Partially Linear Models under Data Combination*

Xavier D'Haultfoeuille[†] Christophe Gaillac[‡] Arnaud Maurel[§]

April 11, 2022

Abstract

We consider the identification of and inference on a partially linear model, 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. Finally, we apply our methodology to study intergenerational income mobility over the period 1850-1930 in the United States. Our method allows 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 Federico Bugni, Nathael Gozlan, Jinyong Hahn, 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.

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. This type of data combination environment, sometimes also referred to as data fusion, arises very frequently in a number of subfields of empirical microeconomics. This includes, among others, health (Bhattacharya, 2013; Davillas and Pudney, 2020), income and consumption (Buchinsky et al., 2022), education and returns to skills (Piatek and Pinger, 2016), as well as early childhood development (Garcia et al., 2020). A common example is one where the researcher seeks to combine experimental data with another observational dataset (Athey et al., 2020), although data combination issues are also pervasive when working with observational data only (see Ridder and Moffitt, 2007 for a survey).

We consider the following partially linear model:

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

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 variables X_c are thus common to both datasets, whereas the variables X_{nc} are 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 $k \in \{1, ..., p\}$.

We first provide a tractable characterization of the sharp identified set. Using in particular 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},

[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 $S(q) = \max_{\lambda q \in \mathcal{C}} \lambda$.

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 $Y - f(X_c) - X_{nc}'\beta_0$. 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), for instance, relies on the assumption $f(X_c) = X_{c,i}'\gamma_0$ for some γ_0 and $X_c = (X_{c,e}', X_{c,i}')'$.[2] Point identification then follows. But the exclusion restriction that $E(Y|X)$ does not depend on $X_{c,e}$ may not be credible. We show instead that shape restrictions on f may be sufficient to, e.g. identify the sign of β_{0k}. Finally, we prove that our results are robust to measurement errors on Y and X: under some conditions on these errors, β_0 belongs to the identified set we obtain using the error-ridden covariates and outcome instead of the true variables.

Our identification result is constructive, and readily leads to a simple, plug-in estimator of the identified sets for β_0 or β_{0k}. The estimator of the radial function, however, is not asymptotically normal in general. To construct asymptotically valid confidence regions on β_0 or confidence intervals on β_{0k}, we show that we can use the numerical bootstrap (Hong and Li, 2018, 2020).

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 be used instead. We show through simulations the advantages of relying on the targeted method we propose. First, it often leads to a power gain, that is to say shorter confidence intervals. Second, our method is fast. With a univariate X_{nc}, confidence regions are typically computed in seconds, whereas they take less than seven minutes with a bivariate X_{nc} in the hardest case we consider. Compared to the method of Andrews and Shi (2017), this corresponds to a reduction by a factor of more than 200 of the computational time.

Finally, we apply our method to study intergenerational income mobility over the

[2] In this context, $X_{c,e}$ (resp. $X_{c,i}$) corresponds to the excluded (resp. included) instruments.

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 two-sample 2SLS approach used in Olivetti and Paserman (2015).

Related literature

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} that are not observed jointly with the outcome 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 consider a model that is linear in the covariates X_{nc}. This linearity assumption plays an important role in our ability to derive a tractable inference method.

Our paper is also related to Pacini (2019), which constructs 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.

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.

From a technical point of view and within the data combination literature, our paper is 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.

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.

Finally, our method can 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, one can see our paper as expanding the range of data environments in which unconfoundedness is a credible assumption, complementing a small set of papers that focus on evaluating its reasonableness in the absence of data combination (see, e.g., Altonji et al., 2005).

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, while Section 4 illustrates the finite sample performances of our inference method through Monte Carlo simulations. In Section 5 we apply our method to revisit the analysis of Olivetti and Paserman (2015) about intergenerational income mobility over the period 1850 to 1930 in the United States. Section 6 concludes. The Appendix gathers details on inference with weights and different sample sizes, additional material on the application, and all the proofs. Finally, our inference method can be implemented using our companion R package, `RegCombin`, available at github.com/cgaillac/RegCombin.

2 Identification

Before presenting our main identification results, we introduce some notation that will be used throughout the paper. We let 0_p and $\mathcal{S}_p$ denote respectively 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. For any random variable A, we let $A_0 = A - E(A)$, $\text{Supp}(A)$ be its support and F_A denote its cdf. We also let $\succ_{\text{cv}}$ denote the convex ordering, namely, for two random variables A and B, $A \succ_{\text{cv}} B$ if $E[\phi(A_0)] \geq E[\phi(B_0)]$ for all convex functions ϕ. 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 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 $E(X_0 X_0')$ is non-singular. Moreover, $E(Y_0|X_0) = X_0'\beta_0$ for some $\beta_0 \in \mathbb{R}^p$.*

The identified set of β is defined as the set of all vectors in $\mathbb{R}^p$ that are compatible with the model and the marginal distributions of Y and X, namely

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

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:

$$R(\alpha, F, G) = \frac{\int_\alpha^1 F^{-1}(t)dt}{\int_\alpha^1 G^{-1}(t)dt}, \tag{2}$$
$$S(F,G) = \inf_{\alpha \in (0,1)} R(\alpha, F, G).$$

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 $R(\alpha, F, G)$ 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*

$$\mathcal{B} = \left\{\lambda q : \ q \in \mathcal{S}, \ 0 \leq \lambda \leq S(F_{Y_0}, F_{X_0'q})\right\}. \tag{3}$$

$\mathcal{B}$ 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 first give a sketch of the proof of (3). Let $\mathcal{B}'$ denote the set on the right-hand side of (3). First, by definition of $S(F_{Y_0}, F_{X_0'q})$,

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

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

$$\mathcal{B}' = \{\beta \in \mathbb{R}^p : Y \succ_{\text{cv}} X'\beta\}.$$

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:

$$\mathcal{B}' = \{\beta \in \mathbb{R}^p : \ E\left[\max(0, Y_0 - t)\right] \geq E\left[\max(0, X_0'\beta - t)\right] \ \forall t \in \mathbb{R}\}. \tag{4}$$

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 will return to this alternative characterization of the identified set in Section 4, where we will document the computational advantages of using our characterization instead.

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

$$\inf_{(\widetilde{Y},\widetilde{X}^{\beta}):\widetilde{Y}\overset{d}{=}Y,\,\widetilde{X}^{\beta}\overset{d}{=}X'\beta} E\left[\left|\widetilde{X}_0^{\beta} - E[\widetilde{Y}_0|\widetilde{X}_0^{\beta}]\right|\right] = 0. \tag{5}$$

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, (5) 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

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

Together, (5), (6), and the existence of a minimizer on the left-hand side of (6) (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 $Y \succ_{\text{cv}} X'\beta$ 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.

A point worth mentioning with the expression (3) of the identified set is that we characterize the convex set $\mathcal{B}$ using S, which corresponds to the inverse of the Minkowski gauge function of $\mathcal{B}$ (see, e.g., 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.

[3]We thank Nathael Gozlan for his help in obtaining (6).

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. A key practical advantage of using the radial function S instead of the support function σ in our context lies in its computational simplicity, as this only requires minimizing a simple function over the interval $(0, 1)$. In contrast, the support function approach requires evaluating a supremum over the whole set $\mathcal{B}$, which can be computationally demanding.

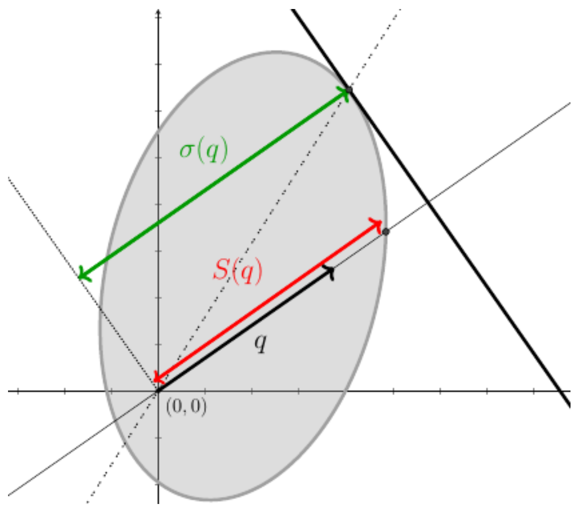


Figure 1: Two characterizations of a closed convex set including the origin, either through its support function σ (green), or through the radial function S (red).

On the other hand, the support function 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,*

$$\sigma(e_k, F_{Y_0}, F_{X_0}) = \frac{1}{\inf_{q \in \mathbb{R}^p : q_k = 1} 1/S(F_{Y_0}, F_{X_0'q})}. \tag{7}$$

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

Expression (7) is appealing because $q \mapsto 1/S(F_{Y_0}, F_{X_0'q})$ is convex, as shown in the proof of Proposition 1 below. Thus, we can recover the support function σ, and in turn the sharp bounds on β_k, by simply minimizing a convex function over $\mathbb{R}^{p-1}$.

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. Then, 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. Specifically, we define, for any $\varepsilon \in (0, 1/2)$,

$$S_\varepsilon(F, G) = \min_{\alpha \in [\varepsilon, 1-\varepsilon]} R(\alpha, F, G), \tag{8}$$

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

Note that for all F, G, $\alpha \mapsto R(\alpha, F, G)$ is continuous on $[\varepsilon, 1-\varepsilon]$. Thus, the minimum in (8) is well-defined. Proposition 1 below describes some properties of $\mathcal{B}_\varepsilon$ and relates it with the sharp identified set $\mathcal{B}$.

Proposition 1. *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 there exist $0 < \underline{\alpha} \leq \overline{\alpha} < 1$ such that, for all $q \in \mathcal{S}$, $\alpha \mapsto F_{Y_0}^{-1}(\alpha)/F_{X_0'q}^{-1}(\alpha)$ is (weakly) increasing on $[\overline{\alpha}, 1)$ and (weakly) decreasing on $(0, \underline{\alpha}]$. Then, there exists $0 < \varepsilon_0 < 1/2$ such that $\mathcal{B}_{\varepsilon_0} = \mathcal{B}$.*

The first part of Proposition 1 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 under a regularity condition on the quantile ratio function $\alpha \mapsto F_{Y_0}^{-1}(\alpha)/F_{X_0'q}^{-1}(\alpha)$, the set $\mathcal{B}_\varepsilon$ coincides with the sharp set $\mathcal{B}$ for ε small enough. As a leading special case, suppose that Y and X are both normally distributed, with $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$ and $X \sim \mathcal{N}(\mu_X, \Sigma)$, where Σ is nonsingular. Then, it is easy to check that for all $q \in \mathcal{S}$, $\alpha \mapsto F_{Y_0}^{-1}(\alpha)/F_{X_0'q}^{-1}(\alpha)$ is constant and equal to $\sigma/(q'\Sigma q)^{1/2}$. In this particular case, we actually have $\mathcal{B}_\varepsilon = \mathcal{B}$ for all $\varepsilon \in (0, 1/2)$. In general though, $\mathcal{B}_\varepsilon$ will be a strict superset of $\mathcal{B}$ for ε large enough.

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:

$$\begin{cases} X = X^* + \xi_X, & \xi_X \perp\!\!\!\perp X^*, \\ Y = Y^* + \xi_Y, & \xi_Y \perp\!\!\!\perp (X^*, Y^*). \end{cases} \tag{9}$$

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 2. *If Assumption 1 is satisfied with (X, Y) replaced by (X^*, Y^*), (9) holds and for all $\beta \in \mathcal{B}^*$, $\xi_Y \succ_{cv} \xi_X' \beta$, then $\mathcal{B}^* \subset \mathcal{B}$.*

This proposition establishes that the identified set is robust to measurement errors in the following sense: if measurement errors on the outcome Y^* second-order stochastically dominate those on the linear index $X^{*\prime}\beta$ for all $\beta \in \mathcal{B}^*$, the identified set $\mathcal{B}$ based on the observed covariates and outcome X and Y always contains the true value of the parameter of interest. 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 $\xi_Y \succ_{\text{cv}} \xi_X \beta$ 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 2 is likely to hold in this context.

2.1.2 Point identification

We now show that under some additional restrictions, our approach yields point identification of the parameters of interest.

Proposition 3. *Suppose that Assumption 1 holds. Let $U := Y - X'\beta_0$. Then:*

1. *If $\beta_0 \neq 0_p$ and for all $\lambda > 0$, $U \not\succ_{cv} (X'\beta_0)\lambda$, then $\beta_0 \in \partial\mathcal{B}$.*

2. *If $\beta_0 = 0_p$ and for all $\beta \neq 0_p$, $Y \not\succ_{cv} X'\beta$, then $\mathcal{B} = \{0_p\}$.*

A consequence of the first point is that β_0 is point identified if $X \in \mathbb{R}$ and in addition to $E(Y|X) = X\beta_0$, we impose (i) $\beta_0 > 0$ (say) and (ii) $U \not\succ_{\text{cv}} X\lambda$, for all $\lambda > 0$. Then, β_0 is equal to the upper bound of $\mathcal{B}$. Condition (ii) basically imposes that the tails of $X\lambda$ are fatter than those of U, as we shall see more precisely below. Thus, the relative thickness of the tails of the distributions of the observable index, $X\lambda$, and of the distribution of the unobservables U, plays an important role in attaining point identification of β_0. Note that, for the case where the unobservables U are normally distributed, this condition is satisfied under the relatively mild restriction that the linear index $X\beta_0$ has thicker tails than a normal distribution.

The second point of Proposition 3 further establishes point identification when the true coefficient satisfies $\beta_0 = 0_p$ and for all $\beta \neq 0_p$ $Y \not\succ_{\text{cv}} X'\beta$, so that Y has lighter tails than any linear index of X. In contrast to the first part of the proposition, point identification holds for such DGP without using any restrictions on β_0, since in this case the identified set is reduced to a singleton ($\mathcal{B} = \{0_p\}$).

The following lemma allows us to derive sufficient conditions for the key restrictions we have imposed in Proposition 3, namely $U \not\succ_{\text{cv}} (X'\beta_0)\lambda$ for all $\lambda > 0$ or $Y \not\succ_{\text{cv}} X'\beta$ for all $\beta \neq 0_p$.

Lemma 1. *Let $S \in \mathbb{R}$ and $T \in \mathbb{R}^p$ be two random variables and suppose that there exists ϕ_1 and ϕ_2 functions from $\mathbb{R}^+$ to $\mathbb{R}^+$ such that (i) ϕ_1 is increasing and convex; (ii) $\lim_{x\to\infty} \phi_2(x)/x = \infty$; (iii) $E[\phi_1 \circ \phi_2(|S_0|)] < \infty$ and $E[\phi_1(|T_0'\gamma|)] = \infty$ for all $\gamma \in \mathcal{S}$. Then, for all $\beta \neq 0_p$, $S \not\succ_{cv} T'\beta$.*

This lemma shows for instance (taking $p = 1$) that $S \not\succ_{\text{cv}} T\lambda$ for all $\lambda \neq 0$ if for some $a > b > 0$, $E[|S|^a] < E[|T|^b] = \infty$. Alternatively, the condition holds if $E[\exp(a|S|^b)] < E[\exp(c|T|)] = \infty$ for some $b > 1$ and $a, c > 0$. The lemma also applies for any $p \in \mathbb{N}$ if $E[\exp(|S|^{2+\eta})] < \infty$ for some $\eta > 0$ and T is a Gaussian vector with nonsingular variance matrix.

To illustrate Point 2 of Proposition 3 and Lemma 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 Lemma 1, $Y \not\succ_{\text{cv}} X\lambda$ for all $\lambda \neq 0$. It then follows from Point 2 of Proposition 3 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 $\mathcal{B}^V = [-1/\sqrt{2}, 1/\sqrt{2}] \simeq [-0.707,\, 0.707]$.

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:

$$\begin{aligned} X^x &= X_{nc} - E(X_{nc}|X_c = x), \\ Y^x &= Y - E(Y|X_c = x). \end{aligned}$$

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 4. *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:*

$$\begin{aligned} \mathcal{B}^c &= \left\{\lambda q : q \in \mathcal{S},\ 0 \le \lambda \le \overline{S}(F_{Y,X_c}, F_{X_{nc}'q,X_c})\right\}, \\ \mathcal{F} &= \left\{x \mapsto E(Y|X_c = x) - E(X_{nc}|X_c = x)'\beta :\ \beta \in \mathcal{B}^c\right\}, \end{aligned}$$

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.

The set $\mathcal{B}^c$ is simply the intersection over all $x_c \in \text{Supp}(X_c)$ of the identified sets obtained on the subpopulations for whom $X_c = x_c$. Clearly, this identified set can be reduced if one imposes some constraints on $f(\cdot)$. We consider below a few important special cases.

Consider first the restriction $f(X_c) = f_1(X_{i,c})$, with $X_c = (X'_{i,c}, X'_{e,c})'$. This restriction is implicit in, and central to the two-sample two-stage least squares strategy. Under this restriction, the model is rewritten as:

$$E\left[Y - E(Y|X_{i,c})|X_c\right] = R'\beta_0,$$

with $R = E\left[X_{nc} - E(X_{nc}|X_{i,c})|X_c\right]$. Hence, under the relevance condition that $E(RR')$ is nonsingular, β_0 is point identified.

Another restriction one may consider is linearity of $f(\cdot)$, namely $f(X_c) = X'_c\gamma_0$. It follows from this restriction that:

$$E(Y|X_c) = X'_c\gamma_0 + E(X_{nc}|X_c)'\beta_0.$$

If X_c and $E(X_{nc}|X_c)$ are not collinear, which implies that $E(X_{nc}|X_c)$ is a nonlinear function of X_c, β_0 is again point identified. 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.

Third and most importantly, shape constraints on $f(\cdot)$ arise naturally in many empirical applications. For concreteness, we consider the case of monotonicity constraints, where we assume that X_c is finitely supported. Specifically, assume that $\text{Supp}(X_c) = \{x_{c,1}, ..., x_{c,K}\}$, $f(X_c) = \sum_{k=1}^{K} \gamma_{0,k}\mathbb{1}\{X_c = x_{c,k}\}$ and let us consider (weak) sign constraints on some of the $\gamma_{0,k}$. We let $s_k = 1$ if the constraint is $\gamma_{0,k} \geq 0$, $s_k = -1$ if the constraint is $\gamma_{0,k} \leq 0$. Finally, let $s_k = 0$ if we do not impose any constraint on $\gamma_{0,k}$. We denote by $\mathcal{B}^{\text{con}}$ the identified set under these constraints.

We first express $\mathcal{B}^{\text{con}}$ using the unconstrained set $\mathcal{B}^c$. Denoting by $\overline{Y}_k = E\left(Y|X_c = x_{c,k}\right)$, $\overline{X}_k = E(X_{nc}|X_c = x_{c,k})$, we have

$$s_k\left(\overline{Y}_k - \overline{X}'_k\beta_0\right) \geq 0 \quad \forall k \in \{1, ..., K\}. \tag{10}$$

Because these are the only constraints on $f(\cdot)$, $\mathcal{B}^{\text{con}}$ satisfies

$$\mathcal{B}^{\text{con}} = \{b \in \mathcal{B}^c : (10) \text{ holds.}\} \tag{11}$$

Since the set of parameter values b satisfying the constraints (10) is convex and closed, $\mathcal{B}^{\text{con}}$ is still convex and compact. On the other hand, it does not include the origin if $s_k\overline{Y}_k < 0$ for some k.

It is possible to express $\mathcal{B}^{\text{con}}$ in a way that is closer to the characterizations of $\mathcal{B}$ and $\mathcal{B}^c$ in Theorem 1 and Proposition 4, respectively. Let $q \in \mathcal{S}^+$ denote a direction in the upper hemisphere. From (11) and Proposition 4, any $\lambda q \in \mathcal{B}^{\text{con}}$ should satisfy

$$\begin{aligned} -\overline{S}(F_{Y,X_c}, F_{-X_{nc}'q,X_c}) \leq \lambda \leq \overline{S}(F_{Y,X_c}, F_{X_{nc}'q,X_c}), \\ \lambda\left(s_k \overline{X}_k' q\right) \leq s_k \overline{Y}_k \quad \forall k \in \{1, ..., K\}. \end{aligned}$$

This shows that λ satisfies additional inequalities. To write these, we introduce the sets $\mathcal{K}^+(q) = \left\{k \in \{1, ..., K\} : \ s_k \overline{X}_k' q > 0\right\}$ and $\mathcal{K}^-(q) = \left\{k \in \{1, ..., K\} : \ s_k \overline{X}_k' q < 0\right\}$. Then, define

$$\underline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) = \max\left(-\overline{S}(F_{Y,X_c}, F_{-X_{nc}'q,X_c}), \max_{k \in \mathcal{K}^-(q)} \frac{\overline{Y}_k}{\overline{X}_k' q}\right) \tag{12}$$

$$\overline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) = \min\left(\overline{S}(F_{Y,X_c}, F_{X_{nc}'q,X_c}), \min_{k \in \mathcal{K}^+(q)} \frac{\overline{Y}_k}{\overline{X}_k' q}\right). \tag{13}$$

By what precedes, we can rewrite the identified set as follows:

$$\mathcal{B}^{\text{con}} = \left\{\lambda q : q \in \mathcal{S}^+, \ \underline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) \leq \lambda \leq \overline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c})\right\}. \tag{14}$$

The identified set for $f(\cdot)$ is then

$$\mathcal{F}^{\text{con}} = \left\{\sum_{k=1}^{K} \gamma_k \mathbb{1}\left\{X_c = x_{c,k}\right\} : \exists \beta \in \mathcal{B}^{\text{con}} : \ \gamma_k = \overline{Y}_k - \overline{X}_k' \beta\right\}.$$

Note that we may have $\underline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c}) > \overline{S}^{\text{con}}(q, F_{Y,X_c}, F_{X_{nc},X_c})$, in which case the intersection of $\mathcal{B}^{\text{con}}$ with the linear span of q is empty. If $\mathcal{B}^{\text{con}}$ itself is empty, one would reject the model (1) together with the constraints (10).

2.3 Numerical illustration

We conclude this section with a numerical illustration of the identification results above. We consider the following model:

$$Y = \gamma_{0,0} + \gamma_{0,1}\mathbb{1}\{X_c = 1\} + X_{nc,1}\beta_{nc,1} + X_{nc,2}\beta_{nc,2} + U, \ U|X \sim \mathcal{N}(0, 4).$$

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

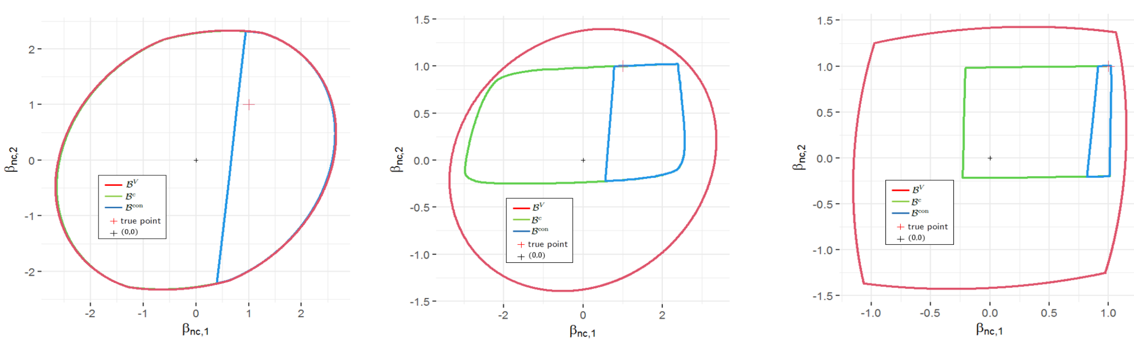
$$\Sigma = \begin{pmatrix} 1 & 0.8 & -0.1 \\ 0.8 & 1 & -0.2 \\ -0.1 & -0.2 & 1 \end{pmatrix}.$$

Specifically, the common regressor is given by $X_c = \mathbb{1}\{N_1 \leq 1.2\}$, and we consider three cases for the regressors that are observed in one of the datasets only, X_{nc}. In the first case, $(X_{nc,1}, X_{nc,2}) = (N_2, N_3)$, in the second, $(X_{nc,1}, X_{nc,2}) = (N_2, \exp(N_3))$ and in the third, $(X_{nc,1}, X_{nc,2}) = (\exp(N_2), \exp(N_3))$.

Figure 2 displays several identified sets for each of the three data-generating processes 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 4 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 the additional sign constraint $\gamma_{0,1} \geq 0$.



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

Note: the sets are obtained using a sample of 100,000 observations and taking the convex hull of the set obtained from 6,000 directions sampled uniformly on the 2 dimensional sphere.

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 sets $\mathcal{B}^V$ and $\mathcal{B}^c$ are equal. This is expected because conditional on X_c, the distributions of Y and X_{nc} are close to normal distributions, in which case the variance restrictions are sufficient to rationalize the linear conditional expectation restriction. The sign restriction on X_c nearly cuts by half the area of the identified set and excludes 0_2 from $\mathcal{B}^{\text{con}}$. In case (b) on the other hand, 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 3 applied conditional on $X_c = 0$ and $X_c = 1$. In this case also, the restriction $\gamma_{0,1} \geq 0$ is sufficient to imply that $0_2 \notin \mathcal{B}^{\text{con}}$. Here, when projecting on the x-axis, we can also reject that $\beta_{nc,1} = 0$.

Finally, the identified set $\mathcal{B}^c$ is reduced further in case (c), as a result of the fatter tails of both $X_{nc,1}$ and $X_{nc,2}$. Like in case (b), under the sign constraint on $\gamma_{0,1}$, we reject not only that $0_2 \in \mathcal{B}^{\text{con}}$, but also that $\beta_{nc,1} = 0$. Overall, that the sharp identified sets $\mathcal{B}^c$ are, for the cases (b) and (c), 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.

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 and then show how to incorporate such regressors. We conclude this section by discussing some computational and implementation details of our procedure.

3.1 Inference without common regressors

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

Assumption 2. *We observe $(Y_1, ..., Y_n)$ and $(X_1, ..., X_n)$, two independent samples of i.i.d. variables with the same distribution as Y and X, respectively.*

To simplify the exposition, we assume in the following that the two samples have equal size; the general case is described in Appendix A.

3.1.1 Estimation of the identification region and confidence region

For any $q \in \mathcal{S}$, we let $(X'q)_{(1)} < ... < (X'q)_{(n)}$ and $Y_{(1)} < ... < Y_{(n)}$ denote the order statistics associated with the $X'q$ and Y, respectively. For any $\alpha \in (1/n, 1-1/n)$,

$$\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q}) = \frac{\sum_{i=\lceil n\alpha \rceil}^{n} [Y_{(i)} - \overline{Y}]}{\sum_{i=\lceil n\alpha \rceil}^{n} [(X'q)_{(i)} - \overline{X}'q]},$$

where $\lceil x \rceil = \min\{j \in \mathbb{N} : j \geq x\}$. $\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q})$ is a modified plug-in estimator of $R(\alpha, F_{Y_0}, F_{X_0'q})$, in the following sense. 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 $\widehat{F}_{X_0'q}(t) = \widehat{F}_{X'q}(t + \overline{X}'q)$. Then, for any $\alpha \in (1/n, 1-1/n)$, the plug-in estimator of $R(\alpha, F_{Y_0}, F_{X_0'q})$ satisfies, in view of (2),

$$R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) = \frac{\sum_{i=\lceil n\alpha \rceil}^{n} [Y_{(i)} - \overline{Y}] + (\lceil n\alpha \rceil - n\alpha)(Y_{\lceil n\alpha \rceil - 1} - \overline{Y})}{\sum_{i=\lceil n\alpha \rceil}^{n} [(X'q)_{(i)} - \overline{X}'q] + (\lceil n\alpha \rceil - n\alpha)(X'_{\lceil n\alpha \rceil - 1}q - \overline{X}'q)}.$$

Compared to this plug-in estimator, we thus neglect in $\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q})$ the second terms appearing in the numerator and denominator of $R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q})$.

Then, for any $\varepsilon \in (1/n, 1/2(1-1/n))$ we estimate the regularized radial function, $S_\varepsilon(F_{Y_0}, F_{X_0'q})$, by

$$\begin{aligned} \widehat{S}_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right) &= \min_{\alpha \in [\varepsilon, 1-\varepsilon]} \widehat{R}(\alpha, F_{Y_0}, F_{X_0'q}) \\ &= \min_{\lceil \varepsilon n \rceil \leq j \leq n - \lfloor \varepsilon n \rfloor} \frac{\sum_{i=j}^{n} [Y_{(i)} - \overline{Y}]}{\sum_{i=j}^{n} [(X'q)_{(i)} - \overline{X}'q]}, \end{aligned} \tag{15}$$

with $\lfloor x \rfloor = \max\{j \in \mathbb{N} : j \leq x\}$. Equation (15) makes it clear that the computation of $\widehat{S}_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right)$ is almost immediate.

With $\widehat{S}_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right)$ at hand, we estimate the identified set $\mathcal{B}_\varepsilon$ simply by:

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

Next, we build confidence regions on β_0. As the asymptotic distribution of $\widehat{S}_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right)$ is not Gaussian in general, we rely on a particular bootstrap scheme called the numerical bootstrap (see, e.g., Hong and Li, 2018, 2020). This method is well-suited here because the map S_ε is Hadamard directionally differentiable, but not Hadamard differentiable unless strong conditions hold.[4] The main idea of this method is to approximate the local behavior of S_ε by taking finite differences. Specifically, let $(W_1, ..., W_n)$ (resp. $(W_{q,1}, ..., W_{q,n})$) be standard (namely, multinomial) bootstrap weights associated with the sample $(Y_{(1)}, ..., Y_{(n)})$ (resp. $((X'q)_{(1)}, ..., (X'q)_{(n)})$). Then, let

$$\widehat{R}^*(\alpha, F_{Y_0}, F_{X_0'q}) = \frac{\sum_{i=\lceil n\alpha\rceil}^n W_i[Y_{(i)} - \overline{Y}]}{\sum_{i=\lceil n\alpha\rceil}^n W_{q,i}[(X'q)_{(i)} - \overline{X}'q]}$$

denote the bootstrap counterpart of $\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q})$ and define

$$\mathbb{Z}_{n,q}^*(\alpha) = \sqrt{n}\left(\widehat{R}^*(\alpha, F_{Y_0}, F_{X_0'q}) - \widehat{R}(\alpha, F_{Y_0}, F_{X_0'q})\right).$$

We let $\widehat{c}_{\beta,\varepsilon}(q)$ denote the quantile of order $\beta \in (0,1)$ of

$$\frac{\min_{\alpha\in[\varepsilon,1-\varepsilon]}(\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q}) + \delta_n \mathbb{Z}_{n,q}^*) - \widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q})}{\delta_n}, \tag{16}$$

for some δ_n such that $\delta_n \to 0$ and $\sqrt{n}\delta_n \to \infty$. For a nominal coverage of $1-\alpha$, the confidence region on β_0 we consider is given by

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

The idea behind the use of $\widehat{c}_{\alpha,\varepsilon}(q)$ is that the distribution of (16) consistently estimates the asymptotic distribution of

$$\sqrt{n}\left(\widehat{S}_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right) - S_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right)\right). \tag{17}$$

An alternative for consistently estimating the asymptotic distribution of (17) would be subsampling, see Politis et al. (1999).

In practice, one is often interested in conducting inference on subcomponents of β_0. In view of (7), the identified (outer) set $\mathcal{B}_{k,\varepsilon}$ of $\beta_{0,k}$ corresponding to $\mathcal{B}_\varepsilon$ satisfies

$$\mathcal{B}_{k,\varepsilon} = [-\sigma_\varepsilon(-e_k, F_{Y_0}, F_{X_0}), \ \sigma_\varepsilon(e_k, F_{Y_0}, F_{X_0})], \tag{18}$$

[4]We refer to Fang and Santos (2019) for the difference between these two notions of differentiability, the proof that the standard bootstrap generally fails for Hadamard directionally differentiable functions, and the validity of the numerical bootstrap in this context (see in particular their discussion pp. 390-91).

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 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

$$\widehat{\sigma}_\varepsilon(e, F_{Y_0}, F_{X_0}) = \frac{1}{\inf_{q\in\mathbb{R}^p:q'e=1} 1/\widehat{S}_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right)}, \tag{19}$$

see Corollary 1. Then, we estimate the asymptotic distribution of $\widehat{\sigma}_\varepsilon(\cdot, F_{Y_0}, F_{X_0})$, in a similar way as above. Let $\widetilde{c}_{\beta,\varepsilon}(e)$ denote the quantile of order $\beta \in (0,1)$ of

$$\frac{1}{\delta_n}\left[\frac{1}{\inf_{q\in\mathbb{R}^p:q'e=1} 1/\min_{\alpha\in[\varepsilon,1-\varepsilon]}(\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q}) + \delta_n \mathbb{Z}^*_{n,q})} - \widehat{\sigma}_\varepsilon(e, F_{Y_0}, F_{X_0})\right], \tag{20}$$

where δ_n is as above. Then, the confidence interval we consider for $\beta_{0,k}$ is

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

where $x^- = \min(0,x)$ and $x^+ = \max(0,x)$. The rationale for using $(\cdot)^-$ and $(\cdot)^+$ is to ensure that $0 \in \text{CI}_{1-\alpha}(\beta_{0,k})$; 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 shown formally in Theorem 3 below.

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,*

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

Next, we establish the asymptotic validity of $\text{CR}_{1-\alpha}(\beta_0)$ and $\text{CI}_{1-\alpha}(\beta_{0,k})$, under the following regularity conditions.

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*

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

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. 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 the last part of Assumption 4 (on X, say) is that, for all $q \in \mathcal{S}$, $X'q$ admits a density $f_{X'q}$ with respect to the Lebesgue measure and $\inf_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} f_{X'q}(\alpha) > 0$.

Theorem 3. *Fix $(\varepsilon, \alpha) \in (0, 1/2)^2$ and suppose that Assumptions 1-2 hold. Then:*

1. *If Assumption 3 also holds,*

$$\inf_{\beta\in\mathcal{B}} \liminf_{n\to\infty} P\left(\beta \in CR_{1-\alpha}(\beta_0)\right) \geq \inf_{\beta\in\mathcal{B}_\varepsilon} \liminf_{n\to\infty} P\left(\beta \in CR_{1-\alpha}(\beta_0)\right) = 1 - \alpha. \quad (21)$$

2. *If Assumption 4 also holds and the asymptotic distributions of $\sqrt{n}\left(\widehat{\sigma}_\varepsilon(e, F_{Y_0}, F_{X_0})\right.$ $\left.-\sigma_\varepsilon(e, F_{Y_0}, F_{X_0})\right)$ for $e = \pm e_k$ are continuous at their α-th quantile,*

$$\liminf_{n\to\infty} \inf_{\beta_k\in\mathcal{B}_k} P\left(\beta_k \in CI_{1-\alpha}(\beta_{0,k})\right) \geq \liminf_{n\to\infty} \inf_{\beta_k\in\mathcal{B}_{k,\varepsilon}} P\left(\beta_k \in CI_{1-\alpha}(\beta_{0,k})\right) = 1-\alpha. \quad (22)$$

Note that we get equalities in (21)-(22) when $\mathcal{B}_\varepsilon = \mathcal{B}$. This occurs in particular under the conditions displayed in Point 3 of Proposition 1, a leading example of which being when both Y and X are normally distributed.

To prove the result, we first show the weak convergence of

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

seen as a process indexed by either α or (α, q). Then, to prove the consistency of the numerical bootstrap and in turn (21)-(22), we check the conditions of Theorem 3.2 in

Fang and Santos (2019). To this end, we use in particular the Hadamard directional differentiability of the minimum and maximin maps, shown respectively by Cárcamo et al. (2020) and Firpo et al. (2021).

Finally, note that to obtain (22), we impose the continuity of the asymptotic distribution ($F_{\infty,e}$, say, for $e = \pm e_k$) of $\sqrt{n}\,(\widehat{\sigma}_\varepsilon(e, F_{Y_0}, F_{X_0}) - \sigma_\varepsilon(e, F_{Y_0}, F_{X_0}))$. This regularity condition can be shown to hold if $R(\cdot, F_{Y_0}, F_{X_0'q})$ admits a unique minimizer for all $q \in \mathcal{S}$, and we conjecture that it holds more generally.[5]

3.2 Inference with common regressors

We now turn to inference on β_0 with common regressors X_c. Recall from Proposition 4 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. We follow the same logic as above. For any $x \in \text{Supp}(X_c)$, define $\widehat{S}_\varepsilon(F_{Y^x|X_c=x}, F_{X^{x\prime}q|X_c=x})$ as in (15), but restricted to the subsamples of (Y, X_c) and (X_{nc}, X_c) for which $X_c = x$ (in case these subsamples are of different size, see Appendix A). Then, let

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

Next, denote by $\widehat{c}^c_{\beta,\varepsilon}(q)$ the quantile of order $\beta \in (0,1)$ of

$$\frac{1}{\delta_n}\left\{\min_{x \in \text{Supp}(X_c)}\left[\min_{\alpha \in [\varepsilon, 1-\varepsilon]}\left(\widehat{R}(\alpha, F_{Y^x|X_c=x}, F_{X^{x\prime}q|X_c=x}) + \delta_n \mathbb{Z}^*_{n,q,x}\right)\right] - \widehat{\overline{S}}(q, F_{Y,X_c}, F_{X_{nc},X_c})\right\}, \tag{23}$$

with $\mathbb{Z}^*_{n,q,x} = \sqrt{n}\left(\widehat{R}^*(\alpha, F_{Y^x|X_c=x}, F_{X^{x\prime}q|X_c=x}) - \widehat{R}(\alpha, F_{Y^x|X_c=x}, F_{X^{x\prime}q|X_c=x})\right)$. For a nominal coverage of $1-\alpha$, the confidence region on β_0 we consider is

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

[5]The methods developed by Davydov et al. (1998), often used to prove absolute continuity of functional of Gaussian processes (as we do in Part 1 of Theorem 3), do not apply here: $F_{\infty,e}$ is the distribution of a functional of a Gaussian process, but this functional is neither smooth nor convex.

The validity of this confidence region can be obtained as in the case without common regressors. Specifically, using Theorem 3 and $\text{Supp}(X_c) = \{x_1, ..., x_K\}$, we have the weak convergence of

$$\sqrt{n}\begin{pmatrix} \widehat{R}(\alpha, F_{Y^{x_1}|X_c=x_1}, F_{X^{x_1\prime}q|X_c=x_1}) - R(\alpha, F_{Y^{x_1}|X_c=x_1}, F_{X^{x_1\prime}q|X_c=x_1}) \\ \vdots \\ \widehat{R}(\alpha, F_{Y^{x_K}|X_c=x_K}, F_{X^{x_K\prime}q|X_c=x_K}) - R(\alpha, F_{Y^{x_K}|X_c=x_K}, F_{X^{x_K\prime}q|X_c=x_K}) \end{pmatrix},$$

seen as a process indexed by either α or (α, q). Since the map $\overline{S}$ can be viewed as the composition of two maps, $R \mapsto (\inf_{\alpha\in[\varepsilon,1-\varepsilon]} R_1(\alpha), \ldots, \inf_{\alpha\in[\varepsilon,1-\varepsilon]} R_K(\alpha))$ and $\phi : \theta \in \mathbb{R}^K \mapsto \min(\theta_1, \ldots, \theta_K)$, the numerical derivative given by (23) satisfies the assumptions of Theorem 3.2 in Fang and Santos (2019) (see the proof of Theorem 3 and Example 2.2 in Hong and Li, 2018). This ensures the consistency of the numerical bootstrap and, in turn, the asymptotic validity of $\text{CR}^c_{1-\alpha}(\beta_0)$.

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]

3.3 Tuning parameters and computational aspects

Our estimator of the identified set and confidence regions rely on the choice of a regularization parameter ε. We distinguish the cases $p = 1$, where we can adapt the choice to the direction $q \in \mathcal{S}$ while preserving the convexity of $\mathcal{B}_\varepsilon$ and $\widehat{\mathcal{B}}_\varepsilon$, from the case $p > 1$. When $p = 1$, we suggest the following selection rule for a direction $q \in \mathcal{S} = \{-1, 1\}$,

$$\varepsilon(q) = \underset{\varepsilon\in[\varepsilon_0,0.5]}{\text{argmin}}\ \widehat{S}_\varepsilon(F_Y, F_{X'q}) - \widehat{c}_{\alpha,\varepsilon}(q)n^{-1/2}, \tag{24}$$

and we let $\varepsilon_0 = C\ln(n)/n$ with $C = 1$ without common regressors and $C = 3$ with common regressors. Hence, $\varepsilon(q)$ simply minimizes the boundary value of the

[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.

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, instead, we are interested in the set $\mathcal{B}$ itself, we recommend using $\varepsilon = \min_{q\in\mathcal{Q}} \varepsilon(q)$, where $\mathcal{Q}$ is a set of vectors in $\mathcal{S}$, for instance $\pm e_k$ for $k = 1, ..., p$.

Another tuning parameter we have to choose is δ_n, which appears in the numerical bootstrap, see Eq. (16). Its choice seems to matter less in practice than that of ε. Without common regressors, we fix it to $n^{-0.3}$, so that it satisfies $\delta_n \to 0$ and $n^{1/2}\delta_n \to \infty$. With common regressors, we use instead $\delta_n = n^{-0.35}$.

To compute $\sigma_\varepsilon(\pm e_k, F_{Y_0}, F_{X_0})$, we solve (19), in which $q \mapsto 1/\widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q})$ 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 $\text{CR}_{1-\alpha}(\beta_0)$ requires the computation of $\widehat{S}_\varepsilon(F_Y, F_{X'q})$ 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 $\widehat{S}_\varepsilon(F_Y, F_{X'q})$ and $\widehat{c}_{\alpha,\varepsilon}(q)$ for each $q \in \widetilde{\mathcal{S}}$; (iii) construct an approximation of $\widehat{\mathcal{B}}_\varepsilon$ and $\text{CI}_{1-\alpha}$ by computing the convex hulls of $\{\widehat{S}_\varepsilon(F_Y, F_{X'q})q : \; q \in \widetilde{\mathcal{S}}\}$ and $\left\{\left(\widehat{S}_\varepsilon(F_Y, F_{X'q}) - \widehat{c}_{\alpha,\varepsilon}(q)n^{-1/2}\right) q : \; q \in \widetilde{\mathcal{S}}\right\}$, respectively.[7] The resulting sets, $\widetilde{\mathcal{B}}_\varepsilon$ and $\widetilde{\text{CR}}_{1-\alpha}(\beta_0)$ say, are convex, inner approximations of $\widetilde{\mathcal{B}}_\varepsilon$ and $\text{CR}_{1-\alpha}(\beta_0)$, 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 $d_H(\widetilde{\text{CR}}_{1-\alpha}(\beta_0), \text{CR}_{1-\alpha}(\beta_0)) \to 0$.

The computation of the parameters (ε, δ_n), the estimated set, confidence regions and confidence intervals are implemented in our companion R package `RegCombin`.

4 Monte Carlo simulations

In this section we study the finite sample performances of our proposed inference method through Monte Carlo simulations. We first consider the baseline case where

[7]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.

no common regressor is available to the econometrician, both in the univariate and multivariate cases, 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.

4.1 Univariate case without common regressors

We first explore the finite sample performances of our inference method in the baseline case ($p = 1$) considered in Section 3.1. Namely, we consider a sample of i.i.d. observations drawn from the model

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

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 (4) 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.[8]

In Table 1 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 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

[8]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.

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 estimated identified set $\mathcal{B}_{\varepsilon(q)}$, where $\varepsilon(q)$ is given by (24) and thus varies from one simulation to another.

	DGM				**AS**		
Sample size	Bounds	Ex. length	Coverage	$B_{\varepsilon(q)}$	Bounds	Ex. length	Coverage
Normal							
Identified set	[-1.202,1.202]				[-1.202,1.202]		
400	[-1.317,1.316]	0.229	0.946	[-1.202,1.202]	[-1.374,1.367]	0.337	0.983
800	[-1.279,1.28]	0.155	0.954	[-1.202,1.202]	[-1.329,1.328]	0.253	0.985
1,200	[-1.271,1.27]	0.138	0.96	[-1.202,1.202]	[-1.301,1.301]	0.198	0.978
2,400	[-1.249,1.248]	0.093	0.946	[-1.202,1.202]	[-1.268,1.27]	0.134	0.975
4,800	[-1.235,1.235]	0.067	0.966	[-1.202,1.202]	[-1.251,1.25]	0.097	0.98
Gamma							
Identified set	[-0.025,1.046]				[-0.025,1.046]		
400	[-0.441,1.36]	0.729	0.996	[-0.170, 1.282]	[-0.538,1.343]	0.809	1
800	[-0.376,1.309]	0.613	0.984	[-0.148, 1.276]	[-0.466,1.313]	0.707	1
1,200	[-0.353,1.291]	0.571	0.986	[-0.138, 1.266]	[-0.438,1.302]	0.668	1
2,400	[-0.302,1.255]	0.486	0.994	[-0.123, 1.251]	[-0.391,1.277]	0.596	1
4,800	[-0.267,1.23]	0.425	0.994	[-0.110, 1.224]	[-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 "$B_{\varepsilon(q)}$" displays the average, across all simulations, of the estimated identified set $\mathcal{B}_{\varepsilon(q)}$, where $\varepsilon(q)$ is given by (24). We use 1,000 bootstrap replications to compute the confidence intervals for both DGM and AS methods.

Table 1: 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 9.5% (for $n = 400$) and 2.8% (for $n = 4,800$) wider than the identified set. In the gamma case, the confidence set ranges between 50.2% and 31.9% larger than the (regularized) identified set ($\mathcal{B}_\varepsilon$).

Second, the coverage of our confidence intervals is good: coverage rates are always larger than 94.6%. Third, our inference method generally performs better than AS, delivering consistently tighter confidence sets. 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. Gains are smaller but remain quite sizable in the gamma case. 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 (24) 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 in $\widehat{S}_\varepsilon$ and thus larger quantiles $\widehat{c}_{\alpha,\varepsilon}$).

4.2 Multivariate case without common regressor

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

$$Y \quad = \quad \gamma_0 + X_{nc}'\beta_0 + U, \;\; X_{nc} \perp\!\!\!\perp U, \;\; U \sim \mathcal{N}(0, 4). \tag{25}$$

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 2 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, even for small sample sizes, and declines as n gets larger.

Average	Bounds	Excess length	Coverage
Identified set	[-2.367,2.367]		
Sample size			
400	[-2.5951,2.592]	0.454	0.96
800	[-2.5479,2.551]	0.366	0.968
1,200	[-2.5256,2.525]	0.318	0.976
2,400	[-2.493,2.493]	0.254	0.986
4,800	[-2.470,2.47]	0.207	0.998

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 $\beta_1 \in \text{CI}_{1-\alpha}(\beta_{0,1})$. We use 200 bootstrap replications to compute the confidence intervals.

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

4.3 Case with a common regressor

We now examine the performances of our inference method in the presence of a common regressor. Namely, we consider the data-generating process

$$Y = \gamma_0 \mathbb{1}\{X_c = 1\} + X_{nc}\beta_0 + U, \ X \perp\!\!\!\perp U, \ U \sim \mathcal{N}(0, 4).$$

Where 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 3 below the performances of our inference method applied to the parameter β_0 along with the identified sets. With the exception of the smallest sample size ($n = 400$), for which we obtain a coverage rate of 93.4%, coverage ranges between 95% and 96.4% for sample sizes $n \geq 800$. Besides, and similar to the baseline case without common regressor, 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.

Average	Bounds	Excess length	Coverage
Identified set	[-2.121,2.121]		
Sample size			
400	[-2.357,2.358]	0.473	0.934
800	[-2.298,2.3]	0.356	0.95
1,200	[-2.278,2.278]	0.315	0.958
2,400	[-2.237,2.237]	0.233	0.962
4,800	[-2.206,2.208]	0.172	0.964

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 $\beta \in \mathcal{B}$, of the estimated probability that $\beta \in \text{CR}_{1-\alpha}(\beta_0)$. We use 1,000 bootstrap replications to compute the confidence intervals.

Table 3: Finite sample performances for β_0 with one common regressor

4.4 Computational time

We examine the computational time of our 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 4.1 and 4.2, respectively, and for the five different sample sizes considered above. Table

4 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 $p = 2$.[9]

In the univariate case ($p = 1$), the computational gains of our method range from a factor of 185 to 282 compared to AS, for $n = 4,800$ and $n = 2,400$, respectively. While the computational time associated with our method increases with the sample size, it remains modest (around 6 seconds) 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 $1/\inf_{q\in\mathbb{R}^p:q_k=1} 1/\widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q})$), 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 6.5 minutes 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 200 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 bootstrapping of the convex optimization in (19).

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

[9] 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.

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)	1.2	1.5	1.9	2.9	6.1
$p = 2$					
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 (min)	1.6	1.9	2.3	3.5	6.5

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 of 3.2%, for $n = 1, 200$. As in Sections 4.1-4.2, we use 1,000 bootstrap replications when $p = 1$ and 200 replications when $p = 2$ for both DGM and AS methods. The CPU times are obtained using our companion R package, on an Intel Xeon Gold 6130 CPU 2.10GHz with 382Gb of RAM.

Table 4: CPU time as function of sample size and dimension p of X_{nc}.

5 Application to intergenerational mobility in the United States, 1850–1930

We 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).[10]

[10]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.

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), which, beyond this paper, is very frequently used in the presence of data combination.[11] For the periods 1860-1880 and 1880-1900 only, the IPUMS Linked Representative Samples links fathers and sons using information on first and last names, which allows us to estimate more directly the father-son elasticity using OLS.

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 dummies 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 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 that is associated

[11] 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.

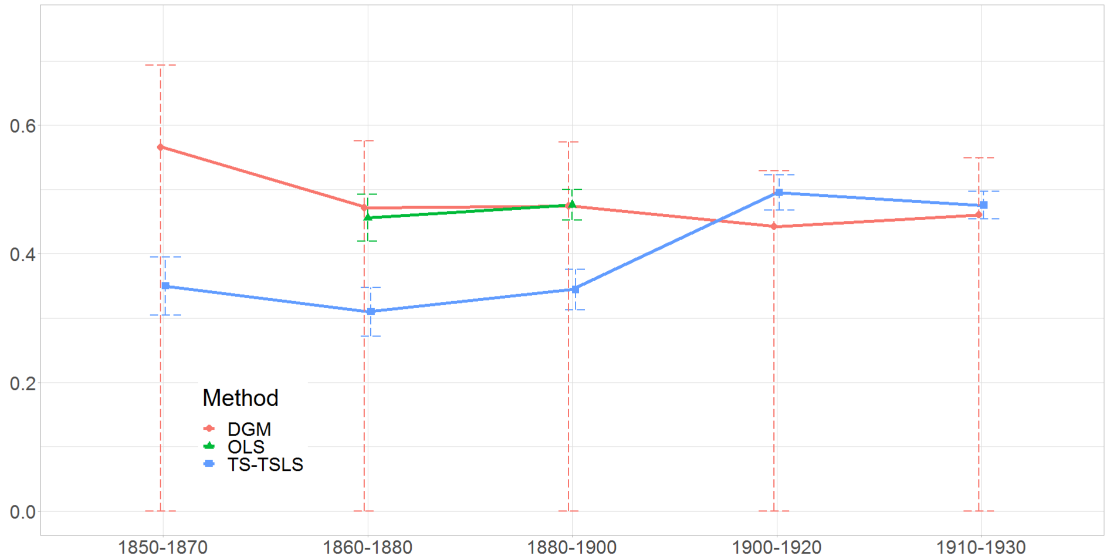
with a much larger number of observations (with $n > 100,000$ for both samples of Y and X_{nc}). Nonetheless, our inference procedure remains tractable in this case too, with a computational time of about 14 minutes.[12] Overall, this illustrates the applicability of our method, which can be easily implemented even in this type of rich and fairly high-dimensional data environment.

Figures 3(a)-3(b) and Table 5 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.[13] Specifically, we report in Figures 3(a)-3(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, for 1860-1880 and 1880-1900 and the sample of sons only, the OLS estimates and confidence intervals (solid and dashed green, resp.).

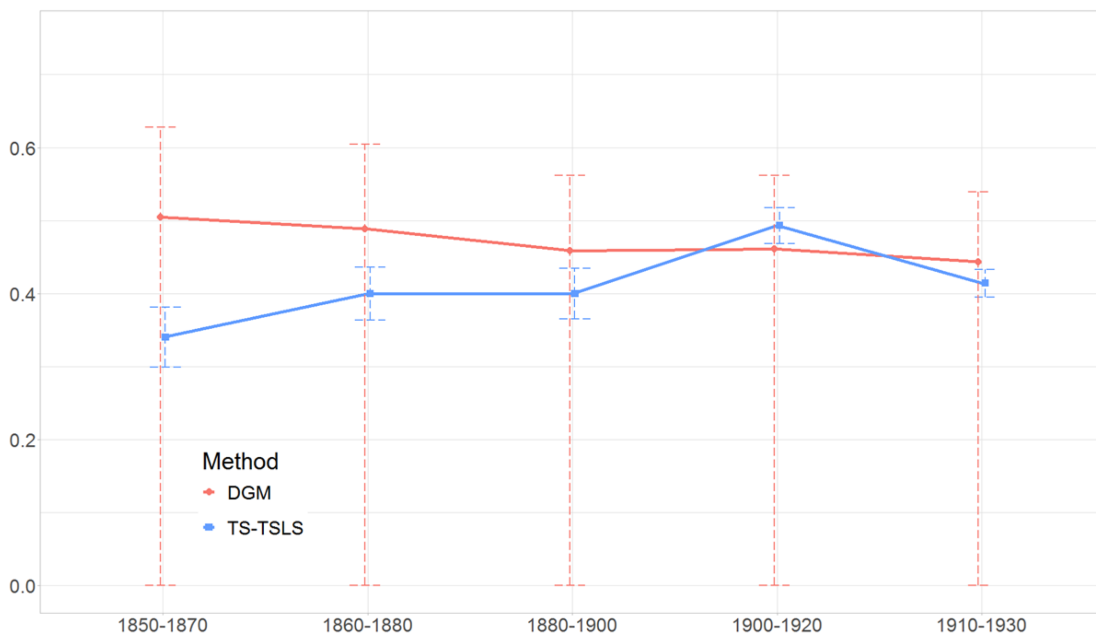
A first conclusion from these results is that the upper bounds of the confidence intervals associated with our method range, depending on the periods, between 0.53 and 0.69 (0.54 and 0.63) for the sample of sons (sons-in-law). These values of the intergenerational correlation coefficient are all well below the natural upper bound of 1. Also, even if the estimates vary depending on the data and econometric specification 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 indicates that our method does lead to informative inference on the parameter of interest.

[12]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.

[13]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 below the robustness of our results to alternative cutoffs.



(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 are negative (see Table 5).

Figure 3: Intergenerational income correlation using either TSTSLS, OLS (on matched data only when available), or our method (DGM).

Sample: **Sons**	**1850-1870**	**1860-1880**	**1880-1900**	**1900-1920**	**1910-1930**
DGM, pt.	[-0.501,0.565]	[-0.399,0.472]	[-0.491,0.474]	[-0.405,0.442]	[-0.388,0.46]
DGM, CI.	[-0.607,0.693]	[-0.481,0.576]	[-0.588,0.573]	[-0.479,0.529]	[-0.457,0.549]
TSTSLS, pt.	0.3403	0.4001	0.3998	0.4931	0.4143
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	0.000	0.029	0.167	0.971	0.797
(Stat.; critical val. 95%)	(29.96; 22.04)	(26.40; 24.88)	(26.17; 29.52)	(12.14; 37.23)	(16.71; 32.53)
Number of names X_c	224	259	380	512	596
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, pt.	[-0.436,0.504]	[-0.367,0.489]	[-0.287,0.459]	[-0.421,0.461]	[-0.384,0.443]
DGM, CI.	[-0.539,0.628]	[-0.451,0.605]	[-0.35,0.562]	[-0.504,0.561]	[-0.453,0.539]
TSTSLS, pt.	0.3403	0.4001	0.3998	0.4931	0.4143
TSTSLS, CI.	[0.2995,0.3811]	[0.3638,0.4364]	[0.365,0.4345]	[0.4687,0.5176]	[0.3954,0.4331]
Test of equality, p-value	0.012	0.126	0.645	0.989	0.951
(Stat.; critical val. 95%)	(19.95; 17.17)	(25.53; 28.04)	(22.64; 31.87)	(6.65; 28.90)	(6.54; 21.77)
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, pt." 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.

Table 5: Intergenerational income correlation for sons using either TSTSLS or our method (DGM).

Second, consider the two cases where the linked data is available (1860-1880 and 1880-1900 for the sample of sons). 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 3 in Section 2.1.2, 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 $\overline{\theta}_0$.[14] In other words, the results from these two periods support

[14]Proposition 3 is obtained without X_c. Yet, it can be combined with Proposition 4 to show that

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. If so, 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 that our estimates point to a mild decrease, both for sons and sons-in-law, in this elasticity coefficient over 1850-1930.

Third, the TSTSLS estimates are included in the confidence intervals associated with our method, for all five periods and for both of the samples. In addition, the results from the test of equality reported in Table 5 indicate that the TSTSLS estimates are in most cases not statistically distinguishable from the estimated upper bounds of our identified sets. The first two periods, 1850-1870 (for both samples) and 1860-1880 (for the sample of sons) are, however, notable exceptions. Besides, for the sample of sons, 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.

Finally, we provide in Tables 6 and 7 in Appendix B several robustness checks that relate to the set of first names we are including as controls in our estimation procedure (Panel A), the choice of the parameter ε (Panel B), the use of subsampling rather than numerical bootstrap (Panel C), and restriction of the sample to the set of individuals whose first name is included in the set of controls X_c (Panel D). Throughout these tables we focus on the upper bound of the estimated identified set ("DGM, pt") and of the confidence interval ("DGM, CI.").

The main takeaway from Table 6 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 0.53 to 0.61, the results remaining, however, stable for the other periods. The results for the sample of sons-in-law (Table 7) are also, for most periods, qualitatively, and in

β_0, and in turn γ_0 (and thus θ_0 here) are point identified with such X_c.

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.

6 Conclusion

We study the identification of and inference on partially linear models, in a data combination 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 frequently in economics, including in situations where 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 than the main outcome and regressor of interest. Focusing on the coefficients of the regressors that are not jointly observed with the outcome, we use recent insights from optimal transport to derive a constructive characterization of the sharp identified set. We then build on this result and develop a novel inference method that exploits the geometric properties of the identified set. The resulting procedure is very tractable, and as a result can be widely applied. We establish the asymptotic validity of the confidence region and show that our inference method exhibits good finite sample properties.

We apply our method to study intergenerational mobility over the period 1850 to 1930 in the United States, revisiting the analysis of Olivetti and Paserman (2015) on this question. Our method allows us to relax the exclusion restrictions underlying the TSTSLS approach implemented in Olivetti and Paserman (2015). Overall our confidence sets are informative, excluding for some periods values of the correlation coefficients larger than .55. While the TSTSLS estimates are included in our confidence intervals, our results suggest 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.

References

Altonji, J. G., Elder, T. E. and Taber, C. R. (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 Shi, X. (2017), 'Inference based on many conditional moment inequalities', *Journal of Econometrics* **196**(2), 275–287.

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

Backhoff-Veraguas, J., Beiglböck, M. and Pammer, G. (2019), 'Existence, duality, and cyclical monotonicity for weak transport costs', *Calculus of Variations and Partial Differential Equations* **58**(6), 203.

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

Bhattacharya, D. (2013), 'Evaluating treatment protocols using data combination', *Journal of Econometrics* **173**, 160–174.

Bontemps, C. and Magnac, T. (2017), 'Set identification, moment restrictions, and inference', *Annual Review of Economics* **9**, 103–129.

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

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

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

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

Davillas, A. and Pudney, S. (2020), 'Using biomarkers to predict healthcare costs: Evidence from a UK household panel', *Journal of Health Economics* **73**, 1–13.

Davydov, Y. A., Lifshits, M. A. et al. (1998), *Local properties of distributions of stochastic functionals*, American Mathematical Society.

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

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

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

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

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

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

Fang, Z. and Santos, A. (2019), 'Inference on directionally differentiable functions', *Review of Economic Studies* **86**(1), 377–412.

Firpo, S., Galvao, A. F. and Parker, T. (2021), 'Uniform inference for value functions', *arXiv preprint arXiv:1911.10215* .

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

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

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

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

Hong, H. and Li, J. (2018), ‘The numerical delta method’, *Journal of Econometrics* **206**(2), 379–394.

Hong, H. and Li, J. (2020), ‘The numerical bootstrap’, *The Annals of Statistics* **48**(1), 397–412.

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’, Vol. 7 of *Handbook of Econometrics*, Elsevier, pp. 355–486.

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

Olivetti, C. and Paserman, M. D. (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.

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

Piatek, R. and Pinger, P. (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., Romano, J. P. and Wolf, M. (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 Moffitt, R. (2007), 'The econometrics of data combination', *Handbook of Econometrics* **6**, 5469–5547.

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

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

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

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

Van der Vaart, A. W. and Wellner, J. A. (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 Inference with weights and different sample sizes

We describe here how to handle weights and different sample sizes. Consider $(Y_1, \dots, Y_{n_y})$ and $(X_1, \dots, X_{n_x})$ two independent samples of i.i.d variables with associated weights $(W_1^y, \dots, W_{n_y}^y)$ and $(W_1^x, \dots, W_{n_x}^x)$, which can represent either sampling or bootstrap weights. Let us denote by F_{Y,W^y} and F_{X,W^x} (resp. $\widehat{F}_{Y,W^y}$ and $\widehat{F}_{X,W^x}$) the weighted cdf (resp. weighted empirical cdf) of Y and X. Let $\overline{Y}_w$ and $\overline{X'q}_w$ be the corresponding weighted sample averages. Let us consider

$$I_{W,n} := \left\{ \sum_{i=1}^{j} W_i^y : \quad j = 1, \dots, n_y \right\} \cup \left\{ \sum_{i=1}^{j} W_i^x : \; j = 1, \dots, n_x \right\}$$

and $I_{W,n,\varepsilon} := \{\alpha \in I_{W,n} \cap [\epsilon, 1-\epsilon]\}$. Then, we define

$$\widehat{R}_W\left(\alpha, F_{Y,W}, F_{X'q,W^x}\right) = \frac{\sum_{i=1}^{n_y} W_i^y \left(Y_i - \overline{Y}_w\right) \mathbb{1}\left\{\widehat{F}_{Y,W^y}(Y_i) > \alpha\right\}}{\sum_{i=1}^{n_x} W_i^x \left(X_i'q - \overline{X'q}_w\right) \mathbb{1}\left\{\widehat{F}_{X'q,W^x}(X_i'q) > \alpha\right\}},$$

$$\widehat{S}_{\varepsilon,W}(F_{Y,W^y}, F_{X'q,W^x}) = \min_{\alpha \in I_{W,n,\varepsilon}} \widehat{R}_W\left(\alpha, F_{Y,W}, F_{X'q,W^x}\right).$$

Consider $Y_{(1)} < \dots < Y_{(n_y)}$ and $(X'q)_{(1)} < \dots < (X'q)_{(n_x)}$ and let the associated weights be $(W_{(1)}^y, \dots, W_{(n_y)}^y)$ and $(W_{(1)}^x, \dots, W_{(n_x)}^x)$. We also have

$$\widehat{S}_{\varepsilon,W}(F_{Y,W^y}, F_{X'q,W^x}) = \min_{\alpha \in I_{W,n,\varepsilon}} \frac{\sum_{i:\ \sum_k^i W_{(k)}^y \le \alpha} W_{(i)}^y \left(Y_{(i)} - \overline{Y}_w\right)}{\sum_{i:\ \sum_k^i W_{(k)}^x \le \alpha} W_{(i)}^x \left((X'q)_{(i)} - \overline{X'q}_w\right)}.$$

Hence, $\widehat{S}_{\varepsilon,W}(F_{Y,W^y}, F_{X'q,W^x})$ can still be computed quickly in this case.

B Additional results on the application

Sample:	**1850-1870**	**1860-1880**	**1880-1900**	**1900-1920**	**1910-1930**
Baseline specification					
DGM, pt.	0.565	0.472	0.474	0.442	0.46
DGM, CI.	0.693	0.576	0.573	0.529	0.549
Number of names X_c	224	259	380	512	596
Panel A: Robustness to the set of first names					
Threshold 0.005%					
DGM, pt.	0.565	0.472	0.474	0.442	0.46
DGM, CI.	0.693	0.576	0.574	0.529	0.549
Number of names X_c	224	259	380	513	624
Threshold 0.02%					
DGM, pt.	0.565	0.472	0.492	0.507	0.46
DGM, CI.	0.693	0.575	0.595	0.61	0.549
Number of names X_c	224	259	330	376	415
Panel B: Robustness to the choice of ε					
$\varepsilon/2$					
DGM, pt.	0.559	0.448	0.474	0.442	0.45
DGM, CI.	0.684	0.547	0.572	0.529	0.537
Number of names X_c	224	259	380	512	596
2ε					
DGM, pt.	0.565	0.501	0.474	0.442	0.463
DGM, CI.	0.693	0.609	0.574	0.529	0.554
Number of names X_c	224	259	380	512	596
Panel C: Using subsampling					
DGM, pt.	0.565	0.472	0.474	0.442	0.46
DGM, CI.	0.704	0.583	0.581	0.536	0.557
Number of names X_c	224	259	380	512	596
Panel D: Restricting the sample to the selected first names					
DGM, pt.	0.566	0.472	0.472	0.436	0.451
DGM, CI.	0.695	0.578	0.571	0.517	0.533
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 (24). Panel B presents the results when using 0.5 or 2 times this choice of ε. Panel C presents results with the subsampling instead of the numerical bootstrap, and the subsample size (calibrated on the simulations of Section 4) $b_n = 0.25n - 0.05\max(n-800,0) - 0.05\max(n-1200,0) - 0.05\max(n-1800,0) - 0.05\max(n-2000,0) - 0.05(1-\log(2400)/\log(n))\max(n-2400,0)$, with $n = n_y n_x/(n_y+n_x)$, n_y and n_x being the respective sample sizes of Y and X_{nc}. Panel D 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 6: Robustness checks for the upper bound on intergenerational income correlation for sons using our method (DGM).

Sample:	**1850-1870**	**1860-1880**	**1880-1900**	**1900-1920**	**1910-1930**
Baseline specification					
DGM, pt.	0.504	0.489	0.459	0.461	0.443
DGM, CI.	0.628	0.605	0.562	0.561	0.539
Number of names X_c	155	212	323	468	545
Panel A: Robustness to the set of first names					
Threshold 0.02%					
DGM, pt.	0.504	0.489	0.459	0.583	0.442
DGM, CI.	0.628	0.605	0.562	0.711	0.537
Number of names X_c	155	212	316	397	430
Panel B: Robustness to the choice of ε					
$\varepsilon/2$					
DGM, pt.	0.464	0.382	0.394	0.455	0.415
DGM, CI.	0.574	0.47	0.48	0.55	0.504
Number of names X_c	155	212	323	468	545
2ε					
DGM, pt.	0.504	0.489	0.459	0.461	0.454
DGM, CI.	0.627	0.605	0.563	0.561	0.553
Number of names X_c	155	212	323	468	545
Panel C: Using subsampling					
DGM, pt.	0.504	0.489	0.459	0.461	0.443
DGM, CI.	0.637	0.61	0.566	0.567	0.544
Number of names X_c	224	259	380	512	596
Panel D: Restricting the sample to the selected first names					
DGM, pt.	0.502	0.496	0.459	0.456	0.437
DGM, CI..	0.626	0.617	0.561	0.555	0.531
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 (24). Panel B presents the results when using 0.5 or 2 times this choice of ε. Panel C presents results with the subsampling instead of the numerical bootstrap, and the subsample size (calibrated on the simulations of Section 4) $b_n = 0.25n - 0.05\max(n-800,0) - 0.05\max(n-1200,0) - 0.05\max(n-1800,0) - 0.05\max(n-2000,0) - 0.05(1-\log(2400)/\log(n))\max(n-2400,0)$, with $n = n_y n_x/(n_y+n_x)$, n_y and n_x being the respective sample sizes of Y and X_{nc}. Panel D 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 7: Robustness checks for the upper bound on intergenerational income correlation for sons-in-law using our method (DGM).

C Proofs

C.1 Notation

We let $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote respectively the usual Euclidean norm in $\mathbb{R}^p$ and the supremum norm. We denote by $\mathcal{P}(\mathbb{R}^p)$ the set of Borel probability measures on $\mathbb{R}^p$ and by $\mathcal{P}_q(\mathbb{R}^p)$ the subset of $\mathcal{P}(\mathbb{R}^p)$ whose elements have q finite absolute moments, namely $\mathcal{P}_q(\mathbb{R}^p) = \{\mu \in \mathcal{P}(\mathbb{R}^p),\ \int_{\mathbb{R}^p} |x|^q \mu(dx) < \infty\}$. We assimilate herafter probability measures on $\mathbb{R}^p$ with their cdf, so we may write for instance $F \in \mathcal{P}_q(\mathbb{R}^p)$. Finally, we let W_1 denote the 1-Wasserstein distance and recall that for $(F, G) \in \mathcal{P}_1(\mathbb{R})^2$,

$$W_1(F,G) = \inf_{U \sim F, V \sim G} E\left[|U - V|\right] = \int_0^1 |F^{-1}(t) - G^{-1}(t)| dt = \int_{-\infty}^{\infty} |F(t) - G(t)| dt. \tag{26}$$

Denote also by $\ell^\infty(\mathcal{X})$ the space of bounded functions on $\mathcal{X}$ for the uniform metric, by $\ell^\infty_+(\mathcal{X})$ the space of positive such functions.

C.2 Theorem 1

Let $\mathcal{B}'$ denote the set on the right-hand side of (3). 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 $g(\alpha) = \int_\alpha^1 F^{-1}(t)dt$. 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 $F^{-1}(\alpha) = 0$ for all $\alpha \in (0,1)$, 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,

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

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 $(F, G, H) \in \mathcal{P}$, let us define

$$\begin{aligned} W_w(F,G) &:= \inf_{F_{U,V}: F_U = F, F_V = G} E\left[|V - E[U|V]|\right], \\ W_c(F,G) &:= \inf_{F_{U,V,W}:\; F_U = F,\; F_{V,W} = G} E\left[|V - E(U|V,W)|\right]. \end{aligned}$$

Now, let $\beta \in \mathcal{B}'$. By Strassen's theorem (Theorem 8 in Strassen, 1965), we have $W_w(F_Y, F_{X'\beta}) = 0$. Then, given the definition of $\mathcal{B}$, it suffices to prove that $W_c(F,G) \leq W_w(F, G_1)$, where G_1 is the first marginal distribution of G.

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 c satisfies the assumptions of Theorem 1.3. in Backhoff-Veraguas et al. (2019), we have

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

where $\mathcal{F}$ is the space of continuous and bounded real functions on $\mathbb{R}$ and

$$R_c(f)(x_1,x_2) = \inf_{H \in \mathcal{P}_1(\mathbb{R})} \int f(y) dH(y) + \left| x_1 - \int y dH(y) \right|.$$

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,

$$\int R_c(f)(x_1,x_2) dG(x_1,x_2) - \int f(y) dF(y)$$

$$\leq E[f(U)] + E\left[|V_1 - E(U|V_1)|\right] - \int f(y)dF(y)$$
$$= E\left[|V_1 - E(U|V_1)|\right].$$

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 $f \in \mathcal{F}$, we obtain $W_c(F, G) \leq W_w(F, G_1)$. Because c satisfies the assumptions of Theorem 1.2 in Backhoff-Veraguas et al. (2019), there exists a minimizer of the problem $W_c(F, G)$. 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 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,

$$\begin{aligned} E\left[\phi(X_0'\beta)\right] &= E\left[\liminf_n \phi(X_0'\beta_n)\right] \\ &\leq \liminf_n E\left[\phi(X_0'\beta_n)\right] \\ &\leq E\left[\phi(Y_0)\right]. \end{aligned}$$

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

C.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 $\sigma(-e_k, F_{Y_0}, F_{X_0})$.

C.4 Proposition 1

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

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

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

$$\int_\alpha^1 F^{-1}_{\lambda U+(1-\lambda)V}(t)dt \leq \int_\alpha^1 F^{-1}_{\lambda U}(t)dt + \int_\alpha^1 F^{-1}_{(1-\lambda)V}(t)dt$$

$$= \lambda \int_\alpha^1 F_U^{-1}(t)dt + (1-\lambda)\int_\alpha^1 F_V^{-1}(t)dt,$$

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 $q \mapsto \int_\alpha^1 F_{X_0'q}^{-1}(t)dt$ 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 $\mathcal{B}_\varepsilon = \{q \in \mathbb{R}^p : p_\varepsilon(q) \le 1\}$ 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, $S_\varepsilon(F,G) \le S_{\varepsilon'}(F,G)$ for any $0 < \varepsilon < \varepsilon' < 1/2$. Now,

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

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, $\inf_{\varepsilon\in(0,1/2)} S_\varepsilon(F,G) \ge S(F,G)$ since $S_\varepsilon(F,G) \ge S(F,G)$ 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

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

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

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

The function $R(\cdot, F, G)$ is differentiable on $(0,1)$, with

$$\frac{\partial R}{\partial \alpha}(\alpha, F_{Y_0}, F_{X_0'q}) = \frac{-F_{Y_0}^{-1}(\alpha)\int_\alpha^1 F_{X_0'q}^{-1}(t)dt + F_{X_0'q}^{-1}(\alpha)\int_\alpha^1 F_{Y_0}^{-1}(t)dt}{\left(\int_\alpha^1 F_{X_0'q}^{-1}(t)dt\right)^2}. \tag{27}$$

Assume that $\alpha \mapsto F_{Y_0}^{-1}(\alpha)/F_{X_0'q}^{-1}(\alpha)$ is increasing on $[\overline{\alpha}, 1]$ and suppose without loss of generality that $\overline{\alpha}$ is such that $F_{Y_0}^{-1}(\overline{\alpha}) \wedge F_{X_0'q}^{-1}(\overline{\alpha}) > 0$. Fix $\alpha \in [\overline{\alpha}, 1)$. By the mean value theorem, there exists $\alpha' \in [\alpha, 1)$ such that

$$\int_\alpha^1 F_{Y_0}^{-1}(t)dt = \int_\alpha^1 \left[\frac{F_{Y_0}^{-1}(t)}{F_{X_0'q}^{-1}(t)}\right] F_{X_0'q}^{-1}(t)dt = \frac{F_{Y_0}^{-1}(\alpha')}{F_{X_0'q}^{-1}(\alpha')} \int_\alpha^1 F_{X_0'q}^{-1}(t)dt.$$

Thus, because $\alpha \mapsto F_{Y_0}^{-1}(\alpha)/F_{X_0'q}^{-1}(\alpha)$ is increasing on $[\overline{\alpha}, 1]$,

$$F_{X_0'q}^{-1}(\alpha) \int_\alpha^1 F_{Y_0}^{-1}(t)dt \geq F_{Y_0}^{-1}(\alpha) \int_\alpha^1 F_{X_0'q}^{-1}(t)dt.$$

In view of (27), this ensures that $\partial R/\partial\alpha(\alpha, F_{Y_0}, F_{X_0'q}) \geq 0$. Hence,

$$\inf_{\alpha \in [\overline{\alpha}, 1)} R(\alpha, F_{Y_0}, F_{X_0'q}) = R\left(\overline{\alpha}, F_{Y_0}, F_{X_0'q}\right).$$

Using a similar argument on $[0, \underline{\alpha}]$, we finally obtain

$$\inf_{\alpha \in (0,1)} R\left(\alpha, F_{Y_0}, F_{X_0'q}\right) = \min_{\alpha \in [\underline{\alpha}, \overline{\alpha}]} R\left(\alpha, F_{Y_0}, F_{X_0'q}\right).$$

The result follows, with $\varepsilon_0 = \min(\underline{\alpha}, 1 - \overline{\alpha})$.

C.5 Proposition 2

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, $\xi_Y \succ_{\text{cv}} \xi_X'\beta$. Thus,

$$X^{*\prime}\beta + \xi_Y \succ_{\text{cv}} X^{*\prime}\beta + \xi_X'\beta = X'\beta. \tag{28}$$

Now, because $\beta \in \mathcal{B}^*$, we also have, by Theorem 1, $Y^* \succ_{\text{cv}} X^{*\prime}\beta$. Hence, by independence, $Y^* + \xi_Y \succ_{\text{cv}} X^{*\prime}\beta + \xi_Y$. Combined with (28), this yields $Y \succ_{\text{cv}} X'\beta$. Hence, $\beta \in \mathcal{B}$ and $\mathcal{B}^* \subset \mathcal{B}$.

C.6 Proposition 3

By the proof of Theorem 1, we have

$$\mathcal{B} = \{\beta \in \mathbb{R}^p : Y \succ_{\text{cv}} X'\beta\}. \tag{29}$$

The second result follows. To prove the first result, let $\lambda > 0$. Because $U \not\geq_{\text{cv}} (X'\beta_0)\lambda$, there exists $p_0 \in (0,1)$ such that

$$\int_{p_0}^1 \left(F_{U_0}^{-1}(t) - F_{(X_0'\beta_0)\lambda}^{-1}(t)\right) dt < 0. \tag{30}$$

The subadditivity of the superquantiles (see, e.g. Embrechts and Wang, 2015) yields that, for all $p \in (0,1)$,

$$\int_p^1 F_{Y_0}^{-1}(t)dt \leq \int_p^1 F_{X_0'\beta_0}^{-1}(t)dt + \int_p^1 F_{U_0}^{-1}(t)dt.$$

Thus, we have

$$\begin{aligned}\int_{p_0}^1 \left(F_{Y_0}^{-1}(t) - F_{X_0'\beta_0(1+\lambda)}^{-1}(t)\right) dt &= \int_{p_0}^1 \left(F_{Y_0}^{-1}(t) - F_{X_0'\beta_0}^{-1}(t) - F_{(X_0'\beta_0)\lambda}^{-1}(t)\right) dt \\ &\leq \int_{p_0}^1 \left(F_{U_0}^{-1}(t) - F_{X_0'\beta_0\lambda}^{-1}(t)\right) dt < 0.\end{aligned}$$

This shows that, for all $\lambda > 0$, $Y \not\geq_{\text{cv}} X'\beta_0(1+\lambda)$. Then, in view of (29), $\beta_0(1+\lambda) \notin \mathcal{B}$. The first result follows.

C.7 Proof of Lemma 1

Fix $\beta \neq 0_p$ and let $\beta = \lambda\gamma$ with $\lambda > 0$ and $\gamma \in \mathcal{S}$. Then, let $\phi_\lambda(x) = \phi_1(x/\lambda)$. We have $E[\phi_\lambda(|T_0'\beta|)] = E[\phi_1(|T_0'\gamma|)] = \infty$. On the other hand, by condition (ii), there exists $c_0 > 0$ such that for all $x \geq c_0$, $x/\lambda \leq \phi_2(x)$. Then, since ϕ_1 is increasing,

$$\begin{aligned}E[\phi_\lambda(|S_0|)] &= E[\phi_\lambda(|S_0|)\mathbb{1}\{|S_0| \leq c_0\}] + E[\phi_\lambda(|S_0|)\mathbb{1}\{|S_0| \geq c_0\}] \\ &\leq \phi_1(c_0/\lambda) + E[\phi_1 \circ \phi_2(|S_0|)] \\ &< \infty.\end{aligned}$$

Thus, $E[\phi_\lambda(|S_0|)] < E[\phi_\lambda(|T_0'\beta|)]$. Since $x \mapsto \phi_\lambda(|x|)$ is convex, we get $S \not\geq_{\text{cv}} T'\beta$.

C.8 Theorem 2

Recall that $\widehat{F}_{Y_0}(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\left\{Y_i - \overline{Y} \leq t\right\}$ and $\widehat{F}_{X_0'q}(t)$ is defined similarly. The proof proceeds in three steps. First, we prove that for all $q \in \mathcal{S}$, $S_\varepsilon(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) \xrightarrow{\mathbb{P}} S_\varepsilon(F_{Y_0}, F_{X_0'q})$. Next, we prove that $\widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q}) \xrightarrow{\mathbb{P}} S_\varepsilon(F_{Y_0}, F_{X_0'q})$. Finally, 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 (26) 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. We first prove that 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

$$d'((\alpha, F, G), (\alpha', F', G')) = |\alpha' - \alpha| + W_1(F,F') + W_1(G,G'). \tag{31}$$

Remark that for all $a, a', b, b' > 0$, we have

$$\left|\frac{a'}{b'} - \frac{a}{b}\right| \leq \frac{1}{b}\left[|a'-a| + \left|\frac{a'}{b'} - \frac{a}{b}\right| |b'-b| + \frac{a}{b}|b'-b|\right]. \tag{32}$$

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 (32), we get

$$\begin{aligned}
|R(\alpha',F,G) - R(\alpha,F,G)| &\leq \frac{\left|\int_\alpha^{\alpha'} F^{-1}(t)dt\right| + R(\alpha,F,G)\left|\int_\alpha^{\alpha'} G^{-1}(t)dt\right|}{\int_\alpha^1 G^{-1}(t)dt - \left|\int_\alpha^{\alpha'} G^{-1}(t)dt\right|} \\
&\leq \frac{|\alpha'-\alpha|\left(|F^{-1}(\alpha)| \vee |F^{-1}(\alpha')| + R(\alpha,F,G)|G^{-1}(\alpha)| \vee |G^{-1}(\alpha')|\right)}{1/2\int_\alpha^1 G^{-1}(t)dt} \\
&\leq \frac{2|\alpha'-\alpha|\left(c_F + R(\alpha,F,G)c_G\right)}{\int_\alpha^1 G^{-1}(t)dt}. \qquad (33)
\end{aligned}$$

Next, for any F', using again (32),

$$\begin{aligned}
|R(\alpha',F',G') - R(\alpha',F,G)| &\leq \frac{\left|\int_{\alpha'}^1 F^{-1}(t) - F'^{-1}(t)dt\right| + R(\alpha',F,G)\left|\int_{\alpha'}^1 G^{-1}(t) - G'^{-1}(t)dt\right|}{\int_{\alpha'}^1 G^{-1}(t)dt - \left|\int_{\alpha'}^1 G'^{-1}(t) - G^{-1}(t)dt\right|} \\
&\leq \frac{W_1(F,F') + R(\alpha',F,G)W_1(G,G')}{1/4\int_\alpha^1 G^{-1}(t)dt} \\
&\leq \frac{4}{\int_\alpha^1 G^{-1}(t)dt}\left[W_1(F,F') + \left(\frac{2|\alpha'-\alpha|\left(c_F + R(\alpha,F,G)c_G\right)}{\int_\alpha^1 G^{-1}(t)dt}\right.\right. \\
&\qquad \left.\left. + R(\alpha,F,G)\right)W_1(G,G')\right]. \qquad (34)
\end{aligned}$$

Inequalities (33) and (34) and the triangle inequality imply that R is continuous for the metric d' defined by (31).

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 by Berge maximum theorem (see, e.g., Theorem 9.14 in Sundaram et al., 1996) that S_ε is also continuous with respect to the metric d. The result follows.

Step 2: $\widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q}) \xrightarrow{\mathbb{P}} S_\varepsilon(F_{Y_0}, F_{X_0'q})$ **for** $q \in \mathcal{S}$**.**

Let us define $I_{n,\varepsilon} = \{\alpha_{1,n}, ..., \alpha_{J_n,n}\} := \{[n\varepsilon]/n, ([n\varepsilon]+1)/n, ..., [n(1-\varepsilon)]/n\}$ and remark that

$$\widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q}) = \min_{\alpha \in I_{n,\varepsilon}} R\left(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right). \tag{35}$$

As a result,

$$\begin{aligned}
&\left|\widehat{S}_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right) - S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right)\right| \\
\leq & \max_{i=1,...,J_n-1} \sup_{\alpha \in [\alpha_{i,n}, \alpha_{i+1,n})} \left|R\left(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right) - R\left(\alpha_{i,n}, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right)\right| \\
\leq & \frac{2}{n} \max_{i=1,...,J_n-1} \frac{c_{\widehat{F}_{Y_0}} + R(\alpha_{i,n}, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) c_{\widehat{F}_{X_0'q}}}{\int_{\alpha_{n_i}}^1 \widehat{F}^{-1}_{X_0'q}(t)dt} \\
\leq & \frac{2}{n \int_{\alpha_{J_n-1,n}}^1 \widehat{F}^{-1}_{X_0'q}(t)dt} \left[c_{\widehat{F}_{Y_0}} + \left(\max_{\alpha \in [\varepsilon, 1-\varepsilon]} R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q})\right) c_{\widehat{F}_{X_0'q}}\right],
\end{aligned}$$

where we have used (33) in the second inequality. By definition of $\widehat{F}_{Y_0}$ and c_F,

$$c_{\widehat{F}_{Y_0}} = \left|\widehat{F}_Y^{-1}(\varepsilon) - \overline{Y}\right| \vee \left|\widehat{F}_Y^{-1}(1-\varepsilon) - \overline{Y}\right|.$$

By convergence of $\overline{Y}$ and empirical quantiles, $c_{\widehat{F}_{Y_0}} = O_P(1)$. Similarly, $c_{\widehat{F}_{X_0'q}} = O_P(1)$. Also,

$$\begin{aligned}
\left|\int_{\alpha_{J_n-1,n}}^1 \widehat{F}^{-1}_{X_0'q}(t)dt - \int_{1-\varepsilon}^1 F^{-1}_{X'q}(t)dt\right| &\leq \left|\int_{\alpha_{J_n-1,n}}^1 \widehat{F}^{-1}_{X_0'q}(t) - F^{-1}_{X'q}(t)dt\right| + \left|\int_{\alpha_{J_n-1,n}}^{1-\varepsilon} F^{-1}_{X'q}(t)dt\right| \\
&\leq W_1(\widehat{F}_{X_0'q}, F_{X'q}) + o(1),
\end{aligned}$$

where the last term tends to 0 since $\alpha_{J_n-1,n} \to 1-\varepsilon$. We proved in Step 1 that $W_1(\widehat{F}_{X_0'q}, F_{X'q}) = o_P(1)$. Since $\int_{1-\varepsilon}^1 F^{-1}_{X'q}(t)dt > 0$, we have $1/\int_{\alpha_{J_n-1,n}}^1 \widehat{F}^{-1}_{X_0'q}(t)dt = O_P(1)$. Finally, by the same reasoning as in Step 1,

$$\max_{\alpha \in [\varepsilon, 1-\varepsilon]} R(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}) \xrightarrow{\mathbb{P}} \max_{\alpha \in [\varepsilon, 1-\varepsilon]} R(\alpha, F_Y, F_{X'q}).$$

Hence, at the end of the day,

$$\widehat{S}_\varepsilon\left(F_{Y_0}, F_{X_0'q}\right) - S_\varepsilon\left(\widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right) = O_P\left(\frac{1}{n}\right).$$

The result follows by the first step.

Step 3: Convergence of the set $\widehat{\mathcal{B}}_\varepsilon$.

We showed in the proof of Proposition 1 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 $\widehat{p}_\varepsilon(q) = 1/\widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q})$. By Step 2 and the continuous mapping theorem, for all $q \in \mathcal{S}$,

$$\widehat{p}_\varepsilon(q) \xrightarrow{\mathbb{P}} p_\varepsilon(q).$$

Now, (35) implies that

$$\widehat{p}_\varepsilon(q) = \max_{\alpha \in I_{n,\varepsilon}} 1/R\left(\alpha, \widehat{F}_{Y_0}, \widehat{F}_{X_0'q}\right). \tag{36}$$

Note that for any (F_Y, F_X), $q \mapsto 1/R(F_{Y_0}, F_{X_0'q})$ is convex (see the proof of Point 1 in Proposition 1). Then, (36) implies that $\widehat{p}_\varepsilon$ is also convex. As a result, by the convexity lemma of Pollard (1991),

$$\sup_{q \in \mathcal{S}} |\widehat{p}_\varepsilon(q) - p_\varepsilon(q)| \xrightarrow{\mathbb{P}} 0. \tag{37}$$

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 p149 in Hiriart-Urruty and Lemaréchal, 2012). Thus, using Theorem 3.3.6 p155 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 (37), $d_H\left(\widehat{\mathcal{B}}_\varepsilon^\circ, \mathcal{B}_\varepsilon^\circ\right) \xrightarrow{\mathbb{P}} 0$. The result follows because convergence of polar sets for the Hausdorff distance implies convergence of the sets themselves, see Theorem 7.2 in Wijsman (1966).

C.9 Proof of Theorem 3

Before establishing the validity of the confidence regions and confidence intervals, we show the weak convergence of $\mathbb{F}_n := n^{1/2}\left(\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q}) - R(\alpha, F_{Y_0}, F_{X_0'q})\right)$, seen as a process indexed by either $\alpha \in [\varepsilon, 1-\varepsilon]$, for a fixed $q \in \mathcal{S}$, or $(q, \alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]$.

1. Weak convergence of $\mathbb{F}_n$

We first show the weak convergence of $\mathbb{F}_n$ to $\mathbb{F}$ as a process indexed by $(q, \alpha) \in \mathcal{S} \times [\varepsilon, 1-\varepsilon]$, under Assumption 4 . We then 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.

First, remark that $R(\alpha, F_{Y_0}, F_{X_0'q}) = \theta_1(q,\alpha)/\theta_2(q,\alpha)$, 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, by Assumption 3, F_{Y_0} and F_{X_0} are continuous. Then, using e.g. Lemma 21.1 in Van der Vaart (2000), we obtain

$$\begin{aligned}\theta_1(q,\alpha) =& E\left[(Y - E(Y))\mathbb{1}\left\{F_Y(Y) \geq \alpha\right\}\right],\\ \theta_2(q,\alpha) =& E\left[(X'q - E(X'q))\mathbb{1}\left\{F_{X'q}(X'q) \geq \alpha\right\}\right].\end{aligned}$$

Similarly, $\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q}) = \widehat{\theta}_1(q,\alpha)/\widehat{\theta}_2(q,\alpha)$, where

$$\widehat{\theta}_1(q,\alpha) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)\mathbb{1}\left\{\widehat{F}_Y(Y_i) > \alpha\right\}, \tag{38}$$

$$\widehat{\theta}_2(q,\alpha) = \frac{1}{n}\sum_{i=1}^n \left(X_i'q - \overline{X'q}\right)\mathbb{1}\left\{\widehat{F}_{X'q}(X_i'q) > \alpha\right\}. \tag{39}$$

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 1). $\theta_2(q,\cdot)$ 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 soon 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. To this end, remark first that

$$n^{1/2}\left(\widehat{\theta}_2(q,\alpha) - \theta_2(q,\alpha)\right) = \mathbb{G}_n g_{q,\alpha} + R_n(q,\alpha),$$

where $\mathbb{G}_n$ denotes the empirical process associated to $(X_1, ..., X_n)$ and

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

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 $R_n(q,\alpha)$. 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 $R_n(q,\alpha) = R_{1n} + R_{2n} + R_{3n}$, with

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

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

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} R_{kn}(q,\alpha) = o_p(1). \tag{40}$$

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

$$\left|\sum_{i=1}^{n}\left(\widehat{I}_i(q,\alpha)-\alpha\right)\right| = \lceil n\alpha\rceil - n\alpha < 1. \tag{41}$$

As a result,

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

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, (40) holds for $k=2$.

Next, consider R_{3n}. We have

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

The first term is an $O_p(1)$. Recall that the class $\mathcal{I}_1$ is is Donsker; hence it is also Glivenko-Cantelli. Therefore, the second term is an $o_p(1)$. Therefore, (40) holds for $k=3$.

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

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

That R_{11n} 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} and using (41). For R_{12n}, remark that by definition of $I_i(q,\alpha)$ and $\widehat{I}_i(q,\alpha)$,

$$|R_{12n}| \leq \left|\widehat{F}^{-1}_{X'q}(\alpha)-F^{-1}_{X'q}(\alpha)\right| \times \left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}(I_i(q,\alpha)-\widehat{I}_i(q,\alpha))\right|.$$

By (41) and the fact that $\mathcal{I}_1$ is a Donsker class,

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

Thus, the result holds as long as

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} \left|\widehat{F}^{-1}_{X'q}(\alpha) - F^{-1}_{X'q}(\alpha)\right| = o_p(1). \tag{42}$$

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,

$$\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} \left|F_{X'q}(\alpha) - \widehat{F}_{X'q}(\alpha)\right| = o_p(1). \tag{43}$$

Now, let $U_q = F_{X'q}(X'q)$ and $U_{q,1} < ... < U_{q,n}$ denote the corresponding order statistic. Remark that $\widehat{F}^{-1}_{X'q}(\alpha) = F^{-1}_{X'q}(U_{q,\lceil n\alpha\rceil})$. Also, note that $\inf_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} U_{q,\lceil n\alpha\rceil} < \varepsilon'$ implies that for some $q_0 \in \mathcal{S}$, $\widehat{F}_{X'q_0}(F^{-1}_{X'q_0}(\varepsilon')) \geq \lceil n\alpha\rceil/n$ 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 (43), this occurs with probability approaching zero. The same is true for the event $\sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} U_{q,\lceil n\alpha\rceil} > 1-\varepsilon'$. Hence, with probability approaching one,

$$\varepsilon' \leq \inf_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} U_{q,\lceil n\alpha\rceil} \leq \sup_{(q,\alpha)\in\mathcal{S}\times[\varepsilon,1-\varepsilon]} U_{q,\lceil n\alpha\rceil} \leq 1-\varepsilon'. \tag{44}$$

Moreover, under this event,

$$\begin{aligned}
\left|\widehat{F}^{-1}_{X'q}(\alpha) - F^{-1}_{X'q}(\alpha)\right| =& \left|F^{-1}_{X'q}(U_{q,\lceil n\alpha\rceil}) - F^{-1}_{X'q}(\alpha)\right| \\
<& m\left(|U_{q,\lceil n\alpha\rceil} - \alpha|\right) \\
\leq& m\Big(|F_{X'q}((X'q)_{\lceil n\alpha\rceil}) - \widehat{F}_{X'q}((X'q)_{\lceil n\alpha\rceil})| \\
&\quad + \left|\widehat{F}_{X'q}((X'q)_{\lceil n\alpha\rceil}) - \alpha\right|\Big) \\
<& m\left(\sup_{q\in\mathcal{S}} \|F_{X'q} - \widehat{F}_{X'q}\|_\infty + \frac{1}{n}\right).
\end{aligned}$$

Using (43) and the continuity of m finally yields (42).

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 (42). Now, given that q is fixed, we only need to establish the weaker

$$\sup_{\alpha\in[\varepsilon,1-\varepsilon]} \left|\widehat{F}^{-1}_{X'q}(\alpha) - F^{-1}_{X'q}(\alpha)\right| = o_p(1). \tag{45}$$

Because $F_{X'q}^{-1}$ 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

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

Moreover, $\sup_{\alpha\in[\varepsilon,1-\varepsilon]}\left|U_{q,\lceil n\alpha\rceil} - \alpha\right| = o_p(1)$. This implies that (45) holds.

2. Asymptotic validity of the confidence region

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

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

Moreover, by Theorem 2.1 of Cárcamo et al. (2020), ι is Hadamard directionally differentiable. Then, by the functional delta method for Hadamard directionally differentiable functions (see, e.g., Proposition 2.1 in Cárcamo et al., 2020), we have

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

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

Note that for all (G_1, G_2) in $\ell^\infty([\varepsilon, 1-\varepsilon])$, $|\iota(G_1) - \iota(G_2)| \leq \|G_1 - G_2\|_\infty$. Thus, ι is Lipschitz as a function from $\ell^\infty([\varepsilon, 1-\varepsilon])$ to $\mathbb{R}$. Then, by Theorem 3.2 in Hong and Li (2018), the numerical derivative $\iota'_{G,n} : h \mapsto (\iota(G + \delta_n h) - \iota(G))/\delta_n$ is also Lipschitz, uniformly in δ_n. Then, by Lemma 3.1 and Theorem 3.3 in Hong and Li (2018), Assumption 4 in Fang and Santos (2019) holds. By Theorem 3.2 in Fang and Santos (2019), the numerical bootstrap based on (16) is valid: conditional on the data and with probability approaching one,

$$\frac{\min_{\alpha\in[\varepsilon,1-\varepsilon]}(\widehat{R}(\alpha, F_{Y_0}, F_{X_0'q}) + \delta_n\mathbb{Z}^*_{n,q}) - \widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q})}{\delta_n} \xrightarrow{d} H,$$

where H is the cdf of $\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}(\mathbb{F})$.

Now, let us show that $\widehat{c}_{\alpha,\varepsilon} \xrightarrow{\mathbb{P}} c_{\alpha,\varepsilon}$. 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 $-\iota'_{R(\cdot,F_{Y_0},F_{X_0'q})}(\mathbb{F})$, it follows from Lemma 21.2 in Van der Vaart (2000) 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, $\beta \in \text{CR}_{1-\alpha}(\beta_0)$ if and only if

$$n^{1/2}\left(\widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q}) - \lambda\right) - \widehat{c}_{\alpha,\varepsilon} \geq 0. \tag{46}$$

Suppose first that $\lambda < S_\varepsilon(F_{Y_0}, F_{X_0'q})$. Since $\widehat{S}_\varepsilon(F_{Y_0}, F_{X_0'q})$ is consistent for $S_\varepsilon(F_{Y_0}, F_{X_0'q})$ and $\widehat{c}_{\alpha,\varepsilon} = O_p(1)$, (46) holds with probability approaching one and $\liminf_{n\to\infty} P(\beta \in \text{CR}_{1-\alpha}(\beta_0)) = 1$. Now, suppose that $\lambda = S_\varepsilon(F_{Y_0}, F_{X_0'q})$. Then, by what precedes,

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

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, $\liminf_{n\to\infty} P(\beta \in \text{CR}_{1-\alpha}(\beta_0)) = 1 - \alpha$. The result follows.

3. Asymptotic validity of the confidence interval

First, remark that for any $e \in \mathbb{R}^p$,

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

Let us define $\kappa(G) := \sup_{q\in\mathcal{S}} \inf_{\alpha\in[\varepsilon,1-\varepsilon]} G(q,\alpha)$ and $G_e(q,\alpha) := R(\alpha, F_{Y_0}, F_{X_0'q}) \times q'e$. By Lemma A.1 in Firpo et al. (2021), κ is Hadamard directionally differentiable. Moreover, by Step 1, the process $(q,\alpha) \mapsto \mathbb{F}_n(q,\alpha) \times q'e$ converges weakly to a Gaussian process. Then, as above,

$$n^{1/2}\left(\widehat{\sigma}(e, F_{Y_0}, F_{X_0}) - \sigma(e, F_{Y_0}, F_{X_0})\right) \xrightarrow{d} \kappa'_{G_e}(\mathbb{F}),$$

where the expression of κ' is given by (3.10) in Firpo et al. (2021).

The proof of the validity of the numerical bootstrap based on (20) follows the same steps as for the validity of the numerical bootstrap for the confidence region, as the function $\kappa : \ell^\infty(\mathcal{S} \times [\varepsilon, 1-\varepsilon]) \to \mathbb{R}$ is also Lipschitz as a composition of two Lipschitz functions. Then, by continuity of the cdf of $\kappa'_{G_e}(\mathbb{F})$ at its quantile $c^s_{\alpha,\varepsilon}(e)$ (for $e = \pm e_k$), we have $\widetilde{c}_{\alpha,\varepsilon}(e) \xrightarrow{\mathbb{P}} c^s_{\alpha,\varepsilon}(e)$ for such e.

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

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

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

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

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

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

where we let $\overline{E}_n := \left\{ n^{1/2} \left(\widehat{\sigma}_\varepsilon(e_k, F_{Y_0}, F_{X_0}) - \sigma_\varepsilon(e_k, F_{Y_0}, F_{X_0}) \right) < \widetilde{c}_{\alpha,\varepsilon}(e_k) \right\}$. By what precedes, we have $\Pr(\underline{E}_n) \to \alpha$ and $\Pr(\overline{E}_n) \to \alpha$. The result follows.