor,...,\lceil n(1-\varepsilon)\rceil}\left|(X'q)_{(i)} - \overline{X'q}\right|}{n_X^{1/2}} \\ \leq & \frac{\left|(X'q)_{(n)} - (X'q)_{(1)}\right|}{n_X^{1/2}} \\ \xrightarrow{\mathbb{P}} & 0, \end{aligned}$$

where the convergence follows by, e.g., Problem 2.3.4 in Van der Vaart and Wellner (1996). Hence, it suffices to show the weak convergence of $n_X^{1/2}(\widetilde{\theta}_2(q,\alpha) - \theta_2(q,\alpha))$. By, e.g. Lemma 21.1 in Van der Vaart (2000),

$$\theta_2(q,\alpha) = E\left[(X'q - E(X'q))\mathbb{1}\left\{F_{X'q}(X'q) \geq \alpha\right\}\right].$$

As a result,

$$n_X^{1/2}\left(\widetilde{\theta}_2(q,\alpha) - \theta_2(q,\alpha)\right) = \mathbb{G}_{n_X} g_{q,\alpha} + R_{n_X}(q,\alpha),$$

where $\mathbb{G}_{n_X}$ denotes the empirical process associated to $(X_1, ..., X_{n_X})$ and

$$\begin{aligned} g_{q,\alpha}(x) =& \left[F^{-1}_{X'q}(\alpha) - E(X'q)\right] \mathbb{1}\left\{F_{X'q}(x'q) \leq \alpha\right\} - (1-\alpha)x'q \\ &+ (x'q - E(X'q))\mathbb{1}\left\{F_{X'q}(x'q) > \alpha\right\}, \\ R_{n_X}(q,\alpha) =& \frac{1}{n_X^{1/2}} \sum_{i=1}^{n_X} \Big\{ \left(X_i'q - \overline{X'q}\right) \left[\mathbb{1}\left\{F_{X'q}(X_i'q) \leq \alpha\right\} - \mathbb{1}\left\{\widehat{F}_{X'q}(X_i'q) \leq \alpha\right\}\right] \\ &- \left[F^{-1}_{X'q}(\alpha) - E(X'q)\right] \left(\mathbb{1}\left\{F_{X'q}(X_i'q) \leq \alpha\right\} - \alpha\right) \Big\} \\ &+ \frac{n_X^{1/2}\left(\overline{X'q} - E(X'q)\right)}{n_X} \sum_{i=1}^{n_X} \left(\mathbb{1}\left\{F_{X'q}(X_i'q) \leq \alpha\right\} - \alpha\right). \end{aligned}$$

We first prove that the class $\mathcal{G} = \{g_{q,\alpha} : (q,\alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]\}$ is Donsker. The class $\mathcal{I}_0 = \{x \mapsto \mathbb{1}\{x'q \leq u\} : (q,u) \in \mathcal{S} \times \mathbb{R}\}$ is Donsker by Problem 2.6.14 and Theorem 2.6.8 in Van der Vaart and Wellner (1996). Then, $\mathcal{I}_1 = \{x \mapsto \mathbb{1}\{F_{X'q}(x'q) \leq \alpha\} : (q,\alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]\} \subset \mathcal{I}_0$ is also Donsker (see, e.g., Theorem 2.10.1 in Van der Vaart and Wellner, 1996). Similarly, $\mathcal{I}_2 = \{x \mapsto \mathbb{1}\{F_{X'q}(x'q) > \alpha\} : (q,\alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]\}$ is Donsker. $\mathcal{I}_2$ also has a finite integral entropy and an envelope of 1. Since $\{x \mapsto x'q : q \in \mathcal{S}\}$ also has a finite integral entropy with envelope $x \mapsto \|x\|$, and $E[\|X\|^2] < \infty$, the class $\mathcal{I}_3 = \{x \mapsto (x'q)\mathbb{1}\{F_{X'q}(x'q) > \alpha\} : (q,\alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]\}$ is also Donsker (see Example 19.19 in Van der Vaart, 2000). Because $\{x \mapsto (1-\alpha)x'q : (q,\alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]\}$ is also Donsker and sums of Donsker classes are also Donsker, we finally get that $\mathcal{G}$ is Donsker.

Next, we consider the remainder term $R_{n_X}(q,\alpha)$. Let $I_i(q,\alpha) = \mathbb{1}\{F_{X'q}(X_i'q) \leq \alpha\}$ and $\widehat{I}_i(q,\alpha) = \mathbb{1}\left\{\widehat{F}_{X'q}(X_i'q) \leq \alpha\right\}$. We have $R_{n_X}(q,\alpha) = R_{1n_X} + R_{2n_X} + R_{3n_X}$, with

$$\begin{aligned} R_{1n_X}(q,\alpha) &= \frac{1}{n_X^{1/2}} \sum_{i=1}^{n_X} (I_i(q,\alpha) - \widehat{I}_i(q,\alpha)) \left[\left(X_i'q - \overline{X'q}\right) - \left(F^{-1}_{X'q}(\alpha) - E(X'q)\right)\right], \\ R_{2n_X}(q,\alpha) &= \frac{\left(F^{-1}_{X'q}(\alpha) - E(X'q)\right)}{n_X^{1/2}} \sum_{i=1}^{n_X} \left[\alpha - \widehat{I}_i(q,\alpha)\right], \\ R_{3n_X}(q,\alpha) &= \frac{n_X^{1/2}\left(\overline{X'q} - E(X'q)\right)}{n_X} \sum_{i=1}^{n_X} \left(I_i(q,\alpha) - \alpha\right). \end{aligned}$$

We now prove that for all $k \in \{1,2,3\}$,

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} R_{kn_X}(q,\alpha) = o_P(1). \tag{51}$$

Consider R_{2n_X} first. By definition of the empirical cdf., we have, for all (q, α),

$$\left|\sum_{i=1}^{n_X}\left(\widehat{I}_i(q,\alpha)-\alpha\right)\right| = \lceil n_X\alpha\rceil - n_X\alpha < 1. \tag{52}$$

As a result,

$$\begin{aligned}\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]}|R_{2n_X}(q,\alpha)| \leq& \frac{F^{-1}_{\|X\|}(1-\varepsilon)+E(\|X\|)}{n_X^{1/2}} \times \sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]}\left|\sum_{i=1}^{n_X}\left(\widehat{I}_i(q,\alpha)-\alpha\right)\right| \\ \leq& \frac{F^{-1}_{\|X\|}(1-\varepsilon)+E(\|X\|)}{n_X^{1/2}},\end{aligned}$$

where the first inequality follows from the triangle and Cauchy-Schwarz inequalities and $|F^{-1}_{X'q}(\varepsilon)| \vee |F^{-1}_{X'q}(1-\varepsilon)| \leq F^{-1}_{\|X\|}(1-\varepsilon)$. Hence, (51) holds for $k = 2$.

Next, consider R_{3n_X}. We have

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]}|R_{3n_X}(q,\alpha)| \leq n_X^{1/2}\|\overline{X}-E(X)\| \times \sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]}\left|\frac{1}{n_X}\sum_{i=1}^{n_X}\left(I_i(q,\alpha)-\alpha\right)\right|.$$

The first term is an $O_P(1)$. Recall that the class $\mathcal{I}_1$ is Donsker; hence it is also Glivenko-Cantelli. Therefore, the second term is an $o_P(1)$. Therefore, (51) holds for $k = 3$.

Finally, consider R_{1n_X}. We first decompose it further into $R_{11n_X} + R_{12n_X}$, with

$$\begin{aligned}R_{11n_X}(q,\alpha) &= \frac{-n_X^{1/2}(\overline{X'q}-E(X'q))}{n_X}\sum_{i=1}^{n_X}[I_i(q,\alpha)-\widehat{I}_i(q,\alpha)], \\ R_{12n_X}(q,\alpha) &= \frac{1}{n_X^{1/2}}\sum_{i=1}^{n_X}(I_i(q,\alpha)-\widehat{I}_i(q,\alpha))\left(X_i'q - F^{-1}_{X'q}(\alpha)\right).\end{aligned}$$

That R_{11n_X} is uniformly negligible follows by writing $I_i(q,\alpha)-\widehat{I}_i(q,\alpha) = I_i(q,\alpha) - \alpha + \alpha - \widehat{I}_i(q,\alpha)$, reasoning as for R_{3n_X} and using (52). For R_{12n_X}, remark that by definition of $I_i(q,\alpha)$ and continuity of $X_i'q$, $I_i(q,\alpha) = \mathbb{1}\left\{X_i'q \leq F^{-1}_{X'q}(\alpha)\right\}$. Similarly, but accounting for the discontinuity of $\widehat{F}_{X'q}$, we have

$$\widehat{I}_i(q,\alpha) = \left|\begin{array}{ll}\mathbb{1}\left\{X_i'q < \widehat{F}^{-1}_{X'q}(\alpha)\right\} & \text{if } n_X\alpha \notin \mathbb{N}, \\ \mathbb{1}\left\{X_i'q \leq \widehat{F}^{-1}_{X'q}(\alpha)\right\} & \text{otherwise.}\end{array}\right.$$

As a result,

$$\sum_{i=1}^{n_X}\left|I_i(q,\alpha)-\widehat{I}_i(q,\alpha)\right| = \left(2\mathbb{1}\left\{F^{-1}_{X'q}(\alpha) \geq \widehat{F}^{-1}_{X'q}(\alpha)\right\}-1\right)\left(\sum_{i=1}^{n_X}I_i(q,\alpha)-\widehat{I}_i(q,\alpha)\right)$$

$$= \left| \sum_{i=1}^{n_X} I_i(q,\alpha) - \widehat{I}_i(q,\alpha) \right|.$$

Moreover, $|I_i(q,\alpha) - \widehat{I}_i(q,\alpha)| = 1$ only if $X_i'q \in J$, the interval $[\widehat{F}^{-1}_{X'q}(\alpha), F^{-1}_{X'q}(\alpha)]$ if $\widehat{F}^{-1}_{X'q}(\alpha) < F^{-1}_{X'q}(\alpha)$ and $[F^{-1}_{X'q}(\alpha), \widehat{F}^{-1}_{X'q}(\alpha)]$ otherwise. As a result,

$$\begin{aligned} |R_{12n_X}| \leq & \frac{1}{n_X^{1/2}} \sum_{i=1}^{n_X} \left| I_i(q,\alpha) - \widehat{I}_i(q,\alpha) \right| \left| X_i'q - F^{-1}_{X'q}(\alpha) \right| \\ & \leq \left| \widehat{F}^{-1}_{X'q}(\alpha) - F^{-1}_{X'q}(\alpha) \right| \times \left| \frac{1}{n_X^{1/2}} \sum_{i=1}^{n_X} (I_i(q,\alpha) - \widehat{I}_i(q,\alpha)) \right|. \end{aligned} \tag{53}$$

By (52) and the fact that $\mathcal{I}_1$ is a Donsker class,

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} \left| \frac{1}{n_X^{1/2}} \sum_{i=1}^{n_X} (I_i(q,\alpha) - \widehat{I}_i(q,\alpha)) \right| = O_P(1).$$

Thus, the result holds as long as

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} \left| \widehat{F}^{-1}_{X'q}(\alpha) - F^{-1}_{X'q}(\alpha) \right| = o_P(1). \tag{54}$$

To prove this, note first that the class $\{x \mapsto \mathbb{1}\{x'q \leq \alpha\} : (q,\alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]\}$ is Glivenko-Cantelli (as it is Donsker). Hence,

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} \left| F_{X'q}(\alpha) - \widehat{F}_{X'q}(\alpha) \right| = o_P(1). \tag{55}$$

Now, let $U_q = F_{X'q}(X'q)$ and $U_{q,1} < ... < U_{q,n_X}$ denote the corresponding order statistic. Remark that $\widehat{F}^{-1}_{X'q}(\alpha) = F^{-1}_{X'q}(U_{q,\lceil n_X\alpha\rceil})$. Also, note that $\inf_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} U_{q,\lceil n_X\alpha\rceil} < \varepsilon'$ implies that for some $q_0 \in \mathcal{S}$, $\widehat{F}_{X'q_0}(F^{-1}_{X'q_0}(\varepsilon')) \geq \lceil n_X\alpha\rceil/n_X$ and thus

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} \left| F_{X'q}(\alpha) - \widehat{F}_{X'q}(\alpha) \right| > \varepsilon - \varepsilon'.$$

In view of (55), this occurs with probability approaching zero. The same is true for the event $\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} U_{q,\lceil n_X\alpha\rceil} > 1-\varepsilon'$. Hence, with probability approaching one,

$$\varepsilon' \leq \inf_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} U_{q,\lceil n_X\alpha\rceil} \leq \sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} U_{q,\lceil n_X\alpha\rceil} \leq 1-\varepsilon'. \tag{56}$$

Moreover, under this event,

$$\left| \widehat{F}^{-1}_{X'q}(\alpha) - F^{-1}_{X'q}(\alpha) \right| = \left| F^{-1}_{X'q}(U_{q,\lceil n_X\alpha\rceil}) - F^{-1}_{X'q}(\alpha) \right|$$

$$
\begin{aligned}
&< m\left(\left|U_{q,\lceil n_X\alpha\rceil} - \alpha\right|\right) \\
&\leq m\left(\left|F_{X'q}((X'q)_{\lceil n_X\alpha\rceil}) - \widehat{F}_{X'q}((X'q)_{\lceil n_X\alpha\rceil})\right|\right. \\
&\qquad \left. + \left|\widehat{F}_{X'q}((X'q)_{\lceil n_X\alpha\rceil}) - \alpha\right|\right) \\
&< m\left(\sup_{q\in\mathcal{S}}\sup_{t\in\mathbb{R}}\left|F_{X'q}(t) - \widehat{F}_{X'q}(t)\right| + \frac{1}{n_X}\right).
\end{aligned}
$$

Using (55) and the continuity of m finally yields (54).

Finally, let us prove the weak convergence of $\mathbb{F}_n$ as a process indexed by $\alpha \in [\varepsilon, 1-\varepsilon]$ only, but under the weaker Assumption 3. It suffices to remark that all steps above still hold, except (54). Now, given that q is fixed, we only need to establish the weaker

$$
\sup_{\alpha\in[\varepsilon,1-\varepsilon]}\left|\widehat{F}^{-1}_{X'q}(\alpha) - F^{-1}_{X'q}(\alpha)\right| = o_P(1). \tag{57}
$$

Because $F^{-1}_{X'q}$ is continuous on $[\varepsilon, 1-\varepsilon]$ (as the inverse of $F_{X'q}$ is strictly increasing on its support by Assumption 3), it is uniformly continuous on $[\varepsilon, 1-\varepsilon]$. Now, note that

$$
\left|\widehat{F}^{-1}_{X'q}(\alpha) - F^{-1}_{X'q}(\alpha)\right| = \left|F^{-1}_{X'q}(U_{q,\lceil n_X\alpha\rceil}) - F^{-1}_{X'q}(\alpha)\right|.
$$

Moreover, $\sup_{\alpha\in[\varepsilon,1-\varepsilon]}\left|U_{q,\lceil n_X\alpha\rceil} - \alpha\right| = o_P(1)$. This implies that (57) holds.

4. Proof of Proposition 9

Our proof heavily draws on and use the same notation as in Theorem 3. It proceeds in four steps. First, we show that $h(\beta,\alpha) := E\left[X_{v0}\mathbb{1}\left\{F_{X'_{v0}\beta}(X'_{v0}\beta) \geq \alpha\right\}\right]$ is continuous. Second, we prove that $\theta_3(t,\alpha) := \theta_2(t\widehat{\beta}_v + (1-t)\beta_v)$ $(t \in [0,1])$ is differentiable as a function of $t \in (0,1)$. Third, we show that $\sqrt{n}\left(\widehat{\theta}(\widehat{\beta}_v,\alpha) - \theta(\beta_v,\alpha)\right)$ converges to a Gaussian process. Finally, we prove the two points of the proposition.

Step 1: Continuity of h.

More precisely, we prove below that h is continuous at any $(\beta_1,\alpha_1) \in K\times[\varepsilon,1-\varepsilon]$, with K convex compact including β_v in its interior and such that $\{\beta_1/\|\beta_1\| : \beta_1 \in K\} \subset \mathcal{V}$. By the triangle inequality, for any $(\beta_1,\alpha_1) \in K\times[\varepsilon,1-\varepsilon]$ and $(\beta_2,\alpha_2) \in K\times[\varepsilon,1-\varepsilon]$,

$$
\|h(\beta_1,\alpha_1) - h(\beta_2,\alpha_2)\| \leq \|h(\beta_1,\alpha_1) - h(\beta_1,\alpha_2)\| + \|h(\beta_1,\alpha_2) - h(\beta_2,\alpha_2)\|. \tag{58}
$$

Regarding the first term, and assuming without loss of generality that $\alpha_1 \leq \alpha_2$, we have

$$\begin{aligned}\|h(\beta_1,\alpha_1) - h(\beta_1,\alpha_2)\| =& \left\| E\left[X_{v0}\mathbb{1}\left\{\alpha_2 \geq F_{X'_{v0}\beta_1}(X'_{v0}\beta_1) \geq \alpha_1\right\}\right]\right\| \\ \leq& E\left[\|X_{v0}\|^2\right]^{1/2}(\alpha_2-\alpha_1)^{1/2}. \end{aligned} \tag{59}$$

Turning to the second term, we have

$$\begin{aligned}\|h(\beta_1,\alpha_2) - h(\beta_2,\alpha_2)\| \leq& E\left[\|X_{v0}\|^2\right]^{1/2}\Big[P\left(F_{X'_{v0}\beta_1}(X'_{v0}\beta_1) \geq \alpha_2 > F_{X'_{v0}\beta_2}(X'_{v0}\beta_2)\right) \\ &+P\left(F_{X'_{v0}\beta_2}(X'_{v0}\beta_2) \geq \alpha_2 > F_{X'_{v0}\beta_1}(X'_{v0}\beta_1)\right)\Big]^{1/2}.\end{aligned}$$

As both probabilities are similar, we only consider the first one, P_1 say. To simplify notation, let $\delta = \beta_2 - \beta_1$, $U_k := X'_{v0}\beta_k$, $F_k := F_{X'_{v0}\beta_k}(k=1,2)$ and $F_\delta := F_{X'_{v0}\delta}$. Fix $\eta \in (0, 1-\alpha_2)$ and let δ be such that

$$\|\delta\| \leq \frac{c\|\beta_1\|\eta^2}{2(\eta F^{-1}_{\|X_{v0}\|}(1-\eta/2) + E\left[\|X_{v0}\|\right])}, \tag{60}$$

where c is defined in Assumption 7. Then, we have

$$\begin{aligned}P_1 &\leq P\left(U_1 \in [\alpha_2, \alpha_2+\eta)\right) + P\left(U_1 \geq F_1^{-1}(\alpha_2+\eta), U_2 < F_2^{-1}(\alpha_2)\right) \\ &= \eta + P\left(U_1 \geq F_1^{-1}(\alpha_2+\eta), X'_{v0}\delta < F_2^{-1}(\alpha_2) - F_1^{-1}(\alpha_2+\eta)\right) \\ &\leq \eta + P\left(X'_{v0}\delta < F_1^{-1}(\alpha_2+\eta/2) + F_\delta^{-1}(1-\eta/2) - F_1^{-1}(\alpha_2+\eta)\right) \\ &\leq \eta + P\left(X'_{v0}\delta < -c\|\beta_1\|\eta/2 + F^{-1}_{\|X_{v0}\|}(1-\eta/2)\|\delta\|\right) \\ &\leq \eta + P\left(\|X_{v0}\| > c\|\beta_1\|\eta/(2\|\delta\|) - F^{-1}_{\|X_{v0}\|}(1-\eta/2)\right) \\ &\leq \eta + \frac{E\left[\|X_{v0}\|\right]}{c\|\beta_1\|\eta/(2\|\delta\|) - F^{-1}_{\|X_{v0}\|}(1-\eta/2)} \\ &\leq 2\eta. \end{aligned} \tag{61}$$

The second inequality follows from Lemma 1. The third uses $F_\delta(x) \leq F_{\|X_{v0}\|}(x/\|\delta\|)$, which implies $F_\delta^{-1}(1-\eta/2) \leq F^{-1}_{\|X_{v0}\|}(1-\eta/2)\|\delta\|$, and $F_1^{-1}(y) - F_1^{-1}(x) > c\|\beta_1\|(y-x)$ for $y > x$, which follows from Assumption 7 and $\beta_1/\|\beta_1\| \in \mathcal{V}$. The fourth inequality follows from the Cauchy-Schwarz inequality, and the fifth uses Markov's inequality and the fact that by (60), $c\|\beta_1\|\eta/(2\|\delta\|) - F^{-1}_{\|X_{v0}\|}(1-\eta/2) > 0$. The last inequality follows from (60). By combining (58), (59) and (61), we obtain that h is continuous.

Step 2: Differentiability of $t \mapsto \theta_3(t,\alpha)$ on $(0,1)$.

Specifically, we prove the result with probability approaching one. We show it by applying the envelope theorem in Milgrom and Segal (2002). To this end, first remark that by Proposition 3 in Horowitz and Manski (1995),

$$\theta_2(\beta,\alpha) = \max_{F_{X_{v0},W}:W\sim \text{Be}(1-\alpha)} E[(X_{v0}'\beta)W],$$

where Be denotes Bernoulli distributions. As a result,

$$\theta_3(t,\alpha) = \max_{F_{X_{v0},W}:W\sim \text{Be}(1-\alpha)} \int x'(t\widehat{\beta}_v + (1-t)\beta_v) w dF_{X_{v0},W}(x,w).$$

By the dominated convergence theorem, the function $t \mapsto f_\alpha(t, F_{X_{v0},W}) := \int x'(t\widehat{\beta}_v + (1-t)\beta_v) w dF_{X_{v0},W}(x,w)$ is differentiable and

$$\frac{\partial f_\alpha}{\partial t}(t, F_{X_{v0},W}) = \left[\int x w dF_{X_{v0},W}(x,w)\right]'(\widehat{\beta}_v - \beta_v).$$

Since $t \mapsto \partial f_\alpha/\partial t(t, F_{X_{v0},W})$ is constant, the family $\{\partial f_\alpha/\partial t(\cdot, F_{X_{v0},W}) : W \sim \text{Be}(1-\alpha)\}$ is equicontinuous and thus the family of functions $\{f_\alpha(\cdot, F_{X_{v0},W}) : W \sim \text{Be}(1-\alpha)\}$ is equidifferentiable at any $t \in (0,1)$ (see Milgrom and Segal, 2002, p.587). Moreover, by the Cauchy-Schwarz inequality,

$$\sup_{F_{X_{v0},W}:W\sim \text{Be}(1-\alpha)} \left|\frac{\partial f_\alpha}{\partial t}(t, F_{X_{v0},W})\right| \leq \left(E[\|X_{v0}\|^2](1-\alpha)\right)^{1/2} \|\widehat{\beta}_v - \beta_v\|.$$

Because $\widehat{\beta}_v$ is consistent, with probability approaching one, $\widehat{\beta}_v \in K$ and since K is convex, $\{t\widehat{\beta}_v + (1-t)\beta_v\} \subset K$. Then, the first step above implies that

$$t \mapsto \left[\int x \mathbb{1}\left\{F_{X_{v0}'(t\widehat{\beta}_v+(1-t)\beta_v)}[x'(t\widehat{\beta}_v + (1-t)\beta_v)] \geq \alpha\right\} dF_{X_{v0}}(x)\right]'\left(\widehat{\beta}_v - \beta_v\right).$$

is continuous on $[0,1]$. Hence, the conditions in Theorem 3 of Milgrom and Segal (2002) hold. Combined with Theorem 1 therein, this implies that $t \mapsto \theta_3(t,\alpha)$ is differentiable and

$$\frac{\partial \theta_3}{\partial t}(t,\alpha) = \left[\int x \mathbb{1}\left\{F_{X_{v0}'(t\widehat{\beta}_v+(1-t)\beta_v)}[x'(t\widehat{\beta}_v + (1-t)\beta_v)] \geq \alpha\right\} dF_{X_{v0}}(x)\right]'\left(\widehat{\beta}_v - \beta_v\right).$$

Step 3: Convergence to a Gaussian process of $\sqrt{n}\left(\widehat{\theta}(\widehat{\beta}_v,\alpha)-\theta(\beta_v,\alpha)\right)$.

First, note that

$$n^{1/2}\left(\widehat{\beta}_v-\beta_v\right)=V(X_v)^{-1}\left(\frac{1}{n^{1/2}}\sum_{i=1}^{n}X_{vi}\varepsilon_{vi}\right)+o_P(1), \tag{62}$$

where $\varepsilon_{vi}:=Y_{v0i}-X_{v0i}'\beta_v$. Let $\theta(q,\alpha)=(\theta_1(q,\alpha),\theta_2(q,\alpha))$ and define $\widehat{\theta}(q,\alpha)$ accordingly. By (62), the Cramér-Wold device, stability of Donsker classes by addition and the first part of the proof of Theorem 3, the process $\mathbb{G}_n:=\sqrt{n}\left(\widehat{\theta}(.,.)-\theta(.,.),\widehat{\beta}_v-\beta_v\right)$ converges weakly to a Gaussian process on $\mathcal{V}\times[\varepsilon,1-\varepsilon]$. Then, when $\left\|\widehat{\beta}_v\right\|\neq 0$, which occurs with probability approaching one, we have

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}(\widehat{\beta}_v,\alpha)-\theta(\beta_v,\alpha)\right)=&\left\|\widehat{\beta}_v\right\|\sqrt{n}\left(\widehat{\theta}(\widehat{q},\alpha)-\theta(\widehat{q},\alpha)\right)\\&+\sqrt{n}\left(\theta(\widehat{\beta}_v,\alpha)-\theta(\beta_v,\alpha)\right),\end{aligned} \tag{63}$$

where we let $\widehat{q}=\widehat{\beta}_v/\left\|\widehat{\beta}_v\right\|$. First, consider the second term. By the second step and the mean value theorem,

$$\begin{aligned}\theta_2(\widehat{\beta}_v,\alpha)-\theta_2(\beta_v,\alpha)=&\theta_3(1,\alpha)-\theta_3(0,\alpha)\\=&h(\widetilde{\beta},\alpha)'\left(\widehat{\beta}_v-\beta_v\right),\end{aligned}$$

with $\widetilde{\beta}=t\widehat{\beta}_v+(1-t)\beta_v$ for some $t\in[0,1]$. Now, by the first step, h is continuous on the compact set $K\times[\varepsilon,1-\varepsilon]$, which includes $\widetilde{\beta}$ with probability approaching one. Thus, by the maximum theorem and the continuous mapping theorem, $\sup_{\alpha\in[\varepsilon,1-\varepsilon]}|h(\widetilde{\beta},\alpha)-h(\beta_v,\alpha)|\xrightarrow{\mathbb{P}}0$. As a result,

$$\sqrt{n}\left(\theta_2(\widehat{\beta}_v,\alpha)-\theta_2(\beta_v,\alpha)\right)=h(\beta_v,\alpha)'\sqrt{n}\left(\widehat{\beta}_v-\beta_v\right)+\varepsilon_n'(\alpha), \tag{64}$$

where $\sup_{\alpha\in[\varepsilon,1-\varepsilon]}|\varepsilon_n'(\alpha)|\xrightarrow{\mathbb{P}}0$.

Now let us turn to the first term in (63). We show below that

$$\sup_{\alpha\in[\varepsilon,1-\varepsilon]}\int[g_{\widehat{\beta}_v,\alpha}(x)-g_{\beta_v,\alpha}(x)]^2dF_X(x)\xrightarrow{\mathbb{P}}0. \tag{65}$$

Then, $\left\|\widehat{\beta}_v\right\|\xrightarrow{\mathbb{P}}\|\beta_v\|$ and the proof of Theorem 19.26 in Van der Vaart (2000) imply that

$$\left\|\widehat{\beta}_v\right\|\sqrt{n}\left(\widehat{\theta}(\widehat{q},\alpha)-\theta(\widehat{q},\alpha)\right)=\|\beta_v\|\sqrt{n}\left(\widehat{\theta}(q_0,\alpha)-\theta(q_0,\alpha)\right)+\varepsilon_n(\alpha), \tag{66}$$

where $\sup_{\alpha\in[\varepsilon,1-\varepsilon]}|\varepsilon_n(\alpha)| \xrightarrow{\mathbb{P}} 0$. Convergence of $\mathbb{G}_n$ combined with equations (63), (64) and (66) imply that $\sqrt{n}\left(\widehat{\theta}(\widehat{\beta}_v,\alpha)-\theta(\beta_v,\alpha)\right)$ converges in distribution to a Gaussian process $\mathbb{G}$.

To prove (65), given the definition of $g_{\beta,\alpha}$, it suffices to prove

$$\sup_{\alpha\in[\varepsilon,1-\varepsilon]}\left[F^{-1}_{X_v'\widehat{\beta}_v}(\alpha)-F^{-1}_{X_v'\beta_v}(\alpha)\right]^2 \xrightarrow{\mathbb{P}} 0, \tag{67}$$

$$\sup_{\alpha\in[\varepsilon,1-\varepsilon]}\int\left|\mathbb{1}\left\{F_{X_v'\widehat{\beta}_v}(x_v'\widehat{\beta}_v)\le\alpha\right\}-\mathbb{1}\left\{F_{X_v'\beta_v}(x_v'\beta_v)\le\alpha\right\}\right|dF_X(x) \xrightarrow{\mathbb{P}} 0, \tag{68}$$

$$\sup_{\alpha\in[\varepsilon,1-\varepsilon]}\int\left(x_v'\widehat{\beta}_v\mathbb{1}\left\{F_{X_v'\widehat{\beta}_v}(x_v'\widehat{\beta}_v)>\alpha\right\}-x_v'\beta_v\mathbb{1}\left\{F_{X_v'\beta_v}(x_v'\beta_v)>\alpha\right\}\right)^2 dF_X(x) \xrightarrow{\mathbb{P}} 0. \tag{69}$$

We prove that the three terms inside the three suprema are continuous as functions of $(\widehat{\beta}_v,\alpha)$. The results then follow by the maximum and continuous mapping theorems. First remark that since $F_{X_v'\beta}$ is strictly increasing that for all $(\beta,\alpha)\in K\times[\varepsilon,1-\varepsilon]$,

$$F^{-1}_{X_v'\beta}(\alpha)=\text{argmin}_{a\in[-M,M]}E[\rho_\alpha(X_v'\beta-a)],$$

for some $M>0$ large enough and $\rho_\alpha(x)=(\alpha-\mathbb{1}\{x\le 0\})x$. By the dominated convergence theorem, the function $(\beta,\alpha,a)\mapsto E[\rho_\alpha(X_v'\beta-a)]$ is continuous. Hence, by the maximum theorem, $(\beta,\alpha)\mapsto F^{-1}_{X_v'\beta}(\alpha)$ is continuous. Then, let $\lambda(\beta):=\max_{\alpha\in[\varepsilon,1-\varepsilon]}(F^{-1}_{X_v'\beta}(\alpha)-F^{-1}_{X_v'\beta_v}(\alpha))^2$. By what precedes, $(\beta,\alpha)\mapsto(F^{-1}_{X_v'\beta}(\alpha)-F^{-1}_{X_v'\beta_v}(\alpha))^2$ is continuous, which implies (67).

The continuity of $(\beta,\alpha)\mapsto E\left[\left|\mathbb{1}\left\{F_{X_v'\beta}(X_v'\beta)\le\alpha\right\}-\mathbb{1}\left\{F_{X_v'\beta_v}(X_v'\beta_v)\le\alpha\right\}\right|\right]$ follows from the exact same reasoning as the continuity of h. Finally, we prove the continuity of

$$j:(\beta,\alpha)\mapsto E\left[\left(X_v'\beta\mathbb{1}\left\{F_{X_v'\beta}(X_v'\beta)>\alpha\right\}-X_v'\beta_v\mathbb{1}\left\{F_{X_v'\beta_v}(X_v'\beta_v)>\alpha\right\}\right)^2\right]$$

on $K\times[\varepsilon,1-\varepsilon]$. Using $a^2-b^2=(a-b)(a+b)$, the Cauchy-Schwarz inequality and $(\sum_{i=1}^k a_i)^2\le k\sum_{i=1}^k a_i^2$, we obtain

$$\begin{aligned}
&|j(\beta_1,\alpha_1)-j(\beta_2,\alpha_2)|\\
\le&6^{1/2}\left\{E\left[\left(X_v'\beta_1\mathbb{1}\left\{F_{X_v'\beta_1}(X_v'\beta_1)>\alpha_1\right\}-X_v'\beta_2\mathbb{1}\left\{F_{X_v'\beta_2}(X_v'\beta_2)>\alpha_2\right\}\right)^2\right]\right.\\
&\left.+E\left[(X_v'\beta_v)^2\left|\mathbb{1}\left\{F_{X_v'\beta_v}(X_v'\beta_v)>\alpha_1\right\}-\mathbb{1}\left\{F_{X_v'\beta_v}(X_v'\beta_v)>\alpha_2\right\}\right|\right]\right\}^{1/2}\\
&\times\left\{E\left[(X_v'\beta_1)^2+(X_v'\beta_2)^2+2(X_v'\beta_v)^2\right]\right\}^{1/2}.
\end{aligned}$$

Thus, it suffices to bound the first and second terms, corresponding to the first and second lines. Regarding the second, by applying Hölder's inequality and using $E[\|X\|^{2+\delta}] < \infty$, we just need to bound

$$E\left[\left|\mathbb{1}\left\{F_{X_v'\beta_v}(X_v'\beta_v) > \alpha_1\right\} - \mathbb{1}\left\{F_{X_v'\beta_v}(X_v'\beta_v) > \alpha_2\right\}\right|\right],$$

which can be done as in Step 1 above. Regarding the first term, we also reason as in Step 1, with the sole difference that because of the square, we use again Hölder's inequality and $E[\|X\|^{2+\delta}] < \infty$.

Step 4: Conclusion.

Because $(F_1, F_2) \mapsto F_1/F_2$ is Hadamard differentiable for all (F_1, F_2) such that F_2 does not vanish, the functional delta method implies that the process

$$\mathbb{H}_n(\alpha) := n^{1/2}\left(R(\alpha, \widehat{F}_{Y_{v0}}, \widehat{F}_{X_{v0}'\widehat{\beta}_v}) - R(\alpha, F_{Y_{v0}}, F_{X_{v0}'\beta_v})\right)$$

defined on $[\varepsilon, 1-\varepsilon]$, also converges to a Gaussian process $\mathbb{H}$. By the directional Hadamard differentiability of ι, we obtain

$$n^{1/2}\left(S_\varepsilon(\widehat{F}_{Y_{v0}}, \widehat{F}_{X_{v0}'\widehat{\beta}_v}) - S_\varepsilon(F_{Y_{v0}}, F_{X_{v0}'\beta_v})\right) \xrightarrow{d} L := \iota'_{R(\cdot, F_{Y_{v0}}, F_{X_{v0}'\beta_v})}(\mathbb{H}).$$

Moreover, by the same argument as in the proof of Theorem 3, the distribution of L is continuous. Combined with Theorem 2.2.1 in Politis et al. (1999), this implies that $q_{1-\alpha}(T^*) \xrightarrow{\mathbb{P}} c_{1-\alpha}$, the quantile of order $1-\alpha$ of L. Finally, under the null hypothesis, because $S_\varepsilon(F_{Y_{v0}}, F_{X_{v0}'\beta_v}) = S(F_{Y_{v0}}, F_{X_{v0}'\beta_v}) = 1$, we have

$$T = n^{1/2}\left(S_\varepsilon(\widehat{F}_{Y_{v0}}, \widehat{F}_{X_{v0}'\widehat{\beta}_v}) - S_\varepsilon(F_{Y_{v0}}, F_{X_{v0}'\beta_v})\right).$$

As a result, $P(T > q_{1-\alpha}(T^*)) \to P(L > c_{1-\alpha}) = \alpha$. The second result also follows since $T \to \infty$ under the alternative.

5. Proof of Proposition 8

Remark that for any random variables A, B and C such that $A \succ_{\text{cv}} B$, $A \perp\!\!\!\perp C$ and $B \perp\!\!\!\perp C$, we have $A + C \succ_{\text{cv}} B + C$. Fix $\beta \in \mathcal{B}^*$. By assumption, $\xi_{Y_0} \succ_{\text{cv}} \xi_{X_0}'\beta$. Thus,

$$X_0^{*\prime}\beta + \xi_{Y_0} \succ_{\text{cv}} X_0^{*\prime}\beta + \xi_{X_0}'\beta = X_0'\beta. \tag{70}$$

Now, because $\beta \in \mathcal{B}^*$, we also have, by Theorem 1, $Y_0^* \succ_{\text{cv}} X_0^{*\prime}\beta$. Hence, by independence, $Y_0^* + \xi_{Y_0} \succ_{\text{cv}} X_0^{*\prime}\beta + \xi_{Y_0}$. Combined with (70), this yields $Y_0 \succ_{\text{cv}} X_0'\beta$. Hence, $\beta \in \mathcal{B}$ and $\mathcal{B}^* \subset \mathcal{B}$.